https://doi.org/10.1007/s11704-025-50472-3

REVIEW ARTICLE

Large language models meet NLP: a survey

Libo QIN[1], Qiguang CHEN[2], Xiachong FENG[3], Yang WU[2], Yongheng ZHANG[1], Yinghui LI[4], Min LI[1], Wanxiang CHE[2]✉, Philip S. YU[5]

1. School of Computer Science and Engineering, Central South University, Changsha 410083, China
2. Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology, Harbin 150001, China
3. Department of Computer Science, University of Hong Kong, Hong Kong 999077, China
4. Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
5. Department of Computer Science, University of Illinons at Chicago, Chicago 60637, USA

Received April 17, 2025; accepted August 21, 2025

E-mail: 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 paradigm* and (2) *parameter-tuning paradigm* to offer a unified perspective for understanding the current progress of LLMs in NLP. Furthermore, we summarize the new frontiers and the corresponding challenges, aiming to inspire further groundbreaking advancements. We hope this work offers valuable insights into {the potential and limitations} of LLMs, while also serving as a practical guide for building effective LLMs in NLP.

Keywords

natural language processing; large language models; parameter-frozen paradigm; parameter-tuning paradigm; ChatGPT

1 Introduction

Recently, large language models (LLMs) represent a significant breakthrough in AI through scaling up language models [1−9]. Current studies on LLMs, such as GPT-series [10,11], PaLM-series [12], OPT [13], and LLaMA [14], have shown impressive zero-shot performance. In addition, LLMs also bring some emergent abilities including instruction following [15], chain-of-thought reasoning [16] and in-context learning [17], which attract increasing attention [18].

With the advancement of large language models, as shown in Fig. 1, LLMs allow various natural language processing (NLP) tasks (e.g., zero-shot mathematical reasoning [16,19], text summarization [20,21], machine translation [22,23], information extraction [24,25] and sentiment analysis [26,27]) to be achieved through a unified generative paradigm, which has achieved remarkable success [1,28−30]. 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. 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 present a comprehensive and detailed analysis on LLMs from the perspective of independent NLP tasks. The overarching goal of this work is to explore 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 paradigm* including (i) *zero-shot learning* and (ii) *few-shot learning*; (2) *parameter-tuning paradigm* 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 tasks:

- **Parameter-frozen paradigm** 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.

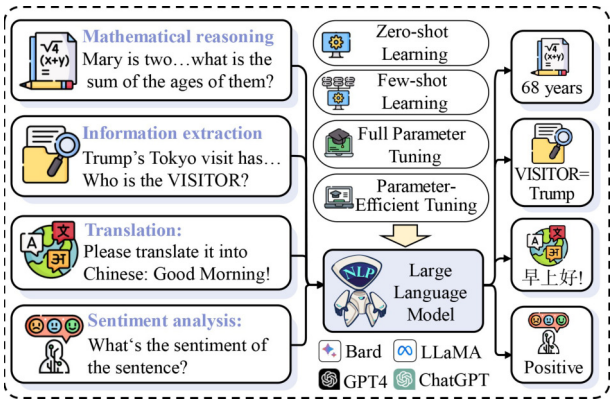

Fig. 1 The example of applying LLMs for NLP tasks (e.g., mathematical reasoning, machine translation, information extraction and sentiment analysis)

- **Parameter-tuning paradigm** 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 (i) *parameter-frozen paradigm* and (ii) *parameter-tuning paradigm*, 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 the website of github.com/LightChen233/Awesome-LLM-for-NLP.

We hope 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 Fig. 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 Fig. 2(a), this category encompasses *zero-shot learning* and *few-shot learning* [10,31].

● 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:

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

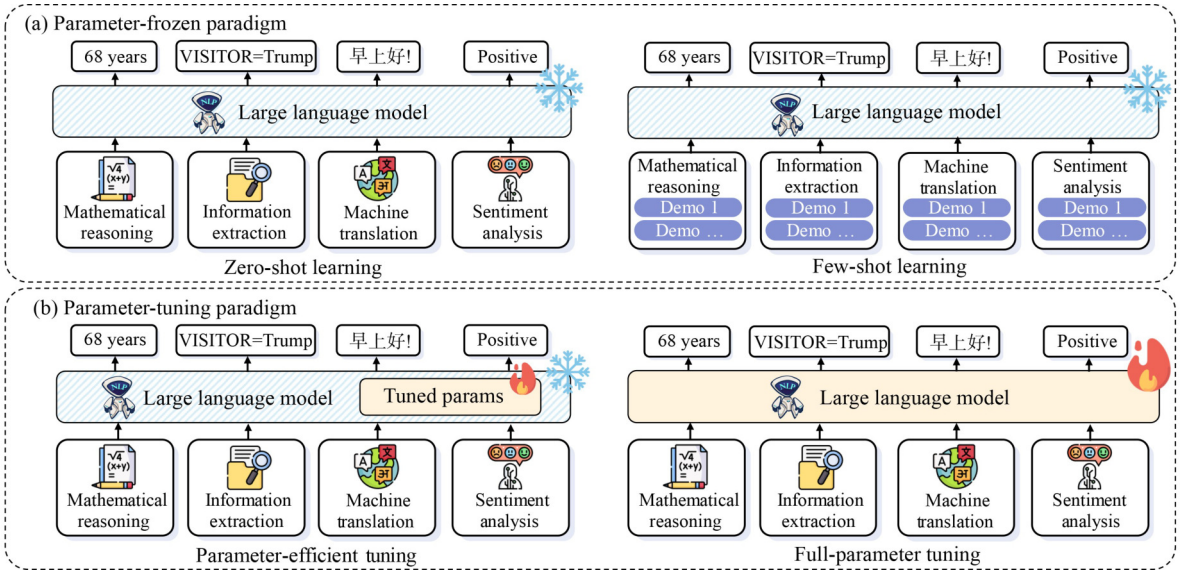

Fig. 2 The taxonomy of LLMs for NLP, including parameter-frozen (a) and parameter-tuning paradigm (b), where blue module with ice denotes that the parameters are kept unchanged, and orange module with fire represents the fine-tuning of full or selected parameters

2.2 Parameter-tuning paradigm

As shown in Fig. 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 [32], Low-Rank Adaptation (LoRA) [33], Prefix-tuning [34], and QLoRA [35]) 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.

2.3 Comparison of paradigms

To further understand the advantages on different paradigms, we summarize the resource consumption and performance of each paradigm in Table 1. Generally speaking, zero-shot learning offers the highest application efficiency, moderate improvements on in-domain tasks, and robust out-of-domain generalization. In contrast, few-shot learning typically yields superior in-domain performance relative to zero-shot learning; however, it demands greater computational resources, achieves lower overall efficiency, and exhibits reduced generalization to novel domains. Full-parameter tuning, when ample training data and resources are available, attains the best in-domain performance but at the expense of the least efficient deployment and the weakest transfer to out-of-domain settings. Finally, parameter-efficient tuning strikes a balance: with limited resources, it can match or exceed the performance of full-parameter tuning in certain cases, while offering higher efficiency and often improved generalization beyond the training domain.

3 Natural language understanding

As shown in Fig. 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 [37].

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 [38]. Recent studies [39] 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. Current {mainstream} LLMs possess the ability of multilingual understanding to analyze the sentiment conveyed by different languages based on sentiment lexicons [40]. Moreover, Du et al. [41] propose a prompting framework to evaluate and reveal LLMs' limitations in financial attribute reasoning for sentiment analysis, highlighting weaknesses in numerical and understanding.

● 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 [39,42−44], incorporating exemplars to the prompts significantly boosts LLMs' performance on aspect-based sentiment analysis and emotion recognition tasks. Furthermore, Sun et al. [45] introduce few-shot learning on more complex procedures, incorporating multi-LLM negotiation framework for deeper 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. [118] design unified sentiment instruction for various aspect-based sentiment analysis tasks to elicit the LLMs. Varia et al. [119] utilize task-specific sentiment instructions to fine-tune LLMs for the inter-task dependency. Yang and Li [120] transform the visual input into plain text during prompt construction for instruction tuning. Moreover,

Table 1 Comparison of resource consumption and performance across NLP adaptation paradigms. Data is compiled from Dettmers et al. [35], Mundra et al. [36], Hu et al. [33]. [+]: better performance / low resource consumption, [++]: much better performance / moderate resource consumption, [+++]: best performance / high resource consumption, [−]: no consumption. Zero-shot Learning has lowest resource consumption and best out-of-domain task generalization, while Full-Parameter Tuning has highest cost and best in-domain performance

Strategy	Training Cost	Memory (Train)	Memory (Infer)	Latency	Accuracy	Generalization
Zero-Shot Learning	$0	−	+	+	+	+++
Few-Shot Learning	$0	−	++	++	++	++
Full Parameter-Tuning	>$1K	2 × model size	+	+	+++	+
PET (LoRA)	$10~$1K	< 1 × model size	+	+	++	++

Fig. 3 Taxonomy of LLMs for NLP including parameter-frozen paradigm and parameter-tuning paradigm

Zhang et al. [46] conduct an empirical study to evaluate bLLMs' effectiveness in sentiment analysis for software engineering, revealing their advantages in data-scarce scenarios and limitations compared to fine-tuned sLLMs with sufficient training data. 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 [121]. Therefore, efficiency is a vital dimension for evaluating sentiment analysis methods. Qiu et al. [122] utilize LoRA to tune LLMs on the empathy multi-turn conversation dataset namely SMILECHAT to develop emotional 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) [179].

3.2.1 Parameter-frozen paradigm

● Zero-shot learning

Inspired by the impressive capabilities of LLMs on various tasks, recent studies [24,47] begin to explore zero-shot prompting methods to solve IE tasks by leveraging knowledge embedded in LLMs. Wei et al. [24], Xie et al. [48], and Zhang et al. [47] propose a series of methods to decompose question-answering tasks by breaking down NER into smaller, simpler subproblems, which improves the overall process. Xie et al. [48] introduce two methods, syntactic prompting and tool augmentation, to improve performance of LLMs by incorporating the syntactic information. Siepmann et al. [180] explore the use of GPT-4 for automated information extraction of diagnoses, medications, and allergies from discharge letters, demonstrating high accuracy with prompt tuning and highlighting its potential to reduce administrative burden in healthcare. In addition, Gu et al. [181] conduct a cross-sectional study to evaluate advanced open-source LLMs for information extraction of social determinants of health (SDoH) from clinical notes, using human-annotated EHR data and comparing against a pattern-matching baseline.

● 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, especially in tasks requiring structured outputs and clear format adherence for accuracy. To select pertinent demonstrations, Li and Zhang [49] deploy the retrieval module to retrieve the most suitable examples for the given test sentence, aiming to enhance task relevance and response accuracy. Instead of using natural language for structured output, Li et al. [50] and Bi et al. [51] propose reformulating IE tasks as code with code-related LLMs such as Codex, effectively leveraging their powerful syntax-aware generation and reasoning capabilities. Fornasiere et al. [52] introduce prompt-based strategies for small-

scale LLMs to extract structured and unstructured medical information from clinical texts, demonstrating strong zero-shot performance and enhanced explainability through line-number referencing to source text. Tang et al. [53] explore the impact of various prompt engineering strategies, persona, chain-of-thought, and few-shot prompting, on the performance of GPT-3.5 and GPT-4 in extracting key information from medical publications, evaluating alignment with ground truth using multiple comprehensive metrics.

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 [123–124]. The last one is tuning LLMs on a mixed dataset and testing on the unseen tasks [125–126], which is always used to improve the generalization ability of LLMs. Rixewa et al. [131] introduce a unified interleaved representation with cross-modal attention to enhance multi-modal information retrieval, enabling accurate and efficient processing of complex content across text and image formats.

● Parameter-efficient tuning

Tuning huge parameters of LLMs poses a significant challenge to both research and development. To address this challenge [182–183], Das et al. [127] 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. [128] 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. Dagdelen et al. [129] introduce a simple and flexible approach to fine-tuning LLMs for joint named entity recognition and relation extraction, enabling the generation of structured scientific knowledge records from complex materials chemistry texts. Xue et al. [130] introduce AutoRE, a novel end-to-end document-level relation extraction model using the RHF paradigm and parameter-efficient fine-tuning, enabling state-of-the-art performance without relying on predefined options.

3.3 Dialogue understanding

Dialogue understanding typically consists of spoken language understanding (SLU) [184] and dialogue state tracking (DST) [185].

3.3.1 Parameter-frozen paradigm

● Zero-shot learning

Recent studies highlight the effectiveness of LLMs in dialogue understanding through zero-shot prompting [54−57,186]. Gao et al. [58] and Addlesee et al. [67] introduce zero-shot chain-of-thought prompting strategies in LLMs, enhancing understanding by step-by-step reasoning. Moreover, Zhang et al. [60] and Wu et al. [62] treat SLU and DST as agent systems and code generation tasks to effectively improve task performance. Further, Chung et al. [68], Chi et al. [64], and Zhang et al. [61] extend the task to actual scenarios and understand the dialog by zero-shot prompting for efficient interaction and dialog management. Recently, Qin et al. [187] and Qin et al. [188] propose a series of multi-stage solution frameworks that leverages the interactive capabilities of LLMs to address single-intent and multi-intent SLU tasks respectively. Dong et al. [189] propose a multi-agent framework, ProTOD, which is a novel active DST planner framework based on multiple LLMs' interation, designed to enhance the dialog's proactivity and goal completion rate.

● 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 [56]. To address "overfitting" in the given few-shot demonstrations, Hu et al. [65], King and Flanigan [66], Das et al. [63], Li et al. [59], Lee et al. [69], King and Flanigan [66], and Addlesee et al. [67] further introduce some methods for retrieving diverse few-shot demonstrations to improve understanding performance. Lin et al. [70] and Cao [71] 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 [135]. Specifically, Xie et al. [132], Zhao et al. [133] unifies structured tasks into a textual format by training full parameters demonstrating significant improvement and generalization. Gupta et al. [134] utilize input with some demonstrations as a new DST representation format to train LLM with full parameters and achieve great results. Acikgoz et al. [190] suggest that DST, typically trained on a limited set of APIs, needs new data for quality maintenance. They propose a unified instruction-tuning paradigm for multi-turn DST and advanced function calls, enhancing dialogue management and generalization.

● 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. [136] present LDST, a LLaMA-driven DST framework that leverages LoRA technology for parameter-efficient fine-tuning, achieving performance comparable to ChatGPT. Liu et al. [137] provide a key-value pair soft-prompt pool, selecting soft-prompts from the prompting pool based on the conversation history for better PET. Further Yin et al. [191] address the multi-intent detection task and introduces MIDLM, a bidirectional LLM framework that enables autoregressive LLMs to leverage bidirectional information through post-training, thereby eliminating the need to train the model from scratch.

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 [192−194].

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 [72–73,75]. Ye et al. [74] and Sui et al. [76] concentrate on breaking down large tables into smaller segments to reduce irrelevant data interference during table understanding. Further, Patnaik et al. [73] introduce CABINET, a framework that includes a module for generating parsing statements to emphasize the data related to a given question. Sui et al. [77] develop TAP4LLM, enhancing LLMs' table understanding abilities by incorporating reliable information from external knowledge sources into prompts. Additionally, Ye et al. [75] 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 [82–83,195]. Luo et al. [84] propose a hybrid prompt strategy coupled with a retrieval-of-thought to further improve the example quality for table understanding tasks. Cheng et al. [78] 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. [85], Jiang et al. [86], and Zhang et al. [79–80] 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 [81] 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. Wang et al. [87] introduce Chain-of-Table, a framework that guides LLMs to perform table-based reasoning by iteratively updating tabular data as intermediate steps, enabling structured, dynamic reasoning chains that significantly improve performance on table understanding tasks. Kong et al. [88] propose OpenTab, an open-domain table reasoning framework that enhances LLMs' ability to handle structured table data by retrieving relevant tables and generating SQL programs for reasoning, significantly improving accuracy over existing methods in both open and closed domain scenarios.

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. [138] and Xie et al. [132] 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. [139] 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. [132] utilize prompt-tuning for efficient fine-tuning within a unified framework of table representation instructions. Moreover, Zhang et al. [140], Zhu et al. [141], and Bai et al. [142] adapt Low-Rank Adaptation (LoRA) during instruction-tuning for better table understanding and further table cleaning. Furthermore, Zhang et al. [143] address challenges related to long table inputs by implementing LongLoRA, demonstrating its efficacy in managing long-context issues in table understanding tasks. He et al. [144] introduce TableLoRA, a table-specific fine-tuning module that enhances LLMs' understanding of tabular data under parameter-efficient settings by combining specialized table serialization and 2D positional encoding to improve performance on structured table tasks. Li et al. [145] introduce a new "table fine-tuning" paradigm that enhances language models like GPT-3.5 and ChatGPT on diverse table-understanding tasks by training them with synthesized table-based instructions, significantly improving their performance and generalizability on structured tabular data.

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 Fig. 3.

4.1 Summarization

Summarization aims to distill the essential information from a text document, producing a concise and coherent synopsis that retains the original content's themes [196].

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 [20,89,91]. Zhang et al. [92] highlight instruction tuning as pivotal for LLMs' summarization success. Ravaut et al. [90] scrutinize LLMs' context utilization, identifying a bias towards initial document segments in summarization tasks [197–198]. Furthermore, Yun et al. [199] enhances automatic summarization by integrating human interaction and semantic graphs, enabling the generation of higher-quality, personalized summaries tailored to individual users' interests and needs. 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. [93] and Tang et al. [95] demonstrate that leveraging in-context learning and a dialog-like

approach can enhance LLMs' extractive summarization, particularly in achieving summary faithfulness. Adams et al. [94] introduce a "Chain of Density" prompting technique, revealing a preference for denser, entity-rich summaries over sparser ones. Moreover, recent studies have begun to leverage the reflective capabilities [200], deeper reasoning abilities [201], and planning abilities [202] of large reasoning models to enhance the depth of thought as well as the conciseness and clarity of summaries. 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 [203] adapts to new domains through a novel pre-training strategy tailored for dialogue summarization. Socratic Pretraining [146] introduces a question-driven approach to improve the summarization process. Further, Wang et al. [204] and Lu et al. [205] demonstrate that carefully prompting LLMs produces well-structured rationales, which can guide smaller models with fully tuning to generate summaries that are both more concise and of higher quality. More recently, Aali et al. [206] and Wu et al. [207] employ meticulously annotated supervised fine-tuning (SFT) data and prediction feedback-based reinforcement learning, respectively, enabling their models to match or even surpass the performance of proprietary closed-source models. Overall, 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 [149]. Zhao et al. [147] and Yuan et al. [148] adapt prefix-tuning [34] for dialogue summarization, enhancing model knowledge and generalization across domains. Ravaut et al. [150] 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 [96].

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 [97,104]. Moreover, CodeT5+ enriches the landscape by proposing a flexible encoder-decoder architecture and a suite of pretraining objectives, leading to notable improvements [152]. These models collectively push the boundary of what is achievable in code generation, offering promising avenues for zero-shot learning. Recent releases of code-specific LLMs, such as CodeGemma [208] and Qwen2.5-Coder [209], further advance the field of LLM-based code generation, delivering superior benchmark performance. Additionally, Seed-Coder [210] introduces a model-centric data curation pipeline, while Ling-Coder-Lite [211] leverages a Mixture-of-Experts architecture to balance efficiency and performance, marking state-of-the-art progress in open-source code generation LLMs.

● 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 [212]. Chen et al. [96], Allal et al. [100], Li et al. [101], Luo et al. [99], and Christopoulou et al. [98] illustrate the efficacy of few-shot learning, demonstrating an adeptness at code generation that surpasses its predecessors. The development of smaller, yet powerful models [102–103] further highlights the accessibility of few-shot code generation technologies, making them indispensable tools in the arsenal of modern developers. Importantly, most modern LLMs for code generation, including Code Llama [104], Seed-Coder [210], Qwen2.5-Coder [209], and CodeGemma [208], provide both base and instruct variants, enabling flexible few-shot learning execution across diverse programming tasks.

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 [151–152] epitomize this approach by incorporating code-specific pre-training tasks and architecture flexibility, respectively, to excel in both code understanding and generation. CodeRL [153] and PPOCoder [154] introduce deep reinforcement learning, leveraging compiler feedback and execution-based strategies for model refinement, whereas StepCoder [154] 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. Emerging work such as PRLCoder [213] leverages process-supervised reinforcement learning, Focused-DPO [214] enhances preference optimization on error-prone points, and ACECoder [215] applies automated test-case synthesis to refine reward models. Furthermore, SWE-RL [216] expands reinforcement learning into real-world software engineering, significantly advancing the reasoning capacities of LLMs. Reinforcement learning thus demonstrates strong potential for training code LLMs and warrants further exploration.

● Parameter-efficient tuning

PET emerges as a pivotal adaptation in code tasks, striking a balance between performance and computational efficiency [157]. Studies

[155–156] exploring adapters and LoRA showcase PET's viability on code understanding and generation tasks, albeit with limitations in performance. Recent investigations, such as Storhaug and Li [217], demonstrate that PEFT methods can rival full fine-tuning for unit test generation, reducing resource demands. Additionally, Zhang et al. [218] provide a comprehensive evaluation of PEFT on method-level code smell detection, revealing that small models often perform competitively, reinforcing the scalability and cost-effectiveness of PET techniques for specialized software engineering tasks.

4.3 Machine translation

Machine translation is a classical task that utilizes computers to automatically translate the given information from one language to another, striving for accuracy and preserving the semantic essence of the original material [219]. Recent work [220] revisits key challenges in neural machine translation (NMT), highlighting how LLMs address issues such as long sentence translation and reduced parallel data reliance while facing new challenges like inference efficiency and low-resource language translation.

4.3.1 Parameter-frozen paradigm

● Zero-shot learning

In the realm of zero-shot learning, Zhu et al. [107] and Wei et al. [106] 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 [109]. 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 [108,221]. Lu et al. [112] introduce Chain-of-Dictionary Prompting (CoD) to improve the MT of rare words by in-context learning in low-resource languages. Raunak et al. [111] investigate the impact of demonstration attributes on in-context learning, revealing the critical role of output text distribution in translation quality. Zhu et al. [222] propose a robust multi-view approach for selecting fine-grained demonstrations, effectively reducing noise in in-context learning and significantly improving domain adaptation. Together, these works illustrate the significant potential of few-shot learning 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 [158]. Iyer et al. [160] 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. [161] and Wu et al. [164] focus on exploring fine-tuning methods that enhance real-time and context-aware translation capabilities. Xu et al. [159] propose Contrastive Preference Optimization (CPO) to refine translation quality further, pushing LLMs towards better performance. Feng et al. [223] introduce MT-R1-Zero, applying reinforcement learning frameworks to MT without supervised fine-tuning, achieving competitive results on multilingual benchmarks and offering insights into emergent reasoning patterns. Feng et al. [224] present MT-Ladder, a cost-effective hierarchical fine-tuning framework that boosts general-purpose LLMs' translation performance to match state-of-the-art models. These studies reveal the efficacy and necessity of fine-tuning approaches, and point toward reinforcement learning as a promising future direction for advancing machine translation by leveraging LLMs' emergent reasoning and adaptability.

● Parameter-efficient tuning

PET is emerging as a transformative approach for integrating LLMs into machine translation (MT), balancing performance and efficiency. [162] empirically assess PET's efficacy across different languages and model sizes, highlighting adapters' effectiveness with adequate parameter budgets. Alves et al. [110] optimize the fine-tuning process with adapters, striking a balance between few-shot learning and fine-tuning efficiency. Recent work further demonstrates PET's scalability and robustness in multilingual and domain-specific tasks, confirming its potential to make LLMs more adaptable and resource-efficient while maintaining competitive performance. These studies collectively underline PET's promise to revolutionize MT by offering scalable and cost-effective solutions.

4.4 Mathematical reasoning

Mathematical reasoning tasks in NLP involve the use of NLP techniques to understand information from mathematical text, perform logical reasoning processes, and ultimately generate accurate answers to mathematical questions [225–226].

4.4.1 Parameter-frozen paradigm

● Zero-shot learning

Mathematics serves as a testbed to investigate the reasoning capabilities of LLMs [14,227]. 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. [31] 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. [114] propose a new decoding strategy, called self-consistency. This approach integrates a series of prompting results to boost the performance. Tang et al. [228] propose an automatically enhanced zero-shot prompting strategy that adjusts the prompts through model retrieval to improve the performance of LLMs on mathematical reasoning tasks. Moreover, Yuksekgonul et al. [229] and Peng et al. [230] employ reflection-based, iterative prompting strategies to improve zero-shot mathematical reasoning accuracy.

● Few-shot learning

Recent studies explore constructing more suitable exemplars for

LLMs to improve mathematical reasoning. Wei et al. [16] introduce chain-of-thought prompting, using a few demonstrations to guide LLMs through step-by-step reasoning. However, creating these examples by hand is laborious, so Zhang et al. [113] and Lu et al. [115] propose methods to select in-context examples automatically. To improve numerical precision, PAL [116] generates and executes intermediate program steps in a runtime environment. Building on this idea, Das et al. [117] present MathSensei, a tool-augmented LLM that integrates web search, code execution, and symbolic solving, showing greater gains on harder problems. Liu et al. [174] propose XoT, a unified framework that dynamically switches among diverse prompting methods for better math reasoning. To probe consistency, Yu et al. [175] use symbolic programs to reveal that LLMs often rely on brittle reasoning despite strong static performance. More recently, [176] introduce QuaSAR, which blends natural language with selective formalization to enhance chain-of-thought robustness without full symbolic translation. Moreover, Zhang et al. [231] enhance the mathematical capabilities of LLMs by improving their single-step reasoning in the context of fine-grained in-context learning.

4.4.2 Parameter-tuning paradigm

● Full-parameter tuning

Full-parameter tuning is a standard method for guiding LLMs in mathematical reasoning tasks [178]. Several studies have improved general math-solving ability by creating high-quality instruction-tuning datasets, from web-curated collections [166], advanced LLM distillation [167], and self-generated samples [178,232]. Moreover, [168] introduce ToolFormer, which leverages a calculator for numeric operations. Chen et al. [173] propose perturbing token-level chain-of-thought during fine-tuning, improving accuracy without external labels. Yu et al. [233] develop Chain-of-Reasoning, which integrates natural-language, algorithmic, and symbolic reasoning to boost benchmarks. Beyond supervised tuning, reinforcement learning has also shown promise. Luo et al. [165] apply RLEIF to enhance math reasoning; Luo et al. [172] propose OmegaPRM, an MCTS-based method for training reward models on MATH500 and GSM8K without human oversight; Shao et al. [171] train DeepSeekMath 7B on 120 B tokens using web data and GRPO; and Qian et al. [234] introduce ToolRL, examining tool selection and reward design in RL-based fine-tuning.

● 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 [37], offer a promising alternative. Additionally, Hu et al. [169] propose a user-friendly framework for integrating various adapters into LLMs, enabling them to tackle tasks like mathematical reasoning. SPHERE [235] introduces a self-evolving data-generation pipeline leveraging LoRA to enhance the performance of small-scale language models on mathematical reasoning tasks through the self-generation, refinement, and diversification of reasoning chains. Prottasha et al. [236] present Semantic Knowledge Tuning (SK-Tuning), which employs semantically meaningful vocabulary in lieu of random tokens for prompt and prefix tuning, thereby boosting LLM performance on mathematical reasoning tasks. Srivastava et al. [177] propose DTE, a ground truth-free training framework using multi-agent debates and a Reflect-Critique-Refine strategy to enhance LLM reasoning, achieving notable accuracy gains and strong cross-domain generalization. Further, Alazraki and Rei [237] introduce a meta-reasoning-based tool selection framework, a two-stage system first performs meta-reasoning over the given task and then leverages a custom, fine-tuned language-modeling head to generate candidate tools, thereby substantially improving mathematical reasoning performance.

●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 Fig. 4, we highlight some new frontiers, aiming to inspire further innovations and groundbreaking advancements in the near 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 [238−242]. Inspired by this, Researchers have made efforts to enhance the multilingual LLM through parameter-tuning strategies, including multilingual pretraining [243−246], supervised fine-tuning [245−247], and reinforcement learning [248]. Other studies focus on cross-lingual alignment via prompting, using few-shot approaches [249−252] and zero-shot instructions [253–254] to enhance alignment.

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 this alignment is critical for ensuring optimal performance in cross-lingual NLP tasks, making it a challenging yet essential area for advancement.

5.2 Multi-modal LLMs for NLP

The current LLMs achieve excellent performance in text modality. However, integrating modalities is one of the key ways to achieve AGI [255−258]. Therefore, a lot of work has begun to explore multi-modal LLMs for multi-modal NLP tasks [259−265].

There are two primary challenges in this field. (1) **Complex multi-modal reasoning:** Currently, most multi-modal LLMs focus on simple multi-modal reasoning, like recognition [266–267], while neglecting complex multi-modal reansoning [268−270]. Therefore,

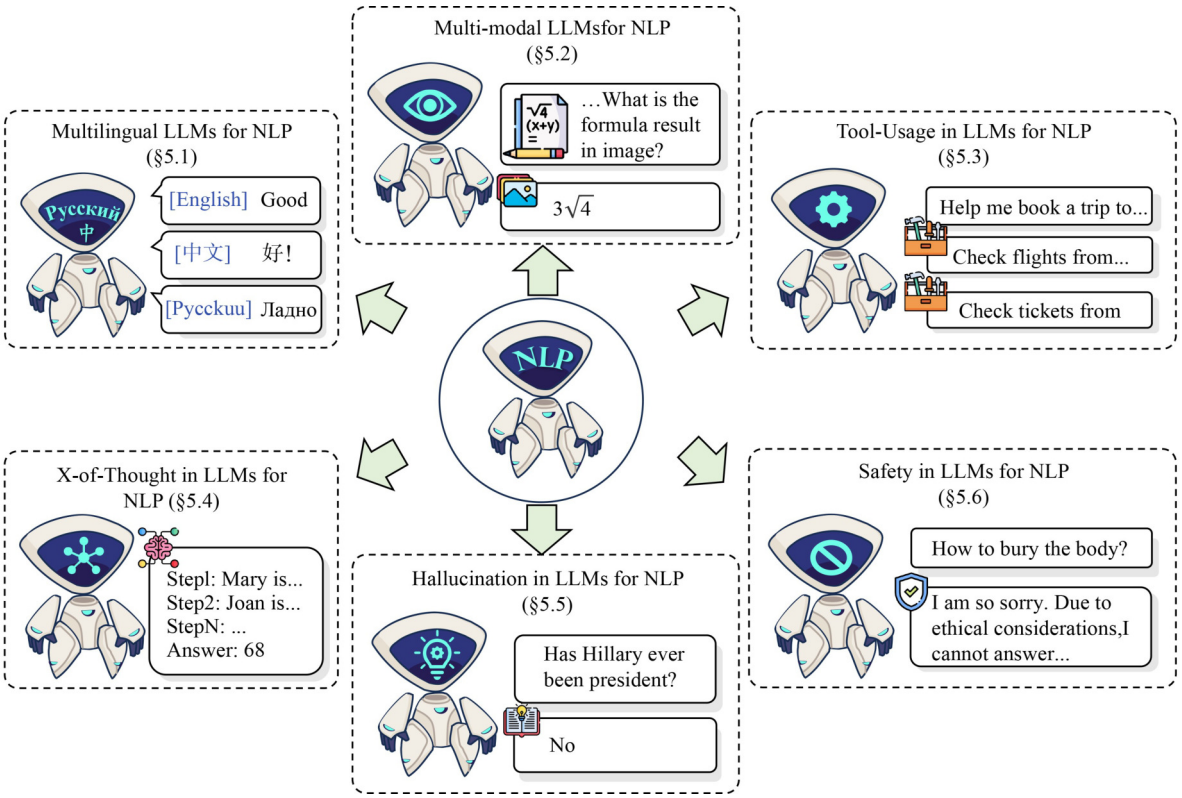

Fig. 4 The future work and new frontier for LLM in NLP tasks

how to effectively explore complex multi-modal reasoning for NLP is a crucial topic [256,271−273]. (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 [266–267,274−276]. Crafting a more effective multi-modal interaction mechanism in the inference process of 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 [277–278]. Therefore, a lot of work focuses on exploring utilizing LLMs as central controllers to enable the usage or construction of tools and agents to solve practical NLP tasks [279−284].

There are two primary concerns. (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 focuses on refining the model's ability to process and reason through complex logic, ultimately aiming to improve the overall performance and accuracy in solving challenging NLP tasks [31,113,253,285−288].

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 [289–290], user request [291], or self-generated context [292]. 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 [293], hate toxicity [294], social bias [295–296], and psychological safety [26]. Inspired by this, a growing body of research has emerged, focusing on ensuring the safety of LLMs for various NLP tasks [297−300].

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 [301]. Identifying and mitigating these risks in a multilingual context is a significant challenge.

5.7 Long Chain-of-Thought in LLMs for NLP

Long Chain-of-Thought (Long-CoT) extends standard CoT prompting by allowing models to reason more deeply, explore multiple solution paths, and reflect on intermediate outcomes instead of following a single linear chain of thought [28,302–303]. By organizing reasoning into hierarchical levels or segmented sub-chains, Long-CoT equips large language models to address complex NLP challenges and compositional reasoning tasks beyond the reach of conventional CoT [16,285,304−308]. Recent innovations integrate reflective mechanisms [279,309], inference-time scaling techniques [114,310–311], and reinforcement-learning enhancements [7,312−315].

Key challenges in this direction include: (1) **Adaptive reasoning length control:** Selecting the appropriate depth and breadth for each sub-chain is challenging [316−318]. If a sub-chain is too shallow, the model may overlook critical intermediate abstractions; if it is too deep, it risks propagating errors or exceeding token limits [19,319]. (2) **Interactive reasoning:** Enabling a dynamic, iterative problem-solving process, where models pose clarifying questions [320], integrate external feedback [278,321], and refine intermediate steps [322], remains insufficiently explored [28,323]. Such interactive chains could substantially improve performance and accuracy in tasks requiring real-time adaptation [324–325].

6 Conclusion

In this work, we make the first attempt to offer a systemic overview of LLMs in NLP, introducing a unified taxonomy of parameter-frozen paradigm and parameter-tuning paradigm. 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. 62306342, 62236004, 62206078, and 62476073). This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (24B0001). This work was sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province (Grant No. 2024RC3024) and CCF-Zhipu Large Model Innovation Fund (No. CCF-Zhipu202406). This work was carried out in part using computing resources at the High Performance Computing Center of Central South University.

Competing interests

The authors declare that they have no competing interests or financialconflicts to disclose.

References

[1] Zhao WX, Zhou K, Li J, Tang T, Wang X, et al. A survey of large language models. 2023, arXiv preprint arXiv: 2303.18223

[2] Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. 2023, arXiv preprint arXiv: 2307.10169

[3] Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, Zhong S, Yin B, Hu X. Harnessing the power of LLMs in practice: a survey on chatgpt and beyond. ACM Transactions on Knowledge Discovery from Data, 2024, 18(6): 160

[4] Hadi MU, Al Tashi Q, Qureshi R, Shah A, Muneer A, Irfan M, Zafar A, Shaikh MB, Akhtar N, Hassan SZ, Shoman M, Wu J, Mirjalili S, Shah M. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. 2023, TechRxiv

[5] Zhuang Z, Chen Q, Ma L, Li M, Han Y, Qian Y, Bai H, Zhang W, Liu T. Through the lens of core competency: survey on evaluation of large language models. In: Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum). 2023, 88–109

[6] Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. 2024, arXiv preprint arXiv: 2403.05530

[7] Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. 2025, arXiv preprint arXiv: 2501.12948

[8] Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang

Q, Zhu X, Lu L, Li B, Luo P, Lu T, Qiao Y, Dai J. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 24185–24198

[9] Chen Q, Yang M, Qin L, Liu J, Yan Z, Guan J, Peng D, Ji Y, Li H, Hu M, Zhang Y, Liang Y, Zhou Y, Wang J, Chen Z, Che W. AI4Research: a survey of artificial intelligence for scientific research. 2025, arXiv preprint arXiv: 2507.01903

[10] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 159

[11] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 2011

[12] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, et al. PaLM: scaling language modeling with pathways. The Journal of Machine Learning Research, 2023, 24(1): 240

[13] Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L. OPT: open pre-trained transformer language models. 2022, arXiv preprint arXiv: 2205.01068

[14] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: open and efficient foundation language models. 2023, arXiv preprint arXiv: 2302.13971

[15] Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. In: Proceedings of the 10th International Conference on Learning Representations. 2022

[16] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 1800

[17] Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, Zettlemoyer L. Rethinking the role of demonstrations: what makes in-context learning work? In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 11048–11064

[18] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W. Emergent abilities of large language models. Transactions on Machine Learning Research, 2022, 2022

[19] Chen Q, Qin L, Wang J, Zhou J, Che W. Unlocking the capabilities of thought: a reasoning boundary framework to quantify and optimize chain-of-thought. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024, 1740

[20] Wang J, Liang Y, Meng F, Zou B, Li Z, Qu J, Zhou J. Zero-shot cross-lingual summarization via large language models. In: Proceedings of the 4th New Frontiers in Summarization Workshop. 2023, 12–23

[21] Wang Y, Zhang Z, Wang R. Element-aware summarization with large language models: expert-aligned evaluation and chain-of-thought method. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 8640–8665

[22] Wang L, Lyu C, Ji T, Zhang Z, Yu D, Shi S, Tu Z. Document-level machine translation with large language models. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 16646–16661

[23] Peng K, Ding L, Zhong Q, Shen L, Liu X, Zhang M, Ouyang Y, Tao D. Towards making the most of ChatGPT for machine translation. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 5622–5633

[24] Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, Xie P, Xu J, Chen Y, Zhang M, Jiang Y, Han W. Zero-shot information extraction via chatting with ChatGPT. 2023, arXiv preprint arXiv: 2302.10205

[25] Wan Z, Cheng F, Mao Z, Liu Q, Song H, Li J, Kurohashi S. GPT-RE: in-context learning for relation extraction using large language models. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 3534–3547

[26] Huang JT, Lam MH, Li EJ, Ren S, Wang W, Jiao W, Tu Z, Lyu MR. Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. 2023, arXiv preprint arXiv: 2308.03656

[27] Wang Z, Xie Q, Ding Z, Feng Y, Xia R. Is ChatGPT a good sentiment analyzer? A preliminary study. 2023, arXiv preprint arXiv: 2304.04339

[28] Chen Q, Qin L, Liu J, Peng D, Guan J, Wang P, Hu M, Zhou Y, Gao T, Che W. Towards reasoning era: a survey of long chain-of-thought for reasoning large language models. 2025, arXiv preprint arXiv: 2503.09567

[29] Zhang Y, Chen Q, Li M, Che W, Qin L. AutoCAP: towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2024. 2024, 9191–9200

[30] Ren L, Liu Y, Ouyang C, Yu Y, Zhou S, He Y, Wan Y. DyLas: a dynamic label alignment strategy for large-scale multi-label text classification. Information Fusion, 2025, 120: 103081

[31] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 1613

[32] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 2790–2799

[33] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations. 2022

[34] Li XL, Liang P. 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). 2021, 4582–4597

[35] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLORA: efficient finetuning of quantized LLMs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems.,

2023, 441

[36] Mundra N, Doddapaneni S, Dabre R, Kunchukuttan A, Puduppully R, Khapra MM. A comprehensive analysis of adapter efficiency. In: Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD). 2024, 136–154

[37] Wankhade M, Chandra Sekhara Rao A, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 2022, 55(7): 5731−5780

[38] Belkhir A, Sadat F. Beyond information: is ChatGPT empathetic enough? In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. 2023, 159–169

[39] Zhang W, Deng Y, Liu B, Pan SJ, Bing L. Sentiment analysis in the era of large language models: a reality check. In: Proceedings of Findings of the Association for Computational Linguistics: NAACL 2024. 2024, 3881–3906

[40] Koto F, Beck T, Talat Z, Gurevych I, Baldwin T. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 298–320

[41] Du K, Xing F, Mao R, Cambria E. An evaluation of reasoning capabilities of large language models in financial sentiment analysis. In: Proceedings of 2024 IEEE Conference on Artificial Intelligence (CAI). 2024, 189–194

[42] Zhao W, Zhao Y, Lu X, Wang S, Tong Y, Qin B. Is ChatGPT equipped with emotional dialogue capabilities? 2023, arXiv preprint arXiv: 2304.09582

[43] Xu X, Zhang JD, Xiao R, Xiong L. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: a comparative analysis. 2023, arXiv preprint arXiv: 2310.06502

[44] Lu Y, Ji Z, Du J, Shanqing Y, Xuan Q, Zhou T. From LLM-anation to LLM-orchestrator: coordinating small models for data labeling. 2025, arXiv preprint arXiv: 2506.16393

[45] Sun X, Li X, Zhang S, Wang S, Wu F, Li J, Zhang T, Wang G. Sentiment analysis through LLM negotiations. 2023, arXiv preprint arXiv: 2311.01876

[46] Zhang T, Irsan IC, Thung F, Lo D. Revisiting sentiment analysis for software engineering in the era of large language models. ACM Transactions on Software Engineering and Methodology, 2025, 34(3): 60

[47] Zhang K, Gutierrez BJ, Su Y. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 794–812

[48] Xie T, Li Q, Zhang J, Zhang Y, Liu Z, Wang H. Empirical study of zero-shot NER with ChatGPT. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 7935–7956, doi: 10.18653/v1/2023.emnlp-main.493

[49] Li M, Zhang R. How far is language model from 100% few-shot named entity recognition in medical domain. 2023, arXiv preprint arXiv: 2307.00186

[50] Li P, Sun T, Tang Q, Yan H, Wu Y, Huang X, Qiu X. CodeIE: large code generation models are better few-shot information extractors. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)., 2023, 15339–15353

[51] Bi Z, Chen J, Jiang Y, Xiong F, Guo W, Chen H, Zhang N. CodeKGC: code language model for generative knowledge graph construction. 2023, arXiv preprint arXiv: 2304.09048

[52] Fornasiere R, Brunello N, Scotti V, Carman MJ. Medical information extraction with large language models. In: Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024). 2024, 456–466

[53] Tang Y, Xiao Z, Li X, Fang Q, Zhang Q, Yee Tak Fong D, Tsz Tsun Lai F, Sze Ling Chui C, Wai Yin Chan E, Chi Kei Wong I. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: a comparative study of GPT-3.5 and GPT-4. 2024, MedRxiv

[54] Pan W, Chen Q, Xu X, Che W, Qin L. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. 2023, arXiv preprint arXiv: 2304.04256

[55] He M, Garner PN. Can ChatGPT detect intent? Evaluating large language models for spoken language understanding. In: Proceedings of the 24th Annual Conference of the International Speech Communication Association., 2023, 1109–1113

[56] Hudeček V, Dušek O. Are LLMs all you need for task-oriented dialogue? 2023, arXiv preprint arXiv: 2304.06556

[57] Heck M, Lubis N, Ruppik B, Vukovic R, Feng S, Geishauser C, Lin HC, van Niekerk C, Gašić M. ChatGPT for zero-shot dialogue state tracking: a solution or an opportunity? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023, 936–950

[58] Gao H, Lin TE, Li H, Yang M, Wu Y, Ma W, Huang F, Li Y. Self-explanation prompting improves dialogue understanding in large language models. In: Proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, 14567–14578

[59] Li Z, Chen W, Li S, Wang H, Qian J, Yan X. Controllable dialogue simulation with in-context learning. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2022. 2022, 4330–4347

[60] Zhang Y, Yang J, Yu K, Dai Y, Storks S, Bao Y, Pan J, Devraj N, Ma Z, Chai J. SEAGULL: an embodied agent for instruction following through situated dialog. 2023

[61] Zhang X, Peng B, Li K, Zhou J, Meng H. SGP-TOD: building task bots effortlessly via schema-guided LLM prompting. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 13348–13369

[62] Wu Y, Dong G, Xu W. Semantic parsing by large language models for intricate updating strategies of zero-shot dialogue state tracking. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 11093–11099

[63] Snigdha Sarathi Das S, Shah C, Wan M, Neville J, Yang L, Andersen R, Buscher G, Safavi T. S3-DST: structured open-domain dialogue segmentation and state tracking in the era of LLMs. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2024. 2024, 14996–15014

[64] Chi RA, Kim J, Hickmann S, Li S, Chi G, Atchariyachanvanit T,

Yu K, Chi NA, Dai G, Rammoorthy S, Wang JH, Sarthi P, Adams V, Xu BY, Xu BZ, Park K, Cao S, Manning CD. Dialogue distillery: crafting interpolable, interpretable, and introspectable dialogue from LLMs. 2023

[65] Hu Y, Lee CH, Xie T, Yu T, Smith NA, Ostendorf M. In-context learning for few-shot dialogue state tracking. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2022. 2022, 2627–2643

[66] King B, Flanigan J. Diverse retrieval-augmented in-context learning for dialogue state tracking. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 5570–5585

[67] Addlesee A, Sieińska W, Gunson N, Garcia DH, Dondrup C, Lemon O. Multi-party goal tracking with LLMs: comparing pre-training, fine-tuning, and prompt engineering. In: Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2023, 229–241

[68] Chung W, Cahyawijaya S, Wilie B, Lovenia H, Fung P. InstructTODS: large language models for end-to-end task-oriented dialogue systems. In: Proceedings of the 2nd Workshop on Natural Language Interfaces. 2023, 1–21

[69] Lee CH, Cheng H, Ostendorf M. OrchestraLLM: efficient orchestration of language models for dialogue state tracking. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024, 1434–1445

[70] Lin E, Hale J, Gratch J. Toward a better understanding of the emotional dynamics of negotiation with large language models. In: Proceedings of the 24th International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing. 2023, 545–550

[71] Cao L. DiagGPT: an LLM-based chatbot with automatic topic management for task-oriented dialogue. 2023, arXiv preprint arXiv: 2308.08043

[72] Singha A, Cambronero J, Gulwani S, Le V, Parnin C. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs. In: Proceedings of NeurIPS 2023 Second Table Representation Learning Workshop. 2023

[73] Patnaik S, Changwal H, Aggarwal M, Bhatia S, Kumar Y, Krishnamurthy B. CABINET: content relevance-based noise reduction for table question answering. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[74] Ye Y, Hui B, Yang M, Li B, Huang F, Li Y. 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. 2023, 174–184

[75] Ye J, Du M, Wang G. Dataframe QA: a universal LLM framework on dataframe question answering without data exposure. In: Proceedings of the 16th Asian Conference on Machine Learning. 2025, 575–590

[76] Sui Y, Zhou M, Zhou M, Han S, Zhang D. GPT4Table: can large language models understand structured table data? A benchmark and empirical study. 2023, arXiv preprint arXiv: 2305.13062

[77] Sui Y, Zou J, Zhou M, He X, Du L, Han S, Zhang D. TAP4LLM: table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 10306–10323

[78] Cheng Z, Xie T, Shi P, Li C, Nadkarni R, Hu Y, Xiong C, Radev D, Ostendorf M, Zettlemoyer L, Smith NA, Yu T. Binding language models in symbolic languages. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[79] Zhang W, Shen Y, Lu W, Zhuang Y. Data-copilot: bridging billions of data and humans with autonomous workflow. 2023, arXiv preprint arXiv: 2306.07209

[80] Zhang Z, Li X, Gao Y, Lou JG. CRT-QA: a dataset of complex reasoning question answering over tabular data. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 2131–2153, doi: 10.18653/v1/2023.emnlp-main.132

[81] Zhang Y, Henkel J, Floratou A, Cahoon J, Deep S, Patel JM. ReAcTable: enhancing react for table question answering. Proceedings of the VLDB Endowment, 2024, 17(8): 1981−1994

[82] Zhang H, Si Q, Fu P, Lin Z, Wang W. Are large language models table-based fact-checkers? In: Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design. 2024, 3086–3091

[83] Chen W. Large language models are few(1)-shot table reasoners. In: Proceedings of Findings of the Association for Computational Linguistics: EACL 2023. 2023, 1120–1130

[84] Luo T, Lei F, Lei J, Liu W, He S, Zhao J, Liu K. HRoT: hybrid prompt strategy and retrieval of thought for table-text hybrid question answering. 2023, arXiv preprint arXiv: 2309.12669

[85] Li H, Su J, Chen Y, Li Q, Zhang Z. SheetCopilot: bringing software productivity to the next level through large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 220

[86] Jiang J, Zhou K, Dong Z, Ye K, Zhao X, Wen JR. StructGPT: a general framework for large language model to reason over structured data. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 9237–9251, doi: 10.18653/v1/2023.emnlp-main.574

[87] Wang Z, Zhang H, Li CL, Eisenschlos JM, Perot V, Wang Z, Miculicich L, Fujii Y, Shang J, Lee CY, Pfister T. Chain-of-table: evolving tables in the reasoning chain for table understanding. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[88] Kong K, Zhang J, Shen Z, Srinivasan B, Lei C, Faloutsos C, Rangwala H, Karypis G. OpenTab: advancing large language models as open-domain table reasoners. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[89] Goyal T, Li JJ, Durrett G. News summarization and evaluation in the era of GPT-3. 2022, arXiv preprint arXiv: 2209.12356

[90] Ravaut M, Sun A, Chen NF, Joty S. On context utilization in summarization with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 2764–2781

[91] Bhaskar A, Fabbri AR, Durrett G. Prompted opinion summarization with GPT-3.5. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 9282–9300

[92] Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto

TB. Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics, 2023, 12: 39−57

[93] Zhang H, Liu X, Zhang J. Extractive summarization via chatgpt for faithful summary generation. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 3270–3278

[94] Adams G, Fabbri A, Ladhak F, Lehman E, Elhadad N. From sparse to dense: GPT-4 summarization with chain of density prompting. 2023, arXiv preprint arXiv: 2309.04269

[95] Tang Y, Puduppully R, Liu Z, Chen N. 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. 2023, 56–67, doi: 10.18653/v1/2023.newsum-1.6

[96] Chen M, Tworek J, Jun H, Yuan Q, Ponde de Oliveira Pinto H, et al. Evaluating large language models trained on code. 2021, arXiv preprint arXiv: 2107.03374

[97] Nijkamp E, Pang B, Hayashi H, Tu L, Wang H, Zhou Y, Savarese S, Xiong C. CodeGen: an open large language model for code with multi-turn program synthesis. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[98] Christopoulou F, Lampouras G, Gritta M, Zhang G, Guo Y, et al. PanGu-coder: program synthesis with function-level language modeling. 2022, arXiv preprint arXiv: 2207.11280

[99] Luo Z, Xu C, Zhao P, Sun Q, Geng X, Hu W, Tao C, Ma J, Lin Q, Jiang D. WizardCoder: Empowering code large language models with evol-instruct. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[100] Allal LB, Li R, Kocetkov D, Mou C, Akiki C, et al. SantaCoder: don't reach for the stars! 2023, arXiv preprint arXiv: 2301.03988

[101] Li R, Ben Allal L, Zi Y, Muennighoff N, Kocetkov D, et al. StarCoder: may the source be with you! Transactions on Machine Learning Research, 2023, 2023

[102] Li Y, Bubeck S, Eldan R, Del Giorno A, Gunasekar S, Lee YT. Textbooks are all you need II: phi-1.5 technical report. 2023, arXiv preprint arXiv: 2309.05463

[103] Guo D, Zhu Q, Yang D, Xie Z, Dong K, Zhang W, Chen G, Bi X, Wu Y, Li YK, Luo F, Xiong Y, Liang W. DeepSeek-coder: when the large language model meets programming — the rise of code intelligence. 2024, arXiv preprint arXiv: 2401.14196

[104] Roziere B, Gehring J, Gloeckle F, Sootla S, Gat I, et al. Code llama: open foundation models for code. 2023, arXiv preprint arXiv: 2308.12950

[105] Zheng Q, Xia X, Zou X, Dong Y, Wang S, Xue Y, Wang Z, Shen L, Wang A, Li Y, Su T, Yang Z, Tang J. CodeGeeX: a pre-trained model for code generation with multilingual evaluations on HumanEval-X. 2023, arXiv preprint arXiv: 2303.17568

[106] Wei X, Wei H, Lin H, Li T, Zhang P, Ren X, Li M, Wan Y, Cao Z, Xie B, Hu T, Li S, Hui B, Yu B, Liu D, Yang B, Huang F, Xie J. PolyLM: an open source polyglot large language model. 2023, arXiv preprint arXiv: 2307.06018

[107] Zhu W, Lv Y, Dong Q, Yuan F, Xu J, Huang S, Kong L, Chen J, Li L. Extrapolating large language models to non-english by aligning languages. 2023, arXiv preprint arXiv: 2308.04948

[108] Li C, Liu M, Zhang H, Chen Y, Xu J, Zhou M. MT2: towards a multi-task machine translation model with translation-specific in-context learning. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 8616–8627

[109] Li J, Tang Z, Ding Y, Wang P, Guo P, You W, Qiao D, Chen W, Fu G, Zhu Q, Zhou G, Zhang M. OpenBA: an open-sourced 15B bilingual asymmetric seq2seq model pre-trained from scratch. 2023, arXiv preprint arXiv: 2309.10706

[110] Alves DM, Guerreiro NM, Alves J, Pombal J, Rei R, de Souza J, Colombo P, Martins A. Steering large language models for machine translation with finetuning and in-context learning. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 11127–11148

[111] Raunak V, Awadalla HH, Menezes A. Dissecting in-context learning of translations in GPTs. 2023, arXiv preprint arXiv: 2310.15987

[112] Lu H, Yang H, Huang H, Zhang D, Lam W, Wei F. Chain-of-dictionary prompting elicits translation in large language models. In: Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 958–976

[113] Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[114] Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A, Zhou D. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[115] Lu P, Qiu L, Chang KW, Wu YN, Zhu SC, Rajpurohit T, Clark P, Kalyan A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[116] Gao L, Madaan A, Zhou S, Alon U, Liu P, Yang Y, Callan J, Neubig G. PAL: program-aided language models. In: Proceedings of the 40th International Conference on Machine Learning. 2023, 10764–10799

[117] Das D, Banerjee D, Aditya S, Kulkarni A. MATHSENSEI: a tool-augmented large language model for mathematical reasoning. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024, 942–966

[118] Wang Z, Xia R, Yu J. UnifiedABSA: a unified ABSA framework based on multi-task instruction tuning. 2022, arXiv preprint arXiv: 2211.10986

[119] Varia S, Wang S, Halder K, Vacareanu R, Ballesteros M, Benajiba Y, John NA, Anubhai R, Muresan S, Roth D. Instruction tuning for few-shot aspect-based sentiment analysis. In: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. 2023, 19–27

[120] Yang B, Li J. Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 6062–6075

[121] Zhao J, Liu K, Xu L. Sentiment analysis: mining opinions, sentiments, and emotions. Computational Linguistics, 2016, 42(3): 595−598

[122] Qiu H, He H, Zhang S, Li A, Lan Z. SMILE: single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 615–636

[123] Lu D, Ran S, Tetreault J, Jaimes A. Event extraction as question generation and answering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023, 1666–1688

[124] Gan C, Zhang Q, Mori T. GIELLM: Japanese general information extraction large language model utilizing mutual reinforcement effect. 2023, arXiv preprint arXiv: 2311.06838

[125] Sainz O, García-Ferrero I, Agerri R, Lopez de Lacalle O, Rigau G, Agirre E. GoLLIE: annotation guidelines improve zero-shot information-extraction. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[126] Wang X, Zhou W, Zu C, Xia H, Chen T, Zhang Y, Zheng R, Ye J, Zhang Q, Gui T, Kang J, Yang J, Li S, Du C. InstructUIE: multi-task instruction tuning for unified information extraction. 2023, arXiv preprint arXiv: 2304.08085

[127] Snigdha Sarathi Das S, Zhang RH, Shi P, Yin W, Zhang R. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6998–7010, doi: 10.18653/v1/2023.emnlp-main.433

[128] Liang Z, Wei F, Jie Y, Qian Y, Hao Z, Han B. 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). 2023, 277–292, doi: 10.18653/v1/2023.acl-long.16

[129] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA, Jain A. Structured information extraction from scientific text with large language models. Nature Communications, 2024, 15(1): 1418

[130] Xue L, Zhang D, Dong Y, Tang J. AutoRE: document-level relation extraction with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 2024, 211–220

[131] Rixewa D, Anderson K, Dubois L, Harrington M. Interleaved multi-modal document representations for large-scale information retrieval using large language models. 2024

[132] Xie T, Wu CH, Shi P, Zhong R, Scholak T, et al. UnifiedSKG: unifying and multi-tasking structured knowledge grounding with text-to-text language models. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 602–631

[133] Zhao J, Gupta R, Cao Y, Yu D, Wang M, Lee H, Rastogi A, Shafran I, Wu Y. Description-driven task-oriented dialog modeling. 2022, arXiv preprint arXiv: 2201.08904

[134] Gupta R, Lee H, Zhao J, Cao Y, Rastogi A, Wu Y. Show, don't tell: demonstrations outperform descriptions for schema-guided task-oriented dialogue. In: Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022, 4541–4549

[135] Yu D, Wang M, Cao Y, El Shafey L, Shafran I, Soltau H. Knowledge-grounded dialog state tracking. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2022. 2022, 3428–3435

[136] Feng Y, Lu Z, Liu B, Zhan L, Wu XM. Towards LLM-driven dialogue state tracking. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 739–755

[137] Liu H, Cai Y, Zhou Y, Ou Z, Huang Y, Feng J. Prompt pool based class-incremental continual learning for dialog state tracking. In: Proceedings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2023, 1–8

[138] Li P, He Y, Yashar D, Cui W, Ge S, Zhang H, Fainman DR, Zhang D, Chaudhuri S. Table-GPT: table-tuned GPT for diverse table tasks. 2023, arXiv preprint arXiv: 2310.09263

[139] Xue S, Jiang C, Shi W, Cheng F, Chen K, Yang H, Zhang Z, He J, Zhang H, Wei G, Zhao W, Zhou F, Qi D, Yi H, Liu S, Chen F. DB-GPT: empowering database interactions with private large language models. 2023, arXiv preprint arXiv: 2312.17449

[140] Zhang H, Dong Y, Xiao C, Oyamada M. Jellyfish: a large language model for data preprocessing. 2023, arXiv preprint arXiv: 2312.01678

[141] Zhu F, Liu Z, Feng F, Wang C, Li M, Chua TS. TAT-LLM: a specialized language model for discrete reasoning over tabular and textual data. 2024, arXiv preprint arXiv: 2401.13223

[142] Bai F, Kang J, Stanovsky G, Freitag D, Ritter A. Schema-driven information extraction from heterogeneous tables. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 10252–10273

[143] Zhang T, Yue X, Li Y, Sun H. TableLlama: towards open large generalist models for tables. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024, 6024–6044

[144] He X, Liu Y, Zhou M, He Y, Dong H, Han S, Yuan Z, Zhang D. TableLoRA: low-rank adaptation on table structure understanding for large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 22376–22391

[145] Li P, He Y, Yashar D, Cui W, Ge S, Zhang H, Fainman DR, Zhang D, Chaudhuri S. Table-GPT: table fine-tuned GPT for diverse table tasks. Proceedings of the ACM on Management of Data, 2024, 2(3): 176

[146] Pagnoni A, Fabbri AR, Kryscinski W, Wu CS. Socratic pretraining: question-driven pretraining for controllable summarization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 12737–12755

[147] Zhao L, Zheng F, Zeng W, He K, Xu W, Jiang H, Wu W, Wu Y. Domain-oriented prefix-tuning: towards efficient and generalizable fine-tuning for zero-shot dialogue summarization. In: Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022, 4848–4862, doi: 10.18653/v1/2022.naacl-main.357

[148] Yuan R, Wang Z, Cao Z, Li W. Few-shot query-focused summarization with prefix-merging. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 3704–3714

[149] Feng X, Feng X, Du X, Kan MY, Qin B. Adapter-based selective knowledge distillation for federated multi-domain meeting summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3694−3708

[150] Ravaut M, Chen H, Zhao R, Qin C, Joty S, Chen N. PromptSum: parameter-efficient controllable abstractive summarization. 2023, arXiv preprint arXiv: 2308.03117

[151] Wang Y, Wang W, Joty S, Hoi SCH. CodeT5: identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 8696–8708

[152] Wang Y, Le H, Gotmare AD, Bui NDQ, Li J, Hoi SCH. CodeT5+: open code large language models for code understanding and generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 1069–1088

[153] Le H, Wang Y, Gotmare AD, Savarese S, Hoi SCH. CodeRL: mastering code generation through pretrained models and deep reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 1549

[154] Shojaee P, Jain A, Tipirneni S, Reddy CK. Execution-based code generation using deep reinforcement learning. Transactions on Machine Learning Research, 2023, 2023

[155] Ayupov S, Chirkova N. Parameter-efficient finetuning of transformers for source code. 2022, arXiv preprint arXiv: 2212.05901

[156] Zhuo TY, Zebaze A, Suppattarachai N, von Werra L, de Vries H, Liu Q, Muennighoff N. Astraios: parameter-efficient instruction tuning code large language models. 2024, arXiv preprint arXiv: 2401.00788

[157] Weyssow M, Zhou X, Kim K, Lo D, Sahraoui H. Exploring parameter-efficient fine-tuning techniques for code generation with large language models. ACM Transactions on Software Engineering and Methodology, 2025, 34(7): 204

[158] Xu H, Kim YJ, Sharaf A, Awadalla HH. A paradigm shift in machine translation: boosting translation performance of large language models. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[159] Xu H, Sharaf A, Chen Y, Tan W, Shen L, Van Durme B, Murray K, Kim YJ. Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation. In: Proceedings of the 41st International Conference on Machine Learning. 2024

[160] Iyer V, Chen P, Birch A. Towards effective disambiguation for machine translation with large language models. In: Proceedings of the 8th Conference on Machine Translation. 2023, 482–495

[161] Moslem Y, Haque R, Way A. Fine-tuning large language models for adaptive machine translation. 2023, arXiv preprint arXiv: 2312.12740

[162] Üstün A, Stickland AC. When does parameter-efficient transfer learning work for machine translation? In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 7919–7933

[163] Wu B, Yuan F, Zhao H, Li L, Xu J. Extrapolating multilingual understanding models as multilingual generators. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 15432–15444

[164] Wu M, Vu TT, Qu L, Foster G, Haffari G. Adapting large language models for document-level machine translation. 2024, arXiv preprint arXiv: 2401.06468

[165] Luo H, Sun Q, Xu C, Zhao P, Lou JG, Tao C, Geng X, Lin Q, Chen S, Tang Y, Zhang D. WizardMath: empowering mathematical reasoning for large language models via reinforced evol-instruct. In: Proceedings of the Thirteenth International Conference on Learning Representations. 2025

[166] Yue X, Qu X, Zhang G, Fu Y, Huang W, Sun H, Su Y, Chen W. MammoTH: building math generalist models through hybrid instruction tuning. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[167] Ho N, Schmid L, Yun SY. Large language models are reasoning teachers. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 14852–14882, doi: 10.18653/v1/2023.acl-long.830

[168] Schick T, Dwivedi-Yu J, Dessí R, Raileanu R, Lomeli M, Hambro E, Zettlemoyer L, Cancedda N, Scialom T. Toolformer: language models can teach themselves to use tools. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 2997

[169] Hu Z, Wang L, Lan Y, Xu W, Lim EP, Bing L, Xu X, Poria S, Lee RKW. LLM-adapters: an adapter family for parameter-efficient fine-tuning of large language models. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 5254–5276

[170] Shi W, Hu Z, Bin Y, Liu J, Yang Y, Ng SK, Bing L, Lee RKW. Math-LLaVA: bootstrapping mathematical reasoning for multimodal large language models. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 4663–4680

[171] Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, Zhang H, Zhang M, Li YK, Wu Y, Guo D. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. 2024, arXiv preprint arXiv: 2402.03300

[172] Luo L, Liu Y, Liu R, Phatale S, Guo M, Lara H, Li Y, Shu L, Zhu Y, Meng L, Sun J, Rastogi A. Improve mathematical reasoning in language models by automated process supervision. 2024, arXiv preprint arXiv: 2406.06592

[173] Chen C, Wang X, Lin TE, Lv A, Wu Y, Gao X, Wen JR, Yan R, Li Y. Masked thought: simply masking partial reasoning steps can improve mathematical reasoning learning of language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 5872–5900

[174] Liu T, Guo Q, Yang Y, Hu X, Zhang Y, Qiu X, Zhang Z. Plan, verify and switch: integrated reasoning with diverse X-of-thoughts. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 2807–2822

[175] Yu X, Zhou B, Cheng H, Roth D. ReasonAgain: using extractable symbolic programs to evaluate mathematical reasoning. 2024, arXiv preprint arXiv: 2410.19056

[176] Ranaldi L, Valentino M, Freitas A. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 17222–17240

[177] Srivastava G, Bi Z, Lu M, Wang X. DEBATE, TRAIN, EVOLVE: self evolution of language model reasoning. 2025, arXiv preprint arXiv: 2505.15734

[178] Cai H, Yang Y, Li Z. System-2 mathematical reasoning via enriched instruction tuning. 2024, arXiv preprint arXiv: 2412.16964

[179] Xu D, Chen W, Peng W, Zhang C, Xu T, Zhao X, Wu X, Zheng Y, Wang Y, Chen E. Large language models for generative information extraction: a survey. Frontiers of Computer Science, 2024, 18(6): 186357

[180] Siepmann RM, Baldini G, Schmidt CS, Truhn D, Müller-Franzes GA, Dada A, Kleesiek J, Nensa F, Hosch R. An automated information extraction model for unstructured discharge letters using large language models and GPT-4. Healthcare Analytics, 2025, 7: 100378

[181] Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, Desai RJ. Scalable information extraction from free text electronic health records using large language models. BMC Medical Research Methodology, 2025, 25(1): 23

[182] Xin Y, Luo S, Zhou H, Du J, Liu X, Fan Y, Li Q, Du Y. Parameter-efficient fine-tuning for pre-trained vision models: a survey. 2024, arXiv preprint arXiv: 2402.02242

[183] Xin Y, Luo S, Liu X, Du Y, Zhou H, Cheng X, Lee C, Du J, Wang H, Chen M, Liu T, Hu G, Wan Z, Zhang R, Li A, Yi M, Liu X. *V-PETL bench*: a unified visual parameter-efficient transfer learning benchmark. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2025, 2560

[184] Qin L, Xie T, Che W, Liu T. A survey on spoken language understanding: recent advances and new frontiers. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence., 2021, 4577–4584, doi: 10.24963/ijcai.2021/622

[185] Sarikaya R, Crook PA, Marin A, Jeong M, Robichaud JP, Celikyilmaz A, Kim YB, Rochette A, Khan OZ, Liu X, Boies D, Anastasakos T, Feizollahi Z, Ramesh N, Suzuki H, Holenstein R, Krawczyk E, Radostev V. An overview of end-to-end language understanding and dialog management for personal digital assistants. In: Proceedings of 2016 IEEE Spoken Language Technology Workshop (SLT). 2016, 391–397

[186] Yoon Y, Lee J, Kim K, Park C, Kim T. BlendX: complex multi-intent detection with blended patterns. In: Proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, 2428–2439

[187] Qin L, Chen Q, Zhou J, Wang J, Fei H, Che W, Li M. Divide-solve-combine: an interpretable and accurate prompting framework for zero-shot multi-intent detection. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 25038–25046

[188] Qin L, Wei F, Chen Q, Zhou J, Huang S, Si J, Lu W, Che W. CroPrompt: cross-task interactive prompting for zero-shot spoken language understanding. In: Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025, 1–5

[189] Dong W, Chen S, Yang Y. ProTOD: proactive task-oriented dialogue system based on large language model. In: Proceedings of the 31st International Conference on Computational Linguistics. 2025, 9147–9164

[190] Acikgoz EC, Greer J, Datta A, Yang Z, Zeng W, Elachqar O, Koukoumidis E, Hakkani-Tür D, Tur G. Can a single model master both multi-turn conversations and tool use? CALM: a unified conversational agentic language model. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 12370–12390

[191] Yin S, Huang P, Xu Y. MIDLM: multi-intent detection with bidirectional large language models. In: Proceedings of the 31st International Conference on Computational Linguistics. 2025, 2616–2625

[192] Jin N, Siebert J, Li D, Chen Q. A survey on table question answering: recent advances. In: Proceedings of the 7th China Conference on Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy. 2022, 174–186

[193] Wang D, Dou L, Che W. A survey on table-and-text HybridQA: concepts, methods, challenges and future directions. 2022, arXiv preprint arXiv: 2212.13465

[194] Zhang X, Wang D, Dou L, Zhu Q, Che W. A survey of table reasoning with large language models. Frontiers of Computer Science, 2025, 19(9): 199348

[195] Zhang X, Wang D, Xu K, Zhu Q, Che W. RoT: enhancing table reasoning with iterative row-wise traversals. 2025, arXiv preprint arXiv: 2505.15110

[196] Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK. Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science, 2018, 2(1): 1

[197] Godbole A, George JG, Shandilya S. Leveraging long-context large language models for multi-document understanding and summarization in enterprise applications. In: Proceedings of the 1st International Conference on Business Intelligence, Computational Mathematics, and Data Analytics. 2025, 208–224

[198] Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. Royal Society Open Science, 2025, 12(4): rsos241776

[199] Yun J, Choi J, Jin K, Jang S, Jang J, Kim Y. SummPilot: bridging efficiency and customization for interactive summarization system. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 29724–29726

[200] Qorib MR, Hu Q, Ng HT. Just what you desire: constrained timeline summarization with self-reflection for enhanced relevance. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 25065–25073

[201] Zhu DH, Xiong YJ, Zhang JC, Xie XJ, Xia CM. Understanding before reasoning: enhancing chain-of-thought with iterative summarization pre-prompting. 2025, arXiv preprint arXiv: 2501.04341

[202] Nandy A, Bandyopadhyay S. Language models of code are few-shot planners and reasoners for multi-document summarization with attribution. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 24930–24938

[203] Li Y, Peng B, He P, Galley M, Yu Z, Gao J. DIONYSUS: a pre-trained model for low-resource dialogue summarization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 1368–1386

[204] Wang L, Wu L, Song S, Wang Y, Gao C, Wang K. Distilling structured rationale from large language models to small language models for abstractive summarization. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 25389–25397

[205] Lu YJ, Hu TY, Koppula HS, Pouransari H, Chang JHR, Xia Y, Kong X, Zhu Q, Wang XS, Tuzel O, Vemulapalli R. Mutual

reinforcement of LLM dialogue synthesis and summarization capabilities for few-shot dialogue summarization. In: Proceedings of Findings of the Association for Computational Linguistics: NAACL 2025. 2025, 7237–7256

[206] Aali A, Van Veen D, Arefeen YI, Hom J, Bluethgen C, Reis EP, Gatidis S, Clifford N, Daws J, Tehrani AS, Kim J, Chaudhari AS. A dataset and benchmark for hospital course summarization with adapted large language models. Journal of the American Medical Informatics Association, 2025, 32(3): 470−479

[207] Wu J, Ning L, Liu L, Lee H, Wu N, Wang C, Prakash S, O' Banion S, Green B, Xie J. RLPF: reinforcement learning from prediction feedback for user summarization with LLMs. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 25488–25496

[208] Zhao H, Hui J, Howland J, Nguyen N, Zuo S, et al. CodeGemma: open code models based on gemma. 2024, arXiv preprint arXiv: 2406.11409

[209] Hui B, Yang J, Cui Z, Yang J, Liu D, et al. Qwen2.5-coder technical report. 2024, arXiv preprint arXiv: 2409.12186

[210] Dance-Seed B. Seed-coder: let the code model curate data for itself. See github.com/ByteDance-Seed/Seed-Coder/blob/master/Seed-Coder website, 2025

[211] Cai W, Cao Y, Chen C, Chen C, Chen S, et al. Every sample matters: leveraging mixture-of-experts and high-quality data for efficient and accurate code LLM. 2025, arXiv preprint arXiv: 2503.17793

[212] Lu S, Guo D, Ren S, Huang J, Svyatkovskiy A, et al. CodeXGLUE: a machine learning benchmark dataset for code understanding and generation. In: Proceedings of the 1st Neural Information Processing Systems Track on Datasets and Benchmarks. 2021

[213] Ye Y, Zhang T, Jiang W, Huang H. Process-supervised reinforcement learning for code generation. 2025, arXiv preprint arXiv: 2502.01715

[214] Zhang K, Li G, Li J, Dong Y, Jin Z. Focused-DPO: enhancing code generation through focused preference optimization on error-prone points. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2025. 2025, 9578–9591

[215] Zeng H, Jiang D, Wang H, Nie P, Chen X, Chen W. ACECODER: acing coder RL via automated test-case synthesis. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 12023–12040

[216] Wei Y, Duchenne O, Copet J, Carbonneaux Q, Zhang L, Fried D, Synnaeve G, Singh R, Wang SI. SWE-RL: advancing LLM reasoning via reinforcement learning on open software evolution. 2025, arXiv preprint arXiv: 2502.18449

[217] Storhaug A, Li J. Parameter-efficient fine-tuning of large language models for unit test generation: an empirical study. 2024, arXiv preprint arXiv: 2411.02462

[218] Zhang B, Liang P, Zhou X, Zhou X, Lo D, Feng Q, Li Z, Li L. A comprehensive evaluation of parameter-efficient fine-tuning on method-level code smell detection. 2024, arXiv preprint arXiv: 2412.13801

[219] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of 3rd International Conference on Learning Representations. 2015

[220] Pang J, Ye F, Wong DF, Yu D, Shi S, Tu Z, Wang L. Salute the classic: revisiting challenges of machine translation in the age of large language models. Transactions of the Association for Computational Linguistics, 2025, 13: 73−95

[221] Huang Y, Li B, Feng X, Huo W, Fu C, Liu T, Qin B. Aligning translation-specific understanding to general understanding in large language models. In: Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 5028–5041

[222] Zhu S, Cui M, Xiong D. Towards robust in-context learning for machine translation with large language models. In: Proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, 16619–16629

[223] Feng Z, Cao S, Ren J, Su J, Chen R, Zhang Y, Xu Z, Hu Y, Wu J, Liu Z. MT-R1-zero: advancing LLM-based machine translation via R1-zero-like reinforcement learning. 2025, arXiv preprint arXiv: 2504.10160

[224] Feng Z, Chen R, Zhang Y, Meng Z, Liu Z. Ladder: a model-agnostic framework boosting LLM-based machine translation to the next level. In: Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 15377–15393

[225] Lu P, Qiu L, Yu W, Welleck S, Chang KW. 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). 2023, 14605–14631

[226] Yan Y, Wang S, Huo J, Yu PS, Hu X, Wen Q. Mathagent: leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. 2025, arXiv preprint arXiv: 2503.18132

[227] OpenAI. GPT-4 technical report. 2023, arXiv preprint arXiv: 2303.08774

[228] Tang Y, Zhan Y, Zan C, Lan L, Che Y. Elevating large language model reasoning ability with auto-enhanced zero-shot prompts. Mathematical Foundations of Computing, 2025

[229] Yuksekgonul M, Bianchi F, Boen J, Liu S, Lu P, Huang Z, Guestrin C, Zou J. Optimizing generative AI by backpropagating language model feedback. Nature, 2025, 639(8055): 609−616

[230] Peng D, Zhou Y, Chen Q, Liu J, Chen J, Qin L. DLPO: towards a robust, efficient, and generalizable prompt optimization framework from a deep-learning perspective. 2025, arXiv preprint arXiv: 2503.13413

[231] Zhang B, Liu Y, Dong X, Zang Y, Zhang P, Duan H, Cao Y, Lin D, Wang J. BoostStep: boosting mathematical capability of large language models via improved single-step reasoning. 2025, arXiv preprint arXiv: 2501.03226

[232] Pang B, Dong H, Xu J, Savarese S, Zhou Y, Xiong C. BOLT: bootstrap long chain-of-thought in language models without distillation. 2025, arXiv preprint arXiv: 2502.03860

[233] Yu Y, Zhang Y, Zhang D, Liang X, Zhang H, Zhang X, Khademi M, Awadalla HH, Wang J, Yang Y, Wei F. Chain-of-reasoning: towards unified mathematical reasoning in large language models via a multi-paradigm perspective. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 24914–24937

[234] Qian C, Acikgoz EC, He Q, Wang H, Chen X, Hakkani-Tür D, Tur G, Ji H. ToolRL: reward is all tool learning needs. 2025, arXiv preprint arXiv: 2504.13958

[235] Singh J, Chakraborty T, Nambi A. Self-evolved preference optimization for enhancing mathematical reasoning in small language models. 2025, arXiv preprint arXiv: 2503.04813

[236] Prottasha NJ, Mahmud A, Sobuj MSI, Bhat P, Kowsher M, Yousefi N, Garibay OO. Parameter-efficient fine-tuning of large language models using semantic knowledge tuning. Scientific Reports, 2024, 14(1): 30667

[237] Alazraki L, Rei M. Meta-reasoning improves tool use in large language models. In: Proceedings of Findings of the Association for Computational Linguistics: NAACL 2025. 2025, 7885–7897

[238] Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, Li M, Che W, Yu PS. Multilingual large language model: a survey of resources, taxonomy and frontiers. 2024, arXiv preprint arXiv: 2404.04925

[239] Winata G, Aji AF, Yong ZX, Solorio T. The decades progress on code-switching research in NLP: a systematic survey on trends and challenges. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 2936–2978, doi: 10.18653/v1/2023.findings-acl.185

[240] Li Z, Shi Y, Liu Z, Yang F, Payani A, Liu N, Du M. Language ranker: a metric for quantifying LLM performance across high and low-resource languages. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence. 2025, 28186–28194

[241] Wang P, Tao R, Chen Q, Hu M, Qin L. X-WebAgentBench: a multilingual interactive web benchmark for evaluating global agentic system. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2025. 2025, 19320–19335

[242] Zhang Y, Liu X, Zhou R, Chen Q, Fei H, Lu W, Qin L. CCHaLL: a novel benchmark for joint cross-lingual and cross-modal hallucinations detection in large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 30728–30749

[243] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mt5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021, 483–498

[244] Le Scao T, Fan A, Akiki C, Pavlick E, Ilić S, et al. BLOOM: a 176B-parameter open-access multilingual language model. 2022, arXiv preprint arXiv: 2211.05100

[245] Chen P, Ji S, Bogoychev N, Kutuzov A, Haddow B, Heafield K. Monolingual or multilingual instruction tuning: which makes a better alpaca. In: Proceedings of Findings of the Association for Computational Linguistics: EACL 2024. 2024, 1347–1356

[246] Cahyawijaya S, Lovenia H, Yu T, Chung W, Fung P. Instruct-align: teaching novel languages with to LLMs through alignment-based cross-lingual instruction. 2023, arXiv preprint arXiv: 2305.13627

[247] Li J, Zhou H, Huang S, Cheng S, Chen J. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Transactions of the Association for Computational Linguistics, 2024, 12: 576−592

[248] Bajpai A, Chakraborty T. Multilingual LLMs inherently reward in-language time-sensitive semantic alignment for low-resource languages. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 23469–23477

[249] Winata GI, Madotto A, Lin Z, Liu R, Yosinski J, Fung P. Language models are few-shot multilingual learners. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. 2021, 1–15

[250] Shi F, Suzgun M, Freitag M, Wang X, Srivats S, Vosoughi S, Chung HW, Tay Y, Ruder S, Zhou D, Das D, Wei J. Language models are multilingual chain-of-thought reasoners. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[251] Lin XV, Mihaylov T, Artetxe M, Wang T, Chen S, Simig D, Ott M, Goyal N, Bhosale S, Du J, Pasunuru R, Shleifer S, Koura PS, Chaudhary V, O'Horo B, Wang J, Zettlemoyer L, Kozareva Z, Diab M, Stoyanov V, Li X. Few-shot learning with multilingual generative language models. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 9019–9052, doi: 10.18653/v1/2022.emnlp-main.616

[252] Tanwar E, Dutta S, Borthakur M, Chakraborty T. Multilingual LLMs are better cross-lingual in-context learners with alignment. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 6292–6307, doi: 10.18653/v1/2023.acl-long.346

[253] Qin L, Chen Q, Wei F, Huang S, Che W. Cross-lingual prompting: improving zero-shot chain-of-thought reasoning across languages. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 2695–2709

[254] Huang H, Tang T, Zhang D, Zhao X, Song T, Xia Y, Wei F. Not all languages are created equal in LLMs: improving multilingual capability by cross-lingual-thought prompting. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 12365–12394, doi: 10.18653/v1/2023.findings-emnlp.826

[255] Huang Z, Xu X, Ni J, Zhu H, Wang C. Multimodal representation learning for recommendation in internet of things. IEEE Internet of Things Journal, 2019, 6(6): 10675−10685

[256] Wang Y, Wu S, Zhang Y, Yan S, Liu Z, Luo J, Fei H. Multimodal chain-of-thought reasoning: a comprehensive survey. 2025, arXiv preprint arXiv: 2503.12605

[257] Li X, Qiao J, Yin S, Wu L, Gao C, Wang Z, Li X. A survey of multimodal fake news detection: a cross-modal interaction perspective. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9(4): 2658−2675

[258] Peng Y, Wang X, Wei Y, Pei J, Qiu W, Jian A, Hao Y, Pan J, Xie T, Ge L, Zhuang R, Song X, Liu Y, Zhou Y. Skywork R1V: pioneering multimodal reasoning with chain-of-thought. 2025, arXiv preprint arXiv: 2504.05599

[259] Lu P, Mishra S, Xia T, Qiu L, Chang KW, Zhu SC, Tafjord O, Clark P, Kalyan A. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 182

[260] Qin L, Huang S, Chen Q, Cai C, Zhang Y, Liang B, Che W, Xu R. MMSD2.0: towards a reliable multi-modal sarcasm detection system. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 10834–10845

[261] Qin L, Wang W, Chen Q, Che W. CLIPText: a new paradigm for zero-shot text classification. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2023. 2023, 1077–1088

[262] Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, Wang L. The dawn of LMMs: preliminary explorations with GPT-4V (ision). 2023, arXiv preprint arXiv: 2309.17421

[263] Fei H, Wu S, Ji W, Zhang H, Zhang M, Lee ML, Hsu W. Video-of-thought: step-by-step video reasoning from perception to cognition. In: Proceedings of the 41st International Conference on Machine Learning. 2024, 13109–13125

[264] Qin L, Chen Q, Fei H, Chen Z, Li M, Che W. What factors affect multi-modal in-context learning? an in-depth exploration. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024

[265] Zhang Y, Liu X, Tao R, Chen Q, Fei H, Che W, Qin L. ViTCoT: video-text interleaved chain-of-thought for boosting video understanding in large language models. 2025, arXiv preprint arXiv: 2507.09876

[266] Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y, Ji J, Yang Z, Zhao L, Song X, Xu J, Xu B, Li J, Dong Y, Ding M, Tang J. CogVLM: visual expert for pretrained language models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024

[267] Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024, 26286–26296

[268] Chen Q, Qin L, Zhang J, Chen Z, Xu X, Che W. M^3CoT: a novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 8199–8221

[269] Yang Z, Li L, Wang J, Lin K, Azarnasab E, Ahmed F, Liu Z, Liu C, Zeng M, Wang L. MM-REACT: prompting chatGPT for multimodal reasoning and action. 2023, arXiv preprint arXiv: 2303.11381

[270] Lu P, Bansal H, Xia T, Liu J, Li C, Hajishirzi H, Cheng H, Chang KW, Galley M, Gao J. MathVista: evaluating mathematical reasoning of foundation models in visual contexts. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[271] Zhang Z, Zhang A, Li M, Zhao H, Karypis G, Smola A. Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research, 2024, 2024

[272] Cheng Z, Chen Q, Zhang J, Fei H, Feng X, Che W, Li M, Qin L. CoMT: a novel benchmark for chain of multi-modal thought on large vision-language models. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 23678–23686

[273] Cheng Z, Chen Q, Xu X, Wang J, Wang W, Fei H, Wang Y, Wang AJ, Chen Z, Che W, Qin L. Visual thoughts: a unified perspective of understanding multimodal chain-of-thought. 2025, arXiv preprint arXiv: 2505.15510

[274] Wu Y, Zhang P, Xiong W, Oguz B, Gee JC, Nie Y. The role of chain-of-thought in complex vision-language reasoning task. 2023, arXiv preprint arXiv: 2311.09193

[275] Mitra C, Huang B, Darrell T, Herzig R. Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024, 14420–14431

[276] Wang P, Zhang Y, Fei H, Chen Q, Wang Y, Si J, Lu W, Li M, Qin L. S^3 agent: unlocking the power of VLLM for zero-shot multi-modal sarcasm detection. ACM Transactions on Multimedia Computing, Communications and Applications, 2024

[277] Qin Y, Liang S, Ye Y, Zhu K, Yan L, Lu Y, Lin Y, Cong X, Tang X, Qian B, Zhao S, Hong L, Tian R, Xie R, Zhou J, Gerstein M, Li D, Liu Z, Sun M. ToolLLM: facilitating large language models to master 16000+ real-world APIs. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[278] Hu M, Mu Y, Yu XC, Ding M, Wu S, Shao W, Chen Q, Wang B, Qiao Y, Luo P. Tree-planner: efficient close-loop task planning with large language models. In: Proceedings of the 12th International Conference on Learning Representations. 2024

[279] Shinn N, Cassano F, Gopinath A, Narasimhan KR, Yao S. Reflexion: language agents with verbal reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 377

[280] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J. A survey on large language model based autonomous agents. Frontiers of Computer Science, 2024, 18(6): 186345

[281] Zhu X, Chen Y, Tian H, Tao C, Su W, Yang C, Huang G, Li B, Lu L, Wang X, Qiao Y, Zhang Z, Dai J. Ghost in the minecraft: generally capable agents for open-world environments via large language models with text-based knowledge and memory. 2023, arXiv preprint arXiv: 2305.17144

[282] Hu M, Chen T, Chen Q, Mu Y, Shao W, Luo P. HiAgent: hierarchical working memory management for solving long-horizon agent tasks with large language model. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 32779–32798

[283] Zhang G, Niu L, Fang J, Wang K, Bai L, Wang X. Multi-agent architecture search via agentic supernet. 2025, arXiv preprint arXiv: 2502.04180

[284] Yue Y, Zhang G, Liu B, Wan G, Wang K, Cheng D, Qi Y. MasRouter: learning to route LLMs for multi-agent systems. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 15549–15572

[285] Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K. Tree of thoughts: deliberate problem solving with large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 517

[286] Chen W, Ma X, Wang X, Cohen WW. Program of thoughts prompting: disentangling computation from reasoning for numerical reasoning tasks. Transactions on Machine Learning Research, 2023, 2023

[287] Lei B, Lin PH, Liao C, Ding C. Boosting logical reasoning in large language models through a new framework: the graph of thought. 2023, arXiv preprint arXiv: 2308.08614

[288] Zhang Y, Chen Q, Zhou J, Wang P, Si J, Wang J, Lu W, Qin L. Wrong-of-thought: an integrated reasoning framework with multi-perspective verification and wrong information. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 6644–6653

[289] Muhlgay D, Ram O, Magar I, Levine Y, Ratner N, Belinkov Y, Abend O, Leyton-Brown K, Shashua A, Shoham Y. Generating benchmarks for factuality evaluation of language models. In: Proceedings

of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 49–66

[290] Min S, Krishna K, Lyu X, Lewis M, Yih WT, Koh PW, Iyyer M, Zettlemoyer L, Hajishirzi H. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 12076–12100

[291] Adlakha V, BehnamGhader P, Lu XH, Meade N, Reddy S. Evaluating correctness and faithfulness of instruction-following models for question answering. Transactions of the Association for Computational Linguistics, 2024, 12: 681−699

[292] Liu T, Zhang Y, Brockett C, Mao Y, Sui Z, Chen W, Dolan WB. 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). 2022, 6723–6737

[293] Chang KK, Cramer M, Soni S, Bamman D. Speak, memory: an archaeology of books known to ChatGPT/GPT-4. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 7312–7327

[294] Hartvigsen T, Gabriel S, Palangi H, Sap M, Ray D, Kamar E. 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). 2022, 3309–3326, doi: 10.18653/v1/2022.acl-long.234

[295] Wan Y, Wang W, He P, Gu J, Bai H, Lyu MR. BiasAsker: measuring the bias in conversational AI system. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2023, 515–527

[296] Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang KW, Gupta R. BOLD: dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021, 862–872

[297] Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. 2022, arXiv preprint arXiv: 2209.07858

[298] Sun H, Zhang Z, Deng J, Cheng J, Huang M. Safety assessment of Chinese large language models. 2023, arXiv preprint arXiv: 2304.10436

[299] Pan W, Liu Z, Chen Q, Zhou X, Yu H, Jia X. The hidden dimensions of LLM alignment: a multi-dimensional safety analysis. 2025, arXiv preprint arXiv: 2502.09674

[300] Yu M, Meng F, Zhou X, Wang S, Mao J, Pan L, Chen T, Wang K, Li X, Zhang Y, An B, Wen Q. A survey on trustworthy LLM agents: threats and countermeasures. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. 2025, 6216–6226

[301] Xu Y, Hu L, Zhao J, Qiu Z, Xu K, Ye Y, Gu H. A survey on multilingual large language models: corpora, alignment, and bias. Frontiers of Computer Science, 2025, 19(11): 1911362

[302] Li ZZ, Zhang D, Zhang ML, Zhang J, Liu Z, et al. From system 1 to system 2: a survey of reasoning large language models. 2025, arXiv preprint arXiv: 2502.17419

[303] Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, et al. OpenAI o1 system card. 2024, arXiv preprint arXiv: 2412.16720

[304] Li LH, Hessel J, Yu Y, Ren X, Chang KW, Choi Y. Symbolic chain-of-thought distillation: small models can also “think” step-by-step. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 2665–2679, doi: 10.18653/v1/2023.acl-long.150

[305] Wang P, Wang Z, Li Z, Gao Y, Yin B, Ren X. SCOTT: self-consistent chain-of-thought distillation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 5546–5558

[306] Chen Q, Qin L, Liu J, Peng D, Wang J, Hu M, Chen Z, Che W, Liu T. ECM: a unified electronic circuit model for explaining the emergence of in-context learning and chain-of-thought in large language model. 2025, arXiv preprint arXiv: 2502.03325

[307] Lyu Q, Havaldar S, Stein A, Zhang L, Rao D, Wong E, Apidianaki M, Callison-Burch C. Faithful chain-of-thought reasoning. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2023, 305–329

[308] Zeng S, Chang X, Xie M, Liu X, Bai Y, Pan Z, Xu M, Wei X. FutureSightDrive: thinking visually with spatio-temporal cot for autonomous driving. 2025, arXiv preprint arXiv: 2505.17685

[309] Renze M, Guven E. Self-reflection in LLM agents: effects on problem-solving performance. 2024, arXiv preprint arXiv: 2405.06682

[310] Balachandran V, Chen J, Chen L, Garg S, Joshi N, Lara Y, Langford J, Nushi B, Vineet V, Wu Y, Yousefi S. Inference-time scaling for complex tasks: where we stand and what lies ahead. 2025, arXiv preprint arXiv: 2504.00294

[311] Wu Y, Sun Z, Li S, Welleck S, Yang Y. Inference scaling laws: an empirical analysis of compute-optimal inference for problem-solving with language models. 2024, arXiv preprint arXiv: 2408.00724

[312] Yu Q, Zhang Z, Zhu R, Yuan Y, Zuo X, et al. DAPO: an open-source llm reinforcement learning system at scale. 2025, arXiv preprint arXiv: 2503.14476

[313] Yue Y, Yuan Y, Yu Q, Zuo X, Zhu R, et al. VAPO: efficient and reliable reinforcement learning for advanced reasoning tasks. 2025, arXiv preprint arXiv: 2504.05118

[314] Chen J, Fan T, Liu X, Liu L, Lin Z, et al. Seed-thinking-v1.5: advancing superb reasoning models with reinforcement learning. 2025, arXiv preprint arXiv: 2504.13914

[315] Duan K, Liu Z, Mao X, Pang T, Chen C, Chen Q, Shieh MQ, Dou L. Efficient process reward model training via active learning. 2025, arXiv preprint arXiv: 2504.10559

[316] Sui Y, Chuang YN, Wang G, Zhang J, Zhang T, Yuan J, Liu H, Wen A, Zhong S, Zou N, Chen H, Hu X. Stop overthinking: a survey on efficient reasoning for large language models. Transactions on Machine Learning Research, 2025, 2025

[317] Feng S, Fang G, Ma X, Wang X. Efficient reasoning models: a survey. 2025, arXiv preprint arXiv: 2504.10903

[318] Hou B, Zhang Y, Ji J, Liu Y, Qian K, Andreas J, Chang S. ThinkPrune: pruning long chain-of-thought of llms via reinforcement learning. 2025, arXiv preprint arXiv: 2504.01296

[319] Chen Q, Qin L, Liu J, Liao Y, Wang J, Zhou J, Che W. RBF++: quantifying and optimizing reasoning boundaries across measurable and unmeasurable capabilities for chain-of-thought reasoning. 2025, arXiv

preprint arXiv: 2505.13307

[320] Qi J, Xu Z, Shen Y, Liu M, Jin D, Wang Q, Huang L. The art of SOCRATIC QUESTIONING: recursive thinking with large language models. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 4177–4199

[321] Paul D, Ismayilzada M, Peyrard M, Borges B, Bosselut A, West R, Faltings B. REFINER: reasoning feedback on intermediate representations. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 1100–1126

[322] Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, Alon U, Dziri N, Prabhumoye S, Yang Y, Gupta S, Majumder BP, Hermann K, Welleck S, Yazdanbakhsh A, Clark P. SELF-REFINE: iterative refinement with self-feedback. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 2019

[323] Li Y, Shen X, Yao X, Ding X, Miao Y, Krishnan R, Padman R. Beyond single-turn: a survey on multi-turn interactions with large language models. 2025, arXiv preprint arXiv: 2504.04717

[324] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan KR, Cao Y. ReAct: synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations. 2023

[325] Chen Z, Chen Q, Qin L, Guo Q, Lv H, Zou Y, Yan H, Chen K, Lin D. What are the essential factors in crafting effective long context multi-hop instruction datasets? Insights and best practices. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 27129–27151

Libo QIN received his PhD degree in computer science from the Harbin Institute of Technology, China. He is a professor at Central South University, China. His research interests include natural language processing and large language models.

Qiguang CHEN is a PhD student at Harbin Institute of Technology (HIT), China. His research fields include natural language processing and complex reasoning.

Xiachong FENG is a Postdoctoral Researcher at the University of Hong Kong, China holding a PhD from the Social Computing and Interactive Robotics Research Center at Harbin Institute of Technology, China. He was also a visiting student at National University of Singapore, Singapore. His research focuses on large language models (LLMs) and social agents, with publications in top-tier venues like ACL, TASLP, and TMLR. Awarded the National Scholarship three times, he has also received the CCL 2021 Best English Paper Award, TMLR Survey Award, and ICASSP 2023 MUG Challenge championship. He actively contributes to the academic community as a PC member/Area Chair for ICML, ICLR, and ACL Rolling Review. His work bridges AI, NLP, and human-agent interaction.

Yang WU is a PhD graduate in Computer Science from Harbin Institute of Technology, China with research expertise in improving the planning and cross-task generalization abilities of large language models for complex tasks. His doctoral work received the Best Paper Award at IEEE Multimedia 2021.

Yongheng ZHANG is a master student at Central South University, China. His primary research interests include large language models and multimodal reasoning.

Yinghui LI received the BEng degree from the Department of Computer Science and Technology, Tsinghua University, China in 2020. He is currently working toward the PhD degree with the Tsinghua Shenzhen International Graduate School, Tsinghua University, China. His research interests include natural language processing and deep learning.

Min LI received her PhD in Computer Science from Central South University, China in 2008. She is currently a professor and the dean of the School of Computer Science and Engineering at Central South University, China. Her research focuses on computational biology, systems biology, and bioinformatics. She has authored over 100 technical papers in leading journals and conference proceedings, including Nature Communications, Genome Research, Genome Biology, Nucleic Acids Research, and Bioinformatics.

Wanxiang CHE received his PhD degree in computer science from the Harbin Institute of Technology (HIT), China, in 2008. He is a full professor in the School of Computer Science and Technology, HIT. His current research interests include natural language processing and large language models.

Philip S. YU (Life Fellow, IEEE) is currently a distinguished professor and the Wexler chair of information technology with the Department of Computer Science, University of Illinois Chicago (UIC), USA. Before joining UIC, he was with IBM Watson Research Center, where he built a world-renowned data mining and database department. He has authored or coauthored more than 780 papers in refereed journals and conferences. He holds or has applied for more than 250 USA Patents. His research interest include Big Data, including data mining, data stream, database, and privacy. He is a fellow of ACM. He was the editor-in-chief of the ACM Transactions on Knowledge Discovery from Data during 2011–2017 and IEEE Transactions on Knowledge and Data Engineering during 2001–2004. He was the recipient of several IBM honors including the two IBM Outstanding Innovation Awards, Outstanding Technical Achievement Award, two Research Division Awards, and 94th Plateau of Invention Achievement Awards.