

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

JINGFENG YANG*, Amazon, USA

HONGYE JIN*, Department of Computer Science and Engineering, Texas A&M University, USA

RUIXIANG TANG*, Department of Computer Science, Rice University, USA

XIAOTIAN HAN*, Department of Computer Science and Engineering, Texas A&M University, USA

QIZHANG FENG*, Department of Computer Science and Engineering, Texas A&M University, USA

HAOMING JIANG, Amazon, USA

BING YIN, Amazon, USA

XIA HU, Department of Computer Science, Rice University, USA

This paper presents a comprehensive and practical guide for practitioners and end-users working with Large Language Models (LLMs) in their downstream natural language processing (NLP) tasks. We provide discussions and insights into the usage of LLMs from the perspectives of models, data, and downstream tasks. Firstly, we offer an introduction and brief summary of current GPT- and BERT-style LLMs. Then, we discuss the influence of pre-training data, training data, and test data. Most importantly, we provide a detailed discussion about the use and non-use cases of large language models for various natural language processing tasks, such as knowledge-intensive tasks, traditional natural language understanding tasks, natural language generation tasks, emergent abilities, and considerations for specific tasks. We present various use cases and non-use cases to illustrate the practical applications and limitations of LLMs in real-world scenarios. We also try to understand the importance of data and the specific challenges associated with each NLP task. Furthermore, we explore the impact of spurious biases on LLMs and delve into other essential considerations, such as efficiency, cost, and latency, to ensure a comprehensive understanding of deploying LLMs in practice. This comprehensive guide aims to provide researchers and practitioners with valuable insights and best practices for working with LLMs, thereby enabling the successful implementation of these models in a wide range of NLP tasks. A curated list of practical guide resources of LLMs, regularly updated, can be found at <https://github.com/Mooler0410/LLMsPracticalGuide>.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Natural language generation**; **Machine translation**.

Additional Key Words and Phrases: Large Language Models, Neural Language Processing, Practical Guide, ChatGPT

1 INTRODUCTION

In recent years, the rapid development of Large language Models has been revolutionizing the field of natural language processing [12, 128, 131]. These powerful models have shown great potential in addressing a variety of NLP tasks, ranging from natural language understanding (NLU) to generation tasks, even paving the way to Artificial General Intelligence (AGI). However, utilizing these models effectively and efficiently requires a practical understanding of their capabilities and limitations, as well as the data and tasks involved in NLP.

To provide a guide for practitioners and end-users, this work focuses on the practical aspects of working with LLMs in downstream NLP tasks. This guide aims to provide practical advice on why or why not to choose LLMs for a given

*These authors contributed equally.

Authors' addresses: Jingfeng Yang, jingfengyangpku@gmail.com, Amazon, USA; Hongye Jin, jhy0410@tamu.edu, Department of Computer Science and Engineering, Texas A&M University, USA; Ruixiang Tang, rt39@rice.edu, Department of Computer Science, Rice University, USA; Xiaotian Han, han@tamu.edu, Department of Computer Science and Engineering, Texas A&M University, USA; Qizhang Feng, qf31@tamu.edu, Department of Computer Science and Engineering, Texas A&M University, USA; Haoming Jiang, jhaoming@amazon.com, Amazon, USA; Bing Yin, alexbyin@amazon.com, Amazon, USA; Xia Hu, xia.hu@rice.edu, Department of Computer Science, Rice University, USA.

task, as well as guidance on how to select the most suitable LLM, taking into account factors such as model sizes, computational requirements, and the availability of domain-specific pre-trained models. This work offers a thorough understanding of LLMs from a practical perspective, therefore, empowers practitioners and end-users with the practical knowledge needed to successfully leverage the power of LLMs for their own NLP tasks.

Our work is structured as follows. First, our work offers a brief introduction to LLMs by discussing the most important models, such as GPT-style and BERT-style architectures. Then, we delve into the critical factors that influence model performance from the data perspective, including pre-training data, training/tuning data, and test data. Last and most importantly, we dive deep into various concrete NLP tasks, offering insights into the applicability of LLMs for knowledge-intensive tasks, traditional NLU tasks, and generation tasks, along with the emergent abilities that these models possess and challenging real-world scenarios. We provide detailed examples to highlight both the successful use cases and the limitations of LLMs in practice.

To analyze the abilities of large language models, we compare them with fine-tuned models. As of present, there is no universally recognized definition for LLMs and fine-tuned models. With consideration to practical utility, in our article, the definitions of them are proposed as: LLMs are huge language models pretrained on large amounts of datasets without tuning on data for specific tasks; fine-tuned models are typically smaller language models which are also pretrained and then further tuned on a smaller, task-specific dataset to optimize their performance on that task¹.

This work summarizes the following main practical guides for using LLMs:

- **Natural language understanding.** Employ the exceptional generalization ability of LLMs when facing out-of-distribution data or with very few training data.
- **Natural language generation.** Utilize LLMs’ capabilities to create coherent, contextually relevant, and high-quality text for various applications.
- **Knowledge-intensive tasks.** Leverage the extensive knowledge stored in LLMs for tasks requiring domain-specific expertise or general world knowledge.
- **Reasoning ability.** Understand and harness the reasoning capabilities of LLMs to improve decision-making and problem-solving in various contexts.

2 PRACTICAL GUIDE FOR MODELS

This section provides a brief introduction to state-of-the-art LLMs. These models differ in their training strategies, model architectures, and use cases. To provide a clearer understanding of the LLM landscape, we categorize them into two types: encoder-decoder or encoder-only language models and decoder-only language models. In Figure 1, we show the detailed evolution process of language models. From the evolutionary tree, we make the following interesting observations:

- Decoder-only models have been gradually dominating the development of LLMs. At the early stage of LLMs development, **decoder-only** models were not as popular as **encoder-only** and **encoder-decoder** models. However, after 2021, with the introduction of game-changing LLMs - GPT-3, decoder-only models experienced a significant boom. Meanwhile, after the initial explosive growth brought about by BERT, encoder-only models gradually began to fade away.

¹From a practical standpoint, we consider models with less than 20B parameters to be fine-tuned models. While it’s possible to fine-tune even larger models like **PlAM** (540B), in reality, it can be quite challenging, particularly for academic research labs and small teams. Fine-tuning a model with 3B parameters can still be a daunting task for many individuals or organizations.

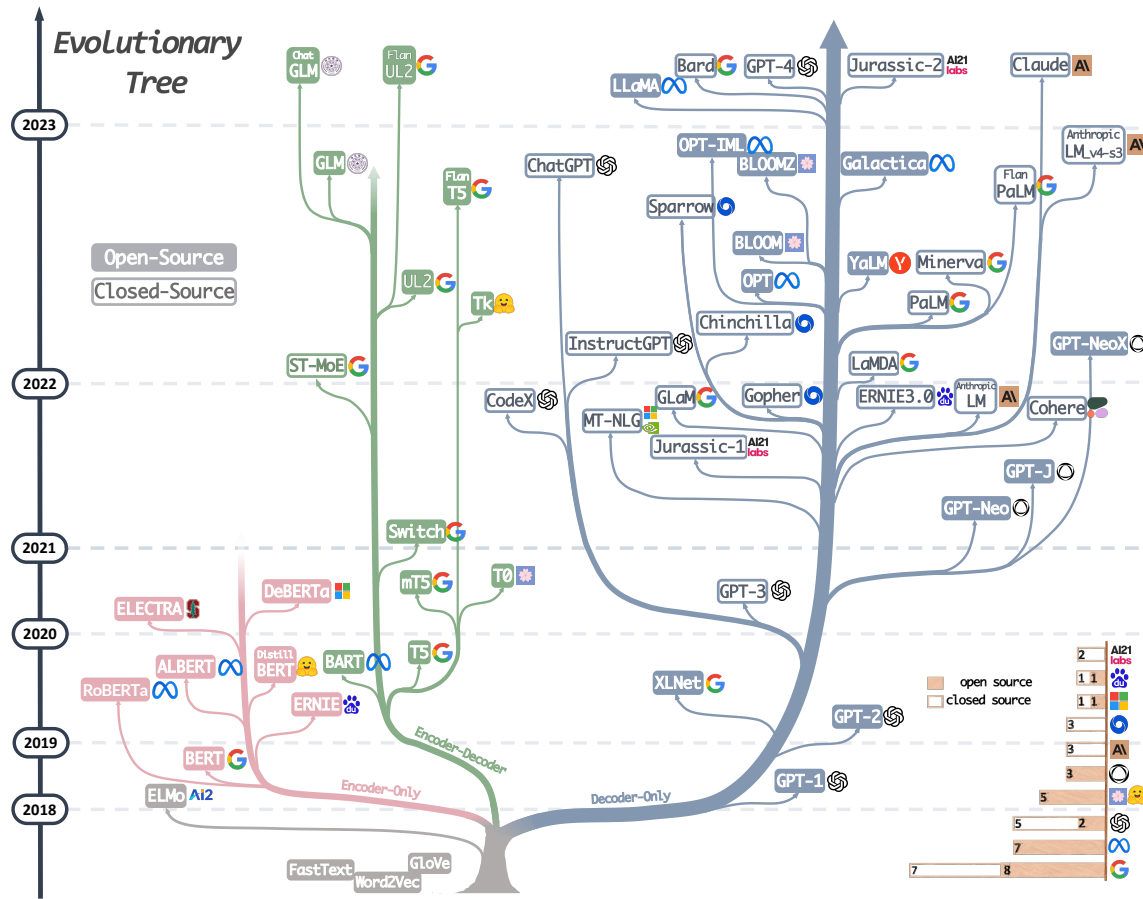


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

- b) OpenAI consistently maintains its leadership position in LLM, both currently and potentially in the future. Other companies and institutions are struggling to catch up with OpenAI in developing models comparable to GPT-3 and the current GPT-4. This leadership position may be attributed to OpenAI's steadfast commitment to its technical path, even when it was not widely acknowledged initially.
- c) Meta contributes significantly to open-source LLMs and promotes research of LLMs. When considering contributions to the open-source community, particularly those related to LLMs, Meta stands out as one of the most generous commercial companies, as all the LLMs developed by Meta are open-sourced.
- d) LLMs exhibit a tendency towards closed-sourcing. In the early stages of LLM development (before 2020), the majority of models were open-sourced. However, with the introduction of GPT-3, companies have increasingly

Table 1. Summary of Large Language Models.

	Characteristic	LLMs
Encoder-Decoder or Encoder-only (BERT-style)	Training: Masked Language Models Model type: Discriminative Pretrain task: Predict masked words	ELMo [80], BERT [28], RoBERTa [65], DistilBERT [90], BioBERT [57], XLM [54], Xlnet [119], ALBERT [55], ELECTRA [24], T5 [84], XLM-E [20], ST-MoE [133], AlexaTM [95]
Decoder-only (GPT-style)	Training: Autoregressive Language Models Model type: Generative Pretrain task: Predict next word	GPT-3 [16], OPT [126], PaLM [22], BLOOM [92], GLM [123], MT-NLG [93], GLaM [32], Gopher [83], chinchilla [41], LaMDA [102], GPT-J [107], LLaMA [103], GPT-4 [76], BloombergGPT [117]

opted to close-source their models, such as PaLM, LaMDA, and GPT-4. Consequently, it has become more difficult for academic researchers to conduct experiments on LLM training. As a result, API-based research could become the predominant method in the academic community.

- e) Encoder-decoder models remain promising, as this type of architecture is still being actively explored, and most of them are open-sourced. Google has made substantial contributions to open-source encoder-decoder architectures. However, the flexibility and versatility of decoder-only models seem to make Google’s insistence on this direction less promising.

We also briefly summarize the characteristics and the representative LLMs of each type in Table 1.

2.1 BERT-style Language Models: Encoder-Decoder or Encoder-only

As natural language data is readily available and unsupervised training paradigms have been proposed to better utilize extremely large datasets, this motivates the unsupervised learning of natural language. One common approach is to predict masked words in a sentence while considering the surrounding context. This training paradigm is known as the Masked Language Model. This type of training allows the model to develop a deeper understanding of the relationships between words and the context in which they are used. These models are trained on a large corpus of texts using techniques such as the Transformer architecture and have achieved state-of-the-art results in many NLP tasks, such as sentiment analysis and named entity recognition. Notable examples of Masked Language Models include BERT [28], RoBERTa [65], and T5 [84]. MLMs have become an important tool in the field of natural language processing due to their success in a wide range of tasks.

2.2 GPT-style Language Models: Decoder-only

Although language models are typically task-agnostic in architecture, these methods require fine-tuning on datasets of the specific downstream task. Researchers found that scaling up language models significantly improves the few-shot, even zero-shot performance [16]. The most successful models for better few-shot and zero-shot performance are Autoregressive Language Models, which are trained by generating the next word in a sequence given the preceding words. These models have been widely used for downstream tasks such as text generation and question answering. Examples of Autoregressive Language Models include GPT-3 [16], OPT [126], PaLM [22], and BLOOM [92]. The game changer, GPT-3, for the first time, demonstrated reasonable few-/zero-shot performance via prompting and in-context learning, thus showing the superiority of autoregressive language models. There are also models such as CodeX [2]

that are optimized for specific tasks such as code generation, BloombergGPT [117] for the financial domain. The recent breakthrough is ChatGPT, which refines GPT-3 specifically for conversational tasks, resulting in more interactive, coherent, and context-aware conversational for various real-world applications.

3 PRACTICAL GUIDE FOR DATA

In this section, we'll be discussing the critical role that data plays in selecting appropriate models for downstream tasks. The impact of data on the models' effectiveness starts during the pre-training stage and continues through to the training and inference stages.

Remark 1

- (1) LLMs generalize better than fine-tuned models in downstream tasks facing out-of-distribution data, such as adversarial examples and domain shifts.
- (2) LLMs are preferable to fine-tuned models when working with limited annotated data, and both can be reasonable choices when abundant annotated data is available, depending on specific task requirements.
- (3) It's advisable to choose models pre-trained on fields of data that are similar to downstream tasks.

3.1 Pretraining data

Pre-training data plays a pivotal role in the development of large language models. As the foundation of remarkable capabilities [5, 47] of LLMs, the quality, quantitative, and diversity of pre-training data influence the performance of LLMs significantly [124]. The commonly used pretraining data consists of a myriad of text sources, including books, articles, and websites. The data is carefully curated to ensure a comprehensive representation of human knowledge, linguistic nuances, and cultural perspectives. The importance of pretraining data lies in its capacity to inform the language model with a rich understanding of word knowledge, grammar, syntax, and semantics, as well as the ability to recognize context and generate coherent responses. The diversity of pretraining data also plays a crucial role in shaping the model's performance, and the selection of LLMs highly depends on the components of the pretraining data. For example, PaLM [22] and BLOOM [92] excel in multilingual tasks and machine translation with an abundance of multilingual pretraining data. Moreover, PaLM's performance in Question Answering tasks is enhanced by incorporating a considerable amount of social media conversations and Books corpus [22]. Likewise, code execution and code completion capabilities of GPT-3.5 (code-davinci-002) are amplified by the integration of code data in its pretraining dataset. In brief, when selecting LLMs for downstream tasks, it is advisable to choose the model pre-trained on a similar field of data.

3.2 Finetuning data

When deploying a model for downstream tasks, it is essential to consider three primary scenarios based on the availability of annotated data: zero, few, and abundant. In this section, we provide a succinct overview of the appropriate models to employ for each scenario.

Zero annotated data: In scenarios where annotated data is unavailable, utilizing LLMs in a zero-shot setting proves to be the most suitable approach. LLMs have been shown to outperform previous zero-shot methods [120]. Additionally, the absence of a parameter update process ensures that catastrophic forgetting [49] is avoided since the language model parameters remain unaltered.

Few annotated data: In this case, the few-shot examples are directly incorporated in the input prompt of LLMs, which is named as in-context learning, and these examples can effectively guide LLMs to generalize to the task. As reported in [16], one-shot and few-shot performance make significant gains, even matching the performance of the SOTA fine-tuned open-domain models. And LLMs’ zero/few-shot ability can be improved further by scaling [16]. Alternatively, some few-shot learning methods are invented to enhance fine-tuned models, such as meta-learning [56] or transfer learning [88]. However, performance might be inferior compared to using LLMs due to fine-tuned models’ smaller scale and overfitting.

Abundant annotated data: With a substantial amount of annotated data for a particular task available, both fine-tuned models and LLMs can be considered. In most cases, fine-tuning the model can fit the data pretty well. Although, LLMs can be used to meet some constraints such as privacy [99]. In this scenario, the choice between using a fine-tuned model or a LLM is task-specific and also depends on many factors, including desired performance, computational resources, and deployment constraints.

In a brief summary: LLMs are more versatile w.r.t. the data availability, while fine-tuned models can be considered with abundant annotated data.

3.3 Test data/user data

When deploying LLMs for downstream tasks, we often face challenges stemming from distributional differences between the test/user data and that of the training data. These disparities may encompass domain shifts [132], out-of-distribution variations [31], or even adversarial examples [82]. Such challenges significantly hinder fine-tuned models’ effectiveness in real-world applications. They fit into a specific distribution and have a poor ability to generalize to OOD data. However, LLMs perform quite well facing such scenarios because they do not have an explicit fitting process. Moreover, recent advancements have further enhanced the ability of language models in this regard. The Reinforcement Learning from Human Feedback (RLHF) method has notably enhanced LLMs’ generalization capabilities [77]. For example, InstructGPT demonstrates proficiency in following various instructions for a wide range of tasks and occasionally complying with instructions in different languages, even though such instructions are scarce. Similarly, ChatGPT exhibits consistent advantages on most adversarial and out-of-distribution (OOD) classification and translation tasks [109]. Its superiority in understanding dialogue-related texts led to an impressive performance on the DDXPlus dataset [101], a medical diagnosis dataset designed for OOD evaluation.

4 PRACTICAL GUIDE FOR NLP TASKS

In this section, we discuss in detail the use cases and no use cases for LLMs in various downstream NLP tasks and the corresponding model abilities. And in Figure 2, we summarize all discussions into a decision flow. It can be a guide for a quick decision while facing a task.

4.1 Traditional NLU tasks

Traditional NLU tasks are some fundamental tasks in NLP including text classification, named entity recognition (NER), entailment prediction, and so on. Many of them are designed to serve as intermediate steps in larger AI systems, such as NER for knowledge graph construction.

¹As we mention in Section 1, LLMs are pretrained on large and diverse datasets without fine-tuning, while fine-tuned models are typically pretrained on a large dataset and then further trained on a smaller, task-specific dataset to optimize their performance on that task.

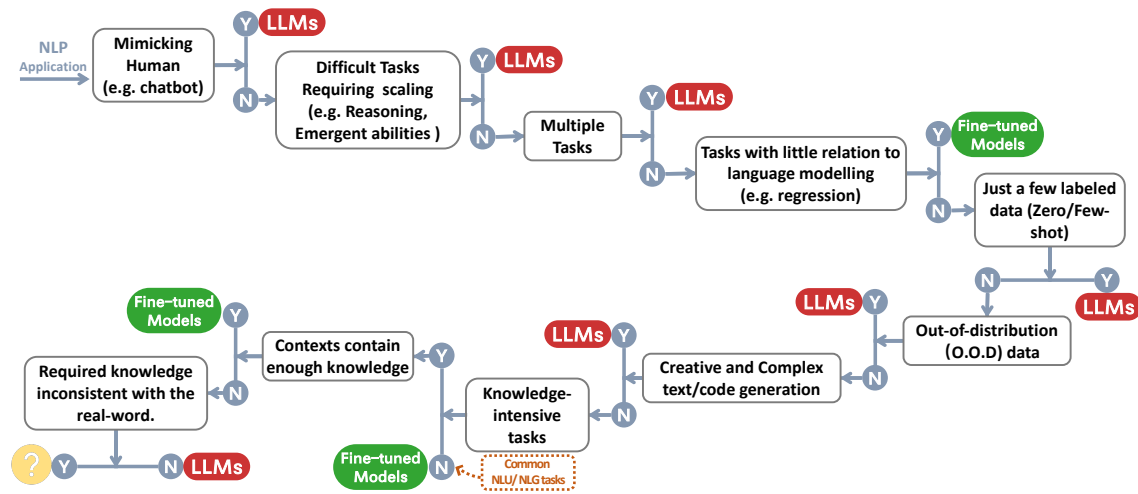


Fig. 2. The decision flow for choosing LLMs or fine-tuned models² for user’s NLP applications. The decision flow helps users assess whether their downstream NLP applications at hand meet specific conditions and, based on that evaluation, determine whether LLMs or fine-tuned models are the most suitable choice for their applications. During the decision process in the figure, **Y** means meeting the condition, and **N** means not meeting the condition. The yellow circle for **Y** of the last condition means there’s no model working well on this kind of application.

Remark 2

Fine-tuned models generally are a better choice than LLMs in traditional NLU tasks, but LLMs can provide help while requiring strong generalization ability.

4.1.1 No use case. In most natural language understanding tasks, such as tasks in GLUE[106] and SuperGLUE[105], fine-tuned models still have better performance, if such tasks come with rich well-annotated data and contain very few out-of-distribution examples on test sets. For different tasks and datasets, the gap between small fine-tuned models and LLMs varies.

In text classification, on most datasets, LLMs perform slightly worse than fine-tuned models. For sentiment analysis, such as on IMDB [69] and SST [94], fine-tuned models and LLMs perform equally well. For toxicity detection, which is another iconic text classification task, the gap is much larger. All LLMs cannot perform well on this task, and on CivilComments [13] even the best one is only better than random guessing [59]. On the other hand, most popular fine-tuned models can obtain much better performance [33]. and the Perspective API³ is still one of the best for detecting toxicity. This API is powered by a multilingual BERT-based model, which is tuned on publicly available toxicity data and several smaller single-language CNNs distilled from this model. This might be due to the fact that toxicity is defined by subtle nuances in linguistic expressions, and large language models are unable to accurately comprehend this task solely based on the provided input.

The trend of performance gaps is similar in some other tasks. For natural language inference (NLI) tasks, on most datasets, such as on RTE [106] and SNLI [14], fine-tuned models perform better than LLMs, while on some data such as CB [105], LLMs have obtained comparable performance with fine-tuned models [22]. For question answering (QA), on

³<https://perspectiveapi.com>

SQuADv2 [86], QuAC [21] and many other datasets, fine-tuned models have superior performance, while on CoQA [87], LLMs perform as well as fine-tuned models [22].

In information retrieval (IR) tasks, LLMs are not widely exploited yet. One major reason is that IR tasks are fundamentally different from others. There's no natural way to transform the thousands of candidate texts into a few/zero-shot form which is required by LLMs. The existing evaluation results on MS MARCO(regular/TREC) [73] show that methods based on fine-tuned models have better performance [59]. In this evaluation, the LLMs rank passages in an unorthodox way, which requires the LLMs to produce probabilities for passages one by one.

For some low-level intermediate tasks, which are not intended for regular users but rather for high level tasks, such as named entity recognition (NER) and dependency parsing, there's not enough result coming from LLMs, because the most current evaluation of LLMs focuses on practical tasks. According to available evaluation results, for the NER task, CoNLL03 [89] is still a challenge for LLMs [81], where the performance of fine-tuned models is around as twice as LLMs. These intermediate tasks may vanish soon because LLMs can take over high-level tasks without the help of those intermediate tasks (e.g. dependency parsing for coding tasks; NER for some text generation tasks).

In brief, for most traditional NLU tasks, a fine-tuned model is a better choice in terms of the performance on benchmark datasets and the computational cost. The scale of LLMs is usually 10× or even 100× larger than fine-tuned models. One possible cause for the inferior performance of LLMs on certain tasks can be the design of instructions/prompts. Transforming input from tasks like IR and sentence labeling into a few/zero-shot instruction form is non-trivial. There may be better ways to adapt language models to traditional NLP tasks in the future. On the other hand, the upper limit of capabilities of fine-tuned models is not reached, and some methods like FLAN-tuning [67] can further boost the performance on NLU tasks. Another interesting finding is that on NLU tasks, after fine-tuning, masked language models, like T5[85], are better than most auto-regressive language models at the same scale, while some recent results imply that this gap can be bridged by scaling[22].

4.1.2 Use case. However, there are still some NLU tasks suitable for LLMs.

One of the representative tasks is miscellaneous text classification [59]. In contrast to classic domain-specific text classification tasks such as sentiment analysis, miscellaneous text classification deals with a diverse range of topics and categories that may not have a clear or strong relationship with one another. It's closer to real-world cases and hard to be formatted for using fine-tuned models. Another is the Adversarial NLI (ANLI)[74]. It is a difficult dataset composed of adversarially mined natural language inference questions in three rounds (R1, R2, and R3). LLMs have shown superior performance on ANLI, especially on the R3 and R2. Both examples demonstrate the exceptional ability of LLMs to generalize well on out-of-distribution and sparsely annotated data in traditional NLP tasks, surpassing that of fine-tuned models. We've discussed this in the section above 3.3.

4.2 Generation tasks

Natural Language Generation broadly encompasses two major categories of tasks, with the goal of creating coherent, meaningful, and contextually appropriate sequences of symbols. The first type focuses on converting input texts into new symbol sequences, as exemplified by tasks like paragraph summarization and machine translation. The second type, "open-ended" generation, aims to generate text or symbols from scratch to accurately match input descriptions such as crafting emails, composing news articles, creating fictional stories and writing code.

Remark 3

Due to their strong generation ability and creativity, LLMs show superiority at most generation tasks.

4.2.1 Use case. Generation tasks require models to have a comprehensive understanding of the input contents or requirements and a certain level of creativity. This is what LLMs excel at.

For summarization tasks, although LLMs do not have an obvious advantage over fine-tuned models under traditional automatic evaluation metrics, such as ROUGE [60], human evaluation results indicate that humans tend to prefer the results generated by LLMs [38, 127] compared to that of fine-tuned models. For example, on CNN/DailyMail [71] and XSUM [72], fine-tuned models like Brio [66] and Pegasus [125] have much better performance than any LLMs w.r.t. ROUGE, but LLMs like OPT [126] perform far better in human evaluation considering all aspects including faithfulness, coherence, and relevance [127]. This demonstrates the superiority of LLMs in summarization tasks. On the other hand, it implies that current summarization benchmarks don't contain summaries with high quality or the automatic metrics are not proper for the evaluation of summarization.

In machine translation (MT), LLMs can perform competent translation, although the average performance is slightly worse than some commercial translation tools [45] considering some automatic metrics like BLEU[78]. LLMs are particularly good at translating some low-resource language texts to English texts, such as in the Romanian-English translation of WMT'16 [11], zero-shot or few-shot LLMs can perform better than SOTA fine-tuned model[22]. This is mainly due to the fact that English resources compose the main part of the pre-training data. BLOOM [92] is pre-trained on more multi-lingual data, leading to better translation quality in both rich-resource and low-resource translation. Another interesting finding is that BLOOM achieves good translation quality among Romance languages, even for translation from Galician, which is not included in the pre-training data. One reasonable explanation is that texts from some languages in the same language group can help the LLMs learn more from the similarity. If more multi-lingual texts can be added to the pre-training data, the translation capability may be improved further.

Additionally, LLMs are highly skilled in open-ended generations. One example is that the news articles generated by LLMs are almost indistinguishable from real news articles by humans [16]. LLMs are remarkably adept at code synthesis as well. Either for text-code generation, such as HumanEval [18] and MBPP [7], or for code repairing, such as DeepFix [39], LLMs can perform pretty well. GPT-4 can even pass 25% problems in Leetcode, which are not trivial for most human coders [76]. With training on more code data, the coding capability of LLMs can be improved further [22]. While performing well on such tasks, the codes generated by LLMs should be tested carefully to figure out any subtle bugs, which is one of the main challenges for applying LLMs in code synthesis.

4.2.2 No use case. Fine-tuned models, such as DeltaLM+Zcode [118], still perform best on most rich-resource translation and extremely low-resource translation tasks. In rich resource machine translation, fine-tuned models slightly outperform LLMs [22, 92]. And in extremely low-resource machine translation, such as English-Kazakh translation, fine-tuned models significantly perform better than LLMs.

4.3 Knowledge-intensive tasks

Knowledge-intensive NLP tasks refer to a category of tasks that have a strong reliance on background knowledge, domain-specific expertise, or general real-world knowledge. These tasks go beyond simple pattern recognition or syntax analysis. And they are highly dependent on memorization and proper utilization of knowledge about specific entities, events, and common sense of our real world.

Remark 4

- (1) LLMs excel at knowledge-intensive tasks due to their massive real-world knowledge.
- (2) LLMs struggle when the knowledge requirements do not match their learned knowledge, or when they face tasks that only require contextual knowledge, in which case fine-tuned models can work as well as LLMs.

4.3.1 Use case. In general, with billions of training tokens and parameters, LLMs have much more real-world knowledge than fine-tuned models.

Closed-book question-answering tasks require the model to answer a given question about factual knowledge without any external information. It does require the memorization of real-world knowledge in the model. LLMs perform better on nearly all datasets, such as on NaturalQuestions [52], WebQuestions [9], and TriviaQA [46]. On TriviaQA, even zero-shot LLMs is still much better [22].

The massive multitask language understanding (MMLU) [40] is also highly knowledge-intensive. It contains multiple-choice questions spanning over 57 different subjects and requires general knowledge of the model. It's pretty challenging even for LLMs, although the newly released GPT-4 [76] outperforms existing models by a considerable margin in English with a satisfactory 86.5% accuracy.

Also, some tasks in Big-bench[96], which are designed to probe LLMs and extrapolate their future capabilities, heavily relied on the memorization of real-world knowledge. In such tasks, the performance of some LLMs is better than the average level of humans, and even comparable to the best human performance. For example, the task *Hindu_knowledge* requires models to give facts about Hindu mythology, *Periodic Elements* require the capability of predicting the element name from the periodic table and *Physics* tests the physics knowledge of models by asking for the formula needed to solve a given physics problem.

4.3.2 No use case. There are some other tasks requiring knowledge different from that learned by LLMs. The required knowledge is not that learned by LLMs about the real world. In such tasks, LLMs are not notably superior.

Some tasks only require the model to capture the self-contained knowledge in the contexts. The knowledge in the contexts from the input is enough for the model to make predictions. For these tasks, small fine-tuned models can work pretty well. One such task is machine reading comprehension (MRC). An MRC task provides several paragraphs and requires the model to predict the answer to questions based on these paragraphs. We've discussed MRC in the previous section because it's also a traditional NLU task.

Another scenario is that the knowledge within LLMs about real world is useless to the task, or even the required knowledge is counterfactual to the real world. As a result, the LLMs cannot work well on such tasks. In some cases, inconsistent knowledge may even make the LLMs worse than random guessing. For example, in Big-Bench, the Mnist ascii task requires the model to tell the digit represented by an ASCII art. The capability required by this task is nothing about real-world knowledge. Also, in the Inverse Scaling Phenomenon competition [70], the task *redefine math* redefines a common symbol and requires the model to choose between the original meaning and the meaning derived from the redefinition. What it requires contrasts to the LLMs' knowledge, thus LLMs even perform worse than random guessing.

As an alternative to real-world knowledge in LLMs, access to extra knowledge is allowed, and models can thus get enough knowledge for a task via retrieval augmentation. The basic idea of retrieval augmentation is to add an extra information retrieval step prior to making predictions, in which, some useful texts related to the task will be retrieved from a large corpus. Then, the model will make predictions based on both the input contexts and the retrieved texts. With retrieved additional information, the closed-book task can become "open-book". In such a scenario, fine-tuned

models are pretty good with much smaller sizes, because the required knowledge can be obtained by retrieving. For example, on NaturalQuestions [52], with extra corpus, retrieval augmented models [44, 48] are much better than any other methods.

4.4 Abilities Regarding Scaling

Scaling of LLMs (e.g. parameters, training computation, etc.) can greatly empower pretrained language models. With the model scaling up, a model generally becomes more capable in a range of tasks. Reflected in some metrics, the performance shows a power-law relationship with the model scale. For example, the cross-entropy loss which is used to measure the performance for language modeling decreases linearly with the exponential increase in the model scale, which is also called 'scaling-law' [41, 47]. For some crucial abilities, such as reasoning, scaling the model has gradually transformed these abilities from a very low state to a usable state, and even approaching human capabilities. In this section, we provide an overview of the usage of LLMs in terms of the abilities and behaviors of LLMs along with scaling.

Remark 5

- (1) With the exponential increase of model scales, LLMs become especially capable of reasoning like arithmetic reasoning and commonsense reasoning.
- (2) Emergent abilities become serendipity for uses that arise as LLMs scale up, such as ability in word manipulation and logical ability.
- (3) In many cases, performance does not steadily improve with scaling due to the limited understanding of how large language models' abilities change as they scale up.

4.4.1 Use Case with Reasoning. Reasoning, which involves making sense of information, drawing inferences, and making decisions, is one of the essential aspects of human intelligence. It is challenging for NLP. Many existing reasoning tasks can be classified into commonsense reasoning and arithmetic reasoning.

Arithmetic reasoning/problem solving. The arithmetic reasoning capability of LLMs benefits greatly from the scaling of model size. For GPT-3, the ability of two-digit addition only becomes apparent when the number of parameters exceeds 13B [16]. Tasks to test arithmetic reasoning are trivial for humans and designed to challenge the capability of transferring natural language into mathematical symbols and multi-step inference. On GSM8k [26], SVAMP [79] and AQuA [61], LLMs, as generalists, have competitive performance with most methods which have task-specific designs. And GPT-4 overperforms any other methods [76], even some huge models particularly tuned for arithmetic problems [104]. Nevertheless, it should be noted that, without the intervention of external tools, LLMs may occasionally make mistakes in performing basic calculations, although chain-of-thought (CoT) prompting [115] can significantly improve LLMs' ability in calculations.

Commonsense reasoning. Commonsense reasoning not only requires LLMs to remember factual knowledge but also requires LLMs to do several inference steps about the facts. Commonsense reasoning increases gradually with the growth of model size. Compared to fine-tuned models, LLMs keep the superiority on most datasets, such as StrategyQA [36] and ARC-C [25]. Especially on ARC-C, which contains difficult questions in science exams from grade 3 to grade 9, GPT-4 has been close to the performance of 100% (96.3%) [76].

4.4.2 Use Cases with Emergent Abilities. Scaling of models also endows the model with some unprecedented, fantastic abilities that go beyond the power-law rule. These abilities are called "emergent ability". As defined in [113], *emergent abilities of LLMs are abilities that are not present in smaller-scale models but are present in large-scale models*. This means

such abilities cannot be predicted by extrapolating the performance improvements on smaller-scale models and the model suddenly gains good performance on some tasks once the scale exceeds a certain range. The emergent ability is typically unpredictable and surprising, leading to tasks that emerge randomly or unexpectedly. We examine concrete examples of the emergent abilities of LLMs and provide them as an important reference for deciding whether to leverage LLMs' emergent abilities.

Handling word manipulation is a typical emergent ability. It refers to the ability to learn symbolic manipulations, such as the reversed words [16], in which the model is given a word spelled backwards, and must output the original word. For example, GPT-3 [16] shows the emergent ability for word sorting, and word unscrambling tasks. PaLM [22] exhibits the emergent ability on ASCII word recognition⁴ and hyperbaton⁵ task. The logical abilities of language models tend to emerge as the model scales up, such as logical deduction, logical sequence, and logic grid puzzles. Additionally, other tasks, such as advanced coding (e.g., auto debugging, code line description), and concept understanding (e.g., novel concepts, simple Turing concepts), are also use cases with the emergent abilities of large language models.

4.4.3 No-Use Cases and Understanding. Although in most cases, as discussed above, larger models bring better performance, there are still many exceptions that should be considered when choosing the appropriate model.

On certain tasks, with the size of LLMs increasing, the performance begins to decrease, such as Redefine-math: tests whether language models are able to work with common symbols when they are redefined to mean something else; Into-the-unknown: requires the model to choose which piece of information would help answer a question; Memo-trap: asks an LM to write a phrase in a way that starts like a famous quote but ends differently⁶. This is also called *Inverse Scaling Phenomenon*. Another interesting phenomenon observed in the scaling of LLMs is called the *U-shaped Phenomenon* [114]. As the name implies, This phenomenon refers to that as LLM size increases, their performance on certain tasks initially improves but then starts to decline before eventually improving again, such as on: Hindsight-neglect: it tests whether language models are able to assess whether a bet was worth taking based on its expected value; NegationQA: this task takes an existing multiple-choice dataset and negates a part of each question to see if language models are sensitive to negation; Quote-repetition: it asks models to repeat back sentences given in the prompt, with few-shot examples to help it recognize the task. Hence the risk of diminishing performance should be noted and if the task is similar to those we just discussed, careful consideration should be given to whether or not to use huge LLMs.

Gaining a deeper understanding of emergent abilities, inverse scaling phenomenon and U-shape phenomenon in LLMs is essential for advancing research in this field. In a certain sense, the U-shape phenomenon suggests that small-scale models and huge-scale models make predictions with different internal mechanisms. From this perspective, the U-shape phenomenon can be seen as a transformation of the inverse-scaling phenomenon due to some emergent abilities from sufficiently large models [114]. GPT-4 [76] exhibits a reversal of the inverse scaling phenomenon in some cases, such as on a task called Hindsight Neglect. The explanation for these behaviors of LLMs during scaling is still an open problem. Several hypotheses have been proposed. For emergent abilities, one explanation is that there may be multiple key steps for a task and the LLM cannot handle this task until it's large enough to handle every step, and another explanation is focused on the granularity of evaluation metrics [113]. For inverse-scaling phenomenon and

⁴Asking models to identify the word displayed as ASCII art, https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/ascii_word_recognition

⁵Asking models to choose the English sentence with adjectives in the "correct" order within two choices, https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/hyperbaton

⁶More such tasks include: modus-tollens, pattern-matching-suppression, prompt-injection, repetitive-algebra and sig-figs. You can check them on: <https://github.com/inverse-scaling/prize>

u-shape phenomenon, the explanations mainly focus on the model's over-reliance on information from its prior rather than the input prompts, valid but misleading few-shot examples, and distracting easier tasks within a hard task [114].

4.5 Miscellaneous tasks

This section explores miscellaneous tasks which cannot be involved in previous discussions, to better understand LLMs' strengths and weaknesses.

Remark 6

- (1) Fine-tuned models or specified models still have their space in tasks that are far from LLMs' pretraining objectives and data.
- (2) LLMs are excellent at mimicking human, data annotation and generation. They can also be used for quality evaluation in NLP tasks and have bonuses like interpretability.

4.5.1 No use case. LLMs generally struggle with some tasks due to differences in objectives and training data.

Although LLMs have achieved remarkable success in various natural language processing tasks, their performance in regression tasks has been less impressive. For example, ChatGPT's performance on the GLUE STS-B dataset, which is a regression task evaluating sentence similarity, is inferior to a fine-tuned RoBERTa performance [130]. The Regression tasks typically involve predicting a continuous value rather than a discrete label, posing unique challenges for LLMs. One primary reason for their subpar performance is the inherent difference between the language modeling objective and the regression task objective. LLMs are designed to predict the next word in a sequence or generate coherent text, with their pre-training focused on capturing linguistic patterns and relationships. Consequently, their internal representations may not be well-suited for modeling continuous numerical outputs. Besides, LLMs have predominantly been trained on text data, focusing on capturing the intricacies of natural language processing. As a result, their performance on multimodal data, which involves handling multiple data types such as text, images, audio, video, actions, and robotics, remains largely unexplored. And fine-tuned multimodal models, like BEiT[110] and PaLI [19], still dominate many tasks such as visual question answering (VQA) and image captioning. Nonetheless, the recently introduced GPT-4 [76] has taken the step in multimodal fusion, but there is still a lack of detailed evaluation of its capabilities.

4.5.2 Use case. LLMs are particularly suitable for certain tasks.

LLMs are very good at mimicking humans, acting as a chatbot, and performing various kinds of tasks. The LLMs-powered ChatGPT⁷ is surprising for its consistency, reliability, informativeness, and robustness during multiple utterances with humans. The human-feedback procedure plays an important role in acquiring such abilities

LLMs can both act as a good annotator and data generator for data augmentation, such as in[27, 29, 99, 121, 122]. Some LLMs have been found as good as human annotators [37] in some tasks. And the collected texts from GPT-3.5 (text-davinci-003) have been used as human-like instruction-following demonstrations to train other language models [100].

LLMs can also be used for quality assessment on some NLG tasks, such as summarization and translation. On summarization tasks, GPT-4 as an evaluator achieves a higher correlation with humans than other methods with a large margin [64]. Some other evaluators based on LLMs [34, 50, 64, 108] also show good human alignment in more

⁷<https://chat.openai.com>

NLG tasks, especially compared with traditional automatic metrics. But the LLM evaluator may have a bias towards the LLM-generated texts [64].

Also, as we discussed above, some abilities of LLMs bring bonuses in addition to performance improvement, such as interpretability. The CoT reasoning ability of LLMs can show how an LLM reaches the prediction, which is a good interpretation on the instance level, while it also improves the performance.

4.6 Real world "tasks"

In the last part of this section, we would like to discuss the usage of LLMs and fine-tuned models in real-world "tasks". We use the term "tasks" loosely, as real-world scenarios often lack well-formatted definitions like those found in academia. Many requests to models even cannot be treated as NLP tasks. Models face challenges in the real world from three perspectives:

- **Noisy/Unstructured input.** Real-world input comes from real-world non-experts. They have little knowledge about how to interact with the model or even cannot use texts fluently. As a result, real-world input data can be messy, containing typos, colloquialisms, and mixed languages, unlike those well-formed data used for pre-training or fine-tuning.
- **Tasks not formalized by academia.** In real-world scenarios, tasks are often ill-defined by academia and much more diverse than those in academic settings. Users frequently present queries or requests that do not fall neatly into predefined categories, and sometimes multiple tasks are in a single query.
- **Following users' instructions.** A user's request may contain multiple implicit intents (e.g. specific requirement to output format), or their desired predictions may be unclear without follow-up questions. Models need to understand user intents and provide outputs that align with those intents.

Essentially, these challenges in the real world come from that users' requests deviate significantly from the distribution of any NLP datasets designed for specific tasks. Public NLP datasets are not reflective of how the models are used [77].

Remark 7

LLMs are better suited to handle real-world scenarios compared to fine-tuned models. However, evaluating the effectiveness of models in the real world is still an open problem.

Handling such real-world scenarios requires coping with ambiguity, understanding context, and handling noisy input. Compared to fine-tuned models, LLMs are better equipped for this because they have been trained on diverse data sets that encompass various writing styles, languages, and domains. Additionally, LLMs demonstrate a strong ability to generate open-domain responses, making them well-suited for these scenarios. Fine-tuned models, on the other hand, are often tailored to specific, well-defined tasks and may struggle to adapt to new or unexpected user requests. They heavily rely on clear objectives and well-formed training data that specify the types of instructions the models should learn to follow. Fine-tuned models may struggle with noisy input due to their narrower focus on specific distributions and structured data. An additional system is often required as an assistant for fine-tuned models to process unstructured context, determine possible intents, and refine model responses accordingly.

Additionally, some mechanics such as instruction tuning [91, 112] and human alignment tuning [77] further boost the capabilities of LLMs to better comprehend and follow user instructions. These methods improve the model's ability to generate helpful, harmless, and honest responses while maintaining coherence and consistency [77, 91, 112]. While both methods can make LLMs better generalize to unseen tasks and instructions, it has been noticed that while human

labelers prefer models tuned for human alignment [77] to models tuned with instructions from public NLP tasks, such as FLAN [112] and T0 [91]. The reason may be similar to reasons for fine-tuned models’ inferiority: public NLP tasks/datasets are designed for easy and automatic evaluation, and they can only cover a small part of real-world usage.

One of the main issues when it comes to real-world scenarios is how to evaluate whether the model is good or not. Without any formalized tasks or metrics, the evaluation of model effectiveness can only rely on feedback from human labelers. Considering the complexity and cost of human evaluation, there’s no massive and systematic comparison between fine-tuned models and LLMs yet. Nevertheless, the huge success and popularity of LLMs such as chatGPT, have confirmed the superiority of LLMs to some extent.

5 OTHER CONSIDERATIONS

Despite LLMs are suitable for various downstream tasks, there are some other factors to consider, such as efficiency and trustworthiness. Our discussion of efficiency encompasses the training cost, inference latency, and parameter-efficient tuning strategies for LLMs. Meanwhile, our examination of trustworthiness includes robustness & calibration, fairness & biases, potential spurious correlations, and the safety challenges in LLMs.

Remark 8

- (1) Light, local, fine-tuned models should be considered rather than LLMs, especially for those who are sensitive to the cost or have strict latency requirements. Parameter-Efficient tuning can be a viable option for model deployment and delivery.
- (2) The zero-shot approach of LLMs prohibits the learning of shortcuts from task-specific datasets, which is prevalent in fine-tuned models. Nevertheless, LLMs still demonstrate a degree of shortcut learning issues.
- (3) Safety concerns associated with LLMs should be given utmost importance as the potentially harmful or biased outputs, and hallucinations from LLMs can result in severe consequences. Some methods such as human feedback have shown promise in mitigating these problems.

5.1 Efficiency

In real-world deployment, performance, cost, and latency are all important considerations, not just the performance of the models. While some parameter-efficient methods have been developed, practitioners must balance efficiency with effectiveness in the practice.

Cost. LLMs have grown increasingly larger in recent years, with models such as GPT-1, GPT-2, and GPT-3 featuring 117 million, 1.5 billion, and 175 billion parameters, respectively. The cost of training an LLM is heavily influenced by its size, with estimates suggesting that training the 11B parameter variant of T5 costs well over \$1.3 million for a single run, while a single training run of GPT-3 175B requires \$4.6 million [3]. The energy consumption for training large models is equally impressive. The total energy consumption for training a transformer model with 6B parameters to completion is estimated to be around 103.5 MWh [30]. Google reports that training PaLM consumed about 3.4 GWh in about two months [6]. Furthermore, the dataset size also scales rapidly with the size of the model, with GPT-3 175B trained on 499 billion tokens [16]. Another key metric that reflects the computing cost is Flops, with GPT-3 175B requiring 3.14×10^{23} Flops, while a T5 11B model only requires 3.30×10^{22} , which is 10 times less. In addition to these costs, hardware requirements are also substantial. OpenAI has collaborated with Microsoft on a supercomputer hosted in the Microsoft Azure cloud, consisting of 285k CPU cores and 10k high-end GPUs to support the training of large models. For users of

the OpenAI API, pricing varies based on the model and usage, with options such as GPT-3.5-turbo charging \$0.002 per 1k tokens for chat service. However, for users who require custom models, training costs \$0.03 per 1k tokens, while usage costs \$0.12 per 1k tokens [4]. Therefore, for users who cannot afford such a large cost, such as small startups, individual users, etc., a small, fine-tuned model is a better and more reasonable choice.

Latency. Latency is a crucial factor to consider in real-world applications of LLMs. Inference time is a commonly used metric to measure latency, which is highly dependent on the model size, architecture, and token size. For instance, the inference time for the GPT-J 6B model is 0.077s, 0.203s, and 0.707s when the max token size is set to 2, 8, and 32, respectively. Additionally, when the max token size is fixed at 32, the inference time for the InstructGPT model (davinci v2) is 1.969s. As LLMs are often too large to be run on a single user’s machine, companies provide LLM services via APIs. The API latency can vary depending on the user’s location, and the average latency of the OpenAI API service for a single request can range from a few hundred milliseconds to several seconds. In scenarios where high latency is not acceptable, large LLMs may not be appropriate. For example, scalability is critical in many information retrieval applications. To deploy information retrieval systems on the web, search engines require very efficient inference for systems to be useful. The idealized denoised inference time for the InstructGPT davinci v2 (175B*) model is 0.21s per request (i.e., a query-passage pair to be scored), which is too slow for web search engines.

Parameter-Efficient Tuning. In practice, we may tune the model on some specific datasets. Parameter-Efficient Tuning (PET) is an efficient technique to tune a small portation of model parameters (or extra parameters) while freezing most parameters of the pre-trained LLMs. The main goal of PEFT is to greatly decrease the computational and storage costs while keeping the performance of the original models. The common techniques for PET are LoRA [42], Prefix Tuning [58], P-Tuning [62, 63]. As an illustration, the LoRA method maintains the weights of the pre-trained model and incorporates low-rank matrices into every layer of the Transformer architecture. This approach considerably minimizes the number of parameters that require training for subsequent tasks, thereby increasing overall efficiency. Alpaca-LoRA⁸ proposes integrating Low-Rank Adaptation (LoRA) into LLaMA-Alpaca, which enables runs LLaMA within hours on a single RTX 4090. All these PFT methods can be helpful either for fine-tuning a model to a specific task or tuning LLMs to meet special requirements like human alignment.

5.2 Trustworthiness

Given that LLMs are now involved in sensitive areas such as healthcare, finance, and law, it is crucial to ensure that they are trustworthy and capable of producing reliable output.

Robustness and Calibration. The accuracy and robustness of the LLMs are shown to have a very strong correlation [59]. The models that have high accuracy on the scenario also have good robustness. However, the robustness of the zero-shot becomes worse after being tuned on extra application-specific tasks data [116]. This may due to overfitting, which leads to poor generalizability due to the extremely high complexity of the model and the limited training samples from downstream tasks [43]. In a similar vein, it has been observed that fine-tuning a model can result in significant miscalibrations, owing to over-parameterization [51]. Therefore, fine-tuned models may not be an optimal choice when robustness and calibration are critical considerations. However, human alignment has been found as a potential solution for enhancing model robustness. InstructGPT davinci v2 (175B*) has been shown to outperform other models in terms

⁸<https://github.com/tloen/alpaca-lora>

of robustness. On the other hand, achieving optimal calibration of the model depends on the scenario and adaptation procedure employed.

Fairness and Bias. LLMs have been shown to exhibit disparate treatment and impact, perpetuating societal biases and potentially leading to discrimination [10, 17]. To ensure fairness and equity for all users, it is crucial to address these issues in the development and deployment of NLP models. Disparities in performance between demographic groups can serve as an indicator of fairness problems. LLMs are particularly susceptible to fairness issues, as significant performance disparities have been observed across demographic categories such as dialect, religion, gender, and race [59]. However, research has shown that aligning models with human instructions can improve LLM performance regardless of their size, with the InstructGPT model (davinci v2) exhibiting smaller performance disparities than other LLMs [23].

Spurious Biases. The shortcut learning problem has been observed in various natural language understanding tasks under the pretraining and fine-tuning paradigm, where models heavily rely on spurious correlations between input and labels in the fine-tuning data for prediction [31, 35, 98]. For example, in reading comprehension tasks, fine-tuned models tend to focus on the lexical matching of words between the question and the original passage, neglecting the intended reading comprehension task itself [53]. In contrast, large language models are not directly trained on fine-tuned datasets, which makes it less likely for them to learn shortcut features present in the fine-tuned dataset, thereby enhancing the model's generalization capabilities. However, LLMs are not infallible and may exhibit some shortcut learning during in-context learning. For example, recent preliminary studies have begun investigating the robustness of prompt-based methods in large-scale language models [111, 129]. One such study evaluates the few-shot learning performance of GPT-3 on text classification and information extraction tasks [129], and reveal that the examined LLMs are susceptible to majority label bias and position bias, where they tend to predict answers based on the frequency or position of the answers in the training data. Moreover, these LLMs exhibit common token bias, favoring answers that are prevalent in their pre-training corpus. Recent studies show that this positional bias can be mitigated by selecting proper prompts [68]. In summary, while LLMs significantly reduce the shortcut learning problem prevalent in fine-tuned models, they still exhibit some shortcut learning issues and should be approached with caution when deploying them in downstream applications.

5.3 Safety challenges

LLMs have demonstrated their extremely strong capabilities in many areas such as reasoning, knowledge retention, and coding. As they become more powerful and human-like, their potential to influence people's opinions and actions in significant ways grows. As a result, some new safety challenges to our society should be considered and have caught lots of attention in recent works [75, 76].

Hallucinations. The potential for LLMs to "hallucinate," or generate nonsensical or untruthful content, can have significant negative impacts on the quality and reliability of information in various applications. As LLMs become increasingly convincing and believable, users may develop an overreliance on them and trust them to provide accurate information in areas with which they are somewhat familiar. This can be particularly dangerous if the model produces content that is entirely false or misleading, leading to incorrect decisions or actions taken based on that information. Such outcomes can have serious consequences in many domains, such as healthcare, finance, or public policy, where the accuracy and reliability of information are critical. To mitigate these issues, reinforcement learning from human feedback (RLHF) is widely used [75, 77] and LLMs themselves have been integrated into the loop [75].

Harmful content. Due to the high coherence, quality, and plausibility of texts generated by LLMs, harmful contents from LLMs can cause significant harm, including hate speech, discrimination, incitement to violence, false narratives, and even social engineering attack. The implementation of safeguards to detect and correct those contents can be mitigation [97]. These LLMs can also have dual-use potential by providing required illicit information, leading to risks such as the proliferation of weapons [75] and even terrorism attack planning. It is crucial to ensure using these LLMs responsibly, with safeguards in place to prevent harm. Also, in existing work, feedback from humans plays an important role in getting rid of harmful outputs.

Privacy. LLMs can face serious security issues. An example is the issue of user privacy. It is reported that Samsung employees were using ChatGPT to process their work when they inadvertently leaked top-secret data, including the source code proper of the new program, internal meeting minutes related to the hardware, etc. The Italian data protection agency declared that OpenAI, the developer of ChatGPT, illicitly gathered personal user data, leading Italy to become the first government to prohibit ChatGPT over privacy concerns [1].

6 CONCLUSION AND FUTURE CHALLENGES

Recent advances in large language models have been revolutionizing the field of natural language processing. Effectively using LLMs requires understanding their capabilities, and limitations for various NLP tasks. This work presents a practical guide to working with LLMs for downstream NLP tasks. We first discuss prominent models like GPT-style and BERT-style architectures and the factors influencing their performance. We then explore using LLMs for downstream tasks, including knowledge-intensive tasks, NLU, and NLG tasks, as well as providing concrete examples of successes and limitations. This practical guide offers insights into LLMs and best practices for harnessing LLMs across NLP tasks. We hope it would enable researchers and practitioners to leverage their potential and drive innovation in language technologies.

In the following, we figure out the future challenges of the LLMs:

- **Evaluation of proposed models on real-world “datasets”.** While existing deep learning models are primarily evaluated on standard academic datasets, such as ImageNet, which have been milestones in deep learning development. However, the limitations of standard academic datasets can not exactly reflect real-world performance. As models advance, it is crucial to assess them on more diverse, complex, and realistic data that reflect real-world needs. Evaluating models on real-world “datasets”, in addition to academic ones, will provide a more rigorous test of their capabilities, as well as a better understanding of their effectiveness in real-world applications. This ensures that the models are capable of addressing real-world challenges and delivering practical solutions.
- **Model Alignment.** Ensuring that increasingly powerful and autonomous models align with human values and priorities is essential. Methods must be developed to guarantee that these models behave as intended and do not optimize for undesirable outcomes. It is crucial to integrate alignment techniques from the start of the model development process. Model transparency and interpretability are also important factors for evaluating and ensuring alignment. Additionally, as we look toward the future, an even more daunting challenge looms: aligning superhuman systems. While this task is currently beyond our demands, it is important to consider and prepare for the potential implications of aligning such advanced systems, as they may present unique complexities and ethical concerns [8, 15].
- **Safety Alignment.** While discussion of AI existential risks is important, concrete research is needed to guarantee the safe development of advanced AI. This includes techniques for interpretability, scalable oversight and governance,

and formal verification of model properties. Safety should be considered not just an add-on but an integral part of the model-building process.

- **Performance Prediction with Scaling.** It is difficult to anticipate how model performance will change as model size and complexity increases dramatically. Developing methods to better predict model performance after scaling up or as new architectures are developed would allow for more efficient use of resources and accelerated progress. Some possibilities include: training a smaller 'seed' model and extrapolating its growth, simulating the effects of increased scale or model tweaks, and benchmarking iterations of the model at different scales to build scaling laws. These could provide insight into the performance of models even before they are built.

REFERENCES

- [1] Chatgpt is banned in Italy over privacy concerns - the New York Times. <https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html>. (Accessed on 04/23/2023).
- [2] OpenAI Codex. <https://openai.com/blog/openai-codex>.
- [3] OpenAI's GPT-3 language model: A technical overview. <https://lambdalabs.com/blog/demystifying-gpt-3#1>. (Accessed on 03/02/2023).
- [4] Pricing. <https://openai.com/pricing>. (Accessed on 03/02/2023).
- [5] Ahmed Alajlami and Nikolaos Aletras. How does the pre-training objective affect what large language models learn about linguistic properties? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–147, 2022.
- [6] Anil Ananthaswamy. In AI, is bigger always better? *Nature*, 615(7951):202–205, 2023.
- [7] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [9] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.
- [10] Camiel J Beukeboom and Christian Burgers. How stereotypes are shared through language: a review and introduction of the social categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37, 2019.
- [11] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névoul, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [13] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.
- [14] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
- [15] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.
- [16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- [18] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [19] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [20] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. XLM-e: Cross-lingual language model pre-training via electra. *arXiv preprint arXiv:2106.16138*, 2021.
- [21] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

- [22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. [arXiv preprint arXiv:2204.02311](#), 2022.
- [23] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. [arXiv preprint arXiv:2210.11416](#), 2022.
- [24] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. [arXiv preprint arXiv:2003.10555](#), 2020.
- [25] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. [arXiv preprint arXiv:1803.05457](#), 2018.
- [26] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. [arXiv preprint arXiv:2110.14168](#), 2021.
- [27] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. [arXiv preprint arXiv:2302.13007](#), 2023.
- [28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. [arXiv preprint arXiv:1810.04805](#), 2018.
- [29] Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. Is gpt-3 a good data annotator? [arXiv preprint arXiv:2212.10450](#), 2022.
- [30] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. Measuring the carbon intensity of ai in cloud instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894, 2022.
- [31] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding: A survey. [arXiv preprint arXiv:2208.11857](#), 2022.
- [32] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.
- [33] Corentin Duchene, Henri Jamet, Pierre Guillaume, and Reda Dehak. A benchmark for toxic comment classification on civil comments dataset. [arXiv preprint arXiv:2301.11125](#), 2023.
- [34] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. [arXiv preprint arXiv:2302.04166](#), 2023.
- [35] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [36] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.
- [37] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. [arXiv preprint arXiv:2303.15056](#), 2023.
- [38] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. [arXiv preprint arXiv:2209.12356](#), 2022.
- [39] Rahul Gupta, Soham Pal, Aditya Kanade, and Shirish Shevade. Deepfix: Fixing common c language errors by deep learning. In *Proceedings of the aaai conference on artificial intelligence*, volume 31, 2017.
- [40] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. [arXiv preprint arXiv:2009.03300](#), 2020.
- [41] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. [arXiv preprint arXiv:2203.15556](#), 2022.
- [42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. [arXiv preprint arXiv:2106.09685](#), 2021.
- [43] Hang Hua, Xingjian Li, Dejing Dou, Cheng-Zhong Xu, and Jiebo Luo. Fine-tuning pre-trained language models with noise stability regularization. [arXiv preprint arXiv:2206.05658](#), 2022.
- [44] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot Learning with Retrieval Augmented Language Models. 2022.
- [45] Wenxiang Jiao and WenxuanWang Jen-tseHuang XingWang ZhaopengTu. Is chatgpt a good translator? yes with gpt-4 as the engine.
- [46] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. [arXiv preprint arXiv:1705.03551](#), 2017.
- [47] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. [arXiv preprint arXiv:2001.08361](#), 2020.
- [48] Akhil Kedia, Mohd Abbas Zaidi, and Haejun Lee. Fie: Building a global probability space by leveraging early fusion in encoder for open-domain question answering. [arXiv preprint arXiv:2211.10147](#), 2022.
- [49] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

- [50] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. [arXiv preprint arXiv:2302.14520](#), 2023.
- [51] Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in-and out-of-distribution data. [arXiv preprint arXiv:2010.11506](#), 2020.
- [52] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. [Transactions of the Association for Computational Linguistics](#), 7:453–466, 2019.
- [53] Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. Why machine reading comprehension models learn shortcuts? In [Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021](#), pages 989–1002, 2021.
- [54] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. [arXiv](#), 2019.
- [55] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. [arXiv](#), 2019.
- [56] Hung-Yi Lee, Shang-Wen Li, and Thang Vu. Meta learning for natural language processing: A survey. In [Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 666–684, 2022.
- [57] Jinhuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. [Bioinformatics](#), 36(4):1234–1240, 2020.
- [58] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. [arXiv preprint arXiv:2101.00190](#), 2021.
- [59] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. [arXiv preprint arXiv:2211.09110](#), 2022.
- [60] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In [Text summarization branches out](#), pages 74–81, 2004.
- [61] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. [arXiv preprint arXiv:1705.04146](#), 2017.
- [62] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. [arXiv preprint arXiv:2110.07602](#), 2021.
- [63] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In [Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics \(Volume 2: Short Papers\)](#), pages 61–68, 2022.
- [64] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment, 2023.
- [65] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. [arXiv preprint arXiv:1907.11692](#), 2019.
- [66] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio: Bringing order to abstractive summarization. [arXiv preprint arXiv:2203.16804](#), 2022.
- [67] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. [arXiv preprint arXiv:2301.13688](#), 2023.
- [68] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In [Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 8086–8098, 2022.
- [69] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In [Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies](#), pages 142–150, 2011.
- [70] Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. Inverse scaling prize: Second round winners, 2023.
- [71] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. [arXiv preprint arXiv:1602.06023](#), 2016.
- [72] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. [arXiv preprint arXiv:1808.08745](#), 2018.
- [73] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. [choice](#), 2640:660, 2016.
- [74] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. [arXiv preprint arXiv:1910.14599](#), 2019.
- [75] OpenAI. Gpt-4 system card.
- [76] OpenAI. Gpt-4 technical report, 2023.
- [77] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. [Advances in Neural Information Processing Systems](#), 35:27730–27744, 2022.

- [78] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [79] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.
- [80] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv*, 2018.
- [81] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.
- [82] Shilin Qiu, Qihe Liu, Shijie Zhou, and Wen Huang. Adversarial attack and defense technologies in natural language processing: A survey. *Neurocomputing*, 492:278–307, 2022.
- [83] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- [84] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [85] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [86] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- [87] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.
- [88] Sebastian Ruder, Matthew Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing tutorial. *NAACL HTL 2019*, page 15, 2019.
- [89] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.
- [90] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv*, 2019.
- [91] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.
- [92] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [93] Shaden Smith, Mostofa Patwary, Brandon Norrick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [94] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [95] Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*, 2022.
- [96] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- [97] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*, 2023.
- [98] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645, 2021.
- [99] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.
- [100] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [101] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Proceedings of the Neural Information Processing Systems-Track on Datasets and Benchmarks*, 2, 2022.
- [102] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Llama: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [103] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2023.

- [104] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. [arXiv preprint arXiv:2211.14275](#), 2022.
- [105] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.
- [106] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. [arXiv preprint arXiv:1804.07461](#), 2018.
- [107] Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [108] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. [arXiv preprint arXiv:2303.04048](#), 2023.
- [109] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. [arXiv preprint arXiv:2302.12095](#), 2023.
- [110] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. [arXiv preprint arXiv:2208.10442](#), 2022.
- [111] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, 2022.
- [112] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. [arXiv preprint arXiv:2109.01652](#), 2021.
- [113] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.
- [114] Jason Wei, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. [arXiv preprint arXiv:2211.02011](#), 2022.
- [115] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. [arXiv preprint arXiv:2201.11903](#), 2022.
- [116] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.
- [117] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjanjan Kambadur, David Rosenberg, and Gideon Mann. Bloombergpt: A large language model for finance. [arXiv preprint arXiv:2303.17564](#), 2023.
- [118] Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online, November 2021. Association for Computational Linguistics.
- [119] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [120] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, 2019.
- [121] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. Gpt3mix: Leveraging large-scale language models for text augmentation. [arXiv preprint arXiv:2104.08826](#), 2021.
- [122] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. [arXiv preprint arXiv:2303.16756](#), 2023.
- [123] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. [arXiv preprint arXiv:2210.02414](#), 2022.
- [124] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. [arXiv preprint arXiv:2303.10158](#), 2023.
- [125] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.
- [126] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. [arXiv preprint arXiv:2205.01068](#), 2022.
- [127] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. [arXiv preprint arXiv:2301.13848](#), 2023.
- [128] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. [arXiv preprint arXiv:2303.18223](#), 2023.
- [129] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

- [130] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. [arXiv preprint arXiv:2302.10198](#), 2023.
- [131] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. [arXiv preprint arXiv:2302.09419](#), 2023.
- [132] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. [IEEE Transactions on Pattern Analysis and Machine Intelligence](#), 2022.
- [133] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. URL <https://arxiv.org/abs/2202.08906>.