

The Implications of Open Generative Models in Human-Centered Data Science Work: A Case Study with Fact-Checking Organizations

Robert Wolfe, Tanushree Mitra

University of Washington
rwolfe3@uw.edu, tmitra@uw.edu

Abstract

Calls to use open generative language models in academic research have highlighted the need for reproducibility and transparency in scientific research. However, the impact of generative AI extends well beyond academia, as corporations and public interest organizations have begun integrating these models into their data science pipelines. We expand this lens to include the impact of open models on *organizations*, focusing specifically on fact-checking organizations, which use AI to observe and analyze large volumes of circulating misinformation, yet must also ensure the reproducibility and impartiality of their work. We wanted to understand where fact-checking organizations use open models in their data science pipelines; what motivates their use of open models or proprietary models; and how their use of open or proprietary models can inform research on the societal impact of generative AI. To answer these questions, we conducted an interview study with $N=24$ professionals at 20 fact-checking organizations on six continents. Based on these interviews, we offer a five-component conceptual model of where fact-checking organizations employ generative AI to support or automate parts of their data science pipeline, including Data Ingestion, Data Analysis, Data Retrieval, Data Delivery, and Data Sharing. We then provide taxonomies of fact-checking organizations' motivations for using open models and the limitations that prevent them from further adopting open models, finding that they prefer open models for Organizational Autonomy, Data Privacy and Ownership, Application Specificity, and Capability Transparency. However, they nonetheless use proprietary models due to perceived advantages in Performance, Usability, and Safety, as well as Opportunity Costs related to participation in emerging generative AI ecosystems. Finally, we propose a research agenda to address limitations of both open and proprietary models. Our research provides novel perspective on open models in data-driven organizations.

Introduction

Generative AI models have rapidly become a component of organizational infrastructure, with more than 90% of Fortune 500 companies now using ChatGPT (Porter 2023). Such models promise to transform information work by providing approachable conversational interfaces for performing complex tasks involving large quantities of text and data (Hassani

and Silva 2023; Eloundou et al. 2023). Recent research indicates that organizational integration of generative AI can complement the skills of educated professionals, especially early in their careers, increasing productivity and job satisfaction by automating repetitive tasks and making know-how of experienced workers more available to entry-level staff (Noy and Zhang 2023; Brynjolfsson, Li, and Raymond 2023).

Despite this potential for positive impact, however, many scholars have voiced concerns over the growing reliance on closed, proprietary models (Birhane et al. 2022). Concerns have arisen from scholars in both Natural Language Processing (NLP) and the social sciences (Rogers et al. 2023; Ollion et al. 2023), responding to a growing body of research contending that ChatGPT and similar proprietary models can be used as a substitute for human subjects, both for labeling data in scientific studies (Gilardi, Alizadeh, and Kubli 2023; Amin, Cambria, and Schuller 2023), and simulating the behavior of human subjects (Park et al. 2023; Shanahan, McDonnell, and Reynolds 2023), ignoring the paucity of technical information available about proprietary models and the uncertain reproducibility of results obtained. Palmer, Smith, and Spirling (2023) contend that academic researchers should prioritize the use of *open models* for which the weights are available for download, and training data is specified to the end user, unless they can provide an explicit, study-specific justification for choosing a proprietary model (*e.g.*, studying the impact of OpenAI's DALL-E models on artists due to their widespread adoption (Jiang et al. 2023b)).

While these studies address the importance of open models for scientific integrity, they do not consider the impact that choosing open models can have on *organizations* that adopt generative language models as technological infrastructure. In the present work, we seek to better understand the societal implications of open models by studying their use at fact-checking organizations, a group that shares several characteristics that render them worthy of consideration in this context. First, fact-checking organizations routinely employ state-of-the-art language models in their work, lest they find themselves overwhelmed by large volumes of misinformation (Das et al. 2023; Guo, Schlichtkrull, and Vlachos 2022). Second, they must ensure the reproducibility, reliability, and impartiality of their work, or they will compromise both trust with their audiences and their membership in organizations such as the International Fact-Checking Network (IFCN)

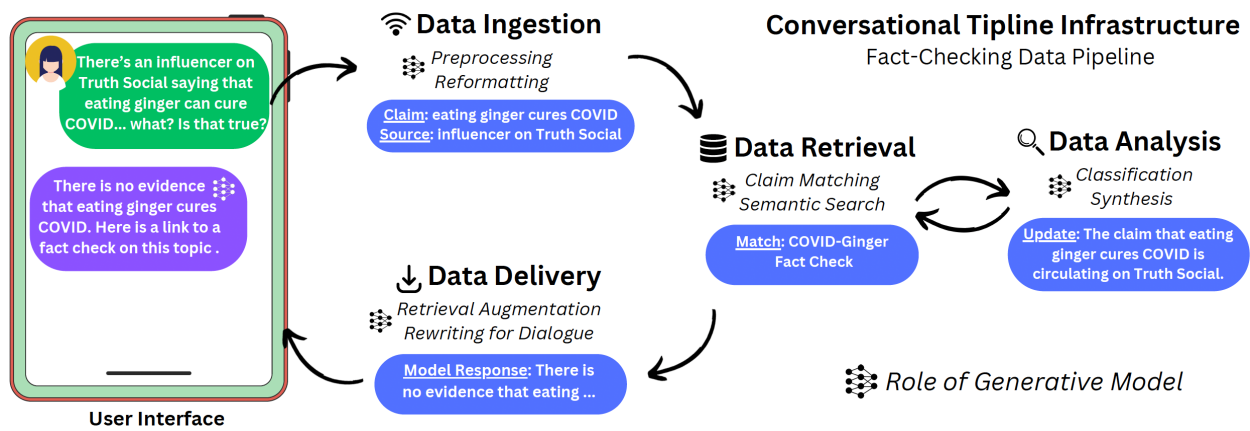


Figure 1: Conversational tiplines are novel data science pipelines for fact-checking accelerated by the advent of chat-based language models. Four components leverage generative AI: Data Ingestion, Data Retrieval, Data Analysis, and Data Delivery.

(Poynter 2024a; Walter et al. 2020). And third, they play a vital role in maintaining the health of information ecosystems around the world (Li and Chang 2023). Understanding use of open models at fact-checking organizations can thus provide insight into the experiences of public interest organizations leveraging generative AI in an impactful sociotechnical context. In this work, we address three research questions:

- **RQ1:** Where do fact-checking organizations employ generative AI models in their data science pipelines?
- **RQ2:** What motivates the adoption of open models by fact-checking organizations?
- **RQ3:** What prevents fact-checking organizations from further employing open models in their work?

To answer these questions, we conducted an interview study with $N=24$ professionals working at 20 fact-checking organizations across six continents. Adopting a human-centered approach to contextualize fact-checker perspectives on generative AI within the context of its use by practitioners, we found that fact-checking organizations reported employing generative models for Data Ingestion, Data Analysis, Data Retrieval, Data Delivery, and Data Sharing. Most participants preferred open models over proprietary models due to concerns related to organizational autonomy, data privacy and ownership, application and domain specificity, and model capability transparency. However, with a few exceptions, their use of open generative models was largely aspirational, as participants cited significant perceived shortcomings in the performance, usability, and safety of open models, as well as opportunity costs associated with not participating in emerging generative AI ecosystems offered by companies like OpenAI and Google. We make three contributions:

- **We offer a five-component conceptual model to describe where fact-checker organizations employ generative models in sociotechnical fact-checking data science pipelines.** We offer two concrete examples of in-use pipelines that employ generative models: media monitoring pipelines, and retrieval-augmented conversational tiplines (illustrated in Figure 1, the latter of which

has seen significant improvements since the advent of general-purpose conversational models such as ChatGPT.

- **We offer taxonomies of 1) motivations of fact-checking organizations for preferring open models, and 2) limitations that prevent further adoption of open models.** We contextualize motivations and limitations by identifying the components of the data science pipeline wherein participants most located their impact, providing a grounded view of the relationship between the model itself and its organizational and societal impacts.
- **We propose a research agenda for addressing the concerns of fact-checking organizations with both open and proprietary generative models.** We offer concrete suggestions for research addressing the performance, usability, and safety of open models, which we suggest can help further their adoption. Given that general-purpose performance of proprietary models will mostly exceed open models, and the revenue of fact-checking organizations may be dependent on producing custom models that integrate with proprietary ecosystems, we also offer directions for research addressing transparency, agency, privacy, and specificity in proprietary models.

Rather than offering a prescriptive approach to open models or an empirical study of their effectiveness, we contribute an understanding of open models in a consequential setting, including how open and proprietary models are valued in practice. We believe these insights can inform the perspectives and research agenda of the AI ethics community.

Related Work

We review related work on open models, fact-checking organizations, and human-centered approaches to data science.

Open and Proprietary Models

While scholars have observed a spectrum of relative openness in AI releases (Solaiman 2023; Langenkamp and Yue 2022), we adopt the definition of open models proposed by Palmer, Smith, and Spirling (2023) and Rogers et al. (2023).

Specifically, open models can be: 1) downloaded locally; 2) run without a call to an API; 3) shared with others; and 4) versioned. The contents of open models' training data must be disclosed, even if the data itself is not available (Palmer, Smith, and Spirling 2023). This definition is satisfied by recent releases of generative language models such as Meta's LLaMA-2 (Touvron et al. 2023) and its finetuned variants such as Stable Vicuna (Chiang et al. 2023). Some generative models, such as the popular Mistral-7B, are better characterized as "open weight" models (Jiang et al. 2023a), as they make the model's weights available, but do not disclose training data to preserve competitive advantage. We note while many transformer-based models, such as Google's BERT (Devlin et al. 2019), would qualify as open models under these definitions, we study specifically *generative* models.

The dominance of OpenAI's ChatGPT (OpenAI 2022) has engendered studies comparing its performance against that of finetuned open models. Kocoń et al. (2023) find that ChatGPT's few-shot performance lags that of task-specific models such as fine-tuned transformers. Moreover, Wolfe et al. (2024) find that small, open generative models can match or exceed ChatGPT after fine-tuning, while Thalken et al. (2023) show that fine-tuned Legal-BERT outperforms generalist models on classifying legal reasoning. While open generative chatbot models without fine-tuning have not matched the performance of GPT-4, which regularly tops evaluation leaderboards such as Stanford's Holistic Evaluation of Language Models (HELM) (Liang et al. 2023), recent open-weight models such as 01.ai's Yi (AI et al. 2024) and Mistral AI's Mixtral 8x7B (Jiang et al. 2024) have surpassed proprietary competitors including Anthropic's Claude 2 (Anthropic 2024) and OpenAI's GPT-3.5-Turbo (OpenAI 2024b), indicating that the best open chat models are comparable to proprietary models six months earlier. Our interviews show that perceived performance shortcomings of open models play a role in fact-checking organizations' use of proprietary models.

Most recently, proprietary model providers have allowed users to create customized versions of generative models, incorporating specialized prompts to govern behavior, and equipping models with external sources of data and programmatic actions. For example, OpenAI refers to customized versions of ChatGPT as "GPTs" (OpenAI 2023), and allows ChatGPT Premium subscribers to build and share GPTs via OpenAI's "GPT Store" web platform, through which OpenAI has promised that builders of custom GPTs will be able to generate revenue (OpenAI 2024a). The GPT Store is an example of an emerging AI *ecosystem*, wherein users select from a range of generative models based on their needs. Our work offers insight into how fact-checkers foresee such AI ecosystems impacting their financial models.

Sociotechnical Infrastructure of Fact-Checking

Fact-checking refers to assessing of the veracity of ostensibly factual information circulating in an information ecosystem that has the potential to harmfully mislead its audience (Graves 2017). While a version of fact-checking has long existed in the form of investigative journalism (Dickey 2019), modern fact-checking coincides with the rise of the internet and social media in particular (Graves and Amazeen 2019),

which provided new conduits for the spread of misinforming content among vast networks of people. Fact-checking is typically described as a sociotechnical task (Graves 2018), meaning in this case that it is successfully accomplished by human investigators working with technological tools that support their expertise (Chopra and Singh 2018; Radiya-Dixit and Neff 2023). Fact-checking organizations mostly embrace language models as essential infrastructure, expressing optimism about their potential to reduce manual workloads and exposure to harmful content, though they remain skeptical of AI that promises full automation of work that relies on human expertise (Juneja and Mitra 2022). Many modern fact-checkers adhere to the principles set forth by the International Fact Checking Network (IFCN) (Poynter 2024a), an organization created to establish common standards for fact-checking (Poynter 2024b). We interview professionals working primarily at signatory organizations of the IFCN to understand how they perceive the merits of open and proprietary models.

Human-Centered Data Science

As noted by Berman (2023), interaction between data practitioners and the tools they use constitutes a social context that shapes the ethics of AI practices in organizations. Much work in social computing seeks to describe these tools and in interaction with human practitioners in data science *pipelines* (Wang et al. 2019). For example, in a study of 183 data scientists, Zhang, Muller, and Wang (2020) describe a common data science pipeline consisting of three high-level steps, including preparation, modeling, and deployment. However, as described by Hopkins and Booth (2021), studies of data practitioners typically center on participants from big tech and academia, and may overlook the challenges faced by smaller organizations facing resource constraints, such as tensions between user privacy and organizational growth, a finding also echoed in Bessen, Impink, and Seamans (2022), who note that AI startups may face tradeoffs between building more competitive and more ethical products. Human-centered data science centers the context in which data practitioners perform their work, acknowledging that data work may be undertaken by domain experts or other workers not traditionally considered data scientists (Muller et al. 2019), a perspective that can yield domain-specific understandings of data pipelines. For example, Rothschild et al. (2022) note that civic workers at public and non-profit institutions exhibit skill with data contextualization that provides value far in excess of their sometimes less-developed computational abilities.

Adopting a human-centered approach is well-motivated for fact checking, because despite the many benchmarks and techniques to support detection of misinformation (Russo, Tekiroğlu, and Guerini 2023; Raj, Mukherjee, and Zhu 2023; Choi and Ferrara 2023), fact-checking organizations often view academic research as too detached from the real world (Juneja and Mitra 2022; Wolfe and Mitra 2024), and recent work argues that even core NLP research on fact-checking should also study human factors (Das et al. 2023). We privilege the views of fact-checking professionals, surfacing where generative models fit into fact-checking data pipelines, and contextualizing the value fact-checking organizations see in open and proprietary models within those pipelines.

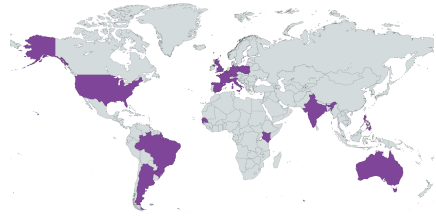
20 Organizations in 15 Countries on 6 Continents

Australian Associated Press (Press 2024), Africa Check (Check 2024), Aos Fatos (Fatos 2024), Chequeado (Chequeado 2024), Code for Africa (for Africa 2024), Der Spiegel (Spiegel 2024), Factly (Factly 2024), India Today (Today 2024), Lead Stories (Stories 2024), logically.ai (logically.ai 2024), Maldita.es (Maldita 2024), Meedan (Meedan 2024), MindaNews (MindaNews 2024), Newtral (Newtral 2024), Pagella Politica (Politica 2024), Pravda (Pravda 2024), Rappler (Rappler 2024), Science Feedback (Feedback 2024), Snopes (Snopes 2024)

Table 1: Participants in 15 countries, 20 organizations.

Approach

We conducted an interview study with $N=24$ professionals at the 20 organizations shown in Table 1 to better understand the use of open models at fact-checking organizations. The study was approved by our university’s IRB.

Participants

We reached out to 92 organizations via cold email, explaining our research and asking for an interview. We employed primarily purposive sampling (Etikan et al. 2016) in emailing member organizations of the International Fact Checking Network (IFCN) (Poynter 2024b) and their partner organizations, and snowball sampling (Naderifar, Goli, and Ghaljaie 2017) when individuals at these organizations offered to connect us with another organization well-positioned for participation. Five participants at five organizations enrolled in our study as a result of snowball sampling; the rest enrolled as a result of purposive sampling. Individuals at ten additional fact-checking organizations responded to our emails but lacked technical knowledge needed to respond to our questions about open and proprietary models, as their roles were related to editing or upper management. We thus excluded them from this study. Most participants were engineers, research scientists, or department managers, with experience ranging from two years to 18 years in their current role. We refer to participants using a randomly assigned number between 1 and N (e.g., P24 said “...”).

Interview Process

We developed a semi-structured interview protocol that asked participants about their organization’s use of generative language models; the opportunities and challenges of generative AI in fact-checking; their organization’s use of open models, and their motivations for adopting them; their reasons for using proprietary models; and what research could support the use of language models in fact-checking work. Where

participants raised topics germane to the research but not covered in our interview protocol, we asked follow-up questions; for example, we asked P4 clarifying questions about their organization’s beta release of a generative chatbot to collect misinformation circulating on platforms such as WhatsApp (WhatsApp 2024). Interviews lasted between twenty-five and ninety minutes and averaged approximately forty-five minutes. All interviews were conducted in English.

Data Analysis

Interviews were recorded over Zoom (Zoom 2024), transcribed using Rev Max AI (Rev 2024), and manually corrected as necessary by the first author. To answer the study’s research questions, we adopted a deductive-inductive approach to coding the interview transcripts. We employed the following deductive codes: Uses of Open Models, Motivations for Using Open Models, Limitations of Open Models, Motivations for Using Proprietary Models, and Implications for Research. The authors first coded four transcripts, and the first author created an initial codebook that included inductively generated themes. The last author reviewed the codebook and the authors jointly revised the codes. The authors then coded four additional transcripts at a time until all transcripts had been coded, reviewing and revising the codebook after each round of coding. Finally, the authors followed a thematic analysis process (Braun and Clarke 2022) to generate themes that answered the study’s research questions, using shared memos to precisely define the themes.

Data Pipelines in Fact-Checking

Consistent with prior work in human-centered data science, we found during thematic analysis that understanding the motivations of fact-checking organizations for using open or proprietary models requires understanding the ways in which they collect, analyze, and exchange fact-checking data - and specifically where they use generative models in these processes. To that end, we begin by providing a conceptual model of five components of fact-checking data science pipelines in which participants described using generative models. We assign each component an icon subsequently used in the Motivations and Limitations sections to associate participant perspectives with components of the pipeline.

📡 Data Ingestion

Participants reported using generative AI to collect and preprocess data, whether via *media monitoring* efforts employing AI-driven tools designed by social media companies or by the organizations themselves, or via *tipline interfaces* wherein a user can submit misinformation for fact-checking. We refer to this component of the data science pipeline as **📡 Data Ingestion**, and denote it using an RSS icon 📡 to suggest the role of monitoring novel information.

Most participants described media monitoring pipelines like that illustrated in Figure 2 as an essential means of observing circulating misinformation. P3 noted that “social media listening is the main point of entry” to their data pipeline, while P11 said that they focus on monitoring WhatsApp because “the coverage is so massive” in their

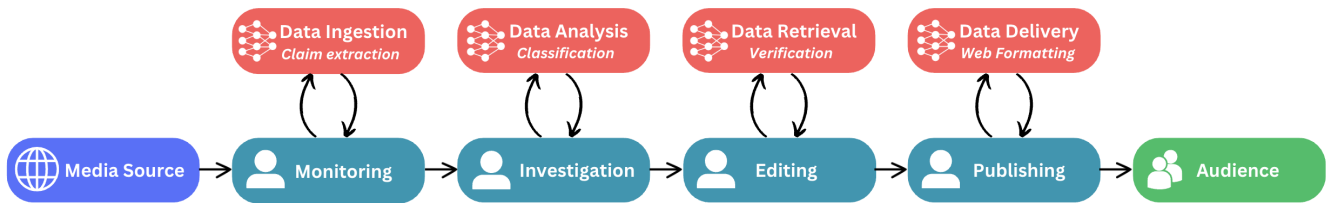


Figure 2: A sociotechnical media monitoring pipeline, with generative AI in red, and human processes in teal.

country. P17 said that automated approaches including generative AI were necessary for media monitoring “given the volume of production and how much content we can reasonably digest.” P2 noted that one of their primary uses of generative AI was automating the data ingestion stages of their data pipeline: “we’re totally focused on this pipeline, and we’re capable of automating the monitoring and the detection phases.” Participants also used generative AI to preprocess data for other pipeline stages. P16 said that they use generative AI to “clean up data” and that “those things [language models] are a time saver” when ingesting data. P19 described using generative AI such that “content can be synthesized and reformatted” for analysis.

Participants also described gathering information using conversational tiplines (*i.e.*, chatbot interfaces) via which audiences can submit potential misinformation. Such interfaces are novel in comparison to media monitoring: P24 described an internal conversational interface that existed as early as 2016, but this interface was never used for data collection. P11 said they first developed a user-facing tipline in 2019, and that they had to significantly scale the tipline during the COVID-19 pandemic, as user interactions increased more than tenfold. Conversational models utilizing modern generative AI first saw release in 2023, as P4 released a beta for a tipline utilizing ChatGPT for its back end, and P11 improved their existing framework with generative models. Generative tiplines serve several purposes, including engaging audiences and bringing in data from sources not easily observed through media monitoring. P24 noted conversational tiplines help to observe “especially provincial media that we are not that aware of or that we’re not looking at regularly,” while P14 noted they use a tipline to bring in new claims for investigation. P20 expressed interest in adapting their existing tipline to utilize generative AI for “constraining the conversation with [the user] to elicit more data that’s actionable.”

🔍 Data Analysis

Participants reported using generative AI to analyze large volumes of potentially misinforming content, leveraging its capabilities to parse highly contextual information, and utilizing few-shot approaches to avoid fine-tuning additional models. We denote 🔍 **Data Analysis** with a magnifying glass 🔍 to suggest the closer study of information.

Participants reported using generative AI to support *classification* and *synthesis* of text and multimodal content. P4 used generative AI to classify text based on constructs like urgency that they previously captured using proxies like message for-

matting: “instead of counting how many exclamation points a post has or how many caps it uses... Gen AI... give[s] us a score from zero to a hundred of the sense of urgency.” P15 noted using generative models as part of an ensemble of deep learning and rule-based NLP classifiers. P2 used generative AI to extract and classify “patterns of manipulative messages... like the government doesn’t want you to know this or share it,” noting that generative AI can be “something like an anti-spam filter” for misinformation. P3 said they use generative AI for classifying multimodal misinformation, including memes: “the text itself isn’t disinformation. The image without the text isn’t disinformation. The image plus the text can feed very clearly into a disinformation narrative.” P2 noted that user-friendly generative AI enables less technical fact checkers to create classifiers: “Generative AI... democratizes who can work with AI... with an API and a little magic with a prompt, you can have something really powerful.”

Participants also embraced uses of generative AI for synthesizing data. P22 said that “one of the values of generative AI is really synthesis, and looking and combing through tons of material, which... [we] will not have time for.” P1 noted using ChatGPT to synthesize hundreds of documents collected every day via media monitoring, and to structure the data “in a tabular format... in 90% of the cases, it gives me a nice table.” P3 used generative AI to synthesize component claims into narratives, improving the scalability of fact-checking work: “large language models are super good at basically clustering claims into a narrative. Fact checking just individual claims is whack-a-mole, a losing proposition... you’ll never be able to scale it.” P11 demoed a GPT-4 driven narrative system for us, explaining that it synthesizes new claims into overarching narratives already associated with fact checks, pending human review: “this summary is linked to these four different contexts that we have already received... generative AI here has proposed to us something that is a bit overarching... [it] is seeing what we produce, and the evidence that accompanies [our] debunks, and is proposing to us [a fact check] that has already been verified by a human.”

🗄️ Data Retrieval

Participants reported using generative AI to facilitate 🗄️ **Data Retrieval** from catalogues of past fact checks or other verified sources of information maintained by the organization. We denote Data Retrieval with a database 🗄️ to suggest the retrieval of stored data.

Many participants described using *Retrieval-Augmented Generation* (RAG) (Lewis et al. 2020) to allow GPT models

to incorporate factual data. P4 described “a RAG pipeline that connects OpenAI’s GPT-4 with our database of fact checks... we have all the articles and fact checks that we ever published stored as embeddings. And then... we perform semantic search using cosine similarity, and we take the most relevant results.” P19 described retrieving data from a catalogue of past fact checks to support fact-checker investigations. P13 noted using third-party reliability metrics to determine what external content can be accessed using RAG. P11 noted that RAG allowed them “to go beyond very basic keyword searches... so that the search in the database was actually fruitful and accurate.” Some participants adopted more complex methods. P22 described creating an internal knowledge graph from which generative models could retrieve content: “We started building our own knowledge graph, our own ontology, and using that, the structured data from that, to generate content.” P23 created custom GPTs to retrieve data: “we built it on the claim review database... only our factcheck articles... so interacting with that search persona would give results only from the database along with a source link.” P23 also produced custom models to retrieve external data: “Parliament data is completely public, so we had the tech team... scrape the entire database... and the persona only picks up responses from this dataset.”

📄 Data Delivery

Participants reported using generative AI to support 📄 **Data Delivery** to end users on websites or social media, as well as by providing automatic responses to users via conversational tiplines. We denote Data Delivery with a download icon 📄 to suggest the transfer of data to the end user.

Among the most common uses of generative AI reported was to format content or generate metadata prior to sharing it with audiences. P1 described using GPT models to “generate hashtags for... mini FactCheck videos that we publish on TikTok.” P8 said that “in most daily use cases, in terms of generative AI use, I would say it’s help with promotion. So all of the social media content, coming up with summaries for SEO purposes for article publishing, title generation.” P5 noted that they use a generative AI-backed tool that “suggests times the best times for us to post our content on social media based on the type of demographic our subscribers are, or our audiences are, and when they’re using social media.”

Participants who used conversational tiplines reported using generative AI to *deliver* information to end users, in addition to *ingesting* misinforming content. P4 said that incorporating ChatGPT into their tipline was “an obvious use case to improve a product that was already relevant for our readers,” noting their conversational tipline had “over 70,000 users, which for an organization our size is quite a lot.” Several participants described an evolving information ecosystem wherein users sought information from specialized conversational agents, rather than traditional search applications. P21 said that they see generative AI as “a preferred medium for somebody to get at the work that we have done... it is summarizing or reporting on work that was done by the trusted fact checkers.” P24 said that they use a WhatsApp chatbot that “allows us to answer to a high volume of messages,” noting that “if you could actually ask

the [chatbot], can you please tell me what the inflation was in the last five years? And it could actually answer you with information that comes from a reliable source, which we know is one of the big problems of generative AI, we think that’s an enormous leap forward in the way that we can actually reach people with verified information.” P19 contended that conversational agents can assist users “even if we don’t have a fact check... explain this persuasion technique that’s being used or the trope that is being repeated.” P9 envisioned reaching younger audiences with an in-progress tool allowing a user to “chat with our archive... this is a hurdle for young people to get into the discourse, that some background knowledge is missing. Maybe the AI could help.”

👥 Data Sharing

Participants noted that generative AI assists in 👥 **Data Sharing** between organizations, in that they structure data for sharing, and serve as shared computational infrastructure. We denote Data Sharing with a Users icon 👥 suggesting transfer of data between organizations.

P18 described generative AI as a tool for exchanging fact check data and collectively scaling audiences to address problematic information: “especially during elections, we try at (anonymized) to bring together other fact checkers so that people are not working in their own little silos... that’s one area in which generative AI can really help. If fact-checkers are working together, whatever data they have, they can help scale the impact of their fact check to different segmented audiences that they serve.” P3 described a prototype generative AI system operating on shared data, noting that especially in the case of elections, “fact checkers are banding together to offer a united response... European checkers will put their claims in a common database, and we will build systems in which, when we detect a new narrative... [if] it’s present more than one or two countries, there will be a special task force that will be tasked with producing a debunk... we could not do it with a BERT or a Sentence BERT... early tests are promising with a [generative] large language model. This is also where the multilingual aspect comes in very handy.” P11 described a similar multi-organization “project where we have installed [generative] technology... at fact-checkers in all Russia bordering countries. And we are also looking at... Spanish speaking Latin America. And we can see how common threats appear... where there are common narratives that point to particular actors.”

Motivations for Open Models

We found that four primary concerns motivate the use of open models: Organizational Autonomy, Data Privacy and Ownership, Task Specificity, and Capability Transparency. We describe these concerns in turn, making reference to the components of the data pipeline with which they intersect.

Organizational Autonomy

Participants expressed concern that dependence on proprietary generative models could compromise the autonomy of their organizations. They noted that open models offer more 🔍 **Reliable Availability** than proprietary models, which

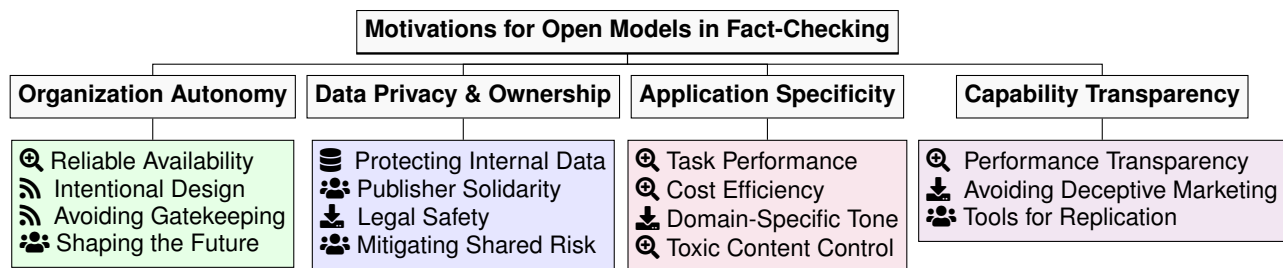


Figure 3: Motivations of participants for preferring open models over proprietary models in fact-checking organizations.

could be affected by unexpected deprecation or provider instability. Several participants described uncertainty about using OpenAI’s models in the wake of its CEO’s firing and reinstatement. P4 said “the whole OpenAI drama was an eye opener. If OpenAI goes bankrupt tomorrow, then it’s really bad to build products that depend on their software and on their API. We’d rather just host all the models that we use.” P2 said that their organization is building next-generation content moderation and misinformation detection tools to circumvent this dependence, while P7 noted that their company builds on open models, mitigating issues of deprecation: “We primarily rely on open source technologies. When Facebook releases their models and they open source it, that’s what we use. We don’t primarily rely on any [closed] corporate models.” Participants also preferred open models because they facilitated **📡 Intentional Design** of tools for fact-checking use cases. P8 said that proprietary tools offered by social media companies were often inadequate because they “are targeting brands. And brands like McDonald’s will have certain sets of keywords that will not change over time. . . we want to have this additional aspect, which is discovering new keywords because we want to stay on top of narratives.” P24 said “for example, social media monitoring or social listening tools. . . we end up developing a lot of things out of need. . . because we don’t have the same needs as a marketing agency.” Similarly, participants noted that **🚫 Avoiding Gatekeeping** by corporations motivates open models. P10 said that “we have to prove ourselves to the social media companies” to gain access to the only tools to monitor and address misinformation on their platforms, noting this process is onerous for resource-challenged local organizations. Finally, participants said open models could help fact-checking organizations in **👥 Shaping the Future** of fact-checking. P9 argued for accelerating adoption of open generative models: “This all brings chances and holds a lot of potential for us, and we should be the one who shape this development, and not let others be the one who dictate how we have to deal with this at some point because we waited too long.” Similarly, P22 remarked that “fact-checkers need to be part of the creation of these tools. . . a lot of the tools that we are seeing don’t really quite fit our use cases. . . it’s about the process by which things are created.”

Data Privacy and Ownership

Participants described concerns surrounding data privacy and ownership as a central motivation for using open models. P9

preferred using open models for **🔒 Protecting of Internal Data**, saying that “we have very sensitive material that we’re working with here, investigative reporting and investigative stories, and we don’t want this to be used in [corporate] models and as training material,” and further noting that they use cloud instances hosted only in Europe for “a sort of legal safety” due to stricter European data protection laws. P4 said that, due to data privacy concerns, “we have as a policy to always prioritize using open source software where we can.” P21 noted “I blocked [OpenAI’s] crawlers from being able to train on our content until some type of commercial compensation becomes available. I think most of us are kind of waiting, holding our breaths for the New York Times case lawsuit to play out because otherwise individual orgs the size of the fact checkers, we don’t really have the leverage to accomplish what that lawsuit stands to do in setting a precedent.” Participants also said that **👥 Publisher Solidarity** motivated use of open models over proprietary models that profit from journalistic organizations’ data without consent. P4 noted their discomfort with “the notion that those companies are profiting using other companies’ and other people’s work.” Participants also preferred open models that disclosed their training datasets, providing a sense of **⚖️ Legal Safety**, especially for user-facing applications. P4 said that “the issue of copyright is a big one, especially for image generation. . . we could never use anything, any tool that generate images in our workflow, because we don’t know how most models were trained.” P8 noted that they delayed using generative models due to fears of copyright infringement: “With generative AI, we were. . . scared to use it, because of the fact that we don’t want to feel like we are plagiarizing. . . because of the possibility of. . . [copying] other articles.” Participants noted that open models at least disclose their training datasets, offering some clarity concerning the risk of infringement. Finally, participants preferred open models for **👥 Mitigating Shared Risk** when building technologies for the fact-checking community. P4 created a transcription tool for the community using OpenAI’s models, but noted is not in production because “a lot of people have concerns about sending their data - interviews, important interviews - to OpenAI.”

Application Specificity

Participants preferred open models for specific applications that demanded high performance, domain-specific tone, and

control over toxic content. P14 said that their organization preferred small, fine-tuned open models for internal data analysis, noting that they exhibit stronger **🔍 Task Performance** and **🔍 Cost Efficiency** than proprietary generative models. P6 echoed this, noting that their organization still prefers thoroughly vetted, task-specific models for many tasks, despite hype about replacing these methods with proprietary models. P10 considered GPT models one of many tools, not the sole solution to any problem involving language, despite the marketing of proprietary models. Participants also noted that proprietary models intended for general-purpose use couldn't achieve **⬇️ Domain-Specific Tone** for end user applications. P4 said "out of the box, GPT-4, for instance, you get very decent results, but it's also very generic." P8 noted that, as a result of RLHF, GPT-4 "sounds completely unnatural," rendering it difficult to incorporate in their user-facing applications. Participants also reported using open models to gain more granular **🔍 Toxic Content Control**. Participants including P17 took issue with corporate models' one-size-fits-all approach to alignment with human values, which they noted could hamper the model's ability to respond to toxic and hateful content that fact-checking organizations handle during their work. P5 noted that the OpenAI's RLHF process makes it difficult for models "to unlearn things that you've already taught it through human feedback," which are not advantageous for many fact-checking applications.

Capability Transparency

While participants agreed that GPT-4 outperformed open models, they also said that the capabilities of open models were presented with more **🔍 Performance Transparency**. P3 expressed surprise that GPT-4 performed poorly for restructuring their data, a task they thought fell within the model's capabilities, noting that while it looked reasonable "on the surface... when you actually dug into whether it was structured coherently, it wasn't as good." Participants also said that **⬇️ Avoiding Deceptive Marketing** motivated use of open models in settings involving user interaction. P17 said the presentation of models like ChatGPT encouraged inappropriate trust by non-experts: "Yes, there is a popup on ChatGPT, and a line at the ending that tells you the results could be inaccurate, so beware. But how the interface is built, and how it is marketed, how it is presented, and the fact that it answers you in a confident way... there is a constant behavioral trick... that what you have in front of you, what a machine is telling you, answering to your prompt, is the truth." P1 echoed this sentiment, noting that "There's a problem with people assigning too much credibility to large language models." Finally, participants said that providing **🛠️ Tools for Replication** motivated open models. P8 noted that "we want to give a reader the ability to replicate our research in total... every source, everything." P9 said that, while all models created problems of explainability, API-gated proprietary models introduced more "black box problems" than open models, rendering reproducibility uncertain for end users.

Limitations of Open Models

Despite the many motivations participants gave for preferring open models, most nonetheless used primarily proprietary

models, especially OpenAI's GPT models (OpenAI 2024b), citing their Performance, Usability, Safety, and the Opportunity Costs of not participating in proprietary AI ecosystems.

Performance

While fine-tuned open models may achieve the strongest performance on a given task, all participants said that GPT-4 was the best-performing *general-purpose* generative model, motivating its adoption over open models in many settings where general-purpose reasoning is preferable to task specialization. P13 noted that they use GPT-4 over LLaMA models due to **🔍 Few-Shot Reasoning** disparities, while P9 echoed this in noting that "if you want to work with a narrative generative AI model, then I think there is, at the moment, no alternative to GPT-4." Participants also noted that open models lag OpenAI significantly in **🔍 Multilingual Performance**. P4, who reported creating primarily non-English models, said that "we tried the largest [LLaMA models]... usually the models tend to perform poorly in languages that are not English. And in the case of OpenAI, it's pretty good." P4 also said that performance disparities affected the **⬇️ End User Readiness** of open models: "when we were developing [our conversational tipline], we tested a bunch of models, especially LLaMA-2, but in terms of performance, it's behind OpenAI significantly... I'd love to use open source models." P2 said that the **🔍 Pace of Innovation** in generative AI made it untenable to try to build open models: "the pace of innovation nowadays is so quick that it's very difficult to keep the pace... you don't know which is the new tech that you need to use or which is the model that is going to work... even research institutions, they don't know." In some cases, participants reported that GPT-3.5-Turbo's inexpensive **📊 Average Performance** was good enough. P1 said that, for retrieving data internally, they needed high recall and not necessarily high accuracy, since human fact-checkers would see the data. They noted "OpenAI is... so easy and cheap. We pay a couple of dollars per month... I can use the old (anonymized) servers to run an open language model and see what it can do. But the savings would be minimal, like a couple of dollars, and the results would probably be worse."

Usability

Most participants said the ease of calling proprietary APIs motivated their use of proprietary generative models. P19 said that, even though they prefer to build on open source models, **🔍 API Simplicity** motivated them to OpenAI: "OpenAI models are convenient for prototyping because it's just an API call, and you don't have to worry too much about it." P2 said, "what is the good part of OpenAI? Everything with an API is much easier." However, P2 also pointed to the possibility that OpenAI's perceived edge in usability is actually due more to its market dominance, noting that "there are a lot of different frameworks now... [such as] LangChain, there are a lot of frameworks that make it easier to work with [open models], and they are the solution." Participants noted the **🔍 Demands of Fine-Tuning** open models drove them to use few-shot proprietary models. Even if task-specific performance exceeding GPT-4 is achievable with fine-tuned open models, data scientists at fact-checking organizations

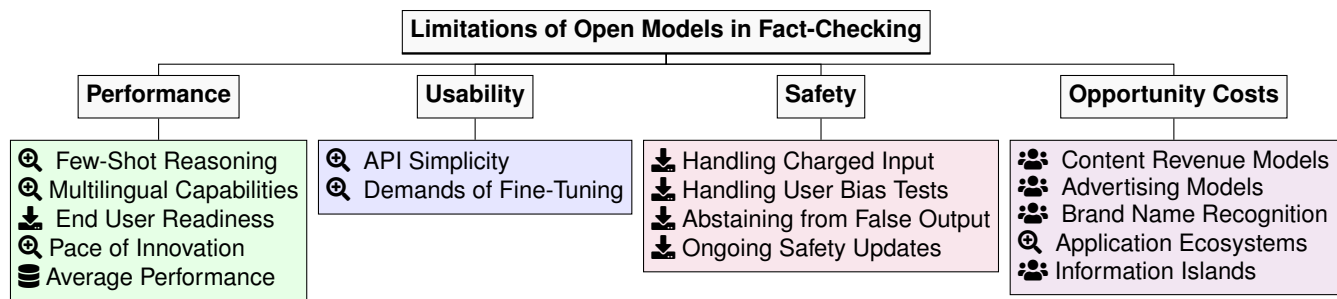


Figure 4: Limitations of open models described by participants as preventing their further adoption in fact-checking.

may not have the time to invest in fine-tuning. P2 said “the other issue for typical fact-checking organizations, is we are not Microsoft, we are not big technology companies, is the cost of the systems . . . to fine tune your own language model it is too. . . time demanding.” P4 said it was hard to find time to understand the opportunities of fine-tuning, noting they only used LLaMA “out of the box, we didn’t try to fine tune it. Maybe in the future we will, but. . . resources are limited, so we opted not to explore more.”

Safety

Participants said that proprietary models offer advantages for user-facing applications because of the safety features built into them by larger technology corporations. P13 noted that because user-facing fact-checking technologies typically involved **⬇ Handling Charged User Input**, reliably adhering to ethical guardrails was essential for maintaining user trust, and open models could not always accomplish this. P4 said that user-facing technologies also had to be prepared for **⬇ Handling Bias Stress Tests**, noting that users actively attempt to probe their in-beta conversational model (leveraging GPT-4) for political biases: “one of the most common types of questions that I think people were asking the bot. . . [was] just swapping the name of the [politician] you’re asking about. . . and so far we haven’t noticed anything that would be concerning for us. . . it’s not symmetric, but I think it usually gives you a quite nuanced answer.” P4 also noted that proprietary models tend to outperform open models in **⬇ Abstaining from False Output**, noting that users most frequently complained that their RAG-enabled GPT-4 chatbot couldn’t answer a question, but that “in most cases, we didn’t have the answer. So the bot behaved as expected as it should, which is to say that it doesn’t know the answer to a question instead of trying to come up an invented answer.” Finally, participants said that proprietary model developers were better positioned to perform **⬇ Ongoing Safety Updates**. P21 said that “fact check organizations are underequipped to really test everything and keep up with everything on an ongoing basis”; accomplishing this with open models, they said, would require “revenue streams. . . that’s still trying to be figured out.”

Opportunity Costs

Participants worried that using open models would entail foregone opportunities for integration into an emerging in-

formation ecosystem, potentially lessening the relevance of their content and precluding them from taking advantage of new streams of revenue. P9 said that **👤 Content Revenue Models** would “totally define how we will work with Gen AI models, and if we will work in close partnerships with these companies, or if we will work with open source models, or if they will be state funded projects. All the journalistic companies say we need it to save democracy, and we push for open source models; or we say, no, we’re fine, we’re getting millions and millions from Google, and OpenAI, and Amazon, and so on. So yeah, we’re sort of at a crossroads.” P21 noted that, in addition to direct compensation for content, proprietary models might be preferred due to **👤 Advertising Models**, noting that “the value exchange that we’re familiar with today with Google search is you structure your metadata the right way, and you’ll show up in Search, and you’ll get clicked on, and you’ll generate ad revenue. And perhaps Google’s even the one who buys those ads, and the revenue comes from them anyway. I think that [Google is] more ready as an organization to think about paying publishers. With ChatGPT that’s not really established yet.” P21 also said that they hoped to leverage the **👤 Brand Name Recognition** of emerging AI ecosystems to increase their reach, noting “ultimately ChatGPT is the biggest brand name in chatbots, and they are also already integrated and backing Bing. . . to reach all of the users who are using chatbots, it’s not realistic to think that we’ll make the biggest splash just by having our own private code base and onsite chat experience. I do think that we have to be able to play into the bigger arena.” P10 further discussed the possibility of integrating their internal models into broader **🔍 Application Ecosystems**, noting that “we use the Google workspace workflow, so it really helps to incorporate the Bard features.” Finally, P21 raised broader concerns about the information ecosystem, describing the possibility of **👤 Information Islands** if internet users privately interact with open chat models: “if people choose to search or engage with a chat bot that. . . isn’t trained on your content, we stand to have these kind of information islands. . . someone who searches on Google today or Bing today, they’re going to get [our content] as results that were fully enabled to be crawled by search engines. But as more and more chat bots emerge. . . that is kind of a threat, I think, to information integrity overall.”

Research Directions for Use of Open Models in Fact Checking		
Concern	Research Question	Research Directions
Performance of Open Models	How can open models offer competitive performance to proprietary models while maintaining an approachable conversational interface?	Developing scale-efficient open models; Developing more suitable evaluation suites for fact-checking; Developing evaluations specifically for open models.
Usability of Open Models	What kinds of open application interfaces can help fact-checkers feel as comfortable with open models as proprietary models for inference and fine-tuning?	Creating open source or public APIs of comparable simplicity to OpenAI; Decreasing the time and expertise demands of fine-tuning.
Safety of Open Models	How can open models achieve the actual and perceived safety of proprietary models without incurring significant time and cost burdens?	Community standards for lightweight red-teaming of open models; Technologies to monitor the fairness of model responses in user-facing conversational tiplines.
Opportunity Costs of Open Models	How can open model ecosystems facilitate reliable revenue streams for fact-checking organizations similar to those in proprietary model ecosystems?	Fostering community and collaboration between fact-checking organizations and AI developers; Developing a revenue model for open model ecosystems.
Research Directions for Use of Proprietary Models in Fact Checking		
Lack of Autonomy	How can proprietary model developers assure clients of access to models integrated in data pipelines?	Developing approaches to allowing selective access to deprecated models.
Lack of Data Privacy and Ownership	How can proprietary model developers guarantee clients that their data won't be used inappropriately? How can they guarantee clients that using their products will not put them in legal jeopardy?	Developing models of compensation for publishers; Supporting legal standards for client data privacy; Clearly communicating when data will be retained or used outside of initial context.
Lack of Application Specificity	How can developers afford clients more control over tone in user-facing applications and more contextually appropriate means of processing toxic content?	Developing approaches for personalized models that meet fact-checking use cases without compromising model safety.
Lack of Transparency	How can developers provide transparency about model capabilities and afford fact-checking organizations the ability to explain their use to audiences?	Communicating capabilities of domain-specific GPTs through systematic quantitative & qualitative evaluation; more explainable outputs like token-level probabilities.

Table 2: Research directions for addressing limitations of both open and proprietary models proprietary models in fact-checking.

Discussion

Our findings add new perspective to the evolving conversation about the use of open or proprietary generative AI, surfacing the tradeoffs that face *organizations* as they consider whether and where to adopt open models. Some tradeoffs would seem contradictory without being contextualized within the fact-checking data science pipeline. For example, participants chafed at limitations imposed by toxicity guardrails when using OpenAI models for internal analysis tasks, yet reported *relying* on those guardrails for user-facing applications. Moreover, while participants preferred fine-tuned open models for strong performance on internal tasks, they also accepted “good enough” performance via the cheap, easy-to-use GPT-3.5-Turbo API for high-recall data retrieval tasks, over the time cost incurred for an open model.

Table 2 outlines research questions and directions arising from this work. Fact-checking organizations appear at first to face diverging futures: one in which they adopt open models and control their AI infrastructure; and another operating profitably in the emerging ecosystems of chat-driven interfaces now collecting and delivering information in interactions with users. However, our findings surface that the future may resemble the present, with organizations employing specialized open models for mission-critical tasks, proprietary generalist models to handle interactions with users, and the least expensive option when performance just needs to be good enough. Even so, proprietary models face challenges around data ownership. Participants voiced unease with their work and that of their colleagues being used without consent

to create a lucrative competing product. Some participants also avoided generative AI for fear of inadvertently committing plagiarism. These concerns echo those of many journalists, as well as domains like visual art (Jiang et al. 2023b). Addressing data privacy, ownership, and compensation may prove paramount for proprietary model providers to establish relationships with producers of factual and creative content.

Limitations and Future Work

While we sought to capture perspectives of fact-checking organizations globally, our participants all spoke English, and most used models in English. Aside from noting that OpenAI’s models outperform open models in multilingual settings for generalist conversational tasks, our work does not speak to the additional intricacies of multilingual NLP, including NLP in low-resource languages. Additionally, our work is concerned with open generative models, rather than with all open models. Future work might study perspectives on open models and proprietary models of any architecture.

Conclusion

We contribute an organizational perspective on open and proprietary models, finding that while fact-checking organizations prefer open models for Autonomy, Data Privacy and Ownership, Application Specificity, and Capability Transparency, they nonetheless use proprietary models for Performance, Usability, and Safety. We suggest open models could benefit from increased usability, while proprietary model providers must address concerns over data ownership.

Researcher Positionality

The authors are researchers at a university in the United States. The second author has long-standing relationships with fact-checking organizations around the world, which they have fostered throughout their career. The first author has had informal experience working with fact-checking organizations in research contexts, and has technical expertise with generative AI and related NLP and computer vision technologies. We sought to include a diversity of perspectives in this research by interviewing individuals working at fact-checking organizations around the world; nonetheless, we note that our positionality as researchers located in the U.S. prevents us from comprehensively describing the contexts in which our diverse pool of participants operate.

Adverse Impacts

We sought in this research to provide a balanced perspective representative of the views of our participants. However, in suggesting a research agenda that accounts for both open and proprietary models, our study may seem to place these technologies on equal footing. The harms arising from proprietary models are likely to be very different than those arising from open models, and potentially greater in scale due to their much larger end user base at the present time. Thus, this research should not be read as concluding that open and proprietary models have *similar* societal implications, but rather that challenges faced by these models are both sufficiently important as to warrant concurrent research agendas. Moreover, we have focused not on the question of *whether* use of generative AI is desirable in fact-checking organizations, and what the drawbacks of that might be, but on the tradeoffs of open and proprietary models. We note that there may be notable harms in use of any form of generative AI in fact-checking, particularly when the technology is positioned as a replacement for human expertise (Juneja and Mitra 2022).

Ethical Considerations

We centered the role played by experts at fact-checking organizations, as well as the data pipelines employed by these organizations. However, we note that there are many human roles that we have not managed to surface in this work, including roles held by individuals who participate in data pipelines in ways that would not traditionally be recognized as data science (Muller et al. 2019). This is a consequence of our research questions, which focus on uses of open and proprietary language models, rather than on the less recognized data labor performed by individuals and organizations that seek to counter misinformation. Future work might more directly address these less obvious *human* contributions to the data pipelines surfaced in the present work

References

AI, .; Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652.

Amin, M. M.; Cambria, E.; and Schuller, B. 2023. Will affective computing emerge from foundation models and general AI. *A first evaluation on ChatGPT*. ArXiv, abs/2303.03186.

Anthropic. 2024. Claude 2. <https://www.anthropic.com/news/claude-2>. [Accessed 09-03-2024].

Berman, G. 2023. Machine Learning practices and infrastructures. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 466–481. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Bessen, J.; Impink, S. M.; and Seamans, R. 2022. The cost of ethical AI development for AI startups. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 92–106.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Braun, V.; and Clarke, V. 2022. Everything changes. . . well some things do: Reflections on, and resources for, reflexive thematic analysis. *QMIP Bulletin*.

Brynjolfsson, E.; Li, D.; and Raymond, L. R. 2023. Generative AI at work. Technical report, National Bureau of Economic Research.

Check, A. 2024. Africa Check. <https://africacheck.org/>. [Accessed 22-01-2024].

Chequeado. 2024. Chequeado. <https://chequeado.com/>. [Accessed 22-01-2024].

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Choi, E. C.; and Ferrara, E. 2023. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. *arXiv preprint arXiv:2310.09223*.

Chopra, A. K.; and Singh, M. P. 2018. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 48–53.

Das, A.; Liu, H.; Kovatchev, V.; and Lease, M. 2023. The state of human-centered NLP technology for fact-checking. *Information processing & management*, 60(2): 103219.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dickey, C. 2019. The Rise and Fall of Facts. https://www.cjr.org/special_report/rise-and-fall-of-fact-checking.php. [Accessed 09-03-2024].

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. Gpts are gpts: An early look at the labor market

- impact potential of large language models. *arXiv preprint arXiv:2303.10130*.
- Etikan, I.; Musa, S. A.; Alkassim, R. S.; et al. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1): 1–4.
- Factly. 2024. Factly. <https://factly.in/>. [Accessed 22-01-2024].
- Fatos, A. 2024. Aos Fatos. <https://www.aosfatos.org/>. [Accessed 22-01-2024].
- Feedback, S. 2024. Science Feedback. <https://science.feedba ck.org/>. [Accessed 22-01-2024].
- for Africa, C. 2024. Code for Africa. <https://github.com/Cod eForAfrica/>. [Accessed 22-01-2024].
- Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.
- Graves, L. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3): 518–537.
- Graves, L. 2018. Factsheet: Understanding the promise and limits of automated fact-checking. *Reuters Inst. Study of Journalism, Univ. Oxford, Oxford*.
- Graves, L.; and Amazeen, M. A. 2019. Fact-checking as idea and practice in journalism. In *Oxford research encyclopedia of communication*.
- Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.
- Hassani, H.; and Silva, E. S. 2023. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing*, 7(2): 62.
- Hopkins, A.; and Booth, S. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21*, 134–145. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023a. Mistral 7B. *arXiv preprint arXiv:2310.06825*, (): .
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. *ArXiv*, abs/2401.04088.
- Jiang, H. H.; Brown, L.; Cheng, J.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebu, T. 2023b. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.
- Juneja, P.; and Mitra, T. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–36.
- Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, (): 101861.
- Langenkamp, M.; and Yue, D. N. 2022. How open source machine learning software shapes ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 385–395.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- Li, J.; and Chang, X. 2023. Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information systems frontiers*, 25(4): 1479–1493.
- Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. *arXiv:2211.09110*.
- logically.ai. 2024. logically.ai. <https://www.logically.ai/>. [Accessed 22-01-2024].
- Maldita. 2024. Maldita. <https://maldita.es/>. [Accessed 22-01-2024].
- Meedan. 2024. Meedan. <https://meedan.com/>. [Accessed 22-01-2024].
- MindaNews. 2024. MindaNews. <https://www.mindanews.com/>. [Accessed 22-01-2024].
- Muller, M.; Feinberg, M.; George, T.; Jackson, S. J.; John, B. E.; Kery, M. B.; and Passi, S. 2019. Human-centered study of data science work practices. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, 1–8.
- Naderifar, M.; Goli, H.; and Ghaljaie, F. 2017. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education*, 14(3).
- Newtral. 2024. Newtral. <https://www.newtral.es/>. [Accessed 22-01-2024].
- Noy, S.; and Zhang, W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654): 187–192.
- Ollion, E.; Shen, R.; Macanovic, A.; and Chatelain, A. 2023. Chatgpt for Text Annotation? Mind the Hype! *SocArXiv*. October, 4.

- OpenAI. 2022. Introducing ChatGPT. <https://openai.com/index/chatgpt/>. [Accessed 19-01-2024].
- OpenAI. 2023. Introducing GPTs. <https://openai.com/index/introducing-gpts/>. [Accessed 19-01-2024].
- OpenAI. 2024a. Introducing the GPT Store. <https://openai.com/index/introducing-the-gpt-store/>. [Accessed 19-01-2024].
- OpenAI. 2024b. Models. <https://platform.openai.com/docs/models/>. [Accessed 19-01-2024].
- Palmer, A.; Smith, N. A.; and Spirling, A. 2023. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, (0): 1–2.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
- Politica, P. 2024. Pagella Politica. <https://pagellapolitica.it/>. [Accessed 22-01-2024].
- Porter, J. 2023. ChatGPT continues to be one of the fastest-growing services ever. *The Verge*, (0): .
- Poynter. 2024a. Commit to transparency — sign up for the International Fact-Checking Network’s code of principles. <https://ifcncodeofprinciples.poynter.org/>. [Accessed 09-03-2024].
- Poynter. 2024b. Verified signatories of the IFCN code of principles. <https://ifcncodeofprinciples.poynter.org/signatories/>. [Accessed 22-01-2024].
- Pravda. 2024. Pravda. <https://pravda.org.pl/>. [Accessed 22-01-2024].
- Press, A. A. 2024. Australian Associated Press. <https://www.aap.com.au/>. [Accessed 22-01-2024].
- Radiya-Dixit, E.; and Neff, G. 2023. A sociotechnical audit: Assessing police use of facial recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346.
- Raj, C.; Mukherjee, A.; and Zhu, Z. 2023. True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 962–963.
- Rappler. 2024. Rappler. <https://www.rappler.com/>. [Accessed 22-01-2024].
- Rev. 2024. Rev. <https://www.rev.com/>. [Accessed 10-03-2024].
- Rogers, A.; Balasubramanian, N.; Derczynski, L.; Dodge, J.; Koller, A.; Luccioni, S.; Sap, M.; Schwartz, R.; Smith, N. A.; and Strubell, E. 2023. Closed AI Models Make Bad Baselines.
- Rothschild, A.; Meng, A.; Disalvo, C.; Johnson, B.; Shapiro, B. R.; and Disalvo, B. 2022. Interrogating Data Work as a Community of Practice. *Proceedings of the ACM on Human-Computer Interaction*, 6: 1 – 28.
- Russo, D.; Tekiroğlu, S. S.; and Guerini, M. 2023. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics*, 11: 1250–1264.
- Shanahan, M.; McDonnell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.
- Snopes. 2024. Snopes. <https://www.snopes.com/>. [Accessed 09-03-2024].
- Solaiman, I. 2023. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 111–122.
- Spiegel, D. 2024. Der Spiegel. <https://www.spiegel.de/international/>. [Accessed 22-01-2024].
- Stories, L. 2024. Lead Stories. <https://leadstories.com/>. [Accessed 22-01-2024].
- Thalke, R.; Stiglitz, E.; Mimno, D.; and Wilkens, M. 2023. Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9252–9265. . .
- Today, I. 2024. India Today. <https://www.indiatoday.in/>. [Accessed 22-01-2024].
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, (0): .
- Walter, N.; Cohen, J.; Holbert, R. L.; and Morag, Y. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political communication*, 37(3): 350–375.
- Wang, D.; Weisz, J. D.; Muller, M.; Ram, P.; Geyer, W.; Dugan, C.; Tausczik, Y.; Samulowitz, H.; and Gray, A. 2019. Human-AI collaboration in data science: Exploring data scientists’ perceptions of automated AI. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–24.
- WhatsApp. 2024. WhatsApp. <https://www.whatsapp.com/>. [Accessed 10-03-2024].
- Wolfe, R.; and Mitra, T. 2024. The Impact and Opportunities of Generative AI in Fact-Checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1531–1543.
- Wolfe, R.; Slaughter, I.; Han, B.; Wen, B.; Yang, Y.; Rosenblatt, L.; Herman, B.; Brown, E.; Qu, Z.; Weber, N.; et al. 2024. Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1199–1210.
- Zhang, A. X.; Muller, M.; and Wang, D. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).
- Zoom. 2024. Zoom. <https://zoom.us/>. [Accessed 10-03-2024].