Large Language Models Meet NLP: A Survey

Libo Qin[♣] **Qiguang Chen**[♠] **Xiachong Feng**[♢] **Yang Wu**[♠] **Yongheng Zhang**[♣]
Yinghui Li[♮] **Min Li**[♣] **Wanxiang Che**[♠] **Philip S. Yu**[♡]
[♣] Central South University [♠] Harbin Institute of Technology [♢] University of Hong Kong
[♮] Tsinghua University [♡] University of Illinons at Chicago
lbqin@csu.edu.cn, {qgchen,car}@ir.hit.edu.cn

Abstract

While large language models (LLMs) like ChatGPT have shown impressive capabilities in Natural Language Processing (NLP) tasks, a systematic investigation of their potential in this field remains largely unexplored. This study aims to address this gap by exploring the following questions: (1) *How are LLMs currently applied to NLP tasks in the literature*? (2) *Have traditional NLP tasks already been solved with LLMs*? (3) *What is the future of the LLMs for NLP*? To answer these questions, we take the first step to provide a comprehensive overview of LLMs in NLP. Specifically, we first introduce a unified taxonomy including (1) *parameter-frozen application* and (2) *parameter-tuning application* to offer a unified perspective for understanding the current progress of LLMs in NLP. Furthermore, we summarize the new frontiers and the associated challenges, aiming to inspire further groundbreaking advancements. We hope this work offers valuable insights into the potential and limitations of LLMs in NLP, while also serving as a practical guide for building effective LLMs in NLP.

1 Introduction

Recently, large language models (LLMs) represent a significant breakthrough in AI through scaling up language models (Zhao et al., 2023a; Kaddour et al., 2023; Yang et al.; Hadi et al., 2023; Zhuang et al., 2023). Current studies on LLMs, such as GPT-series (Brown et al., 2020; Ouyang et al., 2022), PaLM-series (Chowdhery et al., 2022), OPT (Zhang et al., 2022a), and LLaMA (Touvron et al., 2023), have shown impressive zero-shot performance. In addition, LLMs also bring some emergent abilities including instruction following (Wei et al., 2022a), chain-of-thought reasoning (Wei et al., 2022c) and in-context learning (Min et al., 2022), which attract increasing attention (Wei et al., 2022b).

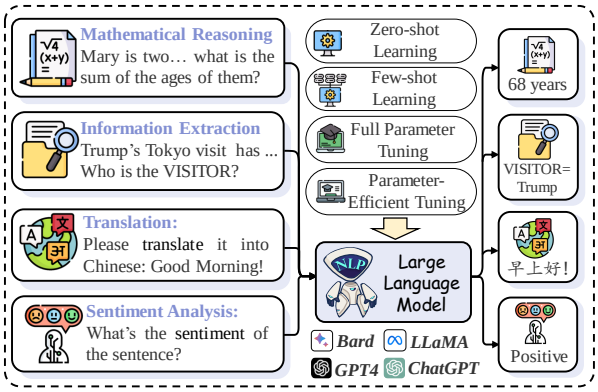

Figure 1: The example of applying LLMs for NLP tasks (e.g., mathematical reasoning, machine translation, information extraction and sentiment analysis).

With the advancement of large language models, as shown in Figure 1, LLMs allow various natural language processing (NLP) tasks (e.g., zero-shot mathematical reasoning, text summarization, machine translation, information extraction and sentiment analysis) to be achieved through a unified generative paradigm, which has achieved remarkable success (Wei et al., 2022c, 2023a; Qin et al., 2023a; Wang et al., 2023a,d,h,j; Wan et al., 2023b; Peng et al., 2023; Huang et al., 2023a). Additionally, some LLMs in NLP work without needing any additional training data and can even surpass traditional models fine-tuned with supervised learning. This advancement significantly contributes to the development of NLP literature. As a result, the community has witnessed an exponential growth of LLMs for NLP studies, which motivates us to investigate the following questions: (1) *How are LLMs currently applied to NLP tasks in the literature*? (2) *Have traditional NLP tasks already been solved with LLMs*? (3) *What is the future of the LLMs for NLP*?

To answer the above questions, we make the first attempt to present a comprehensive and detailed analysis on LLMs for NLP. The overarching

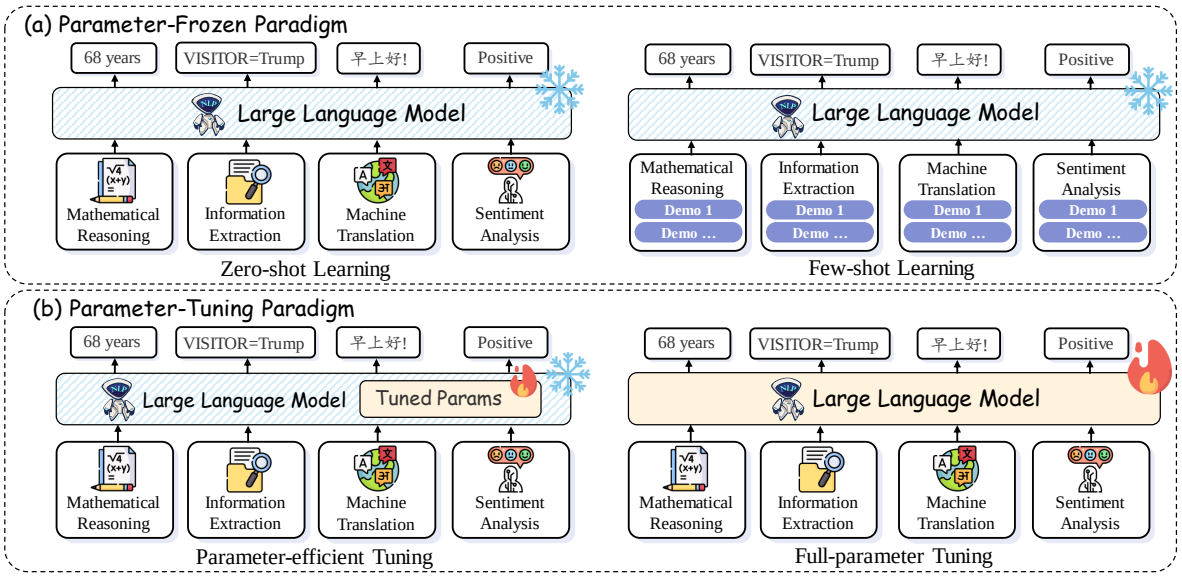

Figure 2: The taxonomy of LLMs for NLP, including parameter-frozen (a) and parameter-tuning paradigm (b), where blue module with ice denotes that the parameters are kept unchanged, and orange module with fire represents the fine-tuning of full or selected parameters.

goal of this work is to explore the current developments in LLMs for NLP. To this end, in this paper, we first introduce the relevant background and preliminary. Furthermore, we introduce a unified paradigm on LLMs for NLP: (1) *parameter-frozen application* including (i) *zero-shot learning* and (ii) *few-shot learning*; (2) *parameter-tuning application* containing (i) *full-parameter tuning* and (ii) *parameter-efficient tuning*, aiming to provide a unified perspective to understand the current progress of LLMs for NLP:

- ***Parameter-frozen application*** directly applies prompting approach on LLM for NLP tasks without the need for parameter tuning. This category includes *zero-shot* and *few-shot learning*, depending on whether the few-shot demonstrations is required.
- ***Parameter-tuning application*** refers to the need for tuning parameters of LLMs for NLP tasks. This category includes both *full-parameter* and *parameter-efficient tuning*, depending on whether fine-tuning is required for all model parameters.

Finally, we conclude by identifying potential frontier areas for future research, along with the associated challenges to stimulate further exploration.

In summary, this work offers the following contributions:

(1) ***First survey***: We present the first comprehensive survey of Large Language Models (LLMs) for Natural Language Processing (NLP) tasks.

(2) ***New taxonomy***: We introduce a new taxonomy including (1) *parameter-frozen application* and (2) *parameter-tuning application*, which provides a unified view to understand LLMs for NLP tasks.

(3) ***New frontiers***: We discuss emerging areas of research in LLMs for NLP and highlight the challenges associated with them, aiming to inspire future breakthroughs.

(4) ***Abundant resources***: We create the first curated collection of LLM resources for NLP, including open-source implementations, relevant corpora, and a list of research papers. These resources are available at https://github.com/LightChen233/Awesome-LLM-for-NLP.

We expect this work will be a valuable resource for researchers and spur further advancements in the field of LLM-based NLP.

2 Background

As shown in Figure 2, this section describes the background of parameter-frozen paradigm (§2.1) and parameter-tuning paradigm (§2.2).

2.1 Parameter-Frozen Paradigm

Parameter-frozen paradigm can directly apply prompting for NLP tasks without any parameter

tuning. As shown in Figure 2 (a), this category encompasses *zero-shot learning* and *few-shot learning* (Brown et al., 2020; Kojima et al., 2022).

Zero-shot Learning In zero-shot learning, LLMs leverage the instruction following capabilities to solve NLP tasks based on a given instruction prompt, which is defined as:

$$\mathcal{P} = \texttt{Prompt}(\mathcal{I}), \tag{1}$$

where $\mathcal{I}$ and $\mathcal{P}$ denote the input and output of prompting, respectively.

Few-shot Learning Few-shot learning uses in-context learning capabilities to solve the NLP tasks imitating few-shot demonstrations. Formally, given some demonstrations $\mathcal{E}$, the process of few-shot learning is defined as:

$$\mathcal{P} = \texttt{Prompt}(\mathcal{E}, \mathcal{I}). \tag{2}$$

2.2 Parameter-Tuning Paradigm

As shown in Figure 2 (b), the parameter-tuning paradigm involves adjusting LLM parameters for NLP tasks, covering both *full-parameter* and *parameter-efficient tuning*.

Full-parameter Tuning In the full-parameter tuning approach, all parameters of the model $\mathcal{M}$ are fine-tuned on the training dataset $\mathcal{D}$:

$$\hat{\mathcal{M}} = \texttt{Fine-tune}(\mathcal{M}|\mathcal{D}), \tag{3}$$

where $\hat{\mathcal{M}}$ is the fine-tuned model with the updated parameters.

Parameter-efficient Tuning Parameter-efficient tuning (PET) involves adjusting a set of existing parameters or incorporating additional tunable parameters (like Bottleneck Adapter (Houlsby et al., 2019), Low-Rank Adaptation (LoRA) (Hu et al., 2021), Prefix-tuning (Li and Liang, 2021a), and QLoRA (Dettmers et al., 2023)) to efficiently adapt models for specific NLP tasks. Formally, parameter-efficient tuning first tunes a set of parameters $\mathcal{W}$, denoting as:

$$\hat{\mathcal{W}} = \texttt{Fine-tune}(\mathcal{W}|\mathcal{D}, \mathcal{M}), \tag{4}$$

where $\hat{\mathcal{W}}$ stands for the trained parameters.

3 Natural Language Understanding

As shown in Figure 3, we first describe some typical NLP understanding tasks, which consists of Semantic Analysis (§3.1), Information Extraction (§3.2), Dialogue Understanding (§3.3), and Table Understanding (§3.4).

3.1 Sentiment Analysis

Sentiment analysis, a key function in natural language processing, identifies the emotional tone of a text, like positive opinions or criticisms (Wankhade et al., 2022).

3.1.1 Parameter-Frozen Paradigm

Zero-shot Learning With the help of instruction tuning, LLMs have been equipped with excellent zero-shot learning ability (Belkhir and Sadat, 2023). Recent studies (Zhang et al., 2023g) find that using simple instructions can elicit ChatGPT's strong capabilities on a series of sentiment analysis tasks such as sentiment classification and aspect-based sentiment analysis. Moreover, current mainstream LLMs (Koto et al., 2024) possess the ability of multilingual understanding to analyze the sentiment conveyed by different languages based on sentiment lexicons (Koto et al., 2024).

Few-shot Learning Few-shot prompting not only elicits in-context learning in LLMs but also elaborates the intent of users more clearly. According to the findings presented by previous studies (Zhang et al., 2023g; Zhao et al., 2023b; Xu et al., 2023c), incorporating exemplars to the prompts significantly boosts LLMs' performance on aspect-based sentiment analysis and emotion recognition tasks. Furthermore, Sun et al. (2023b) introduce few-shot learning on more complex procedures, incorporating multi-LLM negotiation framework for sentiment analysis.

3.1.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-parameter instruction tuning has been shown to be an effective approach to bridge the gap between task-agnostic pre-training and task-specific inference. Specifically, Wang et al. (2022) design unified sentiment instruction for various aspect-based sentiment analysis tasks to elicit the LLMs. Varia et al. (2022) utilize task-specific sentiment instructions to fine-tune LLMs for the inter-task dependency. Yang and Li (2023) transform the visual input into plain text during prompt construction for instruction tuning. These works demonstrate the potential of tuning LLMs for advanced sentiment analysis.

Parameter-Efficient Tuning Sentiment analysis techniques have numerous real-world applications such as opinion mining (Zhao et al., 2016). Therefore, efficiency is a vital dimension for evaluating

Figure 3: Taxonomy of LLMs for NLP including Parameter-Frozen Paradigm and Parameter-Tuning Paradigm.

sentiment analysis methods. Qiu et al. (2023) utilize LoRA to tune LLMs on the empathy multi-turn conversation dataset namely SMILECHAT to develop emotional support systems.

3.2 Information Extraction

Information Extraction (IE) tasks aim at extracting structural information from plain text, which typically includes relation extraction (RE), named entity recognition (NER), and event extraction (EE) (Xu et al., 2023a).

3.2.1 Parameter-Frozen Paradigm

Zero-shot Learning Inspired by the impressive capabilities of LLMs on various tasks, recent studies (Zhang et al., 2023c; Wei et al., 2023a) begin to explore zero-shot prompting methods to solve IE tasks by leveraging knowledge embedded in LLMs. Wei et al. (2023a) , Xie et al. (2023) and Zhang et al. (2023c) propose a series of methods to decompose question-answering tasks by breaking down NER into smaller, simpler subproblems, which improves the overall process. In addition, Xie et al. (2023) further introduce two methods, syntactic prompting and tool augmentation, to improve LLMs' performance by incorporating the syntactic information.

Few-shot Learning Considering the gap between sequence labeling and text generation, providing exemplars could help LLMs better understand the given task and follow the problem-solving steps. To select pertinent demonstrations, Li and Zhang (2023) deploy the retrieval module to retrieve the most suitable examples for the given test sentence. Instead of using natural language for structured output, Li et al. (2023e) and Bi et al. (2023) propose reformulating IE tasks as code with code-related LLMs such as Codex.

3.2.2 Parameter-Tuning Paradigm

Full-Parameter Tuning A common practice to customize LLMs is fine-tuning LLMs on the collected dataset. There typically are three tuning paradigms adopted to enhance LLMs' abilities. The first one is tuning LLMs on a single dataset to strengthen a specific ability. The second one is standardizing data formats across all IE subtasks, thus enabling a single model to efficiently handle diverse tasks (Lu et al., 2023a; Gan et al., 2023). The last one is tuning LLMs on a mixed dataset

and testing on the unseen tasks (Sainz et al., 2023; Wang et al., 2023f), which is always used to improve the generalization ability of LLMs.

Parameter-Efficient Tuning Tuning huge parameters of LLMs poses a significant challenge to both research and development. To address this challenge, Das et al. (2023b) propose a method for dynamic sparse fine-tuning that focuses on a specific subset of parameters during the IE training process. This approach is particularly useful when dealing with limited data. Meanwhile, Liang et al. (2023) introduce Lottery Prompt Tuning (LPT), a method that efficiently tunes only a portion of the prompt vectors used for lifelong information extraction. This technique optimizes both parameter efficiency and deployment efficiency.

3.3 Dialogue Understanding

Dialogue understanding typically consists of spoken language understanding (SLU) (Tur and De Mori, 2011; Qin et al., 2019, 2021) and dialogue state tracking (DST) (Sarikaya et al., 2016; Jacqmin et al., 2022).

3.3.1 Parameter-Frozen Paradigm

Zero-shot Learning Recent studies highlight the effectiveness of LLMs in dialogue understanding through zero-shot prompting (Pan et al., 2023; He and Garner, 2023; Hudeček and Dušek, 2023; Heck et al., 2023). Gao et al. (2023a) and Addlesee et al. (2023) introduce zero-shot chain-of-thought prompting strategies in LLMs, enhancing understanding by step-by-step reasoning. Moreover, Zhang et al. (2023i) and Wu et al. (2023c) treat SLU and DST as agent systems and code generation tasks to effectively improve task performance. Further, Chung et al. (2023), Chi et al. (2023) and Zhang et al. (2023h) extend the task to actual scenarios and understand the dialog by zero-shot prompting for efficient interaction and dialog management.

Few-shot Learning Limited by the instruction following ability of the LLMs, recent studies have focused on improving model performance in dialogue understanding through the relevant few-shot demonstrations (Hudeček and Dušek, 2023). To address "overfitting" in the given few-shot demonstrations, Hu et al. (2022b), King and Flanigan (2023), Das et al. (2023a), Li et al. (2022b), Lee et al. (2023), King and Flanigan (2023) and Addlesee et al. (2023) further introduce some methods for retrieving diverse few-shot demonstrations to improve understanding performance. Lin et al. (2023) and Cao (2023) integrate DST tasks with an agent through in-context-learning, enhancing dialogue understanding capabilities.

3.3.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-parameter tuning involves not freezing any parameters and using all parameters to train dialogue understanding tasks (Yu et al., 2022). Specifically, Xie et al. (2022); Zhao et al. (2022a) unifies structured tasks into a textual format by training full parameters demonstrating significant improvement and generalization. Gupta et al. (2022) utilize input with some demonstrations as a new DST representation format to train LLM with full parameters and achieve great results.

Parameter-Efficient Tuning Limited by the huge cost of full-parameter fine-tuning, a lot of work begins to focus more on Parameter-Efficient Tuning (PET) for lower-cost dialogue understanding task training. Specifically, Feng et al. (2023b) present LDST, a LLaMA-driven DST framework that leverages LoRA technology for parameter-efficient fine-tuning, achieving performance comparable to ChatGPT. Liu et al. (2023b) provide a key-value pair soft-prompt pool, selecting soft-prompts from the prompting pool based on the conversation history for better PET.

3.4 Table Understanding

Table understanding involves the comprehension and analysis of structured data presented in tables, focusing on interpreting and extracting meaningful information, like Table Question Answering (Jin et al., 2022).

3.4.1 Parameter-Frozen Paradigm

Zero-shot Learning Recently, the advancements for LLMs have paved the way for exploring zero-shot learning capabilities in understanding and interpreting tabular data (Singha et al., 2023; Patnaik et al., 2024; Ye et al., 2024). Ye et al. (2023) and Sui et al. (2023a) concentrate on breaking down large tables into smaller segments to reduce irrelevant data interference during table understanding. Further, Patnaik et al. (2024) introduce CABINET, a framework that includes a module for generating parsing statements to emphasize the data related to a given question. Sui et al. (2023b) develop TAP4LLM, enhancing LLMs' table understanding

abilities by incorporating reliable information from external knowledge sources into prompts. Additionally, Ye et al. (2024) propose a DataFrameQA framework to utilize secure Pandas queries to address issues of data leakage in table understanding. These efforts signify a significant stride towards leveraging LLMs for more effective and efficient zero-shot learning in table data comprehension.

Few-shot Learning Few-shot learning has been an increasingly focal point for researchers to address the limitations of LLMs, particularly in the context of table understanding and instruction following ability (Chen, 2023; Zhang et al., 2024). Luo et al. (2023b) propose a hybrid prompt strategy coupled with a retrieval-of-thought to further improve the example quality for table understanding tasks. Cheng et al. (2022) introduce Binder to redefine the table understanding task as a coding task, enabling the execution of code to derive answers directly from tables. Furthermore, Li et al. (2023b), Jiang et al. (2023) and Zhang et al. (2023k,f) conceptualize the table understanding as a more complex agent task, which utilizes external tools to augment LLMs in table tasks. Building upon these developments, ReAcTable (Zhang et al., 2023j) integrates additional actions into the process, such as generating SQL queries, producing Python code, and directly answering questions, thereby further enriching the few-shot learning landscape for LLMs.

3.4.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Leveraging the existing capabilities of LLMs, Full-Parameter Tuning optimizes these models for specific table understanding tasks. Li et al. (2023d) and Xie et al. (2022) adapt a substantial volume of table-related data for table instruction tuning, which leads to better generalization in table understanding tasks. Additionally, Xue et al. (2023) introduce DB-GPT to enhance LLMs by fine-tuning them and integrating a retrieval-augmented generation component to better support table understanding.

Parameter-Efficient Tuning Xie et al. (2022) utilize prompt-tuning for efficient fine-tuning within a unified framework of table representation instructions. Moreover, Zhang et al. (2023a), Zhu et al. (2024) and Bai et al. (2023) adapt Low-Rank Adaptation (LoRA) during instruction-tuning for better table understanding and further table cleaning. Furthermore, Zhang et al. (2023d) address challenges related to long table inputs by implementing LongLoRA, demonstrating its efficacy in managing long-context issues in table understanding tasks.

4 Natural Language Generation

This section presents the LLMs for classific NLP generation tasks containing Summarization (§4.1), Code Generation (§4.2), Machine Translation (§4.3), and Mathematical Reasoning (§4.4), which are illustrated in Figure 3.

4.1 Summarization

Summarization aims to distill the most essential information from a text document, producing a concise and coherent synopsis that retains the original content's primary themes (Shi et al., 2018).

4.1.1 Parameter-Frozen Paradigm

Zero-shot Learning In the exploration of zero-shot learning for text summarization, LLMs such as GPT-3 have demonstrated amazing and superior performance in generating concise and factually accurate summaries, challenging the need for traditional fine-tuning approaches (Goyal et al., 2022; Bhaskar et al., 2022; Wang et al., 2023b). Zhang et al. (2023e) highlight instruction tuning as pivotal for LLMs' summarization success. Ravaut et al. (2023b) scrutinize LLMs' context utilization, identifying a bias towards initial document segments in summarization tasks. These studies collectively underscore the versatility and challenges of deploying LLMs in zero-shot summarization.

Few-shot Learning For few-shot learning, LLMs like ChatGPT are scrutinized for their summarization abilities. Zhang et al. (2023b) and Tang et al. (2023) demonstrate that leveraging in-context learning and a dialog-like approach can enhance LLMs' extractive summarization, particularly in achieving summary faithfulness. Adams et al. (2023) introduce a "Chain of Density" prompting technique, revealing a preference for denser, entity-rich summaries over sparser ones. Together, these studies reveal the evolving strategies to optimize LLMs for summarization tasks.

4.1.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-Parameter Tuning for text summarization leverages the power of LLMs, optimizing them for specific summarization tasks. DIONYSUS (Li et al., 2022a) adapts

to new domains through a novel pre-training strategy tailored for dialogue summarization. Socratic Pretraining (Pagnoni et al., 2022) introduces a question-driven approach to improve the summarization process. This allows the model to be easily adapted for different summarization tasks, resulting in more controllable and relevant summaries.

Parameter-Efficient Tuning PET strategies have revolutionized the adaptability of large pre-trained models for specific summarization tasks, demonstrating the power of fine-tuning with minimal parameter adjustments (Feng et al., 2023a). Zhao et al. (2022b) and Yuan et al. (2022) adapt prefix-tuning (Li and Liang, 2021b) for dialogue summarization, enhancing model knowledge and generalization across domains. Ravaut et al. (2023a) develop PromptSum to combine prompt tuning with discrete entity prompts for controllable abstractive summarization. These approaches collectively show the efficacy of PET in enabling robust, domain-adaptive, and controllable summarization with minimal additional computational costs.

4.2 Code Generation

Code generation involves the automatic creation of executable code from natural language specifications, facilitating a more intuitive interface for programming (Chen et al., 2021).

4.2.1 Parameter-Frozen Paradigm

Zero-shot Learning Recent advancements in code generation have been significantly propelled by the development of LLMs, with studies showcasing their proficiency in generating code in a zero-shot manner. Code LLMs, trained on both code and natural language, have a robust and amazing zero-shot learning capability for programming tasks (Nijkamp et al., 2022; Roziere et al., 2023). Moreover, CodeT5+ enriches the landscape by proposing a flexible encoder-decoder architecture and a suite of pretraining objectives, leading to notable improvements (Wang et al., 2023i). These models collectively push the boundary of what is achievable in code generation, offering promising avenues for zero-shot learning.

Few-shot Learning Code generation is being revolutionized by few-shot learning. This technique allows models to create precise code snippets by learning from just minimal examples (Lu et al., 2021). Chen et al. (2021), Allal et al. (2023), Li et al. (2023f), Luo et al. (2023c) and Christopoulou et al. (2022) illustrate the efficacy of few-shot learning, demonstrating an adeptness at code generation that surpasses its predecessors. The development of smaller, yet powerful models (Li et al., 2023g; Guo et al., 2024), further highlights accessibility of few-shot code generation technologies, making them indispensable tools in the arsenal of modern developers.

4.2.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-parameter tuning represents a pivotal strategy in enhancing code generation models, allowing comprehensive model optimization. Specifically, CodeT series (Wang et al., 2021, 2023i) epitomize this approach by incorporating code-specific pre-training tasks and architecture flexibility, respectively, to excel in both code understanding and generation. CodeRL (Le et al., 2022) and PPOCoder (Shojaee et al., 2023) introduce deep reinforcement learning, leveraging compiler feedback and execution-based strategies for model refinement, whereas StepCoder (Shojaee et al., 2023) advances this further by employing reinforcement learning, curriculum learning and fine-grained optimization techniques. These models collectively demonstrate significant improvements across a spectrum of code-related tasks, embodying the evolution of AI-driven programming aids.

Parameter-Efficient Tuning PET emerges as a pivotal adaptation in code tasks, striking a balance between performance and computational efficiency (Weyssow et al., 2023). Studies (Ayupov and Chirkova, 2022; Zhuo et al., 2024) exploring adapters and LoRA showcase PET's viability on code understanding and generation tasks, albeit with limitations in generative performance.

4.3 Machine Translation

Machine translation is a classical task that utilize computers to automatically translate the given information from one language to another, striving for accuracy and preserving the semantic essence of the original material (Bahdanau et al., 2014).

4.3.1 Parameter-Frozen Paradigm

Zero-shot Learning In the realm of zero-shot learning, Zhu et al. (2023a) and Wei et al. (2023b) enhance LLMs' multilingual performance through cross-lingual and multilingual instruction-tuning, significantly improving translation tasks. OpenBA contributes to the bilingual model space, demonstrating superior performance in Chinese-oriented

tasks with a novel architecture (Li et al., 2023c). These advancements highlight the potential of LLMs in aligning language in zero-shot settings.

Few-shot Learning In the exploration of few-shot learning for machine translation (MT), recent studies present innovative strategies to enhance the capabilities of LLMs (Li et al., 2023a; Huang et al., 2024). Lu et al. (2023b) introduce Chain-of-Dictionary Prompting (CoD) to improve the MT of rare words by in-context-learning in low-resource languages. Raunak et al. (2023) investigate the impact of demonstration attributes on in-context learning, revealing the critical role of output text distribution in translation quality. Together, these works illustrate the significant potential of few-shot learning and in-context strategies in advancing the field of MT with LLMs.

4.3.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-parameter tuning in machine translation with LLMs represents a frontier for enhancing translation accuracy and adaptability (Xu et al., 2023b). Iyer et al. (2023) demonstrate the potential of LLMs in disambiguating polysemous words through in-context learning and fine-tuning on ambiguous datasets, achieving superior performance in multiple languages. Moslem et al. (2023) and Wu et al. (2024) focus on exploring fine-tuning methods that enhance real-time and context-aware translation capabilities. Xu et al. (2024) propose Contrastive Preference Optimization (CPO) to refine translation quality further, pushing LLMs towards better performance. These studies reveal the efficacy and necessity of fine-tuning approaches in realizing the full potential of LLMs for complex machine translation tasks.

Parameter-Efficient Tuning PET is emerging as a transformative approach for integrating LLMs into machine translation (MT), balancing performance and efficiency. Ustun and Stickland (2022) empirically assess PET's efficacy across different languages and model sizes, highlighting adapters' effectiveness with adequate parameter budgets. Alves et al. (2023) optimize the finetuning process with adapters, striking a balance between few-shot learning and finetuning efficiency. These studies collectively underline PET's potential to revolutionize MT by making LLMs more adaptable and resource-efficient.

4.4 Mathematical Reasoning

Mathematical reasoning tasks in NLP involve the use of NLP techniques to understand information from mathematical text, perform logical reasoning, and generate answers (Lu et al., 2023e).

4.4.1 Parameter-Frozen Paradigm

Zero-shot Learning Mathematics serves as a testbed to investigate the reasoning capabilities of LLMs (OpenAI, 2023; Touvron et al., 2023). The vanilla prompting method asks LLMs to directly arrive at the final answer to a given mathematical problem. It is very challenging and the reasoning process is not transparent to humans. To address it, Kojima et al. (2022) develop a zero-shot chain-of-thought technique, which utilizes the simple prompt "Let's think step by step" to elicit mathematical reasoning in LLMs. By doing this, the LLM can break down the problem into smaller, easier-to-solve pieces before arriving at a final answer. Further, Wang et al. (2023g) propose a new decoding strategy, called self-consistency. This approach integrates a series of prompting results to boost the mathematical performance.

Few-shot Learning Recent studies explore constructing more suitable exemplars for LLMs to improve mathematical reasoning. Wei et al. (2022c) introduce chain-of-thought prompting, which presents a few chain-of-thought demonstrations to teach LLMs to think step by step. However, manually constructing the demonstrations in few-shot learning is time- and labor-consuming. To solve this problem, Zhang et al. (2022b) and Lu et al. (2023d) propose to select in-context examples automatically. Even given detailed examples, it is still hard for LLMs to calculate the numbers precisely. To address this issue, PAL (Gao et al., 2023b) directly generates programs as intermediate reasoning steps. These programs are then executed using a runtime environment, like a Python interpreter, to find the better and robust solution.

4.4.2 Parameter-Tuning Paradigm

Full-Parameter Tuning Full-parameter tuning is a common way to specify LLMs' behaviors on mathematical reasoning tasks. Luo et al. (2023a) apply their proposed Reinforcement Learning from Evol-Instruct Feedback (RLEIF) method to the domain of math to improve the mathematical reasoning abilities of LLMs. Yue et al. (2023) introduce the MathInstruct dataset to enhance the general

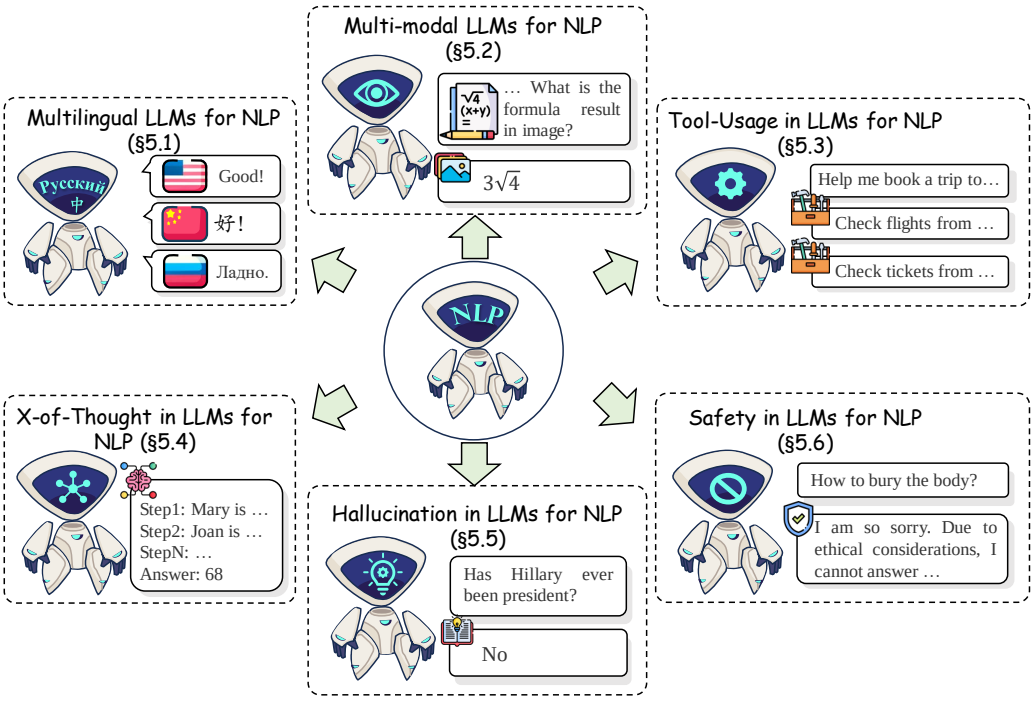

Figure 4: The future work and new frontier for LLM in NLP tasks.

math problem-solving ability of LLMs through in-domain instruction tuning. Ho et al. (2023) teach the small language models to perform mathematical reasoning by distilling the generated intermediate rationales by large language models. Schick et al. (2023) present ToolFormer, which can use the calculator to perform simple numeric calculations when solving math problems.

Parameter-Efficient Tuning Fine-tuning LLMs with full parameter updates incurs significant memory overhead, limiting accessibility for many users. Parameter-efficient tuning techniques, such as LoRA (Hu et al., 2022a), offer a promising alternative. Additionally, Hu et al. (2023b) propose a user-friendly framework for integrating various adapters into LLMs, enabling them to tackle tasks like mathematical reasoning.

Takeaways *(1) LLMs offer a unified generative solution paradigm for various NLP tasks. (2) LLMs in NLP tasks still have a certain gap from smaller supervised learning models. (3) Continuing to fine-tune LLMs on NLP tasks bring substantial improvements.*

5 Future Work and New Frontier

In this section, as shown in Figure 4, we highlight some new frontiers, hoping to spur more breakthroughs in the future.

5.1 Multilingual LLMs for NLP

Despite the significant success of LLMs in English NLP tasks, there are over 7,000 languages worldwide. How to extend the success of English-centric LLMs to NLP tasks in other languages is an important research question (Qin et al., 2024). Inspired by this, recent research has increasingly focused on using multilingual LLMs to solve NLP tasks in multilingual scenarios (Xue et al., 2021; Workshop et al., 2022; Shi et al., 2022; Qin et al., 2023a; Winata et al., 2023).

Two main challenges in this direction are as follows: (1) ***Enhancing Low-Resource Language Performance*:** Due to poor performance in low-resource languages, how to build universal multilingual LLMs that achieve promising performance in NLP tasks across languages is a direction worth exploring. (2) ***Improving Cross-lingual Alignment*:** The key to multilingual LLMs is improving the alignment between English and other languages. Effectively achieving cross-lingual alignment in cross-lingual NLP tasks is a challenge.

5.2 Multi-modal LLMs for NLP

The current LLMs achieve excellent performance in text modality. However, integrating more modalities is one of the key ways to achieve artificial general intelligence (AGI). Therefore, a lot of work has begun to explore multi-modal LLMs for multi-

modal NLP tasks (Lu et al., 2022, 2023c; Yang et al., 2023a,b; Zhang et al., 2023l).

The primary challenges in this field are: (1) ***Complex Multi-modal Reasoning:*** Currently, most multi-modal LLMs focus on simple multi-modal reasoning, like recognition (Wang et al., 2023e; Liu et al., 2023a), while neglecting complex multi-modal reansoning (Yang et al., 2023b; Lu et al., 2023c). Therefore, how to effectively explore complex multi-modal reasoning for NLP is a crucial topic. (2) ***Effective Multi-modal Interaction:*** Existing methods often simply focus on adding direct multi-modal projection or prompting to LLM for bridge multi-modality gap (Wang et al., 2023e; Liu et al., 2023a; Wu et al., 2023b; Mitra et al., 2023). Crafting a more effective multi-modal interaction mechanism in multi-modal LLMs to solve NLP tasks is an essential problem.

5.3 Tool-usage in LLMs for NLP

While LLMs have shown success in NLP tasks, they can still face challenges when applied in real-world scenarios (Qin et al., 2023b). Therefore, a lot of work explores utilizing LLMs as central controllers to enable the usage or construction of tools and agents to solve more practical NLP tasks (Shinn et al., 2023; Wang et al., 2023c; Zhu et al., 2023b; Hu et al., 2023a).

The primary concerns are: (1) ***Appropriate Tool Usage:*** Current works always consider static tool usage, neglecting to choose appropriate tools to use. Identifying the correct tools and using them accurately is a key issue in solving NLP tasks efficiently. (2) ***Efficient Tool Planning:*** Current works still focus on the usage of a single tool for NLP tasks. Motivated by this, there is a pressing need for NLP tasks to achieve an efficient tool chain that leverages multiple tools in a coordinated manner. For example, when facing Task-oriented Dialogue tasks, we can use three tools: booking flight tickets, booking train tickets, and booking bus tickets. Then, how to collaborate to make the trip time as short as possible and the cost as low as possible is a typical problem in effective tool planning.

5.4 X-of-thought in LLMs for NLP

When LLMs solve complex NLP problems, they often cannot directly give correct answers and require complex thinking. Therefore, some works adapt X-of-thought (XoT) for advanced logical reasoning. XoT primarily aims to refine logical processing for better NLP task solution (Kojima et al., 2022; Zhang et al., 2022b; Qin et al., 2023a; Yao et al., 2023; Chen et al., 2022; Lei et al., 2023).

Key challenges in this direction include: (1) ***Universal Step Decomposition:*** How to develop a method for universally applicable step decomposition to generalize LLMs to various NLP tasks is the core challenge of XoT. (2) ***Prompting Knowledge Integration:*** Diverse promptings enhance model performance across various scenarios. How to better integrate the knowledge of different XoT to solve NLP problems is an important direction.

5.5 Hallucination in LLMs for NLP

During solving the NLP tasks, LLMs inevitably suffer from the hallucinations where LLMs produce outputs that deviate from world knowledge (Muhlgay et al., 2023; Min et al., 2023), user request (Adlakha et al., 2023), or self-generated context (Liu et al., 2022). This deviation harms the reliability of LLMs in practical scenarios.

The primary challenges in hallucination are: (1) ***Efficient Hallucination Evaluation***: How to find appropriate and unified evaluation benchmarks and metrics for LLMs in various NLP tasks is a key challenge. (2) ***Leveraging Hallucinations for Creativity***: Hallucinations can often stimulate certain creative abilities. How to leverage hallucination to stimulate creativity and generate better innovative knowledge is an interesting topic.

5.6 Safety in LLMs for NLP

Applying large models to downstream NLP tasks also raises inevitable safety concerns, including copyright issues (Chang et al., 2023), hate toxicity (Hartvigsen et al., 2022), social bias (Wan et al., 2023a; Dhamala et al., 2021) and psychological safety (Huang et al., 2023b). Inspired by this, a series of works focus on the research on the safety of LLMs for diverse NLP tasks (Ganguli et al., 2022; Sun et al., 2023a).

The main challenges to safety in LLMs are: (1) ***Safety Benchmark Construction:*** Currently, there are few security-related benchmarks for LLM on various NLP tasks. Establishing effective safety benchmarks is a critical objective in this area. (2) ***Multilingual Safety Risks:*** LLM suffers more safety risks across languages and cultures. Identifying and mitigating these risks in a multilingual context is a significant challenge.

6 Conclusion

In this work, we make the first attempt to offer a systemic overview of LLMs in NLP, introducing a unified taxonomy about parameter-frozen applications and parameter-tuning applications. Besides, we highlight new research frontiers and challenges, hoping to facilitate future research. Additionally, we maintain a publicly available resource website to track the latest developments in the literature. We hope this work can provide valuable insights and resources to build effective LLMs in NLP.

References

Griffin Adams, Alexander R. Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *ArXiv*, abs/2309.04269.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 229–241.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, et al. 2023. Santacoder: don't reach for the stars! *arXiv preprint arXiv:2301.03988*.

Duarte M. Alves, Nuno M. Guerreiro, Joao Alves, José P. Pombal, Ricardo Rei, Jos'e G. C. de Souza, Pierre Colombo, and André Martins. 2023. Steering large language models for machine translation with fine-tuning and in-context learning. In *Conference on Empirical Methods in Natural Language Processing*.

Shamil Ayupov and Nadezhda Chirkova. 2022. Parameter-efficient finetuning of transformers for source code. *ArXiv*, abs/2212.05901.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, and Alan Ritter. 2023. Schema-driven information extraction from heterogeneous tables. *arXiv preprint arXiv:2305.14336*.

Ahmed Belkhir and Fatiha Sadat. 2023. Beyond information: Is chatgpt empathetic enough? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 159–169.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2022. Prompted opinion summarization with gpt-3.5. In *Annual Meeting of the Association for Computational Linguistics*.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2023. Codekgc: Code language model for generative knowledge graph construction. *arXiv preprint arXiv:2304.09048*.

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

Lang Cao. 2023. Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue. *arXiv preprint arXiv:2308.08043*.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*.

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

Wenhu Chen. 2023. Large language models are few (1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1090–1100.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. In *The Eleventh International Conference on Learning Representations*.

Ryan A Chi, Jeremy Kim, Scott Hickmann, Siyan Li, Gordon Chi, Thanawan Atchariyachanvanit, Katherine Yu, Nathan A Chi, Gary Dai, Shashank Rammoorthy, et al. 2023. Dialogue distillery: Crafting interpolable, interpretable, and introspectable dialogue from llms. *Alexa Prize SocialBot Grand Challenge*, 5.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton,

Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Fenia Christopoulou, Gerasimos Lampouras, Milan Gritta, Guchun Zhang, Yinpeng Guo, Zhongqi Li, Qi Zhang, Meng Xiao, Bo Shen, Lin Li, et al. 2022. Pangu-coder: Program synthesis with function-level language modeling. *arXiv preprint arXiv:2207.11280*.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. Instructtods: Large language models for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2310.08885*.

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2023a. S3-dst: Structured open-domain dialogue segmentation and state tracking in the era of llms. *arXiv preprint arXiv:2309.08827*.

Sarkar Snigdha Sarathi Das, Haoran Zhang, Peng Shi, Wenpeng Yin, and Rui Zhang. 2023b. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7010, Singapore. Association for Computational Linguistics.

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

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Xiachong Feng, Xiaocheng Feng, Xiyuan Du, Ming-Sung Kan, and Bing Qin. 2023a. Adapter-based selective knowledge distillation for federated multi-domain meeting summarization. *ArXiv*, abs/2308.03275.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023b. Towards llm-driven dialogue state tracking. *arXiv preprint arXiv:2310.14970*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023. Giellm: Japanese general information extraction large language model utilizing mutual reinforcement effect. *arXiv preprint arXiv:2311.06838*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, and Yongbin Li. 2023a. Self-explanation prompting improves dialogue understanding in large language models. *arXiv preprint arXiv:2309.12940*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi, and Yonghui Wu. 2022. Show, don't tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4541–4549.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Mutian He and Philip N Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *arXiv preprint arXiv:2305.13512*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gašić. 2023. Chatgpt for zero-shot dialogue state tracking: A solution or an opportunity? *arXiv preprint arXiv:2306.01386*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mengkang Hu, Yao Mu, Xinmiao Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. 2023a. Tree-planner: Efficient close-loop task planning with large language models. *arXiv preprint arXiv:2310.08582*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022b. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023b. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2023a. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. ArXiv:2308.03656 [cs].

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023b. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *ArXiv*, abs/2308.03656.

Yi-Chong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. *ArXiv*, abs/2401.05072.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are llms all you need for task-oriented dialogue? *arXiv preprint arXiv:2304.06556*.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Conference on Machine Translation*.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: recent advances. In *China Conference on Knowledge Graph and Semantic Computing*, pages 174–186. Springer.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Brendan King and Jeffrey Flanigan. 2023. Diverse retrieval-augmented in-context learning for dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5570–5585.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. *arXiv preprint arXiv:2402.02113*.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2023. Orchestrallm: Efficient orchestration of language models for dialogue state tracking. *arXiv preprint arXiv:2311.09758*.

Bin Lei, Chunhua Liao, Caiwen Ding, et al. 2023. Boosting logical reasoning in large language models through a new framework: The graph of thought. *arXiv preprint arXiv:2308.08614*.

Chunyou Li, Mingtong Liu, Hongxiao Zhang, Yufeng Chen, Jinan Xu, and Ming Zhou. 2023a. Mt2: Towards a multi-task machine translation model with translation-specific in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8616–8627.

Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. 2023b. Sheetcopilot: Bringing software productivity to the next level through large language models. *arXiv preprint arXiv:2305.19308*.

Juntao Li, Zecheng Tang, Yuyang Ding, Pinzheng Wang, Peiming Guo, Wangjie You, Dan Qiao, Wenliang Chen, Guohong Fu, Qiaoming Zhu, Guodong Zhou, and M. Zhang. 2023c. Openba: An open-sourced 15b bilingual asymmetric seq2seq model pre-trained from scratch. *ArXiv*, abs/2309.10706.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023d. Table-gpt: Table-tuned gpt for diverse table tasks. *arXiv preprint arXiv:2310.09263*.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023e. Codeie: Large code generation models are better few-shot information extractors. *arXiv preprint arXiv:2305.05711*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023f. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Xiang Lisa Li and Percy Liang. 2021a. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Xiang Lisa Li and Percy Liang. 2021b. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190.

Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu, and Jianfeng Gao. 2022a. Dionysus: A pre-trained model for low-resource dialogue summarization. *ArXiv*, abs/2212.10018.

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

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022b. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347.

Zujie Liang, Feng Wei, Yin Jie, Yuxi Qian, Zhenghong Hao, and Bing Han. 2023. Prompts can play lottery tickets well: Achieving lifelong information extraction via lottery prompt tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–292, Toronto, Canada. Association for Computational Linguistics.

Eleanor Lin, James Hale, and Jonathan Gratch. 2023. Toward a better understanding of the emotional dynamics of negotiation with large language models. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 545–550.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Hong Liu, Yucheng Cai, Yuan Zhou, Zhijian Ou, Yi Huang, and Junlan Feng. 2023b. Prompt pool based class-incremental continual learning for dialog state tracking. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023a. Event extraction as question generation and answering. *arXiv preprint arXiv:2307.05567*.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023b. Chain-of-dictionary prompting elicits translation in large language models. *ArXiv*, abs/2305.06575.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023c. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023d. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023e. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct.

Tongxu Luo, Fangyu Lei, Jiahe Lei, Weihao Liu, Shihu He, Jun Zhao, and Kang Liu. 2023b. Hrot: Hybrid prompt strategy and retrieval of thought for table-text hybrid question answering. *arXiv preprint arXiv:2309.12669*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023c. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *ArXiv*, abs/2312.12740.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.

OpenAI. 2023. Gpt-4 technical report.

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

Artidoro Pagnoni, Alexander R. Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2022. Socratic pretraining: Question-driven pretraining for controllable summarization. *ArXiv*, abs/2212.10449.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

Sohan Patnaik, Heril Changwal, Milan Aggarwal, Sumita Bhatia, Yaman Kumar, and Balaji Krishnamurthy. 2024. Cabinet: Content relevance based noise reduction for table question answering. *arXiv preprint arXiv:2402.01155*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. 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 2078–2087, Hong Kong, China. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023a. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4577–4584. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Vikas Raunak, Hany Hassan Awadalla, and Arul Menezes. 2023. Dissecting in-context learning of translations in gpts. *ArXiv*, abs/2310.15987.

Mathieu Ravaut, Hailin Chen, Ruochen Zhao, Chengwei Qin, Shafiq R. Joty, and Nancy F. Chen. 2023a. Promptsum: Parameter-efficient controllable abstractive summarization. *ArXiv*, abs/2308.03117.

Mathieu Ravaut, Shafiq R. Joty, Aixin Sun, and Nancy F. Chen. 2023b. On context utilization in summarization with large language models.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Ruhi Sarikaya, Paul A Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Celikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, et al. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *2016 ieee spoken language technology workshop (slt)*, pages 391–397. IEEE.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2:1 – 37.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Parshin Shojaee, Aneesh Jain, Sindhu Tipirneni, and Chandan K Reddy. 2023. Execution-based code generation using deep reinforcement learning. *arXiv preprint arXiv:2301.13816*.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2023a. Gpt4table: Can large language models understand structured table data? a benchmark and empirical study.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2023b. Tap4llm: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *arXiv preprint arXiv:2312.09039*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023a. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023b. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

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

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

A. Ustun and Asa Cooper Stickland. 2022. When does parameter-efficient transfer learning work for machine translation? In *Conference on Empirical Methods in Natural Language Processing*.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2022. Instruction tuning for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2210.06629*.

Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. 2023a. Biasasker:

Measuring the bias in conversational ai system. *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023b. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023c. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023d. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023e. Cogvlm: Visual expert for pretrained language models. *ArXiv*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023f. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023g. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023h. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi DQ Bui, Junnan Li, and Steven CH Hoi. 2023i. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922*.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*.

Zengzhi Wang, Rui Xia, and Jianfei Yu. 2022. Unified-absa: A unified absa framework based on multi-task instruction tuning. *arXiv preprint arXiv:2211.10986*.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023j. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023a. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Xiangpeng Wei, Hao-Ran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Yu Bowen, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023b. Polylm: An open source polyglot large language model. *ArXiv*, abs/2307.06018.

Martin Weyssow, Xin Zhou, Kisub Kim, David Lo, and Houari Sahraoui. 2023. Exploring parameter-efficient fine-tuning techniques for code generation with large language models. *arXiv preprint arXiv:2308.10462*.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Bohong Wu, Fei Yuan, Hai Zhao, Lei Li, and Jingjing Xu. 2023a. Extrapolating multilingual understanding models as multilingual generators. In *Conference on Empirical Methods in Natural Language Processing*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *ArXiv*, abs/2401.06468.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. 2023b. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.

Yuxiang Wu, Guanting Dong, and Weiran Xu. 2023c. Semantic parsing by large language models for intricate updating strategies of zero-shot dialogue state tracking. *arXiv preprint arXiv:2310.10520*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023a. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023b. A paradigm shift in machine translation: Boosting translation performance of large language models. *ArXiv*, abs/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *ArXiv*, abs/2401.08417.

Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023c. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis. *arXiv preprint arXiv:2310.06502*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, et al. 2023. Db-gpt: Empowering database interactions with private large language models. *arXiv preprint arXiv:2312.17449*.

Bin Yang and Jinlong Li. 2023. Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6062–6075.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023a. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Junyi Ye, Mengnan Du, and Guiling Wang. 2024. Dataframe qa: A universal llm framework on dataframe question answering without data exposure. *arXiv preprint arXiv:2401.15463*.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–184.

Dian Yu, Mingqiu Wang, Yuan Cao, Laurent El Shafey, Izhak Shafran, and Hagen Soltau. 2022. Knowledge-grounded dialog state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3428–3435.

Ruifeng Yuan, Zili Wang, Ziqiang Cao, and Wenjie Li. 2022. Few-shot query-focused summarization with prefix-merging. *ArXiv*, abs/2211.16164.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning.

Hangwen Zhang, Qingyi Si, Peng Fu, Zheng Lin, and Weiping Wang. 2024. Are large language models table-based fact-checkers? *arXiv preprint arXiv:2402.02549*.

Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2023a. Jellyfish: A large language model for data preprocessing. *arXiv preprint arXiv:2312.01678*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Extractive summarization via chatgpt for faithful summary generation. In *Conference on Empirical Methods in Natural Language Processing*.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023c. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *ArXiv*, abs/2305.11159.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023d. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023e. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2023f. Data-copilot: Bridging billions of data and humans with autonomous workflow. *arXiv preprint arXiv:2306.07209*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023g. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023h. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. *arXiv preprint arXiv:2305.09067*.

Yichi Zhang, Jianing Yang, Keunwoo Yu, Yinpei Dai, Shane Storks, Yuwei Bao, Jiayi Pan, Nikhil Devraj, Ziqiao Ma, and Joyce Chai. 2023i. Seagull: An embodied agent for instruction following through situated dialog.

Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M Patel. 2023j. Reactable: Enhancing react for table question answering. *arXiv preprint arXiv:2310.00815*.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023k. CRT-QA: A dataset of complex reasoning question answering over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2153, Singapore. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023l. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022a. Description-driven task-oriented dialog modeling. *arXiv preprint arXiv:2201.08904*.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. Sentiment analysis: mining opinions, sentiments, and emotions.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Weiran Xu, Huixing Jiang, Wei Wu, and Yanan Wu. 2022b. Domain-oriented prefix-tuning: Towards efficient and generalizable fine-tuning for zero-shot dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4848–4862, Seattle, United States. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023b. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over tabular and textual data. *arXiv preprint arXiv:2401.13223*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023a. Extrapolating large language models to non-english by aligning languages. *ArXiv*, abs/2308.04948.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. 2023b. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. *arXiv preprint arXiv:2308.07902*.

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788*.