Data-Centric AI in the Age of Large Language Models

Xinyi Xu[1,2], Zhaoxuan Wu[1,3], Rui Qiao[1], Arun Verma[1], Yao Shu[4], Jingtan Wang[1,2], Xinyuan Niu[1,2], Zhenfeng He[1], Jiangwei Chen[1,2], Zijian Zhou[1,3], Gregory Kang Ruey Lau[1,5], Hieu Dao[1], Lucas Agussurja[1], Rachael Hwee Ling Sim[1], Xiaoqiang Lin[1], Wenyang Hu[1], Zhongxiang Dai[6], Pang Wei Koh[7,8], Bryan Kian Hsiang Low[1]

[1]National University of Singapore, [2]Agency for Science, Technology and Research,
[3]Singapore-MIT Alliance for Research and Technology, [4]Guangdong Lab of AI and Digital Economy (SZ),
[5]CNRS@CREATE, [6]The Chinese University of Hong Kong, Shenzhen, [7]Allen Institute for AI, [8]University of Washington

Correspondence: xinyi.xu@u.nus.edu

Abstract

This position paper proposes a data-centric viewpoint of AI research, focusing on large language models (LLMs). We start by making a key observation that *data* is instrumental in the developmental (e.g., pretraining and fine-tuning) and inferential stages (e.g., in-context learning) of LLMs, and advocate that data-centric research should receive more attention from the community. We identify four specific scenarios centered around data, covering data-centric benchmarks and data curation, data attribution, knowledge transfer, and inference contextualization. In each scenario, we underscore the importance of data, highlight promising research directions, and articulate the potential impacts on the research community and, where applicable, the society as a whole. For instance, we advocate for a suite of data-centric benchmarks tailored to the scale and complexity of data for LLMs. These benchmarks can be used to develop new data curation methods and document research efforts and results, which can help promote openness and transparency in AI and LLM research.

1 Introduction

The latest large language models (LLMs) (Alayrac et al., 2022; Anil et al., 2023b; OpenAI, 2023; Radford et al., 2021; Ramesh et al., 2021, 2022; Rombach et al., 2022; Touvron et al., 2023) are typically trained on extensive corpora of raw data scrapped from the Internet and then fine-tuned on specialized domain data. These LLMs have demonstrated not only incredible performance on benchmarks (Lee et al., 2023b; Liang et al., 2023), but also remarkable abilities to follow and execute human instructions (Ouyang et al., 2022; Wang et al., 2022), and to learn "in-context" (Dong et al., 2023) from the contextual data given by the user along with the query. At the core of these impressive achievements, we identify that data, in different forms, scales, and usages, is a common denominator.

However, the bulk of research to date has focused on modeling improvements, and little is known about how to best use data for the developmental stages (i.e., pretraining and fine-tuning) and the inferential stage (using LLMs for inference or generation). For pretraining, the exact composition of pretraining datasets used by many leading foundation models is proprietary (Anil et al., 2023a; Chen et al., 2021; Li et al., 2022b; OpenAI, 2023), while data scrapped from the Internet is often noisy and can pose legal and security risks (Barrett et al., 2023; Carlini et al., 2023; Henderson et al., 2023; Min et al., 2024). Moreover, since pretraining large models is expensive (e.g., GPT-4 costs over $100 million to build (Knight, 2023)), it is prohibitively costly to evaluate different choices of pretraining data. These characteristics raise the difficulties of identifying the factors that underlie an effective pretraining dataset. Then, for fine-tuning, compared to the array of modeling techniques (Zhang et al., 2023), the methods for data curation are under-explored (Chen and Mueller, 2024) and most prior works adopt manual approaches (Honovich et al., 2023; Wang et al., 2023e; Wei et al., 2022; Ye et al., 2021) which are difficult to generalize and costly to deploy at scale.

It is yet unclear how to push the LLMs' limits beyond what is achievable solely by better modeling techniques. Specifically, we propose to identify a generalizable and cost-effective approach to designing pretraining and fine-tuning datasets to complement the model-centric techniques. Separately, for the inferential stage, there are model-centric efforts "optimizing the instructions" for LLMs to improve how they utilize the user-provided contextual data (Peng et al., 2023; Wang et al., 2023a) but only relatively limited data-centric research on improving the user-supplied contextual data itself, even though the LLM's performance is shown to be sensitive to the contextual data's quality (Liu et al., 2023a) and ordering (Liu et al., 2023d; Lu

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11895–11913
November 12-16, 2024

et al., 2022a).

We advocate for data-centric research that can turn the art of using data into science and unlock the next generation of more effective and compact LLMs. Our position is framed within the following four scenarios of different interactions between LLMs and data; refer to Fig. 1 for a diagrammatic overview. For each scenario, we highlight the unique characteristics and challenges, identify motivating use cases and promising research directions, and discuss potential impacts. We do not claim to be the first to propose these directions, but rather aim to underscore the importance of the data-centric perspective and its impacts. While our exposition is not exhaustive, we hope our "first cut" at a holistic viewpoint of data-centric research can generate more discussion and inspire innovation.

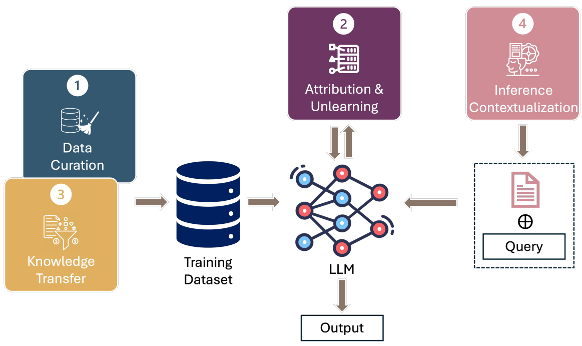

Figure 1: Sec. 2 (indexed 1 in the figure) underscores the importance of the training data (for both pretraining and fine-tuning) and the data curation techniques. Sec. 3 (indexed 2 in the figure) highlights that the LLMs' outputs depend on the training data. Sec. 4 (indexed 3 in the figure) describes the "knowledge" of the LLMs to be transferred from some training data. Sec. 5 (indexed 4 in the figure) demonstrates the usage of data by the LLMs at inference (i.e., response to a query).

Benchmarks and curation for training data. The recent successes of LLMs such as ChatGPT (OpenAI, 2023), PALM 2 (Anil et al., 2023b), and LLaMA 2 (Touvron et al., 2023), as well as vision-language models including CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), Stable Diffusion (Rombach et al., 2022) and DALL-E (Ramesh et al., 2021, 2022), are powered by large, heterogeneous datasets rather than solely by advanced modeling techniques. CLIP is trained on 400 million image-text pairs (roughly 300× greater than the size of ImageNet (Deng et al., 2009)), InstructGPT is trained on thousands of user-supplied and diverse prompts (Ouyang et al., 2022), and LLaVA's instruction dataset contains over 100 thousand image-text pairs (Liu et al., 2023b).

These examples underscore the critical role of better designed and curated training data in further advancing the capabilities of LLMs. However, the heterogeneity, scale, and proprietary nature (Bommasani et al., 2023) of the training data for most of the currently best-performing LLMs significantly impede the progress in developing and training LLMs through curating better training data. To advance the research on data curation, we advocate for building towards rigorous data-centric benchmarks (Sec. 2) on the foundation of existing efforts like DataComp (Gadre et al., 2023).

Data attribution. The training data is a "source" for the outputs generated by LLMs (Keskar et al., 2019). The ability to support source attribution and trace the generated outputs back to the specific training data is imperative for legal and safety purposes: (i) To respect the copyright/intellectual property rights, by correctly accrediting the creators of writings (Eldan and Russinovich, 2023; Rahman and Santacana, 2023), datasets (Li et al., 2022a; Liu et al., 2023e), or code (Lee et al., 2023a). (ii) To mitigate the issue of problematic outputs of the LLMs (e.g., hateful, toxic, harmful messages (Sap et al., 2019; Shelby et al., 2023; Weidinger et al., 2022) or dangerous information (Bommasani et al., 2022)), by identifying and removing the source. Hence, we describe the promising directions for data attribution and removal (Sec. 3).

Knowledge transfer. The costs of developing and deploying LLMs make it challenging to democratize the benefits of LLMs: GPT-4 costs over $100 million to build (Knight, 2023) and is estimated to cost over $21,000 a month for a small business to use for customer service support (Chen et al., 2023b). Hence, a smaller model distilled from its larger counterparts for a specialized domain or task presents a cost-effective alternative (Jiang et al., 2023; Taori et al., 2023; Yu et al., 2024). The Zephyr 7B beta outperforms the 70B Llama 2 in coding, math, and roleplay (Tunstall et al., 2023) while MiniLLM matches the performance in instruction following of an LLM twice its parameter count (Gu et al., 2024). These results open up promising avenues for transferring the knowledge of trained LLMs to compact and specialized models, and we discuss existing efforts and new opportunities where the outputs of a trained LLM are treated as (synthesized) data (Sec. 4).

Inference contextualization with data. In contrast to standard ML models, LLMs have a unique capability of flexibly using data at inference to augment the outputs' factuality (Wang et al., 2023b) or quality (Borgeaud et al., 2022). For example, an LLM can "acquire" a skill on the fly for a user's task via some user-provided examples (Brown et al., 2020). As another example, when queried, an LLM can search through a user-prepared datastore for relevant information as supplementary information for generating a response (Lewis et al., 2020). This capability enables the user to establish the right *context* for the LLM at inference through the data (examples or datastore) and gives rise to an inference contextualization paradigm that can significantly streamline the applications of LLMs. We elaborate on this paradigm w.r.t. two prevalent technical frameworks and highlight how it can improve the personalization of LLMs (Sec. 5).

2 Rigorous Data-centric Benchmarks

There are increasing data-centric efforts on quantitatively understanding how the training data affects LLMs' performance via identifying and improving the scaling laws (DeepSeek-AI, 2024; Hoffmann et al., 2022; Hu et al., 2024; Kaplan et al., 2020; Sardana and Frankle, 2023). However, the datasets that train the state-of-the-art LLMs are often proprietary and closed-source while public datasets do not seem to achieve comparable scaling behavior to their proprietary counterparts (Cherti et al., 2023). Moreover, even for public datasets like C4 (Raffel et al., 2020) or LAION-2B (Schuhmann et al., 2022), the critical factors underlying effective training datasets remain unclear. Indeed, different training data compositions (i.e., proportions of different sources) can lead to vastly different properties of the trained CLIP models (Nguyen et al., 2022) and language models (Anil et al., 2023b; Xie et al., 2023a), while data filtering and pruning can sometimes even outperform the standard power-law scaling (Abbas et al., 2023; Sorscher et al., 2022; Toneva et al., 2018). There are also promising results by sourcing for "clean" data (Gunasekar et al., 2023) or low-perplexity data (Marion et al., 2023).

These observations inspire several questions: What factors (besides the scale) are important to a training dataset (Sachdeva et al., 2024)? How do the data compositions affect the performance (Xie et al., 2023a)? What is a principled methodology to reliably outperform the power-law scaling trends (Sorscher et al., 2022)? While the above works provide excellent starting points, comprehensively addressing these questions requires a series of well-documented results and a systematic approach to identifying and quantitatively analyzing the key underlying factors of LLMs' performance. By building on the foundations in (Gadre et al., 2023; Mazumder et al., 2023) which primarily target conventional ML, we advocate for rigorous data-centric benchmarks catering to LLMs' scale and complexity. We also identify directions (that leverage existing non-LLM-specific techniques) for designing effective data curation methods.

2.1 Benchmarks and Data Curation

A cornerstone towards more efficient and effective LLM training powered by new data curation methods is rigorous and large-scale benchmarks for evaluation and results documentation. The conventional ML benchmarking paradigm is completely flipped in these data-centric benchmarks (Gadre et al., 2023) where the training code and computational budget are held constant so that participants innovate by proposing new training sets (e.g., new sources (Gunasekar et al., 2023) or new filtering techniques (Sachdeva et al., 2024)). We describe two specialized benchmarks, respectively, for designing training datasets and adapting to downstream domains and tasks, and further elaborate how they can be leveraged to design better methods for dataset design and curation.

Benchmarks for heterogeneous and large-scale pretraining data. Two key characteristics of the pretraining data of LLMs are *heterogeneity* (e.g., multi-domain, multi-modality, multi-source) and the unprecedented *scale*. They induce not only an intricate interplay among the different domains, modalities, and sources but also a high complexity and cost of comprehensive evaluations (Lee et al., 2023b; Liang et al., 2023), thus making designing effective curation techniques challenging. Hence, instead of tackling the problem of curating pretraining data outright, we advocate for laying the foundations first by building benchmarks for heterogeneous and large-scale pretraining data based upon existing efforts such as DataComp (Gadre et al., 2023), as a future direction. DataComp is a benchmark for multimodal image-text dataset design for contrastive training of CLIP-like models. Importantly, it spans several orders of magnitude in compute and data scale and includes the largest publicly available collection of over 6 billion image-text pairs, making it a suitable testbed for testing hypotheses and drawing insights w.r.t. the pretraining data for LLMs. For example, one initial finding reports that changing how the training data is filtered led to significant improvements in CLIP-like models over OpenAI's original CLIP models (Gadre et al., 2023). Compared to the few existing benchmarking efforts (often at a smaller data scale (Ng et al., 2021)), DataComp is a more suitable starting point due to its scale and the promising initial findings. Additionally, another related future direction is to investigate the efficiency of pretraining data such as by building on (Warstadt et al., 2023).

Such benchmarking efforts can be complemented by the efforts on open-source pretraining datasets (Lozhkov et al., 2024; Penedo et al., 2024; Soldaini et al., 2024) and can serve as a platform for documenting model performance on specific datasets for the purpose of analysis and comparison. This can further aid the researchers in understanding the overall "landscape" (of data and models) and draw generalizable conclusions, for instance about a quantitative relationship between the perplexity (PPL) of model w.r.t. vocab size, diversity size and other key factors of a dataset. We acknowledge there are challenges with holistic evaluation of LLM performance (Lee et al., 2023b; Liang et al., 2023) and believe these further motivate the suggested benchmarking efforts (possibly paired with existing evaluation frameworks) to implement new evaluation metrics that may additionally depend on independent components (e.g., a hold-out validation dataset).

Benchmarks for adapting to downstream domains and tasks. Users usually want to apply LLMs to their downstream domains or tasks, motivating the investigation of how best to construct domain- or task-specific datasets to fine-tune an LLM pretrained on certain data. For example, if we want to fine-tune a general-purpose LLM for medical tasks, does that general-purpose LLM need to have been pretrained on medical data (and if so, in what proportion), or does it suffice to fine-tune the LLM on a small amount of medical data? As another example, to obtain an LLM for low-resource languages such as Southeast Asian (IMDA, 2023) or African languages (Nguyen et al., 2023), should we fine-tune an LLM pretrained on a mix of languages or one pretrained only on the target language? Due to the specialized nature of these tasks, it is beneficial for future endeavours to explore more specialized adaptations of the existing benchmarking efforts. We suggest the following start points for future works in this direction. For multi-lingual adaptations (e.g., to adapt an LLM pretrained on English text to other languages), both Xtreme (Hu et al., 2020) and TyDi QA (Clark et al., 2020) benchmarks provide the resources for adequate evaluation and are thus suitable potential options. For medical use cases, the CME (Liu et al., 2023c) and MedEval (He et al., 2023) benchmarks provide viable starting points.

Dataset design and curation. The next step is developing methods for curating datasets for training LLMs and adapting them to downstream domains and tasks (i.e., fine-tuning). While there is on-going research in this direction, we further highlight the importance and potential, by describing some possible avenues of exploration.

For training general-purpose LLMs, the data needs to be diverse and spanning multiple distinct domains (e.g., books, Wikipedia, code, academic papers, etc.) (Chowdhery et al., 2022) such that each domain is sufficiently well-represented in the training data to avoid overfitting (Xie et al., 2023b).

The inter-domain and intra-domain curation processes have different requirements, so our proposed future directions (below) have correspondingly different emphases. The inter-domain curation process should maximize heterogeneity, for instance, by incrementally selecting fine-grained domains (Xie et al., 2023a) and adding in a new domain only if it adds to the heterogeneity of the pool of added domains. Statistical testing (Gretton et al., 2012; Wei et al., 2021) or distributional divergence (Ben-David et al., 2010; Wu et al., 2022) are principled methods to determine if a domain adds to the heterogeneity. On the other hand, the intra-domain curation should maximize diversity (Sachdeva et al., 2024), for instance, by integrating classic approaches such as determinantal point processes (Kulesza and Taskar, 2012) and coreset selection (Sener and Savarese, 2018) with existing ML-based data valuation methods (Amiri et al., 2023; Sim et al., 2022).

For adapting to downstream target domains or tasks, a core objective is to address the distribution shift between the target domain and the available training data; otherwise, the model learns irrelevant information about the target domain. In this regard, a "good" data source has a high distributional similarity to the target domain. Hence, as possible future direction is to extend prior data valuation works in standard, unimodal ML settings (Amiri et al., 2023; Just et al., 2023) to efficiently handle multi-modal data at scale. For selecting individual data points, prior works demonstrate the usefulness of influence scores (Choe et al., 2024; Grosse et al., 2023; Guo et al., 2021; Kwon et al., 2024; Xia et al., 2024).

In the proposed directions above, the underlying principle centers around removing duplicates and maximizing diversity, but they have different emphases because of the different intended scenarios. The design of specific techniques (e.g., diversity maximization, de-duplication and coreset selection) should take into consideration of the characteristics of the intended scenarios (e.g., the required scale of data and computational resources for pretraining often implies a much more efficient approach than for fine-tuning). We thus suggest separate benchmarks for pretraining and fine-tuning.

2.2 Data-centric Open LLM Research

With the benchmarking efforts and data curation methods, we hope to initiate a new brand of data-centric LLM research, welcoming openness and transparency. While many efforts have been made to open-source the LLMs such as BLOOM (Scao et al., 2023)[1] and LLaMA 2 (Touvron et al., 2023) and open-source model-centric benchmarks Biderman et al., most of the training data is held closed-source (Bommasani et al., 2023) with a few recent exceptions: Groeneveld et al. (2024) completely open-sourced their training data and pipelines while (NVIDIA, 2024) have made public the data generation pipeline for their aligning process. With this new brand of data-centric open research, we hope to encourage more transparency in future research, which goes beyond the technological advancement itself but is also of great importance towards responsible adoptions of the technology and management of the ensuing socio-economical implications (Bommasani et al., 2022, 2023). For instance, the recently launched National AI Research Resource (NAIRR) by the U.S. National Science Foundation (Alexandria, 2024) lists open research (i.e., NAIRR open) as one of the four focus areas.

[1]At time of writing, the efforts to open-sourcing the training data of BLOOM are underway: BigScience LM data.

3 Data Attribution

For copyright/intellectual property rights considerations, data attribution is primarily motivated by the need for credit attribution. For ensuring safe applications of LLMs, the goal of attribution is to trace (and then remove) the sources of potentially problematic outputs. Notably, data attribution and unlearning are useful to both these use cases.

Since most of the training data for the popular LLMs is scraped from the Internet, it is almost inevitable that the training data contains certain copyrighted data (e.g., writing, code, or even entire datasets). Then, it is important to design techniques to mitigate potential copyright infringements, especially when the data owners or creators request takedowns. This process involves first correctly identifying the source through data attribution and then removing it via unlearning (Eldan and Russinovich, 2023). For sources that lead to problematic outputs by the LLMs, we first identify sources through attribution and then remove (the effects of) the sources through unlearning (Si et al., 2023). The challenge in the unlearning step is to ensure its effectiveness without compromising the performance of the LLM (Chen and Yang, 2023), incurring prohibitive costs from iterative retraining (Si et al., 2023) or needing additional training data (Yao et al., 2023b).

3.1 Data Attribution and Unlearning

We describe data attribution followed by unlearning, which depends on data attribution.

Data attribution. We highlight two proposed approaches where the first targets attribution to individual training data (i.e., more granular), and the second aims to identify a data source among several data sources (i.e., less granular). For the first approach, attribution is by tracing the influence (Koh and Liang, 2017) or determining the value (Ghorbani and Zou, 2019) of individual training data to LLMs. While there have been successes in applying the influence function to attribute the prediction of an ML model to its training data (Koh et al., 2019), there remain challenges in extending it to LLMs. The increasing complexity and size of model architectures significantly raise the computational cost (Grosse et al., 2023) and deteriorate the influence scores' accuracy and utility (Bae et al., 2022; Basu et al., 2021). Two promising approaches to address these computational challenges are efficient approximations (Guo et al., 2021; Grosse et al., 2023), and direct empirical estimators (Guu et al., 2023; Ilyas et al., 2022; Pruthi et al., 2020). Preliminary results demonstrate a computational speedup by reducing the original problem to a much smaller subproblem (Guo et al., 2021) or exploiting certain training structures (Choe et al., 2024; Kwon et al., 2024).

The existing data valuation methods (Ghorbani and Zou, 2019; Jia et al., 2019b; Schoch et al., 2022; Sim et al., 2022; Yoon et al., 2020) can provide attribution by identifying the "most valuable" training data of a model (e.g., LLM). However, a similar scaling issue is encountered when applying these methods to LLMs, especially if they require multiple re-training of the LLM (Schoch et al., 2023). Similarly, potential solutions include efficient approximations (Jia et al., 2019a; Schoch et al., 2023) and training-free surrogates (Just et al., 2023; Nohyun et al., 2023; Wu et al., 2022) for designing scalable data valuation methods for LLMs.

For the second approach, source attribution differs from data attribution in being less granular and aiming to identify a data source instead of individual data. This approach is particularly relevant in use cases involving copyrights or intellectual property rights, where the data source is the intellectual creator. For source attribution, a natural idea is to adopt watermarking as a unique identifier for a piece of writing or design. For LLMs, watermarking techniques are used to identify or pinpoint the data sources that contribute most significantly to a given output (Marra et al., 2018; Yu et al., 2019, 2021). Conceptually, a unique watermark is first assigned to each data source and then inserted into the training data from this source during training. Subsequently, given a generated output during inference, the most influential sources can be identified and correctly attributed by observing which of these watermarks are present in the output. Some specific types of watermarks include linguistic watermarks (Kirchenbauer et al., 2023; Kuditipudi et al., 2023) and (non-linguistic) Unicode character-based watermarks (Wang et al., 2023c).

Unlearning of data. To remove (the effects of) certain identified training data (called target data), the set of unlearning techniques is suitable. The gold standard is to remove the target data and retrain the entire model from scratch on the remaining data, but it is prohibitively expensive for large models (Cao and Yang, 2015; Si et al., 2023) and infeasible when regulations stipulate a short execution time (Graves et al., 2021). Then, one alternative direction is to perform additional fine-tuning of the LLMs using only the remaining data to erase the effect of the target data (Mehta et al., 2022; Neel et al., 2021). Another more directed approach is to leverage the knowledge of the target data to design cost-effective and efficient solutions, e.g., target data-oriented fine-tuning (Yao et al., 2023b) and in-context unlearning to "mimic" unlearning (the knowledge of specific tokens) via careful contextualization at inference time (Pawelczyk et al., 2023).

3.2 Safe and Responsible Deployment of LLM Technologies

The ex-post data attribution and removal are useful for the safe and responsible deployment of LLMs by respecting the copyrights/intellectual property rights and mitigating problematic outputs. These ex-post methods are complementary to possible ex-ante data-centric approaches (e.g., conditioning on certain types of data (Keskar et al., 2019)) or other ex-post approaches (e.g., mitigation at decoding or inference time (Krause et al., 2021; Liu et al., 2021)). Importantly, these methods target different stages of the LLM pipeline (i.e., before training, after training, and during inference) and collectively form "multiple layers of defense" against problematic outputs. Hence, we hope to inspire research towards "multi-layered" approaches for the safe and responsible deployments of LLM technologies.

4 Knowledge Transfer

Given the prohibitive costs of deploying full-fledged LLMs (Chen et al., 2023b; Patterson et al., 2021), and that most users may not need such powerful general-purpose LLMs, the cost-effective adaptations of LLMs to users' specialized tasks are more appealing. In many cases, the general-purpose LLM already has the necessary "knowledge" to perform the specialized task (Li et al., 2023; Xu et al., 2024), which can be transferred to a more compact and specialized model. Knowledge transfer can be performed by first distilling the knowledge from the LLM as *synthesized data*, then instilled into the specialized model by training it on the synthesized data. Since data synthesis is a niche setting arising from the generative capabilities of LLMs and its quality is key to effective knowledge transfers, we focus on data synthesis, specifically label and input syntheses.

4.1 Cost-effective Data Synthesis

We elaborate on label and input syntheses, focusing on the cost-effectiveness (i.e., the size/quantity and quality of the synthesized data).

Label synthesis. The simpler case is where the user starts with a large pool of unlabeled data (e.g., performing sentiment analysis for public comments) and requires the LLM to synthesize the labels. This case resembles the setting of active learning (Gal et al., 2017) where the goal is cost-effectiveness (i.e., using a small number of synthesized labels to achieve a high learning performance). The emphasis on cost-effectiveness (which is similar to in active learning) stems from that many current approaches incur a considerable manual involvement to inspect and ensure the quality of the labels, evidenced by methods specifically designed to minimize such manual involvement (Honovich et al., 2023; Wang et al., 2023e; Wei et al., 2022; Ye et al., 2021). The core idea of this proposed direction is to select and annotate only the most "useful" data, which can be implemented via unsupervised data valuation techniques such as feature-based diversity (Amiri et al., 2023; Xu et al., 2021), uncertainty modeling (Lewis and Catlett, 1994), and optimized heuristics (Bairi et al., 2015). Additionally and different from the conventional unimodal settings, multi-modal classifiers like CLIP (Ilharco et al., 2021; Radford et al., 2021) can be leveraged to perform cross-modal (e.g., image to text) or multi-modal (e.g., image-text to text) label synthesis.

Moreover, the unique explanatory capabilities of LLMs can be exploited (i) to augment the synthesis with additional generated explanations and rationales (Hsieh et al., 2023), and (ii) to be used, not as a "label generator" for direct label synthesis (as above), but as a labeled data "selector". Specifically, from a pool of labeled data (with labels possibly synthesized by an LLM), the LLM is asked to select the high-quality ones. It is useful when the original LLM cannot synthesize labels very accurately but is able to filter out the low-quality, noisy, or incorrect labels (Sachdeva et al., 2024).

Input synthesis. The more challenging scenario arises when no initial data is available, not even unlabeled data, possibly because the specialized task is niche or less well-established and the user does not know what unlabeled data to collect. In this direction, we propose to fully utilize the generative capabilities of LLMs to synthesize coherent and diverse inputs (Ding et al., 2024), such as via prompt engineering and fine-tuning procedures (Li and Liang, 2021) and sophisticated prompting techniques (Naseh et al., 2024). Then, the aforementioned label synthesis techniques can be applied, making label and input syntheses complementary to each other and suggesting it is possible to develop integrated treatments, such as jointly using existing unlabeled input and the generation of new input or using LLMs to complete a partial input with a randomly generated label (Xu and nad Wenpeng Yin, 2022). Notably, the 1.3B phi-1.5 trained (almost) exclusively on synthesized data can outperform models $5\times$ larger (Li et al., 2023) and the recently released Nemotron-4 family (NVIDIA, 2024) further showcase the potential of synthesized data where over 98% of data in their alignment process is synthesized. Nevertheless, we note the importance of identifying and investigating the limitations of LLM-generated/synthesized data (Dohmatob et al., 2024), presenting an opportunity for research.

4.2 Democratization of the LLM Technologies

The true testament to the impact of LLMs lies not in the streak of impressive metrics they score (OpenAI, 2023; Srivastava et al., 2023) but rather in the concrete real-life successes (Carbonell, 1992; Wagstaff, 2012, Impact Challenges). To do so requires the technology to be democratized and made accessible, not only through online API function calls but also in offline and resource-constrained environments, which is important to level the playing field for small organizations and individuals. We envision that the research directions of knowledge transfer can further widen the adoptions of LLMs (i) into different specialized domains including healthcare (Savova et al., 2010; Yang et al., 2023), law (Dahl et al., 2024), and education (Mind, 2024), (ii) at different scales, including consumer-grade hardware such as laptops (Hannun et al., 2023) and smart-phones (Sreeraman, 2023), and (iii) in different scenarios where internet accessibility, data security and privacy concerns can present obstacles to users making use of the API function calls online (Hao et al., 2022; Liu and Liu, 2023).

5 Inference Contextualization with Data

As in the two examples in Sec. 1 on how data can be used to contextualize the inference process of

LLMs (Brown et al., 2020; Lewis et al., 2020), LLMs have the demonstrated remarkable ability to utilize information "in-context" (Dong et al., 2023) where the context here is often in the form of a few example data points for demonstration or supplementary information (Brown et al., 2020). Such unique and unseen abilities present exciting use cases of data as an "anchor" to establish the right context at inference and enable the users to make certain specifications with flexibility and ease.

We illustrate such contextualization as follows: (i) If a user prompts the LLM to generate a piece of writing while providing writings from Shakespeare, then the LLM's generated output can appear "Shakespearean" even though the LLM is not necessarily (extensively) trained on the writings from Shakespeare. (ii) If a user asks the LLM to solve a mathematical question while providing data containing similar questions and the reasoning steps, then the generated output can also contain reasoning steps, even though the LLM might not have been explicitly trained to do so.

5.1 Data Selection for the Right Context

For two technical frameworks that enable an LLM to utilize data at inference, namely retrieval-augmented generation (RAG) and in-context learning (ICL), we outline how LLMs utilize the data and then describe the corresponding research directions of data selection for contextualization.

Retrieval-augmented generation. RAG consists of two main components: the datastore and the retriever. The datastore is a collection of unstructured data (e.g., documents and their chunks), and structured data (e.g., as databases or knowledge graphs). Given a user query, (i) the retriever selects the most relevant and informative data from the datastore to (ii) contextualize the query for the LLM to generate an output (Asai et al., 2024). These two steps can be targeted as follows.

For (i), a more effective data selection (i.e., better relevance and informativeness) can be achieved by improving the indexing system of the datastore. Currently, the data (e.g., documents) in the datastore each has an indexing "key" (typically a vector in some embedding space (Lewis et al., 2020; Salton et al., 1975) containing some of the data's semantic meaning). However, for a Q&A task, this indexing system can be ineffective for the retriever to identify the correct answer (i.e., data) to the question (i.e., query) since typically questions and answers have different semantic meanings. A future direction is to improve its effectiveness, such as by developing vector embeddings with built-in relevance in addition to the semantic meaning of data (Formal et al., 2021; Zamani et al., 2018) and pair them with the more classic approach of inverted index (Zobel and Moffat, 2006) based on keywords and metadata.

For (ii), there are promising avenues of improvements targeting the different ways of how LLMs contextualize the query (i.e., utilizing the retrieved data) such as by improving the augmentation of the query with the retrieved data (Shi et al., 2024; Yao et al., 2023a), and designing more effective retrieval methods and ways of utilizing the retrieved information (e.g., Borgeaud et al. (2022) use local similarity of consecutive document chunks to improve retrieval and model predictions, Wang et al. (2023a,b) demonstrate the effectiveness of pretraining LLMs/decoder-only LMs with retrieval).

In-context learning. From a few user-provided demonstrations (i.e., data) in the query alone, LLMs can learn the hidden patterns and respond accordingly (Dong et al., 2023). For example, to teach an LLM to solve mathematical questions, a user can query the LLM following this template:

`Your goal is to solve math problems. Here are some examples: [EXAMPLES]. Now solve [QUESTION].`

The demonstration data, denoted as `[EXAMPLES]`, establishes the right context for the LLM. Indeed the choice and quality of this demonstration data have a significant impact on the LLM's response quality (i.e., the correctness of the LLM's solution to `[QUESTION]`) (Lu et al., 2022b; Zhang et al., 2022). Existing methods have shown the effectiveness of heuristics, including similarity (Liu et al., 2022), uncertainty (Diao et al., 2023) and entropy (Lu et al., 2022b). These results suggest opportunities for integrated frameworks with provable guarantees as future directions. For instance, optimization-based techniques have achieved preliminary successes in instruction optimization of LLMs (e.g., reinforcement learning (Deng et al., 2022), Bayesian optimization (Lin et al., 2024) and evolutionary algorithms (Guo et al., 2024)), but have yet to be adopted for demonstration optimization or selection in ICL. Recently, Wu et al. (2024) utilize neural bandits for the joint optimization of instructions and demonstrations while Zhou et al. (2024) exploit the internal mechanism of transformers to optimize the selection of the demonstrations via their influences.

Note that RAG and ICL are not competing but rather complementary frameworks. With RAG, the user can leverage the size of the datastore for keeping more information while with ICL the user has an on-the-fly flexibility to direct specify the data with the query.

5.2 Personalized Usages of LLMs

Such contextualization (i.e., setting the context via specifying the data such as in RAG or ICL) has two hallmark practical benefits of being (i) simple and flexible via specifying the data and (ii) lightweight (i.e., no or minimal training/fine-tuning). For a user, the data need not be static. For instance, a company using a RAG-powered Q&A agent would, from time to time, update its product or service-related information. To ensure the Q&A agent has updated information, updating the datastore would suffice. In contrast, updating the LLM via either training or fine-tuning can be time-consuming, costly, and technically complex, so personalization approaches (e.g., via RAG or ICL) that minimize or sidestep updating the LLM are more appealing in practice.

Such features can simplify and make feasible the personalization of LLM technologies, which can have a significant impact on domains such as education (Alqahtani et al., 2023a; Gan et al., 2023; Latif et al., 2023) and healthcare (Abbasian et al., 2023; Belyaeva et al., 2023). LLMs-powered personalized curriculum designs can cater to the different needs of the students and educators can use LLMs to help prepare personalized feedback with significant time-saving benefits (Alqahtani et al., 2023b). LLMs-based chatbots can provide timely personalized health assessments (Cascella et al., 2024).

6 Conclusion and Future Outlook

This position paper has outlined a data-centric approach towards AI research with a focus on large language models (LLMs). We highlight the multifaceted role of data in the different developmental (e.g., pretraining, fine-tuning) and inferential (e.g., data synthesis, inference contextualization) stages of LLMs. In particular, we have identified four scenarios centered around data: rigorous data-centric benchmarks and data curation, data attribution, knowledge transfer, and inference contextualization with data. They each have unique challenges that require careful consideration, and present opportunities for innovation.

The impacts are described within each scenario for concreteness and clarity, but they are certainly *not* restricted to each of the scenarios and can sometimes "cross over". For instance, while we have identified democratization of the LLM technologies as an impact of Sec. 4, it is also applicable to Sec. 5, which has highlighted the practical viability of personalized usages of LLMs. Similarly, these scenarios (and the research directions therein) should not be viewed in isolation because there are indeed relationships and connections between the components. For instance, to mitigate problematic outputs by LLMs, a holistic treatment comprising both *ex-ante* and *ex-post* data-centric methods can perhaps be most effective (e.g., a more targeted data curation method from Sec. 2 paired with attribution and unlearning methods from Sec. 3).

This initial exploration into a data-centric AI research paradigm in the age of LLMs is necessarily non-exhaustive and intended to catalyze broader discussions, stimulate further inquiry, and spark innovation that will expand the current limits of LLMs and, more broadly, AI, and build toward deployment of such technologies that promote greater democratization.

7 Limitations and Impact Statement

This section organizes the limitations and alternative viewpoints following the same organization as the main paper.

On Sec. 2. One limitation of the outlined research directions is that these directions do not specifically account for the interplay between different steps (e.g., pretraining data and fine-tuning data) or between model (e.g., architecture and size) and data (Hoffmann et al., 2022; Sardana and Frankle, 2023). It is an appealing next step to develop integrated pipelines covering data curation methods for different steps and jointly leverage model-centric and data-centric insights.

On Sec. 3. One specific limitation of using fine-tuning to achieve unlearning is that its effectiveness is limited if there are only a small number of fine-tuning iterations due to a short stipulated execution time or a small fine-tuning dataset (Golatkar et al., 2020). As a result, after unlearning via fine-tuning, the model might still contain traces of the "deleted" target data. This limitation can be mitigated by adopting more directed unlearning techniques such as those described in Sec. 3.

We differentiate our described data-centric watermarking approaches (for data attribution) from existing model-oriented watermarking methods (Huang et al., 2024; Kuditipudi et al., 2023; Zhao et al., 2023a,b) (for determining whether a given output is generated by LLMs or a specific LLM). Additionally, we differentiate our described unlearning approaches (for removing or erasing certain target data) from knowledge unlearning (Si et al., 2023), whose goal is to forget an abstract definition of knowledge (Chen and Yang, 2023; Jang et al., 2023; Wang et al., 2023d).

On Sec. 4. A key requirement for effective knowledge transfer is that the general-purpose LLM has the "necessary" knowledge. This requirement is not always satisfied as there are areas where even the most advanced LLMs are lacking (e.g., reasoning and planning (Dziri et al., 2023; Valmeekam et al., 2023)). Nevertheless, there are many areas and use cases for which existing open-sourced LLMs are very capable (Groeneveld et al., 2024; NVIDIA, 2024) and can be used for knowledge transfer, and data synthesis in general. Furthermore, even if the LLM is not able to perform label synthesis optimally, it can still be useful for filtering out low-quality labels and leaving the good labels for training, as in "impossible distillation" (Jung et al., 2023). We note that our discussion on data synthesis has a specific focus on the quality of synthesized data w.r.t. the learning performance of ML models or LLMs and there are other important considerations not covered here due to page limits (e.g., safety considerations).

Another possible limitation in practice is due to the possible legal restrictions of how/whether existing proprietary and closed-source LLMs can be used, especially for commercial purposes.[2][3][4] Nevertheless, there are more efforts underway to open-source and democratize LLM technologies (Chiang et al., 2023; Liu et al., 2023b; Taori et al., 2023; Touvron et al., 2023). For instance, Groeneveld et al. (2024) completely open-sourced their LLM, including the pretraining data, model architecture, and trained weights, and the entire training logs, under the Apache-2.0 license, permitting a "free" use of this trained model, such as for knowledge transfer. As another example, NVIDIA (2024) released the Nemotron-4 family and their entire synthetic data generation pipeline under the NVIDIA Open Model License,[5] allowing the distribution, modification, and use of the models and its outputs.

On Sec. 5. One limitation of the inference contextualization is that it is difficult to design foolproof techniques or guarantees due to the complexity and the intricate black-box internal working mechanism of LLMs. It may require additional future investigation to understand and then leverage the mechanism of LLMs to design techniques with provable guarantees. Our position is to highlight a practically simple and technically viable approach for personalizing LLMs, as well as the promising research directions and techniques.

Impact Statement

This position paper presents a data-centric viewpoint towards AI research with a focus on LLMs, outlining specific scenarios for future research and highlighting the respective impacts therein.

Acknowledgement

This research is supported by the National Research Foundation Singapore and the Singapore Ministry of Digital Development and Innovation, National AI Group under the AI Visiting Professorship Programme (award number AIVP-2024-001). Xinyi Xu is supported by the Institute for Infocomm Research of Agency for Science, Technology and Research (A*STAR).

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv:2303.09540*.

Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2023. Conversational health agents: A personalized LLM-powered agent framework. *arXiv:2310.02374*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proc. NeurIPS*, pages 23716–23736.

Virginia Alexandria. 2024. Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot.

[2] Terms of use, OpenAI.
[3] Terms of Service, Anthropic.
[4] Generative AI APIs Additional Terms of Service, Google.
[5] NVIDIA Open Model License Agreement.

Tariq Alqahtani, Hisham A Badreldin, Mohammed Alrashed, Abdulrahman I Alshaya, Sahar S Alghamdi, Khalid Bin Saleh, Shuroug A Alowais, Omar A Alshaya, Ishrat Rahman, Majed S Al Yami, and Abdulkareem M Albekairy. 2023a. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, 19(8):1236—1242.

Tariq Alqahtani, Hisham A. Badreldin, Mohammed Alrashed, Abdulrahman I. Alshaya, Sahar S. Alghamdi, Khalid bin Saleh, Shuroug A. Alowais, Omar A. Alshaya, Ishrat Rahman, Majed S. Al Yami, and Abdulkareem M. Albekairy. 2023b. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, 19(8):1236–1242.

Mohammad Mohammadi Amiri, Frédéric Berdoz, and Ramesh Raskar. 2023. Fundamentals of task-agnostic data valuation. In *Proc. AAAI/IAAI/EAAI*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, et al. 2023a. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, et al. 2023b. PaLM 2 technical report. *arXiv:2305.10403*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv:2403.03187*.

Juhan Bae, Nathan Hoyen Ng, Alston Lo, Marzyeh Ghassemi, and Roger Baker Grosse. 2022. If influence functions are the answer, then what is the question? In *Proc. NeurIPS*.

Ramakrishna Bairi, Rishabh Iyer, Ganesh Ramakrishnan, and Jeff Bilmes. 2015. Summarization of multi-document topic hierarchies using submodular mixtures. In *Proc. ACL-IJCNLP*, pages 553–563.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative AI. *Foundations and Trends® in Privacy and Security*, 6(1):1–52.

Samyadeep Basu, Phil Pope, and Soheil Feizi. 2021. Influence functions in deep learning are fragile. In *Proc. ICLR*.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y. McLean, and Nicholas A. Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. *arXiv:2307.09018*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proc. ICML*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, et al. 2022. On the opportunities and risks of foundation models. *arXiv:2108.07258*.

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proc. ICML*, pages 2206–2240.

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. In *Proc. NeurIPS*, pages 1877–1901.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *Proc. SP*, pages 463–480.

Jaime Carbonell. 1992. Machine learning: A maturing field. *Machine Learning*, 9:5–7.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *arXiv:2302.10149*.

Marco Cascella, Federico Semeraro, Jonathan Montomoli, Valentina Bellini, Ornella Piazza, and Elena Bignami. 2024. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *Journal of Medical Systems*, 48(22).

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proc. EMNLP*, pages 12041–12052.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *arXiv:2403.12776*.

Lingjiao Chen, Bilge Acun, Newsha Ardalani, Yifan Sun, Feiyang Kang, Hanrui Lyu, Yongchan Kwon, Ruoxi Jia, Carole-Jean Wu, Matei Zaharia, and James Zou. 2023a. Data acquisition: A new frontier in data-centric AI. *arXiv:2311.13712*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv:2305.05176*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv:2107.03374*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proc. CVPR*, pages 2818–2829.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is your data worth to GPT? LLM-scale data valuation with influence functions. *arXiv:2405.13954*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*.

J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Hallucinating law: Legal mistakes with large language models are pervasive.

DeepSeek-AI. 2024. Deepseek LLM: Scaling open-source language models with longtermism. *arXiv:2401.02954*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proc. EMNLP*, pages 3369–3391.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv:2302.12246*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. *arXiv:2403.02990*.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: Model collapse as a change of scaling laws. In *Proc. ICML*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *arXiv:2301.00234*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Proc. NeurIPS*, volume 36, pages 70293–70332.

Ronen Eldan and Mark Russinovich. 2023. Who's Harry Potter? approximate unlearning in LLMs. *arXiv:2310.02238*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proc. SIGIR*, pages 2288–2292.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In search of the next generation of multimodal datasets. *arXiv:2304.14108*.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proc. ICML*, pages 1183–1192.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *Proc. IEEE BigData*.

Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable valuation of data for machine learning. In *Proc. ICML*, pages 2242–2251.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proc. CVPR*, pages 9301–9309.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. 2021. Amnesiac machine learning. In *Proc. AAAI*, pages 11516–11524.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, et al. 2024. OLMo: Accelerating the science of language models.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv:2308.03296*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *Proc. ICLR*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv:2306.11644*.

Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *Proc. EMNLP*, pages 10333–10350.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *Proc. ICLR*.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. 2023. Simfluence: Modeling the influence of individual training examples by simulating training runs. *arXiv:2303.08114*.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. MLX: Efficient and flexible machine learning on Apple silicon. Open-sourced software. Version 0.0.

Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. Iron: Private inference on transformers. In *Proc. NeurIPS*.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y. Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In *Proc. EMNLP*.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. 2023. Foundation models and fair use. *arXiv:2303.15715*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Proc. NeurIPS*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proc. ACL*, volume 1, pages 14409–14428.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv:2305.02301*.

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proc. ICML*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv:2404.06395*.

Baihe Huang, Banghua Zhu, Hanlin Zhu, Jason D. Lee, Jiantao Jiao, and Michael I. Jordan. 2024. Towards optimal statistical watermarking. *arXiv:2312.07930*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Understanding predictions with data and data with predictions. In *Proc. ICML*, pages 9525–9587.

IMDA. 2023. Singapore pioneers S$70m flagship AI initiative to develop Southeast Asia's first large language model ecosystem catering to the region's diverse culture and languages. *IMDA Press Releases*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proc. ACL*, pages 14389–14408.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. In *Proc. VLDB*, pages 1610–1623.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019b. Towards efficient data valuation based on the Shapley value. In *Proc. AISTATS*, volume 89, pages 1167–1176.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of proprietary large language models. In *Proc. EMNLP*, pages 3134–3154.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible Distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *arXiv:2305.16635*.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. 2023. LAVA: Data valuation without pre-specified learning algorithms. In *Proc. ICLR*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv:2001.08361*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv:1909.05858*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proc. ICML*, pages 17061–17084.

Will Knight. 2023. OpenAI's CEO says the age of giant AI models is already over. *WIRED*.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. 2019. On the accuracy of influence functions for measuring group effects. In *Proc. NeurIPS*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proc. ICML*, pages 1885–1894.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of EMNLP*, pages 4929–4952.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv:2307.15593*.

Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *Proc. ICLR*.

Ehsan Latif, Luyang Fang, Ping Ma, and Xiaoming Zhai. 2023. Knowledge distillation of LLM for education. *arXiv:2312.15842*.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2023a. Who wrote this code? watermarking for code generation. *arXiv:2305.15060*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023b. Holistic evaluation of text-to-image models. In *Proc. NeurIPS (Track on Datasets and Benchmarks)*.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*, pages 9459–9474.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proc. ACL*.

Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. 2022a. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In *Proc. NeurIPS*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need II: phi-1.5 technical report. *arXiv:2309.05463*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022b. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

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 Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad 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 S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew 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. *Transactions on Machine Learning Research*.

Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Use Your INSTINCT: INSTruction optimization usIng Neural bandits Coupled with Transformers. In *Proc. ICML*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proc. IJCNLP*, pages 6691–6706.

Hanxi Liu, Xiaokai Mao, Haocheng Xia, Jian Lou, and Jinfei Liu. 2023a. Prompt valuation based on Shapley values. *arXiv:2312.15395*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proc. NeurIPS*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Zhu Lei, and Michael Lingzhi Li. 2023c. Benchmarking large language models on CMExam - a comprehensive chinese medical exam dataset. In *Proc. NeurIPS*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023d. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

Xuanqi Liu and Zhuotao Liu. 2023. LLMs can understand encrypted prompt: Towards privacy-computing friendly transformers. *arXiv:2305.18396*.

Yixin Liu, Hongsheng Hu, Xun Chen, Xuyun Zhang, and Lichao Sun. 2023e. Watermarking classification dataset for copyright protection. *arXiv:2305.13257*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu. Open-source dataset.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022a. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. ACL*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022b. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. ACL*, pages 8086–8098.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv:2309.04564*.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2018. Do GANs Leave Artificial Fingerprints? In *Proc. MIPR*, pages 506–511.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Lilith Bat-Leah, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. DataPerf: Benchmarks for data-centric AI development. In *Proc. NeurIPS*. Track on Datasets and Benchmarks.

Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. 2022. Deep unlearning via randomized conditionally independent Hessians. In *Proc. CVPR*, pages 10422–10431.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2024. SILO language models: Isolating legal risk in a nonparametric datastore. In *Proc. ICLR*.

Merlyn Mind. 2024. First-ever education-specific language models open door to trustworthy generative AI for teachers and students.

Ali Naseh, Katherine Thai, Mohit Iyyer, and Amir Houmansadr. 2024. Iteratively prompting multimodal llms to reproduce natural and AI-generated images. *arXiv:2404.13784*.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proc. ALT*, pages 931–962.

Andrew Ng, Dillon Laird, and Lynn He. 2021. Data-centric AI competition.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In *Proc. NeurIPS*, pages 21455–21469.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. *arXiv:2306.11372*.

Ki Nohyun, Hoyong Choi, and Hye Won Chung. 2023. Data valuation without training of a model. In *Proc. ICLR*.

NVIDIA. 2024. Nemotron-4 340B technical report.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

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. In *Proc. NeurIPS*, pages 27730–27744.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv:2104.10350*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv:2310.07579*.

Guilherme Penedo, Hynek Kydlíček, Leandro von Werra, and Thomas Wolf. 2024. FineWeb. Open-source dataset.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *arXiv:2304.03277*. Work in progress.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Proc. NeurIPS*, pages 19920–19930.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, pages 5485–5551.

Noorjahan Rahman and Eduardo Santacana. 2023. Beyond fair use: Legal risk evaluation for training LLMs on copyrighted text. In *ICML 2023 Workshop on Generative AI and Law*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proc. ICML*, pages 8821–8831.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient LLMs. *arXiv:2402.09668*.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. ACL*, pages 1668–1678.

Nikhil Sardana and Jonathan Frankle. 2023. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. *arXiv:2401.00448*.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv:2211.05100*.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. 2023. Data selection for fine-tuning large language models using transferred Shapley values. In *Proc. ACL*, volume 4, pages 266–275. Student Research Workshop.

Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. 2022. Cs-Shapley: Class-wise Shapley values for data valuation in classification. In *Proc. NeurIPS*, pages 34574–34585.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proc. NeurIPS (Track on Datasets and Benchmarks)*, pages 25278–25294.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *Proc. ICLR*.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proc. AIES*, page 723–741.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. *arXiv:2310.10638*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for LLMs: Tasks, methods, and challenges. *arXiv:2311.15766*.

Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. 2022. Data valuation in machine learning:"ingredients", strategies, and open challenges. In *Proc. IJCAI*, pages 5607–5614.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv:2402.00159*.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Proc. NeurIPS*, pages 19523–19536.

Rupesh Sreeraman. 2023. Alpaca.cpp. Open-sourced software.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford Alpaca: An instruction-following LLAMA model.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. In *Proc. ICLR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 2023. LLAMA 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *arXiv:2310.16944*.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models - a critical investigation. In *Proc. NeurIPS*, volume 36, pages 75993–76005.

Kiri L. Wagstaff. 2012. Machine learning that matters. In *Proc. ICML*, page 1851–1856.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023a. InstructRetro: Instruction tuning post retrieval-augmented pretraining. *arXiv:2310.07713*.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023b. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. In *Proc. EMNLP*, pages 7763–7786.

Jingtan Wang, Xinyang Lu, Zitong Zhao, Zhongxiang Dai, Chuan-Sheng Foo, See-Kiong Ng, and Bryan Kian Hsiang Low. 2023c. WASA: Watermark-based source attribution for large language model-generated data. *arXiv:2310.00646*.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023d. KGA: A general machine unlearning framework based on knowledge gap alignment. In *Proc. ACL*, pages 13264–13276.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023e. Self-instruct: Aligning language models with self-generated instructions. In *Proc. ACL*, pages 13484–13508.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *In Proc. EMNLP*, pages 5085–5109.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *Proc. ICLR*.

Jiaheng Wei, Zuyue Fu, Yang Liu, Xingyu Li, Zhuoran Yang, and Zhaoran Wang. 2021. Sample elicitation. In *Proc. AISTATS*, pages 2692–2700.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proc. FAccT*, page 214–229.

Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Prompt optimization with EASE? efficient ordering-aware automated selection of exemplars. In *Proc. NeurIPS*.

Zhaoxuan Wu, Yao Shu, and Bryan Kian Hsiang Low. 2022. DAVINZ: Data valuation using deep neural networks at initialization. In *Proc. ICML*, pages 24150–24176.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. *arXiv:2402.04333*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023a. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Proc. NeurIPS*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023b. Data selection for language models via importance resampling. In *Proc. NeurIPS*.

Hanzi Xu and Slobodan Vuceti nad Wenpeng Yin. 2022. OpenStance: Real-world zero-shot stance detection. In *Proc. CoNLL*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv:2402.13116*.

Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2021. Validation free and replication robust volume-based data valuation. In *Proc. NeurIPS*, pages 10837–10848.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023a. React: Synergizing reasoning and acting in language models. In *Proc. ICLR*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023b. Large language model unlearning. *arXiv:2310.10683*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proc. EMNLP*, pages 7163–7189.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. 2020. Data valuation using reinforcement learning. In *Proc. ICML*, pages 10842–10851.

Ning Yu, Larry Davis, and Mario Fritz. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proc. ICCV*, pages 7555–7565.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proc. ICCV*, pages 14428–14437.

R. Yu, S. Liu, and X. Wang. 2024. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(01):150–170.

Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proc. CIKM*, pages 497–506.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv:2308.10792*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proc. EMNLP*, pages 9134–9148.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023a. Provable robust watermarking for AI-generated text. *arXiv:2306.17439*.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023b. Protecting language generation models via invisible watermarking. In *Proc. ICML*.

Zijian Zhou, Xiaoqiang Lin, Xinyi Xu, Alok Prakash, Daniela Rus, and Bryan Kian Hsiang Low. 2024. DETAIL: Task demonstration attribution for interpretable in-context learning. In *Proc. NeurIPS*.

Justin Zobel and Alistair Moffat. 2006. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6–es.