Thesis Proposal

NLP in the Last Mile: Characterizing and Resisting Incentives Towards Generality

Sireesh Kaivar Gururaja

December 19, 2025

School of Computer Science
Department of Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:

Emma Strubell, Chair
Fernando Diaz
Sherry Tongshuang Wu
Lucy Suchman, *Lancaster University*

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technologies

Abstract

In the past decade, natural language processing (NLP) has gone from being applied to narrow, scoped problems in a handful of domains, to near ubiquity. This spread has been accompanied by broad promises of AI's applicability to nearly any problem, from discovering new materials to automating software engineering entirely. Yet amid persistent narratives of AI's general capability, users of such tools often find a jagged frontier: AI can be useful in some contexts, and fail in seemingly related contexts.

In this thesis proposal, I begin by presenting work in which we interview 26 NLP researchers about the state of the field, and argue that the predominant method of measuring and publicizing NLP systems' capabilities, the static benchmark, is increasingly used to support poorly scoped inductive claims of what AI can and cannot do. This insufficiency leads to widely released technology that, despite its promises of generality, I argue often poorly address the needs of knowledge workers in how they conceptualize and perform their jobs . I propose two lines of work to address this: benchmark datasets that prioritize *ecological validity*, in which models that perform well on benchmarks have immediate practical utility; and interfaces to NLP technologies that allow for users — both with and without NLP engineering assistance — to design and implement NLP-based tools that augment their capability in ways that center their agency.

Table of Contents

Chapter 1

Introduction

As NLP (and more broadly, AI) systems see wide deployment amid a remarkably capital-intensive infrastructure buildout across the tech industry, there remains a persistent tension between narratives of the AI's general-purpose capabilities and impending economic disruption advanced by industry players (Wilkins, 2025) and the much more qualified benefits being seen across the economy. With economists and analysts disagreeing about the potential upside of AI (Acemoglu, 2024; Nathan *et al.*, 2024), reports that only 5% of enterprise AI prototypes reach production (Challapally *et al.*, n.d.), and experts seeing AI as most fit for menial tasks under human supervision (Woodruff *et al.*, 2024), AI is perhaps more properly seen as ”normal technology” (Narayanan & Kapoor, n.d.), i.e. technology whose use and impacts vary enormously with the domain in which it is deployed, even as some of those impacts might well be seriously disruptive.

At the heart of this tension between narratives are the systems of measurement that we use to judge the capabilities of AI systems. In the vast majority of cases even today, this remains the static benchmark: a collection of human-labeled instances against which a system's predictions are measured. Despite longstanding and persistent criticism of benchmarks regarding their inability to inductively measure ”general capability” (Raji *et al.*, 2021), calls for measurement to be and contextual and specific(Bowman & Dahl, 2021a; Saxon *et al.*, 2024) and to handle subject position and disagreements among populations that may use the resulting models (Plank, 2022), benchmarks remain the most common way to frame progress in NLP.

In this thesis proposal, I argue that benchmarks are increasingly used as a rhetorical device to support poorly scoped inductive claims about general-purpose capabilities in models, which in turn facilitates insufficiently grounded, massive continued investment in the technology and a focus on speculative future harms, rather than ongoing issues such as de-skilling of workers (Sambasivan & Veeraraghavan, 2022), the compromise of information ecosystems (Shah & Bender, 2024), and concentration of corporate power (Widder *et al.*, 2024b). I propose two lines of work to address the growing insufficiency of the benchmark paradigm:

narrowly scoped, domain-specific benchmark datasets that prioritize ecological validity, or the ability to immediately transfer benchmark performance into real-world benefit; and model based tools that shift evaluation from static benchmarks to human-centered interactive evaluations that prioritize individual users finding utility in these systems.

1.1 Thesis Structure

In part 1 (chapters 2, 3, and 4) of this thesis, I characterize the role of benchmarks in the modern NLP/AI ecosystem, and the growing issues with that role. I discuss the origins of benchmarking culture in NLP and the current relationship of the field to benchmarks (Gururaja *et al.*, 2023), before discussing how benchmark results and trends are used outside of their academic contexts to justify massive investment (chapter 3, in-progress, Widder *et al.*, 2024a, under review). Finally, I discuss the processes that experts engage in in the materials science, law, and policy domains when performing knowledge work with documents that the benchmarked systems propose to replace. In (Gururaja *et al.*, 2025a), we find that in many cases, accessible systems and the benchmarks used to measure them do not reflect the contingent, social understanding of documents that experts have, and consequently do not support many of their needs.

In part 2 of this thesis (chapters 5-8), I discuss interventions to improve the degree to which benchmarks reflect individual user-level utility of NLP systems. In (Gururaja *et al.*, 2025c, in progress), we propose data-driven design as a benchmark task that both fits the existing paradigm, while being ecologically valid, i.e. a system that performs well at the benchmark represents an immediate utility to non-NLP expert users of that system. We then move on to tools that consider interactive, rather than static evaluation, beginning with Collage (Gururaja *et al.*, 2025b), a system that facilitates co-design of information extraction (IE) between stakeholders and ML engineers. We then discuss several proposed works building up to an in-browser system to allow user-specified, on-device information extraction with flexible schemas.

1.2 Current Status

This section summarizes completed, ongoing, and proposed work.

1.2.1 Completed Work

1. **Sireesh Gururaja**, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in

Natural Language Processing. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023, pages 13310–13325, 2023.

2. **Sireesh Gururaja**, Nupoor Gandhi, Jeremiah Milbauer, and Emma Strubell. Beyond Text: Characterizing Domain Expert Needs in Document Research. In Findings of the Association for Computational Linguistics: ACL 2025, pages 4732–4745, 2025.

3. **Sireesh Gururaja**, Yueheng Zhang, Guannan Tang, Tianhao Zhang, Kevin Murphy, Yu-Tsen Yi, Junwon Seo, Anthony Rollett, and Emma Strubell. Collage: Decomposable Rapid Prototyping for Co-Designed Information Extraction on Scientific PDFs. In Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 72–82, 2025.

1.2.2 In-progress Work

1. David Gray Widder, **Sireesh Gururaja**, and Lucy Suchman. Basic Research, Lethal Effects: Military AI Research Funding as Enlistment, Preprint, November 2024. arXiv:2411.17840 [cs]. Under review at Big Data and Society.

2. **Sireesh Gururaja**, Jeremiah Milbauer, Hung-Yi Lin, Junwon Seo, Anthony Rollett, and Emma Strubell. Data driven design as a challenge task for few-and zero-shot information extraction. Presented non-archivally at the Workshop for AI and Scientific Discovery at NAACL 2025, with a dataset proposal to appear at the AI4Science Workship at NeurIPS 2025.

3. From Glamor-Proof to Proof of Glamor: The Evolving Role of Benchmarks in NLP.

1.2.3 Proposed Work

1. Tractor Beam: a browser extension to facilitate on-device, user-designed information extraction on PDF documents. This project will involve the design and building of the browser extension and user studies to determine the extent to which users from outside NLP are able to effectively customize and evaluate information extraction systems on PDF documents.

2. A broad evaluation of the state-of-the-art in grounded, on-device few-shot information extraction. This work will attempt to answer what combination of model, training, and inference strategy results in the best on-device IE performance given a laptop-sized compute budget and a small number of training instances. Experiments in this paper will involve evaluating several families of on-device models against a variety of

information extraction tasks, ranging from classic datasets like the CoNLL 2003 shared task (Tjong Kim Sang & De Meulder, 2003), to domain-specific extraction in specific domains like materials science (Gururaja *et al.*, 2025c; Mysore *et al.*, 2019). We plan to evaluate not only currently competitive decoder-only transformer models, but also older encoder-only and encoder-decoder approaches, along with zero- and few-shot methods including prompting, in-context learning, and low-rank adaptation (Hu *et al.*, 2022).

1.3 Thesis Timeline

March 2025	TractorBeam Prototype Complete, begin user studies
May 2025	DDD benchmark and benchmarking chapter submitted
September 2025	TractorBeam user studies complete, job search
January 2026	Begin work on on-device IE characterization work.
May 2026	Estimated graduation.

Part I

Characterizing Gaps Left By Generality

Chapter 2

To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing (Complete)

Modified from a paper published in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Gururaja *et al.*, 2023).

2.1 Overview

In this chapter, we present the results of an interview study of 26 participants in NLP, across academia and industry, spanning the 1970s to the present. This chapter attempts to contextualize the then-current moment in NLP, and is the basis for several of the themes we discuss later in the proposal: the focus on benchmark evaluation, the impacts of centralization on particular tools, frameworks, and models, and the effects of NLP technology now seemingly being able to address lay user issues without expert intervention.

2.2 Introduction

Natural language processing (NLP) is in a period of flux. The successes of deep neural networks and large language models (LLMs) in NLP coincides with a shift not only in the nature of our research questions and methodology, but also in the size and visibility of our field. Since the mid-2010s, the number of first-time authors publishing in the ACL Anthology has been increasing exponentially (Figure 2.1). Recent publicity around NLP technology, most notably ChatGPT, has brought our field into the public spotlight, with corresponding (over-)excitement and scrutiny.

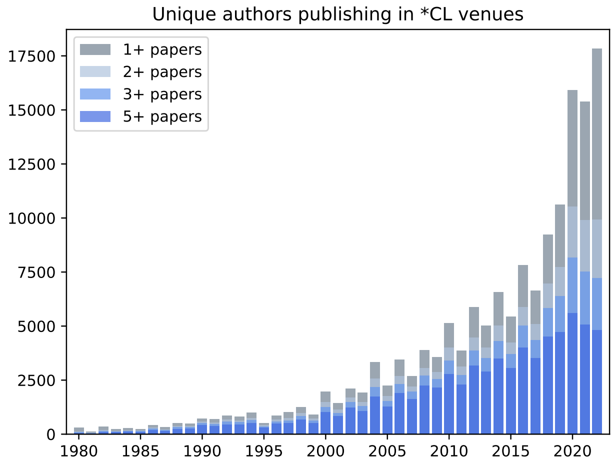


Figure 2.1. The number of unique researchers publishing in ACL venues has increased dramatically, from 715 unique authors in 1980 to 17,829 in 2022.

In the 2022 NLP Community Metasurvey, many NLP practicioners expressed fears that private firms exert excessive influence on the field, that "a majority of the research being published in NLP is of dubious scientific value," and that AI technology could lead to a catastrophic event this century (Michael *et al.*, 2022). More recently, there has been discussion of the increasing prevalence of closed-source models in NLP and how that will shape the field and its innovations (Liao & Vaughan, 2023; Rogers, 2023; Solaiman, 2023). In order to tackle these big challenges, we must understand the factors — norms, incentives, technology and culture — that led to our current crossroads.

We present a study of the community in its current state, informed by a series of long-form retrospective interviews with NLP researchers. Our interviewees identify patterns throughout the history of NLP, describing periods of research productivity and stagnation that recur over decades and appear at smaller scale around prominent methods (§2.4). Interviewees also point out unparalleled shifts in the community's norms and incentives. Aggregating trends across interviews, we identify key factors shaping these shifts, including the rise and persistence of benchmarking culture (§2.5) and the maturation and centralization of software infrastructure (§2.6). Our quantitative analysis of citation patterns, authorship, and language use in the ACL Anthology over time provides a complementary view of the shifts described by interviewees, grounding their narratives and our interpretation in measurable trends. Through our characterization of the current state of the NLP research community and the factors that have led us here, we aim to offer a foundation for informed reflection on the future that we as a community might wish to see.

2.3 Methods

2.3.1 Qualitative methods

We recruited 26 researchers to participate in interviews using *purposive* Campbell *et al.* (2020) and *snowball* sampling Parker *et al.* (2019): asking participants to recommend other candidates, and purposively selecting for diversity in affiliation, career stage, geographic position, and research area (see participant demographics in Appendix 2.12.1). Our sample had a 69-31% academia-industry split, was 19% women, and 27% of participants identified as part of a minoritized group in the field. Of our academic participants, we had a near-even split of early-, mid-, and late-career researchers; industry researchers were 75% individual contributors and 25% managers.

Interviews were semi-structured (Weiss, 1995), including a dedicated notetaker, and recorded with participant consent, lasting between 45-73 minutes (mean: 58 minutes). Interviews followed an interview guide (see section 2.12.3) and began by contrasting the participant's experience of the NLP community at the start of their career and the current moment, then moved to discussion of shifts in the community during their career. Interviews were conducted between November 2022 and May 2023; these interviews were coincidentally contemporaneous with the release of ChatGPT in November 2022 and GPT-4 in March 2023, which frequently provided points of reflection for our participants.

Following procedures of grounded theory Strauss & Corbin (1990), the authors present for the interview produced analytical memos for early interviews Glaser *et al.* (2004). As interviews proceeded, authors began a process of independently *open coding* the data, an *interpretive* Lincoln *et al.* (2011) analytical process where researchers assign conceptual labels to segments of data Strauss & Corbin (1990). After this, authors convened to discuss their open codes, systematizing recurring themes and contrasts to construct a preliminary closed coding frame Miles & Huberman (1994). After this, an author who was not present in the interview applied the closed coding frame to the data. In weekly analysis meetings, new codes arose to capture new themes or provide greater specificity, in which case the closed coding scheme was revised, categories refined, and data re-coded in an iterative process. Analysis reported here emerged first from this coded data, was refined by subsequent review of raw transcripts to check context, and developed in discussion between all authors.

2.3.2 Quantitative Methods

We use quantitative methods primarily as a coherence check on our qualitative results. While our work is largely concerned with the *causes* and *community reception* of changes in the community, our quantitative analyses provide evidence that these changes have occurred.

This includes analyzing authorship shifts (Figure 2.1 and Figure 2.4), citation patterns (Figure 2.2), terminology use (Figure 2.2,Figure 2.3) in the ACL anthology over time; for more details on reproducing this analysis, see Appendix 2.13.

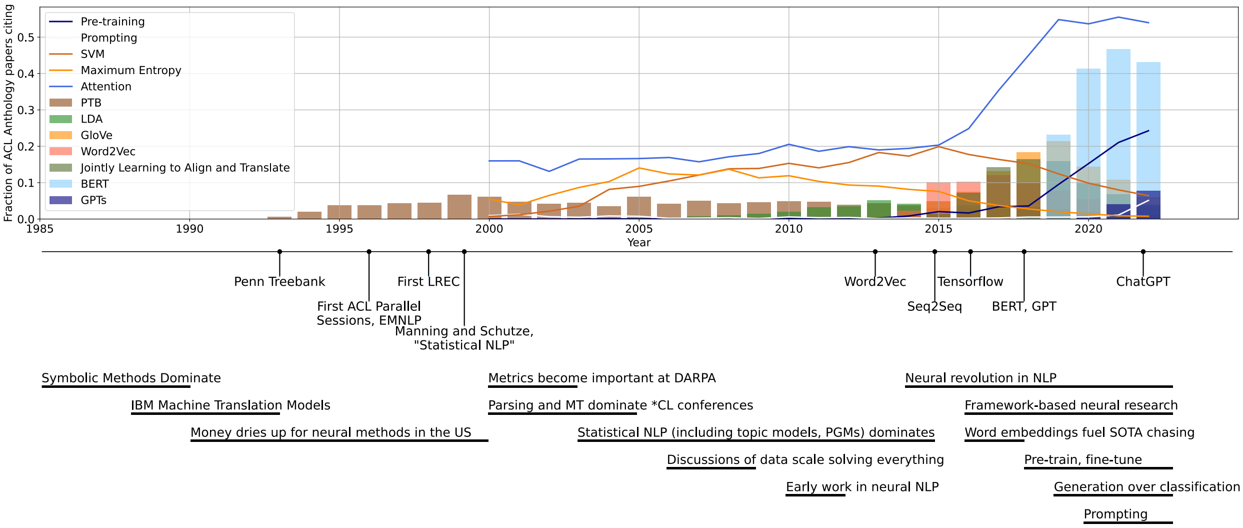


Figure 2.2. Quantitative and qualitative timeline. The lower half of this diagram captures historical information that our participants felt was relevant, along with their reported date ranges. The upper half captures quantitative information that parallels that timeline. Bar charts indicate fraction of papers that cite a given paper, while line charts indicate the fraction of papers that use a particular term.

2.4 Exploit-explore cycles of work

Our participants described cyclical behavior in NLP research following methodological shifts every few years. Many participants referred to these methodological shifts as "paradigm shifts", with similar structure, which we characterize as *explore* and *exploit* phases.[1]

First wave: exploit. Participants suggested that after a key paper, a wave of work is published that demonstrates the utility of that method across varying tasks or benchmarks. Interviewees variously describe this stage as *"following the bandwagon" (17)*, *"land grab stuff" (9)*, or *"picking the low-hanging fruit" (8)*. A researcher with prior experience in computer vision drew parallels between the rise of BERT and the computer vision community after AlexNet was released, where *"it felt like every other paper was, 'I have fine tuned ImageNet trained CNN on some new dataset' " (17)*. However, participants identified benefits of this *"initial wave of showing things work" (17)* in demonstrating the value of techniques across

[1] Inspired by the verbiage of reinforcement learning, e.g. Sutton & Barto (2018).

tasks or domains; in finding the seemingly obvious ideas which do *not* work, thus exposing new areas to investigate; and in developing downstream applications. When one researcher was asked to identify their own exploit work, they joked we could simply *"sort [their] Google Scholar from most cited to least cited" (9)*, describing another incentive to publish the first paper applying a new methodology to a particular task. Participants additionally described doing exploit work early in their careers as a hedge against riskier, longer-term projects. However, most participants ascribed low status to exploit work[2], with participants calling these projects *"low difficulty" (4)* and *"obvious idea[s]" (9)* with *"high probability of success" (4)*. Participants also discussed increasing risk of being *"scooped" (7,9,4)* on exploit work as the community grows larger.

Participants felt that **the exploit phase of work is not sustainable.** Eventually, *"the low hanging fruit has been picked" (8)*; this style of work becomes unsurprising and less publishable. As one researcher put it: *"if I fine tune [BERT] for some new task, it's probably going to get some high accuracy. And that's fine. That's the end." (17)*.

Second wave: explore After some time in the exploit phase, participants described a state where obvious extensions to the dominant methodology have already been proposed, and fewer papers demonstrate dramatic improvements over the state of the art on popular benchmarks.

In this phase, work on identifying the ways that the new method is flawed gains prominence. This work may focus on interpretability, bias, data or compute efficiency. While some participants see this as a time of *"stalled" (8)* progress, others described this as *"the more interesting research after the initial wave of showing things work" (18)*. A mid-career participant identified this as the work they choose to focus on: *"I'm at the stage of my career where I don't want to just push numbers, you know. I'll let the grad students do that. I want to do interesting stuff" (22)*. Participants often saw "pushing numbers" as lower-status work, appropriate for graduate students and important for advancing the field and one's career, but ultimately not what researchers hope to explore.

Some work to improve benchmark performance was also perceived as explore work, particularly when it involved developing new architectures. One participant described a distinction between *"entering a race"* and *"forging in a new direction" (4)* with a project, which focuses the exploit/explore divide more on the perceived surprisingness of an idea rather than the type of contribution made. **Exploration often leads to a new breakthrough, causing the cycle to begin anew.**

[2] Data work is also commonly perceived as low-status Sambasivan *et al.* (2021a); participants agreed data work was previously undervalued in NLP but described a trend of increasing respect, one calling dataset curation *"valorized." (7)*

2.4.1 Where are we now?

Placing the current state of the field along the exploit-explore cycle requires defining the current methodological "paradigm". Participants identified similar patterns at varying scales, with some disagreement on the timing of recent trends.

Prompting as a methodological shift Several participants described prompting as a paradigm shift or a direction that the community found promising, but most participants viewed current work on prompt engineering or *"ChatGPT for X" (9)* as something that people are working on *"instead [...] of something that might make a fundamental difference" (14)*. One participant described both prompt engineering and previous work on feature engineering as *"psuedoscience [...] just poking at the model" (6)*. The current flurry of prompting work was viewed by several participants as lower-status work exploiting a known method.

"Era of scale" For participants who discussed larger-scale cycles, pre-trained models were frequently identified as the most recent methodological shift. Participants disagreed on whether scaling up pre-trained models (in terms of parameter count, training time, and/or pre-training data) was a form of exploiting or exploring this method. Some participants found current approaches to scale to be *"a reliable recipe where we, when we put more resources in, we get [...] more useful behavior and capabilities out" (4)* and relatively easy to perform: *"Once you have that GPU [...] it's like, super simple" (5)*. This perception of scaling up as both high likelihood of success and low difficulty places it as exploit work, and researchers who described scale in this way tended to view it as exploiting "obvious" trends. One researcher described scale as a way of establishing what is possible but *"actually a bad way to achieve our goals." (4)*, with further (explore-wave) work necessary to find efficient ways to achieve the same performance.

A minority of participants argued that, while historical efforts to scale models or extract large noisy corpora from the internet were exploit work, current efforts to scale are different, displaying *"emergence in addition to scale, whereas previously we just saw [...] diminishing returns" (22)*. Some participants also emphasized the engineering work required to scale models, saying that some were *"underestimating the amount of work that goes into training a large model" (8)* and identifying people or engineering teams as a major resource necessary to perform scaling work. The participants who described scaling work as producing surprising results and being higher difficulty also described scaling as higher status, more exploratory work.

"Deep learning monoculture" There was a sense from several participants that the current cycle has changed the field more than previous ones, featuring greater centralization on fewer methods (see §2.6 for more discussion). Some expressed concern: *"a technique* *shows some promise, and then more people investigate it. That's perfectly appropriate and reasonable, but I think it happens a little too much. [...]* ***Everybody collapses on this one approach [...] everything else gets abandoned.*** *" (19).* Another participant described peers from linguistics departments who left NLP because they felt alienated by the focus on machine learning.

Issues with peer review Some felt that peer review was inherently biased toward incremental work because peer reviewers are invested in the success of the current methodological trends, with one participant arguing that *"if you want to break the paradigm and do something different, you're gonna get bad reviews, and that's fatal these days" (21).* Far more commonly, participants did not express inherent opposition to peer review but raised concerns because of the recent expansion of the field, with one senior industry researcher remarking that peer reviewers are now primarily junior researchers who *"have not seen the effort that went into [earlier] papers" (12).* Another participant asserted that ***"my peers never review my papers"*** *(22).* Participants additionally suggested that the pressure on junior researchers to publish more causes an acceleration in the pace of research and reinforcement of current norms, as research that is farther from current norms/methodologies requires higher upfront time investment.

This competitiveness can manifest in harsher reviews, and one participant described a *"deadly combination" (19)* of higher standards for papers and lower quality of reviews. Some participants described this as a reason they were choosing to engage less with NLP conferences; one industry researcher stated that *"I just find it difficult to publish papers in *CL [venues] that have ideas in them." (22).*

2.5 Benchmarking culture

2.5.1 The rise of benchmarks

Senior and emeritus faculty shared a consistent recollection of the ACL community before the prominence of benchmarks as centralized around a few US institutions and characterized by *"patient money" (21)*: funding from DARPA that did not require any deliverables or statements of work. Capabilities in language technologies were showcased with technical *"toy" (26, 19)* demonstrations that were evaluated qualitatively: *"the performance metrics were, 'Oh my God it does that? No machine ever did that before.' " (21).* Participants repeatedly mentioned how small the community was; at conferences, *"everybody knew each other. Everybody was conversing, in all the issues" (26).* **The field was described as *"higher trust" (22)*,** with social mediation of research quality – able to function in the absence of standardization because of the strong interconnectedness of the community.

Many participants recalled the rise of benchmarks in the late 1990s and early 2000s, coinciding with a major expansion in the NLP community in the wake of the "statistical revolution," where participants described statistical models displacing more traditional rules-based work (see Figure 2.2). In the words of one participant, the field became of *"such a big snowballing size that nobody owned the first evers anymore." (26)*. Instead, after the release of the Penn Treebank in 1993 and the reporting of initial results on the dataset, *"the climb started" (25)* to increase performance. Some participants attributed these changes to an influx of methods from machine learning and statistics, while others described them as methods to understand and organize progress when doing this coordination through one's social network was no longer feasible.

Over time, this focus on metrics seems to have overtaken the rest of the field, in part through the operationalization of metrics as a key condition of DARPA funding. One participant credited this to Anthony Tether, who became director of DARPA in 2001: they described earlier DARPA grants as funding *"the crazy [...] stuff that just might be a breakthrough" (21)* and DARPA under Tether as *"show me the metrics. We're going to run these metrics every year." (21)*.[3]

Some participants mourned the risk appetite of a culture that prioritized "first-evers," criticizing the lack of funding for ideas that did not immediately perform well at evaluations (notably leading to the recession of neural networks until 2011). However, there was sharp disagreement here; many other participants celebrated the introduction of benchmarks, with one stating that comparing results on benchmarks between methods *"really brought people together to exchange ideas. [...] I think this really helped the field to move forward." (2)*. Other participants similarly argued that a culture of quantitative measurement was key for moving on from techniques that were appealing for their *"elegance" (14)* but empirically underperforming.

2.5.2 The current state of benchmarks

Roughly twenty years on from the establishment of benchmarks as a field-wide priority, our participants' attitudes towards benchmarks had become significantly more complex. Many of our participants still found benchmarks necessary, but nearly all of them found them increasingly insufficient.

Misaligned incentives Many participants, particular early- and late-career faculty, argued that the field incentivizes the production of benchmark results to the exclusion of all else: *"the*

[3] Other participants named DARPA program managers Charles Wayne and J. Allen Sears as additional key players. A recent tribute to Wayne Church (2018) provides additional context reflecting on DARPA's shift in priorities in the mid-1980s.

typical research paper...their immediate goal has to be to get another 2% and get the boldface black entry on the table." (21). For our participants, **improvements on benchmarks in NLP are the only results that are self-justifying to reviewers.** Some participants felt this encourages researchers to exploit modeling tricks to get state-of-the-art results on benchmarks, rather than explore the deeper mechanisms by which models function (see §2.4).

"We're solving NLP" Some participants perceive a degradation in the value of benchmarks because of the strength of newer models. Participants appreciated both the increased diversity and frequency of new benchmark introduction, but noted that the time for new approaches to reach *"superhuman" (6,22)* levels of performance on any specific benchmark is shortening. One common comparison was between part of speech tagging (*"a hill that was climbed for [...] about 20 years" (15)*) and most modern benchmarks ("solved" within a few years, or even months). Some went further, describing *"solving NLP" (8)* or naming 2020 as the time when *"classification was solved" (15).*

However, when participants were asked for clarification on what it meant to "solve" a problem, most participants hedged in similar ways; that datasets and benchmarks could be solved, with the correct scoping, but problems could rarely or never be solved. Many participants argued that the standard for solving a task should be human equivalency, and that this was not possible without new benchmarks, metrics, or task definition.

NLP in the wild Some participants argued that many benchmarks reflect tasks that *"aren't that useful in the world" (13)*, and that this has led to a situation where *"[NLP], a field that, like fundamentally, is about something about people, knows remarkably little about people" (3).* Industry participants often viewed this as a distinction between their work and the academic community, with one stating that *"most of the academic benchmarks out there are not real tasks" (12).* Many academics articulated a desire for more human-centered NLP, and most participants described feeling pressure over the unprecedented level of outside interest in the field. One participant contrasted the international attention on ChatGPT with the visibility of earlier NLP work: *"It's not like anyone ever went to like parser.yahoo.com to run a parser on something" (3).* Participants argued that, given this outside attention, the benchmark focus of NLP is too narrow, and that **benchmarks fail to capture notions of language understanding that translate to wider audiences**, and that we should move on from benchmarks not when they are saturated but when *"it wouldn't really improve the world to improve this performance anymore" (9).* This echoed a common refrain: many participants, especially early- and mid-career researchers, saw positive social change as a goal of progress in NLP.

2.6 Software lotteries

Hooker (2021) argues that machine learning research has been shaped by a *hardware lottery*: an idea's success is partially tied to its suitability for available hardware. Several participants spoke about software in ways that indicate an analogous *software lottery* in NLP research: as the community centralizes in its software use, it also centralizes in methodology, with researchers' choices of methods influenced by relative ease of implementation. This appeared to be a relatively new phenomenon; participants described previously using custom-designed software from their own research group, or writing code from scratch for each new project.

Centralization on frameworks As deep learning became more popular in NLP, participants described the landscape shifting. As TensorFlow (Abadi *et al.*, 2015) increased support for NLP modeling, PyTorch (Paszke *et al.*, 2019a) was released, along with NLP-specific toolkits such as DyNet Neubig *et al.* (2017) and AllenNLP (Gardner *et al.*, 2018), and *"everything started being [...] simpler to manage, simpler to train" (17)*. Previously, participants described spending *"like 90% of our time re-implementing papers" (12)*; as more papers began releasing code implemented in popular frameworks, the cost of using those methods as baselines decreased. One participant stated that *"things that software makes easy, people are going to do" (18)*; this further compounds **centralization onto the most popular libraries, with little incentive to stray from the mainstream**: *"everybody uses PyTorch, so now I use PyTorch too" (8)*; *"we just use HuggingFace for pretty much everything" (18).* [4] Figure 2.3 visualizes mentions of frameworks across papers in the ACL Anthology, showing both the rise and fall in their popularity. The rising peaks of popularity reflect the centralization over time. While some communities within NLP had previously seen some centralization on toolkits (e.g. machine translation's use of Moses (Koehn *et al.*, 2007)), the current centralization transcends subfields.

Centralization on specific models Participants identified another shift after the release of BERT and subsequent development of Hugging Face. Because of pre-training, participants moved from merely using the same libraries to *"everyone us[ing] the same base models" (9)*. Participants expressed concern that this led to further centralization in the community, with one participant identifying a trend of some people *"not us[ing] anything else than BERT [...] that's a problem" (5)*. This concentration around particular modeling choices has reached **greater heights than any previous concentration on a method**; in 2021, 46.7% of papers in the ACL anthology cited BERT (Devlin *et al.* (2019); Figure 2.2) [5].

[4] In this section, we primarily discuss open source frameworks commonly used by academic and industry researchers. However, many of our participants working in industry also describe affordances and constraints of *internal* toolkits that are used at their respective companies.

[5] While not every paper that cites BERT uses a BERT model, this indicates how central BERT is both as a model and as a frame for the discussion of other work. For comparison, only two other papers have been

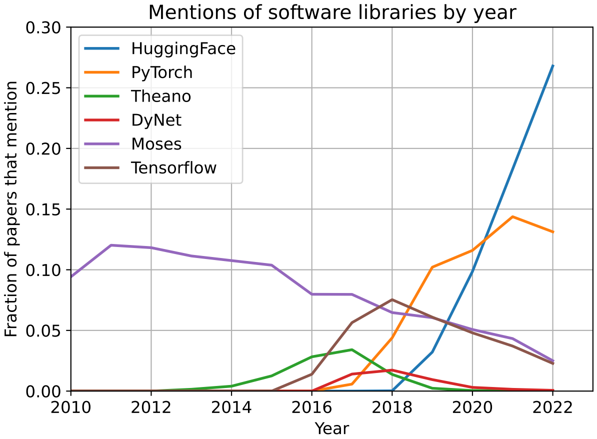


Figure 2.3. Mentions of libraries over time in the ACL Anthology. Note the cyclic pattern and increasing concentration on the dominant framework over time. While some libraries are built on others, the shift in mentions over time captures the primary level of abstraction that researchers consider important. See appendix 2.13 for details on how we handle ambiguity in mentions.

Other large models are only available to researchers via an API. One participant who works on LLMs in industry argued that black-box LLMs, while *"non-scientific" (15)* in many ways, were like large-scale technical tools constructed in other disciplines, drawing a parallel to particle physics: *"computer science is getting to have its CERN moment now [...] there's only one Large Hadron Collider, right?" (15).* This participant argued that NLP has become a field whose frontiers require tools that are beyond most organizations' resources and capabilities to construct, but nonetheless are widely adopted and set bounding parameters for future research as they see wide adoption. In this vision, black-box LLMs take on the same role as the LHC or the Hubble Space Telescope (both notably public endeavors, unlike most LLMs), as tools whose specifications are decided on by a few well-resourced organizations who shape significant parts of the field. But most participants expressed skepticism about the scientific validity of experiments on black-box LLMs, with one participant referencing critiques of early-2000s IR research on Google Kilgarriff (2007).

Centralization on Python While most early-career and late-career participants did not express strong opinions about programming languages, many mid-career participants expressed strong dislike for Python, describing it as *"a horrible language" (22)* with efficiency issues that are *"an impediment to us getting things done" (20)* and data structure implementations that are *"a complete disaster in terms of memory" (9).* One participant described

cited by more than 20% of anthology papers in a single year: "Attention is All You Need" (Vaswani *et al.*, 2017) with 27% in 2021 and GloVe (Pennington *et al.*, 2014) with 21% in 2019.

their ideal next paradigm shift in the field as a shift away from using Python for NLP. Yet even the participants who most vehemently opposed Python used it for most of their research, citing the lack of well-supported NLP toolkits or community use in other languages. This is an instance of a software lottery at a higher level, where the dominance of a single programming language has snowballed with the continued development of research artifacts in that language.

2.6.1 Consequences of centralization

This increasing centralization of the modern NLP stack has several consequences. One of the primary ones, however, is the loss of control of design decisions for the majority of researchers in the community. Practically, researchers can now choose from a handful of well-established implementations, but only have access to software and models once the decisions on how to build them have already been reified in ways that are difficult to change.

Lower barriers Beyond time saved (re-) implementing methods, many participants identified a lower barrier to entry into the field as a notable benefit of centralization on specific software infrastructure. Participants described students getting state of the art results within an hour of tackling a problem; seeing the average startup time for new students decreasing from six months to a few weeks; and teaching students with no computer science background to build NLP applications with prompting.

Obscuring what's "under the hood" One participant recalled trying to convince their earlier students to implement things from scratch in order to understand all the details of the method, but no longer doing so because *"I don't think it's possible [...] it's just too complicated" (11)*; others attributed this to speed more than complexity, stating that *"the pace is so fast that there is no time to properly document, there is no time to properly engage with this code, you're just using them directly" (5).* However, this can cause issues on an operational level; several participants recalled instances where a bug or poor documentation of shared software tools resulted in invalid research results. One participant described using a widely shared piece of evaluation code that made an unstated assumption about the input data format, leading to *"massively inflated evaluation numbers" (3)* on a well-cited dataset. Another participant described working on a paper where they realized, an hour before the paper deadline, that the student authors had used two different tokenizers in the pipeline by mistake: *"we decided that well, the results were still valid, and the results would only get better if [it was fixed]...so the paper went out. It was published that way." (26)* Software bugs in research code are not a new problem,[6] but participants described bugs in toolkits

[6] Tambon *et al.* (2023) describe *silent bugs* in popular deep learning frameworks that escape notice due to undetected error propagation.

as difficult to diagnose because they *"trust that the library is correct most of the time" (8)*, even as they spoke of finding ***"many, many, many" (8)* bugs in toolkits** including HuggingFace and PyTorch.

Software is implicit funding Participants suggested that tools that win the software lottery act as a sort of implicit funding: they enable research groups to conduct work that would not be possible in the tools' absence, and many of our participants asserted that the scope of their projects expanded as a result. However, they also significantly raise the relative cost of doing research that does not fall neatly into existing tools' purview. As one participant stated, *"You're not gonna just build your own system that's gonna compete on these major benchmarks yourself. You have to start [with] the infrastructure that is there" (19)*. This is true even of putatively "open" large language models, which do not necessarily decentralize power, and can often entrench it instead (Widder *et al.*, 2023). This set of incentives pushes researchers to follow current methodological practice, and some participants feared this led toward more incremental work.

2.6.2 Impact on Reproducibility

A common sentiment among participants was that centralization has had an overall positive impact on reproducibility, because using shared tools makes it easier to evaluate and use others' research code. However, participants also expressed concerns that the increasing secrecy of industry research complicates that overall narrative: *"things are more open, reproducible... except for those tech companies who share nothing" (14)*.

Shifts in expectations One participant described a general shift in focus to *"making sure that you make claims that are supported rather than reproducing prior work exactly" (12)* in order to match reviewers' shifting expectations. However, participants also felt that the expectations for baselines had increased: *"[in the past,] everybody knew that the Google system was better because they were running on the entire Internet. But like that was not a requirement [to] match Google's accuracy. But now it is, right?" (8)*.

Disparities in compute access Many felt that building large-scale systems was increasingly out of reach for academics, echoing concerns previously described by Ahmed & Wahed (2020). Participants worried that *"we are building an upper class of AI" (6)* where most researchers must *"build on top of [large models]" (15)* that they cannot replicate, though others expressed optimism that *"clever people who are motivated to solve these problems" (22)* will develop new efficient methods Bartoldson *et al.* (2023). Industry participants from large tech companies also felt resource-constrained: *"modern machine learning expands to fit the available compute." (4)*.

2.7 Related Work

The shifts we explore in this paper have not happened in a vacuum, with adjacent research communities such as computer vision (CV) and machine learning (ML) experiencing similar phenomena, inspiring a number of recent papers discussing norms in AI more broadly. Birhane *et al.* (2022) analyze a set of the most highly cited papers at recent ML conferences, finding that they eschew discussion of societal need and negative potential, instead emphasizing a limited set of values benefiting relatively few entities. Others have noticed that corporate interests have played an increasing role in shaping research, and quantified this with studies of author affiliations over time in machine learning (Ahmed & Wahed, 2020) and NLP (Abdalla *et al.*, 2023). Su & Crandall (2021) study the tangible emotional impact of recent dramatic growth in the CV community by asking community members to write stories about emotional events they experienced as members of their research community.

While we focus on summarizing and synthesizing the views of our participants, some of the over-arching themes identified in this work have been discussed more critically. Fishman & Hancox-Li (2022) critique the unification of ML research around transformer models on both epistemic and ethical grounds. Position papers have critiqued the notion of general purpose benchmarks for AI (Raji *et al.*, 2021), and emphasized the importance of more careful and deliberate data curation in NLP (Bowman & Dahl, 2021b; Rogers, 2021).

The NLP Community Metasurvey (Michael *et al.*, 2022) provides a complementary view to this work, with their survey eliciting opinions from a broad swath of the *CL community on a set of 32 controversial statements related to the field. The survey also asked respondents to guess at what the most popular beliefs would be, eliciting sociological beliefs about the community. While there is no direct overlap between our questions and Metasurvey questions, participants raised the topics of scaling up, benchmarking culture, anonymous peer review, and the role of industry research, which were the subject of Metasurvey questions. Where we can map between our thematic analysis and Metasurvey questions, we see agreement– e.g. many of our participants discussed others valuing scale, but few placed high value themselves on scaling up as a research contribution.

The availability of the ACL Anthology has enabled quantitative studies of our community via patterns of citation, authorship, and language use over time. Anderson *et al.* (2012) perform a topic model analysis over the Anthology to identify different eras of research and better understand how they develop over time, and analyze factors leading authors to join and leave the community. Mohammad (2020) analyze citation patterns in *CL conferences across research topics and paper types, and Singh *et al.* (2023) specifically inspect the phenomenon wherein more recent papers are less likely to cite older work. Pramanick *et al.* (2023) provide a view of paradigm shifts in the NLP community complementary to ours based on a diachronic analysis of the ACL Anthology, inferring causal links between datasets, methods, tasks and metrics.

Shifts in norms and methods in science more broadly has been studied outside computing-related fields. Most notably, Kuhn (1970) coined the term *paradigm shift* in *The Structure of Scientific Revolutions*. His theory of the cyclic process of science over decades or centuries has some parallels with the (shorter timescale) exploit-explore cycles discussed in this work. Note that in this work, we did not prime participants with an *a priori* definition of paradigm shift, allowing each participant to engage with the term according to their own interpretation, which often differed from Kuhn's notion of a paradigm shift.

2.8 The Future

The rise of large language models has coincided with disruptive change in NLP: accelerating centralization of software and methodologies, questioning of the value of benchmarks, unprecedented public scrutiny of the field, and dramatic growth of the community. A shift like this can feel threatening to the fundamental nature of NLP research, but this is not the first period of flux in the field, nor are the fundamental forces enabling LLMs' dominance and other changes entirely new.

Our participants described cycles of change in the NLP community from mid-80s to the present, with common themes of first exploiting and then exploring promising methodologies. Each methodological shift brought corresponding cultural change: the shift from symbolic to statistical methods brought about the rise of benchmark culture and the end of the socially mediated, small-network ACL community. Neural methods began the centralization on software toolkits and the methodologies they support. Pre-training intensified this software lottery, causing unprecedented levels of centralization on individual methods and models. Current models have called into question the value of benchmarks and catapulted NLP into the public eye. Our participants largely agree on the resulting incentives– to beat benchmark results, to do the easiest thing rather than the most fulfilling, to produce work faster and faster – while largely expressing frustration with the consequences.

We hope that this contextualization of the current state of NLP will both serve to inform newer members of the community and stir informed discussion on the condition of the field. While we do not prescribe specific solutions, some topics of discussion emerge from the themes of this work:

- Who holds the power to shape the field? How can a broad range of voices be heard?
- Do the incentives in place encourage the behavior we would like to see? How can we improve reviewing to align with our values?

- What affects the ability to do longer-term work that may deviate from current norms?
- How can the community arrive at an actively mediated consensus, rather than passively being shaped by forces like the ones we discuss?

We personally take great hope for our community from this project. The care with which all participants reflected on the shape of the field suggests to us that many people are concerned about these issues, invested in the community, and hopeful for the future. By sharing publicly what people so thoughtfully articulate privately, we hope to prompt further discussion of what the community can do to build our future.

2.9 Limitations

Western bias The most notably *irrepresentative* sampling bias in our participant pool is the lack of non-Western institutional affiliation (and the strong skew toward North American affiliations). This bias has arisen likely in part due to our own institutional affiliation and conceptions of the community. That being said, given the Association for Computational Linguistics' historically US- and English-centric skews, this allows us to gather historical perspectives. Additionally, considering that Western institutions constitute a citation network largely distinct from Asian networks (Rungta *et al.*, 2022), we believe that our sample allows us to tell a rich and thorough story of factors which have shaped the Western NLP research community, which both informs and is informed by other communities of NLP researchers.

Lack of early career voices Our inclusion criteria for our participants– three or more publications in *CL, IR, or speech venues[7]– necessarily means that we have limited perspectives on and from more junior NLP researchers (such as junior graduate students), those hoping to conduct NLP research in the future, those who have engaged with NLP research in the past and decided not to continue before developing a publication record, and those who have consciously decided *not* to engage with NLP research in the first place. In general, although we gathered perspectives from participants across a variety of demographic backgrounds, our participants represent those who have been successful and persisted in the field. This is especially true for our participants in academia; of our participants' current academic affiliations, only R1 institutions (if in the US) and institutions of comparable research output (if outside the US) are represented. We therefore may be missing perspectives from certain groups of researchers, including those who primarily engage with undergraduate students or face more limited resource constraints than most of the academic faculty we interviewed.

Future research could further examine differences between geographic subcommunities in NLP and more closely examine influences on people's participation in and disengagement

[7] In order to capture perspectives of the community changing over time, and to select for people who are part of these communities.

from the community. Additionally, we leave to future work a more intentional exploration of perspectives from early career researchers and those who have not yet published but are interested in NLP research.

2.10 Ethics Statement

Following Institutional Review Board recommendations, we take steps to preserve the anonymity of our participants, including aggregating or generalizing across demographic information, avoiding the typical practice of providing a table of per-interviewee demographics, using discretion to redact or not report quotes that may be identifying, and randomizing participant numbers. Participants consented to the interview and to being quoted anonymously in this work. This work underwent additional IRB screening for interviewing participants in GDPR-protected zones.

We view our work as having potential for positive impact on the ACL community, as we prompt its members to engage in active reflection. We believe that, given recent developments in the field and the co-occuring external scrutiny, the current moment is a particularly appropriate time for such reflection. Additionally, we hope that our work can serve those currently *external* to the community as an accessible, human-centered survey of the field and factors that have shaped it over the decades, prioritizing sharing of anecdotes and other in-group knowledge that may be difficult for outsiders to learn about otherwise.

2.11 Acknowledgements

This work would not have been possible without our participants, who were generous with their time and engaged deeply with this material. We would additionally like to thank Jill Fain Lehman, Anjalie Field, Dawn Nafus, Momin M. Malik, Graham Neubig and the anonymous reviewers for their helpful comments.

This work was supported in part by a grant from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2140739. Widder gratefully acknowledges the support of the Digital Life Initiative at Cornell Tech. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

2.12 Details of Qualitative Methodologies

2.12.1 Participant demographics

Of the academics interviewed, there was an even split between early, mid, and late career (6/33% each). Of those in industry, 6 (75%) were individual contributors and 2 (25%) were research managers. Our sample is only 19% women, which is likely representative, as women comprise approximately 15% of tenure-track computer science faculty in the US Way *et al.* (2016). For more discussion of the sample characteristics, see 2.9. Our positive response rate was 82%.

Demographic trait	% of sample
In academia	69%
Women	19%
Minoritized group	27%
Born outside US	38%
Currently works outside US	4%

Table 2.1. Self-reported demographic makeup of subjects.

2.12.2 Consent protocol

Interviewees were asked for verbal consent unless they were in a GDPR-protected region at the time of the interview, in which case they provided written consent via DocuSign instead. This consent script was IRB-approved.

> Hi, thanks for taking the time to talk with me! My collaborators on this project and I work for CMU. We can be reached at [emails] should you have any questions for us during the study or after.
> This interview will take between 45 minutes and an hour. There will be no compensation for participation.
> Participation is always voluntary, and you may refuse to participate in the research study or stop participation at any time.
> You will not be identified in any reports we release from this research. This data will be deidentified, which means you will not be identified by name or any other specific characteristic. We may quote you anonymously. Just in case, though, please do not reveal any private or personally-identifiable information about yourself or others in your answer to our questions.

I'd like to record the audio of this interview as a memory aid. You can ask me to stop the recording at any time. Only the members of our research team will have access to these recordings.
We really appreciate your participation, and we hope to publish this research to advance our understanding of the factors shaping the NLP research community.
Do you have any questions? If not, do I have your consent to participate in this study?
[answer any questions; if consented, begin recording]
Alright, I've started the recording. Just to confirm, do I have your consent to participate in this study, and to record this interview?

2.12.3 Interview Guide

These questions are intentionally open-ended, and the interviewers asked non-scripted followups or additional questions where appropriate. Over time, as early themes emerged, additional questions were added, particularly on funding and pace of work.

1. First, I have a few questions about your relationship to the NLP community. You can be as specific or as vague as you'd like with your responses.

    (a) What do you consider to be your main/home/primary research community?

    (b) More specifically, what venues do you follow and/or publish in?

    (c) What subarea(s) or subfields are you most active in?

    (d) Is this different from what you have considered your main community at other points in your career? (Prompt: if so, what changed?)

    (e) What would you define as the start of your NLP research career? (e.g. start of PhD, research as an undergrad, etc). (Prompt: When was this?)

2. I'd like to hear your thoughts on what the field was like near the beginning of your career.

    (a) When you started in your field, what did people generally think were the most promising directions? (Prompt: do you agree?)

    (b) How do you interpret the term "promising"?

    (c) What do you think the research community prioritized when you started?

    (d) What was the scope of the work that your research group did?

    (e) What was your relationship to computing resources at the start of your career?

(f) What did your software workflow look like when you first started doing research? (Prompt: What tools, frameworks, libraries did you use?)

(g) Where did funding for your work come from? (Prompt: what were the major costs involved with your research?)

Now, I'd like to compare this with the current state of the field.

(a) What do you think others in your field would say are the most promising directions? (Prompt: do you agree?)

(b) What do you think the research community prioritizes now?

(c) What is the scope of the work that your research group does?

(d) How do computing resources affect your group's work now?

(e) How does the software workflow look like for you or your students now?

(f) What tools, frameworks, libraries do you or your students use?

(g) Have these tools, frameworks, and libraries made an impact on your (or your students') research?

(h) What impacts do you think these tools, frameworks, and libraries have made on your community's research?

(i) Where does the funding for your work come from? (Prompt: what are the major costs involved in your research?)

(j) When you or your students start a new project, how long on average do you expect it to take, from the start of work to a paper submission?

(k) Has this changed over your career? (Prompt: what has led to this change?)

Now I'd like to hear your thoughts on the changes you've observed in your career.

(a) Are there paradigm shifts that you would identify in the field over the course of your career?

(b) Did your community change at all, as a result of these paradigm shifts?

(c) Prompt: what years would you ascribe to each shift?

(d) How frequently do you feel the community undergoes a paradigm shift? (Prompt: is this frequency changing?)

(e) Are there concerns you have with the direction of the field?

(f) If there were to be a paradigm shift in the next few years, in what direction should it go?

(g) Do you think changes in the research community have changed your teaching? (Prompts: how? How do you feel about this shift?)

Now I have some demographic questions, which will help us understand the range of people that we talk to. If you'd prefer not to answer any of them, just let me know.

(a) With which gender do you identify?

i. Man

ii. Woman

iii. Or, feel free to specify as you wish

(b) I am going to read some age brackets. Can you indicate when I read a bracket that your age falls into?

i. 18-24

ii. 25-34

iii. 35-44

iv. 45-54

v. 55-64

vi. 65+

(c) Which country were you born in?

(d) (if not already known) Which country are you currently based in?

(e) What stage of your academic career would you consider yourself in?

(f) Do you consider yourself a part of a minoritized group in your field?

(g) Anything else in your background that feels relevant or that you want to add?

3. Finally, we'd like to hear from more people about these issues. Is there anyone you could introduce us to who you think would have interesting answers to these questions?

2.13 Details of Quantitative Methodology

We begin with the ACL Anthology and focus on papers between 1980 and 2022. Using the SemanticScholar (S2) API Wade (2022), we obtain author and citation information of papers indexed by S2 (some venues, such as some workshops, and some Findings[8] papers, are not indexed, leaving $77,235$ papers), with a focus on citations of papers identified by our participants as having been influential to the NLP community. We select from this set of influential papers to generate the bar plots for Figure 2.2. Figure 2.1 uses author publication information linked to the individual papers considered.

We rely on S2ORC Lo *et al.* (2020) for full-text PDF parses of a subset of these papers, which we use match for mentions of software toolkits identified by our participants as having been influential to the community, for Figure 2.3, as well as for mentions of influential techniques in the line plot of Figure 2.2. Note that quantifying mentions is noisy: framework/library names can be spelled in a variety of ways, and names like ”Moses” are also used for authors in the ACL Anthology. For Figure 2.3, we normalize by lowercasing all text, and using the most common normalized spelling, e.g. `tensorflow`, or `huggingface`. We estimate that this will overestimate the presence of Moses, due to its other usages. and underestimate the presence of Hugging Face, which is officially spelled “Hugging Face”, but much more often used as “HuggingFace”. Despite this, the incredible growth and popularity of Hugging Face relative to other frameworks is still prominently visible.

We present an alternative view of data, similar to that seen in Figure 2.1, in Figure 2.4. Here, we define members of the community as authors with at least three papers total in *CL venues. Authors are counted as “leaving” the community the year after their last *CL publication. We only consider trends in authorship until 2020, as it is difficult to determine if authors who did not publish in the last few years have left the community indefinitely.

[8] As of October 2023, while some Findings papers, such as from ACL 2021 and EMNLP 2020, are automatically indexed by S2, Findings papers from some conferences such as EMNLP 2022 are not. While some of these papers (or versions of them) may still be indexed by S2 due to also being on ArXiv or a similar service, we do not include them in our set of papers.

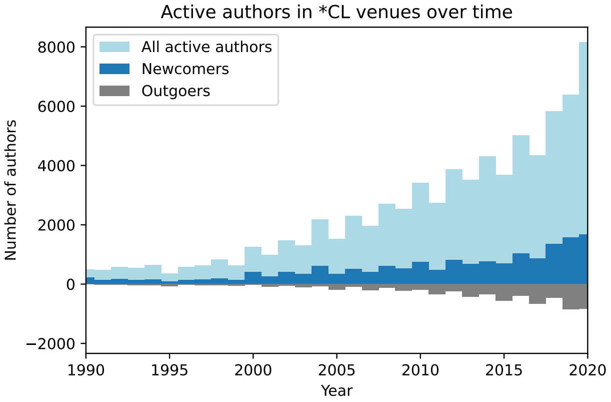


Figure 2.4. The number of “active” researchers publishing in ACL venues has increased dramatically, with more newcomers to the field year over year

Chapter 3

From Glamor-Proof to Proof of Glamor: the Evolving Role of Benchmarks in Natural Language Processing (In-Progress)

> We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.
>
> J.R. Pierce,
> *Whither Speech Recognition?*

3.1 Introduction

Static benchmarking has long been the standard instrument by which progress is measured in NLP. In the past decade, the prominence of benchmark results has risen — from what many consider NLP's "ImageNet moment"of pre-trained models besting previous approaches on many benchmarks to large margins, to current large language model providers relying on tables of comparative benchmark results to advertise their model's capabilities, to increasing feeling in the NLP community that benchmarks are the only type of result that stands on its own, as we discussed in the last chapter.

Even as benchmarks remain the predominant method to report progress, there has been persistent criticism from within the community. While most of these arguments recognize that the paradigm of benchmarking has value, the criticisms range widely in scope. On the more operational side, (Bowman & Dahl, 2021b) argue that benchmarks often lack statistical power, contain numerous annotatation errors, and ultimately may not reflect the constructs they purport to measure; Ethayarajh & Jurafsky (2020) argue that benchmarks and leaderboards fail to capture the interests of stakeholders interested in aspects of model behavior other than raw performance; Plank (2022) and others argue that the way we handle inter-annotator disagreement (which often involves discarding contested data points) is insufficient. More conceptually, Wallach *et al.* (2025) argues that benchmarks could afford to borrow from measurement theory in social science; Saxon *et al.* (2024) argue that benchmark performance often does not provide confidence in the generalized capabilities of a model, and call for a practice of "model metrology"to make model evaluations specific to use contexts; Raji *et al.* (2021) similarly, but more strongly, argue for the insufficiency of benchmarks to provide an inductive notion of model capability at all.

In the years since, evaluation processes for models have evolved: models are increasingly evaluated head-to-head in arena-style evaluations, and as models are increasingly consumed as services using web-based chat and APIs, model providers are doubtlessly running A/B tests on their users. Despite these evolutions, however, many of the critiques of static benchmarking apply equally, if not more, to these new techniques, which if anything are less grounded in specific users and their needs.

In this chapter, I trace the history of benchmarks in NLP to understand their original conception and how their role has shifted, especially in the era of neural NLP. I argue that while benchmarking as a paradigm is still valuable, that that value is primarily found in finitely scoped, specific evaluation for task performance, echoing Raji *et al.* (2021) and Saxon *et al.* (2024). I further argue that contemporary use of benchmark results, rather than specify the performance of a model and allow potential users to effectively reason about its applicability to their tasks, serves instead as a rhetorical move by which to justify the idea of *general capability.* This framing of general capability in turn simultaneously allows broad, indiscriminate marketing of models as tools and companions that are suited to any task, while at the same time allowing potential users of these models, like the military, to justify aggressive investment even without proof of efficacy.

3.2 What is a benchmark, and where do they come from?

In NLP, a static benchmark can be considered an assemblage of a *language task*, a *dataset* or *corpus*, consisting of labeled examples, that is taken to be representative of that language task, and a *metric* that is used to evaluate candidate models. Typically, this will take the form of training a candidate model on a specified "training"subset of the data, and evaluating its performance with the metric on the withheld complementary "test"subset. This benchmark will typically be presented alongside a *baseline* model, which represents a reasonable approach to the language task. New approaches will then be evaluated with the metric on the same test data as the baseline to determine which is the "better"approach.

With some qualification, this definition reflects benchmarks as they were initially constructed. In language technologies, benchmarks largely emerged in the low-trust environment of the mid-1980s, after over a decade of severely curtailed government funding in the 1970s, the so-called first "AI winter". This environment is widely attributed to two pieces of public criticism — 1966's ALPAC committee report (Pierce, 1966), which questioned the efficacy of funding into machine translation, and 1969's *Whither Speech Recognition*(Pierce, 1969), the source of the quote that begins this chapter. Both pieces reflected the vision of John R. Pierce, the coiner of the term "transistor", the chair of the ALPAC committee, and scientist at Bell Labs. While the committee report is relatively diplomatic, *Whither Speech Recognition* is a blistering critique of speech recognition, arguing both for the technical infeasibility of the task of generalized speech recognition given the technology at the time, but also for the unsystematic way in which people working on speech recognition technologies (whom he called "recognizers") conducted their work: "[the typical recognizer] builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment."It was this unsystematic process which leads him to label funders in speech recognition "suckers", taken in with the deceit and glamor of grand, unfulfilled promises of a technology that worked as promised.

This view coheres strongly with the experience of senior academic faculty that we interviewed for the work presented in chapter 1. Participants in that study described a culture of software demos being presented as the primary research artifact: *"We had toy systems. Um, we little toy demos that we recorded on video...that's that's how you would publicize your project." ()* These demos, from another participant's perspective, effectively established that a system pushed forward the frontier of computing, where *"the performance metrics were, 'oh, my god it does that? No machine ever did that before'...and that worked fine." ()* The also discussed a funding model that could understandably have provoked backlash from funders: *“”DARPA money used to go to CMU, MIT, and Stanford because they were the only ones with the computing power...I think I remember one of these blanket DARPA proposals, because it would just get passed around for people to add bullet points..DARPA would say, we're going to give you fifteen million dollars a year. You can divide it up among everybody. However you want...no deliverables, no statement of work.” ()*

Multiple sources (Church, 2018; Liberman, n.d.) argue that benchmarking culture, spearheaded by program managers at DARPA arises as a response to this style of work, which Church (2018) argues is “glamor-and-deceit-proof”. Liberman, in a followup talk at an ACL Workshop focused on Benchmkarking in 2021 [9], argues that there are three components that make up a benchmark: a detailed task definition and evaluation plan, which is developed in collaboration with researchers in the field, relatively inexpensive automatic evaluation code, written by a third party (the National Institute of Standards and Technology (NIST), in the case of early speech datasets) and published at the start of the project, and the combination of public training data and withheld, private testing data, such that each team has access to the same data with which to develop their system. These three components, combined with the public release of reports detailing the techniques used in the systems evaluated by benchmarks, allow Liberman to argue that benchmarks address Pierce's concerns that “no simple, clear, sure knowledge is gained.”

Liberman also outlines a fourth component that he argues makes benchmarks not only suitable for measuring the effectiveness of individual contributions, but also for influencing decade-scale research directions: metrics which are “directionally correct,”i.e. metrics on which improvement indicates a general increase in the ability of a model to perform a given language task. He notes that these directionally correct metrics are often “too crude to validate actual applications”.

These four factors, i.e. careful construction of evaluation, open training and withheld test data, public release of findings, and directionally correct metrics, transformed the field: the culture of repeated, frequent evaluation took hold, and the recognition of the role of large, public datasets led to the creation of institutions like the Linguistic Data Consortium (LDC), and the TREC series of evaluations in information retrieval. These shared resources allowed for progress in language technologies to be framed as improving metric numbers on shared tasks.

[9] https://youtu.be/a-ukPup8gKw

3.3 An Evolution in Five Benchmarks

In this section, we trace the evolving role of benchmarks in language technologies across five instantiations: the Penn Treebank (Marcus *et al.*, 1993), ImageNet (Deng *et al.*, 2009) [10], GLUE and SuperGLUE (Wang *et al.*, 2018, 2019), MMLU (Hendrycks *et al.*, 2020), and τ-Bench (Yao *et al.*, 2024). These benchmarks cover roughly 25 years of NLP history, from the early 2000s to the present. Across this time period, I will discuss the real progress that these benchmarks enabled, alongside three categories of issues: problems within the benchmark paradigm, i.e. where adherence to the four principles from the previous section degrades; problems where the benchmarking paradigm itself fails to measure progress; and problems where the benchmarking paradigm itself makes more difficulty or even precludes alternative kinds of work.

3.3.1 The Penn Treebank (Marcus *et al.*, 1993)

The Penn Treebank (Marcus *et al.*, 1994, 1993, PTB) is a corpus of English text annotated with linguistic structure designed as a benchmark for four tasks: part-of-speech tagging, syntactic parsing, and later for tagging predicate argument structure in text and identifying speech disfluencies. Constructed from a variety of sources, including the Wall Street Journal, IBM computer manuals and the Brown Corpus (Francis, 1964), the corpus consists of millions of words of tagged text across the four tasks (Taylor *et al.*, 2003). The Penn Treebank was one of the first large-scale annotated corpora for language, and as such, occupied a central role as one of the primary research corpora in NLP; multiple interview participants recall better numbers on parsing being a focus of the field for over twenty years.

PTB stands out among the corpora in this section for the degree of specification and quality in its annotations. The creators of the original corpus discuss at length the process of designing the tags that annotators (all with graduate training in linguistics!) applied, and how they differ from existing tag sets at the time; they additionally verify that the mode of annotation used (i.e. correcting automatic annotations, rather than annotating from scratch) produces a corpus that is more consistent among annotators, more correct, and faster to annotate. Over time, the dataset also developed shared conventions for which metrics to use for each task and how to segment the data into train and test splits. The data is made available through the Linguistic Data Consortium, which uses membership fees to cover the cost of hosting and distributing these datasets.

[10] While ImageNet itself is not a language benchmark, it nonetheless cast a long shadow over languge technologies research, and framed much of NLP progress until early pretraining approaches as "the search for NLP's ImageNet moment"

As such, PTB represents a fairly faithful vision of the four principles of benchmarking that Liberman presents. The task is clearly specified and the data is annotated and checked by experts; the data, while not entirely free, was widely available; evaluation standards were largely set and followed by the community, and a robust literature grew around pushing metrics on the test set up. Even almost twenty years on, one of our interviewees recalled the specific setup of the time: *"Try to get the higher F1 on Wall Street Journal Penn Treebank section 23, you know, parsing, and then, once you have that number, it doesn't even matter what you write in the rest of the paper." (12)*

PTB also represents the success of directionally correct metrics at the time. Though the data is only in English, collected from highly specific domains, there was an understanding that systems designed for doing well on PTB tasks were primarily research artifacts that might not translate into practical use. While components or techniques from these systems might otherwise be developed into specific, user-facing technologies, the systems themselves did not themselves represent a colloquiual notion of "language understanding."This gestures at a limitation of directional metrics, however — they support a notion of progress that is more about developmental *potential*, rather than the capability of the system being evaluated. In effect, there is an implicit acknowledgment of the insufficiency of the collected dataset and metric to truly represent the underlying task that the benchmark. In the Penn Treebank's period of popularity, this did not represent a problem, as the artifacts were not intended to be accessed by audiences that would not understand its limitations. As one interview participant put it, *"It's not like anyone ever went to like parser.yahoo.com to like to run a parser on something" (3)* in the way that they do with language models today. In a sense, the success of PTB as a benchmark relied on models effectively not "working"as what would colloquially be understood as language technology.

By the measures that Liberman sets out, PTB was a success - it could reliably be funded year-over-year, and followed the pattern of "error rates declining by a fixed percentage.". Perhaps more importantly, though, PTB can be seen as a success because it allowed for some degree of social coordination at a time when the ACL community had already started to grow exponentially. PTB's popularity ranged over a time perid where ACL conferences went from less than a hundred researchers and a single track to conferences that were several hundreds or thousands of researchers with exclusively parallel tracks, transitioning from a high-trust environment where "everyone knew each other", to one where discussions of abstract ideas like "progress"needed to happen in between the lines of papers and funding proposals, rather than in small group discussions. In a field that was rapidly expanding, benchmarks also provided an easy way to make progress legible: a significant portion of the work of a field could be summarized in a chart that displayed increased benchmark performance.

From a researcher's perspective, this allowed the scope of problem they worked on to be narrow and well-defined; it enabled what people now *hill climbing*, a term that refers to an optimization that seeks to find the highest value of an arbitrary function through iteratively making small changes and keeping those that cause an increase in the function's value. The analogy to an optimization technique speaks to the degree of specification of the problem: a function that can be optimized with hill climbing can only optimize a single value, to the exclusion of alternate criteria. However, the analogy is also illuminating as to the part the researchers play in it — even though they work at different parts of the problem in different ways, their incentives are nonetheless cleanly aligned.

3.3.2 ImageNet (Deng *et al.*, 2009)

ImageNet (Deng *et al.*, 2009) is one of the best known benchmark datasets in existence. Initially released in 2009 (before several subsequent revisions), ImageNet is a dataset designed for the task of object recognition in images across a large number of categories. The categories themselves are a subset of WordNet (University, n.d.), an ontology initially released in the 1980s that defines semantic and hierarchical relationships between conceptual categories of words that they call synonym sets or "synsets". ImageNet was constructed by sampling a subset of synsets (and the related tree structure), searching for online images using a variety of search engines using primarily the words in the synset, and then having crowdworkers verify that the image matched the synset. A version of ImageNet (with a pared down set of categories) was used as the basis of a longstanding competition in computer vision - the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

ImageNet is obviously not a language benchmark — it was designed and used in the computer vision community. However, ImageNet casts a long shadow over much of machine learning because of the so-called "ImageNet moment,"in which a neural network called AlexNet achieved an unexpectedly large performance jump over more popular computer vision techniques of the time. This resulted in the resurgence of neural networks as a research topic, and the common refrain of asking when other fields would have their ImageNet moment, a term that has become common parlance.

What sets ImageNet apart from most of its contemporary datasets is primarily its scale. At the time of the first paper, ImageNet boasted of over 3 million annotated images across 12 subtrees of WordNet and 5427 synsets; they state their aims as being on the order 50 million annotated images, or 500-1000 images per synset. Subsequent ImageNet releases, while not reaching this ambitious goal, still create a massive collection of images, with one more recent release (ImageNet21k, Winter 2021 version) containing over 19,000 categories and over 13 million images (Luccioni & Crawford, 2024). As the original paper points out, even the initial 3 million image scale is two orders of magnitude larger than other common datasets in use at the time. This scale is viewed as one of the key factors that made neural nets a viable approach for computer vision for the first time, while at the same time making more labor-intensive forms of computer vision less workable.

This scale, however, comes with a tradeoff in how much curation of the dataset was possible. In contrast to Penn Treebank, whose annotators all had graduate-level training in linguistics, ImageNet relied on the now-common practice of using crowdworkers to narrow down from candidate images to images included in the final dataset. Given the setup of how the dataset was collected — i.e. by taking a sample of the classes from WordNet without modification, so as to enable the use of crowdworkers to merely verify the images' categorization, rather than to find images representative of that category — ImageNet also inherits issues in the WordNet ontology. These flaws varied widely, from categories that did not reasonably have a visual representation, such as "monolingual", to several more that were offensive, as extensively documented in Crawford & Paglen (2019); Luccioni & Crawford (2024); Prabhu & Birhane (2020). Open datasets, as in the case of WordNet, are often held up as resources that have myriad benefits outside of the intended purpose of the dataset, but especially in cases where the dataset is large or detailed, they also force downstream users to adopt their own assumptions and practices, with little opportunity to opt out. As the expectation of scale from the benchmark datasets we use grows, the scope of a project to create a new dataset from scratch becomes more and more infeasible, leading to the reuse of datasets that may be poorly suited for the new framework.

ImageNet is also an early representative of a problem with directional metrics. As a dataset, ImageNet has several issues with relation to the task that it serves as a proxy measurement for, i.e. recognizing the objects in images. Firstly, ImageNet makes several simplifying assumptions in its annotation process. Most prominently, the annotation process does not produce a comprehensive list of objects in the image, partly as an artifact of the way in which the dataset was created, going from a desired object to candidate images. This resulted in efforts across the computer vision community to develop mutli-label versions of ImageNet. Secondly, later analyses of ImageNet found issue with the initial annotation setup, arguing in the case of Northcutt *et al.* (2021) that almost 6% of images in ImageNet were mislabeled. Thirdly, the project of object detection in images is inherently underspecified and almost necessarily contested. As anyone who has needed to complete a CAPTCHA can attest, the question of whether an image includes enough of an object to count (is just the red light a big enough part of the traffic light?), is hard to answer definitively, let alone to communicate across a diffuse population of crowd workers.

None of these problems are existential for ImageNet's use as a benchmark — if ImageNet's metrics are truly directional, then it is a defensible position to claim that progress on ImageNet, as flawed as it is, still represents progress towards the underlying task of object detection in images, even if a model that performed perfectly on ImageNet would likely not be appropriate as a drop-in component to a downstream use that required object recognition, as outlined in papers such as (Tsipras *et al.*, 2020). Belying this claim, however, is the long history of widely-used models either trained on ImageNet or that are based on models trained on ImageNet (as partially documented in e.g. Luccioni & Crawford (2024)). In effect, ImageNet's utility is not only as a measurement tool — models trained on ImageNet are also seen as powerful models which have some inherent visual faculty from which new models can be trained, whether for other computer vision benchmarks (e.g. Donahue *et al.* (2013)), or for real world use, in ways that cement its issues into any deployed downstream use of it. While it can be argued with some qualification that models that achieve state-of-the-art performance on ImageNet still advance the project of object detection (see e.g. Shankar *et al.*, 2020), this is only a reasonable argument insofar as it acknowledges the narrowly scoped claims that even a dataset as large as ImageNet can hope to support, given the numerous issues with data curation and quality, in addition to the well-documented biases that that curation process propagates down into models trained on the dataset (Paullada *et al.*, 2021). ImageNet has outgrown its status as a benchmark as defined by Liberman — partly because the metrics used to measure models are no longer treated as directional, but as measurements of the models' actual capability on the underlying task.

3.3.3 GLUE (Wang *et al.*, 2018) and SuperGLUE (Wang *et al.*, 2019)

NLP's ImageNet moment finally arrived, several years later, in 2018. As researcher Sebastian Ruder argued (Ruder, 2018), pretraining methods like ULMFit (Howard & Ruder, 2018), ELMo (Peters *et al.*, 2018) and OpenAI's original GPT (Radford & Narasimhan, 2018) achieved for NLP what AlexNet and related approaches had achieved with ImageNet: a large, discontinuous jump in performance on existing benchmarks using neural methods that produced a base model that could then be built upon for other, downstream applications. At the time, benchmarks in NLP focused on "multi-task"learning, in which a single model architecture was evaluated on multiple disparate language tasks. These benchmarks, exemplified by GLUE, therefore were a well-suited platform to demonstrate the effectiveness of pretraining for language tasks, i.e. to argue that pretraining gave models some type of "general"language capability that could then be refined by fine-tuning into strong performance on specific tasks. BERT(Devlin *et al.*, 2019), perhaps the best known of the pretrained models from this time period, is introduced in an abstract that highlights benchmark scores (including GLUE) as its major advancement.

GLUE itself is not one benchmark, but a collection of several existing benchmarks in three categories — single-sentence classification tasks, similarity tasks, in which the model detects a paraphrase, or scores similarity between two sentences, and and inference tasks, in which a model is presented with two sentence, and must determine the relationship between those two sentences. These tasks were, at the time of GLUE's construction, intended to be a stable, directionally correct benchmark; contemporaneous models displayed "low absolute performance". However, within two years of its release (primarily because of pretrained models), state-of-the-art GLUE benchmark scores reached human parity, necessitating the release of a followup dataset, SuperGLUE, which followed the same pattern of combining multiple benchmark datasets into a single-number measurement of model performance, but with harder tasks in more formats. It's also worth noting that while GLUE and SuperGLUE both collapse model performance across ten tasks into an unweighted average score of all the subtasks scores in order to rank models, both of them also present a *diagnostic dataset*: a subset of the dataset that is annotated for linguistic features rather than just a final label, to enable more fine-grained analysis of where models succeed and fail.

Given that GLUE and SuperGLUE benchmarks accompanied NLP's ImageNet moment, it's perhaps unsurprising that many of ImageNet's problems with data quality manifest in GLUE as well. As an example, let us consider Multi-NLI or MNLI (Williams *et al.*, 2018), one of the inference-type tasks discussed above. The task setup of MNLI is that of a classification task which has as input two sentences, a "premise"and a "hypothesis."The expected output is one of three labels: "entails", indicating that the hypothesis logically follows out of the premise; "contradicts,"indicating that the premise and hypothesis cannot both be true, or "neutral,"indicating that neither of the above two relations applies. MNLI is an updated version of an earlier corpus with the same setup, SNLI (Bowman *et al.*, 2015), but diversifies it by pulling from more genres of text. While the setup of this task is simple, the range of language abilities required to perform this task is wide — as the MNLI authors state, to perform well at natural language inference, "a model must handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity."

While that is true of the abstract *task* of natural language inference, the reality on MNLI is often much different. Poliak *et al.* (2018) revealed that because of the way the dataset was collected, i.e. by providing annotators with a premise, and asking them to write hypotheses for each label in turn, oftentimes models that ignored the premise entirely could achieve competitive performance; the hypotheses written by annotators tended to follow patterns that models could pick up on. Further, McCoy *et al.* (2019) demonstrated, by way of a synthetically constructed dataset, that models that succeeded on MNLI, instead of actually learning the skills the authors argued would be necessary, instead learned shallow heuristics that also exploited characteristics of the training dataset. More fundamentally, Gubelmann *et al.* (2023) argue that many of the premise instances in MNLI are not in the correct pragmatic category to support inferences, e.g. questions or sentence fragments. This is easily seen in examples in the dataset such as the premise "Issues in Dataset,"which, in not being propositional, does not entail or contradict anything.

MNLI also shows issues with annotator consistency, in a conflict between what (Schlangen, 2021) calls the intensional definition of the dataset (i.e. the descriptions of what would fall in each class) and the extensional definition of that same dataset (i.e. how classes are defined in the examples of the dataset). (Pavlick & Kwiatkowski, 2019), for instance, show that populations of annotators on NLI even when reading the same prompts are sometimes not well-modeled as a unimodal distribution centered over the same "correct"answer, but can exhibit distribution patterns that indicate support for multiple competing interpretations of the instructions. To their credit, the designers of MNLI attempt to account for this type of variance by having each example's label be the majority vote of five labelers - the original annotator who generated the hypothesis, and four more labelers, but they instruct users to discard any examples that do not reach the three-vote consensus. In the corpus, 88.7% of annotations agree with the final label chosen for the example, and 58.7% of the examples have a unanimous result for the final label, and all the annotator labels are preserved.

More fundamentally, however, this annotation process raises the question of what definitions — and whose — are operationalized in these datasets, and how to account for valid but conflicting disagreement in datasets. As discussed above, this is already an issue for tasks like object recognition, which faces real ontological issues despite being relatively easy to specify; the challenge for language datasets like NLI, or even sentiment classification is much more acute, given the additional subjectivity involved in those tasks. The practice of discarding annotated examples that annotators disagree on is common, and high inter-annotator agreement is often reported as a measure of the "cleanliness"of a dataset. As Fleisig *et al.* (2024); Plank (2022, inter alia) point out, however, variation between annotators should not solely be seen as noise, but rather potentially valuable signal that reflect different understandings of not only what a task is, but normative notions of what a task should be. Processes like majority votes and filtering do not ultimately reflect the variety of valid perspectives on a given instantiation of a language benchmark.

Once again, were we to treat metric results on these benchmarks as purely directional, this would not necessarily be a problem: the claim being made could be argued to be arguing that the model architecture under evaluation is capable of learning *some* notion of the task, given the particular data, task definitions, and annotators. However, this argument rings increasingly hollow for two reasons. Firstly, as in the case of models being trained on ImageNet, models trained on these benchmark datasets are commonly used in downstream applications that often rely on a conflation of benchmark performance and task performance to be viable. Models trained on NLI, for example, are commonly repurposed into models to do zero-shot[11] text classification (Yin *et al.*, 2019), or to verify journalistic or scientific claims (Milbauer *et al.*, 2023). Only much later are issues pointed out with this type of leap (Ma *et al.*, 2021; Mor-Lan & Levi, 2024, respectively). Benchmarks like MNLI (and by extension, GLUE) are seen as indications of model capability, rather than measurements of models on individual datasets.

Secondly, taking a step back from MNLI to GLUE itself, collections of benchmarks are also making an inductive claim: that this collection of benchmark evaluations is an effective proxy of the underlying, "general-purpose"ability of a model to perform "language understanding"writ large. Raji *et al.* (2021) identify and rebut this inductive claim, noting that despite the veneer of "general-purpose"capability these benchmarks intend to measure, they are nonetheless still English-only, narrowly scoped and somewhat arbitrarily chosen tasks that fail to account for the necessarily contextual nature of model deployment. In an illustrative example, they quote (Wagstaff, 2012) who discusses the difference in accuracy needed for a classifier for iris species, as in the famous dataset, and for foragers looking to classify mushroom toxicity.

Raji *et al.* (2021) also note the consolidating effect that benchmarks like ImageNet or GLUE have, in directing progress in a field towards the types of problems that the benchmark represents. In the age of large, pretrained models, however, this issue gains a new dimension. As the top-performing models across datasets are increasingly large models that only a few actors have the resources to train, they are seen as a base on which subsequent developments *must* be built, rather than developing new approaches, which might not achieve good benchmark performance right away. Benchmarks therefore contribute to a software lottery, as we identified in (Gururaja *et al.*, 2023), wherein approaches that are successful on benchmarks occasion further development of infrastructure along that approach's lines, and foreclose alternate development possibilities.

[11] A setting in which the model is given no training examples before being asked to perform the task

3.3.4 MMLU (Hendrycks *et al.*, 2020)

Since BERT, there has been a shift in the research landscape of NLP models: where BERT was published following a fairly standard academic process — details were published in a paper at a high-prestige conference, the paper had enough detail for the work to be replicable, the model weights were widely shared, etc. — models since then, especially after the release of ChatGPT, have looked different. A typical model release at this point is accompanied by a blog post, rather than a formal paper release, details are often alluded to, but strategically hidden for competitive advantage, and while open weight models exist, they are no longer the norm for model releases. Instead, companies — for most of the institutions releasing cutting edge models today are companies — often make their models available through an application programming interface, or API, in which they tightly control what information is available from a model: users may or may not know technical details of how the model was designed or trained, may not have access to the full range of affordances that a model provides, or even know whether it is a single model or several different models in a system that they are using. Benchmarks such as MMLU (which stands for Massive Multitask Language Understanding) occupy a central role in release blog posts for recent models: after a few illustrative examples, blog posts will include a table of benchmark results [12] on which the released model will outperform all of its nearest competitors.

Despite this partial improvement, however, the role of benchmarks must be questioned, for the "models"released by these companies are only questionably research artifacts — the APIs through which the models are provided are often a primary driver of the revenues of the company that releases the model. Even in the case of "open-weight"models, models are typically published in standard formats to allow new models to be dropped in to existing modeling pipelines and even products. In other words, these models are intended to be used in practical contexts, rather than just being objects of study. For models that are intended to see a variety of highly contextual, real-world uses, however, MMLU is an interesting choice. While it is drawn from a wide variety of domains (57), it is a collection of just under 16,000 multiple choice questions, each of which have four answers, and where the correct answer is guaranteed to be one of the four. This stands in stark contrast to the way in which contemporary language models are presented as user-facing software: as open-ended, chat-based interfaces that are intended to do any possible task. While multiple

[12] An interesting note here is that this type of benchmark has returned in one notable way to the original benchmarking paradigm: modern language models are typically evaluated with a few-shot, "in-context learning"approach, in which the models, rather than being trained on the benchmark, are instead given a few representative examples as part of their prompt. In this sense, the inductive capability argument is partially addressed: rather than a particular model that *can* be trained for all of the individual benchmark tasks, the same language model is in fact evaluated on each benchmark without further retraining.

choice questions can serve to measure domain knowledge and language understanding, it's important to distinguish between the types of static understandings of performance that benchmarks can provide, in contrast to the vast promises of model makers.

MMLU also exhibits issues with validity, as in the case of earlier benchmarks. Gema *et al.* (2025), for instance, find significant errors in multiple subsets of MMLU, estimating that over 6

More concerning than annotation issues, however, is the use of the multiple choice format itself: additional work has found that the multiple choice question format, in addition to being mismatched with the actual use cases of these models, in fact introduces new, particular issues to evaluation: Li *et al.* (2024c) and Pezeshkpour & Hruschka (2024), for example, independently find that LLMs are sensitive to the order in which the options in a question are ordered; Chandak *et al.* (2025) find that, similar to NLI, MCQs in popular benchmarks can often be answered without access to the original question, based on the appearance of the answer options. Balepur *et al.* (2025) argue along multiple of these axes — misalignment to use cases, being insufficient to test knowledge, format affordances providing shortcuts — before arguing that theory from education can support more informative use of the MCQ format. Both Chandak *et al.* (2025) and Balepur *et al.* (2025) additionally argue for more generative evaluation of question answering (i.e. where answers are generated freeform.)

Underlying many of these concerns is also a much more existential threat to benchmark-based evaluation: data contamination. Given that the vast majority of models (including putatively "open-source"models) are released with little to no public account of what data was used to train them, let alone access to those datasets, it is nearly impossible to verify that models have not been trained on data contained within these benchmarks, a case that would result in their performance being overestimated significantly. This represents a real breakdown of the principle of withheld test sets, arguably the most important pillar of the benchmarking paradigm. Without this, establishing benchmarks as a useful proxy for model performance is on shaky epistemological grounds at best, and at worst gamifies model measurement beyond utility.

It's worth noting that at this point, MMLU itself is considered an outdated benchmark, and performance on it is not typically reported in new model release announcements. The task paradigm, however, remains ubiquitous. In addition to the updated versions of MMLU discussed above, other tasks that rely on the same multiple choice format with similar levels of uncertainty about contamination dominate. In the recent Gemini 3 release post [13], for instance, Google reports performance on GPQA Diamond (Rein *et al.*, 2023), MMMU-Pro (Yue *et al.*, 2025), an updated version of MMMU (Yue *et al.*, 2024), itself a multimodal take

[13] https://blog.google/products/gemini/gemini-3/#gemini-3

on MMLU, VideoMMMU (Hu *et al.*, 2025), and MMMLU (OpenAI, 2024), a multilingual translation of MMLU.

3.3.5 τ-Bench (Yao *et al.*, 2024)

While multiple choice-style evaluations remain common practice, criticisms of their misalignment with the real use cases for language models have prompted the development of benchmarks that aim to measure models' and systems' performance on tasks that are more closely related to "real-world"tasks, in step with the transition to AI systems being developed as "agents"that are autonomously capable of fulfilling tasks, rather than tools to be prompted and guided in specific ways by the user of the systems. τ-Bench (Yao *et al.*, 2024) is one such example, that aims to assess the ability of models to respond to real-world customer service queries across two domains, airlines and retail. More granularly, the benchmark assesses a model's ability to interact with a simulated "user"(in reality, another model), and reach the intended "end state"of the conversation by interacting with and producing a database state that is identical to the expected outcome. Notably, for a benchmark that aims to study "real-world domains", the data in this benchmark — from the allowable actions that an agent may take, to the data that it operates on — is entirely synthetically generated, either manually or with the aid of another LLM.

This design choice illustrates a key difficulty of designing more difficult benchmarks for evaluation. The ability to measure the capability of a model on a particular task is directly limited by the ability of verifying the correctness of what the model does. To some extent, this has always been a problem with benchmarking, which typically only verifies that a model's output is correct, and does not establish that a model learns a reasonable or correct procedure for carrying out the task. In the MNLI case from above, for instance, models idenfity heuristics as in (McCoy *et al.*, 2019), which allow for high benchmark performance without similar performance on other instances of the same task. In more complex domains, however, verification of the expected answer itself is difficult. Zhu *et al.* (2025), for example, show results on τ-Bench can be overestimated as much as 40% by allowing models to either guess or provide a (correct) empty answer, which confers no predictive power.

Zhu *et al.* (2025) also outline further ways in which the evaluation methods for other agentic tasks also present difficulty in ways that run up with measurement of these tasks in the real world. In discussing SWEBench (Jimenez *et al.*, 2024), for instance, a collection of tasks that aims to measure models' abilities to resolve issues posted to public repositories on GitHub, they point out that many of the model-proposed solutions pass software tests that award points on the task, while leaving the underlying problem unsolved. This is a well-known problem in software engineering; verifying code correctness remains an unsolved problem.

This difficulty — of verifying the correctness of complex or abstract generative tasks — has accelerated the trend of using LLMs as judges, i.e. passing example instances to another model to be scored, rather than using deterministic metrics, which can miss subtle variation in answers, even if semantically equivalent. However, as Bavaresco *et al.* (2025) demonstrate, LLM-based judges vary significantly in performance, even though they are reliable in some cases, and (Thakur *et al.*, 2025) demonstrate that judge models often have unpredictable biases on their performance from factors including response length, the number of instructions, and an overall tendency towards leniency. Both papers recommend caution or evaluation before applying LLMs as judges, but this adds additional layers of evaluation and its attendant problems.

Difficulties in evaluating tasks as they trend closer and closer to real world tasks can also be thought of as a problem of leaving the zone of directionally correct metrics at a time when the technology appears to be capable of solving real problems, instead of narrowly scoped problems that can be effectively measured. Instead of a regime in which the metric is understood to be flawed, but progress on the metric represents progress on an underlying task, current evaluation methods are trapped in an awkward position where the ability of models to produce plausible output outstrips our ability to distinguish plausible from correct at scale.

3.3.6 Where does that leave us?

In this section, I trace through five benchmarks that are representative of the evolution of benchmarks from roughly the turn of the century until the modern day. Across that time, the size, scope, intent, and ultimately the role of benchmarks has shifted significantly: from highly specialized, carefully defined annotation on relatively small datasets for a technology that was seen as being years away from language understanding, to large datasets whose flaws often lead to years of documentary research, built for technology whose capability is hard to disentangle from intuitive notions of language understanding. Underscoring this, however, are the three factors mentioned above: the imperfect use of (static) benchmarking by the community, the ways in which benchmarking is insufficient to measure the qualities of models that might be important in deployed contexts, and how benchmarking as a cultural practice can often preclude alternative ways to think about how models can be built

As I have argued in the previous section, each one of the four principles that underlie the success of benchmarks has been successively eroded as benchmark culture has progressed. Careful construction of shared evaluation metrics has given way to a flood of benchmark datasets of widely varying quality; the principle of withheld test data is harder and harder to assert, given the increasing centralization on proprietary models that release no accounting of their training data; research labs that release models increasingly decline to publish ongoing research. Directional metrics are also increasingly poorly constructed for a world in which models are no longer merely research artifacts to be developed into products, but the products themselves.

Wallach *et al.* (2025) frame this issue as one of an insufficient understanding of what it means to do measurement in the AI context, and argue for borrowing constructs from measurement theory in the social sciences. They argue that AI practitioners must reason more explicitly about the conceptualization and operationalization of the constructs they purport to measure, while also considering different types of validity of these measurements. This framework addresses many of the critiques of benchmarks as constructed today collected in this section.

However, static benchmarks, even were they developed with much more of an eye towards validity, are still constrained by the scope of the data collected, the purpose of that collection, the identities of the annotators, and more. Further, as (Raji *et al.*, 2021) argue, no collection of benchmarks can be sufficient to establish a true idea of the "capability"of a model. This insufficiency can be seen in the wild: even as model developers release larger and larger tables of benchmark results, new paradigms like arena-style evaluations proliferate, often even more poorly grounded in an understanding of concrete tasks. To address this, Saxon *et al.* (2024) call for a new discipline of *model metrology*, in which developers of specific applications build benchmarks that are directly relevant, i.e. *ecologically valid* to their task of interest, instead of general-purpose benchmarks.

Benchmark-based evaluation also serves to narrow the scope of work that the field finds acceptable. Church (2017), for instance, worries that "literature may be turning into a giant leaderboard", echoing one of our interview participants finding it "difficult to publish papers in CL that have ideas in them"(19). Ethayarajh & Jurafsky (2020) also find, for instance, that benchmarks often exclude considerations of other aspects that users of benchmarked models might care about, including computational efficiency or environmental impact. Orr & Kang (2024) also note the ways in which benchmark evaluations both derive from and contribute to a culture of competitive evaluation that undermines the original goal of the benchmarking, using an evocative example of conduct in a cricket match seen as unsportsmanlike and subsequently banned.

3.4 The role of benchmarks today

It's clear that even as benchmarking enabled the funding and progress of AI research for multiple decades, that its utility is increasingly unclear. Critiques of benchmarking and the culture it engenders abound from both inside and outside the field, and even practitioners tire of its pervasiveness. Given all of the qualifications to their effectiveness, why do benchmarks persist in their popularity?

Perhaps the simplest explanation is their initial, early success: benchmarks enable social coordination across a field that has grown multiple orders of magnitude over the past two decades. Procedures of evaluation, especially those set out in concrete terms as code libraries and published datasets allow for a degree of standardization and interoperability that allows participants in the field to acculturate relatively quickly into a field that publishes more than can reasonably be read. However, loose cultural norms mean that benchmarks are not developed or deployed in ways that might make them effective as measurements. Given the lack of centralization, benchmark evaluation *itself* is seen as the necessary component, with less regard towards the benchmarks themselves, leading to an explosion in the number of benchmark datasets created. The NeurIPS Datasets and Benchmarks track, for instance, has seen remarkable growth, from 445 submissions in 2022, to 987 in 2023, to 1820 in 2024 and 1995 in 2025 (Chairs 2025, 2025). While an optimistic view of this is to argue that AI research is finally coming around to doing the data work (Sambasivan *et al.*, 2021b), these conclusions only hold if these datasets are being designed with a degree of care and and attention to downstream use that is atypical of the community. As such, benchmarking represents at best a degenerate solution to the social coordination of a research field, causing a field's worth of research to converge on a solution that few are actually happy with.

In Latour (1986), Latour and Woolgar argue that the process of constructing scientific fact can be seen as an effort to raise the cost of alternative statements — that through labor, equipment, and various processes of inscription, the cost of running an alternative experiment or advancing a rival theory become too demanding to be practical. Benchmarking can be seen through this lens as raising the cost of alternatives in multiple ways. In the most literal reading of Latour and Woolgar, benchmarks make alternate paths of development more difficult when a method is shown to be successful on benchmarks. This is not a new phenomenon; John Makhoul recounts how an early speech benchmark led to Carnegie Mellon researchers switching almost entirely from knowledge-based methods to hidden Markov models in the course of two years, from 1986 to 1988 (Makhoul, 2021). However, as multiple sources argue, the practice of benchmarking itself is biased towards certain types of system. Makhoul, in the same talk argues that benchmarks spurred the use of trainable systems, which by virtue of needing less human work than knowledge-based systems between updates could take better advantage of rapid evaluation; lower-cost evaluation favored more quickly developed methods that could then be repeatedly tested. (Koch & Peterson, 2024) argue that this marriage of rapid testing and rapid development, while productive in the 1980s, set the stage for a monoculture of benchmarking with the advent of deep learning, which advanced trainable system philosophy by further reducing the space of human intervention from manually engineered features to a much narrower space of designing the architecture for models. Benchmarking thus imposes higher costs not only on methods that underperform on benchmarks, but also on entire classes of methods that are less suited to the methodology as a whole.

This unequal effect of benchmarks — prioritizing rapidly developed systems is exacerbated by the degree of centralization that the modern machine learning ecosystem exhibits, as we note in (Gururaja *et al.*, 2023). In successive developments since the advent of neural NLP, the NLP community has increasingly centralized on individual frameworks, libraries, and models, all of which are designed with the benchmarking paradigm in mind. Much of this work is ultimately underwritten and supported by industry actors, who both build the algorithmic and technological frameworks that NLP technology is built on in the form of frameworks such as PyTorch (Paszke *et al.*, 2019b), and also control the resources necessary to build models at the scale that modern degrees of benchmarking progress requires.

Koch & Peterson (2024) also argue that the rise of benchmarking correlates with industry's importance as an actor in AI fields. They cite industrial competitions such as the Netflix Prize (Bennett & Lanning, 2007) as early pushes towards benchmarks and leaderboards, but also discuss how neural networks and trends towards scaling have allowed industry to capture and direct the focus of the field towards industry's own goals. This occurs at multiple levels: industrial actors (including academics jointly affiliated with industry) are frequently the designers of popular benchmarks (Koch *et al.*, 2021), unaffiliated benchmarks that aim towards "real-world"use cases often skew, based on data availability, towards industry priorities, as in the case of τ-bench, and industry actors who release cutting edge models are the major determiners of what counts as an important benchmark, which allows competition among industry labs.

But as Saxon *et al.* (2024) note, benchmarks are poorly suited to actual industrial goals: *"These platforms provide API access to these models as a paid service, so why aren't they benchmarking customer-relevant capabilities?" ().* It's clear that companies at this point do have internal measurements and benchmarks; occasionally, a company will gesture to a model that they release as not being focused on benchmarks, as in the case of Sam Altman's tweet that GPT-4.5 would not "crush benchmarks"[14]. If benchmarks do not model problems as corporate actors see them, however, why is every model release accompanied by a large table of benchmark results?

[14] URL

There is a range of answers to this question. Most charitably, we could argue that the specific benchmarks chosen by AI companies are still seen as close proxies to task performance. I argue, however, that benchmarks are being used for the precise inductive claims that Raji *et al.* (2021) argue against — in effect, that benchmarks are being used as a proxy for general purpose capability. This rhetorical move serves several purposes across the AI ecosystem. For the industry actors themselves, publishing state-of-the art benchmark numbers allows them to continue to claim "progress,"in opposition to other actors and in turn allows them to market their AI tools as the most useful, regardless of the benchmark's degree of correlation to actual end-user tasks.

As Blili-Hamelin *et al.* (2025) also argue, collections of benchmarks aimed at proving "generality"also allow industry actors to distance themselves from specific scientific and societal decisions about how and where their tools should be used (what they term "generality debt") and offload the work of proving efficacy or answering ethical questions downstream, while leaving the entire production process of a model that causes several of these issues unchanged. Benchmark tables also impact the funding landscape, both for the industry labs themselves, but also for other actors in the field. As we explore in (Widder *et al.*, 2024a), military funding for AI research often rests on claims of AI's general capability in order to justify funding priorities; AI has rapidly become a technology that new funding proposals must be compared against.

These factors complicate the role that benchmarks play today — rather than simply being the only form of social coordination that can apply to a rapidly growing research community, we must also reckon with the unintuitive ways in which benchmarking bends AI research towards corporate ends. As benchmarks are not understood to have scope and applicability, as their results are unsupportedly generalized, as they inform wide swathes of social decision making, we should come to understand that rather than being glamour-proof, benchmarks are becoming proof of glamour.

3.5 Conclusion

In this chapter, we have traced the evolutions by which benchmarking has both failed and exceeded the scope of its initial paradigm: benchmark datasets are no longer constructs that produce "simple, clear, sure knowledge", especially as AI systems increasingly exceed our ability to measure them in static contexts.

However, as the numerous sources cited in this article indicate, extant critique from both within and outside the field identifies and engages with many of these concerns. Despite this, and due to a confluence of factors, benchmarks remain by far the most common way to measure AI systems.

Echoing Blili-Hamelin *et al.* (2025); Saxon *et al.* (2024), and others, we argue that evaluation of models must become closer aligned to the actual use contexts of these models, and that there must be a greater pluralism in what counts as NLP progress. As for benchmarks, there must be a more scoped understanding of what a benchmark can and cannot do. In the initial phases of the development of benchmarking practices in language technologies, there was a clear understanding that a benchmark performance demonstrated not ability on a task, but an approach's potential. Just as the early developers of benchmarks understood that a strong benchmark result would take development to become a practical, user-facing technology, so should we, such that language technologies can serve diverse cohorts of users with diverse needs.

Chapter 4

Characterizing Expert Needs in Document Research (Complete)

Modified from a paper published in the 2025 Findings of the Association for Computational Linguistics (Gururaja *et al.*, 2025a).

4.1 Overview

In the previous chapter, we argue that benchmarks insufficiently capture model capabilities, and facilitate a concentration of power in primarily industrial actors. In this chapter, we proceed to interview 16 domain experts in materials science and law and policy, to understand their processes for engaging in document research, and whether the the state of NLP technologies – including language models developed and and pushed by those industrial actors – accurately model or address their processes. We find a tradeoff between models that engage in the deeply social processes of reading that our experts engage in and models that are publicly accessible, and call for models to be accessible, personalizable, iterative, and socially aware.

4.2 Abstract

Working with documents is a key part of almost any knowledge work, from contextualizing research in a literature review to reviewing legal precedent. Recently, as their capabilities have expanded, primarily text-based NLP systems have often been billed as able to assist or even automate this kind of work. But to what extent are these systems able to model these tasks as experts conceptualize and perform them now? In this study, we interview sixteen domain experts across two domains to understand their processes of document research, and compare it to the current state of NLP systems. We find that our participants processes are idiosyncratic, iterative, and rely extensively on the social context of a document in addition

its content; existing approaches in NLP and adjacent fields that explicitly center the document as an object, rather than as merely a container for text, tend to better reflect our participants' priorities, though they are often less accessible outside their research communities. We call on the NLP community to more carefully consider the role of the document in building useful tools that are accessible, personalizable, iterative, and socially aware.

4.3 Introduction

From contextualizing scientific research in literature reviews, to understanding the functioning of complex organizations, experts conduct a wide variety of tasks that depend on document research. Document research, i.e. reading, understanding, and otherwise working with collections of documents is a process that underlies almost all knowledge work. As such, there is a rich body of literature that aims to understand how experts in various fields read documents, characterizing goals, processes, and how their experiences and knowledge inform what and how they read (Bazerman, 1985; Hillesund, 2010; Mysore *et al.*, 2023, *inter alia*).

More recently, primarily text-based NLP tools such as LLMs have been proposed as "solutions" to document research. General purpose commercial models are billed as being able to "understand" both single documents and even whole corpora, context length limits willing, and there are a growing number of purpose built tools targeted at particular professions, from legal document tasks (Merken, 2024; Ravaglia, 2024; Wiggers, 2024) to aiding in the process of scientific discovery (AI4Science & Quantum, 2023), with some claims going as far as to argue that some parts of the scientific process could soon be wholly automated (Lu *et al.*, 2024).

But to what degree are LLMs able to model document research and understanding as currently carried out by experts? In this study, we interview 16 domain experts working in materials science, law, and policy, to understand their processes of document research: Their goals, the uses they have for these documents, and how they evaluate the documents' content for relevance and quality.

In our analysis, we derive tasks common to the experts we interview, and assess the degree to which modern tools from NLP and adjacent fields address those tasks. We find that our experts' processes are highly personal and varied, involve iteratively constructing mental model, and are consistently informed not only by the content of the documents, but by the material and social context of their production. Tools that design around this context, such as citation-aware tools for scientific support (He *et al.*, 2019; Heimerl *et al.*, 2016), better reflect experts' processes, though they tend to be limited to domains that explicitly center publication structure.

By contrast, modern, general-purpose systems from NLP tend to reflect a common, information-centric view: Documents are merely containers for information, and that information within documents can therefore be segmented, decontextualized and displayed without regard for the source document. This is reflected in the affordances of the systems themselves, which often operate either on individual sentences, or on segments whose lengths are determined by an underlying model's context length or a chunking algorithm (Asai *et al.*, 2023), which may or may not align with semantic boundaries (Qu *et al.*, 2024).

The experts we interviewed view documents in ways that cohere much more strongly with theory found in the Science and Technology Studies (STS) literature: that documents are not merely conduits for information, but are traces of social processes whose details are crucial to the documents' interpretation and use. For our experts, these details can inform evaluating a document's provenance, assessing a document against background knowledge of a field, or considering the global context in which a document exists. In other words, a document "serves not simply to communicate, but also to coordinate social practices" (Brown & Duguid, 1996). We therefore call for the NLP community to develop **accessible** systems and methodologies that are more **personalizable**, **iterative** and **socially aware** (§5.13) in order to more accurately reflect the views and priorities of their users, and the rich social context in which documents are produced and consumed.

4.4 Related Work

Understanding Reading. We locate this work in the tradition of work that seeks to understand expert readers and their processes for working with the documents they read, which often take the form of interview studies. Mysore *et al.* (2023), for example, conducts semistructured interviews and think-aloud sessions with data scientists for how they conduct literature review. We also see many similarities with the findings in Bazerman (1985), who finds that physicists rely on "purpose-laden schemas", which include models of both content and authorship and other metadata, similar to what we find with materials scientists.

Document Theory. We draw on the STS literature for theories of document-centric views of knowledge work. Lund (2009) provides a broad overview of document studies beginning in the early 20th century, focusing on the materiality and social production of documents through the digital age. He points out the shift in focus away from documents towards disembodied information in library sciences in the 1960s. Frohmann (2004) corroborates this shift (albeit with slightly different dates) and places it contemporaneously with "discourses of...artificial intelligence and informatics," while arguing for an understanding of the contingent, social role of the scientific publication. These two sources illustrate the a possible origin for the elision of the document in the framing of contemporary NLP. Brown & Duguid (1996) similarly argue that a document, rather than being a "conduit" for information, serve as a mode of social coordination and control in their production and distribution.

Document-aware Reading Support. Document awareness as a principle of system design is an active area of research, especially in human-computer interaction and information retrieval. The Semantic Reader project (Lo *et al.*, 2024) incorporates a number of these features into an reader that shows an enriched view of a PDF document, and works like He *et al.* (2019) and Heimerl *et al.* (2016) allows exploration through citations and other metadata. Work in NLP that accounts for metadata like citations, such as Viswanathan *et al.* (2021), is less common. We note, however, that metadata is seldom considered in reading support work outside of scientific documents, and personalized reading support, i.e. work that considers the context of the reader, is also rare.

Challenges to NLP. There is a growing ambivalence in NLP towards the practice of benchmark-based evaluation (Gururaja *et al.*, 2023). This paper joins a growing number which call for benchmarks to be more closely aligned to end-user needs. Newman-Griffis *et al.* (2021) call for what they term "translational NLP," which proposes an application focus as the driver of scientific progress. Katz *et al.* (2023) propose a new benchmark that presumes the strength of LLMs at traditional NER tasks as the basis for proposing a much more difficult benchmark that better reflects information seeking needs.

4.5 Methods

We recruited 16 participants, beginning with a convenience sampling method (Galloway, 2005), in which the authors began by interviewing existing non-computer science collaborators across the projects they worked on, and then by snowball sampling (Parker *et al.*, 2019), in which participants were asked to recommend other interview candidates. Our participants, six women and ten men, were drawn from collaborations in the materials science, law, and policy communities, were all based in the U.S. and had a wide age distribution, with five participants between ages 25 and 34, six between 35 and 44, three between 45 and 54, and two 55-64. Ten of our participants were associated with materials science, though many identified themselves as belonging to other disciplines, such as chemical or mechanical engineering. All of our participants in this group were either professors or postdoctoral researchers whose primary focus was the synthesis, characterization, or modeling of materials. As such, the document research that they described to us was primarily the process of literature review: keeping up to date with the subfields they already worked in, or learning about new subfields that became interesting to their work. The remaining six participants were academics and professionals whose jobs involve reading, researching, interpreting, or otherwise engaging with law, public policy, or governmental records. This population's document research operated on many more kinds of documents, but the informational goals were largely similar, e.g. staying abreast of relevant policy, legal precedent, or public reactions to policy.

These populations are neither very similar nor dissimilar; we use them as a way of understanding what themes in how experts work with documents might generalize across groups with nominally different tasks, and what themes might be specialized to a single domain. In essence, we aim to establish a loose lower bound of the variety of tasks that professionals across different domains carry out.

We conducted semi-structured interviews (Weiss, 1995) with our participants that lasted between 27 and 73 minutes, with the median interview lasting 53 minutes. Interviews were conducted with a dedicated notetaker, and we recorded the interview audio with participant consent. We followed an interview guide (included in section 4.11), that developed four broad themes: The participant's current work and positionality, their current process for document research including how it fit into their work, how they evaluate documents for relevance and/or quality, and finally what existing tools they use to perform document research. We developed this set of questions in collaboration with materials scientists on one author's project, and was evaluated with a test interview before wider interviewing, with only minor changes to the wording of some questions. The guide did not change between the two subpopulations we interviewed. We conducted the interviews between February and July 2024.

Following the procedures of grounded theory (Strauss & Corbin, 1990), at the conclusion of each interview, authors produced an analytical memo detailing the themes from that interview (Glaser *et al.*, 2004). After a sufficient number of interviews for recurring, coherent themes to emerge, authors began a process of independently open coding the data, developing a thematic taxonomy that generalized across interviews. After this, the authors met regularly to discuss and refine the open codes into a preliminary closed coding frame (Miles & Huberman, 1994). Authors then annotated each interview with the themes from this closed coding frame. In analysis meetings, the authors further refined this closed coding frame by adding, merging, and removing codes, then iteratively re-coding the data. The analysis of this paper emerged from closed-coded versions of the data, and was validated against the original transcripts for appropriate context.

4.6 Summary of Interviews

Document research implied a wide variety of activities to our participants. Despite the variety of tasks, however, several common threads emerged that bridged the disciplinary gap. We conceptualize the tasks carried out by our participants across domains to broadly fall into three categories: Local context tasks, global context tasks, and corpus construction. In the following sections, we first characterize the logistics of the documents that our participants work with, describe some of the examples of each of these three types of task, as well as some broader themes across participants.

4.6.1 Document Characteristics

The documents our participants worked with were primarily, though not always, in PDF format (with exceptions including maps and raw data files), and originated across a wide temporal range. Several participants described having to work with scanned documents which may or may not have had OCR applied to them. In the case of materials scientists, many often consulted technical reports from the 1950s to 70s that were originally typewritten; some of our policy experts looked at digitized government documents that were hand-scanned. Even in cases where documents had been produced digitally, understanding rich visual content, like page layouts, tables, and charts was a key area of focus for nearly all of our participants.

4.6.2 Task types

Local context tasks. Local context tasks, which only involve the content of a single document, resemble common information extraction tasks. In the materials science context, this often manifested as extracting information about how experiments were conducted, the materials that resulted, and their associated properties, in keeping with the principles of data-driven design of materials (Himanen *et al.*, 2019; Olivetti *et al.*, 2020). For instance, participant 21 described how their students *"work on collecting information from the articles, and...build the models that can predict the material properties" (21).* Information extraction tasks were also present in the policy domain, with one participant describing *"trying to extract policy data from these plans, including...different entities, different policy parameters" (1).* Local context tasks are only carried out once a researcher has already developed a mental model for the content of the documents they are searching, and are conducted only in corpora and subfields that the researcher knows well.

Global context tasks. By contrast, global context tasks, which rely on signals from other documents, or the focus document's connections to them, are often much more exploratory.

Common to all of our participants was the task of coming to understand a new corpus or subdomain with which they were previously unfamiliar; our experts were frequently reading and working with documents outside their core area of expertise. In materials science, this was described as *" building my own intuition of a classic material science thing...if I vary this, it goes up or it goes down." (2)*; in the law and policy domain, this could be understanding a corpus of government communications obtained through a freedom of information request, or the ramifications of a new policy through the public response to it. In these cases, information extraction-like approaches are insufficient: as one participant put it, *"it wouldn't be...sufficient to say...we're just searching for a needle in a haystack...we're interested in understanding the haystack." (39).* Researchers described constructing a mental model that was subdomain- or corpus-specific, progressively refining that mental model through the encounter with more documents. Over time, this allowed them to develop intuitions and expectations of the corpus, which also provided signal when they were subverted. This process parallels the background knowledge integral to evaluating the novelty and epistemic status for materials scientists' keeping up-to-date with their existing interests. Global context tasks were seen as a precursor to local context tasks: only after building a reliable mental model for an area did our participants feel comfortable looking at individual documents one at a time.

Corpus construction. While materials scientists consistently described a standard workflow that relied on academic search engines, these types of resources were only available to our participants in law and policy in the case of firm-internal documents or legal precedent. More commonly, participants explained that it was not trivial to collect corpora, beginning with *"putting in phone calls to various libraries...to sort of find out what material is available" (39).* Our participants frequently described having assemble their corpora from documents *"chopped up into different chapters" (8)*, or that contained a *"reference to some other document that contains the relevant information" (1).* Constructing corpora in this way often involves a great deal of expertise, both in knowing what to include, by conducting *"manual verification of the documents we retrieve to understand if they actually are the documents we're seeking." (1)*, and verifying *"the degree to which it is complete or extensive, that's an important consideration." (1).*

4.6.3 Broader themes

Awareness of contemporary technology. We asked about experts' current processes, including the tools they had used, or whether they had incorporated AI tools into their workflow. While many of our participants were technically sophisticated, with some training their own BERT-based models to do topic classification, or writing an emacs-based tool for paper discovery, usage of LLM tools like ChatGPT was largely constrained to non-document tasks like writing or code assistance. By contrast, when attempting to use them for document research tasks, they described a number of pitfalls. Some expressed doubts about the lack of specific background knowledge, or concern about ceding control of the research process; others described trying to to get models to work for their process and facing challenges. One participant, for instance, said *"can we just dump a bunch of PDFs into a GPT and get a summary? And it turns out it wasn't that easy. It wasn't like plug and play. But it also was showing some potential." (17).* Though some of our participants had evaluated contemporary tools, they still considered them models that had to be customized, as in earlier machine learning, rather than useful as drop-in tools.

Access to technology. Of the tasks that we heard described, many are the subjects of active research in the NLP and NLP-adjacent communities. However, very few of our participants had access to these technologies, primarily because only a few of them could or chose to in furtherance of their tasks. This was perhaps nowhere more evident than in the case of digitization and OCR, where one participant described scoping a project based on *"whether the material is digitized or whether we're going to need to digitize it." (39)*, later discussing how a digitization system that considered each page a separate document prevented them from using keyword searches effectively: *" if we had reliable high quality text, and we had our document organization [taken] into consideration...then we could have used a keyword based search as like a candidate classifier." (39)*

Personalization. Regardless of field, a consistent theme that we observed in our interviews was how idiosyncratic each researcher's process of document research was. There seemed to be no agreement, even within fields, on what makes a document relevant to a given search or what cues a researcher might use to assess the epistemic status of individual claims. We view this as a major challenge for NLP systems.

4.7 Task Analysis

In this section, we identify key tasks that our participants shared across fields, and compare the current state-of-the-art NLP tools with the needs that our experts outlined.

4.7.1 Information Extraction

Traditional IE, where we tag entities of concrete, well-defined types, would be useful for many of our participants. However, our participants' needs were for concepts that would be specific to their research objectives (i.e. not available in off-the-shelf models) and difficult to define succinctly, like tagging spans that provide evidence of 20-25 core political values in an argument, like *"equality and justice, liberties, security, safety" (42)*. Participants also described the acceptable granularity of extracted information varying per-project: *"we just used like more bag-of-words-based, keywords-based approaches...but the whole research project in that case assumed a level of bluntness that wouldn't be appropriate for other projects." (39)* This is different to standard benchmark datasets for IE where the types are more concrete and well-defined (Ding *et al.*, 2021; Tjong Kim Sang & De Meulder, 2003). Further, supervised neural systems require a large amount of expensive annotations for each new tag set (Li *et al.*, 2020), though recent work on few-shot IE with LLMs has aimed to reduce the potential annotation burden (Ashok & Lipton, 2023; Hofer *et al.*, 2018; Huang *et al.*, 2020).

Our participants also emphasized that not all IE tasks of interest to them involve only local information, conflicting with traditional IE focused on within-document and short document settings. They describe settings where IE would be applied to long scientific papers and policy briefs: *"I don't think there's a way for me to get the information that I need without having the full document, but once I have that document, I will only use the specific portions that I need." (11)*. These documents may exceed the context window of LLM-based IE systems and result in poorer performance (Dagdelen *et al.*, 2024a).

Evaluating IE tools extrinsically can reveal the significance of existing vulnerabilities. The value of better aligning evaluation principles with user needs was highlighted by one of our participants: *" it would still dramatically reduce the amount of time a researcher would need to spend... But the conclusion was that these tools... were inadequate on their own" (39)* For example, Adams *et al.* (2024) shows that LLMs do not perform well enough on long document clinical notes for reliable clinical use in question-answering. In relation extraction, there has been work to soften evaluation metrics to accommodate for the use of generative models, reflecting a shift towards evaluation methods that are aligned with downstream utility where it is often enough to recover spans with overlapping span boundaries (Jiang *et al.*, 2024). These types of evaluations would better highlight the most impactful open problems in IE.

4.7.2 Multimodality and OCR

Understanding visual and layout features of documents was a priority for nearly all of our participants, the vast majority of whom worked with documents prominently featuring tables, charts, or which conveyed information through layout. One materials scientists characterized their needs as *""looking for a statement backed up by data and the data can be just numbers. It can be graphs. It can be pictures of microstructures or just all" (0)*, and another described extracting *"a ton of data and figures...because they're usually X, Y plots." (5)*. These tasks are inherently multimodal: while some LLM approaches recommend linearizing tables into text, charts, images, and layout information must be handled visually. Compounding this is the tendency for these documents to be scanned or photocopied instances of paper documents, but which still necessitate the processing detailed above, resulting in it being *"tough to really search through them, and so they might not...appear when you're doing a lit review on Google Scholar" (2)*.

While visual understanding of documents is not a solved problem, dedicated multimodal layout understanding models like the LayoutLM series (Huang *et al.*, 2022, *inter alia*), which serve both as visual segmentation models as well as representations for downstream document reasoning tasks remain an active area of research, general-purpose models like Qwen-2-VL(Wang *et al.*, 2024) and LLaVA-Next (Liu *et al.*, 2024) now include visual document understanding and OCR benchmarks like DocVQA (Mathew *et al.*, 2021) and TextVQA (Singh *et al.*, 2019), and proprietary models like Claude now have modes designed explicitly around PDF processing [15].

4.7.3 Iterative Search and Exploration

One common theme repeated across multiple participants was that the search process is inherently iterative. Rather than rely on a single set of results identified for a particular information need, researchers will iteratively expand their search across multiple stages, using results to inform each successive step, progressively building a mental model of the search space.

Researchers iterate for different reasons. Some seek to identify the provenance of the information they find: *"I go and read a current paper and I find where they cited that they got information from. And I read where that person got the information from and I read where that person got the information from. I usually try and find the original sources to everything." (5)*, with one even stating *"sometimes I read a paper and the main use of it is the references in the paper" (4)*. Others iterate precisely to establish the global context surrounding a particular paper: *"you have to like build up this context around the paper. What came before it? What is it citing? And how does the content of that paper relate to the ideas that came before it?" (4)*.

Iterative approaches do exist within the information retrieval literature. Initial retrieved documents can be used as a source of information on related lexical terms Attar & Fraenkel (1977), which additional can help address issues like terminology drift (section 4.7.4), and the process of learning as you search, where *"reading and coding...helps you generate, develop contextual knowledge" (39)*. More recent work has focused on iterative approaches which

[15] https://docs.anthropic.com/en/docs/build-with-claude/pdf-support

search through the structure of documents Hsu *et al.* (2024); Min *et al.* (2019); Zaheer *et al.* (2022), albeit with a focus on the task of multi-hop reasoning Yang *et al.* (2018) rather than information exploration.

Iterative construction of mental models could also be supported by the task of ontology induction. While there has been recent work to induce concepts from individual documents (Matos *et al.*, 2024), our participants highlighted the need for concepts detected at the corpus level that are tailored to a specific research question. Despite initial work in the scientific domain (Katz *et al.*, 2024), abstracting concepts across documents has been shown to be challenging for state-of-the-art LLMs (Guo *et al.*, 2024), even without the per-question adaptation.

4.7.4 Terminology

One of the consistent difficulties that our participants faced across fields, similar to those in Mysore *et al.* (2023), was terminology. Our participants described three types of terminology shift. The first, temporal, is when the meaning of a word or a term drifts over time. This might be because of shifts in community usage, like one materials scientist pointed out, *"in like 2008-ish, the Chinese community decided that the word shock [testing] also means high rate Kolsky bar testing" (5)*. It could also be because the referents of the words themselves have changed, as in the case of committee organization in a local government: *"if you're looking for committee reports from before 2018 and you're looking for hospitals, the committee on hospitals didn't exist until 2018." (40)*. The second, domain-specific change, is when different fields use different words or terms for very similar things: *"I call it a surrogate model. If you ask a statistician what they're going to call it, they're going to call it an emulator. If you ask someone in the reliability community...they're going to call it a response surface." (17)*. These two types of terminology shift are partially addressed in the existing literature. Periti *et al.* (2024) survey approaches to track new usage and senses for existing vocabulary; Lucy *et al.* (2022) quantifies domain-specific terminology usage and synonymy across fields; Head *et al.* (2021) provide context-sensitive definitions of technical terms and mathematical symbols.

However, while existing systems work on either identifying terms as near-synonyms or providing definitions, our participants emphasized that understanding the differences in meaning and why one term might be used instead of another was also important. For example, one materials scientist outlined how while "oxidation" and "aqueous corrosion" meant similar processes, the keywords used to search for one vs the other, and the numbers that would characterize those properties would be different: *"if it's aqueous corrosion...they might care about the atmosphere or basically the liquid concentration a lot more...But if you were then* *going to oxidation in high temperature, they are mostly looking at mass gain data." (2)* The simultaneous focus on the similarities and differences in mostly-synonymous words reflects a process in which our participants tended to jointly model the semantic content of documents alongside the documents' social context, like its authorship or intended audience. This was especially true with the final type of terminology shift, political, where people describe similar things, but may use different language to convey different valence, or signal which aspects of an issue are being prioritized, as one participant gestured to in the case of climate policy: *"rural communities will talk about micro grids, not as a part of climate action, but as a way to get off the dependency on investor owned utilities like PG&E." (8).*

In these cases, our participants not only had to engage in a iterative process to find the synonyms, they also had to reason about the positionality of the authors and why they might use a different term. Understanding identity, its presence in corpora, and its interaction with LLMs is still a new area of research: Kantharuban *et al.* (2024) and Li *et al.* (2024a) demonstrate the sensitivity of models like ChatGPT to implicit markers of identity in responses and refusals, respectively, and Lucy *et al.* (2024) investigated author positionality and community belonging, and Milbauer *et al.* (2021) demonstrated cross-community lexical differences linked with ideology. Reading support that addresses the needs of our participants would need to unify several running themes of work in an accessible interface: providing definitions, enriching those definitions with understanding of where assumptions and practice across communities might differ and the implications of different terminology across domains.

4.7.5 Corpus Construction

For corpus construction, which we describe in section 4.6.2, our participants most often described starting with collections of library resources and Google searches, only eventually moving to write web scrapers if the kinds of documents were similar enough. Automated tools, such as custom scrapers that use LLMs to explore documents (Huang *et al.*, 2024; Ma *et al.*, 2023) would be extremely useful for expanding corpora beyond what is feasible to manually collect. However, our participants stressed the need to reconstruct documents from chapters, understand document versioning, and how to iteratively build *"some checks of what's missed by [a search], false positives, false negatives" (42)* in constructing corpora, implying that document scrapers for constructing corpora would need to accommodate an iterative, exploratory style to truly function for this purpose.

4.8 Conclusion

In this paper, we interview 16 domain experts from the domains of materials science and law/policy to understand how they conduct document research. We find that their self-described processes rely on understanding and actively modeling the social processes by which text is produced, through the construct of a document and its associated metadata. As exemplified by the success and usage of layout understanding models and the HCI work on scientific reading support, document-aware tools match the processes of our experts. More concretely, however, there are four qualities of systems that we call on the NLP community to build into new document research tools for expert users outside the field:

Accessible Though we have characterized the state of existing research as it addresses many of the concerns of our experts, we note that the most accessible NLP tools today by far are primarily text-based and rarely consider documents a first-class citizen. By contrast, NLP work that does center documents is far less accessible. Many of the issues faced by our participants, whether working with older, undigitized documents, understanding terminology drift, or leveraging metadata to assess the provenance of a document are already the focus of existing research, but are not as accessible to them as web-based LLM systems. How can we both promote better modeling of how users engage with large quantities of text in accessible systems like LLMs, and make tools that do this modeling more accessible?

Personalizable We note in several places that our experts had idiosyncratic processes for finding, evaluating, and reading documents. From looking for different types of information both within and across fields, to having vastly different heuristics for assessing the provenance and reliability of a document, our participants, all successful experts in their fields, conduct research in highly personal, specific ways. Amid concerns of LLMs potentially stifling creativity and serving as a homogenizing force (Anderson *et al.*, 2024; Kumar *et al.*, 2025), NLP tools should encourage and support diversity of thought by allowing for experts to customize systems to their existing personal processes.

Iterative Almost all of our participants discussed the process of constructing a mental model as a series of iterative updates as they read new documents and reconciled the content with their expectations. This process of constructing a mental model was essential, and heavily used global context: experts relied extensively on signals like how any given document related to background knowledge, other documents in the corpora, standard practices in the field, as well as social signals like authorship and author positionality. As they read, their understandings of both the signals and how they pertained to the documents co-evolved. By contrast, NLP tools are often presented as either static or occasionally updated artifacts. To better support how experts work, NLP systems should malleably support their users' evolving mental models, including through easy schema updates, flexible data relabeling, and user-friendly retraining and evaluation loops.

Socially aware The social character of document production played an important role for our participants, whether in understanding terminology and why it might be used, modeling authors and participants in spheres of discourse they participated in, and evaluating information for reliability and provenance that included information beyond the propositional content of the text they worked with. Our participants frequently looked beyond what was written in text to how and why it was written, treating documents more as traces of social processes than containers or conduits for pure, disembodied information. NLP tools should be designed around the idea that social context is crucial for understanding text: authorship, audience, communicative intent, and format are all factors that readers already consider; as readers already know, models should be aware that no text is *just* text.

4.9 Limitations

Our interview study was evaluated and approved by an Institutional Review Board as STUDY2023_00000431.

While we constructed our sample deliberately to span two fields, we acknowledge that this was a convenience sample, and that it is neither specific to one field, nor representative of all the ways that people might work with documents. Our study also focuses exclusively on potential users of NLP tools who are already experts in their domains. Their concerns are not likely to be representative of non-domain expert users, especially non-experts reading technical language.

4.10 Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0121. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

4.11 Interview Guide

Begin by defining document research: we're interested in processes you have for finding documents, or things within documents to help you in research tasks in a professional context.

Demographics/Positionality

1. Can you briefly describe your job, covering the kinds of research questions that you encounter in your line of work?

2. In your work, can you describe the cases where you have to look for documents, or things within documents? An example would be great.

How do you do document research now?

3. Can you describe the goals that you have when you do document research? What kinds of documents and information do you search for? If you have different kinds of searches with different goals, please describe them.

4. When searching for documents, are you searching for documents as a whole, or specific pieces of information/facts within those documents? How much of the document's content as a whole do you end up using?

5. What purpose do those documents or facts have once you find them? (reference? Quotation material? Prior approaches to what you're trying to solve?)

6. Is there an existing ontology to the kind of searching that you do? Are the things that you search for in documents part of a well-defined set of things, or is your approach to these documents creative?

7. To what degree is finding documents or facts an iterative process? Is there a mental model that you have of the space of possible documents that you update as you find new documents?

8. To what extent is the structure of documents relevant?

How do you evaluate the documents that you find?

9. When you search for documents or facts, what does it mean for a document or fact to be high quality to your purpose?

10. When evaluating a document or fact for relevance or quality, how much of that depends on the content of the document itself?

11. How much of that depends on knowledge of the field that you have that's not explicitly in the document - other important documents, standard practices, etc? Are there resources for that kind of domain knowledge?

12. How much depends on metadata, like citations, author affiliations, venue, etc?

Existing tools

13. What tools do you currently use to search for documents?

14. When executing a search, how quickly do you usually find the sort of thing you're looking for? Are there specialized keywords that get you to what you're looking for?

15. If you use specialized tools for your domain, what do they do differently from generic-domain tools, like Google search?

16. Do you ever write code to enable better searching? What are the tasks that code helps you with that existing tools are insufficient for?

17. If there was something that your tool could do differently or better, what would it be?

18. Have you used AI-based tools to aid in your work? How well have they suited your workflow and process?

Demographics

19. I am going to read some age brackets. Can you indicate when I read a bracket that your age falls into?

    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65+

20. Is there anything else in your background that you consider relevant?

Part II

Resisting Generality With User-Focused Evaluation

Chapter 5

Data-driven Design as a High-Impact, Ecologically Valid Benchmark for Document Understanding (In-Progress)

In-progress work, presented non-archivally at the AI for Scientific Discovery Workshop at NAACL 2025, with a dataset proposal to be published at the NeurIPS 2025 AI4Science workshop.

5.1 Overview

In part 2 of this proposal, we discuss scoped interventions to improve the degree to which evaluations of NLP systems accurately reflect users' experience of them. In this chapter, we discuss a benchmark designed explicitly around ecological validity. Performance on this benchmark, which evaluates models' ability to generalizably extract information from academic papers for data-driven design, translates to immediate utility for materials scientists wishing to expand into the data-driven design of new materials systems.

5.2 Introduction

Data-driven design (DDD), a process by which materials scientists use information extracted from the literature to inform future experiments, has emerged in the past decade as an important method by which to accelerate the discovery of materials (Olivetti *et al.*, 2020). As NLP methods have evolved, so too has their application to data-driven design problems, from pipeline-based approaches using multiple purpose-trained models and relying heavily on rules-based, handwritten heuristics (Court & Cole, 2018; Jensen *et al.*, 2019; Kim *et al.*, 2017, *inter alia*) to end-to-end approaches involving fine-tuning large language models (LLMs) to act as information extractors and assistants (Zheng *et al.*, 2023), or generate structured output describing properties directly (Dagdelen *et al.*, 2024b).

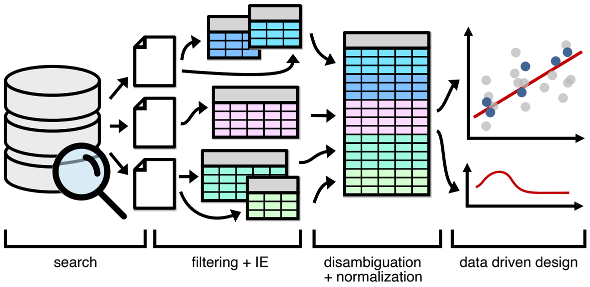


Figure 5.1. The process of data-driven design. Our benchmark focuses on the middle two phases: extraction/filtering and disambiguation/normalization

However, even current, LLM-based data-driven design work relies on laboriously collected annotated data. The method proposed in Dagdelen *et al.* (2024b), for instance, suggests annotating " 100–500 text passages" in order to fine-tune an LLM to produce structured data. This type of data can be difficult to produce: it often requires domain expertise to collect, verify, and postprocess into a format that is appropriate for training such models. This problem is exacerbated when considering that data-driven design efforts often seek to extract information into specific, non-overlapping schemas, limiting the possibility of data sharing or transfer learning between separate DDD efforts.

Given, however, the rapid development of models that can process scientific documents, both in text-only and multimodal formats, we view the possibility of data-driven design projects that require little to no annotated data as both highly desirable and feasible in the near future. The extraction challenges created by typical data-driven design projects also remain at the frontier of the capabilities of even newer models: processing visually-rich documents with information in text, tables, and figures; disambiguating extracted information to a standardized schema; and performing consistent numerical reasoning to normalize scales and units so that extracted information is comparable across papers (Miret & Krishnan, 2024).

In this paper, we propose a dataset to demonstrate DDD's suitability as a challenge task and benchmark for next-generation document understanding models, focused on replicating a subset of the data derived from two prior DDD works: Jensen *et al.* (2019), which focuses on zeolite synthesis, and Pfeiffer *et al.* (2022), on aluminum alloys. Both of these datasets focuses on a different materials system, and the relevant information from each paper and the schema into which they are extracted are also different.

We intend for this benchmark to reflect a realistic subset of the data-driven design process, which often relies on downloading papers from publishers, which are typically provided in XML/HTML or PDF format. We therefore present two settings for this benchmark: a multimodal setting, which presents as input a full paper PDF (or a series of page images, for models that do not accept PDFs directly), and a text-only setting, in which input is XML/HTML. In both cases, the expected output is the disambiguated, normalized information from the corrected versions of the original dataset. We propose zero-shot baselines in both settings, and find that while modern systems perform strongly on common formats of tables, their ability to extract and integrate information varies widely between different sources of information and different table layouts.

Because we cannot republish the content of papers used in this benchmark, we release a script to reconstruct the dataset from the metadata provided in each source paper alongside evaluation code for the benchmark in our repository[16]

5.3 Data Driven Design and Task Scoping

In keeping with the literature, we conceptualize of DDD as a task separated into four phases, which we visualize in figure Figure 5.1, and discuss below:

Search/retrieval In this phase, researchers typically collect a large number of papers using high-recall, low-precision methods like keyword matching. Papers are typically downloaded in a number of different formats, including scraped HTML and XML and PDFs from publisher APIs, then converted to text.

Information extraction and filtering In this phase, researchers will attempt to extract information corresponding to the schema of interest from the retrieved papers. Notably, in this phase, not all extracted information is relevant, necessitating a filtering process. The specific methods by which this phase is carried out have varied over time. Olivetti *et al.* (2020) describe an pipelined approach common at the time; end-to-end approaches have since become more popular.

Disambiguation and normalization In this phase, researchers attempt to make information extracted from retrieved papers comparable. This can be seen as a two-step process: disambiguating extracted information into the intended schema, and normalizing numerical values to be comparable, in both scale and units.

Visualization and Modeling The goal of data-driven design projects is typically not just the extraction and disambiguation of information, but using it to visualize existing literature, in order to plan future experiments, or to serve as a preliminary screen for promising new

[16] https://anonymous.4open.science/r/ddd-benchmark-C3B8/README.md

candidate materials by predicting properties of interest, such as in Zhang *et al.* (2024). We argue that a useful evaluation for document understanding systems is to focus on the second and third phases and their associated tasks, namely information extraction, filtering, disambiguation and normalization. With this scope, we aim to present a system with the content of a paper and the desired schema, and have it output normalized information from the paper in that schema. Systems that perform well at this evaluation would be immensely useful to materials scientists: given a collected set of potentially useful papers and a desired schema, the system could automate the construction of a dataset that allows the modeling and prediction of potentially valuable new materials systems.

5.4 Task Settings

One of our primary focuses in designing this benchmark is *ecological validity*: which would imply that models that are successful at this benchmark would be able to be applied to real-world DDD tasks. For our benchmark to accurately reflect DDD as it is currently carried out, the models that we evaluate must be able to operate effectively on all of the formats in which publishers make their documents available. In our case, we collect PDF and XML/HTML documents from publishers through their APIs, according to our institutions' licensing agreements. We present two settings: a PDF setting and a text-only setting. In the PDF setting, the model receives as input the full PDF of the document, or alternatively a series of PNG images of each page if it does not process PDF files directly. In the text-only setting, the content of the XML document is provided. In neither case do we apply additional preprocessing to the documents being processed. We note that because most publishers do not provide data in both formats, these two settings are not directly comparable; to compensate, we highlight results from the set of papers that we do have available in multiple formats alongside the PDF-only and XML-only results. We envision the setup of this task to be that a model receives a set of input document, and a target schema into which to extract and normalize information, and produces a table in that schema as a result. Preprocessing, prompting, and in-context learning and training a model before the task are all considered parts of acceptable implementations for this task, but fine-tuning per-dataset would not be acceptable.

5.5 Dataset Construction

We begin the construction of our dataset from two prior data-driven studies: Jensen *et al.* (2019) and Pfeiffer *et al.* (2022). In this section, we first describe the process of collecting the paper content from which both studies extracted data, and then the further annotation and collection processes for both datasets. We finally detail the challenges that this dataset presents to models models in order to be successful at this task.

5.6 Data Collection and Distribution

To collect data across both modalities for this dataset, we focused on three publishers with relatively permissive text and data mining (TDM) licenses, for which our institution had an agreement in place. After resolving publisher metadata for each dataset with the CrossRef API,[17] we used publisher APIs from Elsevier and Wiley to download full-text XML and PDF files, respectively, and used the Springer integration with CrossRef to get PDF and XML/HTML if available. We focused on the use of TDM APIs such that this dataset can be replicated at any institution with comparable licensing, because we cannot redistribute the papers directly. Further, framing the task as applying models directly to the outputs of TDM APIs allows models successful at this task to be drop-in augmentations for new or existing DDD projects. For institutions that have less permissive licensing, we tailor our evaluation script to evaluate and compare candidate models with our baselines on subsets of the papers used for the baseline evaluations. See the discussion in 5.9.1. A limitation of this dataset is the scale of the data that is available to us because of the intersection of using manually checked and corrected data and complying with licensing terms. We argue, however, that because of the many-dimensional nature of the information being extracted, that this dataset is still useful as a benchmark.

5.6.1 Zeolites (Derived from Jensen *et al.* (2019))

The original zeolite dataset [18] consists of synthesis parameters and derived products of zeolites, a class of materials with many commercial applications. This paper extracts 1,638 rows of manually verified data from 116 individual papers, looking at content in tables, text, and supplementary information using a combination of learned and rule-based extraction. Zeolite synthesis typically involves creating a gel from several components: the elements that form the crystal, such as silicon and germanium, additional reaction components, such as water, and an organic molecule that directs the crystal formation. This dataset contains 12 columns of these ingredients, as well as several more that represent further normalization of their contents, or the results of corroborating simulations.

For our dataset, we focus on columns that are directly extracted from the papers themselves, removing derived columns. Because publishers often do not provide a way to programatically

[17] https://crossref.org

[18] Available at: https://github.com/olivettigroup/table_extractor/blob/master/zeolite_data/ge_synthesis_data.csv

access supplementary information, we additionally scope down our dataset to information derived from tables and text in the main paper. With our licensing constraints, we have a total of 55 papers, four of which are unavailable through Wiley's API. This results in 51 papers, 22 in PDF format, 39 in XML, and 10 in both formats. Our final dataset contains 414 rows of data, with a total of 4950/4968 non-null values. We provide an example table from this dataset in Figure 5.3, along with a worked example of a subset of the extracted data in Appendix 5.16.1.

5.6.2 Aluminum Alloys (Derived from Pfeiffer *et al.* (2022))

Pfeiffer *et al.* (2022) compiled a dataset of 1278 entries on mechanical properties of aluminum alloys extracted from tables. Because of the limitations of automated extraction tools, this was presented as a separate dataset from the compositions that had those mechanical properties. While this data was validated against documented handbook values for common alloys, only outliers were manually checked. Pfeiffer *et al.* (2022) additionally note that many of the high strength outliers were due to the aluminum alloy being subject to severe plastic deformation (SPD) processing and were flagged but not removed.

In order to create the benchmark dataset all of the original entries were manually inspected by multiple researchers. There were additional entries that were related to SPD processes that were not flagged in the original dataset. Further there were several other categories of issues with extracted data which stray from the purpose of assessing properties of (industrially relevant/viable) aluminum alloys: many entries were mechanical properties of welds (particularly friction stir welds) or aluminum-based composites. The entries that were properties related to SPD, welds, or composites were removed from the dataset. Entries that were references to other works (e.g. referencing a handbook or earlier literature as benchmarks) were also removed.

Once the desired subset of entries from the original mechanical properties dataset was identified, additional information was added to each entry regarding the composition and the processing information, thus unifying the two separate datasets presented in Pfeiffer *et al.* (2022). Room temperature tempering steps, e,g, natural aging, are described at 25°C. All temperatures are in Celsius, all times are in hours. Compositions are in weight percent. This results in 8 columns related to composition and properties, and 28 columns that are weight fractions of individual elements. With our licensing restrictions, we were able to obtain 152 total papers, with 22 in PDF format, 151 in XML format, and 21 in both. This resulted in 330 rows, and 3806/12210 non-null values.

5.7 Diagnostic Datasets

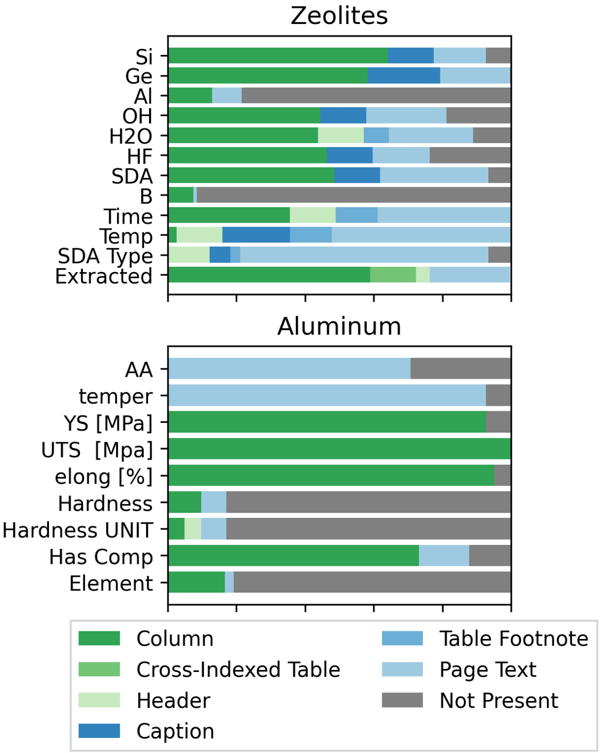


Figure 5.2. Distribution of data locations in the dataset per column type in each dataset. Green bars indicate information found within tables, blue indicates related text, and gray indicates absent information. Note that the aluminum dataset has much more homogenous sources of information.

In order to provide more detail on where models succeed and fail, we additionally annotate subsets of each dataset with location and layout information that indicates where information in each dataset, at a column granularity is found, and where it is expressed. For each datapoint, we annotate if the data was found in the text of the paper, or in several configurations of table. This comprises eight categories in three buckets: (1) Data from the tables (entire columns for that data, information in headers, or information in particular cells under hierarchical indices); (2) Data from text, even if linked to a table (generally text on the page, but also footnotes and table captions); or (3) not present in the paper.
We annotate data from 28 papers found in the zeolite dataset, and 40 papers from the aluminum dataset. We present a visualization of how data are distributed into these buckets in Figure Figure 5.2.

5.8 Task Features

This task presents a number of interesting challenges to information extraction methods. In this section, we discuss these features, using an example of table parsing taken from Lorgouilloux *et al.* (2009, Figure Figure 5.3).

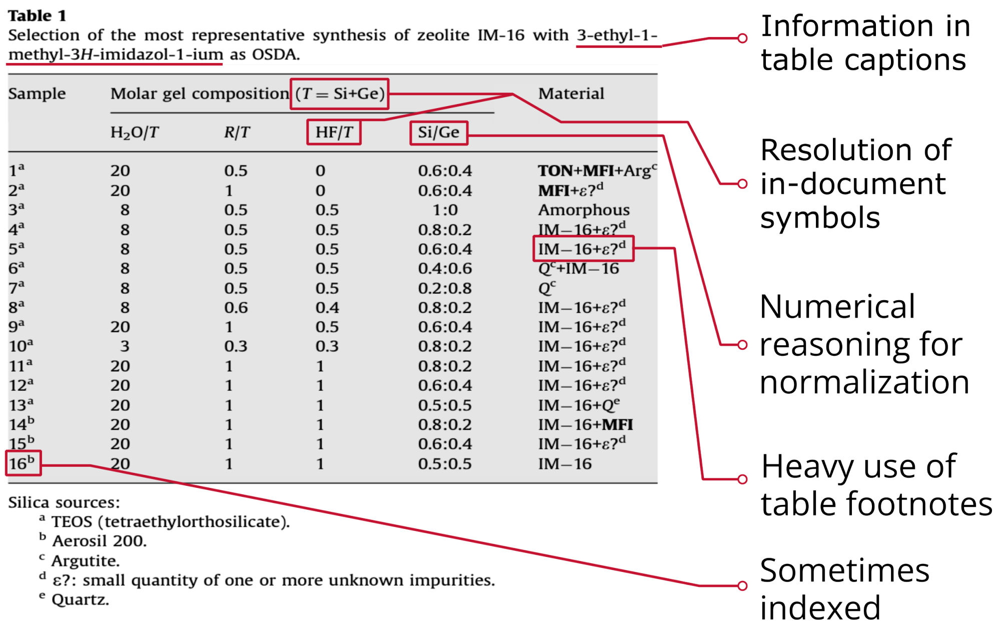


Table 1
Selection of the most representative synthesis of zeolite IM-16 with 3-ethyl-1-methyl-3*H*-imidazol-1-ium as OSDA.

Sample	Molar gel composition (T = Si+Ge)				Material
	H_2O/T	R/T	HF/T	Si/Ge	
1^a	20	0.5	0	0.6:0.4	**TON**+**MFI**+Arg^c
2^a	20	1	0	0.6:0.4	**MFI**+ε?^{d}
3^a	8	0.5	0.5	1:0	Amorphous
4^a	8	0.5	0.5	0.8:0.2	IM−16+ε?^{d}
5^a	8	0.5	0.5	0.6:0.4	IM−16+ε?^{d}
6^a	8	0.5	0.5	0.4:0.6	Q^c+IM−16
7^a	8	0.5	0.5	0.2:0.8	Q^c
8^a	8	0.6	0.4	0.8:0.2	IM−16+ε?^{d}
9^a	20	1	0.5	0.6:0.4	IM−16+ε?^{d}
10^a	3	0.3	0.3	0.8:0.2	IM−16+ε?^{d}
11^a	20	1	1	0.8:0.2	IM−16+ε?^{d}
12^a	20	1	1	0.6:0.4	IM−16+ε?^{d}
13^a	20	1	1	0.5:0.5	IM−16+Q^e
14^b	20	1	1	0.8:0.2	IM−16+**MFI**
15^b	20	1	1	0.6:0.4	IM−16+ε?^{d}
16^b	20	1	1	0.5:0.5	IM−16

Silica sources:
[a] TEOS (tetraethylorthosilicate).
[b] Aerosil 200.
[c] Argutite.
[d] ε?: small quantity of one or more unknown impurities.
[e] Quartz.

Information in table captions
Resolution of in-document symbols
Numerical reasoning for normalization
Heavy use of table footnotes
Sometimes indexed

Figure 5.3. Example table from the dataset, reproduced from Lorgouilloux *et al.* (2009, Table 1). This table is annotated with several of the challenges with table extraction in this dataset, including: (1) Generic table layout understanding; (2) Processing information related to tables, such as captions and footnotes; (3) Understanding and resolving in-document substitutions; and (4) Numerical reasoning to normalize ratios.

Identifying Relevant Information. This task presents an example of long-context information extraction, where models are expected to extract information into a provided schema given the context of an entire scientific paper (which is often longer than the longest context window available), in which relevant text and tables have to be identified before being tagged.

Table Understanding. One of the core challenge of this task is processing tables in the variety of forms in which they occur. Tables expressing synthesis parameters and recipes are difficult to construct: Experiments often involve the systematic variation of several different parameters, leading to a challenge in how to represent hierarchical data in many dimensions in an ultimately two-dimensional table. This results in a number of different formats. Figure Figure 5.3 demonstrates perhaps the most common format, normalized rows per-experiment, but hierarchical representations that involve leaving cells blank to indicate a hierarchical grouping of experiments are also common, and pose a challenge for table understanding models.

Related Information. Information from text can be found in any section of a paper, and information necessary to understanding a table is frequently presented outside the table, whether in text, captions, footnotes, or otherwise. Figure Figure 5.3 specifies the OSDA compound in the caption, and additionally specifies the expansion of several acronyms in table footnotes, which are placed directly below the table. Further, many papers introduce information necessary for table understanding in the text surrounding the tables. We note that table captions can be an edge case for some approaches to understanding document layout: The VILA Shen *et al.* (2022) model, for example, detects the table captions and footnotes as part of the table, which can lead some table understanding models to parse table captions and footnotes as further rows of the table, rather than footnotes. Further, understanding non-table information here requires the resolution of superscripts to their corresponding footnotes.

Numerical Reasoning. Synthesis procedures for zeolites are commonly expressed in terms of molar ratios of the components, and the choice of which element to which to normalize changes interpretation of numerical values in the table. For example, in Figure Figure 5.3, ratios are scaled to the combination of silicon and germanium in the sample. By contrast, several other papers (and the final dataset) scale to only the quantity of silicon, requiring a normalization step that introduces a multiplicative factor to enable direct comparison of results across papers.

Within-document Reference Resolution. Before normalization can occur, tables often require the resolutions of symbols that are defined elsewhere in the document. In this case, the table headers indicate that the H_2O, R, and HF columns are normalized to T, which the upper header declares as the combination of silicon and germanium in the sample.

Sparsity. In many cases in the extracted dataset, columns will have values of 0, because a given element was not used. Systems that attempt this benchmark must not hallucinate non-zero values even when given a comprehensive schema of all items that may or may not be present.

5.9 Evaluation

Given the information extraction-based nature of this task, we consider an F1 metric, with some allowance for what is counted as a match. In the case of numeric columns, to allow for imprecision in normalization of ratios, we consider a “correct” answer to be within 0.1 of the true answer. In the case of the OSDA name column and the extracted products column, we expect an exact match on a lowercased version of the string with all punctuation replaced by an underscore; in the case of the extracted products column, we note that often, several products of a reaction are mentioned; we intend to improve the granularity of our evaluation in ongoing work. Given the large variance of the number of rows/data points extracted from individual articles, we consider a micro-averaged F1 score to be an appropriate choice.

For evaluation, we provide code that accepts a spreadsheet with the same header row as the original dataset (omitting the location rows), and is configurable per-dataset for which columns are to be evaluated. `None` values in predictions indicate that the model is not providing a response, to disambiguate from cases where the correct extraction is zero, or another common placeholder value. For our F1 metric, we consider any data that is available on the page (i.e. not annotated as being "not present") a candidate for extraction. A true positive is any data point that is available to the model and correctly extracted; false negatives are any point that the model fails to extract. False positives include both incorrectly extracted values and values that are not available to the model, but that it provided a value for anyway. True negatives are information not available to the model that it successfully does not provide a value for. We micro-average the F1 across papers, and additionally provide per-location F1 scores to indicate what sources of information models are adept at working with.

However, evaluation does pose additional challenges: While some tables translate straightforwardly between rows in the original table and the dataset, others are structured differently, using hierarchical indices, such that blank cells' content must be inferred, or tables with multiple levels of hierarchy, where one cell and the headers that index it correspond to a row in the final dataset. We call these tables *cross-indexed*. Further, while the table reproduced in Figure Figure 5.3 uses identifiers for individual samples, that is not common in our dataset. As a result, there is no *a priori* alignment between rows in the dataset and rows produced from models solving this task.

To address this, we use a simple heuristic algorithm that attempts to align rows in the dataset with rows produced from the systems under evaluation, with strong priors towards the initial alignment being correct. Our algorithm begins by computing a row-wise score between all rows in the dataset and predictions. This score computes the match discussed above on all columns where information is within the provided context window, to avoid spurious matches on absent information. We then iterate through each row of the dataset, and choose the highest scoring predicted row to align with each row in the dataset. In the case of a score tie, the sequentially following row is assigned. Because of the varying structures of tables in the dataset, we additionally implement fallbacks in the case of a mismatched number of rows between the dataset and predictions. In the case where the model produces more rows than are observed in the dataset, each additional row is penalized as being false positives; in the case where the model produces too few rows, we construct placeholder rows of no predictions to indicate that the model has not provided an answer. We note both that this alignment strategy is not guaranteed to produce the optimal alignment, but also that any similar strategy will end up favoring models by potentially offering mistaken credit.

5.9.1 Permissive Evaluation

As a result of being focused on papers in a field where open-access publishing is not common, our dataset relies on users to reconstruct the dataset. This in turn relies on their access to the same journals we used when constructing the dataset, which we cannot reasonably assume. As a result, we provide a script to evaluate predictions made on a subset of the papers present in our dataset, rather than the whole dataset. In order that metrics computed in this way be comparable to the baselines, we additionally provide a script to re-evaluate the baseline predictions (which we also provide in our repo) on the same subset of data to provide an apples-to-apples comparison on the subset of papers that a given user has access to.

5.10 Baselines

Despite recent work investigating Large Language Models (LLMs) as possible automated scientists (Lu *et al.*, 2024; Si *et al.*, 2024), to our knowledge LLMs have never been systematically evaluated on research processes such as precise multi-document review and synthesis. As a baseline, we evaluated a prompt-based strategy with a variety of Large Language Models, when available comparing unimodal (text only) and multimodal (text + vision) LLMs of comparable sizes. The models used are described in Table Table 5.1.

Model	Size	Open?	Multimodal?
Molmo	7B	✓	✓
Llama3.3	70B	✓	X
Qwen2.5	72B	✓	✓
GPT-4o	?	X	✓
Claude-3.7	?	X	✓

Table 5.1. Statistics of models used for the baseline experiments.

Our goal is for the model to perform both the information extraction and table normalization jointly when provided with either an image (300 dpi PNG) of the PDF page, or the raw underlying XML of the document split into chunks. For the visual modality, a PDF document is separated into individual pages. The VLM is prompted with the page and a text prompt, returning a list of possible data rows discovered in the image. Rows are then aggregated for each PDF document. For the XML modality, the XML corresponding to a document is split using a sliding window with `window size` of 10000 characters, and a `stride` of 5000 characters; this ensures that no data components are split at a window boundary. Each window is passed to the LLM, and extracted rows are aggregated per document. The combined size of XML and prompt text meant that smaller open-weight models had insufficient context size. Prompts were constructed in collaboration between two authors of this paper: One, a graduate student in NLP; the other, a graduate student in materials science. This development process allowed us to leverage insights coming from either NLP or material science expertise. The constructors were provided with three randomly selected articles from the dataset to act as a guide while developing their prompts. We intentionally restricted the prompt constructors' access to the full set of papers so that information and edge cases from the test set would not influence prompt design. The full text of the prompts is included in the Appendix.

Additionally, we define a consistent JSON structure and enforce structured output on the models when available. Outputs were then post-processed so that column names aligned with the evaluation data, and units were equivalent aligned across rows (e.g., converting degrees Kelvin to Celsius).

We used multimodal models to process image inputs, and text-only models to process XML inputs. For GPT-4o and Claude, we tested both image and XML inputs. This resulted in a total of 7 experiments for each dataset.

5.11 Results and Discussion

We summarize our high-level results in Table 5.2. Overall, while many of the models achieve impressive recall scores, the precision of all models is unusably low, with the best models only reaching the scores of 0.2, and with the precision for several models, especially in the aluminum dataset, being within a rounding error of 0. We note that models that accepted the whole paper had significantly higher scores, indicating that naive approaches to splitting documents and aggregating predictions are ultimately insufficient; to improve our baseline results, we anticipate needing better methods to aggregate recipes across different pages, or to rely on models with context lengths long enough to process whole documents.

			precision	recall	f1
Dataset	Model	Modality			
aluminum	claude	pdf	0.044	1.000	0.084
		xml	0.058	0.911	0.108
	gpt4o	pdf	0.002	0.983	0.005
		xml	0.002	0.884	0.005
	llama3.3-70	xml	0.001	0.786	0.001
	molmo	pdf	0.000	0.522	0.001
	qwen2.5VL-72	pdf	0.001	0.759	0.002
zeolite	claude	pdf	0.169	0.766	0.277
		xml	0.204	0.938	0.335
	gpt4o	pdf	0.027	0.728	0.052
		xml	0.013	0.717	0.026
	llama3.3-70	xml	0.012	0.592	0.023
	molmo	pdf	0.005	0.316	0.010
	qwen2.5VL-72	pdf	0.029	0.739	0.056

Table 5.2. Baseline results. We note that in all cases that did not accept a full PDF or XML document, precision is extremely low, indicating that naïve approaches produce significant noise.

We also note that the models achieved impressive recall scores, especially in the aluminum alloy dataset. In annotating that dataset, we found that types of information tended to be expressed in highly formalized, similar ways. Properties were typically expressed in a table, while composition was expressed in page text. This indicates LLMs are highly structure dependent: where the information falls into a suitable structure, LLMs can be effective; this performance can rapidly degrade with variance in information presentation.
To substantiate this hypothesis, we plot the scores from our highest-performing model, Claude-3.7, against the location from which the data was extracted in the Zeolite dataset in Figure Figure 5.4. Page text perhaps suffers the most from the precision issue, which could suggest that the degree of redundancy in page text is challenging for models to disambiguate; information in tables seems to achieve a better balance between precision and recall, though tables in the visual format do not achieve as high scores as in the XML format.

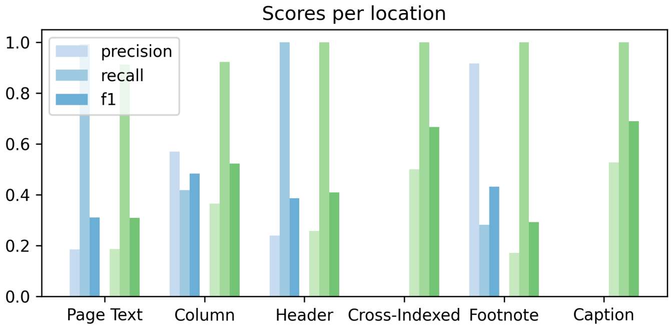


Figure 5.4. Location-based results. Blue bars indicate PDF results, green XML. Bars move from precision to recall to f1 from left to right. Cross-indexed tables and table captions are missing bars for the PDF modality because the PDFs we were able to scrape did not contain information presented in these ways.

We also see differences across modalities. Footnotes seem better parsed in the visual medium, where information in table columns and headers is easier to parse from XML, where semantic markup might assist models' processing.

5.12 Related Work

Visually Rich Document Understanding The proposed benchmark bears many similarities to work in visually rich document understanding (VRDU). Tasks in VRDU, including to answer questions based on financial documents (Chen *et al.*, 2022), or to understand forms (Jaume *et al.*, 2019), receipts (Park *et al.*, n.d.), or to perform information extraction on non-disclosure agreements and financial statements (Stanisławek *et al.*, 2021). Each of these datasets emphasizes the use of visual document features as necessary information to understand the documents' contents. However, while many of these tasks focus on documents that have been scanned and had OCR applied to them, we focus in this paper on scientific documents that are natively digital. We note, however, that scientific literature from before digital typesetting remains of significant interest in many fields.

Scientific Document Understanding Separately from document understanding tasks more generally, work on understanding scientific documents is a growing field. Work like VILA (Shen *et al.*, 2022) implements document structure recognition on scientific publications, and DDD has historically relied on information extraction tools like ChemDataExtractor (Mavracic *et al.*, 2021; Swain & Cole, 2016) and MatSciBERT (Gupta *et al.*, 2022).

5.13 Conclusion

In this paper, we repurpose two tabular datasets for data-driven design in materials science as a benchmark for multimodal document understanding. We scope the problem to be whole-paper understanding tasks, in which models are expected to pull from a variety of information contexts to satisfy the goal, and expose two settings, multimodal and text-only. In evaluating a model on our benchmark, we see that while models attain high recall, potentially due to the redundancy of information across pages and chunks, their precision leaves much to be desired. We argue that benchmarks oriented towards data-driven design should be strong candidates on which to focus effort in information extraction, both to advance the state of the art in NLP, and for the utility to materials science.

5.14 Limitations

We see two major limitations on the scope of work in this paper. Firstly, we were not able to implement alternate paradigms of model prompting in our baselines. While the naive baseline does establish that a reasonable but not especially sophisticated method fails to produce the desired output, this is not a complete characterization of the current state of LM research.

Secondly, while we attempt to cover multiple domains, the range of materials science schemas varies more than we capture in our paper. In future work, we hope to extend this benchmark to further schemas.

5.15 Zeolite Dataset Details

In this section, we provide a more detailed accounting of the zeolite dataset and the columns contained in it. Zeolites are produced by combining reaction components into a gel, which is then heat treated to grow crystals. Typically, the reaction components include precusor materials that act sources of silicon, germanium, and other elements, sources of OH^- ions, an acid, water, and an organic structure directing agent, or OSDA, which encourages crystal growth in specific ways. These make up the columns of the dataset, which we individually describe in table Table 5.3.

Dataset Columns	**Description**
Si	The molar fraction of silicon in the reaction gel. This is the basis for normalization, and is always 1 when silicon is present. If not, the gel is normalized to the quantity of germanium. This quantity is numeric.
Ge, Al, OH, H2O, HF, SDA, B	The molar fraction of each of these components. This molar fraction is calculated based on the precursor materials that are a source for those ions; using e.g. Al_2O_3, for instance, results in twice the quantity of Al^{3+} ions as the quantity of powder used. These quantities are numeric.
Time	The time, in hours, that the gel is processed for. This is usually expressed in days, and needs to be normalized to hours.
Temp	The temperature, in Celsius, that the gel is processed at. This is typically expressed in papers in either Celsius or Kelvin.
SDA Type	The name of the SDA used in the production process. This is typically given as a chemical name, and sometimes an abbreviation.
Extracted	The products extracted from the reaction. Zeolites are typically described by a three-letter code and number provided by the International Zeolite Association (IZA). Where zeolites are not produced in a reaction, the product is usually described as ”amorphous”.

Table 5.3. Descriptions of the columns in the zeolite dataset

5.16 Aluminum Alloy Dataset Details

In this section, we provide a more detailed accounting of the aluminum alloy dataset and the columns contained in it. This aluminum alloy dataset describes the physical characteristics of aluminum alloys relative to their composition. In table Table 5.4, we describe the groups of columns of the dataset.

Dataset Column	Description
AA	The named aluminum alloy series being used in this paper. These series are defined by the Aluminum Association (AA), and are typically a 4-number designations. See e.g. \url{https://www.aluminum.org/industry-standards}
Temper	The tempering process used on the aluminum alloy. This is typically expressed as a suffix of -T<number>on the series designation.
YS [MPa]	The yield strength of the alloy, measured in megapascals (MPa).
UTS [MPa]	The ultimate tensile strength of the alloy, measured in megapascals (MPa).
elong [%]	The degree that the alloy will elongate before fracturing.
Hardness	The measured hardness of the alloy.
Hardness UNIT	The unit in which hardness is measured.
Has comp [True / False / Nominal]	Whether the composition is measured in the paper's experiments (indicated by TRUE), is assumed based on the starting alloy (nominal), or entirely absent (FALSE)
Element columns (Cu, Mn, etc.)	The percentage weight of the given element measured. These will sum up to less than 100; the remainder is aluminum.

Table 5.4. Description of columns in the aluminum dataset.

5.16.1 Worked Example: Zeolite Table

Figure Figure 5.3 represents indices 375-390 from our dataset. We reproduce the first four rows of this table here, and demonstrate how to extract the relevant columns in the first row. For easy comparison, we additionally present an un-annotated version of Figure 5.3 here as

Figure 5.5.

If present, the silicon content is always the basis of normalization, and so receives a value of 1 in the `Si` column. This therefore leads us to normalize the germanium value, in the ratio of Si:Ge 0.4:0.6, to 0.667. This paper uses neither aluminum nor boron, leading to 0 values for both of those. Water and HF content are similarly normalized by dividing by 0.6.

In the table in Figure Figure 5.3, the *R* column is interpreted as the OSDA, even though this is not specified in the paper. This is a common substitution, alongside others, such as using “T” as the basis for normalization. We therefore use the values in the *R* column for the SDA value.

Text found elsewhere on the page provides additional information that must be incorporated. Synthesis paragraph 2.1 implies that the OSDA is also the source of OH^- ions: “and 3-ethyl-1-methyl-3H-imidazol-1-ium bromide (98%, Solvionic), which was transformed into its OH^- form by ion exchange in water.” The time and temperature (170°C for 14 days) are from the same paragraph; 14 days must be normalized to 336 hours.

The name of the OSDA is specified in the table caption. The names of the products are extracted into column S, but must be expanded using the table footnotes to indicate that “Arg” is argutite, and “Q” is quartz.

This table demonstrates several of the challenges in this dataset, from table understanding, to resolving in-table references, having conventional knowledge, and using contextual text that is not explicitly part of the table being considered or extracted.

Si	**Ge**	**Al**	**OH**	**H_2O**	**HF**	**SDA**	**B**	**Time**	**Temp**	**SDA Type**	**Extracted**
1	0.667	0	0.8335	33.34	0	0.8335	0	336	170	3-ethyl-1-meth...	TON+MFI+argutite
1	0.667	0	1.667	33.34	0	1.667	0	336	170	3-ethyl-1-meth...	MFI+unknown
1	0	0	0.5	8	0.5	0.5	0	336	170	3-ethyl-1-meth...	Amorphous
1	0.25	0	0.625	10	0.625	0.625	0	336	170	3-ethyl-1-meth...	IM-16+unknown

Table 5.5. Sample rows from our dataset, filtered from Jensen *et al.* (2019). This table represents the first four rows of the table seen in Figure Figure 5.3. For space, we omit the columns where we describe where the data was located.

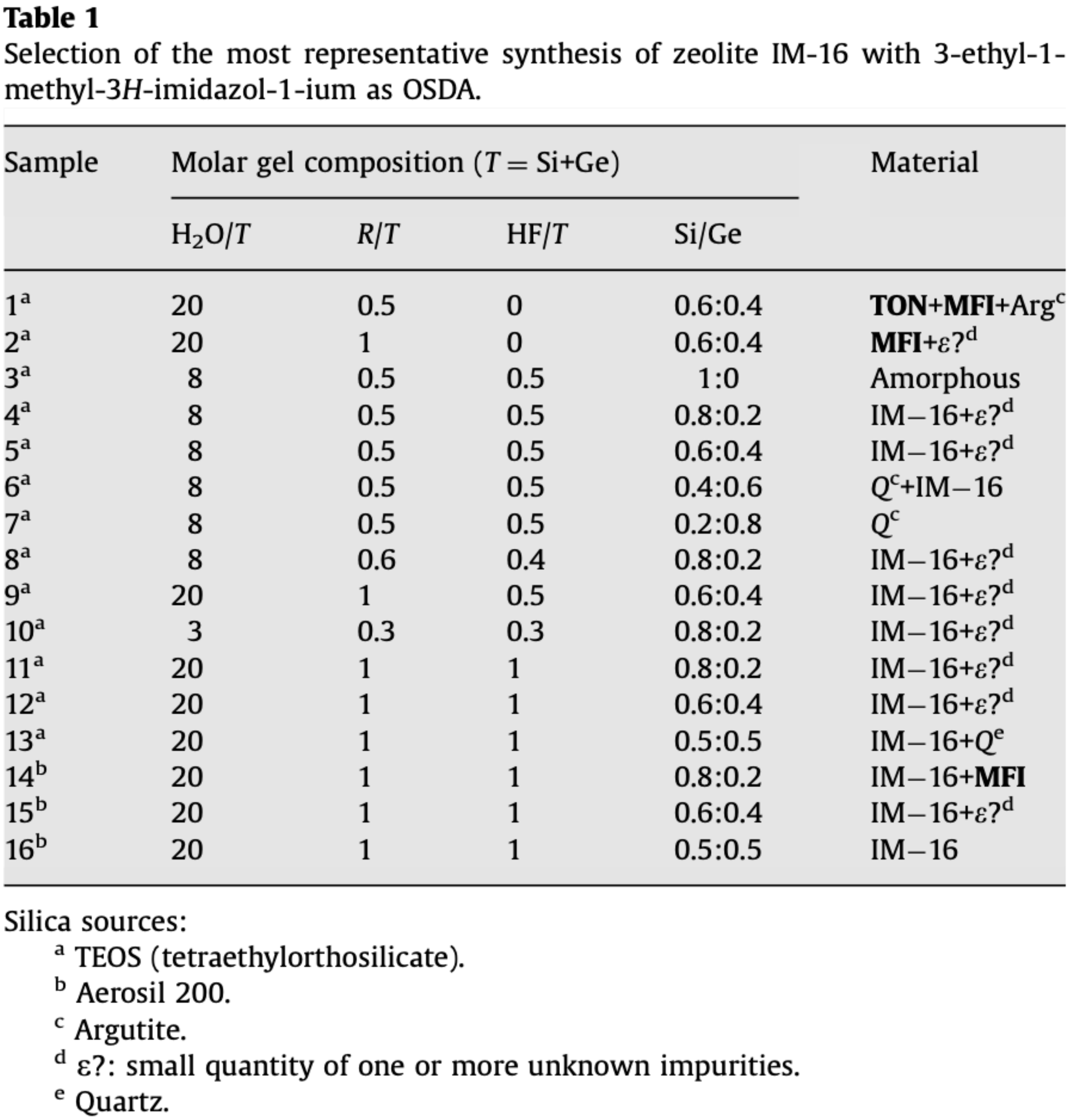


Table 1

Selection of the most representative synthesis of zeolite IM-16 with 3-ethyl-1-methyl-3*H*-imidazol-1-ium as OSDA.

Sample	Molar gel composition (T = Si+Ge)				Material
	H_2O/T	R/T	HF/T	Si/Ge	
1[a]	20	0.5	0	0.6:0.4	**TON**+**MFI**+Arg[c]
2[a]	20	1	0	0.6:0.4	**MFI**+ε?[d]
3[a]	8	0.5	0.5	1:0	Amorphous
4[a]	8	0.5	0.5	0.8:0.2	IM–16+ε?[d]
5[a]	8	0.5	0.5	0.6:0.4	IM–16+ε?[d]
6[a]	8	0.5	0.5	0.4:0.6	*Q*[c]+IM–16
7[a]	8	0.5	0.5	0.2:0.8	*Q*[c]
8[a]	8	0.6	0.4	0.8:0.2	IM–16+ε?[d]
9[a]	20	1	0.5	0.6:0.4	IM–16+ε?[d]
10[a]	3	0.3	0.3	0.8:0.2	IM–16+ε?[d]
11[a]	20	1	1	0.8:0.2	IM–16+ε?[d]
12[a]	20	1	1	0.6:0.4	IM–16+ε?[d]
13[a]	20	1	1	0.5:0.5	IM–16+*Q*[e]
14[b]	20	1	1	0.8:0.2	IM–16+**MFI**
15[b]	20	1	1	0.6:0.4	IM–16+ε?[d]
16[b]	20	1	1	0.5:0.5	IM–16

Silica sources:

[a] TEOS (tetraethylorthosilicate).

[b] Aerosil 200.

[c] Argutite.

[d] ε?: small quantity of one or more unknown impurities.

[e] Quartz.

Figure 5.5. Example table from the dataset, reproduced from Lorgouilloux *et al.* (2009, Table 1).

Chapter 6

Collage: Decomposable Rapid Prototyping for Co-Designed Information Extraction on Scientific PDFs (Complete)

Modified from a paper published in Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025) (Gururaja *et al.*, 2025b)

6.1 Overview

In this chapter, we present Collage, a system designed for the interactive evaluation and co-design of information extraction algorithms on PDF content. Collage is a web-based frontend tool that visualizes each step of an information extraction pipeline, and provides a shared representation to both end-users and engineers of the pipeline that allows for fine-grained feedback and collaboration. Collage has been extensively used in our collaborations with the Army Research Laboratory, which funded the work.

6.2 Introduction

In recent years, systems based on large language models (LLMs) have broadened the public visibility of developments in NLP. With the advent of tools that have publicly accessible, user-friendly interfaces, experts in specialized domains outside NLP are empowered to use and evaluate these models inside their domains, for example to automatically mine insights from scientific literature. Further, an increasing number of these tools are multimodal, handling not only text, but frequently images, or even PDFs directly. However, despite the accessibility of these tools, the processing pipelines they employ remain as end-to-end black boxes and provide little interpretability or debuggability in case of failure. Further, these

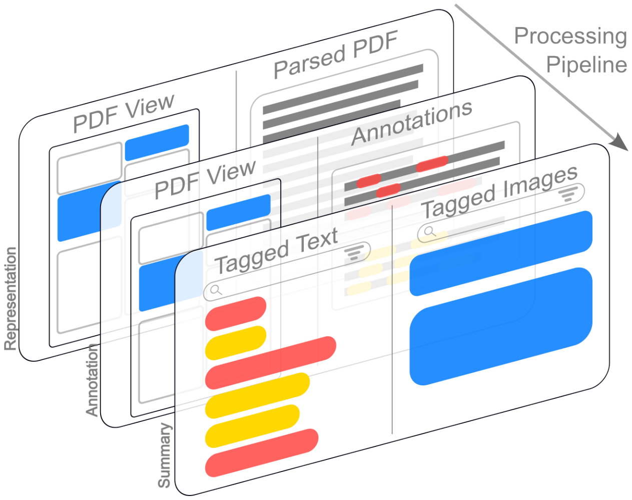


Figure 6.1. Collage allows users to inspect multiple models in different modalities by presenting a stage-by-stage, decomposed view of the PDF modeling pipeline. Here, we see a PDF composed of text and tables, with entities from different models shown in red and yellow. The summary view shows extracted content, while annotations and inspection views allow the user to step back in the modeling pipeline

systems usually rely only on large, deployed models, potentially leaving other user priorities, such as interpretability, efficiency, or domain specialization, unaddressed.

Domain specific research in domains like clinical (Naumann *et al.*, 2023), legal (Preoțiuc-Pietro *et al.*, 2023), and scientific (Cohan *et al.*, 2022; Knoth *et al.*, 2020) NLP have long histories. Models in these areas remain less accessible; in order to run and evaluate these models on your own data, custom code is often needed. Further, because many of these models are text-only, evaluating their results in the context of their eventual use — for example, directly on a PDF — poses a challenge.

This paper presents Collage, a tool that facilitates the rapid prototyping, visualization, and comparison, of multiple models across modalities on the contents of scientific PDF documents. Collage was designed to address the interface between developers of NLP-based tools for scientific documents and the scientists who are the intended users of those tools. To address scientists' needs, we ground our design in a series of interviews with domain experts in multiple fields, with a particular focus on materials science. Further, in cases where model results may not meet scientists' or developers' expectations, we visualize the intermediate representation at each step, giving the user a granular view of the modeling pipeline, allowing shared debugging processes between developers and users. Collage is domain-agnostic, and can visualize any model that conforms to one of its three interfaces - for token classification models, text generation models, and image/text multimodal models.

We provide implementations of these interfaces that allow the use of any HuggingFace token classifier, multiple LLMs, and several additional models without requiring users to write any code. All of the interfaces are easily implemented, and we provide instructions and reference implementations in our repository [19].

6.3 Motivation

Collage is based on collected themes from interviews with 15 professionals across materials science, law, and policy, in which the authors ask about their practices for working with large collections of documents. For a reasonable scope, we focus on the 9 materials scientists in our sample, whose responses concern their process of literature review. We focus on three themes that emerged consistently from these interviews to inform our design of Collage:

Varied focuses. One of the most prominent themes to emerge in our interviews is the variety of focuses that scientists, even in very closely related subfields, can have when reading a paper and evaluating it for relevance to their purpose. While many participants focused on paper metadata, such as the reputation of the publication venue or citation count, others focused on cues from within the content of the paper. For the design of Collage, we focus on accelerating co-design of models that address specific information extraction needs on paper content, by reducing the burden of deploying new models on PDF content, and providing a shared, user-friendly view of the results upon which scientists and developers can base subsequent efforts.

Information in tables. As pointed out above, many of our participants relied heavily on information provided in tables, rather than solely in the document text. As such, an important concern in the design of Collage would be to allow multimodality in the models that it interfaces with and visualizes.

Older documents. Our participants noted that they regularly work with documents across a wide time range. Several participants noted that the work that they relied on most frequently were technical reports from the 1950s to the 1970s. These reports are now digitized, but are otherwise highly variable in their accessibility to modern processing tools: The OCR used when digitizing them can be inaccurate, they often contain noise in the scanned images, and layouts are less standardized. This can lead to confusion on whether issues with performance are the fault of models themselves, or preprocesing choices that cause that degraded performance. We therefore aim to provide an interface that allows users to inspect intermediate stages of processing, to better understand where a model may have failed, and what subsequent development should target next: whether better performing models, or better

[19] https://github.com/slab-cmu/collage

preprocessing.

6.4 Design and Implementation

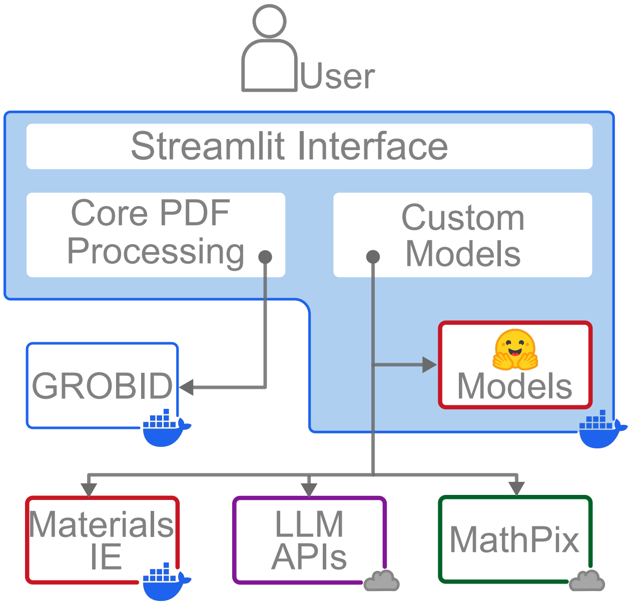


Figure 6.2. System architecture with currently implemented models. All custom models implement our interfaces, outline color indicates which: Token Classification, Text Generation, or Image Processing. 🐳 indicates components running in the same Docker container, and ☁ indicates models running in the cloud. "Materials IE" refers to materials-specific models, like ChemDataExtractor.

We conceptualize our system in three parts: PDF representation, which parses and makes the content of PDFs easily available to downstream usage; modeling, i.e. applying multiple models to that PDF representation, backed by common software interfaces, which facilitate the rapid extension of the set of available models; and a frontend graphical interface that allows users to visualize and compare the results of those models on uploaded PDFs. We discuss the design choices and implementation details of each stage in the following subsections, and show an architectural overview in Figure Figure 6.2.

6.5 PDF Representation

To produce a PDF representation amenable to our later processing, we build a pipeline on top of the PaperMage library (Lo *et al.*, 2023), which provides a convenient set of abstractions for handling multimodal PDF content. PaperMage allows the definition of `Recipes`, i.e. combinations of processing steps that can be reused. We base our pipeline off of its

CoreRecipe pipeline, which identifies visual and textual elements of a paper, such as tables and paragraphs.

We then introduce several new components to the CoreRecipe, to make the paper representation more suitable to our use case. First, we introduce a parser based on Grobid (GRO, 2008–2023), which provides a semantic grouping of paragraphs into structural units, allowing us to segment processing and results by paper section. Second, to address issues with text segmentation in scientific documents, we replace PaperMage's default segmenter (based on PySBD) with a SciBERT Beltagy *et al.* (2019)-based SciSpaCy Neumann *et al.* (2019) pipeline.

At the end of this stage of processing, we have the PaperMage representation of a document, in the form of Entity objects, organized in Layers. Entity objects can be e.g. individual paragraphs by section or index, images of tables, and individual sentences.

6.6 Modeling and Software Interfaces

To facilitate the easy implementation of new information extraction tools, we define common interfaces that simplify the process of adding additional processing to a document's content. These interfaces standardize three kinds of annotation on PDF content, allowing users convenient access to the PDF's content as images or strings (though they can access the PaperMage representation) and automatically handling visualization in several supported formats. This requires users to implement only a few simple functions in the modalities their models already use. All models currently in Collage are implementations of these interfaces. We describe the interfaces, the requirements for implementation, and current implementations below. All interfaces are defined in the papermage_components/interfaces package of our repository. In order to add a new custom processor, users must define a class that extends one of the interfaces specified below, and then register their predictor in the local_model_config.py module.

Token Classification Interface: This interface is intended for any model that produces annotations of spans in text, i.e. most "classical" NER or event extraction models. Users are required to extend the TokenClassificationPredictorABC class and override the tag_entities_in_batch method, which takes a list of strings to tag, and produces a list of lists of tagged entities per-sentence. Tagged entities are expected to have the start and end character offsets, and the interface's code automatically handles mapping indices from the sentence level to the document level, and visualizing annotated text using the displaCy visualizer [20].

[20] https://demos.explosion.ai/displacy-ent

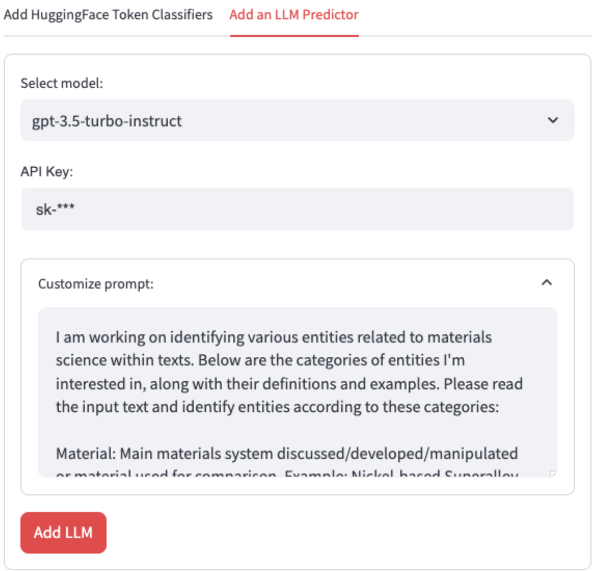
```
class LiteLlmCompletionPredictor(TextGenerationPredictorABC):
  def __init__(
    self,
    model_name: str,
    api_key: str,
    prompt_generator_function: Callable[[str], List[LLMMessage]],
    entity_to_process="reading_order_sections",
  ):
    super().__init__(entity_to_process)
    self.model_name = model_name
    self.api_key = api_key
    self.generate_prompt = prompt_generator_function

  def generate_from_entity_text(self, entity: Entity) -> str:
    messages = [asdict(m) for m in self.generate_prompt(entity.text)]
    llm_response = completion(
      model=self.model_name, api_key=self.api_key, messages=messages,
         max_tokens=2500
    )
    response_text = llm_response.choices[0].message.content
    return response_text
```

Figure 6.3. Partial implementation of the `TextGenerationPredictor` to allow LLM predictions given an `Entity` extracted from the PDF. `LLMMessage` is a data class wrapper around the system and user messages for LLMs in the OpenAI format. Not shown are the property declarations; full listing can be found in our code repository.

To demonstrate this interface, we provide two implementations: one with a common materials information extraction system, ChemDataExtractor2 (Mavracic *et al.*, 2021; Swain & Cole, 2016), which we wrap in a simple REST API and Dockerize to streamline environment and setup, as well as a predictor that can apply any HuggingFace model that conforms to the `TokenClassification` task on the HuggingFace Hub[21].

Text Generation Interface: Given the prominence of large language model-based approaches, this interface is designed to allow for text-to-text prediction. Users are required to extend the `TextGenerationPredictorABC` class, and to implement the `generate_from_entity_text()` method, which takes and returns a string. This basic setup allows users to e.g. prompt an LLM and display the raw response. A popular prompting method, however, is to request structured data e.g. in the form of JSON. To accommodate this, and to allow for aggregating LLM predictions into a table, users can also implement the `postprocess_text_to_dict()`

[21] Model list available here.



Select model: gpt-3.5-turbo-instruct

API Key: sk-***

Customize prompt:

I am working on identifying various entities related to materials science within texts. Below are the categories of entities I'm interested in, along with their definitions and examples. Please read the input text and identify entities according to these categories:

Material: Main materials system discussed/developed/manipulated or material used for comparison. Example: Nickel-based Superalloy

Figure 6.4. LLM Selector, as it appears in the File Upload view. Users specify an LLM to query, enter their API key, customize the prompt for an LLM, and repeat for any number of LLMs and prompts.

method. The default implementation of this method attempts to deserialize the entirety of the LLM response into a dictionary, but users can implement custom logic.

Our implementation of this interface uses LiteLLM[22], a package that allows accessing multiple commercial LLM services behind the same API. We allow users to specify the endpoint or model, their own API key, and a prompt, and display predictions from that model. We show a partial implementation of this predictor in Figure Figure 6.3, and a sample of its results in Figure Figure 6.5.

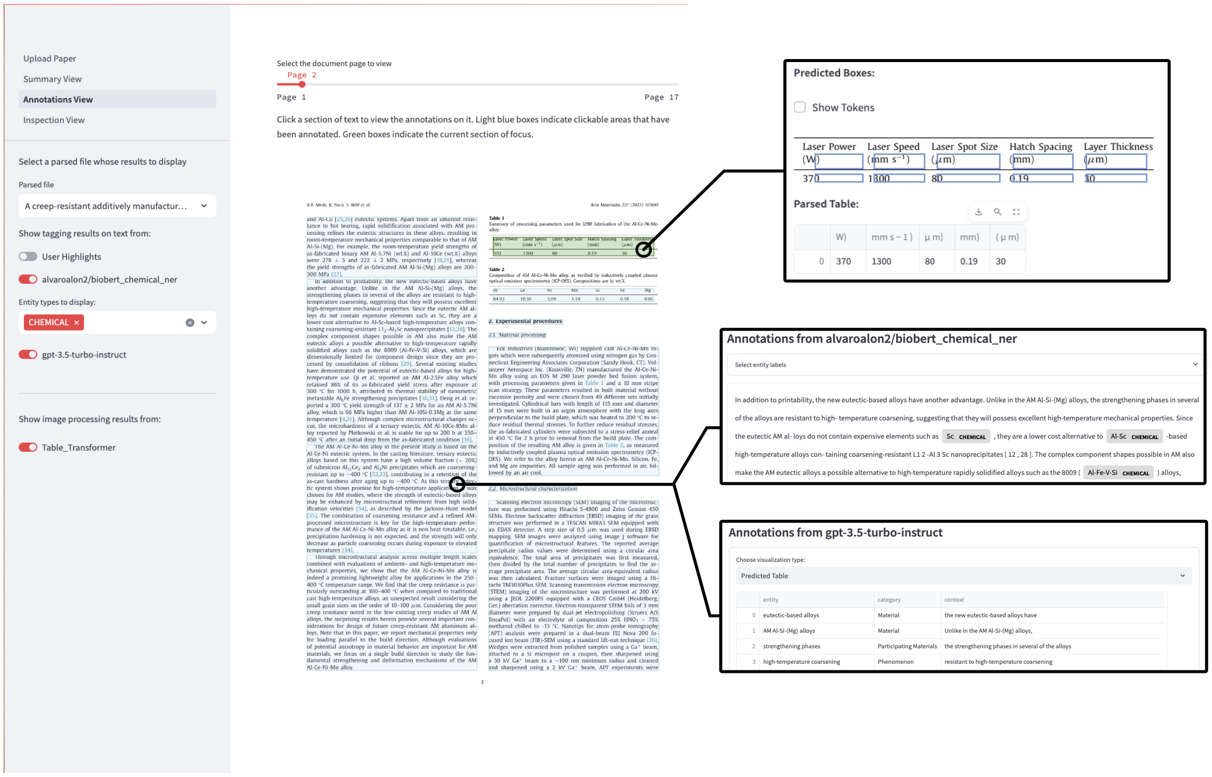
Image Prediction Interface: Given the focus on tables and charts that many of our interview participants discussed, and the fact that table parsing is an active research area, we additionally provide an interface for models that parse images, the `ImagePredictorABC` in order to handle multimodal processing, including tables. This interface allows users two options of method to override: In cases where only image inputs are needed (e.g. if a table extractor performs its own OCR), the `process_image()` method; in cases where the method is inherently multimodal, implementors can instead override the `process_entity()` method, which allows them full access to PaperMage's multimodal `Entity` representation. This interface requires implementors to return at least one of three types of data: a raw string representation, which we view as useful for e.g. image captioning tasks; a tabular

[22] https://docs.litellm.ai/

dictionary representation, for the case of table parsing; or a list of bounding boxes, in the case of models that segment images. Implementations of this interface are free to return more than one type of output; all of them will be visualized in the frontend.

We demonstrate implementations of both types. For raw image outputs, we implement a predictor that calls the MathPix API[23], a commercial service for PDF understanding. For the multimodal approach, we implement a predictor that builds on the Microsoft Table Transformer model (Smock *et al.*, 2023). This model predicts bounding boxes around table cells, which we then cross-reference with extracted PDF text in the PaperMage representation to provide parsed table output. An example of parsed table output from this predictor can be seen in figure Figure 6.5.

6.7 Visualization Frontend



Select a parsed file whose results to display

Parsed file: A creep-resistant additively manufactur…

Show tagging results on text from:

- User Highlights (off)
- alvaroalon2/biobert_chemical_ner (on)

Entity types to display: CHEMICAL ×

- gpt-3.5-turbo-instruct (on)

Show image processing results from:

- Table_Transformer (on)

Select the document page to view: Page 2 (Page 1 – Page 17)

Click a section of text to view the annotations on it. Light blue boxes indicate clickable areas that have been annotated. Green boxes indicate the current section of focus.

Predicted Boxes:

□ Show Tokens

Laser Power (W)	Laser Speed (mm s^{-1})	Laser Spot Size (μm)	Hatch Spacing (mm)	Layer Thickness (μm)
370	1300	80	0.19	30

Parsed Table:

	W)	mm s − 1)	µ m)	mm)	(µ m)
0	370	1300	80	0.19	30

Annotations from alvaroalon2/biobert_chemical_ner

Select entity labels

In addition to printability, the new eutectic-based alloys have another advantage. Unlike in the AM Al-Si-(Mg) alloys, the strengthening phases in several of the alloys are resistant to high- temperature coarsening, suggesting that they will possess excellent high-temperature mechanical properties. Since the eutectic AM al- loys do not contain expensive elements such as Sc CHEMICAL , they are a lower cost alternative to Al-Sc CHEMICAL -based high-temperature alloys con- taining coarsening-resistant L1 2 -Al 3 Sc nanoprecipitates [12 , 28]. The complex component shapes possible in AM also make the AM eutectic alloys a possible alternative to high-temperature rapidly solidified alloys such as the 8009 (Al-Fe-V-Si CHEMICAL) alloys,

Annotations from gpt-3.5-turbo-instruct

Choose visualization type: Predicted Table

	entity	category	context
0	eutectic-based alloys	Material	the new eutectic-based alloys have
1	AM Al-Si-(Mg) alloys	Material	Unlike in the AM Al-Si-(Mg) alloys,
2	strengthening phases	Participating Materials	the strengthening phases in several of the alloys
3	high-temperature coarsening	Phenomenon	resistant to high-temperature coarsening

Figure 6.5. The annotations view. On the left, a screenshot showing the sidebar, allowing file and model selection, and the left pane, a visualization of the PDF with clickable regions highlighted. On the right, screenshots showing visualizations from the Table Transformer model with bounding boxes and parsed table (top), a HuggingFace transformer model with token-level tags (middle), and GPT-3.5 Turbo, with JSON output parsed into a table (bottom).

[23] https://mathpix.com/

We present the results of the PDF processing in an interactive tool built using Streamlit[24] that allows the user – whether scientist or developer – to upload a PDF, define a processing pipeline, and inspect the results of that processing pipeline at each stage. More concretely, after the paper is uploaded and processed, we present the results of the pipeline in three views, in decreasing order of abstraction from the paper. The intention of this is to first show the user the potential output of their chosen pipeline for a given paper, then allow them to inspect each step of the pipeline that led to that final output. Each view is described in more detail below.

File Upload and Processing. The first view we present to a user allows them to upload a file, and to define the processing pipeline applied to that file. Basic PDF processing is always performed, and users can then toggle which custom models will be run. Users can additionally specify any number of HuggingFace token classification models or LLMs with the provided widget, which allows users to search the HuggingFace Hub, select LLMs, and customize the prompts for them. We show a view of the LLM model selector in Figure Figure 6.4.

File Overview. This view presents the high-level extracted information from the paper, as candidates for what could be shown to the user as part of their search process. In particular, we show a two-column view, with tables of tagged entities from both token-level predictors and LLMs on the left, and the processed content of images on the right. Users can filter based on sections, to e.g. find materials mentioned in the methods section of a paper. If the user finds the content extracted with the pipeline useful, the model and processing pipeline could be further developed into a more integrated prototype. If not, the user can proceed to the succeeding views, to see where models may have failed.

Annotations. This view allows the user to compare the results of models in the context of the PDF. We present another two-column view, in which the PDF is visualized on the left, and allows the user to select a paragraph or table at a time, and visualize the results of each model on it. In the case of text annotation, we visualize the entities identified by token prediction models as well as predictions from LLMs. In the case of images, all of the available output types from the image processing interface are visualized. We show a composite screenshot of this interface in Figure Figure 6.5.

Representation Inspection. This view presents visualization of the PDF representation available to any downstream processing that the user might select. In the sidebar, users can choose to visualize any PaperMage `Layer`, i.e. set of `Entity` objects, tagged by the basic processing steps. Then, in a view similar to the raw annotations view, they can see all of those entities highlighted on the PDF in the left-side column. Once the user selects

[24] https://streamlit.io

an object, they see the raw content extracted from that object in the right-side column, in the form of its image representation and the text extracted from it, along with the option to view how the text is segmented into sentences. This view allows users to inspect how the PDF processing choices may have affected the text they send to models, which often have significant effects on their downstream performance (Camacho-Collados & Pilehvar, 2018).

6.8 Addressing Needs from Interviews

Our system is specifically designed to respond to the concerns raised in our interviews. First, to accommodate the varied processes of materials scientists, we design interfaces that allow for easy implementation of new models into our framework; our existing implementations of those interfaces also allow for the application of multiple LLMs and HuggingFace models directly in the context of the PDFs under review. This allows users to search for and evaluate models that suit their existing workflows. For tables, we both provide an interface and implementations that allow the comparison of proprietary and open-source table parsing systems. Extending this work to new table models and evaluating them is simplified by our software and visualization interfaces. Our inspection view is designed to address concerns about older PDFs: in being able to inspect the results of processing, users and engineers of this system can identify failure modes in both the upstream and downstream processing.

6.9 Co-design with Collage

In this section, we walk through an example of how Collage might facilitate the development of an information extraction pipeline for a materials scientist. In this scenario, Bob, a materials scientist, wishes to extract the synthesis parameters of a class of materials called zeolites from a dataset of PDFs from the 1980s to the 2010s. Papers discussing Zeolite synthesis often report parameters both in the text of the paper as well as in tables, so multimodal extraction is crucial. He has worked with Alice, an NLP developer, before but they have not yet collaborated on this project.

Evaluating off-the-shelf models. Bob begins in Collage by trying to see if there is an existing model that already works for his case. Using the HuggingFace model selector in the upload paper view, he searches for tagging models, but only finds models trained on general scientific or biomedical text, not materials. He is, however, able to write a prompt for an LLM model to extract this information, and he adds predictors that call out to two popular commercial models to extract the information that he's interested in. He uploads a recent paper that he's been reading, and waits for Collage to process it.

Finding modeling opportunities. Once Collage has processed the paper, Bob heads to the **summary view**, and compares the results from the two commercial models. He's able to view the parameters that they extracted, filtered by section, to develop an understanding of what heuristics might get him the information he wants: parameters identified in the related works section, for example, are frequently irrelevant to his search. In the summary view, he's also able to see the tables that Collage has identified and parsed with the TableTransformer and MathPix models, along with their labels and captions, and the tagged bounding boxes for the table cells.

To make sure those annotations are reasonable, he heads to the **annotations view**, where he can visualize the extracted information side-by-side with the original PDF content, and compare the annotations from his two LLMs. He's also able to check whether the table detection model has predicted sensible bounding boxes that both don't exclude content like table footnotes, but also don't include irrelevant, non-table content. He notes that while the table parsing from both models is reasonable, the paper he's reading reports values in ratios that may not be comparable across papers. To have a single pipeline that produces normalized results, he'd like to use a multimodal LLM, but in Collage currently, LLMs can only be applied to text. He decides to get in touch with Alice, to see if she can develop an LLM-based table information extraction model.

Prototype model development. Alice begins work on a table information extraction tool, but there are a lot of possible options to evaluate: should she use a multimodal model and process the table in image format? Should she linearize the table into text, and have a text-only LLM work with it? In Collage, both options involve little more than implementing the LLM call, so it's easy to do both and then compare. For the multimodal case, Alice extends the **image predictor interface**, which allows her to receive as input the cropped image of any element on the page and pass that to an LLM; for the text-only case, Alice can easily access the underlying document representation use the already identified and parsed tables (which are in a DataFrame-compatible format) and convert them into markdown for her linearization. She is able to return a dictionary in the same schema for both predictors, which will automatically be visualized in the frontend as a Pandas dataframe. She commits her code, registers the predictors, and asks Bob to take a look in the Collage interface.

In-context evaluation. Bob then re-processes his paper through Collage, making sure to check the boxes for Alice's new table parsing predictors. In the summary view, he's able to compare the predicted, normalized tables to the original PDF, to verify that the models are performing the normalization correctly. He then picks the better performing model, and asks Alice to create a pipeline that can process his entire dataset. Alice is able to take the predictor, add it to the PaperMage recipe that underlies Collage, and run it over Bob's set of PDF documents, adding a step to export the parsed tables that Bob saw in the Collage interface.

Diagnosing errors. Bob looks through the parsed tables from processing all of the PDFs, and notes that for the older PDFs, the parsed content doesn't look right. He'd like to diagnose the problem. Because the processing that Alice and Bob run on these documents is the same as that underlying Collage, the results can be visualized in the tool, even if they were not directly processed through it. Bob loads the representation of the parsed older document, and is able to view the results from the model that didn't look right. While the bounding boxes for the table look correct in the annotations view, he's also able to see in the **inspection view** that the text detected within the table has not been correctly OCR'd. He can now contact Alice to see if there's a fix for that problem, but in the meantime, he can examine the visualizations for his PDFs to understand how the publication year might affect whether the deployed suite of models can correctly extract and normalize information, and what the cutoff year might be for the results to be trustworthy.

In this case, Collage enables Bob to self-serve cutting-edge NLP for his own use case, requiring that he involve Alice only when Collage's functionality needs extension. When that happens, Bob and Alice can both see results in the same interface, and can discuss errors and how to prioritize new work. When Alice develops new predictors to address Bob's needs, she is required to do no PDF processing or visualization, which are built into the tool, and Bob can evaluate and compare the results of these new predictors in the same interface he's been using the whole time. For debugging, both Bob and Alice have access to the same representation and visualization as a shared source of truth, and collaborate to involve both NLP and subject matter expertise in how to fix the problem. Collage can accelerate the process of collaboration between NLP developers and scientists, allowing for co-design and rapid prototyping with a shared representation.

6.10 Related Work

Collage situates itself at the intersection of tools that offer reading assistance for scientific PDFs and tools that partially automate the process of literature review by means of information extraction. Tools for scientific PDFs often focuses on interfaces that augment the existing PDF with new information, such as citation contexts (Nicholson *et al.*, 2021; Rachatasumrit *et al.*, 2022), or highlights that aid skimming (Fok *et al.*, 2023). However, most of these works are designed around and purpose-built for specific models. By contrast, Collage draws from projects like PaperMage (Lo *et al.*, 2023), by attempting to be model-agnostic, while at the same time providing a visual interface to prototype and evaluate those models.

Scientific information extraction and literature review automation also have long histories. Collage's focus on materials science was driven by the field's existing investment into data-driven design (Himanen *et al.*, 2019; Olivetti *et al.*, 2020), which focuses on using information extraction tools to build up knowledge graphs to inform future materials research. This adds to the existing body of work in chemical and material information extraction, including works like ChemDataExtractor (Mavracic *et al.*, 2021; Swain & Cole, 2016) and MatSciBERT (Gupta *et al.*, 2022). Works like Dagdelen *et al.* (2024b) showcase the growing interest in LLM-based extraction; as LLMs increasingly become multimodal, this capability is likely to be used for tasks like scientific document understanding. While all of these tools are intended to be applied to documents from the materials science domain, they do not share an interface: most tools expect plain text, some, like ChemDataExtractor allow HTML and XML documents, and some work with images. Collage aims to be a platform on which multiple competing approaches can be evaluated, regardless of the input and output formats they require.

6.11 Conclusion

In this work, we present Collage, a system designed to facilitate co-design and rapid prototyping of mixed modality information extraction on PDF content between scientists and NLP developers. We focus on a case study in the materials science domain, that allows materials scientists to evaluate models for their ability to assist in literature review. We intend for this work to be a platform on which to evaluate further modeling work in this area.

6.12 Ethics and Broader Impacts

Our interview protocol was evaluated and approved by the Carnegie Mellon University Institutional Review Board as STUDY2023_00000431.

In developing a tool to facilitate the automated processing of scientific PDFs, we feel that it is important to acknowledge that that automation may propagate the biases of the underlying models. Particularly in the case of English that does not reflect the training corpora that models were built on top of, models can perform poorly, leading to fewer results from those papers, and the potential to inadvertently exclude them. However, we hope that in providing a tool to inspect model outputs before such automation tools are deployed, that we can encourage critical evaluation and uses of these tools.

Chapter 7

Proposed: Tractor Beam: User-Customizable, On-Device Information Extraction (Proposed)

7.1 Introduction

As AI-based tools have seen greater integration with real-world processes, applications like AI code assistance have seen the development of purpose built UX that takes advantage of application-specific constraints and affordances. AI code assistants, for example, rather than returning code in a chat context or as one-off throwaway artifacts in a browser, now integrate closely with developer tools. Laban *et al.* (2024) argue similarly for new interaction paradigms for writing assistance, replacing chat-based iteration over text with diff views, and tracking LLM-generated content to scaffold error checking by the author.

AI-based tooling for information access, however, remains largely confined to chat-based interfaces. Chat-based tools for information access abound, from new interfaces to search and information synthesis, as in the case of so-called ”deep research” tools (Google, 2024; OpenAI, 2025; Shao *et al.*, 2025, inter alia), to tools that allow you to ask an AI systems questions about individual PDFs[25], each of which purport to synthesize information from one or more papers into summary text.

While the ability to prompt models for behavior is a powerful paradigm for specifying model behavior, text-based specification for information access creates a number of different problems. Most notably, reading a model-based summary of text introduces an information bottleneck, in that the user must either assume the model correctly parsed and returned the information relevant to them, or must manually verify the presence of that information, thereby reducing or eliminating the benefit of the summary; recent models also suffer

[25] e.g. ChatPDF, or Semantic Scholar's Ask this paper feature

from unpredictable personalization behavior, producing stereotypical responses in response to identity cues. By contrast, information extraction systems have historically been grounded in the source document, by means of tags to the original content.

Secondly, chat-based information access systems make iteration difficult. Diagnosing information extraction systems' accuracy when extracted information is presented out of context is difficult, and creates friction in understanding and implementing updates to the prompt or schema. Further, the only affordance by which to specify updates is through the chat interface itself, where purpose-built systems might offer more flexibility, whether allowing editing the response, user highlights, or deliberate reannotation of data.

This chapter therefore presents the preliminary design for TractorBeam, a tool designed to facilitate user-specified information extraction with flexible, editable schemas. This process moves iteration on, and evaluation of models into the hands of individual users with a rapid feedback loop.

7.2 Motivation

This tool is motivated by the four principles outlined in chapter 4. In particular, we wish to focus on easy access, personalizability, iterativity, and social awareness. In this section, we outline how each of these concerns in turn motivate our proposed design, which we outline in the next section.

Accessibility. For this tool to address the specificity-accessibility tradeoff, such that it both addresses specific user needs and is easily accessed, it must not require writing any code, and ideally must be as closely integrated to users' existing sets of software.

Personalizability and Iterativity. For information extraction to be truly personal, users must be able to define their own sets of entities and relations between them, without regard for standard practice in information extraction. Entities may correspond to ”typical” named entity classes, like names or locations, but may also be abstract. Further, users must be able to easily modify the schemas of entities and relations at any time, such that the tool can continue to accurately reflect their mental model, and the tool must also be designed to support iteration as a first-class construct, rather than have users work around the system to determine how updates impact their tools. Additionally, users should be able to specify entity classes in multiple ”modalities” — both by prompting models initially, but also by correcting models and creating original annotations.

Social Awareness. Social awareness is a more complicated concern. Because this envisioned system is downstream of existing models, and is aiming to run on-device, it is necessarily constrained by the capabilities of those models. As established by previous research, models are often incapable of understanding the pragmatic and social aspects of information tasks that expert users engage in (Gururaja *et al.*, 2025a), can be vulnerable to spin unless specifically prompted otherwise (Yun *et al.*, 2025), and can exhibit variable behavior based on subtle user cues (Kantharuban *et al.*, 2025; Li *et al.*, 2024b). The goal for a system like TractorBeam, therefore, is to allow users to attempt to use models for complex, socially aware annotation, evaluate them in-context, and, should the model ultimately not be capable of that annotation, stop using the model. Opting out selectively — i.e. allowing users to configure *which* parts of models they will use is a key design goal for TractorBeam. However, using existing modes of interaction for this type of annotation process — i.e. chat windows that provide citations at best — causes unnecessary friction in evaluating models and arriving at conclusions about their utility. We therefore propose that TractorBeam operate primarily at the level of highlights on PDFs. Highlights provide numerous advantages. They can still speed up reading and comprehension by providing additional levels of indexing on document contents, in the style of the Semantic Reader (Lo *et al.*, 2024) and its skimming assistance features. However, crucially, they do not remove engagement with the original document, allow users to evaluate the quality of the annotations directly in context, as in Collage, and also enable browsing and discovery, where users find information without it already being part of a targeted search.

Additional Concerns. Underlying the four concerns outlined above, a key priority for TractorBeam is to maintain user agency at all times. Rather than being "human-in-the-loop," which tends to assume a loop that exists above the level of the individual person, e.g. at the level of a team or company's representation of the project, TractorBeam aims to center user agency at all points. This includes individual user control of schemas, and annotations; this also extends to user choices about modeling. Given that proprietary and hosted models can often be changed or deprecated with little notice, and this can cause significant behavior changes that will impact downstream modeling choices. As a result, TractorBeam will prioritize being local-first [26], using primarily on-device models, and allowing users to use API-based models for either annotation or synthetic data generation if they wish. Over time, we would like to further extend this system to a collaborative one, where users can share and collaboratively edit schemas in a decentralized fashion.

7.3 Proposed Design

Given the design priorities discussed above, our proposed architecture is as a browser extension consisting of three components: a PDF viewer component that allows visualization and

[26] Drawing inspiration from a growing community, e.g. https://www.inkandswitch.com/essay/local-first/

annotation of highlights, a schema editor, which we place in a side pane, and a background worker that runs annotation and allows in-browser storage. We describe each of these components below, along with additional features we consider necessary for a first version of TractorBeam.

PDF Viewer. For the PDF viewer, we plan to use a modified version of the pdf.js library [27]. We plan to implement an annotation frontend similar to the already validated design of the Hypothesis collaborative annotation tool [28], in which users can both select spans of text within a PDF, or annotate bounding-box style areas for entities of interest. We will further allow connections between boxes to facilitate relation labeling.

Schema Editor. Also similar to hypothesis, we will implement a schema editor that allows users to add and remove schemas, and within them, to add, remove, or edit entity classes and relations at any time. Schemas are defined in this tool as collections of entity classes and relations. Entity classes are defined by a name, a description (that is later used in prompts to extract the entities), and zero or more instances: highlighted text or rectangles that are examples of the class. Relations are defined as directed connections between two entity types. We do not plan to support hyperedge-type relations in the initial version of the tool.

Background Worker. The bacgkround worker is responsible for both storing schemas (and their associated instances), and also for running the annotation processes that provide automatic annotations on PDFs. This annotation process will support multiple backends: local LMs, both in browser, e.g. Chrome's built-in LM API [29], on-device using technologies like Ollama[30] or Llamafile [31], and remote, API-based models, whether proprietary or self-hosted. Part of this project will be determining what methods of model adaptation will be possible in this framework: can we perform prompt optimization or LoRA (Hu *et al.*, 2021) finetuning on-device?

Storage and Data Export. For storage, we plan to use the IndexedDB API, which is consistent across browsers, and provides large object storage in a noSQL-style database. Because user agency remains a first priority, allowing data export is also necessary: users should be able to take the data that they have annotated on PDFs and move it to other platforms seamlessly. We plan to provide data export in JSON, and CSV/Excel formats, as well as facilitating automatic uploading to online hosting solutions like the HuggingFace

[27] https://github.com/mozilla/pdf.js

[28] https://web.hypothes.is/

[29] https://developer.chrome.com/docs/ai/prompt-api

[30] https://ollama.com

[31] https://github.com/mozilla-ai/llamafile

Hub [32].

Data Versioning. To enable iteration, we see the need for versioning as a primary concern. For example, if a user has updated their schema, how can they compare annotations between one version of a schema and another? We envision a dependency-graph style data model, where schemas depend on their constituent components — names, definitions, manual annotations — models depend on schemas, and automatic evaluations depend on models. Any edits to this data model at a lower level cause a new version of the higher-level constructs, which can then be visualized side by side. Users should also be able to specify which parts of the old object should be brought through to the new object.

7.4 Potential Extensions

Full collaborative development. Local-first software, while prioritizing local development, nonetheless maintains a strong emphasis on collaboration, albeit without a central source of truth. A feasible extension to TractorBeam would be allowing users to share schemas, with each user of a schema maintaining their own master copies, that could then be reconciled if users wish. This would allow indiviudal users to share models, while maintaining their own mental model as a source of truth.

Better schema editing. The highest friction aspect of TractorBeam as currently envisioned is schema editing, especially as schemas start to contain appreciable numbers of annotated instances. Rather than requiring users to relabel instances, or decide whether a given instance still fits within an updated class, TractorBeam could be extended to assist users in reconciling updates to their schemas with the data — in (Schlangen, 2021)'s words, reconciling the intensional and extensional definitions of their self-created dataset.

7.5 Evaluation

TractorBeam is intended to be used long-term for information extraction tasks that are repeated over many PDFs. As such, we intend to run both an initial user study, to evaluate whether the tool fulfills its initial design goals, and a longer-term diary study with a population of users to study whether the system remains usable long-term.

[32] https://huggingface.co/docs/hub/en/index

Chapter 8

Proposed: Strategies for On-Device Information Extraction (Proposed)

8.1 Introduction

The previous chapter establishes a proposed system for user-specifiable information extraction. An optional part of the system as conceptualized is the ability to perform the information extraction on the user's device, primarily to ensure the user's full control over the modeling workflow that they define. However, on-device inference presents a number of additional advantages in terms of user privacy, cases where there are strict rules about where data can be processed (academic publishers, for instance, often limit the upload of academic papers to proprietary LLM services), and ongoing cost to operate.

There have been numerous promising developments over the past few years that align with the goal of on-device information extraction: a strong research push towards smaller language models, both in parameter count and through quantization; methods like LoRA (Hu *et al.*, 2021) and QLoRA (Dettmers *et al.*, 2023), which lower the resource demands of fine-tuning; prompt optimization methods like GEPA (Agrawal *et al.*, 2025), alongside whole frameworks for using them (Khattab *et al.*, 2023; Sarmah *et al.*, 2024); software that supports both CPU-only and accelerated inference for desktop operating systems [33]; and finally hardware that has evolved to increasingly support on-device AI capabilities in a variety of form factors.

However, much of the field of on-device machine learning and inference focuses on edge computing devices with stricter resource constraints — mobile phones, embedded systems — than users of systems like TractorBeam might have, and primarily focus on training models for inference on such hardware. By contrast, we anticipate TractorBeam users being primarily desktop computer users needing to use these models in a web browser, and work effectively in a few-shot, cross-domain setting. As such, this proposed work aims to evaluate

[33] e.g. Ollama and llama.cpp

the frontier of cross-domain on-device information extraction with the compute budget and software constraints of a desktop-based browser environment.

We plan for our primary contributions to be a characterization of the compute budget/performance tradeoff for methods increasing in complexity, from zero-shot prompting to full model fine-tuning, across a range of English language datasets from standard datasets like CoNLL-03 (Tjong Kim Sang & De Meulder, 2003) to specialized domain datasets from materials science and bioinformatics. We intend to compare the frontier of performance to the state-of-the-art in proprietary large LM baselines. In performing this evaluation, we also hope to contribute a harness for repeatedly performing this evaluation with newer models and techniques as they are developed, as well as software support for less common methods like in-browser finetuning.

8.2 Planned Experiments

We plan to perform a full factorial experiment setup across three axes: base model, task adaptation, and dataset. For base models, while we will primarily consider decoder-only transformers, as they are the most popular choice for language tasks, we will additionally consider alternate architectures that are within the size range to be locally runnable. In particular, we plan to evaluate ModernBERT (Warner *et al.*, 2024), GLiNER2 (Zaratiana *et al.*, 2025), both encoder-only architectures, and Flan-T5 Large and XL, an encoder-decoder architecture. For each of these architectures, we will apply task adaptations if possible.

8.2.1 Datasets

We intend to run this evaluation across six datasets, which we choose for their diversity both in the domains that they sample text from, but also in the label space, to better measure models' ability to generalize to new, highly specific schemas. For this paper, we consider "information extraction" to be a combination of up to three tasks: named entity recognition (NER), relation extraction (RE), and coreference resolution. To limit complexity, the initial scope of this paper is to evaluate only NER performance; we therefore allow datasets that only contain NER data. Many of these datasets also contain data in multiple langages; we only consider English-language data in this work.

We begin with two general-schema NER datasets that consist primarily of news and internet text in the CoNLL-2003 shared task (Tjong Kim Sang & De Meulder, 2003) and the OntoNotes 5.0 dataset (Weischedel, Ralph *et al.*, 2013). the CoNLL-2003 dataset uses what has become a canonical NER schema: locations (LOC), named people (PER), organizations

(ORG), and a miscellaneous category. OntoNotes uses a much broader schema of 11 entity types, but also annotates 7 ”value” types, including dates, percentages, and monetary quantities. We evaluate performance on the full set of 18 labels.

We then move to the CrossNER (Liu *et al.*, 2021), which consists of Wikipedia text from five domains: politics, natural science, music, literature, and artificial intelligence, each with specific schema of between 9 and 17 classes derived from the DBPedia ontology. For this dataset, the authors attempt to build data subsets with low degrees of vocabulary overlap to make the domain subsets more distinct to represent the process of crosslingual transfer.

For more domain-specific datasets, we draw on three corpora across the medical, legal, and scientific fields, each annotated by domain expert annotators: NeuroTrialNER (Doneva *et al.*, 2024), which annotates diseases and treatments, a dataset that measures information extraction on wills (Kwak *et al.*, 2023), and the MSMentions (O'Gorman *et al.*, 2021) corpus, which annotates procedural information in materials science papers. Each of these is highly domain specific, with little overlap in annotation scheme with the more general-domain datasets. We therefore aim to use them as proxies for the degree to which models can acceptably model idiosyncratic user annotations.

8.2.2 Evaluation

We plan to use a standard F1 measure as the primary metric for this task, as each dataset above does. We plan to average F1 scores across tasks for each combination of model and task adaptation method. Additionally, considering that user needs may vary and that models often end up with sometimes extreme precision and recall scores, we plan to report F0.5 and F2 as supplementary metrics.

Each approach will also be characterized in terms of the computational budget it requires. Naively, this would involve reporting counts of floating point operations involved for each method. However, given the degree of hardware specialization that can result in unintuitive performance bottlenecks and uneven speedup of certain kinds of operations (Fernandez *et al.*, 2023), we will report this both in terms of FLOP count and wall clock time on common, commodity hardware platforms.

8.2.3 Task Adaptation

We plan to test a range of task adaptation methods on each of the base model architectures discussed above, with specific base models to be chosen when we start running experiments for this project. In particular, we plan to experiment with the following adaptations:

Zero-shot prompting (with prompt optimization). For this method, we plan to prompt models zero-shot, while using prompt optimization techniques such as those built into (Khattab *et al.*, 2023).

In-context learning (ICL) and example selection. Building on the zero-shot prompts, we then plan to optimize few-shot ICL prompts, using a few different example selection methods.

Prefix tuning(Li & Liang, 2021). Prefix tuning presents an attractive alternative to full-model fine-tuning methods, instead learning task-specific embeddings that are prepended to a prompt. This necessitates less storage for task adaptation: rather than adapters or full model weights, only task-specific vocabulary vectors need to be stored.

LoRA finetuning While full-model finetuning is likely beyond both the hardware and software capabilities of our setup, we intend to investigate the possibility of LoRA finetuning, which eases the hardware burden.

Synthetic data While data privacy concerns often limit how data can be used with proprietary language model services, generating synthetic data for either ICL or for fine-tuning is a promising avenue for using large language models without relying on them for end-task inference.

8.2.4 Baselines

For our baseline approach, we plan to use the strongest performing of a number of popular proprietary large models, under the strongest performing, zero-shot (i.e. non-finetuned) task adaptation method.

Bibliography

2008–2023. *GROBID*. https://github.com/kermitt2/grobid.

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dandelion, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, & Zheng, Xiaoqiang. 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Abdalla, Mohamed, Wahle, Jan Philip, Ruas, Terry, Névéol, Aurélie, Ducel, Fanny, Mohammad, Saif M., & Fort, Karën. 2023. *The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research*.

Acemoglu, Daron. 2024 (May). *The Simple Macroeconomics of AI*.

Adams, Lisa, Busch, Felix, Han, Tianyu, Excoffier, Jean-Baptiste, Ortala, Matthieu, Löser, Alexander, Aerts, Hugo JWL, Kather, Jakob Nikolas, Truhn, Daniel, & Bressem, Keno. 2024. LongHealth: A Question Answering Benchmark with Long Clinical Documents. *arXiv preprint arXiv:2401.14490*.

Agrawal, Lakshya A., Tan, Shangyin, Soylu, Dilara, Ziems, Noah, Khare, Rishi, Opsahl-Ong, Krista, Singhvi, Arnav, Shandilya, Herumb, Ryan, Michael J., Jiang, Meng, Potts, Christopher, Sen, Koushik, Dimakis, Alexandros G., Stoica, Ion, Klein, Dan, Zaharia, Matei, & Khattab, Omar. 2025 (July). *GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning*. arXiv:2507.19457 [cs].

Ahmed, Nur, & Wahed, Muntasir. 2020. *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*.

AI4Science, Microsoft Research, & Quantum, Microsoft Azure. 2023 (Dec.). *The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4*. arXiv:2311.07361 [cs].

Anderson, Ashton, Jurafsky, Dan, & McFarland, Daniel A. 2012. Towards a Computational History of the ACL: 1980-2008. *Pages 13–21 of: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Jeju Island, Korea: Association for Computational Linguistics.

Anderson, Barrett R., Shah, Jash Hemant, & Kreminski, Max. 2024 (June). Homogenization Effects of Large Language Models on Human Creative Ideation. *Pages 413–425 of: Creativity and Cognition*. arXiv:2402.01536 [cs].

Asai, Akari, Min, Sewon, Zhong, Zexuan, & Chen, Danqi. 2023. Retrieval-based Language Models and Applications. *Pages 41–46 of:* Chen, Yun-Nung (Vivian), Margot, Margot, & Reddy, Siva (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*. Toronto, Canada: Association for Computational Linguistics.

Ashok, Dhananjay, & Lipton, Zachary C. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Attar, Rony, & Fraenkel, Aviezri S. 1977. Local feedback in full-text retrieval systems. *Journal of the ACM (JACM)*, **24**(3), 397–417.

Balepur, Nishant, Rudinger, Rachel, & Boyd-Graber, Jordan Lee. 2025. Which of These Best Describes Multiple Choice Evaluation with LLMs? A) Forced B) Flawed C) Fixable D) All of the Above. *Pages 3394–3418 of:* Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, & Pilehvar, Mohammad Taher (eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics.

Bartoldson, Brian R., Kailkhura, Bhavya, & Blalock, Davis. 2023. Compute-Efficient Deep Learning: Algorithmic Trends and Opportunities. *Journal of Machine Learning Research*, **24**(122), 1–77.

Bavaresco, Anna, Bernardi, Raffaella, Bertolazzi, Leonardo, Elliott, Desmond, Fernández, Raquel, Gatt, Albert, Ghaleb, Esam, Giulianelli, Mario, Hanna, Michael, Koller, Alexander, Martins, Andre, Mondorf, Philipp, Neplenbroek, Vera, Pezzelle, Sandro, Plank, Barbara, Schlangen, David, Suglia, Alessandro, Surikuchi, Aditya K, Takmaz, Ece, & Testoni, Alberto. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *Pages 238–255 of:* Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, & Pilehvar, Mohammad Taher (eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vienna, Austria: Association for Computational Linguistics.

Bazerman, Charles. 1985. Physicists Reading Physics: Schema-Laden Purposes and Purpose-Laden Schema. *Written Communication*, **2**(1), 3–23. Publisher: SAGE Publications Inc.

Beltagy, Iz, Lo, Kyle, & Cohan, Arman. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bennett, James, & Lanning, Stan. 2007. The netflix prize.

Birhane, Abeba, Kalluri, Pratyusha, Card, Dallas, Agnew, William, Dotan, Ravit, & Bao, Michelle. 2022. The Values Encoded in Machine Learning Research. *Page 173–184 of: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery.

Blili-Hamelin, Borhane, Graziul, Christopher, Hancox-Li, Leif, Hazan, Hananel, El-Mhamdi, El-Mahdi, Ghosh, Avijit, Heller, Katherine, Metcalf, Jacob, Murai, Fabricio, Salvaggio, Eryk, Smart, Andrew, Snider, Todd, Tighanimine, Mariame, Ringer, Talia, Mitchell, Margaret, & Dori-Hacohen, Shiri. 2025 (July). *Stop treating 'AGI' as the north-star goal of AI research.* arXiv:2502.03689 [cs].

Bowman, Samuel R., & Dahl, George. 2021a. What Will it Take to Fix Benchmarking in Natural Language Understanding? *Pages 4843–4855 of:* Toutanova, Kristina, Rumshisky, Anna, Zettlemoyer, Luke, Hakkani-Tur, Dilek, Beltagy, Iz, Bethard, Steven, Cotterell, Ryan, Chakraborty, Tanmoy, & Zhou, Yichao (eds), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics.

Bowman, Samuel R., & Dahl, George. 2021b. What Will it Take to Fix Benchmarking in Natural Language Understanding? *Pages 4843–4855 of: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics.

Bowman, Samuel R., Angeli, Gabor, Potts, Christopher, & Manning, Christopher D. 2015. A large annotated corpus for learning natural language inference. *Pages 632–642 of:* Màrquez, Lluís, Callison-Burch, Chris, & Su, Jian (eds), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics.

Brown, John Seely, & Duguid, Paul. 1996. The Social Life of Documents; introduction by Esther Dyson. *First Monday*, May.

Camacho-Collados, Jose, & Pilehvar, Mohammad Taher. 2018. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *Pages 40–46 of:* Linzen, Tal, Chrupa\la, Grzegorz, & Alishahi, Afra (eds), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics.

Campbell, Steve, Greenwood, Melanie, Prior, Sarah, Shearer, Toniele, Walkem, Kerrie, Young, Sarah, Bywaters, Danielle, & Walker, Kim. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing*, **25**(8), 652–661.

Chairs 2025, Communications. 2025 (Sept.). *Reflecting on the 2025 Review Process from the Datasets and Benchmarks Chairs – NeurIPS Blog*.

Challapally, Aditya, Pease, Chris, Raskar, Ramesh, & Chari, Pradyumna. STATE OF AI IN BUSINESS 2025.

Chandak, Nikhil, Goel, Shashwat, Prabhu, Ameya, Hardt, Moritz, & Geiping, Jonas. 2025 (July). *Answer Matching Outperforms Multiple Choice for Language Model Evaluation*. arXiv:2507.02856 [cs].

Chen, Zhiyu, Chen, Wenhu, Smiley, Charese, Shah, Sameena, Borova, Iana, Langdon, Dylan, Moussa, Reema, Beane, Matt, Huang, Ting-Hao, Routledge, Bryan, & Wang, William Yang. 2022 (May). *FinQA: A Dataset of Numerical Reasoning over Financial Data*. arXiv:2109.00122 [cs].

Church, Kenneth Ward. 2017. Emerging trends: I did it, I did it, I did it, but. . . *Natural Language Engineering*, **23**(3), 473–480.

Church, Kenneth Ward. 2018. Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering*, **24**(1), 155–160.

Cohan, Arman, Feigenblat, Guy, Freitag, Dayne, Ghosal, Tirthankar, Herrmannova, Drahomira, Knoth, Petr, Lo, Kyle, Mayr, Philipp, Shmueli-Scheuer, Michal, de Waard, Anita, & Wang, Lucy Lu (eds). 2022. *Proceedings of the Third Workshop on Scholarly Document Processing*. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Court, Callum J., & Cole, Jacqueline M. 2018. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data*, **5**(1), 180111. Publisher: Nature Publishing Group.

Crawford, Kate, & Paglen, Trevor. 2019 (Sept.). *Excavating AI: The Politics of Training Sets for Machine Learning*.

Dagdelen, John, Dunn, Alexander, Lee, Sanghoon, Walker, Nicholas, Rosen, Andrew S, Ceder, Gerbrand, Persson, Kristin A, & Jain, Anubhav. 2024a. Structured information extraction from scientific text with large language models. *Nature Communications*, **15**(1), 1418.

Dagdelen, John, Dunn, Alexander, Lee, Sanghoon, Walker, Nicholas, Rosen, Andrew S., Ceder, Gerbrand, Persson, Kristin A., & Jain, Anubhav. 2024b. Structured information extraction from scientific text with large language models. *Nature Communications*, **15**(1), 1418. Publisher: Nature Publishing Group.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, & Fei-Fei, Li. 2009 (June). ImageNet: A large-scale hierarchical image database. *Pages 248–255 of: 2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919.

Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, & Zettlemoyer, Luke. 2023 (May). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv:2305.14314 [cs].

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Pages 4171–4186 of:* Burstein, Jill, Doran, Christy, & Solorio, Thamar (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Ding, Ning, Xu, Guangwei, Chen, Yulin, Wang, Xiaobin, Han, Xu, Xie, Pengjun, Zheng, Haitao, & Liu, Zhiyuan. 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. *Pages 3198–3213 of:* Zong, Chengqing, Xia, Fei, Li, Wenjie, & Navigli, Roberto (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, & Darrell, Trevor. 2013 (Oct.). *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*. arXiv:1310.1531 [cs].

Doneva, Simona Emilova, Ellendorff, Tilia, Sick, Beate, Goldman, Jean-Philippe, Cannon, Amelia Elaine, Schneider, Gerold, & Ineichen, Benjamin Victor. 2024. NeuroTrialNER: An Annotated Corpus for Neurological Diseases and Therapies in Clinical Trial Registries. *Pages 18868–18890 of:* Al-Onaizan, Yaser, Bansal, Mohit, & Chen, Yun-Nung (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics.

Ethayarajh, Kawin, & Jurafsky, Dan. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. *Pages 4846–4853 of:* Webber, Bonnie, Cohn, Trevor, He, Yulan, & Liu, Yang (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Fernandez, Jared, Kahn, Jacob, Na, Clara, Bisk, Yonatan, & Strubell, Emma. 2023. The Framework Tax: Disparities Between Inference Efficiency in NLP Research and Deployment. *Pages 1588–1600 of:* Bouamor, Houda, Pino, Juan, & Bali, Kalika (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.* Singapore: Association for Computational Linguistics.

Fishman, Nic, & Hancox-Li, Leif. 2022. Should attention be all we need? The epistemic and ethical implications of unification in machine learning. *Pages 1516–1527 of: 2022 ACM Conference on Fairness, Accountability, and Transparency.*

Fleisig, Eve, Blodgett, Su Lin, Klein, Dan, & Talat, Zeerak. 2024 (May). *The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels.* arXiv:2405.05860 [cs].

Fok, Raymond, Kambhamettu, Hita, Soldaini, Luca, Bragg, Jonathan, Lo, Kyle, Hearst, Marti, Head, Andrew, & Weld, Daniel S. 2023. Scim: Intelligent Skimming Support for Scientific Papers. *Pages 476–490 of: Proceedings of the 28th International Conference on Intelligent User Interfaces.* Sydney NSW Australia: ACM.

Francis, Winthrop Nelson. 1964. *A Standard Sample of Present-day English for Use with Digital Computers.* Brown University. Google-Books-ID: KupWAAAAMAAJ.

Frohmann, Bernd. 2004. Introduction: From Information to Documentation. *Pages 3–22 of: Deflating Information.* From Science Studies to Documentation. University of Toronto Press.

Galloway, Alison. 2005. Non-Probability Sampling. *Pages 859–864 of:* Kempf-Leonard, Kimberly (ed), *Encyclopedia of Social Measurement.* New York: Elsevier.

Gardner, Matt, Grus, Joel, Neumann, Mark, Tafjord, Oyvind, Dasigi, Pradeep, Liu, Nelson F., Peters, Matthew, Schmitz, Michael, & Zettlemoyer, Luke. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. *Pages 1–6 of: Proceedings of Workshop for NLP Open Source Software (NLP-OSS).* Melbourne, Australia: Association for Computational Linguistics.

Gema, Aryo Pradipta, Leang, Joshua Ong Jun, Hong, Giwon, Devoto, Alessio, Mancino, Alberto Carlo Maria, Saxena, Rohit, He, Xuanli, Zhao, Yu, Du, Xiaotang, Madani, Mohammad Reza Ghasemi, Barale, Claire, McHardy, Robert, Harris, Joshua, Kaddour, Jean, Krieken, Emile van, & Minervini, Pasquale. 2025 (Jan.). *Are We Done with MMLU?* arXiv:2406.04127 [cs].

Glaser, Barney G, Holton, Judith, *et al.* 2004. Remodeling grounded theory. *In: Forum qualitative sozialforschung/forum: qualitative social research*, vol. 5.

Google. 2024 (Dec.). *Gemini Deep Research — your personal research assistant.*

Gubelmann, Reto, Kalouli, Aikaterini-lida, Niklaus, Christina, & Handschuh, Siegfried. 2023. When Truth Matters - Addressing Pragmatic Categories in Natural Language Inference (NLI) by Large Language Models (LLMs). *Pages 24–39 of:* Palmer, Alexis, & Camacho-collados, Jose (eds), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023).* Toronto, Canada: Association for Computational Linguistics.

Guo, Shaoru, Chen, Yubo, Liu, Kang, Li, Ru, & Zhao, Jun. 2024. NutFrame: Frame-based Conceptual Structure Induction with LLMs. *Pages 12330–12335 of:* Calzolari, Nicoletta, Kan, Min-Yen, Hoste, Veronique, Lenci, Alessandro, Sakti, Sakriani, & Xue, Nianwen (eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* Torino, Italia: ELRA and ICCL.

Gupta, Tanishq, Zaki, Mohd, Krishnan, NM Anoop, & Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, **8**(1), 102.

Gururaja, Sireesh, Bertsch, Amanda, Na, Clara, Widder, David, & Strubell, Emma. 2023. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing. *Pages 13310–13325 of:* Bouamor, Houda, Pino, Juan, & Bali, Kalika (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.* Singapore: Association for Computational Linguistics.

Gururaja, Sireesh, Gandhi, Nupoor, Milbauer, Jeremiah, & Strubell, Emma. 2025a. Beyond Text: Characterizing Domain Expert Needs in Document Research. *Pages 4732–4745 of:* Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, & Pilehvar, Mohammad Taher (eds), *Findings of the Association for Computational Linguistics: ACL 2025.* Vienna, Austria: Association for Computational Linguistics.

Gururaja, Sireesh, Zhang, Yueheng, Tang, Guannan, Zhang, Tianhao, Murphy, Kevin, Yi, Yu-Tsen, Seo, Junwon, Rollett, Anthony, & Strubell, Emma. 2025b. Collage: Decomposable Rapid Prototyping for Co-Designed Information Extraction on Scientific PDFs. *Pages 72–82 of:* Ghosal, Tirthankar, Mayr, Philipp, Singh, Amanpreet, Naik, Aakanksha, Rehm, Georg, Freitag, Dayne, Li, Dan, Schimmler, Sonja, & De Waard, Anita (eds), *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025).* Vienna, Austria: Association for Computational Linguistics.

Gururaja, Sireesh, Seo, Junwon, Lin, Hung-Yi, Milbauer, Jeremiah, Rollett, Anthony, & Strubell, Emma. 2025c (Oct.). Data-driven Design as a High-Impact, Ecologically Valid Benchmark for Document Understanding.

He, Jiangen, Ping, Qing, Lou, Wen, & Chen, Chaomei. 2019. PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology*, **70**(8), 843–857.

Head, Andrew, Lo, Kyle, Kang, Dongyeop, Fok, Raymond, Skjonsberg, Sam, Weld, Daniel S., & Hearst, Marti A. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. *In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery.

Heimerl, Florian, Han, Qi, Koch, Steffen, & Ertl, Thomas. 2016. CiteRivers: Visual Analytics of Citation Patterns. *IEEE Transactions on Visualization and Computer Graphics*, **22**, 190–199.

Hendrycks, Dan, Burns, Collin, Basart, Steven, Zou, Andy, Mazeika, Mantas, Song, Dawn, & Steinhardt, Jacob. 2020 (Oct.). Measuring Massive Multitask Language Understanding.

Hillesund, Terje. 2010. Digital reading spaces: How expert readers handle books, the Web and electronic paper. *First Monday*, Apr.

Himanen, Lauri, Geurts, Amber, Foster, Adam Stuart, & Rinke, Patrick. 2019. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, **6**(21), 1900808.

Hofer, Maximilian, Kormilitzin, Andrey, Goldberg, Paul, & Nevado-Holgado, Alejo. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.

Hooker, Sara. 2021. The Hardware Lottery. *Communications of the ACM*, **64**(12), 58–65.

Howard, Jeremy, & Ruder, Sebastian. 2018 (May). *Universal Language Model Fine-tuning for Text Classification*. arXiv:1801.06146 [cs].

Hsu, Sheryl, Khattab, Omar, Finn, Chelsea, & Sharma, Archit. 2024. *Grounding by Trying: LLMs with Reinforcement Learning-Enhanced Retrieval*.

Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, & Chen, Weizhu. 2021 (Oct.). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685 [cs].

Hu, Edward J, Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, Chen, Weizhu, *et al.* 2022. Lora: Low-rank adaptation of large language models. *ICLR*, **1**(2), 3.

Hu, Kairui, Wu, Penghao, Pu, Fanyi, Xiao, Wang, Zhang, Yuanhan, Yue, Xiang, Li, Bo, & Liu, Ziwei. 2025 (Jan.). *Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos.* arXiv:2501.13826 [cs].

Huang, Jiaxin, Li, Chunyuan, Subudhi, Krishan, Jose, Damien, Balakrishnan, Shobana, Chen, Weizhu, Peng, Baolin, Gao, Jianfeng, & Han, Jiawei. 2020. Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978.*

Huang, Wenhao, Gu, Zhouhong, Peng, Chenghao, Li, Zhixu, Liang, Jiaqing, Xiao, Yanghua, Wen, Liqian, & Chen, Zulong. 2024 (Apr.). *AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation.* arXiv:2404.12753 [cs] version: 1.

Huang, Yupan, Lv, Tengchao, Cui, Lei, Lu, Yutong, & Wei, Furu. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *Page 4083–4091 of: Proceedings of the 30th ACM International Conference on Multimedia.* MM '22. New York, NY, USA: Association for Computing Machinery.

Jaume, Guillaume, Kemal Ekenel, Hazim, & Thiran, Jean-Philippe. 2019 (Sept.). FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. *Pages 1–6 of: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2.

Jensen, Zach, Kim, Edward, Kwon, Soonhyoung, Gani, Terry Z. H., Román-Leshkov, Yuriy, Moliner, Manuel, Corma, Avelino, & Olivetti, Elsa. 2019. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Central Science*, **5**(5), 892–899. Publisher: American Chemical Society.

Jiang, Pengcheng, Lin, Jiacheng, Wang, Zifeng, Sun, Jimeng, & Han, Jiawei. 2024. GenRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models. *Pages 2820–2837 of:* Duh, Kevin, Gomez, Helena, & Bethard, Steven (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* Mexico City, Mexico: Association for Computational Linguistics.

Jimenez, Carlos E., Yang, John, Wettig, Alexander, Yao, Shunyu, Pei, Kexin, Press, Ofir, & Narasimhan, Karthik. 2024 (Nov.). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* arXiv:2310.06770 [cs].

Kantharuban, Anjali, Milbauer, Jeremiah, Strubell, Emma, & Neubig, Graham. 2024. Stereotype or Personalization? User Identity Biases Chatbot Recommendations. *arXiv preprint arXiv:2410.05613.*

Kantharuban, Anjali, Milbauer, Jeremiah, Sap, Maarten, Strubell, Emma, & Neubig, Graham. 2025. Stereotype or Personalization? User Identity Biases Chatbot Recommendations. *Pages 24418–24436 of:* Che, Wanxiang, Nabende, Joyce, Shutova, Ekaterina, & Pilehvar, Mohammad Taher (eds), *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics.

Katz, Uri, Vetzler, Matan, Cohen, Amir, & Goldberg, Yoav. 2023. NERetrieve: Dataset for Next Generation Named Entity Recognition and Retrieval. *Pages 3340–3354 of:* Bouamor, Houda, Pino, Juan, & Bali, Kalika (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics.

Katz, Uri, Levy, Mosh, & Goldberg, Yoav. 2024. Knowledge Navigator: LLM-guided Browsing Framework for Exploratory Search in Scientific Literature. *arXiv preprint arXiv:2408.15836*.

Khattab, Omar, Santhanam, Keshav, Li, Xiang Lisa, Hall, David, Liang, Percy, Potts, Christopher, & Zaharia, Matei. 2023 (Jan.). *Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP*. arXiv:2212.14024 [cs].

Kilgarriff, Adam. 2007. Last Words: Googleology is Bad Science. *Computational Linguistics*, **33**(1), 147–151.

Kim, Edward, Huang, Kevin, Tomala, Alex, Matthews, Sara, Strubell, Emma, Saunders, Adam, McCallum, Andrew, & Olivetti, Elsa. 2017. Machine-learned and codified synthesis parameters of oxide materials. *Scientific data*, **4**(1), 1–9.

Knoth, Petr, Stahl, Christopher, Gyawali, Bikash, Pride, David, Kunnath, Suchetha N., & Herrmannova, Drahomira (eds). 2020. *Proceedings of the 8th International Workshop on Mining Scientific Publications*. Wuhan, China: Association for Computational Linguistics.

Koch, Bernard, Denton, Emily, Hanna, Alex, & Foster, Jacob G. 2021 (Dec.). *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research*. arXiv:2112.01716 [cs].

Koch, Bernard J., & Peterson, David. 2024. From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution. Publisher: arXiv Version Number: 2.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra, & Herbst, Evan. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Pages 177–180 of: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions.* Prague, Czech Republic: Association for Computational Linguistics.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kumar, Harsh, Vincentius, Jonathan, Jordan, Ewan, & Anderson, Ashton. 2025 (Apr.). Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. *Pages 1–18 of: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* arXiv:2410.03703 [cs].

Kwak, Alice, Jeong, Cheonkam, Forte, Gaetano, Bambauer, Derek, Morrison, Clayton, & Surdeanu, Mihai. 2023. Information Extraction from Legal Wills: How Well Does GPT-4 Do? *Pages 4336–4353 of:* Bouamor, Houda, Pino, Juan, & Bali, Kalika (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023.* Singapore: Association for Computational Linguistics.

Laban, Philippe, Vig, Jesse, Hearst, Marti, Xiong, Caiming, & Wu, Chien-Sheng. 2024. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. *Pages 1–23 of: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology.* Pittsburgh PA USA: ACM.

Latour, Bruno. 1986. *Laboratory life: the construction of scientific facts.* Princeton, N.J: Princeton University Press.

Li, Jing, Sun, Aixin, Han, Jianglei, & Li, Chenliang. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, **34**(1), 50–70.

Li, Victoria, Chen, Yida, & Saphra, Naomi. 2024a. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. *Pages 6327–6345 of: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.*

Li, Victoria R, Chen, Yida, & Saphra, Naomi. 2024b. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. *Pages 6327–6345 of:* Al-Onaizan, Yaser, Bansal, Mohit, & Chen, Yun-Nung (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* Miami, Florida, USA: Association for Computational Linguistics.

Li, Wangyue, Li, Liangzhi, Xiang, Tong, Liu, Xiao, Deng, Wei, & Garcia, Noa. 2024c (May). *Can multiple-choice questions really be useful in detecting the abilities of LLMs?* arXiv:2403.17752 [cs].

Li, Xiang Lisa, & Liang, Percy. 2021 (Jan.). *Prefix-Tuning: Optimizing Continuous Prompts for Generation.* arXiv:2101.00190 [cs].

Liao, Q. Vera, & Vaughan, Jennifer Wortman. 2023. *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap.*

Liberman, Mark. Lessons for Responsible Science from DARPA's Programs in Human Language Technology.

Lincoln, Yvonna S, Lynham, Susan A, & Guba, Egon G. 2011. Paradigmatic Controversies, Contradictions, and Emerging Confluences, Revisited. *The Sage handbook of qualitative research*, **4**, 97–128.

Liu, Haotian, Li, Chunyuan, Li, Yuheng, Li, Bo, Zhang, Yuanhan, Shen, Sheng, & Lee, Yong Jae. 2024 (January). *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.*

Liu, Zihan, Xu, Yan, Yu, Tiezheng, Dai, Wenliang, Ji, Ziwei, Cahyawijaya, Samuel, Madotto, Andrea, & Fung, Pascale. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(15), 13452–13460.

Lo, Kyle, Wang, Lucy Lu, Neumann, Mark, Kinney, Rodney, & Weld, Daniel. 2020. S2ORC: The Semantic Scholar Open Research Corpus. *Pages 4969–4983 of: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics.

Lo, Kyle, Shen, Zejiang, Newman, Benjamin, Chang, Joseph Z, Authur, Russell, Bransom, Erin, Candra, Stefan, Chandrasekhar, Yoganand, Huff, Regan, Kuehl, Bailey, *et al.* 2023. PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents. *Pages 495–507 of: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.*

Lo, Kyle, Chang, Joseph Chee, Head, Andrew, Bragg, Jonathan, Zhang, Amy X., Trier, Cassidy, Anastasiades, Chloe, August, Tal, Authur, Russell, Bragg, Danielle, Bransom, Erin, Cachola, Isabel, Candra, Stefan, Chandrasekhar, Yoganand, Chen, Yen-Sung, Cheng, Evie Yu-Yen, Chou, Yvonne, Downey, Doug, Evans, Rob, Fok, Raymond, Hu, Fangzhou, Huff, Regan, Kang, Dongyeop, Kim, Tae Soo, Kinney, Rodney, Kittur, Aniket, Kang, Hyeonsu B., Klevak, Egor, Kuehl, Bailey, Langan, Michael J., Latzke, Matt, Lochner, Jaron, MacMillan, Kelsey, Marsh, Eric, Murray, Tyler, Naik, Aakanksha, Nguyen, Ngoc-Uyen, Palani, Srishti, Park, Soya, Paulic, Caroline, Rachatasumrit, Napol, Rao, Smita, Sayre, Paul, Shen, Zejiang, Siangliulue, Pao, Soldaini, Luca, Tran, Huy, van Zuylen, Madeleine, Wang, Lucy Lu, Wilhelm, Christopher, Wu, Caroline, Yang, Jiangjiang, Zamarron, Angele, Hearst, Marti A., & Weld, Daniel S. 2024. The Semantic Reader Project. *Commun. ACM*, **67**(10), 50–61.

Lorgouilloux, Yannick, Dodin, Mathias, Paillaud, Jean-Louis, Caullet, Philippe, Michelin, Laure, Josien, Ludovic, Ersen, Ovidiu, & Bats, Nicolas. 2009. IM-16: A new microporous germanosilicate with a novel framework topology containing *d4r* and *mtw* composite building units. *Journal of Solid State Chemistry*, **182**(3), 622–629.

Lu, Chris, Lu, Cong, Lange, Robert Tjarko, Foerster, Jakob, Clune, Jeff, & Ha, David. 2024 (Sept.). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.* arXiv:2408.06292 [cs].

Luccioni, Alexandra Sasha, & Crawford, Kate. 2024. The Nine Lives of ImageNet: A Sociotechnical Retrospective of a Foundation Dataset and the Limits of Automated Essentialism. *Journal of Data-centric Machine Learning Research*, Feb.

Lucy, Li, Dodge, Jesse, Bamman, David, & Keith, Katherine A. 2022. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. arXiv. Version Number: 2.

Lucy, Li, Gururangan, Suchin, Soldaini, Luca, Strubell, Emma, Bamman, David, Klein, Lauren, & Dodge, Jesse. 2024. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. *Pages 7393–7420 of:* Ku, Lun-Wei, Martins, Andre, & Srikumar, Vivek (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Bangkok, Thailand: Association for Computational Linguistics.

Lund, Niels Windfeld. 2009. Document theory. *Annual Review of Information Science and Technology*, **43**(1), 1–55. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aris.2009.1440430116.

Ma, Kaixin, Zhang, Hongming, Wang, Hongwei, Pan, Xiaoman, Yu, Wenhao, & Yu, Dong. 2023. Laser: Llm agent with state-space exploration for web navigation. *arXiv preprint arXiv:2309.08172.*

Ma, Tingting, Yao, Jin-Ge, Lin, Chin-Yew, & Zhao, Tiejun. 2021. Issues with Entailment-based Zero-shot Text Classification. *Pages 786–796 of:* Zong, Chengqing, Xia, Fei, Li, Wenjie, & Navigli, Roberto (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics.

Makhoul, John. 2021 (Aug.). *The Dawn of Benchmarking.*

Marcus, Mitchell, Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen, & Schasberger, Britta. 1994. The Penn Treebank: Annotating Predicate Argument Structure. *In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.*

Marcus, Mitchell P., Santorini, Beatrice, & Marcinkiewicz, Mary Ann. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330. Place: Cambridge, MA Publisher: MIT Press.

Mathew, Minesh, Karatzas, Dimosthenis, & Jawahar, CV. 2021. Docvqa: A dataset for vqa on document images. *Pages 2200–2209 of: Proceedings of the IEEE/CVF winter conference on applications of computer vision.*

Matos, Emanuel, Rodrigues, Mário, & Teixeira, António. 2024. Towards the automatic creation of NER systems for new domains. *Pages 218–227 of:* Gamallo, Pablo, Claro, Daniela, Teixeira, António, Real, Livy, Garcia, Marcos, Oliveira, Hugo Gonçalo, & Amaro, Raquel (eds), *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics.

Mavracic, Juraj, Court, Callum J, Isazawa, Taketomo, Elliott, Stephen R, & Cole, Jacqueline M. 2021. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling*, **61**(9), 4280–4289.

McCoy, R. Thomas, Pavlick, Ellie, & Linzen, Tal. 2019 (June). *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.* arXiv:1902.01007 [cs].

Merken, Sara. 2024. AI company Luminance raises $40 mln as contracts tech investment booms. *reuters.com*. Accessed: 2024-12-15.

Michael, Julian, Holtzman, Ari, Parrish, Alicia, Mueller, Aaron, Wang, Alex, Chen, Angelica, Madaan, Divyam, Nangia, Nikita, Pang, Richard Yuanzhe, Phang, Jason, & Bowman, Samuel R. 2022. *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey.*

Milbauer, Jeremiah, Mathew, Adarsh, & Evans, James. 2021. Aligning Multidimensional Worldviews and Discovering Ideological Differences. *Pages 4832–4845 of:* Moens, Marie-Francine, Huang, Xuanjing, Specia, Lucia, & Yih, Scott Wen-tau (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Milbauer, Jeremiah, Ding, Ziqi, Wu, Zhijin, & Wu, Tongshuang. 2023. NewsSense: Reference-free Verification via Cross-document Comparison. *Pages 422–430 of:* Feng, Yansong, & Lefever, Els (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Singapore: Association for Computational Linguistics.

Miles, Matthew B, & Huberman, A Michael. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

Min, Sewon, Chen, Danqi, Zettlemoyer, Luke, & Hajishirzi, Hannaneh. 2019. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*.

Miret, Santiago, & Krishnan, N. M. Anoop. 2024 (Sept.). *Are LLMs Ready for Real-World Materials Discovery?* arXiv:2402.05200 [cond-mat].

Mohammad, Saif M. 2020. Examining Citations of Natural Language Processing Literature. *Pages 5199–5209 of: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Mor-Lan, Guy, & Levi, Effi. 2024. Exploring Factual Entailment with NLI: A News Media Study. *Pages 190–199 of:* Bollegala, Danushka, & Shwartz, Vered (eds), *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*. Mexico City, Mexico: Association for Computational Linguistics.

Mysore, Sheshera, Jensen, Zachary, Kim, Edward, Huang, Kevin, Chang, Haw-Shiuan, Strubell, Emma, Flanigan, Jeffrey, McCallum, Andrew, & Olivetti, Elsa. 2019. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. *Pages 56–64 of:* Friedrich, Annemarie, Zeyrek, Deniz, & Hoek, Jet (eds), *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics.

Mysore, Sheshera, Jasim, Mahmood, Song, Haoru, Akbar, Sarah, Randall, Andre Kenneth Chase, & Mahyar, Narges. 2023. How Data Scientists Review the Scholarly Literature. *Pages 137–152 of: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. CHIIR '23. New York, NY, USA: Association for Computing Machinery.

Narayanan, Arvind, & Kapoor, Sayash. *AI as Normal Technology.*

Nathan, Allison, Grimberg, Jenny, & Rhodes, Ashley. 2024 (June). *Gen AI: too much spend, too little benefit?* Tech. rept.

Naumann, Tristan, Ben Abacha, Asma, Bethard, Steven, Roberts, Kirk, & Rumshisky, Anna (eds). 2023. *Proceedings of the 5th Clinical Natural Language Processing Workshop.* Toronto, Canada: Association for Computational Linguistics.

Neubig, Graham, Dyer, Chris, Goldberg, Yoav, Matthews, Austin, Ammar, Waleed, Anastasopoulos, Antonios, Ballesteros, Miguel, Chiang, David, Clothiaux, Daniel, Cohn, Trevor, Duh, Kevin, Faruqui, Manaal, Gan, Cynthia, Garrette, Dan, Ji, Yangfeng, Kong, Lingpeng, Kuncoro, Adhiguna, Kumar, Gaurav, Malaviya, Chaitanya, Michel, Paul, Oda, Yusuke, Richardson, Matthew, Saphra, Naomi, Swayamdipta, Swabha, & Yin, Pengcheng. 2017. DyNet: The Dynamic Neural Network Toolkit. *arXiv preprint arXiv:1701.03980.*

Neumann, Mark, King, Daniel, Beltagy, Iz, & Ammar, Waleed. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Pages 319–327 of:* Demner-Fushman, Dina, Cohen, Kevin Bretonnel, Ananiadou, Sophia, & Tsujii, Junichi (eds), *Proceedings of the 18th BioNLP Workshop and Shared Task.* Florence, Italy: Association for Computational Linguistics.

Newman-Griffis, Denis, Lehman, Jill Fain, Rosé, Carolyn, & Hochheiser, Harry. 2021 (Apr.). *Translational NLP: A New Paradigm and General Principles for Natural Language Processing Research.* arXiv:2104.07874 [cs].

Nicholson, Josh M., Mordaunt, Milo, Lopez, Patrice, Uppala, Ashish, Rosati, Domenic, Rodrigues, Neves P., Grabitz, Peter, & Rife, Sean C. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, **2**(3), 882–898.

Northcutt, Curtis G., Athalye, Anish, & Mueller, Jonas. 2021 (Nov.). *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks.* arXiv:2103.14749 [stat].

O'Gorman, Tim, Jensen, Zach, Mysore, Sheshera, Huang, Kevin, Mahbub, Rubayyat, Olivetti, Elsa, & McCallum, Andrew. 2021. MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text. *Pages 1337–1352 of:* Moens, Marie-Francine, Huang, Xuanjing, Specia, Lucia, & Yih, Scott Wen-tau (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Olivetti, Elsa A, Cole, Jacqueline M, Kim, Edward, Kononova, Olga, Ceder, Gerbrand, Han, Thomas Yong-Jin, & Hiszpanski, Anna M. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, **7**(4).

OpenAI. 2024 (Oct.). *MMMLU · Datasets at Hugging Face.*

OpenAI. 2025 (Feb.). *Introducing deep research.*

Orr, Will, & Kang, Edward B. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. *Pages 1875–1884 of: The 2024 ACM Conference on Fairness, Accountability, and Transparency.* Rio de Janeiro Brazil: ACM.

Park, Seunghyun, Shin, Seung, Lee, Bado, Lee, Junyeop, Surh, Jaeheung, Seo, Minjoon, & Lee, Hwalsuk. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing.

Parker, Charlie, Scott, Sam, & Geddes, Alistair. 2019. Snowball sampling. *SAGE research methods foundations.*

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, Desmaison, Alban, Köpf, Andreas, Yang, Edward, DeVito, Zach, Raison, Martin, Tejani, Alykhan, Chilamkurthy, Sasank, Steiner, Benoit, Fang, Lu, Bai, Junjie, & Chintala, Soumith. 2019a. *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Red Hook, NY, USA: Curran Associates Inc.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, Desmaison, Alban, Köpf, Andreas, Yang, Edward, DeVito, Zach, Raison, Martin, Tejani, Alykhan, Chilamkurthy, Sasank, Steiner, Benoit, Fang, Lu, Bai, Junjie, & Chintala, Soumith. 2019b (Dec.). *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* arXiv:1912.01703 [cs].

Paullada, Amandalynne, Raji, Inioluwa Deborah, Bender, Emily M., Denton, Emily, & Hanna, Alex. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, **2**(11), 100336. arXiv:2012.05345 [cs].

Pavlick, Ellie, & Kwiatkowski, Tom. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, **7**(Nov.), 677–694.

Pennington, Jeffrey, Socher, Richard, & Manning, Christopher. 2014. GloVe: Global Vectors for Word Representation. *Pages 1532–1543 of: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: Association for Computational Linguistics.

Periti, Francesco, Cassotti, Pierluigi, Dubossarsky, Haim, & Tahmasebi, Nina. 2024. Analyzing Semantic Change through Lexical Replacements. *Pages 4495–4510 of:* Ku, Lun-Wei, Martins, Andre, & Srikumar, Vivek (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics.

Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, & Zettlemoyer, Luke. 2018 (Mar.). *Deep contextualized word representations.* arXiv:1802.05365 [cs].

Pezeshkpour, Pouya, & Hruschka, Estevam. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *Pages 2006–2017 of:* Duh, Kevin, Gomez, Helena, & Bethard, Steven (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics.

Pfeiffer, Olivia P., Liu, Haihao, Montanelli, Luca, Latypov, Marat I., Sen, Fatih G., Hegadekatte, Vishwanath, Olivetti, Elsa A., & Homer, Eric R. 2022. Aluminum alloy compositions and properties extracted from a corpus of scientific manuscripts and US patents. *Scientific Data*, **9**(1), 128. Publisher: Nature Publishing Group.

Pierce, J. R. 1969. Whither Speech Recognition? *The Journal of the Acoustical Society of America*, **46**(4B), 1049–1051.

Pierce, J.R. 1966 (July). *Language and machines.* Tech. rept.

Plank, Barbara. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *Pages 10671–10682 of:* Goldberg, Yoav, Kozareva, Zornitsa, & Zhang, Yue (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Poliak, Adam, Naradowsky, Jason, Haldar, Aparajita, Rudinger, Rachel, & Van Durme, Benjamin. 2018. Hypothesis Only Baselines in Natural Language Inference. *Pages 180–191 of:* Nissim, Malvina, Berant, Jonathan, & Lenci, Alessandro (eds), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics.

Prabhu, Vinay Uday, & Birhane, Abeba. 2020 (July). *Large image datasets: A pyrrhic win for computer vision?* arXiv:2006.16923 [cs].

Pramanick, Aniket, Hou, Yufang, & Gurevych, Iryna. 2023. *A Diachronic Analysis of the NLP Research Paradigm Shift: When, How, and Why?*

Preoțiuc-Pietro, Daniel, Goanta, Catalina, Chalkidis, Ilias, Barrett, Leslie, Spanakis, Gerasimos, & Aletras, Nikolaos (eds). 2023. *Proceedings of the Natural Legal Language Processing Workshop 2023.* Singapore: Association for Computational Linguistics.

Qu, Renyi, Tu, Ruixuan, & Bao, Forrest. 2024. *Is Semantic Chunking Worth the Computational Cost?*

Rachatasumrit, Napol, Bragg, Jonathan, Zhang, Amy X., & Weld, Daniel S. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. *Pages 707–719 of: Proceedings of the 27th International Conference on Intelligent User Interfaces.* IUI '22. New York, NY, USA: Association for Computing Machinery.

Radford, Alec, & Narasimhan, Karthik. 2018. Improving Language Understanding by Generative Pre-Training.

Raji, Inioluwa Deborah, Bender, Emily M., Paullada, Amandalynne, Denton, Emily, & Hanna, Alex. 2021 (Nov.). *AI and the Everything in the Whole Wide World Benchmark.* arXiv:2111.15366 [cs].

Ravaglia, Ray. 2024. DoNotPay: AI Agents Improve Student Financial Aid Applications. *forbes.com.* Accessed: 2024-12-15.

Rein, David, Hou, Betty Li, Stickland, Asa Cooper, Petty, Jackson, Pang, Richard Yuanzhe, Dirani, Julien, Michael, Julian, & Bowman, Samuel R. 2023 (Nov.). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark.* arXiv:2311.12022 [cs].

Rogers, Anna. 2021. Changing the World by Changing the Data. *Pages 2182–2194 of: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Online: Association for Computational Linguistics.

Rogers, Anna. 2023. *Closed AI Models Make Bad Baselines.*

Ruder, Sebastian. 2018 (July). *NLP's ImageNet moment has arrived.*

Rungta, Mukund, Singh, Janvijay, Mohammad, Saif M., & Yang, Diyi. 2022. Geographic Citation Gaps in NLP Research. *Pages 1371–1383 of: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Sambasivan, Nithya, & Veeraraghavan, Rajesh. 2022. The Deskilling of Domain Expertise in AI Development. *Pages 1–14 of: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* CHI '22. New York, NY, USA: Association for Computing Machinery.

Sambasivan, Nithya, Kapania, Shivani, Highfill, Hannah, Akrong, Diana, Paritosh, Praveen, & Aroyo, Lora M. 2021a. “Everyone Wants to Do the Model Work, Not the Data Work”: Data Cascades in High-Stakes AI. *In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. New York, NY, USA: Association for Computing Machinery.

Sambasivan, Nithya, Kapania, Shivani, Highfill, Hannah, Akrong, Diana, Paritosh, Praveen, & Aroyo, Lora M. 2021b. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May, 1–15. Conference Name: CHI ’21: CHI Conference on Human Factors in Computing Systems ISBN: 9781450380966 Place: Yokohama Japan Publisher: ACM.

Sarmah, Bhaskarjit, Dutta, Kriti, Grigoryan, Anna, Tiwari, Sachin, Pasquali, Stefano, & Mehta, Dhagash. 2024 (Dec.). *A Comparative Study of DSPy Teleprompter Algorithms for Aligning Large Language Models Evaluation Metrics to Human Evaluation*. arXiv:2412.15298 [cs] version: 1.

Saxon, Michael, Holtzman, Ari, West, Peter, Wang, William Yang, & Saphra, Naomi. 2024. Benchmarks as Microscopes: A Call for Model Metrology. Publisher: arXiv Version Number: 2.

Schlangen, David. 2021. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. *Pages 670–674 of:* Zong, Chengqing, Xia, Fei, Li, Wenjie, & Navigli, Roberto (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics.

Shah, Chirag, & Bender, Emily M. 2024. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? *ACM Transactions on the Web*, **18**(3), 33:1–33:24.

Shankar, Vaishaal, Roelofs, Rebecca, Mania, Horia, Fang, Alex, Recht, Benjamin, & Schmidt, Ludwig. 2020. Evaluating Machine Accuracy on ImageNet. *Pages 8634–8644 of: Proceedings of the 37th International Conference on Machine Learning*. PMLR. ISSN: 2640-3498.

Shao, Rulin, Asai, Akari, Shen, Shannon Zejiang, Ivison, Hamish, Kishore, Varsha, Zhuo, Jingming, Zhao, Xinran, Park, Molly, Finlayson, Samuel G., Sontag, David, Murray, Tyler, Min, Sewon, Dasigi, Pradeep, Soldaini, Luca, Brahman, Faeze, Yih, Wen-tau, Wu, Tongshuang, Zettlemoyer, Luke, Kim, Yoon, Hajishirzi, Hannaneh, & Koh, Pang Wei. 2025 (Nov.). *DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research*. arXiv:2511.19399 [cs].

Shen, Zejiang, Lo, Kyle, Wang, Lucy Lu, Kuehl, Bailey, Weld, Daniel S., & Downey, Doug. 2022. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. *Transactions of the Association for Computational Linguistics*, **10**, 376–392. Place: Cambridge, MA Publisher: MIT Press.

Si, Chenglei, Yang, Diyi, & Hashimoto, Tatsunori. 2024 (Sept.). *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers.* arXiv:2409.04109 [cs].

Singh, Amanpreet, Natarajan, Vivek, Shah, Meet, Jiang, Yu, Chen, Xinlei, Batra, Dhruv, Parikh, Devi, & Rohrbach, Marcus. 2019 (June). Towards VQA Models That Can Read. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Singh, Janvijay, Rungta, Mukund, Yang, Diyi, & Mohammad, Saif M. 2023. *Forgotten Knowledge: Examining the Citational Amnesia in NLP.*

Smock, Brandon, Pesala, Rohith, & Abraham, Robin. 2023. Aligning benchmark datasets for table structure recognition. *Pages 371–386 of: International Conference on Document Analysis and Recognition.* Springer.

Solaiman, Irene. 2023. *The Gradient of Generative AI Release: Methods and Considerations.*

Stanisławek, Tomasz, Graliński, Filip, Wróblewska, Anna, Lipiński, Dawid, Kaliska, Agnieszka, Rosalska, Paulina, Topolski, Bartosz, & Biecek, Przemysław. 2021. Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts. vol. 12821. arXiv:2105.05796 [cs].

Strauss, Anselm, & Corbin, Juliet. 1990. *Basics of qualitative research.* Sage publications.

Su, Norman Makoto, & Crandall, David J. 2021 (June). The Affective Growth of Computer Vision. *Pages 9291–9300 of: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Sutton, Richard S, & Barto, Andrew G. 2018. *Reinforcement learning: An introduction.* MIT press.

Swain, Matthew C, & Cole, Jacqueline M. 2016. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, **56**(10), 1894–1904.

Tambon, Florian, Nikanjam, Amin, An, Le, Khomh, Foutse, & Antoniol, Giuliano. 2023. *Silent Bugs in Deep Learning Frameworks: An Empirical Study of Keras and TensorFlow.*

Taylor, Ann, Marcus, Mitchell, & Santorini, Beatrice. 2003. The Penn Treebank: An Overview. *Pages 5–22 of:* Abeillé, Anne (ed), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Springer Netherlands.

Thakur, Aman Singh, Choudhary, Kartik, Ramayapally, Venkat Srinik, Vaidyanathan, Sankaran, & Hupkes, Dieuwke. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. *Pages 404–430 of:* Arviv, Ofir, Clinciu, Miruna, Dhole, Kaustubh, Dror, Rotem, Gehrmann, Sebastian, Habba, Eliya, Itzhak, Itay, Mille, Simon, Perlitz, Yotam, Santus, Enrico, Sedoc, João, Shmueli Scheuer, Michal, Stanovsky, Gabriel, & Tafjord, Oyvind (eds), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*. Vienna, Austria and virtual meeting: Association for Computational Linguistics.

Tjong Kim Sang, Erik F., & De Meulder, Fien. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Pages 142–147 of: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.*

Tsipras, Dimitris, Santurkar, Shibani, Engstrom, Logan, Ilyas, Andrew, & Madry, A. 2020 (May). From ImageNet to Image Classification: Contextualizing Progress on Benchmarks.

University, Princeton. *WordNet.*

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Ł ukasz, & Polosukhin, Illia. 2017. Attention is All you Need. *In:* Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (eds), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.

Viswanathan, Vijay, Neubig, Graham, & Liu, Pengfei. 2021. CitationIE: Leveraging the Citation Graph for Scientific Information Extraction. *Pages 719–731 of:* Zong, Chengqing, Xia, Fei, Li, Wenjie, & Navigli, Roberto (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Wade, Alex D. 2022. The Semantic Scholar Academic Graph (S2AG). *Companion Proceedings of the Web Conference 2022.*

Wagstaff, Kiri. 2012 (June). *Machine Learning that Matters.* arXiv:1206.4656 [cs].

Wallach, Hanna, Desai, Meera, Cooper, A. Feder, Wang, Angelina, Atalla, Chad, Barocas, Solon, Blodgett, Su Lin, Chouldechova, Alexandra, Corvi, Emily, Dow, P. Alex, Garcia-Gathright, Jean, Olteanu, Alexandra, Pangakis, Nicholas, Reed, Stefanie, Sheng, Emily, Vann, Dan, Vaughan, Jennifer Wortman, Vogel, Matthew, Washington, Hannah, & Jacobs, Abigail Z. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. Publisher: arXiv Version Number: 2.

Wang, Alex, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, & Bowman, Samuel. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Pages 353–355 of: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics.

Wang, Alex, Pruksachatkun, Yada, Nangia, Nikita, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, & Bowman, Samuel. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *In: Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Wang, Peng, Bai, Shuai, Tan, Sinan, Wang, Shijie, Fan, Zhihao, Bai, Jinze, Chen, Keqin, Liu, Xuejing, Wang, Jialin, Ge, Wenbin, Fan, Yang, Dang, Kai, Du, Mengfei, Ren, Xuancheng, Men, Rui, Liu, Dayiheng, Zhou, Chang, Zhou, Jingren, & Lin, Junyang. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Warner, Benjamin, Chaffin, Antoine, Clavié, Benjamin, Weller, Orion, Hallström, Oskar, Taghadouini, Said, Gallagher, Alexis, Biswas, Raja, Ladhak, Faisal, Aarsen, Tom, Cooper, Nathan, Adams, Griffin, Howard, Jeremy, & Poli, Iacopo. 2024 (Dec.). *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference*. arXiv:2412.13663 [cs].

Way, Samuel, Larremore, Daniel, & Clauset, Aaron. 2016. Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks. 02.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, & Houston, Ann. 2013 (Oct.). *OntoNotes Release 5.0*. Artwork Size: 2806280 KB Pages: 2806280 KB.

Weiss, Robert S. 1995. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster.

Widder, David Gray, West, Sarah, & Whittaker, Meredith. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. *Concentrated Power, and the Political Economy of Open AI (August 17, 2023)*.

Widder, David Gray, Gururaja, Sireesh, & Suchman, Lucy. 2024a (Nov.). *Basic Research, Lethal Effects: Military AI Research Funding as Enlistment.* arXiv:2411.17840 [cs].

Widder, David Gray, Whittaker, Meredith, & West, Sarah Myers. 2024b. Why 'open' AI systems are actually closed, and why this matters. *Nature*, **635**(8040), 827–833. Publisher: Nature Publishing Group.

Wiggers, Kyle. 2024. OpenAI-backed legal tech startup Harvey raises $100M. *Techcrunch.com.* Accessed: 2024-12-15.

Wilkins, Joe. 2025 (Sept.). *Exactly Six Months Ago, the CEO of Anthropic Said That in Six Months AI Would Be Writing 90 Percent of Code.*

Williams, Adina, Nangia, Nikita, & Bowman, Samuel R. 2018 (Feb.). *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.* arXiv:1704.05426 [cs].

Woodruff, Allison, Shelby, Renee, Kelley, Patrick Gage, Rousso-Schindler, Steven, Smith-Loud, Jamila, & Wilcox, Lauren. 2024. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. *Pages 1–26 of: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* CHI '24. New York, NY, USA: Association for Computing Machinery.

Yang, Zhilin, Qi, Peng, Zhang, Saizheng, Bengio, Yoshua, Cohen, William W, Salakhutdinov, Ruslan, & Manning, Christopher D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600.*

Yao, Shunyu, Shinn, Noah, Razavi, Pedram, & Narasimhan, Karthik. 2024 (June). *$$-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.* arXiv:2406.12045 [cs].

Yin, Wenpeng, Hay, Jamaal, & Roth, Dan. 2019 (Aug.). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach.* arXiv:1909.00161 [cs].

Yue, Xiang, Ni, Yuansheng, Zhang, Kai, Zheng, Tianyu, Liu, Ruoqi, Zhang, Ge, Stevens, Samuel, Jiang, Dongfu, Ren, Weiming, Sun, Yuxuan, Wei, Cong, Yu, Botao, Yuan, Ruibin, Sun, Renliang, Yin, Ming, Zheng, Boyuan, Yang, Zhenzhu, Liu, Yibo, Huang, Wenhao, Sun, Huan, Su, Yu, & Chen, Wenhu. 2024 (June). *MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI.* arXiv:2311.16502 [cs].

Yue, Xiang, Zheng, Tianyu, Ni, Yuansheng, Wang, Yubo, Zhang, Kai, Tong, Shengbang, Sun, Yuxuan, Yu, Botao, Zhang, Ge, Sun, Huan, Su, Yu, Chen, Wenhu, & Neubig, Graham. 2025 (May). *MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark.* arXiv:2409.02813 [cs].

Yun, Hye Sun, Zhang, Karen Y. C., Kouzy, Ramez, Marshall, Iain J., Li, Junyi Jessy, & Wallace, Byron C. 2025 (Feb.). *Caught in the Web of Words: Do LLMs Fall for Spin in Medical Literature?* arXiv:2502.07963 [cs].

Zaheer, Manzil, Marino, Kenneth, Grathwohl, Will, Schultz, John, Shang, Wendy, Babayan, Sheila, Ahuja, Arun, Dasgupta, Ishita, Kaeser-Chen, Christine, & Fergus, Rob. 2022. Learning to navigate wikipedia by taking random walks. *Advances in Neural Information Processing Systems*, **35**, 1529–1541.

Zaratiana, Urchade, Pasternak, Gil, Boyd, Oliver, Hurn-Maloney, George, & Lewis, Ash. 2025 (July). *GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface.* arXiv:2507.18546 [cs] version: 1.

Zhang, Hengrui, Georgescu, Alexandru B., Yerramilli, Suraj, Karpovich, Christopher, Apley, Daniel W., Olivetti, Elsa A., Rondinelli, James M., & Chen, Wei. 2024 (Dec.). *Emerging Microelectronic Materials by Design: Navigating Combinatorial Design Space with Scarce and Dispersed Data.* arXiv:2412.17283 [cond-mat].

Zheng, Zhiling, Zhang, Oufan, Nguyen, Ha L., Rampal, Nakul, Alawadhi, Ali H., Rong, Zichao, Head-Gordon, Teresa, Borgs, Christian, Chayes, Jennifer T., & Yaghi, Omar M. 2023. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Central Science*, **9**(11), 2161–2170.

Zhu, Yuxuan, Jin, Tengjun, Pruksachatkun, Yada, Zhang, Andy, Liu, Shu, Cui, Sasha, Kapoor, Sayash, Longpre, Shayne, Meng, Kevin, Weiss, Rebecca, Barez, Fazl, Gupta, Rahul, Dhamala, Jwala, Merizian, Jacob, Giulianelli, Mario, Coppock, Harry, Ududec, Cozmin, Sekhon, Jasjeet, Steinhardt, Jacob, Kellerman, Antony, Schwettmann, Sarah, Zaharia, Matei, Stoica, Ion, Liang, Percy, & Kang, Daniel. 2025 (July). *Establishing Best Practices for Building Rigorous Agentic Benchmarks.* arXiv:2507.02825 [cs] version: 3.