What is the Role of Small Models in the LLM Era: A Survey

Lihu Chen[1], Gaël Varoquaux[2]
[1] Imperial College London, UK
[2] Soda, Inria Saclay, France
`lihu.chen@imperial.ac.uk`
`gael.varoquaux@inria.fr`

Abstract

Large Language Models (LLMs) have made significant progress in advancing artificial general intelligence (AGI), leading to the development of increasingly large models such as GPT-4 and LLaMA-405B. However, scaling up model sizes results in exponentially higher computational costs and energy consumption, which makes these models impractical for academic researchers and businesses with limited resources. At the same time, Small Models (SMs) are frequently used in practical settings, although their significance is currently underestimated. This raises important questions about the role of small models in the era of LLMs, a topic that has received limited attention in prior research. In this work, we systematically examine the relationship between LLMs and SMs from two key perspectives: *Collaboration* and *Competition*. We hope this survey provides valuable insights for practitioners, fostering a deeper understanding of the contribution of small models and promoting more efficient use of computational resources. The code is available at ⬡ `https://github.com/tigerchen52/role_of_small_models`

1 Introduction

In recent years, the rapid advancement of Large Language Models (LLMs) has revolutionized natural language processing (NLP). Pre-trained language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have validated the *pre-train and fine-tune* paradigm, which involves learning general language representations through pre-training and subsequently transferring this knowledge to enhance performance on specific NLP tasks via fine-tuning (Min et al., 2023a). This approach has evolved into *prompt-based reasoning*, exemplified by the GPT family (Radford
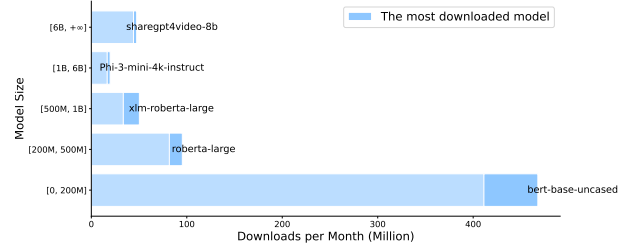

Figure 1: The relationship between model size and monthly downloads. This analysis considers open-source NLP models hosted on HuggingFace and categorizes them into five size groups based on the number of parameters: `[200M, 500M, 1B, 6B]`. The data was collected on August 25, 2024.

et al., 2019; Brown et al., 2020), where a few examples are provided in the prompt before the model performs the task on new inputs (Liu et al., 2023a).

These paradigms have demonstrated exceptional performance across a range of tasks, including language generation (Dong et al., 2023), language understanding (Wang et al., 2019), and domain-specific applications in areas such as coding (Jiang et al., 2024b), medicine (He et al., 2023), and law (Sun, 2023). Moreover, the theory of emergent abilities suggests that certain reasoning capabilities are enhanced by increasing model size, with some abilities only appearing in larger models (Wei et al., 2022a). This has led to a surge in the development of increasingly large models, such as GPT-4 (Achiam et al., 2023), Mixtral 8x22B (Jiang et al., 2024a), PaLM-340B (Anil et al., 2023), and LLaMA-405B (Dubey et al., 2024). As a result, LLMs have become highly prevalent, with data from March 2024 showing that ChatGPT (OpenAI, 2024) reached approximately 180 million users.

While LLMs have made significant strides in artificial general intelligence (AGI), their capabilities come with substantial overhead. Scaling model sizes leads to exponential increases in computational costs and energy consumption (Wan et al., 2023). Additionally, training and deploying LLMs is often unfeasible for academic researchers and businesses with limited resources. As a result, there has been a shift toward smaller language models (SLMs) such as Phi-3.8B (Abdin et al., 2024) and Gemma-2B (Team et al., 2024), which can achieve comparable performance with significantly fewer parameters.

Some may argue that models like Phi-3.8B and Gemma-2B do not qualify as true small models, and that genuinely small models, such as BERT, are no longer prominent. However, our findings suggest that the usage of small models is significantly underestimated in practical settings. As illustrated in Figure 1, we analyzed the number of downloads of open-source models of various sizes from HuggingFace. The results show that smaller models, particularly BERT-base, remain highly popular. This raises important questions about the role of small models in the era of LLMs, a topic that has been largely overlooked in prior research.

As counterparts to LLMs, *Small Models* (SMs) generally refer to models with a relatively lower number of parameters, including not only language models but also simple statistical models and shallow neural networks. However, there is no clear definition distinguishing large models from small ones. In this work, we consider model size in relative terms. For instance, BERT (110M parameters) (Devlin et al., 2019) is considered small compared to LLaMA-8B (Dubey et al., 2024), while LLaMA-8B is small relative to GPT-4 (175B parameters) (Achiam et al., 2023). This relative definition allows for flexibility and ensures that concepts remain relevant as larger models are developed in the future.

To assess the role of SMs, it is essential to compare their strengths and weaknesses relative to LLMs. Table 1 highlights four key dimensions to consider:

Accuracy. LLMs have demonstrated superior performance across a wide range of NLP tasks due to their large number of parameters and ex-

Dimension	LLMs	SMs
Accuracy	state-of-the-art ✓	decent ✗
Generality	general-purpose ✓	task-specific ✗
Efficiency	resource-intensive ✗	resource-efficient ✓
Interpretability	low interpretable ✗	high interpretable ✓

Table 1: Comparisons of different dimensions between LLMs and SMs.

tensive training on diverse datasets (Raffel et al., 2020; Kaplan et al., 2020). Although SMs generally lag behind in overall performance, they can achieve comparable results when enhanced by techniques such as knowledge distillation (Xu et al., 2024a).

Generality. LLMs are highly generalizable, and capable of handling a broad spectrum of tasks with minimal training examples (Dong et al., 2023; Liu et al., 2023a). In contrast, SMs are often more specialized, and research shows that fine-tuning SMs on domain-specific datasets can sometimes outperform general LLMs for specific tasks (Hernandez et al., 2023; Juan José Bucher and Martini, 2024; Zhang et al., 2023a).

Efficiency. LLMs require substantial computational resources for both training and inference (Wan et al., 2023), leading to high costs and latency, making them less practical for real-time applications (e.g. information retrieval (Reimers and Gurevych, 2019)) or resource-constrained environments (e.g. edge devices (Dhar et al., 2024)). In contrast, SMs require less training data and computational power, offering competitive performance while significantly reducing resource demands.

Interpretability. Smaller, shallower models tend to be more transparent and interpretable than their larger, deeper counterparts (Gilpin et al., 2018; Barceló et al., 2020). In fields such as healthcare (Caruana et al., 2015), finance (Kurshan et al., 2021), and law (Eliot, 2021), smaller models are often preferred because their decisions must be easily understandable by non-experts (e.g. doctors, financial analysts).

In this work, we systematically examine the role of small models in the era of LLMs from two key perspectives: (1) ***Collaboration*** (§ 2). LLMs offer superior accuracy and can handle a wide range of tasks, while SMs are more spe-

cialized and cost-effective. In practice, the collaboration between LLMs and SMs can strike a balance between power and efficiency, enabling systems that are resource-efficient, scalable, interpretable, and cost-effective, while maintaining high performance and flexibility. (2) *Competition* (§ 3). SMs possess distinct advantages, such as simplicity, lower cost, and greater interpretability, and they have niche market. It is crucial to carefully assess the trade-offs between LLMs and SMs, depending on the specific requirements of the task or application.

2 Collaboration

In the following, we present how SMs and LLMs can collaborate to optimize resource usage: SMs enhance LLMs (§ 2.1) and LLMs enhance SMs (§ 2.2). The overall collaboration framework is listed in Figure 2.

2.1 Small Models Enhance LLMs

2.1.1 Data Curation

In the following, we present how to use small models to curate data from two aspects: pre-training data and instruction-tuning data.

Curating Pre-training Data The reasoning capabilities of LLMs are largely attributed to their pre-training on extensive and diverse datasets, typically sourced from web scrapes, books, and scientific literature. Since expanding the quantity and diversity of these training datasets enhances the generalization ability of LLMs, significant efforts have been made to compile large-scale and diverse pre-training corpora, such as C4 (Raffel et al., 2020) and Pile (Gao et al., 2021). Furthermore, the scaling law (Kaplan et al., 2020) has demonstrated that model performance is highly dependent on both the scale of model parameters and the size of the training dataset. This may suggest that, to develop more powerful models, one should aim to use as much data as possible for pre-training.

However, this approach faces a significant challenge: data availability is finite, and there is a looming possibility that public human text data could soon be exhausted (Villalobos et al., 2024).

Moreover, not all data contributes equally to model performance; web-scraped content often includes noise and low-quality text. This has led to a paradigm shift from focusing purely on the quantity of data to prioritizing the quality of data. Recent research supports the notion that "less is more" (Marion et al., 2023), advocating for data selection or pruning techniques to curate high-quality subsets from large datasets, thereby enhancing model performance (Albalak et al., 2024).

Prior approaches to data curation often rely on rule-based heuristics, such as blacklist filtering and MinHash deduplication (Raffel et al., 2020; Tirumala et al., 2024; Penedo et al., 2023; Wenzek et al., 2020). However, these manual, rule-based methods are increasingly inadequate given the scale and complexity of raw text data. A common alternative is to employ a small model trained specifically to evaluate text quality, enabling the selection of high-quality subsets. For instance, a simple classifier can be trained to assess content quality, focusing on the removal of noisy, toxic, and private data (Brown et al., 2020; Du et al., 2022; Chowdhery et al., 2023; Xie et al., 2023). Another technique involves using perplexity scores calculated by a proxy language model to select data that is more likely to be of high quality (Wenzek et al., 2020; Marion et al., 2023).

Beyond data selection, data reweighting is a strategy that assigns domain-specific weights, effectively adjusting the sampling probabilities for different text sources. This can be achieved by training a small proxy model to set these domain weights, which in turn can enhance the generalization capabilities of the pre-trained model across various domains (Xie et al., 2024).

Curating Instruction-tuning Data LLMs acquire substantial knowledge through pre-training, and instruction tuning aims to align these model capabilities with human preferences (Ouyang et al., 2022; Bai et al., 2022a). While earlier research has concentrated on tuning LLMs using large-scale instruction datasets, recent findings suggest that strong alignment can be achieved with much smaller datasets. Specifically, the study, *Less is More for Alignment*, demonstrates that fine-tuning on just 1,000 carefully curated instruction examples can yield a well-aligned

3

Collaboration (§ 2)

- **SMs Enhance LLMs (§ 2.1)**
  - **Data Curation (§ 2.1.1)**
    - Curating pre-training data: Data Selection (Xie et al., 2023; Marion et al., 2023); Data Reweighting (Xie et al., 2024)
    - Curating Instruction-tuning Data (Du et al., 2023; Xia et al., 2024b)
  - **Weak-to-Strong Paradigm (§ 2.1.2)**
    - Using weaker (smaller) models to align stronger (larger) models (Burns et al., 2024; Zhou et al., 2024b; Liu and Alahi, 2024; Guo and Yang, 2024; Ji et al., 2024; Guo et al., 2024)
  - **Efficient Inference (§ 2.1.3)**
    - Ensembling different-size models to reduce inference costs: Model Cascading (Kag and Fedorov, 2023; Chen et al., 2023d, 2024a; Madaan et al., 2023); Model Routing (Lu et al., 2023a; Narayanan Hari and Thomson, 2023; Lee et al., 2024a; Ong et al., 2024; Šakota et al., 2024)
    - Speculative Decoding (Leviathan et al., 2023; Xia et al., 2024a; Chen et al., 2023a)
  - **Evaluating LLMs (§ 2.1.4)**
    - Using SMs to evaluate LLM's generations: General Evaluation (Zhang et al., 2020; Yuan et al., 2021); Uncertainty (Kuhn et al., 2023; Manakul et al., 2023); Performance Prediction (Anugraha et al., 2024)
  - **Domain Adaptation (§ 2.1.5)**
    - Using domain-specific SMs to generate knowledge for LLMs at reasoning time (Feng et al., 2024; Li et al., 2024a)
    - Using domain-specific SMs to adjust token probability of LLMs at decoding time (Ormazabal et al., 2023; Lu et al., 2023b; Liu et al., 2024a)
  - **Retrieval Augmented Generation (§ 2.1.6)**
    - Using SMs to retrieve knowledge for enhancing generations: Documents (Lewis et al., 2020b); Knowledge Bases (Wang et al., 2023); Tables (Pan et al., 2022); Codes (Zhou et al., 2023); Tools (Schick et al., 2024); Images (Yasunaga et al., 2023)
  - **Prompt-based Reasoning (§ 2.1.7)**
    - Using SMs to augment prompts for LLMs: Retrieving Prompts (Cheng et al., 2023); Decomposing Complex Problems (Juneja et al., 2023); Generating Pseudo Labels (Xu et al., 2023; Lee et al., 2024b) and Feedback (Hsu et al., 2024; Vernikos et al., 2024)
  - **Deficiency Repair (§ 2.1.8)**
    - Developing SM plugins to repair deficiencies: Hallucinations (Cheng et al., 2024; Chen et al., 2024b); Out-Of-Vocabulary Words (Chen et al., 2022a)
    - Contrasting LLMs and SMs for better generations: Reducing Repeated Texts (Li et al., 2023b); Mitigating Hallucinations (Zhang et al., 2023b); Augmenting Reasoning Capabilities (O'Brien and Lewis, 2023); Safeguarding Privacy (Zhang et al., 2024a)

- **LLMs Enhance SMs (§ 2.2)**
  - **Knowledge Distillation (§ 2.2.1)**
    - Black-box Distillation: Chain-Of-Thought Distillation (Li et al., 2022; Hsieh et al., 2023; Shridhar et al., 2023; Magister et al., 2023; Li et al., 2023a; Fu et al., 2023; Tian et al., 2024); Instruction Following Distillation (Jiang et al., 2023c; Li et al., 2024b)
    - White-box distillation: Logits (Sanh, 2019); Intermediate Features (Yao et al., 2022; Liang et al., 2023; Gu et al., 2024; Liu et al., 2023b)
  - **Data Synthesis (§ 2.2.2)**
    - Data Augmentation: Text Paraphrase (Mi et al., 2022; Witteveen and Andrews, 2019); Query Rewriting (Ma et al., 2023a); Specific Tasks (Hu et al., 2024a; Sahu et al., 2022; Chen et al., 2022b)
    - Training Data Generation: Label Annotation (Wang et al., 2021; Gao et al., 2022); Dataset Generation (Ye et al., 2022; Meng et al., 2022; Chung et al., 2023; Li et al., 2023c; Tang et al., 2023; Josifoski et al., 2023; Hartvigsen et al., 2022)
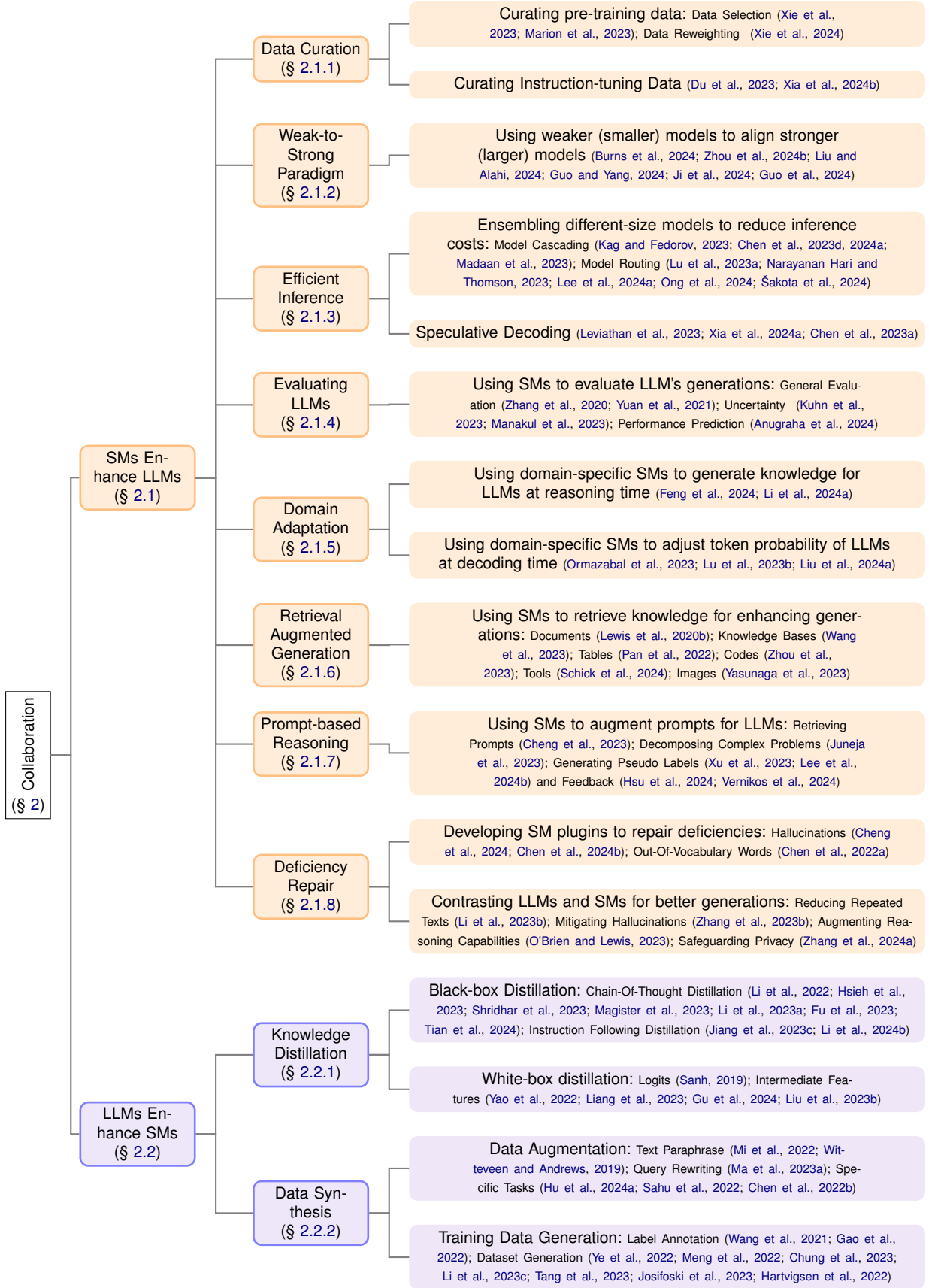
Figure 2: Collaborations between LLMs and SMs

model (Zhou et al., 2024a). This highlights the importance of selecting high-quality data for efficient instruction tuning (Longpre et al., 2023; Chen et al., 2023c).

Model-oriented data selection (MoDS) (Du et al., 2023) is one approach that employs a small language model, DeBERTa (He et al., 2021), to evaluate instruction data based on quality, coverage, and necessity. Additionally, the LESS framework (Xia et al., 2024b) demonstrates that smaller models can be used to select influential data not only for larger models but also for models from different families. This underscores the potential of using targeted data selection techniques to optimize instruction tuning processes.

Summary and Future Directions Given the impending limitations on the amount of data we humans can create, it is crucial to focus on curating existing data and embracing the principle that "less is more". In this section, we explore how small models can play a significant role in data selection and reweighting, both during pre-training and fine-tuning processes.

> **Future Directions**
>
> (,) While data curation offers clear advantages, LLMs still have a tendency to produce hallucinated and toxic content. Moreover, removing low-quality or toxic text could potentially degrade certain capabilities, such as generality (Longpre et al., 2024). Therefore, it is essential to define more nuanced criteria for evaluating data quality, including dimensions like factuality, safety, and diversity (Wettig et al., 2024; Liu et al., 2024b). Investigating the use of small models to develop effective and efficient data selection methods is a valuable area of study.
>
> Synthetic data serves as a valuable supplement to the limited amount of human-generated data (Long et al., 2024), yet the potential of small models in curating synthetic data remains largely unexplored.

2.1.2 Weak-to-Strong Paradigm

LLMs are typically aligned with human values through reinforcement learning with human feedback (RLHF), where behaviors favored by humans are rewarded, and those rated poorly are pe-

nalized (Shen et al., 2023). However, as LLMs continue to evolve and surpass human capabilities in various tasks, they are becoming *superhuman models*, capable of performing complex and creative tasks that may exceed human understanding. For instance, these models can generate thousands of lines of specialized code, engage in intricate mathematical reasoning, and produce lengthy, creative novels. Evaluating the correctness and safety of such outputs poses significant challenges for human evaluators. This scenario introduces a new paradigm for aligning superhuman models, termed *weak-to-strong generalization*, which involves using weaker (smaller) models as supervisors for stronger (larger) models (Burns et al., 2024). In this approach, large, powerful models are fine-tuned on labels generated by smaller, less capable models, enabling the strong models to generalize beyond the limitations of their weaker supervisors.

Building on the concept of weak-to-strong generalization, several variants have recently been proposed. For example, Liu and Alahi (2024) suggests using a diverse set of specialized weak teachers, rather than relying on a single generalist model, to collectively supervise the strong student model. Guo and Yang (2024) introduce an approach that enhances weak-to-strong generalization by incorporating reliability estimation across multiple answers provided by weak models. This method improves the alignment process by filtering out uncertain data or adjusting the weight of reliable data. Beyond data labeling, weak models can also collaborate with large models during the inference phase to further enhance alignment. Aligner (Ji et al., 2024) employs a small model to learn the correctional residuals between preferred and dispreferred responses, enabling direct application to various upstream LLMs for aligning with human preferences. Weak-to-Strong Search (Zhou et al., 2024b) approaches the alignment of a large model as a test-time greedy search, aiming to maximize the log-likelihood difference between small tuned and untuned models, which function as a dense reward signal and a critic, respectively. This weak-to-strong paradigm is not limited to language models but has also been extended to vision foundation models (Guo et al., 2024).

Summary and Future Directions As large models continue to evolve rapidly, we are ap-

proaching a future where superhuman models will emerge, which makes effective human supervision increasingly challenging. The Weak-to-Strong paradigm demonstrates that weak supervisors can be used to draw out knowledge from strong models, enabling the development of superhuman reward models that ensure safe and reliable alignment.

> **Future Directions**
>
> (1) While the weak-to-strong framework is effective in eliciting knowledge from stronger models, it is still far from recovering the full performance gap between weak and strong models. It is crucial to ensure that the strong model has a deep, intuitive understanding of the task at hand, is capable of correcting the weak model's errors, and naturally aligns with the objectives of the task (Burns et al., 2024). Future work should focus on identifying properties and methods that help achieve this goal.
> (2) The current understanding of weak-to-strong generalization is limited. Researchers should develop a deep understanding of the underlying mechanisms that govern the success or failure of alignment methods, e.g. theoretical analysis (Lang et al., 2024), errors in weak supervision (Guo and Yang, 2024), and extrapolating generalization errors using scaling laws (Kaplan et al., 2020).

2.1.3 Efficient Inference

Model Ensembling Larger models are generally more powerful but come with significant costs, including slower inference speed and more expensive prices (APIs). Beyond the financial costs, the use of larger models also has a considerable environmental and energy impact (Wu et al., 2022). In contrast, smaller models, while potentially less performant, offer advantages in terms of lower cost and faster inference. Given that user queries vary widely in complexity—ranging from simple questions that smaller models can handle to more complex ones requiring larger models—it is possible to achieve cost-effective inference by leveraging an ensemble of models of different sizes. This approach to model ensembling

can be divided into two categories: model cascading and model routing.

Model Cascading involves the sequential use of multiple models to make predictions or decisions, where each model in the cascade has a different level of complexity. The output of one model may trigger the activation of the next model in the sequence (Varshney and Baral, 2022; Viola and Jones, 2004; Wang et al., 2011). This approach allows for the collaboration of models of varying sizes, enabling smaller models to handle simpler input queries while transferring more complex tasks to larger models. The critical step in this process is determining whether a given model is capable of addressing the input question. This method effectively optimizes inference speed and reduces financial costs.

Some existing techniques train a small evaluator to assess the correctness (Kag and Fedorov, 2023; Chen et al., 2020, 2023d), confidence (Chen et al., 2024a), or quality (Ding et al., 2024b) of a model's output, thereby deciding whether to escalate the query to a more complex model. Given that LLMs can perform self-verification (Dhuliawala et al., 2023) and provide confidence levels in their responses (Tian et al., 2023), AutoMix (Madaan et al., 2023) employs verification prompts to query the model multiple times, using the consistency of these responses as an estimated confidence score. The framework then determines whether the current model's output should be accepted or if the query should be forwarded to other models for enhanced performance.

Model Routing optimizes the deployment of multiple models of varying sizes by dynamically directing input data to the most appropriate models, thereby enhancing both efficiency and effectiveness in practical applications. The core component of this approach is the development of a router that assigns input to one or more suitable models within the pool.

A straightforward approach is to consider the input-output pairs from all models and select the best-performing one (Jiang et al., 2023a). However, this comprehensive ensemble strategy does not significantly reduce inference costs. To address this, some methods train efficient, reward-based routers that select optimal models without needing to access the models' outputs (Lu

et al., 2023a; Narayanan Hari and Thomson, 2023). OrchestraLLM (Lee et al., 2024a) introduces a retrieval-based dynamic router that assumes instances with similar semantic embeddings share the same difficulty level. This allows for the selection of an appropriate expert based on the embedding distances between the testing instance and those in expert pools. Similarly, RouteLLM (Ong et al., 2024) leverages human preference data and data augmentation to train a small router model, which effectively reduces inference costs and enhances out-of-domain generalization. FORC (Šakota et al., 2024) proposes a meta-model (a regression model) to assign queries to the most suitable model without requiring the execution of any large models during the process. The meta-model is trained on existing pairs of queries and model performance scores. Furthermore, recent benchmarks for model routing have been established (Hu et al., 2024b; Shnitzer et al., 2023), facilitating more accessible and cost-effective deployments of large language models.

Speculative Decoding   This technique aims to speed up the decoding process of a generative model, which often involves using a smaller, faster auxiliary model alongside the main, larger model. The auxiliary model quickly generates multiple token candidates in parallel, which are then validated or refined by the larger, more accurate model. This approach allows for faster initial predictions that are subsequently verified by the more computationally intensive model (Leviathan et al., 2023; Xia et al., 2024a; Chen et al., 2023a).

Summary and Future Directions   The inference costs associated with large models or APIs can be substantial, but the collaboration of heterogeneous models can effectively reduce these monetary expenses and speed up inference. In this section, we introduce model ensembling and speculative decoding as strategies to optimize the inference process.

> **Future Directions**
>
> (1) Existing ensembling methods typically rely on a limited, pre-defined list of models, yet the real world encompasses open-domain and constantly evolving LLMs, such as those available on Hug-

gingFace. Exploring ways to leverage these extensive model libraries to create intelligent and efficient systems holds significant promise (Shen et al., 2024).

(2) In the current approach to speculative decoding, the auxiliary model is typically constrained to be from the same model family as the main model, such as different sizes of GPT. However, it is advantageous to explore collaborations between models from diverse sources.

2.1.4 Evaluating LLMs

Effectively evaluating open-ended text generated by LLMs presents a significant challenge across various NLP tasks (Chang et al., 2024). Traditional evaluation methods, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which focus on surface-form similarity, often fall short in capturing the nuanced semantic meaning and compositional diversity of the generated text (Liu et al., 2016).

To address these limitations, model-based evaluation approaches use smaller models to automatically assess performance. For instance, BERTSCORE (Zhang et al., 2020) employs BERT to compute semantic similarity for evaluating machine translation and image captioning. Similarly, BARTSCORE (Yuan et al., 2021) leverages the encoder-decoder model BART (Lewis et al., 2020a) to evaluate texts from various perspectives, including informativeness, fluency, and factuality. In addition to general text evaluation, some methods use small natural language inference (NLI) models to estimate the uncertainty of LLM responses (Manakul et al., 2023; Kuhn et al., 2023). Another valuable application involves using proxy models to predict LLM performance (Anugraha et al., 2024), which substantially reduces the computational costs associated with fine-tuning and inference during model selection.

Summary and Future Directions   LLMs are capable of generating extensive and complex text, which makes it hard to evaluate the content based on simple heuristic metrics. To overcome this challenge, a small proxy model can be employed to automate the evaluation of generated text from multiple perspectives, including aspects such as factuality and fluency.

Future Directions

As large models advance, they increasingly generate lengthy and complex texts, such as specialized code and scientific papers, which are challenging for humans to evaluate. Consequently, it is essential to develop efficient evaluators to assess various aspects of the generated content, such as factuality (Min et al., 2023b), safety (Zhang et al., 2024b), and uncertainty (Huang et al., 2023).

2.1.5 Domain Adaptation

Despite their growing capabilities, general-purpose large language models (LLMs) still require further customization to achieve optimal performance in specific use cases (*e.g.* coding) and domains (*e.g.* medical tasks). While fine-tuning on specialized data is one approach to adapting LLMs, this process has become increasingly resource-intensive, and in some cases, it is not feasible—especially when access to internal model parameters is restricted, as with models like ChatGPT. Recent research has explored adapting LLMs using smaller models, which can be categorized into two approaches: White-Box Adaptation and Black-Box Adaptation, depending on whether access to the model's internal states is available.

White-Box Adaptation typically involves fine-tuning a small model to adjust the token distributions of frozen LLMs for a specific target domain. For instance, CombLM (Ormazabal et al., 2023) learns a linear function to combine the probability distributions from the large black-box model with those from a smaller domain-specific expert model. IPA (Lu et al., 2023b) introduces a lightweight adapter that tailors a large model toward desired objectives during decoding without requiring fine-tuning. IPA achieves this by optimizing the combined distribution using reinforcement learning. Proxy-tuning (Liu et al., 2024a) fine-tunes a smaller language model, contrasting the probabilities between the tuned model (the expert) and its untuned version (the anti-expert) to guide the larger base model.

These approaches only modify the parameters of small domain-specific experts, allowing LLMs to be adapted to specific domain tasks. However, white-box adaptation is not applicable to API-only modeling services, where access to internal model parameters is restricted.

Black-Box Adaptation involves using a small domain-specific model to guide LLMs toward a target domain by providing textual relevant knowledge. Retrieval Augmented Generation (RAG) can extract query-relevant knowledge from an external document collection or knowledge base, and thus enhance general LLMs by leveraging their in-context learning ability. It involves first using a lightweight retriever to find relevant content from the domain corpus, which is then incorporated into the LLM's input to improve its understanding of domain-specific knowledge (Siriwardhana et al., 2023; Shi et al., 2023; Gao et al., 2023).

Another approach employs small expert models to retrieve background knowledge for the base LLM in a generative manner. For example, BLADE (Li et al., 2024a) and Knowledge Card (Feng et al., 2024) first pre-train a small expert model on domain-specific data, which then generates expertise knowledge in response to a query, thereby enhancing the base LLM's performance.

Summary and Future Directions Tuning large models for specific target domains is resource-intensive. To address this challenge, a more efficient approach is to fine-tune a small model on domain-specific data. This lightweight expert model can then guide the LLM either during decoding (white-box adaptation) or inference (black-box adaptation), offering a cost-effective solution for domain adaptation.

Future Directions

In white-box adaptation, most methods require that the small models and base models belong to the same family, such as the GPT family. To enhance domain adaptation, it is important to develop techniques that leverage a broader range of diverse models (Kasai et al., 2022; Xu et al., 2024b; Remy et al., 2024).

Current methods often necessitate pre-training a domain-specific expert from scratch, which is impractical for resource-constrained tasks. Investigating how to adapt LLMs using a limited number of samples is a valuable area of research (Sun

et al., 2024).

2.1.6 Retrieval Augmented Generation

LLMs exhibit impressive reasoning capabilities, yet their ability to memorize specific knowledge is somewhat limited. Consequently, LLMs may struggle with tasks that require domain-specific expertise or up-to-date information. To address these limitations, Retrieval-Augmented Generation (RAG) enhances LLMs by employing a lightweight retriever to extract relevant document fragments from external knowledge bases, document collections, or other tools (Gao et al., 2023; Lewis et al., 2020b). By incorporating external knowledge, RAG effectively mitigates the issue of generating factually inaccurate content, often referred to as hallucinations (Shuster et al., 2021). RAG methods can be broadly categorized into three types based on the nature of the retrieval source.

Textual Document are the most commonly used retrieval sources in RAG methods, encompassing resources such as Wikipedia (Trivedi et al., 2023; Asai et al., 2023), cross-lingual text (Nie et al., 2023) and domain-specific corpus (e.g. medical (Xiong et al., 2024) and legal (Yue et al., 2023) domains). These approaches generally employ lightweight retrieval models, such as sparse BM25 (Robertson et al., 2009) and dense BERT-based (Izacard et al., 2021) retrievers, to extract relevant text from these sources.

Structured knowledge encompasses sources such as knowledge bases and databases, which are typically verified and can provide more precise information. For example, KnowledgeGPT (Wang et al., 2023) enables LLMs to retrieve information from knowledge bases, while T-RAG (Pan et al., 2022) enhances answers by concatenating retrieved tables with the query. StructGPT (Jiang et al., 2023b) further augments generation by retrieving from hybrid sources, including knowledge bases, tables, and databases. The retriever in these methods can be a lightweight entity linker, query executor, or API.

Other Sources include codes, tools, and even images, which enable LLMs to leverage external information for enhanced reasoning. For instance, DocPrompting (Zhou et al., 2023) employs a BM25 retriever to obtain relevant code

documentation before code generation. Similarly, Toolformer (Schick et al., 2024) demonstrates that LMs can self-learn to use external tools, such as translators, calculators, and calendars, through simple APIs, leading to significant performance improvements.

Summary and Future Directions Retrieval-augmented generation significantly extends the knowledge boundaries of LLMs, with small models primarily serving as retrievers in this process. By employing a lightweight retriever, various types of information—such as documents, structured knowledge, code, and useful tools—can be efficiently accessed to enhance the model's capabilities.

> Future Directions
>
> (1) The retrieval augmented text generation performance is very sensitive to the retrieval quality (Yoran et al., 2023). Therefore, developing robust approaches to integrate noisy retrieved texts is important.
> (2) RAG can be extended to multimodal scenarios beyond text-only information, such as images (Yasunaga et al., 2023), audios (Zhao et al., 2023), etc.

2.1.7 Prompt-based Learning

Prompt-based learning is a prevalent paradigm in LLMs where prompts are crafted to facilitate few-shot or even zero-shot learning, enabling adaptation to new scenarios with minimal or no labeled data (Liu et al., 2023a). This approach leverages the power of In-Context Learning (ICL) (Dong et al., 2023), which operates without performing parameter updates. Instead, it relies on a prompt context that includes a few demonstration examples structured within natural language templates.

In this learning process, small models can be employed to enhance prompts, thereby improving the performance of larger models. For instance, Uprise (Cheng et al., 2023) optimizes a lightweight retriever that autonomously retrieves prompts for zero-shot tasks, thereby minimizing the manual effort required for prompt engineering. Similarly, DaSLaM (Juneja et al., 2023) uses a small model to break down complex problems into subproblems that necessitate fewer reasoning steps, leading to performance improvements in

larger models across multiple reasoning datasets. Other methods involve fine-tuning small models to generate pseudo labels for inputs (Xu et al., 2023; Lee et al., 2024b), which results in better performance than the original ICL. Additionally, small models can be used to verify (Hsu et al., 2024) or rewrite (Vernikos et al., 2024) the outputs of LLMs, thereby achieving performance gains without the need for fine-tuning.

Summary and Future Directions Prompt-based learning is capable of handling a variety of complex tasks by using a few examples embedded within a prompt template. To further enhance this process, small models can be employed to augment prompts by reformulating questions and generating feedback. This efficient augmentation allows for the improvement of LLMs without the need for parameter updates.

> **Future Directions**
>
> Recent research has concentrated on leveraging small models to enhance the reasoning capabilities of large models within the prompt-based learning paradigm. It is also important to explore how small models can be used to develop LLMs that are trustworthy, safe, and fair.

2.1.8 Deficiency Repair

Powerful LLMs may generate repeated, untruthful, and toxic contents, and small models can be used to repair these defects. We introduce two ways to achieve this goal: contrastive decoding and small model plug-ins.

Contrastive Decoding exploits the contrasts between a larger model (expert) and a smaller model (amateur) by choosing tokens that maximize their log-likelihood difference. Existing work has explored the synergistic use of logits from both LLMs and SMs to reduce repeated text (Li et al., 2023b), mitigate hallucinations (Sennrich et al., 2024), augment reasoning capabilities (O'Brien and Lewis, 2023) and safeguarding user privacy (Zhang et al., 2024a). Since fine-tuning LLMs is computing-intensive, proxy tuning proposes fine-tuning a small model and contrasting the difference between the original LLMs and small models to adapt to the target task (Liu et al., 2024a).

Small Model Plugins fine-tune a specialized small model to address some of the shortcomings of the larger model. For example, the performance of LLMs may degrade when encountering unseen words (Out-Of-Vocabulary) words. To address this issue, we can train a small model to mimic the behavior of the large model and impute representations for unseen words (Pinter et al., 2017; Chen et al., 2022a). Through this way, we can make large models robust with little cost. Additionally, LLMs may generate hallucinated texts and we can train a small model to detect hallucinations (Cheng et al., 2024) or calibrate confidence scores (Chen et al., 2024b).

Summary and Future Directions Although language models are very powerful, they have their own weaknesses that need to be addressed, such as hallucinations, toxicity, etc. Here we present the contrastive decoding and developing small model plugins to make LLMs more robust and secure.

> **Future Directions**
>
> We can extend the pattern of using small models to fix flaws of large models to other problems. For example, the mathematical reasoning of LLMs is so fragile that they can collapse in the face of basic math problems, e.g. ChatGPT thinks that 9.11 is bigger than 9.9 [a].
>
> ---
> [a] https://x.com/goodside/status/1812977352085020680

2.2 LLMs Enhance Small Models

2.2.1 Knowledge Distillation

Scaling models to larger sizes is a straightforward method to enhance performance, but it often proves too computationally expensive for widespread deployment to numerous users. To mitigate this challenge, Knowledge Distillation (KD) (Hinton, 2015; Gou et al., 2021; Zhu et al., 2023; Xu et al., 2024a) offers an effective solution. In KD, a smaller student model is trained to replicate the behavior of a larger teacher model. Typically, this process involves the larger model generating a dataset with pseudo labels, which the smaller model then uses for training.

White-box distillation involves using internal

states of the teacher model which provides transparency in the training process of the student model. This approach leverages the output distributions, and intermediate features from the teacher LLMs, collectively referred to as feature knowledge (Liang et al., 2023; Gu et al., 2024; Liu et al., 2023b). It enables the development of cost-effective yet powerful models, exemplified by DistilBERT (Sanh, 2019) and QuantizedGPT (Yao et al., 2022).

In contrast, black-box knowledge distillation typically involves generating a distillation dataset through the teacher LLM, which is then used for fine-tuning the student model. For instance, Chain-of-Thought distillation (Wei et al., 2022b) extracts LLM rationales to provide additional supervision, thereby enhancing the reasoning capabilities of smaller models (Li et al., 2022; Hsieh et al., 2023; Shridhar et al., 2023; Magister et al., 2023; Li et al., 2023a; Fu et al., 2023; Tian et al., 2024). Additionally, Instruction Following Distillation aims to improve the zero-shot performance of LLMs by fine-tuning them with a set of instruction-like prompt-response pairs (Jiang et al., 2023c; Li et al., 2024b). Furthermore, other studies have used distillation to train small models for knowledge-intensive tasks (Li et al., 2024c; Kang et al., 2024; Chen et al., 2024d), intent discovery (Liang et al., 2024), and humor generation (Ravi et al., 2024), demonstrating the versatility and effectiveness of KD in various domains.

Summary and Future Directions Knowledge distillation facilitates the transfer of knowledge from a larger model to a smaller one, enabling the development of more cost-effective and efficient models like DistilBERT. Recent advancements have concentrated on techniques such as Chain-of-Thought distillation and Instruction Following distillation, which enhance the reasoning abilities of smaller models.

> Future Directions
>
> (1) Current knowledge distillation approaches predominantly emphasize using labels and explanations generated by closed-source LLMs to train student models through simple supervised fine-tuning (Xu et al., 2024a). However, expanding the range of knowledge transferred from the teacher model, including feedback on the student's outputs (Lee et al., 2023) and feature knowledge (Gu et al., 2024), can offer additional benefits.
> (2) Efforts in LLM knowledge distillation have largely focused on transferring various skills from LLMs, with relatively limited attention given to trustworthiness (Xu et al., 2024a), such as helpfulness, honesty, and harmlessness (Bai et al., 2022b; Yang et al., 2023; Cui et al., 2023).

2.2.2 Data Synthesis

Human-created data is finite, and there is a concern that publicly available human text may soon be depleted (Villalobos et al., 2024). Additionally, large models are not always necessary for specific tasks. In light of these considerations, using LLMs to generate training data for small model training is both efficient and feasible. In the following sections, we discuss how to leverage LLMs for data synthesis, focusing on two key areas: Training Data Generation and Data Augmentation.

Training Data Generation involves first generating a dataset from scratch using LLMs, such as ChatGPT, in an unsupervised manner, followed by training a small task-specific model on the synthesized dataset. This approach enables highly efficient inference, as the final task model has orders of magnitude fewer parameters compared to the original large model (Ye et al., 2022; Meng et al., 2022; Chung et al., 2023). Subsequent studies have extended this method to various tasks, including text classification (Li et al., 2023c), clinical text mining (Tang et al., 2023), information extraction (Josifoski et al., 2023), and hate speech detection (Hartvigsen et al., 2022). Another approach leverages LLMs solely to generate labels rather than the entire training dataset (Wang et al., 2021; Gao et al., 2022), resembling the process of knowledge distillation.

Data Augmentation in this context refers to the use of LLMs to modify existing data points, thereby increasing data diversity, which can then be directly used to train smaller models (Ding et al., 2024a; Chen et al., 2023b). For instance, LLMs can be employed to paraphrase or rewrite texts to generate additional training samples (Mi et al., 2022; Witteveen and Andrews, 2019). In in-

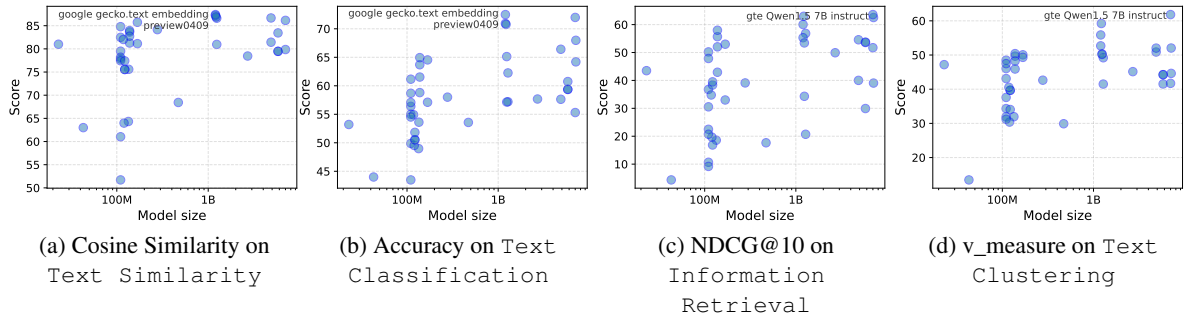|			
(a) Cosine Similarity on `Text Similarity`	(b) Accuracy on `Text Classification`	(c) NDCG@10 on `Information Retrieval`	(d) v_measure on `Text Clustering`

Figure 3: The performance of various models with different sizes on MTEB. We select five datasets for each task. Increasing model sizes only has diminishing returns.

formation retrieval, LLMs can rewrite queries (Ma et al., 2023a) to better align with target documents. Furthermore, data augmentation can be applied to various tasks such as personality detection (Hu et al., 2024a), intent classification (Sahu et al., 2022), and dialogue understanding (Chen et al., 2022b). Fine-tuning smaller models with these augmented samples can significantly enhance their efficacy and robustness.

Summary and Future Directions Synthetic data serves as an effective complement to human-generated data, increasing data diversity and improving coverage of long-tailed samples. This approach can significantly enhance the performance and robustness of smaller models.

Future Directions

(1) Currently, closed-source LLMs remain more powerful than their open-source counterparts. However, using closed-source models for data synthesis may raise privacy and security concerns, particularly in sensitive contexts such as medical scenarios involving patient data (Ollion et al., 2023). Addressing how to protect data privacy during this process is a critical area of concern.

(2) Generating training data with large-scale models is costly, making it essential to explore methods for reducing the expense while still producing high-quality data. For instance, recent research suggests that smaller, less powerful models can sometimes generate better training data points (Bansal et al., 2024).

---

3 Competition

Now, we present three scenarios where smaller models are preferable: computation-constrained environments (§ 3.1), task-specific environments (§ 3.2), and environments requiring interpretability (§ 3.3).

3.1 Computation-constrained Environment

Although LLMs represent a significant milestone in the development of AGI, their impressive capabilities come with substantial computational demands. Scaling model size results in an exponential increase in training time and significantly higher inference latency (Wan et al., 2023). Training and deploying LLMs require more hardware and greater energy consumption, which is often unfeasible for academic researchers and businesses with limited resources. Additionally, this high computational overhead prevents LLMs from being directly applied in computation-constrained environments, such as edge and mobile devices (Dhar et al., 2024).

Moreover, not all tasks require large models; some tasks that are not knowledge-intensive and do not demand complex reasoning can be effectively handled by smaller models. For instance, Figure 3c illustrates the relationship between performance and model size across four tasks in MTEB (Muennighoff et al., 2023), where we observe diminishing returns from increasing model sizes, particularly in tasks like text similarity and classification. The case of information retrieval (Figure 3c), which involves computing similarity between a query and a document collection, faster inference speed is critical. Under these conditions, the lightweight Sentence-BERT (Reimers

and Gurevych, 2019) remains widely used in the IR task.

As a result, there is a growing shift toward smaller, more efficient models, driven by the need for accessibility, efficiency, and democratization. Examples include Phi-3.8B (Abdin et al., 2024), MiniCPM (Hu et al., 2024c), and Gemma-2B (Team et al., 2024).

Small models are increasingly valuable in scenarios where computational resources are limited. Techniques such as knowledge distillation (Xu et al., 2024a) allow the transfer of knowledge from LLMs to smaller models, enabling these smaller models to achieve similar performance while significantly reducing model size. Additionally, smaller models are often preferable for computationally intensive tasks such as information retrieval, due to their lower resource demands.

3.2 Task-specific Environment

Training LLMs require trillions of tokens (Raffel et al., 2020; Kaplan et al., 2020; Gao et al., 2021), but sufficient data is often unavailable for certain specialized domains (e.g., biomedical text) or tasks (e.g., tabular learning). In such cases, pretraining a large foundational model is not feasible, and small models can offer promising returns in this case.

We outline several task-specific scenarios where small models can deliver comparable results:

- Domain-Specific Tasks: Domains such as biomedical or legal fields often have fewer training tokens available. Recent studies have shown that fine-tuning small models on domain-specific datasets can outperform general LLMs on various biomedical (Hernandez et al., 2023; Juan José Bucher and Martini, 2024) and legal (Chalkidis, 2023) tasks.

- Tabular Learning: Tabular datasets are typically smaller than benchmarks in other domains, such as text or image data, and are highly structured, consisting of heterogeneous data types (e.g., numerical, categorical, ordinal). Due to these characteristics, small tree-based models can achieve competitive performance compared to large deep-learning models for tabular data (Grinsztajn

et al., 2022).

- Short Text Tasks: Short text representation and reasoning do not generally require extensive background knowledge. As a result, small models are particularly effective for tasks such as text classification (Zhang et al., 2023a), phrase representation (Chen et al., 2024c), and entity retrieval (Chen et al., 2021).

- Other Specialized Tasks: In certain niche areas, smaller models can surpass larger ones. Examples include machine-generated text detection (Mireshghallah et al., 2023), spreadsheet representation (Joshi et al., 2024), and information extraction (Ma et al., 2023b).

Small models have distinct advantages in specialized areas, and developing lightweight models for these domains or tasks is a promising approach.

3.3 Interpretability-required Environment

The goal of interpretability is to provide a human-understandable explanation of a model's internal reasoning process (Lipton, 2018; Gilpin et al., 2018), i.e., how the model works (*transparency*). Generally, smaller (e.g. shallow) and simpler (e.g. tree-based) models offer better interpretability compared to larger (e.g. deep), more complex models (e.g. neural) (Barceló et al., 2020; Gosiewska et al., 2021).

In practice, industries such as healthcare (Caruana et al., 2015), finance (Kurshan et al., 2021), and law (Eliot, 2021) often favor smaller, more interpretable models because the decisions produced by these models must be understandable to non-experts (e.g., doctors, financial analysts). In high-stakes decision-making contexts, models that can be easily audited and explained are typically preferred.

When selecting LLMs or SMs to use, it is important to make trade-offs for balancing model complexity with the need for human understanding.

4 Conclusion

In this work, we systematically analyze the relationship between LLMs and SMs from two perspectives. First, LLMs and SMs can collaborate to strike a balance between performance and efficiency. Second, they compete under specific conditions: computation-constrained environments, task-specific applications, and scenarios requiring high interpretability. Careful evaluation of the trade-offs between LLMs and SMs is essential when selecting the appropriate model for a given task or application. While LLMs offer superior performance, SMs have notable advantages, including accessibility, simplicity, lower cost, and interoperability. We hope this study provides valuable insights for practitioners, encouraging further research on resource optimization and the development of cost-effective systems.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *ArXiv preprint*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *ArXiv preprint*.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2024. Proxylm: Predicting language model performance on multilingual tasks via proxy models. *ArXiv preprint*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*.

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q Tran, and Mehran Kazemi. 2024. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *ArXiv preprint*.

Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas

Joglekar, Jan Leike, et al. 2024. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *ArXiv preprint*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, (3).

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023a. Accelerating large language model decoding with speculative sampling. *ArXiv preprint*.

Dong Chen, Yueting Zhuang, Shuo Zhang, Jinfeng Liu, Su Dong, and Siliang Tang. 2024a. Data shunt: Collaboration of small and large models for lower costs and better performance. In *Proc. of AAAI*, 10.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023b. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023c. Alpagasus: Training a better alpaca with fewer data. *ArXiv preprint*.

Lihu Chen, Alexandre Perez-Lebel, Fabian M Suchanek, and Gaël Varoquaux. 2024b. Reconfidencing llms from the grouping loss perspective. *arXiv preprint arXiv:2402.04957*.

Lihu Chen, Gael Varoquaux, and Fabian Suchanek. 2022a. Imputing out-of-vocabulary embeddings with love makes languagemodels robust with little cost. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3504.

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2021. A lightweight neural model for biomedical entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12657–12665.

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2024c. Learning high-quality and general-purpose phrase representations. In *EACL 2024-The 18th Conference of the European Chapter of the Association for Computational Linguistics*.

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

Lingjiao Chen, Matei Zaharia, and James Y. Zou. 2020. Frugalml: How to use ML prediction apis more accurately and cheaply. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022b. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Shiyao Cui, Tingwen Liu, Shuohuan Wang, Yu Sun, and Hua Wu. 2024d. Lemon: Reviving stronger and smaller lms from larger lms with linear parameter fusion. In *Proc. of ACL*.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. In *Proc. of EMNLP*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang,

Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. *ArXiv preprint*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, (240).

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proc. of ACL*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv preprint*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Nobel Dhar, Bobin Deng, Dan Lo, Xiaofeng Wu, Liang Zhao, and Kun Suo. 2024. An empirical analysis and resource footprint study of deploying large language models on edge devices. In *Proceedings of the 2024 ACM Southeast Conference*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv preprint*.

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

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. 2024b. Hybrid llm: Cost-efficient and quality-aware query routing. *ArXiv preprint*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. 📒urvey on in-context learning. *ArXiv preprint*.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *ArXiv preprint*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *ArXiv preprint*.

Dr Lance B Eliot. 2021. The need for explainable ai (xai) is especially crucial in the law. *Available at SSRN 3975778*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*. PMLR.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2022. Self-guided noise-free data generation for efficient zero-shot learning. *ArXiv preprint*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2021. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv preprint*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv preprint*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE.

Alicja Gosiewska, Anna Kozak, and Przemysław Biecek. 2021. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, (6).

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. 2024. Vision superalignment: Weak-to-strong generalization for vision foundation models. *ArXiv preprint*.

Yue Guo and Yi Yang. 2024. Improving weak-to-strong generalization with reliability-aware alignment. *ArXiv preprint*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proc. of ACL*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *ArXiv preprint*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *Proc. of ICLR*.

Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*. PMLR.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*.

Cheng-Yu Hsieh, Chun-Liang Li, CHIH-KUAN YEH, Hootan Nakhost, Yasuhisa Fujii, Alex Jason 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. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

I Hsu, Zifeng Wang, Long T Le, Lesly Miculicich, Nanyun Peng, Chen-Yu Lee, Tomas Pfister, et al. 2024. Calm: Contrasting large and small language models to verify grounded generation. *ArXiv preprint*.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024a. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proc. of AAAI*, 16.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024b. Routerbench: A benchmark for multi-llm routing system. *ArXiv preprint*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024c. Minicpm: Unveiling the potential of

small language models with scalable training strategies. *ArXiv preprint*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *ArXiv preprint*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *ArXiv preprint*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024. Aligner: Achieving efficient alignment through weak-to-strong correction. *ArXiv preprint*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *ArXiv preprint*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023a. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Structgpt: A general framework for large language model to reason over structured data. In *Proc. of EMNLP*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024b. Survey on large language models for code generation. *ArXiv preprint*.

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

Harshit Joshi, Abishai Ebenezer, José Cambronero Sanchez, Sumit Gulwani, Aditya Kanade, Vu Le, Ivan Radiček, and Gust Verbruggen. 2024. Flame: A small language model for spreadsheet formulas. In *Proc. of AAAI*, 12.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *Proc. of EMNLP*.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv e-prints*.

Gurusha Juneja, Subhabrata Dutta, Soumen Chakrabarti, Sunny Manchanda, and Tanmoy Chakraborty. 2023. Small language models fine-tuned to coordinate larger language models improve complex reasoning. In *Proc. of EMNLP*.

Anil Kag and Igor Fedorov. 2023. Efficient edge inference by selective query. In *International Conference on Learning Representations*.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*.

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 preprint*.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Hao Peng, Ximing Lu, Dragomir Radev, Yejin Choi, and Noah A. Smith. 2022. Twist decoding: Diverse generators guide each other. In *Proc. of EMNLP*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Eren Kurshan, Jiahao Chen, Victor Storchan, and Hongda Shen. 2021. On the current and emerging challenges of developing fair and ethical ai solutions in financial services. In *Proceedings of the second ACM international conference on AI in finance*.

Hunter Lang, David Sontag, and Aravindan Vi-

jayaraghavan. 2024. Theoretical analysis of weak-to-strong generalization. *ArXiv preprint*.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2024a. Orchestrallm: Efficient orchestration of language models for dialogue state tracking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*.

Jooyoung Lee, Fan Yang, Thanh Tran, Qian Hu, Emre Barut, and Kai-Wei Chang. 2024b. Can small language models help large language models reason better?: Lm-guided chain-of-thought. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Zhijing Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2024a. Blade: Enhancing black-box large language models with small domain-specific models. *ArXiv preprint*.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023a. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proc. of ACL*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024b. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *ArXiv preprint*.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *ArXiv preprint*.

Xiang Li, Shizhu He, Fangyu Lei, JunYang JunYang, Tianhuang Su, Kang Liu, and Jun Zhao. 2024c. Teaching small language models to reason for knowledge-intensive multi-hop question answering. In *Findings of the Association for Computational Linguistics ACL 2024*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proc. of ACL*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023c. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proc. of EMNLP*.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*. PMLR.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In *Findings of the Association for Computational Linguistics ACL 2024*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, (3).

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. 2024a. Tuning language models by proxy. *ArXiv preprint*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of EMNLP*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. -train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, (9).

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Yuejiang Liu and Alexandre Alahi. 2024. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *ArXiv preprint*.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023b. Llm-qat: Data-free quantization aware training for large language models. *ArXiv preprint*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *ArXiv preprint*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*. PMLR.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023a. Routing to the expert: Efficient reward-guided ensemble of large language models. *ArXiv preprint*.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jung, Khyathi Chandu, Abhilasha Ravichander, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, et al. 2023b. erence-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. In *Proc. of EMNLP*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023a. Query rewriting in retrieval-augmented large language models. In *Proc. of EMNLP*.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Aman Madaan, Pranjal Aggarwal, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, et al. 2023. Automix: Automatically mixing language models. *ArXiv preprint*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Dominik Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proc. of EMNLP*.

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 preprint*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023a. ent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, (2).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. of EMNLP*.

Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors. *ArXiv preprint*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Surya Narayanan Hari and Matt Thomson. 2023. Tryage: Real-time, intelligent routing of user prompts to large language models. *arXiv e-prints*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Sean O'Brien and Mike Lewis. 2023. Contrastive

decoding improves reasoning in large language models. *ArXiv preprint*.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype. *SocArXiv preprint*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *ArXiv preprint*.

OpenAI. 2024. Chatgpt [large language model].

Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. 2023. Comblm: Adapting black-box language models through small fine-tuned models. *ArXiv preprint*.

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*.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *ArXiv preprint*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

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. *CoRR*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proc. of EMNLP*.

Alec Radford, Jeffrey Wu, Rewon Child, David

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 preprint*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023a. ent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, (2).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. of EMNLP*.

Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors. *ArXiv preprint*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Surya Narayanan Hari and Matt Thomson. 2023. Tryage: Real-time, intelligent routing of user prompts to large language models. *arXiv e-prints*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *ArXiv preprint*.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype. *SocArXiv preprint*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *ArXiv preprint*.

OpenAI. 2024. Chatgpt [large language model].

Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. 2023. Comblm: Adapting black-box language models through small fine-tuned models. *ArXiv preprint*.

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*.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *ArXiv preprint*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

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. *CoRR*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proc. of EMNLP*.

Alec Radford, Jeffrey Wu, Rewon Child, David

Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, (8).

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. *J. Mach. Learn. Res.*

Sahithya Ravi, Patrick Huber, Akshat Shrivastava, Aditya Sagar, Ahmed Aly, Vered Shwartz, and Arash Einolghozati. 2024. Small but funny: A feedback-driven approach to humor distillation. *ArXiv preprint*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP*.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *ArXiv preprint*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, (4).

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*.

Marija Šakota, Maxime Peyrard, and Robert West. 2024. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*.

V Sanh. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *ArXiv preprint*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *ArXiv preprint*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *ArXiv preprint*.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. 2023. Large language model routing with benchmark datasets. *ArXiv preprint*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*.

Haotian Sun, Yuchen Zhuang, Wei Wei, Chao Zhang, and Bo Dai. 2024. Bbox-adapter: Lightweight adapting for black-box large language models. *ArXiv preprint*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *ArXiv preprint*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *ArXiv preprint*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *ArXiv preprint*.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. 2024. Tinyllm: Learning a small student from multiple large language models. *ArXiv preprint*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2024. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proc. of ACL*.

Neeraj Varshney and Chitta Baral. 2022. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proc. of EMNLP*.

Giorgos Vernikos, Arthur Bražinskas, Jakub Adamek, Jonathan Mallinson, Aliaksei Severyn, and Eric Malmi. 2024. Small language models improve giants by rewriting their outputs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. 2023. Efficient large language models: A survey. *Transactions on Machine Learning Research*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of ICLR*.

Lidan Wang, Jimmy J. Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proc. of SIGIR*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *ArXiv preprint*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language

models. *Advances in neural information processing systems*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *Forty-first International Conference on Machine Learning*.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*.

Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. 2024a. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. *ArXiv preprint*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024b. Less: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2024. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *ArXiv preprint*.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *ArXiv preprint*.

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

Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024b. Bridging the gap between different vocabularies for llm ensemble. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrastive distillation for lm alignment. In *The Twelfth International Conference on Learning Representations*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*. PMLR.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proc. of EMNLP*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant

context. In *The Twelfth International Conference on Learning Representations*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *CoRR*.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. 2024a. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. *ArXiv preprint*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proc. of ICLR*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023a. Sentiment analysis in the era of large language models: A reality check. *ArXiv preprint*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023b. Alleviating hallucinations of large language models through induced hallucinations. *ArXiv preprint*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024b. Safetybench: Evaluating the safety of large language models. In *Proc. of ACL*.

Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Generating synthetic speech from SpokenVocab for speech translation. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*.

Shuyan Zhou, Uri Alon, Frank F Xu, Zhengbao Jiang, and Graham Neubig. 2023. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. 2024b. Weak-to-strong search: Align large language models via searching over small language models. *ArXiv preprint*.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *ArXiv preprint*.