Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings

Robert Wolfe
University of Washington
Seattle, Washington, United States
rwolfe3@uw.edu

Isaac Slaughter
University of Washington
Seattle, Washington, United States
is28@uw.edu

Bin Han
University of Washington
Seattle, Washington, United States
bh193@uw.edu

Bingbing Wen
University of Washington
Seattle, Washington, United States
bingbw@uw.edu

Yiwei Yang
University of Washington
Seattle, Washington, United States
yanyiwei@uw.edu

Lucas Rosenblatt
New York University
New York, New York, United States
lucas.rosenblatt@nyu.edu

Bernease Herman
University of Washington
Seattle, Washington, United States
bernease@uw.edu

Eva Brown
University of Washington
Seattle, Washington, United States
evamxb@uw.edu

Zening Qu
University of Washington
Seattle, Washington, United States
zqu@uw.edu

Nic Weber
University of Washington
Seattle, Washington, United States
nmweber@uw.edu

Bill Howe
University of Washington
Seattle, Washington, United States
billhowe@uw.edu

ABSTRACT

The rapid proliferation of generative AI has raised questions about the competitiveness of lower-parameter, locally tunable, open-weight models relative to high-parameter, API-guarded, closed-weight models in terms of performance, domain adaptation, cost, and generalization. Centering under-resourced yet risk-intolerant settings in government, research, and healthcare, we see for-profit closed-weight models as incompatible with requirements for transparency, privacy, adaptability, and standards of evidence. Yet the performance penalty in using open-weight models, especially in low-data and low-resource settings, is unclear.

We assess the feasibility of using smaller, open-weight models to replace GPT-4-Turbo in zero-shot, few-shot, and fine-tuned regimes, assuming access to only a single, low-cost GPU. We assess value-sensitive issues around bias, privacy, and abstention on three additional tasks relevant to those topics. We find that with relatively low effort, very low absolute monetary cost, and relatively little data for fine-tuning, small open-weight models can achieve competitive performance in domain-adapted tasks without sacrificing generality. We then run experiments considering practical issues in bias, privacy, and hallucination risk, finding that open models offer several benefits over closed models. We intend this work as a case study in understanding the opportunity cost of reproducibility and transparency over for-profit state-of-the-art zero shot performance, finding this cost to be marginal under realistic settings.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Machine learning**; **Artificial intelligence**; • **Applied computing**; • **Human-centered computing**;

KEYWORDS

Generative AI, Language Models, Open Models, Transparency, Chatbots, ChatGPT, GPT-4, qLoRA

ACM Reference Format:
Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and Bill Howe. 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 (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3630106.3658966

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658966

1 INTRODUCTION

According to a November 2023 report by The Verge, OpenAI's ChatGPT boasts more than one hundred million weekly users, two million developers using the API, and more than 80% adoption of among Fortune 500 companies, making it one of fastest growing services in history [51]. Despite the influence of OpenAI's flagship language models on the world's ways of working and seeking information, scientists know little about them: details of the architecture, parameter counts, and training data of GPT-3.5-Turbo and GPT-4-Turbo are omitted or glancingly described in the company's technical reports [41]. Reaffirming the "values encoded in machine

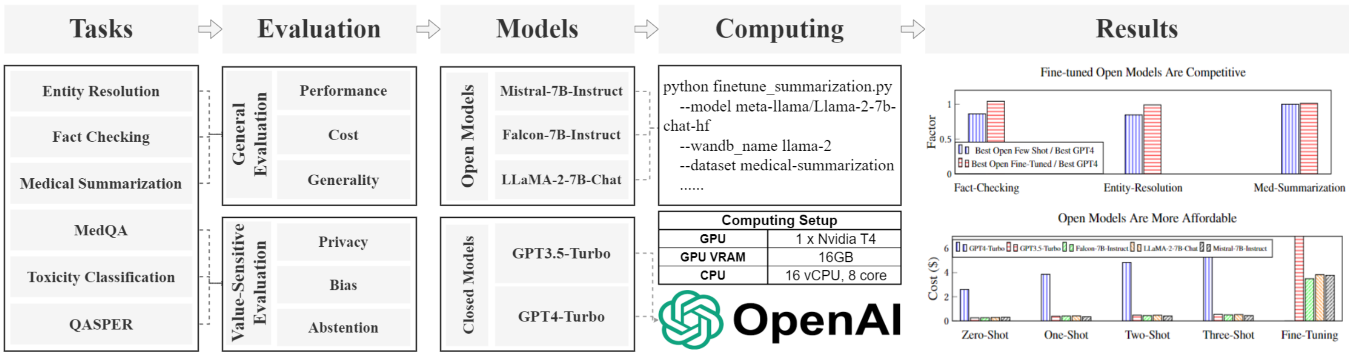

Figure 1: We compared domain-specific performance, general-purpose usability, and amenability to responsible use of three open language models with two dominant closed models. We found that fine-tuning open models renders them competitive with few-shot closed models at low cost.

learning research" described by Birhane et al. [8], transparency has taken a back seat to values that preserve corporate competitive advantage. For many scientists and public interest practitioners, this lack of transparency is at best concerning, and often a reason to avoid such models in their work altogether [38, 46]. At the same time, recent research has enabled the use of accessible and inexpensive hardware to train domain-adapted models. Eight-bit and four-bit quantization allow very large models to run on affordable commercial-grade GPUs [20, 22]. Quantized low-rank adaptation (qLoRA) [21, 32] allow large models to be customized to a domain by adding and tuning a modest number of parameters while allowing most pretrained weights to remain fixed.

These technologies could collectively help enable a future for AI that is not wed to the interests of Big Tech corporations — one that prioritizes transparency, cost-efficiency, and the domain-specific and responsible application of language technologies, in addition to strong performance. In this work, we intend to provide an empirical, practical foundation for this approach, which we call "Laboratory-Scale AI." Concretely, we address the following research questions:

***RQ1**: Do open models offer domain-specific performance competitive with closed models for tasks of scientific and public interest?* We assess open models against closed models on three tasks selected for their scientific or public interest value: government records entity resolution [26], climate misinformation fact-checking [23], and clinical dialogue summarization [5]. We evaluate OpenAI's GPT-3.5-Turbo and GPT-4-Turbo against three open instruction-tuned models: Mistral-7b-Instruct-v.01 [33], Falcon-7b-Instruct [2], and LLaMA-2-Chat-7b [59]. Results show GPT-4-Turbo exceeds the performance of the four other models when using them in a few-shot setting, but GPT-3.5-Turbo and **open models are comparable to GPT-4-Turbo or exceed its performance after fine-tuning for a single dataset epoch**. On fact-checking, fine-tuned Mistral-7b-Instruct achieves accuracy of .75, exceeding the mark of .72 by three-shot GPT-4-Turbo.

***RQ2**: Are open models cost-competitive with closed models?* We find that the cost of running inference on a test dataset with GPT-4-Turbo is comparable to both fine-tuning and inference using an open model. Cost savings achieved by using an open model after fine-tuning are especially notable: on the climate fact-checking test dataset, **inference is almost ten times less costly using an open model** (Mistral-7B-Instruct, $.31) than zero-shot GPT-4-Turbo ($2.65).

***RQ3**: How responsive are small open models to domain-specific fine-tuning data?* We evaluate the performance of LLaMA-2-Chat-7B fine-tuned for clinical dialogue summarization after 0%, 20%, 40%, 60%, 80% and 100% of task training data, and evaluate the performance of the LLaMA-2-Chat-7B, Falcon-7B-Instruct, and Mistral-7B-Instruct every 500 steps of the 4,298-sample fact-checking dataset. After 20% of the fine-tuning dataset (240 samples), the summarization model achieves accuracy of .79 on the test dataset, within .02 of its best. After 2,000 fact-checking samples, Mistral-7B-Instruct achieves accuracy of .71, comparable to fine-tuned GPT-3.5-Turbo. The results show **open models can be adapted with a small amount of data**, without the need for large-scale data collection.

***RQ4**: Can fine-tuned, domain-specific models provide a general-purpose chat-based interface for end users?* Chat-based language models provide an approachable interface to end users. We investigate whether fine-tuning inhibits the utility of this interface by comparing the performance of fine-tuned LLaMA-2-Chat-7B models with the base model on tasks not reflected in the model's fine-tuning dataset. For example, we measure performance of the fine-tuned fact-checking model on the entity resolution task. We find that **fine-tuned open models exhibit performance comparable to general-purpose base chat models**, and in some cases exceed it: for example, the fact-checking LLaMA-2 improves to .85 accuracy on entity resolution, from the base model's .77.

***RQ5**: Can laboratory-scale language models be used in a responsible manner?* We evaluate open and closed models on three tasks of importance for the ethical application of instruction-tuned language models: question answering under differentially private fine-tuning, demographic bias in toxic comment classification, and abstention from answering questions for which insufficient information is available to answer correctly. We find that the performance of open models fine-tuned using a private optimizer approaches non-private fine-tuning, suggesting a better privacy alternative to closed models; that open models exhibit moderate bias that fine-tuning largely fails to mitigate; and that fine-tuning open models can improve their abstention properties: fine-tuned LLaMA-2-7B-Chat achieves an abstention score of .99 (maximum 1.0), exceeding the performance of fine-tuned GPT-3.5-Turbo. **While open models exhibit greater**

bias, they offer greater privacy affordances than closed models, and in some cases abstain more reliably after fine-tuning.

Our experiments demonstrate that a fine-tuned open model running on inexpensive hardware can exceed the performance of GPT-4-Turbo at lower cost. In addition to our core empirical contributions, we offer a practical discussion of the challenges and opportunities of adopting a laboratory-scale approach in the Discussion section.

2 RELATED WORK

We review the related work on generative language models, with attention to model accessibility and evaluation.

Generative Language Models Generative Pretrained Transformers (GPTs) adopt a modified transformer deep learning architecture [62], employing decoder layers to generate an output conditioned on the preceding input [52]. GPT language models [11, 42, 53] are pretrained on the "causal" language modeling objective, taking as input a series of subword tokens and generating the predicted next token [17]. As a result, GPTs can generate freeform output, responsive to a user's input [52]. Early applications of GPTs demonstrated few-shot prompting capabilities: given examples of desired behavior at inference (rather than in training), the model generates a task-relevant output [11]. More sophisticated prompting strategies like chain-of-thought also provide explanations of the reasoning process to derive correct answers [67]. A "zero-shot" setting refers to asking the model to produce correct output with no examples [64].

While remarkably effective for NLP tasks, models that require prompts with examples in context generally serve as a poor interface for lay users, and can result in unpredictable model behavior [18]. To provide a more naturalistic and reliable interface, Ouyang et al. [45] trained GPT models to adhere to the natural language instructions of a human user, with or without examples. Models like OpenAI's ChatGPT [42] and Meta's LLaMA [59] are fine-tuned to engage in dialogue, producing a "chat" based interface wherein the model and a human user take turns, with the user typically providing an instruction or request [55]. Such models are typically trained to be safe and helpful to the end user via reinforcement learning from human feedback (RLHF) [3], or direct optimization using a language model as the reward model [54]. Our research studies decoder-only generative language models fine-tuned to follow instructions of human users.

Improving Model Accessibility Most generative language models use billions of trainable parameters to mimic human language [2, 33, 60]. Training or deploying such models requires vast financial resources, making them difficult to train and access for researchers and public interest practitioners [6]. However, while pretraining LLMs remains prohibitively expensive, recent techniques mitigate the difficulties of using models with low-cost hardware. For example, quantization loads a model in a lower level of precision than used during pretraining [20]. While most generative models are pretrained in 32-bit or mixed 32/16-bit precision, quantization loads the weights in 8-bit [20], 4-bit [22], or even 2-bit precision [14, 22], reducing memory demands. Because transformer models achieve much greater speed using GPUs, the bottleneck for efficiently deploying a model is often the amount of memory (Video-RAM) on the GPU device [21]. However, quantization alone does not permit efficient training on commercial grade hardware [20]. Methods known as *parameter efficient fine-tuning* (PEFT) attempt to preserve the general-purpose functionality of a pretrained model while adapting it for a specific task [24, 74]. Among the most widely used PEFT techniques is low-rank adaptation (LoRA) [32]. LoRA inserts small, trainable weight matrices into a pretrained model, which are fine-tuned while leaving the learned parameters of the pretrained model unchanged [32], reducing fine-tuning memory costs. LoRA weights also require less space than a fully fine-tuned model [32]. Saving a fine-tuned LLaMA-2-7B-Chat model would require about 13.5GB of storage; saving only the LoRA weights — which can later be inserted into the pretrained model — requires only around 260MB [21, 32]. Dettmers et al. [21] introduced qLoRA, a method for allowing trainable LoRA weights to be inserted into quantized models, allowing relatively large models to be mounted on a small GPU and customized using LoRA [21].

Language Model Benchmarking and Evaluation Language models are typically evaluated on specific tasks intended to measure performance on a function of interest, such as sentiment analysis or machine translation. Benchmarks for language models consist of collections of these tasks that assess model ability in a wider domain [16]. For example, the Massive Multitask Language Understanding (MMLU) benchmark measures scientific and world knowledge acquired during pretraining [31]. At a higher level, Stanford's Holistic Evaluation of Language Models (HELM) collects tasks and benchmarks, and asks human users to competitively evaluate model responses to user inputs [37]. Efforts to use human preferences to evaluate models in the wild include LMSYS Chatbot Arena, where users interact with two anonymous language models and vote for a preferred model [77]. Bommasani et al. [10] introduce the Foundation Models Transparency Index, which scores models on transparency-related metrics such as model availability and training details.

Benchmarking ChatGPT The popularity of ChatGPT has prompted researchers to evalute it against traditional NLP methods and models. Kocoń et al. [35] evaluate ChatGPT against state-of-the-art NLP models on 25 NLP tasks, finding that GPT-3.5-Turbo and GPT-4 are outperformed by these models and methods. Thalken et al. [58] show that a fine-tuned LEGAL-BERT [13] is the best-performing model for classifying legal reasoning, outperforming models like GPT-4 and LLaMA-2-Chat. Loukas et al. [39] find that fine-tuned sentence transformer models outperform few-shot GPT-3.5-Turbo and GPT-4 on a financial text classification task. Wang et al. [65] find that a fine-tuned BERT can outperform ChatGPT on sentiment analysis. We build on prior work in benchmarking by comparing open and closed models, but differ by focusing on autonomy, transparency, and responsible use as much as performance.

3 APPROACH

We review the models studied, evaluations employed, and consistent cloud environment used across our experiments.

3.1 Models

The models studied share the following characteristics:

- **Causal (Generative) Pretraining Objective**: All models share the causal language modeling (next-word prediction) objective introduced to the transformer architecture by Radford et al. [52].
- **Instruction-Following**: All models undergo supervised fine-tuning to enable a user to issue instructions in natural language, and receive a natural language response from the model [45].
- **7-Billion Parameters (Open Models)**: The open models each have approximately seven billion trainable parameters, allowing them to be deployed on identical cloud instances. OpenAI has not disclosed parameter counts for GPT-3.5-Turbo and GPT-4-Turbo, but studies suggest they are much larger than open models [48].

We study only generative, instruction-following models for three reasons. First, this accords with the architecture and training regimen of the closed, industry-dominant OpenAI models against which we assess open models. Second, both the closed and open models studied are among the most widely used language models in the world as of this writing, with Meta's LLaMA-2 model and Mistral's Instruct model routinely among the most popular models in the HuggingFace Transformers Python library. Third, these models provide an approachable natural language interface for users who may not be skilled in machine learning but would nonetheless benefit from the use of a domain-aligned language model. In addition to being aligned with our goal of empowering scientists and public interest users, the importance of an accessible interface is borne out by the success of ChatGPT, which far exceeds the userbase of OpenAI's own GPT-3 base models [4]. Finally, studying one group of similar models permits use of consistent infrastructure, allowing us to evaluate cost.

3.1.1 Defining Closed vs. Open Models. We define a *closed* model as a model which is accessible only via a call to an API, and the weights and architecture of the model cannot be accessed. An *open* model is one for which the pretrained weights and architecture are made available and can be modified and built upon. These models are not necessarily licensed to permit any use of the model, as such licenses may still prohibit commercialization or use for unethical purposes as defined by the organization releasing the weights [59, 60]; that is, *open* models are not necessarily fully *open source* models. This definition of "open" aligns with that employed by Palmer et al. [46] and Rogers et al. [56], but omits the requirement that researchers know the data on which the open model was trained, as even in previous definitions, data requirements come with the caveat that such data need not actually be "available for direct inspection" [46].

3.1.2 Closed Models. We study two closed OpenAI models: GPT-3.5-Turbo and GPT-4-Turbo.

- **OpenAI GPT-3.5-Turbo**: OpenAI's cost-efficient and broadly performant model optimized to follow instructions [42, 43]. A GPT-3.5-Turbo fine-tuned with RLHF and Proximal Policy Optimization is the model available to non-paying users who access ChatGPT through the online interface rather than the OpenAI API [42]. We used OpenAI's default GPT-3.5-Turbo at the time of our experiments, which points to "gpt-3.5-turbo-0613" [43].
- **OpenAI GPT-4-Turbo**: OpenAI's state-of-the-art language model, available at greater cost than GPT-3.5-Turbo [43, 44]. GPT-4-Turbo holds the zero-shot state-of-the-art on numerous NLP tasks as of this writing, and achieves first place in human evaluations of chat-based models in Chatbot Arena [77]. GPT-4-Turbo handles much longer text input sequences (128,000 tokens) than GPT-3.5-Turbo, as well as multiple input modalities, such as images [43].

3.1.3 Open Models. We study the following three open models:

- **TII Falcon-7B-Instruct**: A generative model pretrained on 1.5 trillion tokens of the RefinedWeb dataset [49], released under the Apache 2.0 license by the UAE's Technology Innovation Institute (TII) in April 2023 [2]. TII's RefinedWeb dataset consists of filtered web data, and a subset is publicly available [49].
- **Meta LLaMA-2-7B-Chat**: A generative model pretrained on two trillion tokens of publicly available datasets and made available under the LLaMA 2 Community License by Meta AI in July 2023 [60]. The Chat model was fine-tuned for dialogue and underwent RLHF to improve helpfulness and minimize toxic output [60].
- **Mistral AI Mistral-7B-Instruct-v0.1**: A generative model released under the Apache 2.0 license by Mistral AI in September 2023 [33]. Mistral-7B-Instruct-v0.1 is trained on an undisclosed quantity of data from the open internet, and exceeds LLaMA-2-7B-Chat and LLaMA-2-13B-Chat on common benchmarks [33].

3.2 Model Evaluation

We evaluate models in zero-shot, few-shot, and fine-tuned settings.

- **Zero-Shot**: The model is provided with a bare instruction of the task, and given the data to perform the task.
- **Few-Shot**: The model is provided with an instruction and examples of how to respond. We use multi-turn formatting to provide few-shot examples to the model, following the HuggingFace chat template documentation for open models, and OpenAI's documentation for closed models. Falcon-7b-Instruct is not fine-tuned with a defined chat template, and we adhere to the guidance of the model's developers, including examples in a single user prompt.
- **Fine-Tuned**: The model is fine-tuned on a task-specific dataset before evaluation on the task's test dataset. For consistency, we fine-tune for a single dataset epoch, reporting total examples in train and test datasets. We are unable to fine-tune GPT-4-Turbo, for which fine-tuning is available only via an experimental program.

3.2.1 Hyperparameters. We employ four-bit quantization [21] in both inference and fine-tuning. We use qLoRA adapters [21, 32] to fine-tune on domain-specific data, adopting the optimal hyperparameters specified by Dettmers et al. [21]. Specifically, we use qLoRA to tune linear layers, set LoRA matrix rank to 32, and set LoRA dropout to .05, which improves performance in models with fewer than 13-billion parameters [21]. We used gradient checkpointing during fine-tuning to save memory by recomputing activations during the model's backward pass [12, 15]. We set batch size to 1 due to memory limitations. We use the default hyperparameters for training GPT-3.5-Turbo, with the exception of fine-tuning for only one dataset epoch, rather than the OpenAI default of three.

Representative Task	Train Samples	Validation Samples	Test Samples	Eval Metrics
Entity Resolution	700	100	200	Accuracy, F1 Score
Climate Fact Checking	4,298	1,842	1,535	Accuracy, Weighted F1
Clinical Dialogue Summarization	1,201	100	400	BLEU [47], BERTScore F1 [76]

Table 1: Representative tasks with total training, validation, and test samples, as well as evaluation metrics.

3.3 Cloud Infrastructure

We use a consistent cloud environment, allowing comparison of the cost and runtime of open vs. closed models. We defined the environment such that a 7-billion parameter model could be fine-tuned using qLoRA in 4-bit precision with a 1,024-token context window. We chose this setup because 7-billion parameter models are the lowest entry point for the three families of models we study (LLaMA-2-Chat-7B, Falcon-7B-Instruct, and Mistral-7B-Instruct), because fine-tuning in four-bit precision is competitive with fine-tuning in higher precision [21], and because our tasks (*e.g.*, summarization), benefit from a context window of at least 1,000 tokens. Fine-tuning used a $0.32 per hour Google Cloud Platform (GCP) [9] spot instance with the following characteristics: a 16GB Nvidia T4 GPU; 60GB RAM; a 16vCPU, 8-core processor; and 200GB disk. While cost may vary based on region and provider, we found price was generally consistent on GCP and other providers such as AWS and Lambda Labs, within about $.05 per hour. Because we expect that most laboratory-scale AI applications will be fault-tolerant during fine-tuning, we use spot instances, which may be terminated to support higher paying workloads, but are less costly than on-demand resources.

4 MULTIFACETED EVALUATION OF OPEN VS. CLOSED MODELS

We select a practical, representative sample of tasks, including those that 1) reflect real-world uses of generative instruction-tuned models (*e.g.*, fact-checking chatbots, like Aos Fatos' FatimaGPT [25] or Meedan's Check [40]); and 2) reflected consequential work envisioned by other research. For example, Gilardi et al. [28] suggest ChatGPT can be used for data annotation (we consider specifically entity resolution), and Waisberg et al. [63] explore GPT-4 for triaging patients via clinical dialogues. We acknowledge it may not be *desirable* to use an LM in a setting like clinical dialogue summarization or fact-checking, especially without human supervision, and that our tasks are *proxies* to real-world applications.

4.1 Representative General Tasks

We study three tasks to compare performance of open vs. closed models, with sample and evaluation metrics in Table 1.

(1) **Entity Resolution**: We use a custom dataset of public records to evaluate performance on Entity Resolution [26]. Given two pairs of names and addresses, the model determines whether the pairs refer to the same person. One set is derived from home deeds in Mecklenburg County, NC; the other comes from voter records. The dataset contains 1,000 records annotated by three humans (Krippendorff's α of 0.88, 95% CI: 0.85, 0.90 [75]).
(2) **Fact-Checking**: We use the Climate-FEVER dataset [23] to evaluate performance on a fact-checking task. Given a climate-related claim and an associated piece of evidence, the model answers whether the evidence Supports, Refutes, or provides insufficient information to support or refute the claim [23]. For predefined training, validation, and test splits, we use the version of this dataset available at https://huggingface.co/datasets/amandakonet/climate_fever_adopted, used in fine-tuning in-domain climate fact-checking models like a Climate-BERT [66].
(3) **Clinical Dialogue Summarization**: We use the MTS-Dialog dataset [5] to evaluate models on clinical dialogue summarization, following prior work [30, 71]. Given a dialogue between doctor and patient, plus the topic (*e.g.,* medication history, chief complaint), the model must summarize the dialogue, capturing information relevant to the topic.

A simple postprocessing script removed extra words so model output could be measured against labels for tasks 1-2. Given "The answer is Supports" for fact-checking, the script removes "The answer is."

4.2 Performance — Fine-tuned Open Models Can Outperform Closed Models

As shown in Table 2, GPT-4-Turbo outperforms open models in the few-shot setting, and by substantial margins for the entity resolution and fact-checking tasks. Of the open models, only Mistral-7B-Instruct is competitive with GPT-3.5-Turbo in the few-shot setting. Fine-tuning for a single dataset epoch, however, yields open models that are competitive and in some cases even outperform GPT-4-Turbo and fine-tuned GPT-3.5-Turbo. LLaMA-2-7B-Chat achieves no more than 25% accuracy on the fact-checking task in any few-shot setting, yet outperforms GPT-4-Turbo after fine-tuning. GPT-4-Turbo also achieves the best few-shot performance on medical summarization. With fine-tuning, though, Mistral-7B-Instruct outperforms GPT-4-Turbo few shot, achieving higher BLEU score (but not higher BERT score) than GPT-3.5-Turbo, while fine-tuned LLaMA-2-7B-Chat and Falcon-7B-Instruct achieve results competitive with few-shot GPT-4-Turbo.

4.3 Cost Analysis — Open Models Are More Affordable

To better understand the financial cost of customizing and using open models versus using closed models out of the box, we compute the approximate cost of inference and of fine-tuning for the climate fact-checking task. For closed models, we obtain the number of input tokens in our test dataset using the tiktoken tokenizer for OpenAI models. We multiply this total by the per-token costs published by OpenAI. We omit the cost of output tokens in this computation, which we estimate to be less than 1% of the total cost of inference for our tasks. We compute cost for open models by taking the per-hour price of our cloud instance times the runtime logged to our Weights and Biases [7] account. Costs reported are

Model	Scenario	Entity-Resolution		Fact-Checking		Med-Summarization	
		Acc	F1	Acc	F1	BLEU	BERT-F1
GPT-4-Turbo	Zero-Shot	0.93	0.94	0.72	0.72	0.06	0.78
	One-Shot	0.93	0.94	0.72	0.70	0.08	0.79
	Two-Shot	0.97	0.98	0.67	0.68	0.08	0.80
	Three-Shot	0.97	0.97	0.72	0.72	0.08	0.80
GPT-3.5-Turbo	Zero-Shot	0.75	0.78	0.43	0.42	0.05	0.76
	One-Shot	0.85	0.87	0.52	0.52	0.07	0.78
	Two-Shot	0.79	0.79	0.42	0.40	0.08	0.79
	Three-Shot	0.78	0.78	0.52	0.52	0.08	0.79
	Fine-Tuned	0.97	0.97	0.73	0.71	0.07	0.85
Mistral-7B-Instruct	Zero-Shot	0.83	0.86	0.62	0.62	0.06	0.77
	One-Shot	0.69	0.64	0.62	0.62	0.07	0.79
	Two-Shot	0.64	0.58	0.50	0.53	0.07	0.79
	Three-Shot	0.82	0.84	0.59	0.61	0.07	0.80
	Fine-Tuned	0.97	0.98	0.75	0.74	0.10	0.81
Llama-2-7B-Chat	Zero-Shot	0.68	0.79	0.25	0.11	0.02	0.70
	One-Shot	0.60	0.75	0.25	0.11	0.06	0.76
	Two-Shot	0.60	0.75	0.24	0.10	0.06	0.78
	Three-Shot	0.77	0.80	0.24	0.10	0.06	0.79
	Fine-Tuned	0.97	0.98	0.74	0.73	0.08	0.80
Falcon-7B-Instruct	Zero-Shot	0.59	0.75	0.46	0.46	0.07	0.78
	One-Shot	0.59	0.73	0.23	0.29	0.04	0.73
	Two-Shot	0.60	0.75	0.16	0.13	0.05	0.74
	Three-Shot	0.60	0.75	0.16	0.12	0.04	0.74
	Fine-Tuned	0.96	0.97	0.73	0.72	0.09	0.78

Table 2: Performance for three open and two closed models on two classification tasks and one text summarization task. GPT-4 outperforms other models in few-shot settings, but open models are competitive after fine-tuning with modest assumptions.

consistent with billing by OpenAI and GCP. We also report runtime for open and closed models.

If laboratory-scale AI is feasible, we expect open models to be cost-competitive with closed models, and ideally more affordable. Table 3 shows that the few-shot cost of GPT-4-Turbo is approximately ten times that of a few-shot open model or GPT-3.5.-Turbo. The cost of fine-tuning any open model for one dataset epoch and evaluating it once ("Fine-Tuning" in Table 3), a process which Performance results indicate produces a superior fact-checking model to GPT-4-Turbo, is lower than the cost of running inference once using GPT-4-Turbo in the one-shot setting. The most significant savings come when using the model after fine-tuning ("Fine-Tuned" in Table 3). Fine-tuned open models are much less expensive than GPT-4-Turbo, and more performant than few-shot closed models.

Closed models excel on runtime. Fine-tuned GPT-3.5-Turbo is the fastest option, and ten times faster than open models. Few-shot GPT-4-Turbo requires 1.5 times as long as few-shot GPT-3.5-Turbo, but is three times as fast as open models. Our measurements do not include all costs, such as purchasing persistent disk storage, static IPs, and more reliable cloud instances, but provides an empirically grounded analysis of the cost of entry to locally train and deploy a model.

4.4 Data Responsiveness — Modest Fine-tuning Can Make Open Models Competitive

To understand the amount of data needed to produce a domain-specific open model, we study the performance of LLaMA-2-7B-Chat checkpoints for clinical dialogue summarization, entity resolution, and climate fact-checking tasks. We save intermediate model weights at 20%, 40%, 60%, 80%, and 100% of each task-specific training dataset, and assess the intermediate model on the full test dataset. Moreover, for the climate fact-checking task, which has a larger training set of 4,298 samples, we save checkpoints every 500 samples, and assess accuracy using these checkpoints on 150 test samples (approximately 10% of the test dataset). We save these 500-step fact-checking checkpoints for LLaMA-2-7B-Chat; Mistral-7B-Instruct; Falcon-7B-Instruct; and GPT-3.5-Turbo. Because OpenAI does not allow saving model checkpoints during fine-tuning, we submit separate fine-tuning jobs for GPT-3.5-Turbo using subsets of the training dataset.

If laboratory-scale AI is feasible, we would expect that massive data-gathering projects would not be needed to produce a competitive in-domain model. As shown in Figure 2 (left plot), LLaMA-2-Chat-7B achieves BERTScore-F1 of .79 on clinical dialogue summarization after only 20% of training samples (240 samples), and .97 F1 on entity resolution after only 40% of training samples. Similarly (right plot), Mistral-7B-Instruct trained on climate fact-checking achieves accuracy of .71 after 2,000 samples, while LLaMA-2-Chat-7B achieves accuracy comparable to fine-tuned GPT-3.5-Turbo after about 3,500 samples. Fine-tuned laboratory-scale models capable of results comparable to GPT-4-Turbo can be trained using quantities of data feasible for researchers to gather. Variance among open models reflects base model benchmark performance, with Mistral generally outperforming LLaMA-2, and LLaMA-2 outperforming Falcon [31, 37], suggesting pretraining disparities (*e.g.*, LLaMA-2 pretraining on a larger dataset than Falcon) carry over during domain adaptation.

4.5 Model Generality — Fine-tuning Does Not Inhibit the Generality of Open Models

While fine-tuning may improve the performance of a chat-based model for a specific task, it is not clear whether this would compromise the model's general-purpose utility when a user interacts with it via natural language. To study whether the model maintains this utility, we evaluate each of the domain-specific (entity resolution,

Model	Scenario	Input Tokens	1k Token Cost	Runtime Hours	Cloud Cost	Total Cost
GPT4-Turbo	Zero-Shot	260,056	$0.010	0.32	N/A	$2.60
	One-Shot	385,926	$0.010	0.34	N/A	$3.86
	Two-Shot	484,166	$0.010	0.31	N/A	$4.84
	Three-Shot	550,171	$0.010	0.32	N/A	$5.50
GPT3.5-Turbo	Zero-Shot	260,056	$0.001	0.20	N/A	$0.26
	One-Shot	385,926	$0.001	0.23	N/A	$0.39
	Two-Shot	484,166	$0.001	0.20	N/A	$0.48
	Three-Shot	550,171	$0.001	0.20	N/A	$0.55
	Fine-Tuning	260,056	$0.003	1.54	N/A	$6.60
	Fine-Tuned	260,056	$0.003	0.11	N/A	$0.78
Falcon-7B-Instruct	Zero-Shot	N/A	N/A	0.84	$0.32	$0.27
	One-Shot	N/A	N/A	1.24	$0.32	$0.40
	Two-Shot	N/A	N/A	1.37	$0.32	$0.44
	Three-Shot	N/A	N/A	1.58	$0.32	$0.50
	Fine-Tuning	N/A	N/A	9.95	$0.32	$3.18
	Fine-Tuned	N/A	N/A	0.96	$0.32	$0.31
LLaMA-2-7B-Chat	Zero-Shot	N/A	N/A	0.91	$0.32	$0.29
	One-Shot	N/A	N/A	1.27	$0.32	$0.41
	Two-Shot	N/A	N/A	1.48	$0.32	$0.47
	Three-Shot	N/A	N/A	1.64	$0.32	$0.53
	Fine-Tuning	N/A	N/A	10.92	$0.32	$3.49
	Fine-Tuned	N/A	N/A	1.08	$0.32	$0.34
Mistral-7B-Instruct	Zero-Shot	N/A	N/A	0.97	$0.32	$0.31
	One-Shot	N/A	N/A	1.10	$0.32	$0.35
	Two-Shot	N/A	N/A	1.25	$0.32	$0.40
	Three-Shot	N/A	N/A	1.37	$0.32	$0.44
	Fine-Tuning	N/A	N/A	10.90	$0.32	$3.49
	Fine-Tuned	N/A	N/A	0.92	$0.32	$0.29

Table 3: Open models are less costly than GPT-4-Turbo, based on costs computed using fact-checking data. The cost of fine-tuning GPT-3.5-Turbo includes 727,845 Training Tokens, billed at $0.008 per 1,000.

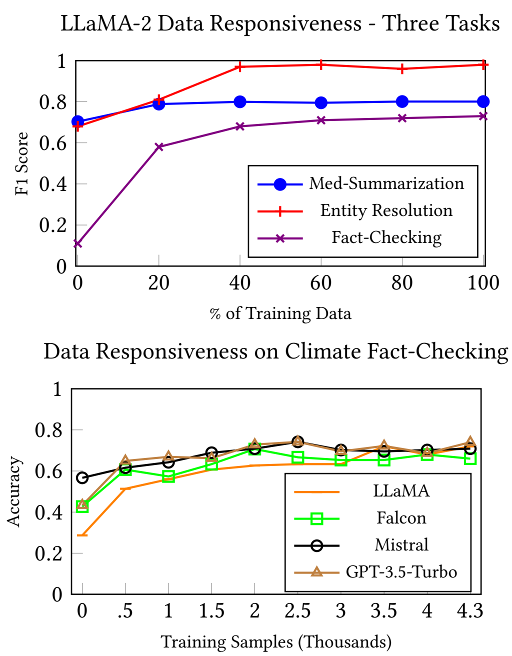

Figure 2: Left: Fine-tuning improvements emerge during the first 50% of the training data, only a few hundred training samples in the case of Medical Summarization and Entity Resolution. Right: Finetuned open models are competitive with finetuned GPT-3.5-Turbo with little data (1,000 fact-checking samples).

fact-checking, and clinical dialogue summarization) LLaMA-2-Chat-7B models on the other tasks for which the model was not fine-tuned. We then compare the domain-specific model's performance on each task against the general-purpose base LLaMA-2.

If the model maintains a general purpose utility, we would expect to see at worst insignificant decreases in the performance of a fine-tuned model when compared with the base model. As illustrated in Figure 3, performance actually increases marginally in most cases when using fine-tuned models on tasks for which they were not fine-tuned. For example, the fine-tuned fact-checking model exceeds base model performance in the one, two, and three shot settings for the entity resolution task. This may not mean that low-rank fine-tuning will always improve performance on related tasks, but our findings suggest that fine-tuning for a specific domain does not degrade the general-purpose utility of an open model.

5 RESPONSIBLE USE OF OPEN MODELS

One of the presumed advantages offered by closed models is the process used to mitigate bias and prevent the closed model from generating harmful or inaccurate output. We thus evaluate three scenarios related to responsible and transparent model use: question answering under differential privacy (privacy), toxicity classification (bias), and abstention, referring to a model refusing to confidently answer questions for which it does not have the answer (transparency).

5.1 Differential Privacy — Privately Fine-tuned Open Models Approach Non-Private Performance

Differentially private (DP) deep learning (using a privatized gradient descent optimizer [1]) has been adopted to protect users and avoid legal risks of sensitive data use [27, 29]. While challenging in the context of language models [50], recent work [70, 73] demonstrates

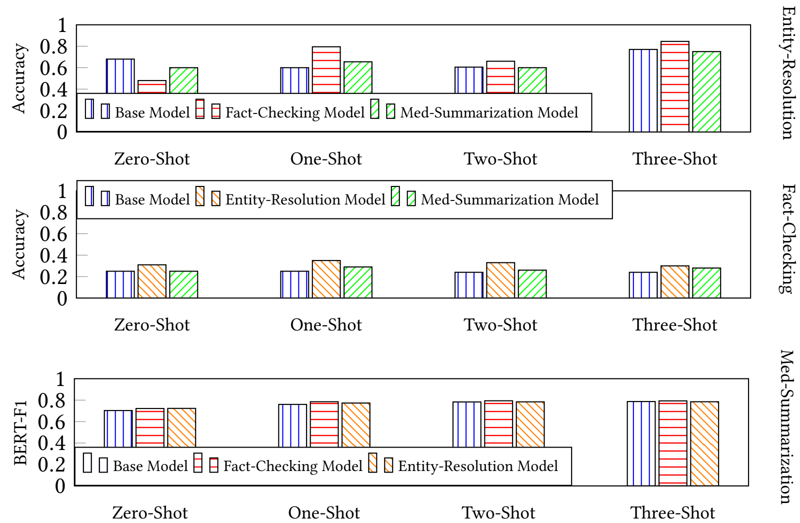

Figure 3: **Models fine-tuned on a task using qLoRA offer strong zero-shot performance on other tasks, often stronger than the base model.**

the potential to train general purpose models using differentially private fine-tuning on sensitive data [72]. We adopt the perspective of a small medical lab with sensitive data, seeking to privately fine-tune an open-source, general purpose medical model. We use the MedQA [34] task as a proxy for this scenario (included in the MultiMedBench [61] benchmark), simplifying it to a binary classification task. We employ private fine-tuning with qLoRA, and report results at five levels of privacy (ϵ = 0.5, 1, 5, 20, ∞, where lower ϵ denotes greater privacy, and $\epsilon = \infty$ is non-private).
Table 4 illustrates how challenging MedQA-TF proved for open models, which performed much lower than the state-of-the-art [61]. However, our results show that private fine-tuning allowed a model like Mistral-7B-Instruct to approach its non-privately fine-tuned performance at ϵ=20. Figure 4 demonstrates how different privacy settings used in Mistral-7B-Instruct impact evaluation loss curves, showing that at lower ϵ, models take longer to converge. A challenge of noisy, privatized updates is that batch size needs to be large, posing issues for lab-scale approaches that use smaller batches.

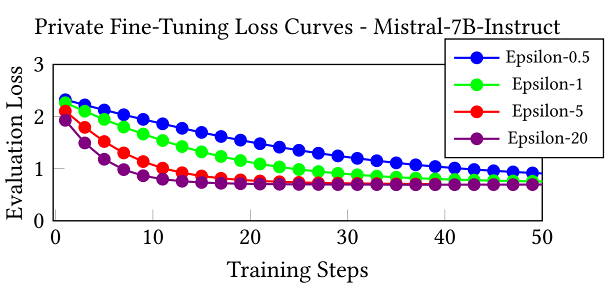

Figure 4: **Increasing privacy (by decreasing ϵ) leads to noisier gradients, delaying convergence; but privately trained open models do learn.**

5.2 Toxicity Bias — Open Models Improve with Fine-Tuning, But Lag Behind Closed Models

We evaluate open and closed models on a subset of CivilComments-WILDS [36], a dataset of real online comments curated from the Civil Comments platform. Dataset labels describe the toxicity of the comment and whether a demographic membership is mentioned in the comment. Models must classify whether a comment is toxic, and their classifications are analyzed through the lens of performance and fairness (whether classifications are incorrect more often for certain demographic groups). We report 1) accuracy on all comments assessed and 2) worst-group accuracy, which represents the lowest accuracy after segmenting the model's output by demographic membership and toxicity label (*e.g.,* worst-group accuracy might refer to accuracy for non-toxic comments and male demographic membership). To ensure a controlled and interpretable experiment, we limited the demographic groups to Male and Female, such that the measurements correspond to gender bias. We used 800 training, 100 validation, and 200 test samples from the dataset. Training, validation, and test data were balanced across the four groups (Male Toxic, Male Non-Toxic, Female Toxic, Female Non-Toxic).

As shown in Table 5, closed models outperform open models on this assessment. Fine-tuning improves overall (Mean) accuracy for Mistral and Falcon, but had no discernable effect for LLaMA-2. Fine-tuning did not increase Worst-Group accuracy over performance in the few-shot setting for any of the open models. The strongest performing model of the group was three-shot GPT-4-Turbo, which exceeded other models in both Mean and Worst-Group accuracy. Fine-tuned GPT-3.5-Turbo matches three-shot GPT-4-Turbo on overall accuracy, but not Worst-Group accuracy. However, the task is difficult, and three-shot Mistral-7B-Instruct surprisingly outperforms zero-shot GPT-4-Turbo on Worst-Group accuracy.

Scenario	Model (finetuned)	Acc. at Privacy Level					F1 at Privacy Level				
		$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 20$	$\epsilon = \infty$	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 20$	$\epsilon = \infty$
MedQA-TF	Falcon-7B-Instruct	0.47	0.51	0.52	0.53	0.52	0.35	0.36	0.36	0.36	0.51
	Llama-2-7B-Chat	0.19	0.41	0.52	0.52	0.56	0.26	0.46	0.53	0.54	0.55
	Mistral-7B-Instruct	0.57	0.58	0.59	0.59	0.65	0.53	0.55	0.56	0.59	0.65

Table 4: **Privately tuned models can approach non-private performance at lower levels of privacy.**

Scenario	GPT-4-Turbo		GPT-3.5-Turbo		Mistral-7B-Instruct		Falcon-7B-Instruct		Llama-2-7B-Chat	
	Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)
Zero-Shot	0.37	0.63	0.61	0.66	0.1	0.53	0	0.51	0.14	0.52
One-Shot	0.37	0.64	0.59	0.63	0.3	0.49	0.4	0.5	0.1	0.5
Two-Shot	0.63	0.68	0.62	0.64	0.41	0.51	0.49	0.53	0.14	0.52
Three-Shot	0.69	0.71	0.54	0.61	0.5	0.56	0.17	0.5	0.09	0.52
Fine-Tuned	-	-	0.66	0.71	0.49	0.57	0.15	0.55	0.11	0.48

Table 5: **Fine-tuning marginally improves toxicity classification accuracy in open models, but closed models still consistently outperform them.**

5.3 Abstention — Fine-tuned Open Models Largely Abstain from Emitting Misinformation

Instruction-tuned language models answer questions based on their parametric knowledge [45] or based on context provided as part of the prompt by the user [57]. If the model has the necessary information in neither its parametric knowledge nor the user-provided context, the model should *abstain* from answering to avoid misinforming a user [68].

We evaluate the ability of open models to abstain by adapting questions from context-dependent scientific knowledge benchmarks, where some questions are designed to be unanswerable if annotators cannot find the answers based on the context. We use the *full* training set from QASPER [19] science question answering dataset to finetune and use the *answerable* questions from the test set to assess abstention in the following way: we remove the context completely, such that the correct answer is to abstain ("Without Context" in Table 7). We use abstention rate to evaluate models' abstention performance following previous work [68]. Ideally, the abstention rate should be 1 if we remove the context completely. In addition to abstention, we evaluate model performance on the full QASPER test set (via F1 score) to assess tradeoffs between overall performance and abstention ability (Table 6). We follow the original split of train, validation, and test sets, resulting in 2,593, 1,005, and 1,451 questions respectively. Table 6 describes task performance on QASPER test set. GPT-4-Turbo excels in few-shot settings. Fine-tuning significantly improves task performance across models, and fine-tuned GPT-3.5-Turbo achieves the highest F1 of 0.74, 0.07 higher than GPT-4-Turbo. Fine-tuning improves Mistral-7B-Instruct, Falcon-7B-Instruct and LLaMA-2-7B-Chat, but performance does not approach GPT-4-Turbo on this challenging task.
Table 7 describes results for the abstention task ("Without Context" means the model is not provided enough information to answer the question and should always abstain) using answerable questions from QASPER test set. Surprisingly, with fine-tuning, abstention performance is reduced for the best question-answering models, suggesting an "overconfidence" effect: Models that are capable of abstaining in the zero-shot setting (GPT3.5Turbo at 0.93 and Mistral-7B-Instruct 0.70) are less likely to abstain in the fine-tuned setting (GPT3.5Turbo at 0.53 and Mistral-7B-Instruct 0.38). However, for models that are *unable* to abstain in the zero-shot setting (Falcon-7B-Instruct at 0.02 and Llama-2-7B-Chat at 0.00), fine-tuning significantly improves this capability (Falcon-7B-Instruct at 0.65 and Llama-2-7B-Chat at 0.99). Results suggest a sweet spot in balancing overall performance with the ability to abstain using ordinary training regimes.

6 DISCUSSION

6.1 The Viability and Implications of Laboratory-Scale AI.

Our work provides empirical support for the viability of adopting a "laboratory-scale" approach to AI that prioritizes user autonomy, privacy, fairness, and transparency while maintaining much of the performance and usability offered by industry-dominant corporate models. With a small GPU card, users can create domain-specific, chat-based language models and deploy them without losing the general-purpose utility and interface that makes such technologies appealing. The laboratory-scale approach intends to address, in a limited capacity, the challenges posed by scholars such as Bender et al. [6], who highlight the dangers of training language models on poorly specified web scraped data generally unrelated to the tasks for which the model will be used; Birhane et al. [8], who describe the performance-centric "values encoded in machine learning research," and highlight the field's capture by big tech companies; and Palmer et al. [46], who contend that scientists and academic researchers must justify the use of proprietary, closed models over open models. Laboratory-scale AI centers the domain-specific, responsible application of small, open models, presenting an option for scientists and public interest technologists who have good reason to avoid closed models that cannot be accessed except via a call to an API.

6.2 Affordances and Challenges of Open Models.

We used the libraries and model ecosystem provided by HuggingFace [69]. The Supervised Fine-Tuning trainer class provided by the TRL library made adapting open language models relatively simple, and primarily dependent on the organization of our data. The

Scenario	GPT-4-Turbo	GPT-3.5-Turbo	Mistral-7B-Instruct	Falcon-7B-Instruct	Llama-2-7B-Chat
Zero-Shot	0.59	0.54	0.36	0.26	0.11
One-Shot	0.56	0.46	0.33	0.31	0.11
Two-Shot	0.60	0.34	0.37	0.21	0.12
Three-Shot	0.67	0.47	0.41	0.16	0.13
Fine-Tuned	-	0.74	0.52	0.45	0.47

Table 6: Fine-tuning significantly improves the performance of open models on the QASPER science question answering dataset, though open models still lag behind few-shot GPT-4-Turbo and finetuned GPT-3.5-Turbo.

Scenario	Model	Without Context
Zero-Shot	GPT3.5-Turbo	0.93
	Falcon-7B-Instruct	0.02
	Llama-2-7B-Chat	0.00
	Mistral-7B-Instruct	0.70
Fine-Tuned	GPT3.5-Turbo	0.53
	Falcon-7B-Instruct	0.65
	Llama-2-7B-Chat	0.99
	Mistral-7B-Instruct	0.38

Table 7: Without context, models that abstain well in the zero-shot setting (GPT3.5 and Mistral) do not abstain well after finetuning. Models that abstain poorly in the zero-shot setting (Falcon and Llama) *improve* after finetuning.

Huggingface ecosystem also supports qLoRA [21], which made customizing quantized models relatively straightforward. However, we nonetheless encountered difficulties with using open models that bear discussion. The most intractable problem we encountered in fine-tuning our own models lay in the difficulty of obtaining cloud instances equipped with even low-cost GPU hardware. We experienced consistent difficulties obtaining results due to lack of available cloud resources. Moreover, we did not expect the quantized open models we tested to run so much more slowly than the closed models we tested. This is related in part to our choice of a low-end GPU, but where inference speed makes a difference, the evidence suggests that cost-efficient, laboratory-scale models still trail closed models.

Open models showed room for improvement on tasks related to responsible use and deployment. While results on differentially private question answering show the potential for privacy-centering open models, they are impeded by small batch sizes required to use low-cost hardware. Fine-tuning has mixed effects for abstention: where a model exhibits strong question answering performance, it is less likely to abstain when it should; but when it exhibits weak question answering performance, it more reliably abstains from answering a question when it should not. Results on the toxicity bias task suggest that open models lag behind closed models on bias mitigation. Though tempting to conclude that the RLHF process used by OpenAI is the right way to address this problem, we note that LLaMA-2-Chat-7B also undergoes RLHF [60], and performs most poorly of any of the models we assess. Future research can contribute by centering these issues.

6.3 Limitations and Future Work.

While our work attempts to provide an open, low-cost approach, we acknowledge that open models have undergone expensive, resource intensive pretraining on large-scale, sometimes opaque datasets. While libraries like qLoRA help to enable adaptations of pretrained models, they cannot equip us with a means of circumventing pretraining, which at this time remains the only reliable means of producing a fluent, general-purpose base model. Future work might explore alternatives that change the pretraining paradigm. We also acknowledge that results from closed models may not be reproducible, should OpenAI change or remove models from its API, potentially without notifying the end user. This is a limitation of closed models that motivates our study, but also necessarily a limitation of our work. Finally, we could not reliably model the carbon cost of closed models due to uncertainties about the exact hardware used to run these models, the location of the data centers on which they run, and practices such as batching user inputs, which may allow for economies of scale.

7 CONCLUSION

We show that small, open, models are competitive with closed models, in that they are cost-efficient, responsive to user data, and robust to fine-tuning. We analyze responsible use of lab-scale models, showing they offer benefits over closed models, particularly in privacy and abstention. We contend that laboratory-scale AI can serve as a basis for future scientific and public interest work, enabling practitioners to customize models without needing to rely on closed, API-based AI.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. , , 308–318.

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867* , (2023), .

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* , (2022), .

[4] Jeff Beckman. 2023. OpenAI Statistics 2023: Growth, Users, and More. https://techreport.com/statistics/openai-statistics/. [Accessed 19-01-2024].

[5] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2291–2302. https://aclanthology.org/2023.eacl-main.168

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. , , 610–623.

[7] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/ Software available from wandb.com.

[8] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. , , 173–184.

[9] Ekaba Bisong and Ekaba Bisong. 2019. An overview of google cloud platform services. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* , (2019), 7–10.

[10] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941* , (2023), .

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[12] Yaroslav Bulatov. 2018. Fitting larger networks into memory. https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9. [Accessed 19-01-2024].

[13] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* , , 2898–2904.

[14] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. 2023. Quip: 2-bit quantization of large language models with guarantees. *arXiv preprint arXiv:2307.13304* , (2023), .

[15] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* , (2016), .

[16] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. 2022. What are the best systems? New perspectives on NLP Benchmarking. *arXiv preprint arXiv:2202.03799* , (2022), .

[17] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *Advances in neural information processing systems* 28, (2015), .

[18] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390* , (2022), .

[19] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4599–4610. https://doi.org/10.18653/v1/2021.naacl-main.365

[20] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339* , (2022), .

[21] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314* , (2023), .

[22] Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*. PMLR, , , 7750–7774.

[23] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614* , (2020), .

[24] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5, 3 (2023), 220–235.

[25] Aos Fatos. 2024. Aos Fatos. https://www.aosfatos.org/. [Accessed 22-01-2024].

[26] Lise Getoor and Ashwin Machanavajjhala. 2012. Entity resolution: theory, practice & open challenges. *Proc. VLDB Endow.* 5, 12 (aug 2012), 2018–2019. https://doi.org/10.14778/2367502.2367564

[27] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. 2021. Deep learning with label differential privacy. *Advances in neural information processing systems* 34 (2021), 27131–27145.

[28] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[29] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and Alex Kai Qin. 2020. A survey on differentially private machine learning. *IEEE computational intelligence magazine* 15, 2 (2020), 49–64.

[30] Bin Han, Haotian Zhu, Sitong Zhou, Sofia Ahmed, M Rahman, Fei Xia, and Kevin Lybarger. 2023. Huskyscribe at mediqa-sum 2023: Summarizing clinical dialogues with transformers. In . CLEF, , , .

[31] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY]

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* , (2021), .

[33] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* , (2023), .

[34] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[35] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* , (2023), 101861.

[36] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, , , 5637–5664.

[37] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[38] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*. , , 1–6.

[39] Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen (Eds.). -, Macao, 74–80. https://aclanthology.org/2023.finnlp-1.7

[40] Meedan. 2024. Meedan. https://meedan.com/. [Accessed 22-01-2024].

[41] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri

Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[42] OpenAI. 2022. Introducing ChatGPT. *OpenAI Blog* , (Nov 2022), .

[43] OpenAI. 2024. Models. https://platform.openai.com/docs/models/. [Accessed 19-01-2024].

[44] OpenAI. 2024. Pricing. https://openai.com/pricing. [Accessed 19-01-2024].

[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[46] Alexis Palmer, Noah A Smith, and Arthur Spirling. 2023. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* , (2023), 1–2.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. , , 311–318.

[48] Dylan Patel and Gerald Wong. 2023. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. https://www.semianalysis.com/p/gpt-4-architecture-infrastructure. [Accessed 19-01-2024].

[49] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* , (2023), .

[50] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. , , 1291–1292.

[51] Jon Porter. 2023. ChatGPT continues to be one of the fastest-growing services ever. *The Verge* , (Nov 2023), .

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. , (2018), .

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[54] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290* , (2023), .

[55] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* , (2023), .

[56] Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed AI Models Make Bad Baselines. https://hackingsemantics.xyz/2023/closed-baselines/

[57] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael McKenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*. , , . https://research.ibm.com/publications/multitask-prompt-tuning-enables-zero-shot-task-generalization

[58] Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 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.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* , (2023), .

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* , (2023), .

[61] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical AI. *arXiv preprint arXiv:2307.14334* , (2023), .

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30, (2017), .

[63] Ethan Waisberg, Joshua Ong, Nasif Zaman, Sharif Amit Kamran, Prithul Sarker, Alireza Tavakkoli, and Andrew G Lee. 2023. GPT-4 for triaging ophthalmic symptoms. *Eye* 37, 18 (2023), 3874–3875.

[64] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–37.

[65] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339* , (2023), .

[66] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010* , (2021), .

[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[68] Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing LLM Abstention Behavior in Science QA with Context Perturbations. *arXiv preprint arXiv:2404.12452* (2024).

[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* , (2019), .

[70] Lukas Wutschitz, Huseyin A. Inan, and Andre Manoel. 2022. dp-transformers: Training transformer models with differential privacy. https://www.microsoft.com/en-us/research/project/dp-transformers.

[71] Wen-wai Yim, A Ben Abacha, N Snider, G Adams, and Meliha Yetisgen. 2023. Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In *CLEF*. , , .

[72] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. 2021. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298* , (2021), .

[73] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500* , (2021), .

[74] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. , , 1–9.

[75] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology* 16, 1 (Aug. 2016), 93. https://doi.org/10.1186/s12874-016-0200-9

[76] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. , , .

[77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]