

Tool learning with large language models: a survey

Changle QU¹, Sunhao DAI¹, Xiaochi WEI², Hengyi CAI³, Shuaiqiang WANG²,
Dawei YIN², Jun XU (✉)¹, Ji-rong WEN¹

¹ Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China

² Baidu Inc., Beijing 100193, China

³ Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100864, China

© Higher Education Press 2025

Abstract Recently, tool learning with large language models (LLMs) has emerged as a promising paradigm for augmenting the capabilities of LLMs to tackle highly complex problems. Despite growing attention and rapid advancements in this field, the existing literature remains fragmented and lacks systematic organization, posing barriers to entry for newcomers. This gap motivates us to conduct a comprehensive survey of existing works on tool learning with LLMs. In this survey, we focus on reviewing existing literature from the two primary aspects (1) why tool learning is beneficial and (2) how tool learning is implemented, enabling a comprehensive understanding of tool learning with LLMs. We first explore the “why” by reviewing both the benefits of tool integration and the inherent benefits of the tool learning paradigm from six specific aspects. In terms of “how”, we systematically review the literature according to a taxonomy of four key stages in the tool learning workflow: task planning, tool selection, tool calling, and response generation. Additionally, we provide a detailed summary of existing benchmarks and evaluation methods, categorizing them according to their relevance to different stages. Finally, we discuss current challenges and outline potential future directions, aiming to inspire both researchers and industrial developers to further explore this emerging and promising area.

Keywords tool learning, large language models, agent

1 Introduction

“Sharp tools make good work.”

—*The Analects: Wei Ling Gong*

Throughout history, humanity has continually sought innovation, utilizing increasingly sophisticated tools to boost efficiency and enhance capabilities [1,2]. These tools, extending both our intellect and physicality, have been crucial in driving social and cultural evolution [3]. From primitive stone tools to advanced machinery, this progression has expanded our potential beyond natural limits, enabling more complex and efficient task management [4].

Today, we are experiencing a new technological renaissance, driven by breakthroughs in artificial intelligence, especially through the development of large language models (LLMs). Pioneering models such as ChatGPT [5] have demonstrated remarkable capabilities, marking significant progress in a range of natural language processing (NLP) tasks, including summarization [6,7], machine translation [8,9], question answering [10,11], etc. However, despite their impressive capabilities, LLMs often struggle with complex computations and delivering accurate, timely information due to their reliance on fixed and parametric knowledge [12,13]. This inherent limitation frequently results in responses that are plausible yet factually incorrect or outdated (often referred to as hallucination) [14,15], posing significant risks and misleading users.

With the continuous enhancement of LLMs capabilities, it is expected that LLMs will become proficient in using tools to solve complex problems as human [16], a concept known as tool learning with LLMs. Tool learning emerges as a promising solution to mitigate these limitations of LLMs by enabling dynamic interaction with external tools [17–20]. This approach not only enhances problem-solving capabilities of LLMs but also broadens their functional scope [21–23]. For instance, LLMs can perform complex calculations using a calculator tool, access real-time weather updates through weather APIs, and execute programming code via interpreters [24,25]. This integration significantly improves their response accuracy to user queries, facilitating more effective and reliable user interactions. As this field continues to evolve, tool-augmented LLMs are expected to play a pivotal role in the future of NLP [26,27], offering more versatile and adaptable solutions [28,29].

As shown in Fig. 1, the past year has witnessed a rapid surge in research efforts on tool learning concurrent with the rise of LLMs. Notably, in practical applications, GPT-4 [5] addresses its knowledge limitations and augments its capabilities by calling on plugins, ultimately integrating the returned results of plugins with its internal knowledge to generate better responses for users. Within the research community, much effort has been made in exploring how to evaluate the tool learning capabilities of LLMs [30–32] and how to enhance it

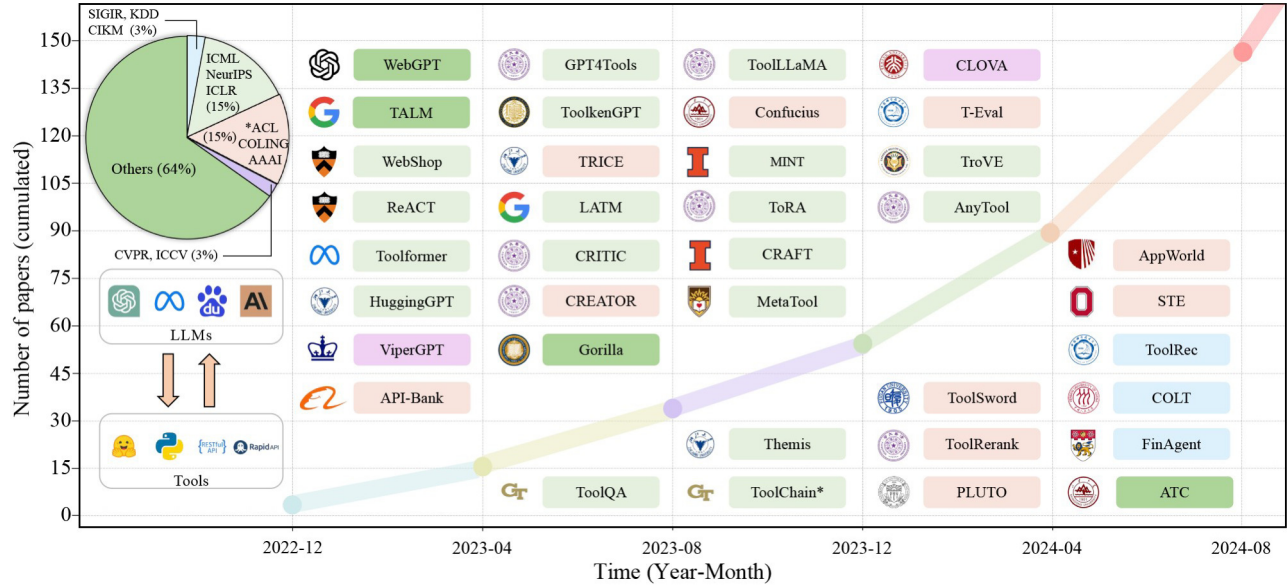


Fig. 1 An illustration of the development trajectory of tool learning. We present the statistics of papers with the publication year and venue, with each venue uniquely represented by a distinct color. For each time period, we have selected a range of representative landmark studies that have significantly contributed to the field. (Note that we use the institution of the first author as the representing institution in the figure.)

to strengthen the capabilities of LLMs [33–35]. Given the increasing attention and rapid development of tool learning with LLMs, it is essential to systematically review the most recent advancements and challenges, so as to benefit researchers and industrial developers in understanding the current progress and inspire more future work in this area.

In this survey, we conduct a systematic exploration of existing studies in two primary dimensions: (1) **why tool learning** is beneficial and (2) **how tool learning** is implemented. Specifically, the “why tool learning” dimension examines both the advantages of tool integration and the inherent benefits of the tool learning paradigm, while the “how tool learning” dimension details the four stages of the entire tool learning workflow: task planning, tool selection, tool calling, and response generation. These dimensions are foundational to understanding tool learning with LLMs. Moreover, we provide a systematic summary of existing benchmarks and evaluation methods, classifying them based on their focus across different stages. Finally, we discuss the current challenges and propose future directions, offering critical insights to facilitate the development of this promising and burgeoning research area. We also maintain a GitHub repository to continually keep track of the relevant papers and resources in this rising area at github.com/quchangle1/LLM-Tool-Survey.

It is worth noting that while other surveys provide comprehensive overviews of techniques and methods used by LLMs [36], applications in planning [37], reasoning [38,39], agents [40–42], and retrieval-augmented generation [43,44], they often mention tools or tool learning but do not extensively explore this aspect. Compared with them, our survey provides a focused and detailed analysis of tool learning with LLMs, especially elucidating the dual aspects of **why** tool learning is essential for LLMs and **how** tool learning can be systematically implemented. Through these two principle aspects, we offer an up-to-date and comprehensive

review of tool learning with LLMs. Meanwhile, we also acknowledge the foundational contributions of earlier perspective papers like those by Mialon et al. (2023) [45] and Qin et al. (2023) [16], which initially highlighted the promising opportunities that tools present to enhance LLMs capabilities. Since the field has seen rapid growth with many new studies emerging, our survey provides a broader introduction to these latest developments. Additionally, a more recent survey [46] discusses various tooling scenarios and approaches employed in language models, serving as an excellent supplement to our comprehensive review.

The remaining part of this paper (as illustrated in Fig. 2) is organized as follows: We begin by introducing the foundational concepts and terminology related to tool learning (Section 2). Following this, we explore the significance of tool learning for LLMs from six specific aspects (Section 3). We then systematically review the recent advancements in tool learning, focusing on four distinct stages of the tool learning workflow (Section 4). Subsequently, we provide a summary of the resources available for tool learning, including benchmarks and evaluation methods (Section 5). Next, we discuss the current challenges in the field and outline open directions for future research (Section 6). Lastly, we conclude the survey by summarizing the key findings (Section 7).

2 Background

In this section, we provide an overview of the key concepts and terminology related to tool learning, offering a clearer understanding of the fundamental aspects of this field.

What is a tool? The definition of a tool is notably broad within the context of augmented LLMs. Mialon et al. (2023) [45] articulate a tool as “*the external employment of an unattached or manipulable attached environmental object to alter more efficiently the form, position, or condition of another object.*” On the other hand, Wang et al. (2024) [46]

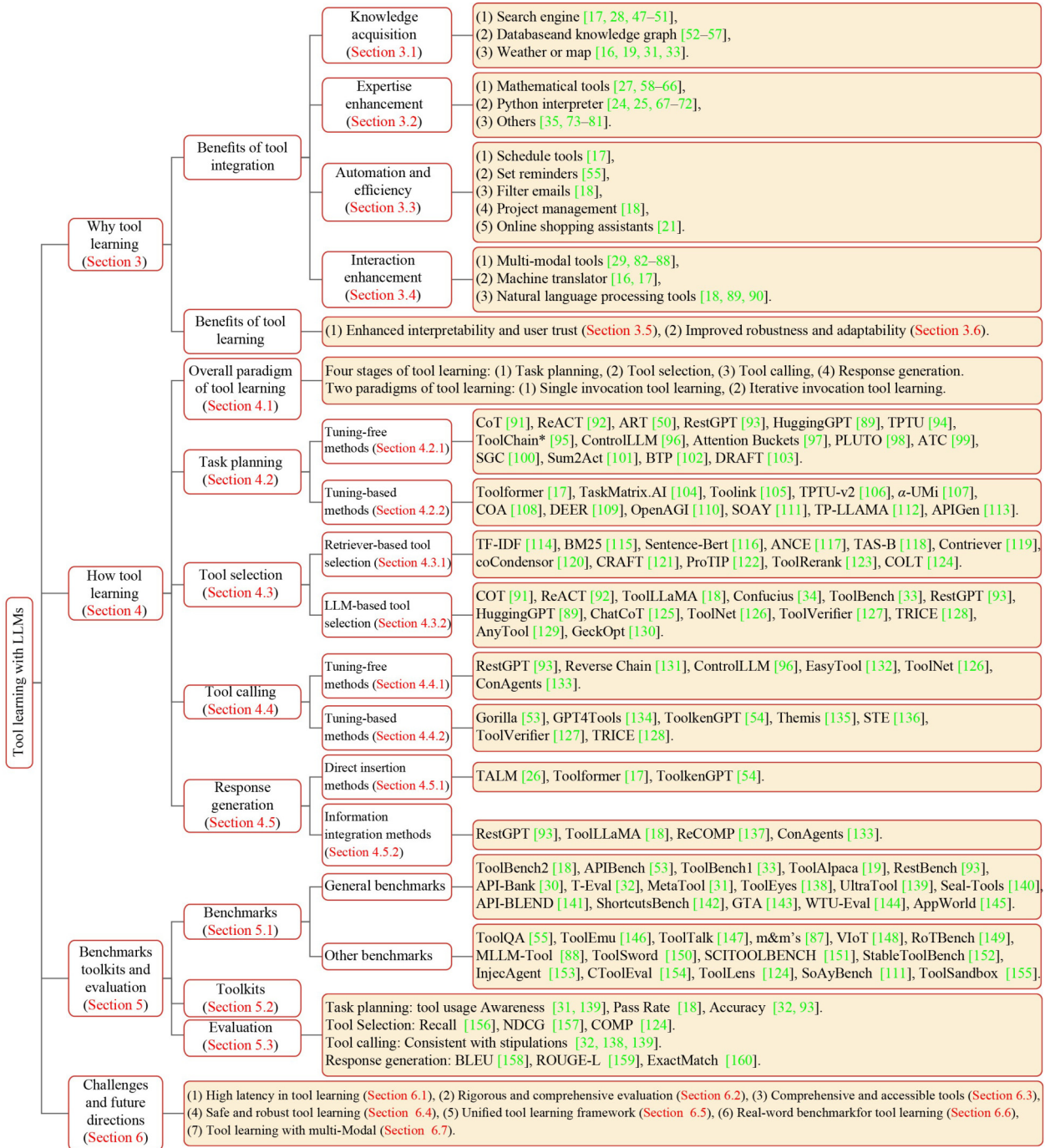


Fig. 2 The overall structure of this paper

define a tool as “An LM-used tool is a function interface to a computer program that runs externally to the LM, where the LM generates the function calls and input arguments in order to use the tool.” Similarly, it is our contention that any method enhancing LLMs through external means qualifies as a tool. Notably, retrieval-augmented generation (RAG) represents a specific instance of tool learning, wherein the search engine is employed as a tool for LLMs. Meanwhile, the definition of “tool” often remains vague and inconsistent across different papers. For example, some studies distinctly define tools and APIs, positing that a tool comprises an aggregation of multiple APIs [18,33,53]. Conversely, other studies treat each API as

an independent tool [19,30,122]. In this survey, adhering to the definitions of tools established earlier in the text, we consider each API as an individual tool.

What is tool learning? Tool learning refers to the process that “aims to unleash the power of LLMs to effectively interact with various tools to accomplish complex tasks” [18]. This paradigm significantly improves the ability of LLMs to solve complex problems. For example, when ChatGPT receives a user query, it evaluates the necessity of calling a specific tool. If a tool is required, ChatGPT will transparently outline the problem-solving process using the tool, explaining the rationale behind its responses, thereby ensuring the user

receives a well-informed answer. Moreover, in instances where the initial solution fails, ChatGPT will reassess its tool selection and employ an alternative to generate a new response.

3 Why tool learning?

In this section, we will delineate the multifaceted importance of tool learning for LLMs from two principal perspectives: the benefits of tool integration and the benefits of the tool learning paradigm itself. On the one hand, tool integration into LLMs enhances capabilities across several domains, namely knowledge acquisition, expertise enhancement, automation and efficiency, and interaction enhancement. On the other hand, the adoption of the tool learning paradigm bolsters the robustness of responses and transparency of generation processes, thereby enhancing interpretability and user trust, as well as improving system robustness and adaptability. Subsequent subsections will elaborate on these six aspects in detail, outlining why tool learning is important for LLMs.

3.1 Knowledge acquisition

Although LLMs have showcased their immense capabilities across various fields [161], their abilities are still bounded by the extent of knowledge learned during pre-training [12]. This embedded knowledge is finite and lacks the ability to acquire updated information. Additionally, the effectiveness of LLMs is further compromised by prompts from users, which may not always be meticulously crafted. Consequently, LLMs are prone to generating contents that seem superficially plausible but may contain factual inaccuracies, which is known as hallucination. A promising approach to mitigate these limitations involves augmenting LLMs with the capability to access external tools, which allows LLMs to acquire and integrate external knowledge dynamically. For example, the employment of search engine tool can enable LLMs to access contemporary information [17,28,47–51], while the integration of database tool allows LLMs to access structured databases to retrieve specific information or execute complex queries, thus expanding their knowledge base [52–57]. Additionally, connections to weather tools allow for real-time updates on weather conditions, forecasts, and historical data [19,31,33], and interfacing with mapping tools enables LLMs to get and provide geographical data, aiding in navigation and location-based queries [16]. Through these enhancements, LLMs can surpass traditional limitations, offering more accurate and contextually relevant outputs.

3.2 Expertise enhancement

Given the fact that LLMs are trained on datasets comprising general knowledge, they often exhibit deficiencies in specialized domains. While LLMs demonstrate robust problem-solving capabilities for basic mathematical problems, excelling in operations such as addition, subtraction, and exhibiting reasonable proficiency in multiplication tasks, their abilities significantly decline when confronted with division, exponentiation, logarithms, trigonometric functions, and other more complex composite functions [162,163]. This limitation extends to tasks involving code generation [164,165] and chemistry and physics problems [73,74], etc., further

underscoring the gap in their expertise in more specialized areas. Consequently, it is feasible to employ specific tools to augment the domain-specific expertise of LLMs [60,61,76,166]. For example, LLMs can use online calculators or mathematical tools to perform complex calculations, solve equations, or analyze statistical data [27,58–66]. Additionally, the integration of external programming resources such as Python compilers and interpreters allows LLMs to receive code execution feedback, which is essential for refining code to align with user requirements and to optimize the code generation [24,25,67–72]. Moreover, LLMs can also leverage tools in fields such as chemistry [73–75], biology [76], economics [77–79], medicine [80,81], and recommendation systems [35] to enhance their domain-specific expertise. This approach not only mitigates the expertise gap in LLMs but also enhances their utility in specialized applications by providing domain-specific knowledge.

3.3 Automation and efficiency

LLMs are fundamentally language processors that lack the capability to execute external actions independently, such as reserving conference rooms or booking flight tickets [46]. The integration of LLMs with external tools facilitates the execution of such tasks by simply populating tool interfaces with the necessary parameters. For example, LLMs can employ task automation tools to automate repetitive tasks such as scheduling [17], setting reminders [55], and filtering emails [18], thereby enhancing their practicality for user assistance. Moreover, by interfacing with project management and workflow tools, LLMs can aid users in managing tasks, monitoring progress, and optimizing work processes [18]. In addition, the integration with online shopping assistants not only simplifies the shopping process [21] but also enhances processing efficiency and user experience. Furthermore, employing data table processing tools enables LLMs to perform data analysis and visualization directly [16], thereby simplifying the data manipulation process of users.

3.4 Interaction enhancement

Due to the diverse and multifaceted nature of user queries in the real-world, which may encompass multiple languages and modalities, LLMs often face challenges in consistently understanding different types of input. This variability can lead to ambiguities in discerning the actual user intent [88]. The deployment of specialized tools can significantly enhance the perceptual capabilities of LLMs. For example, LLMs can utilize multi-modal tools, such as speech recognition and image analysis, to better understand and respond to a broader spectrum of user inputs [29,82–88]. Moreover, by interfacing with machine translator tools, LLMs have the capability to convert languages in which they are less proficient into languages they comprehend more effectively [16,17]. Additionally, the integration of advanced natural language processing tools can augment the linguistic understanding of LLMs, thereby optimizing dialogue management and intent recognition [18,89,90]. Such advancements may include platforms that utilize contextual understanding models to elevate the performance of chatbot systems. Ultimately, improving perceptual input and sensory perception is crucial

for the progression of LLMs capabilities in managing intricate user interactions.

3.5 Enhanced interpretability and user trust

A significant concern with current LLMs is their opaque, “black-box” nature, which does not reveal the decision-making process to users [167,168], thereby severely lacking in interpretability. This opacity often leads to skepticism about the reliability of the response provided by LLMs and makes it challenging to ascertain their correctness [169]. Moreover, interpretability is particularly crucial in high-stakes domains such as aviation, healthcare and finance [16,77], where accuracy is imperative. Therefore, understanding and explaining LLMs is crucial for elucidating their behaviors [168]. Some studies have enhanced the accuracy and interpretability of LLMs by enabling them to generate text with citations [170,171]. In contrast, through the utilization of tool learning, LLMs can exhibit each step of their decision-making process, thereby making their operations more transparent [16]. Even in cases of erroneous outputs, such transparency allows users to quickly identify and understand the source of errors, which facilitates a better understanding and trust in the decisions of LLMs, thus enhancing effective human-machine collaboration.

3.6 Improved robustness and adaptability

Existing research indicates that LLMs are highly sensitive to user inputs within prompts [172–174]. Merely minor modifications to these inputs can elicit substantial changes in the responses, highlighting a lack of robustness in LLMs. In the real world, different users have varying interests and ways of asking questions, leading to a diverse array of prompts. The integration of specialized tools has been proposed as a strategy to reduce reliance on the statistical patterns in the training data [16–18,54,89]. Though the input format from the user is different, the input and output of the tool are the same. This enhancement increases the resistance of LLMs to input perturbations and their adaptability to new environments. Thus, such integration not only stabilizes the models in uncertain conditions but also reduces the risks associated with input errors.

4 How tool learning?

In this section, we will first introduce the overall paradigm of tool learning, which includes four distinct stages and two typical paradigms. Following this framework, we provide a detailed review of each stage within the tool learning workflow, along with the latest advancements associated with each stage. It’s important to note that many works involve multiple stages of tool learning, but we only discuss its core stages here. For each stage, we also present a practical, real-world example utilizing GPT-4 for tool learning to address a specific problem, which are designed to help newcomers better understand what each stage involves and how it is implemented.

4.1 Overall paradigm of tool learning

In this section, we will introduce the entire process of tool learning, including four stages and two paradigms involved in

the utilization of tool-augmented LLMs.

Four stages of tool learning. As illustrated in the left part of Fig. 3, the typical process of tool learning comprises four stages: task planning, tool selection, tool calling, and response generation, which is adopted in numerous works related to tools [89,93,94]. This process outlines the user interaction pipeline with tool-augmented LLMs: given a user question, the preliminary stage involves the LLMs analyzing the requests of users to understand their intent and decompose it into potential solvable sub-questions. Subsequently, the appropriate tools are selected to tackle these sub-questions. This tool selection process is categorized into two types based on whether a retriever is used: retriever-based tool selection and LLM-based tool selection. Recently, there has been an increasing focus on initially using a retriever to filter out the top- k suitable tools [18,34,122]. This necessity stems from the fact real-world systems usually have a vast number of tools, rendering it impractical to incorporate the descriptions of all tools as input for LLMs due to the constraints related to length and latency [124]. Subsequently, the user query along with the selected tools are furnished to the LLMs, enabling it to select the optimal tool and configure the necessary parameters for tool calling. This necessitates that the LLMs possess a keen awareness of using tools and be able to correctly select the tools needed. Moreover, it is imperative for the LLMs to extract the correct tool parameters from the user query, a process that demands not only the accuracy of the parameter content but also adherence to the specific format requirements. Following the invocation of the tool, the LLMs utilizes the results returned by the tool to craft a superior response for the user.

Two paradigms of tool learning. As illustrated in the right part of Fig. 3, the paradigms for employing tool learning can be categorized into two types: tool learning with one-step task solving and tool learning with iterative task solving. These are also referred to as planning without feedback and planning with feedback in Wang et al. (2024) [40], and decomposition-first and interleaved decomposition in Huang et al. (2024) [37]. In earlier studies on tool learning [17,69,89], the primary paradigm is tool learning with one-step task solving: upon receiving a user question, LLMs would analyze the requests of user to understand the user intent and immediately plan all the sub-tasks needed to solve the problem. The LLMs would then directly generate a response based on the results returned by the selected tools without considering the possibility of errors during the process or altering the plan based on tool feedback. Subsequent researches introduce a new paradigm known as tool learning with iterative task solving [18,30,34,93]. This approach does not commit to a complete task plan upfront. Instead, it allows for iterative interactions with the tools, adjusting the sub-tasks progressively based on tool feedback. This enables LLMs to address the problem step-by-step, refining its plan continuously in response to the results returned by tools, thereby improving problem-solving capabilities. This shift from a one-step to an iterative approach marks a significant advancement in the field of tool learning.

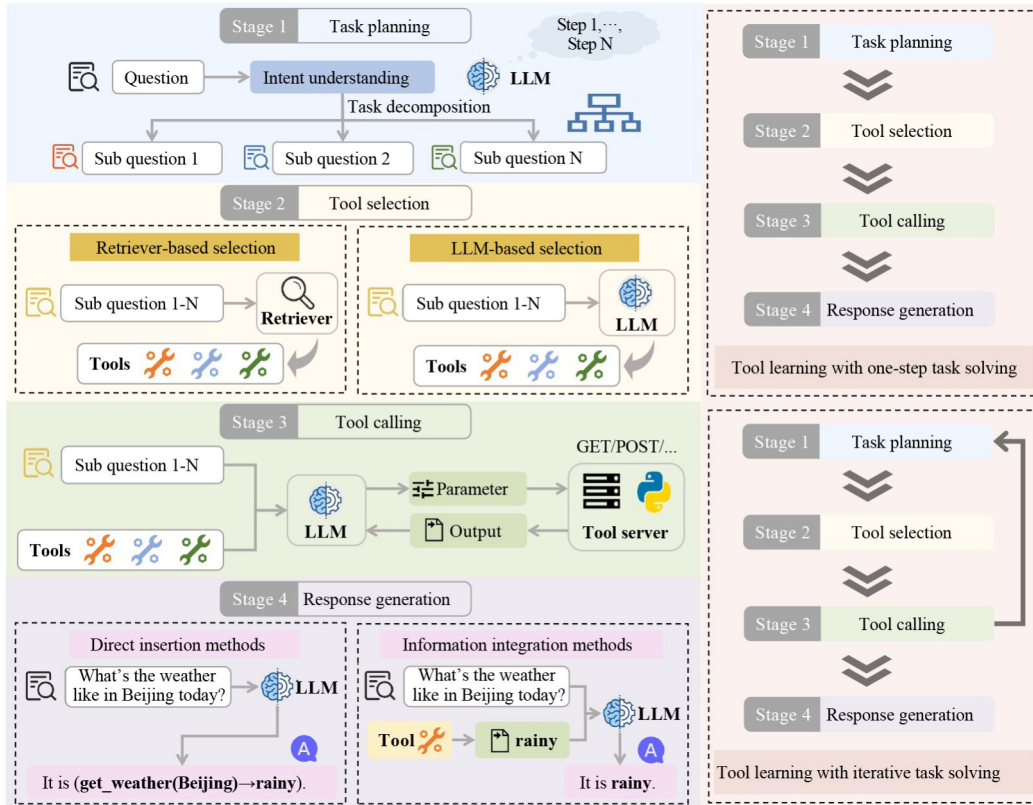


Fig. 3 The overall workflow for tool learning with large language models. The left part illustrates the four stages of tool learning: task planning, tool selection, tool calling, and response generation. The right part shows two paradigms of tool learning: tool learning with one-step task solving and tool learning with iterative task solving

4.2 Task planning

Generally, in real-world scenarios, user queries often embody complex intent. Thus, it is essential to first engage in task planning to conduct a comprehensive analysis of the user intent. This stage involves the decomposition of a user question into multiple sub-questions as required to achieve the specified objective. Additionally, given the dependency relationships inherent among some tasks, the planner is also tasked with delineating the dependencies and execution sequence of the decomposed tasks, thereby facilitating the establishment of interconnections between the sub-questions. For example, a real example for task planning with GPT-4 is shown as follows:

Tool learning with iterative task solving paradigm is not convenient for staged demonstrations, so this example uses the tool learning with one-step task solving paradigm. From this example, we can see that GPT-4 breaks down the user's question into three sub-questions. Since internal knowledge of GPT-4 does not contain the latest information, it is necessary to obtain the current prices of gold and Amazon (AMZN) stock, as well as the latest exchange rate between USD and CNY. By breaking down the query in this manner, GPT-4 can address complex questions step by step, demonstrating its logical analysis capabilities and ability to handle multi-step tasks. Next, we will introduce the latest developments in two categories: tuning-free methods and tuning-based methods.

4.2.1 Tuning-free methods

Existing studies [50,175,176] demonstrate that the innate

An Example for Task Planning with GPT-4

Instruction Prompt: You are currently in the task planning stage. You are given a user query requiring multi-step actions and reasoning. You will break down the user's query into sub-questions, and you only need to output these sub-questions after the breakdown. Ensure that the original problem is comprehensively covered through the minimal number of sub-questions.

User Question: I would like to know the value of 5 ounces of gold plus 1 million AMZN stocks in CNY.

Output: 1. What is the current price of gold per ounce in USD? 2. What is the current stock price of Amazon (AMZN) per share in USD? 3. What is the current exchange rate between USD and CNY (Chinese Yuan)?

abilities of LLMs enable effective planning through methods such as few-shot or even zero-shot prompting. For example, some studies [177–180] leverage prompts to decompose complex tasks into simpler sub-tasks, facilitating a structured

plan of action. ART [50] constructs a task library, from which it retrieves examples as few-shot prompts when encountering real-world tasks. RestGPT [93] introduces a Coarse-to-Fine Online Planning approach, an iterative task planning methodology that enables LLMs to progressively refine the process of task decomposition. HuggingGPT [89] leverages a sophisticated prompt design framework, which integrates specification-based instructions with demonstration based parsing methods. ToolChain* [95] employs a planning mechanism by constructing the entire action space as a decision tree, where each node within the tree represents a potential API function call. TPTU [94] introduces a structured framework specifically designed for LLM-based AI agents, incorporating two distinct types of agents: the One-step agent and the sequential agent. Attention Buckets [97] operates in parallel with unique RoPE angles, forming distinct waveforms that compensate for each other, reducing the risk of LLMs missing critical information. ControlLLM [96] introduces a paradigm known as Thoughts-on-Graph (ToG), which leverages Depth-First Search (DFS) on a pre-constructed tool graph to identify solutions. PLUTO [98] uses an autoregressive planning approach to iteratively improve performance by generating hypotheses, performing cluster analysis, and refining sub-queries until the initial query is satisfied. ATC [99] enables LLMs to independently learn and master new tools by using a chain of tools and a black-box probing method to identify and record tool usage. Tool-Planner [181] organizes tools into toolkits based on API functions with similar functionality, enabling LLMs to plan across different toolkits. SGC [100] enhances task planning by integrating GNNs with LLMs, enabling more efficient and accurate sub-task selection within task graphs. Sum2Act [101] guides LLMs to solve complex tasks by summarizing progress at each step and adjusting actions based on task state and errors. BTP [102] creates an optimal plan for tool usage under budget constraints, allowing LLMs to efficiently manage costs while solving user queries. DRAFT [103] proposes a novel framework aimed at dynamically adjusting and optimizing tool documentation based on the interaction feedback between LLMs and external tools, thereby improving LLMs' comprehension and tool-using capabilities.

4.2.2 Tuning-based methods

Though LLMs demonstrate impressive performance in zero-shot or few-shot settings, they remain less effective compared to models that have been fine-tuned [182]. Toolformer [17] employs API calls that actually assist the model in predicting future tokens to fine-tune GPT-J, which enhances the awareness and capability of LLMs to utilize tools effectively. TaskMatrix.AI [104] leverages Reinforcement Learning from Human Feedback (RLHF) to utilize the knowledge and insights gained through human feedback, thereby enhancing the foundation model. Toolink [105] innovates by decomposing the target task into a toolkit for problem-solving, then employing a model to utilize these tools to answer queries via a chain-of-solving (CoS) approach. TPTU-v2 [106] develops an LLM finetuner to fine-tune a base LLM using a meticulously curated dataset, so that the finetuned LLM can be more capable of task planning and API calls,

especially for domain-specific tasks. α -UMi [107] presents a novel two-phase training paradigm where a foundational large language model is first extensively fine-tuned and then replicated as a planner for further fine-tuning on planning tasks. COA [108] trains LLMs to first decode reasoning chains with abstract placeholders, and then call domain tools to reify each reasoning chain by filling in specific knowledge. DEER [109] stimulates decision-making awareness in LLMs across various scenarios by automatically generating tool usage examples with multiple decision branches, and enhances the generalization ability of LLMs towards unseen tools through proposing novel tool sampling strategies. SOAY [111] first lets the LLM generate a feasible API calling plan, i.e. solution, based on complex user inputs, and then allows the LLM to generate executable API calling code based on the generated solution. TP-LLaMA [112] introduces a method to create preference data from thought trees by using previously ignored failed explorations, creating a dataset for DPO to update the LLM's policy. APiGen [113] creates an automated pipeline for generating diverse, high-quality, and verifiable function-calling datasets, significantly enhancing the performance of models fine-tuned with this data.

Remark. In summary, task planning, as the initial stage of tool learning, is crucial for solving the entire problem. Although there are many methods currently available to enhance the task planning capabilities of LLMs, generating a perfect plan directly when facing complex issues remains challenging. Furthermore, tool learning is a process involving interaction between LLMs and tools. How to better utilize feedback from tools to improve planning is still a question worthy of investigation.

4.3 Tool selection

After the task planning phase, LLMs have already decomposed the user question into multiple sub-questions. In order to better address these sub-questions, it is necessary to select appropriate tools. The tool selection process involves choosing through a retriever or directly allowing LLMs to pick from a provided list of tools. When there are too many tools, a tool retriever is typically used to identify the top- K relevant tools to offer to the LLMs, a process known as retriever-based tool selection. If the quantity of tools is limited or upon receiving the tools retrieved during the tool retrieval phase, the LLMs need to select the appropriate tools based on the tool descriptions and the sub-question, which is known as LLM-based tool selection. For example, an example for tool selection with GPT-4 is shown as follows:

From this example, we can see that for the sub-question about obtaining the price of gold, GPT-4 can correctly select the necessary tools. Specifically, when faced with multiple candidate tools, GPT-4 can analyze the features of each tool and choose the one most suitable for answering the question. In this example, GPT-4 selects the Metals Prices Rates API because it provides real-time updated information on gold prices. This demonstrates accuracy and effectiveness of GPT-4 in tool selection. Next, we will introduce the latest developments in two categories: retriever-based tool selection and LLM-based tool selection.

An Example for Tool Selection with GPT-4

Instruction Prompt: You are currently in the tool selection stage. You are given candidate tools that can be potentially used to solve the sub-question. Among candidate tools, select a list of relevant tools that would help solve the sub-question.

Sub-question 1: *What is the current price of gold per ounce in USD?*

Candidate Tools: *1.Metals Prices Rates API: The latest API endpoint will return real-time exchange rate data updated every 60 seconds. 2.Medium: Get official news from Medium. 3.Cryptocurrency Markets: Recently published cryptocurrencies videos.*

Output: *1.Metals Prices Rates API: The latest API endpoint will return real-time exchange rate data updated every 60 seconds.*

Sub-question 2: ...

⋮

Output: ...

4.3.1 Retriever-based tool selection

Real-world systems often incorporate a wide array of tools, making it impractical to input descriptions of all tools into LLMs due to length limitations and latency constraints. Therefore, to fully exploit the potential of tool-augmented LLMs, it is crucial to develop an efficient tool retrieval system. This system aims to bridge the gap between the broad capabilities of LLMs and the practical limitations of input size by efficiently selecting the top- K most suitable tools for a given query from a vast tool set. State-of-the-art retrieval methods can be categorized into two types: term-based and semantic-based.

Term-based methods. Term-based methods (i.e., sparse retrieval) represent both documents and queries as high-dimensional sparse vectors based on terms, as exemplified by TF-IDF [114] and BM25 [115]. These methods employ exact term matching to achieve efficient alignment between queries and documents. For example, Gorilla [53] employs BM25 and GPT-Index to construct a retriever for implementing tool retrieval.

Semantic-based methods. Conversely, semantic-based methods i.e., dense retrieval) utilize neural networks to learn the semantic relationship between queries and tool descriptions [116–120], and then calculate the semantic similarity using methods such as cosine similarity. Recently, there has been a burgeoning interest in the development and refinement of more efficient tool retrievers. Some studies

[18,34,106] train a Sentence-Bert model as the tool retriever, enabling the high-efficiency retrieval of relevant tools. CRAFT [121] instructs LLMs to generate a fictitious tool description based on the given query and then employs this fabricated tool to conduct a search. Anantha et al. (2023) [122] propose ProTIP based on the concept of task decomposition. Xu et al. (2024a) [183] propose a method that enhances tool retrieval by leveraging feedback from LLMs to progressively refine instructions and align retrieval with tool usage. COLT [124] proposes a novel tool retrieval approach using GNNs, identifying that a critical dimension often overlooked in conventional tool retrieval methodologies is the necessity to ensure the completeness of the tools retrieved. In addition to the recall phase, Zheng et al. (2024) [123] also take into account the re-ranking stage of tool retrieval. They consider the differences between seen and unseen tools, as well as the hierarchical structure of the tool library. Building on these considerations, they propose an adaptive and hierarchy-aware Re-ranking method, ToolRerank. Meanwhile, we can also directly employ off-the-shelf embeddings [184,185] to get the representations of user queries and tool descriptions. In conclusion, constructing an efficient tool retriever is of paramount importance.

Remark. Although traditional information retrieval methods are suitable for tool retrieval scenarios, they still have issues such as focusing solely on semantic similarity and ignoring the hierarchical structure of the tools. Future work should consider the unique needs and characteristics specific to tool retrieval scenarios in order to build a more effective tool retriever.

4.3.2 LLM-based tool selection

In instances where the quantity of tool libraries is limited or upon receiving the tools retrieved from the tool retrieval phase, it is feasible to incorporate the descriptions and parameter lists of these tools into the input context along with the user query provided to LLMs. Subsequently, LLMs are tasked with selecting the appropriate tools from the available tool list based on the user query. Given that the resolution of queries is occasionally sensitive to the order in which tools are invoked, there is a necessity for serial tool calling, where the output of one tool may serve as the input parameter for another. Consequently, this demands a high degree of reasoning capability from the LLMs. It must adeptly select the correct tools based on the information currently at its disposal and the information that needs to be acquired. Existing methods can be similarly categorized into tuning-free and tuning-based approaches.

Tuning-free methods. Tuning-free methods capitalize on the in context learning ability of LLMs through strategic prompting [89,93]. For instance, Wei et al. (2022) [91] introduce the concept of chain of thought (COT), effectively incorporating the directive “let’s think step by step” into the prompt structure. Further advancing this discourse, Yao et al. (2022) [92] propose ReACT, a framework that integrates reasoning with action, thus enabling LLMs to not only justify actions but also to refine their reasoning processes based on

feedback from the environment (e.g., the output of tools). This development marks a significant step forward in enhancing the adaptability and decision-making capabilities of LLMs by fostering a more dynamic interaction between reasoning and action. Building upon these insights, Qin et al. (2024) [18] propose DFSDT method, which addresses the issue of error propagation by incorporating a depth-first search strategy to improve decision-making accuracy. ChatCoT [125] integrates tool use into multi-turn reasoning, allowing LLMs to seamlessly combine conversation-based reasoning with tool manipulation. ToolNet [126] organizes tools into a directed graph, enabling LLMs to navigate from an initial tool node, iteratively selecting tools until the task is completed. AnyTool [129] designs a more efficient tool retriever by leveraging the hierarchical structure of tools. GeckOpt [130] narrows down tool selection by adding intent-driven gating.

Tuning-based methods. Tuning-based methods directly fine-tune the parameters of LLMs on the tool learning dataset to master tool usage. Toolbench [33] analyzes the challenges faced by open-source LLMs during the tool learning process, suggesting that fine-tuning, along with utilizing demonstration retrieval and system prompts, can significantly enhance the effectiveness of LLMs in tool learning. TRICE [128] proposes a two-stage framework, which initially employs behavior cloning for instruct-tuning of the LLMs to imitate the behavior of tool usage, followed by further reinforcing the model through RLEF by utilizing the tool execution feedback. ToolLLaMA [18] employs the instruction solution pairs derived from DFSDT method to fine-tune the LLaMA 7B model. Confucius [34] acknowledges the diversity in tool complexity and proposes a novel tool learning framework. ToolVerifier [127] introduces a self-verification method which distinguishes between close candidates by self-asking contrastive questions during tool selection.

Remark. By comparing the aforementioned methods, we can find that the tuning-based method improves the capability of LLMs in tool selection by modifying model parameters. This approach can integrate extensive knowledge about tools, but it is only applicable to open-source LLMs and incurs substantial computational resource consumption. Conversely, the tuning-free method enhances the capability of LLMs in tool selection using precise prompting strategies or by modifying existing mechanisms, and it is compatible with all LLMs. However, since the possibilities for designing prompts are limitless, finding the ideal way to create the perfect prompt is still a major challenge.

In real-world applications, the sheer volume of available tools, coupled with constraints such as limited context length and high latency, makes it impractical to present all options to LLMs simultaneously. To address this challenge, traditional retrieval methods are typically employed first to filter and narrow down the pool of potential tools. This initial step is crucial for managing the complexity and ensuring that the subsequent LLM-based selection is both efficient and effective. After this retrieval process, the LLM can then refine the selection, making the final choice that aligns with its reasoning and preferences. This two-step approach highlights

the complementary roles of both traditional retrieval methods and LLM-based techniques, demonstrating their collective importance in optimizing the tool selection process for real-world scenarios.

4.4 Tool calling

In the tool calling stage, LLMs need to extract the required parameters from the user query in accordance with the specifications outlined in the tool description and request data from tool servers. This process mandates that the LLMs not only correctly extract the parameters' content and format but also adhere strictly to the prescribed output format to prevent the generation of superfluous sentences. For example, an example for tool calling with GPT-4 is shown as follows:

From this example, we can see that GPT-4 can extract the necessary parameters for calling a tool based on the provided user question and the selected tool's documentation. Specifically, GPT-4 can parse the critical information in the tool description and accurately identify which parameters need to be provided. Next, we will introduce the latest developments in the same way as the previous two stages, dividing them into tuning-free methods and tuning-based methods.

An Example for Tool Calling with GPT-4

Instruction Prompt: You are currently in the tool calling stage. You are given selected tools that can be potentially used to solve the sub-question. Your goal is to extract the required parameters needed to call the tool from the sub-question based on the tool descriptions. Output in the following format: {parameter name: parameter, ..., parameter name: parameter}

Sub-question 1: *What is the current price of gold per ounce in USD?*

Selected Tools: *Tool Name:* {Metals Prices Rates API}. *Tool description:* {The latest API endpoint will return real-time exchange rate data updated every 60 seconds.} *Required params:* { [name: symbols, type: STRING, description: Enter a list of comma-separated currency codes or metal codes to limit output codes., name: base, type: STRING, description: Enter the three-letter currency code or metal code of your preferred base currency.] }

Output: {symbols: "XAU", base: "USD"}

Sub-question 2: ...

⋮

Output: ...

4.4.1 Tuning-free methods

Tuning-free methods predominantly leverage the few-shot approach to provide demonstrations for parameter extraction or rule-based methods, thereby enhancing the capability of LLMs to identify parameters [93,96,126,186]. Reverse Chain [131] utilizes reverse thinking by first selecting a final tool for a task and then having the LLMs populate the necessary parameters; if any are missing, an additional tool is chosen based on the description to complete them and accomplish the task. EasyTool [132] enhances the comprehension of LLMs regarding tool functions and parameter requirements by prompting ChatGPT to rewrite tool descriptions, making them more concise and incorporating guidelines for tool functionality directly within the descriptions. Xu et al. (2024b) [187] propose a method that compresses tool documentation into summary sequences while preserving key information, enabling efficient tool usage in LLMs with minimal performance loss. ConAgents [133] introduces a multi-agent collaborative framework, featuring a specialized execution agent tasked with parameter extraction and tool calling.

4.4.2 Tuning-based methods

Some studies enhance the tool calling capabilities of LLMs using tuning-based methods [53,127,128]. For example, GPT4Tools [134] enhances the open-source LLMs by integrating tool usage capabilities through fine-tuning with LoRA optimization techniques, using a dataset of tool usage instructions generated by ChatGPT. Toolkengpt [54] uses special tokens called “toolkens” to seamlessly call tools, switching to a special mode upon predicting a toolken to generate required input parameters and integrate the output back into the generation process. Themis [135] enhances the interpretability and scoring reliability of RMs by integrating tool usage and reasoning processes in an auto-regressive manner, dynamically determining which tools to call, how to pass parameters, and effectively incorporating the results into the reasoning process. STE [136] coordinates three key mechanisms in biological systems for the successful use of tools: trial and error, imagination, and memory, aiding LLMs in the accurate use of its trained tools.

Moreover, given the frequent occurrence of calling errors during the utilization of tools, such as incorrect formatting of input parameters, input parameters exceeding acceptable ranges of the tool, and tool server error, it is imperative to integrate error handling mechanisms. These mechanisms are designed to refine the action based on the error messages returned upon calling failure. This enables a more resilient and adaptive system, ensuring continuity and efficiency in tool learning even in the face of operational disruptions.

Remark. To sum up, although tuning-based methods can yield better results, they heavily rely on the tools seen in the training set, and perform poorly with new tools not included in the dataset. Additionally, they face challenges such as catastrophic forgetting and a lack of robust generalization. In contrast, tuning-free methods offer significant flexibility and do not require the construction of a specific dataset. So both tuning-based methods and tuning-free methods are crucial for enhancing the tool calling capabilities of LLMs.

4.5 Response generation

Given the diversity and complexity of tool outputs, which can include various formats like text, numbers, code, videos, and images, it is often impractical to present these results directly to users. Upon receiving outputs from tools, LLMs must synthesize information relevant to user queries and integrate their own knowledge to construct comprehensive responses. For example, an example for response generation with GPT-4 is shown as follows:

From this example, we can see that GPT-4 can provide a better response by combining the results from tools and its internal knowledge. Specifically, GPT-4 uses real-time data provided by tools, such as the prices of gold and Amazon stock and the exchange rate, to calculate the final answer to the user’s question. This demonstrates the ability of GPT-4 to integrate multiple information sources and perform complex calculations. We categorize the latest advancements in this stage into two types: direct insertion methods and information integration methods.

4.5.1 Direct Insertion methods

The methods adopted in the early work involved directly inserting the output of tools into the generated response [17,26,46,54]. For instance, if the user query is “How is the weather today?”, LLMs produce a response like “It’s Weather()” (as illustrated in Fig. 3), which is subsequently replaced with the result returned by the tool (e.g., from “It’s Weather()” to “It’s rainy.”). However, given the outputs of

An Example for Response Generation with GPT-4

Instruction Prompt: You are currently in the response generation stage. You are given the output of selected tools that can be potentially used to solve the sub-task. Your goal is to generate a response to the user question based on the results from tools and your internal knowledge.

User Question: *I would like to know the value of 5 ounces of gold plus 1 million AMZN stocks in CNY.*

Return from Tool Calling: *The stock price of AMZN is 130 USD, XAU USD is 1919.7605807522993, and the exchange rate between USD and CNY is 7.24.*

Output: *... Total value in CNY: Total value = Value of gold in CNY + Value of stocks in CNY = 69,495.87 + 941,200,000 = 941,269,495.87 CNY. Therefore, the combined value of 5 ounces of gold and 1 million Amazon stocks in Chinese Yuan is approximately 941,269,495.87 ...*

tools are unpredictable, this method could potentially affect the user experience.

4.5.2 Information integration methods

Most methodologies opt to incorporate the output of tools into the context as input to LLMs, thereby enabling the LLMs to craft a superior reply based on the information provided by the tool [89,189,190]. However, due to the limited context length of LLMs, some tool outputs cannot be directly fed into them. Consequently, various methods have emerged to address this issue. For example, RestGPT [93] simplifies the lengthy results using the pre-created schema, which is a documentation that elaborates on the examples, format, and possible errors. ToolLLaMA [18] resorts to truncation, cutting the output to fit within the length constraints, which potentially loses the required information to solve the user query. Conversely, ReCOMP [137] develops a compressor to condense lengthy information into a more succinct format, which keeps only the most useful information. ConAgents [133] proposes a schema-free method, enabling the observing agent to dynamically generate a function adaptive to extracting the target output following the instruction. And some studies suggest that refining the response generated by LLMs using the tool feedback is more effective than generating the response after invoking the tool [51,191,192].

Remark. In conclusion, direct insertion methods embed tool outputs directly into the generated response. These approaches are straightforward but are only suitable for simple tool outputs. Conversely, information integration methods allow LLMs to process tool results to generate responses. These methods are more powerful and can provide better responses, enhancing user experience. However, future work should consider how to address issues related to overly lengthy tool outputs and the inclusion of multiple other modalities. Meanwhile, some studies highlight that LLMs can generate biased or harmful content and potentially leak sensitive information [193–195]. By incorporating external tools, whether through direct insertion methods or information integration methods, the generated responses are influenced by the tool results, which can help mitigate some of the biases and harmful content originating from the LLM itself. Despite these benefits, the introduction of external tools necessitates rigorous validation of tool outputs to prevent adversarial attacks. Without adequate validation, attackers could manipulate tool results, causing LLMs to generate harmful or malicious responses. Future work should focus on enhancing the ability of LLMs to detect harmful information within tool outputs and develop effective filtering mechanisms to prevent the generation of harmful content.

5 Benchmarks, toolkits, and evaluation

In this section, we systematically summarize and categorize the benchmarks, toolkits, and evaluation methods that are tailored specifically to the various stages of tool learning. This provides a structured overview of the evaluation protocols and

toolkits used to validate the effectiveness of tool learning methods, aiming to make it more convenient for readers to engage with and implement tool learning.

5.1 Benchmarks

With the advancement of research in tool learning, a considerable number of benchmarks have been developed and made available. In our survey, we compile a selection of 33 popular benchmarks¹⁾, as shown in Table 1. Each benchmark evaluates distinct facets of tool learning, offering significant contributions to their respective fields. We categorize these benchmarks into two principal classes: general benchmarks and other benchmarks.

General benchmarks. Given the current uncertainty regarding the capacity of LLMs to effectively utilize tools, a large number of benchmarks have been established to evaluate the tool learning proficiency of LLMs. As tool learning comprises four distinct stages, existing benchmarks focus on evaluating the capabilities of LLMs at different stages. For instance, MetaTool [31] and WTU-Eval [144] benchmarks are designed to assess whether LLMs can recognize the necessity of using tools and appropriately select the most suitable tool to fulfill user requirements. This assessment particularly focuses on the stages of task planning and tool selection. On the other hand, APiBench [53], ToolBench1 [33], API-BLEND [141], and Seal-Tools [140] concentrate on the abilities of LLMs to accurately choose the right tool and configure the correct parameters for its invocation, which correspond to the tool selection and tool calling stages, respectively. Additionally, RestBench [93], TaskBench [188], T-Eval [32], and UltraTool [139] extend their focus to include task planning, tool selection, and tool calling, covering three of the four stages. Subsequent studies such as APi-Bank [30], ToolBench2 [18], and ToolEyes [138] have provided a more comprehensive evaluation of the tool usage capabilities of LLMs, spanning all four stages of tool learning. Notably, ToolBench2 has constructed the existing largest tool learning dataset, comprising 16,464 tools and 126,486 instances. However, the tools included in these benchmarks often suffer from quality issues, such as being inaccessible or non-functional. Additionally, the queries in these benchmarks are typically generated by LLMs, which may not accurately reflect the true needs of users. In contrast, GTA [143], ShortcutsBench [142], and AppWorld [145] address these limitations by collecting real-world tools and generating queries driven by actual user needs.

Other benchmarks. In addition to general benchmarks, there are also benchmarks specifically designed for particular tasks. For example, ToolQA [55] focuses on enhancing the question-answering capabilities of LLMs through the use of external tools, which has developed a dataset comprising questions that LLMs can only answer with the assistance of these external tools. ToolTalk [147] and ToolSandbox [155] concentrate on the ability of LLMs to utilize tools within multi-turn

¹⁾ Given the growing interest in tool learning, this survey may not encompass all benchmarks. We welcome suggestions to expand this list.

Table 1 A detailed list of different benchmarks and their specific configurations. Symbols ①, ②, ③, and ④ represent the four stages in tool learning—task planning, tool selection, tool calling, and response generation, respectively

Benchmark	Focus	# Tools	# Instances	Tool Source	Multi-tool?	Executable?	Time
General benchmarks							
API-Bank [30]	①, ②, ③, ④	73	314	Manual Creation	√	√	2023-04
APIBench [53]	②, ③	1,645	16,450	Public Models	×	×	2023-05
ToolBench1 [33]	②, ③	232	2,746	Public APIs	×	√	2023-05
ToolAlpaca [19]	②, ③, ④	426	3,938	Public APIs	×	×	2023-06
RestBench [93]	①, ②, ③	94	157	RESTful APIs	√	×	2023-06
ToolBench2 [18]	①, ②, ③, ④	16,464	126,486	Rapid API	√	√	2023-07
MetaTool [31]	①, ②	199	21,127	OpenAI Plugins	√	×	2023-10
TaskBench [188]	①, ②, ③	103	28,271	Public APIs	√	√	2023-11
T-Eval [32]	①, ②, ③	15	533	Manual Creation	√	√	2023-12
ToolEyes [138]	①, ②, ③, ④	568	382	Manual Creation	√	√	2024-01
UltraTool [139]	①, ②, ③	2,032	5,824	Manual Creation	√	×	2024-01
API-BLEND [141]	②, ③	–	189,040	Exsiting Datasets	√	√	2024-02
Seal-Tools [140]	②, ③	4,076	14,076	Manual Creation	√	×	2024-05
ShortcutsBench [142]	②, ③	1,414	7,627	Public APIs	√	√	2024-07
GTA [143]	②, ③, ④	14	229	Public APIs	√	√	2024-07
WTU-Eval [144]	①	4	916	BMTTools	√	√	2024-07
AppWorld [145]	①, ②, ③	457	750	FastAPI	√	√	2024-07
Other benchmarks							
ToolQA [55]	QA	13	1,530	Manual Creation	×	√	2023-06
ToolEmu [146]	Safety	311	144	Manual Creation	×	√	2023-09
ToolTalk [147]	Conversation	28	78	Manual Creation	×	√	2023-11
VIoT [148]	VIoT	11	1,841	Public Models	×	√	2023-12
RoTBench [149]	Robustness	568	105	ToolEyes	√	√	2024-01
MLLM-Tool [88]	Multi-modal	932	11,642	Public Models	√	√	2024-01
ToolSword [150]	Safety	100	440	Manual Creation	√	√	2024-02
SciToolBench [151]	Sci-Reasoning	2,446	856	Manual Creation	√	√	2024-02
InjecAgent [153]	Safety	17	1,054	Public APIs	×	√	2024-02
StableToolBench [152]	Stable	16,464	126,486	ToolBench2	√	√	2024-03
m&m’s [87]	Multi-modal	33	4,427	Public Models	√	√	2024-03
GeoLLM-QA [166]	Remote Sensing	117	1,000	Public Models	√	√	2024-04
ToolLens [124]	Tool Retrieval	464	18,770	ToolBench2	√	√	2024-05
SoAyBench [111]	Academic Seeking	7	792	AMiner	√	√	2024-05
ToolSandbox [155]	Conversation	34	1,032	Rapid API	√	√	2024-08
CToolEval [154]	Chinese	398	6,816	Public Apps	√	√	2024-08

dialogues. VIoT [148] focuses on the capability of using Viot tools with LLMs. RoTBench [149], ToolSword [150], and ToolEmu [146] are benchmarks that emphasize the robustness and safety issues in tool learning. These benchmarks highlight the need to improve the robustness and safety of LLMs in tool learning applications. MLLM-Tool [88] and m&m’s [87] extend tool learning into the multi-modal domain, assessing tool usage capabilities of LLMs in multi-modal contexts. Meanwhile, StableToolBench [152] advocates for the creation of a large-scale and stable benchmark for tool learning. SciToolBench [151] introduces a novel task named tool-augmented scientific reasoning, expanding the frontier of tool learning with LLMs applications. GeoLLM-QA [166] is designed to capture complex remote sensing workflows where LLMs handle complex data structures, nuanced reasoning, and interactions with dynamic user interfaces. Moreover, ToolLens [124], acknowledging that user queries in the real world are often concise yet have ambiguous and complex intent, has created a benchmark focused on the tool retrieval stage. Building on this, SoayBench [111] introduces a

specialized focus on academic information seeking, providing a comprehensive benchmark dataset along with the SOAYEval evaluation method. Lastly, CToolEval [154] extends the concept of tool learning to Chinese societal applications. The CToolEval benchmark is crafted to evaluate the performance of LLMs within the unique context of Chinese societal challenges, emphasizing the applicability and relevance of LLMs in this domain.

5.2 Toolkits

Recently, several open-sourced libraries and toolkits for achieving tool learning have been proposed.

LangChain is a framework for developing applications powered by large language models. It enables the creation of complex workflows by allowing LLMs to interact with APIs, databases, and other systems. LangChain is ideal for building intelligent assistants and automated systems. It can be accessed at the website of github.com/langchain-ai/langchain.

Auto-GPT is an open-source application designed to enable large language models to autonomously perform complex tasks with minimal human input. It allows models to break

down goals into subtasks and execute them sequentially, including accessing the internet and interacting with APIs. Auto-GPT is ideal for projects requiring autonomous operation, such as automated content creation and research. It can be accessed at the website of github.com/Significant-Gravitas/Auto-GPT.

BabyAGI is an open-source framework for creating autonomous AI agents that manage and execute tasks with minimal oversight. Its modular design allows developers to extend capabilities by integrating additional tools or APIs, making it ideal for flexible and scalable automation. It can be accessed at the website of github.com/yoheinakajima/babyagi.

BMTTools [16] is an open-source repository designed to enhance LLMs by integrating them with various tools and providing a community-driven platform for developing and sharing these tools. The repository allows for easy plugin creation through simple Python functions and supports the integration of external ChatGPT plugins, making it a powerful resource for extending model capabilities. It can be accessed at the website of github.com/openbmb/BMTTools.

WebCPM [196] is a framework designed for developing Chinese long-form question-answering applications using interactive web search. It allows large language models to simulate human-like web browsing behaviors to retrieve and synthesize information from various sources. WebCPM excels in creating sophisticated workflows where LLMs interact with search engines, extract relevant data, and generate comprehensive answers. It can be accessed at the website of github.com/WebCPM/WebCPM.

5.3 Evaluation

In this section, we will introduce the evaluation methods corresponding to the four stages of tool learning.

Task planning. The task planning capabilities of LLMs can be evaluated in several ways. Firstly, it is crucial to assess whether LLMs correctly identify if a given query requires an external tool, measuring the accuracy of tool usage awareness [31,139]. Next, the effectiveness of the proposed task planning in addressing the query should be evaluated, using metrics like the pass rate provided by ChatGPT [18] or human evaluations [93]. Furthermore, the precision of the plan generated by LLMs can be quantitatively analyzed by comparing it to the gold solution, ensuring its alignment and accuracy [18,32,93].

Tool selection. Existing works employ several metrics to evaluate the effectiveness of tool selection from different perspectives, including Recall, NDCG, and COMP.

Recall@K [156] is measured by calculating the proportion of selected top-K tools that are present in the set of ground-truth tools:

$$\text{Recall}@K = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{|T_q^K \cap T_q^*|}{|T_q^*|},$$

where Q is the set of queries, T_q^* is the set of relevant tools for the query q , and T_q^K is the top-K tools for the query q selected by the model.

NDCG@K [157] metric not only considers the proportion

of positive tools but also takes into account their positions within the list:

$$\text{DCG}_q@K = \sum_{i=1}^K \frac{2^{g_i} - 1}{\log_2(i+1)},$$

$$\text{NDCG}@K = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\text{DCG}_q@K}{\text{IDCG}_q@K},$$

where g_i is the graded relevance score for the i -th selected tool, and $\text{IDCG}_q@K$ denotes ideal discounted cumulative gain at the rank position k .

COMP@K [124] is designed to measure whether the top-K selected tools form a complete set with respect to the ground-truth set:

$$\text{COMP}@K = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \mathbb{I}(\Phi_q \subseteq \Psi_q^K),$$

where Φ_q denotes the set of ground-truth tools for query q , Ψ_q^K represents the top-K tools retrieved for query q , and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the retrieval results include all ground-truth tools within the top-K results for query q , and 0 otherwise.

Tool calling. In the stage of tool calling, LLMs are required to generate requests for tool calling in a specified format. The effectiveness of LLMs in executing tool calling functions can be assessed by evaluating whether the parameters input by LLMs are consistent with the stipulations delineated in the tool documentation [32,138,139]. This assessment entails verifying whether the parameters provided match those required by the specific tool, including confirming if all required parameters are included and whether the output parameters meet the required range and format.

Response generation. The ultimate goal of tool learning is to enhance the capability of LLMs to effectively address downstream tasks. Consequently, the effectiveness of tool utilization is often evaluated based on the performance in solving these downstream tasks [19,138]. This necessitates that the LLMs consolidate information gathered throughout the entire process, providing a direct response to the user query. The quality of the final response can be assessed using metrics such as BLEU score [158], ROUGE-L [159], exact match [160], F1 [141], and other relevant indicators.

6 Challenges and future directions

In this section, we identify current challenges in tool learning with LLMs and propose some promising directions for future research.

6.1 High latency in tool learning

In the reasoning process, LLMs often struggle with high latency and low throughput [197], challenges that become more pronounced when integrating tool learning. For example, even simple queries using ChatGPT with plugins can take 5 seconds to resolve, significantly diminishing the user experience compared to faster search engines. It is essential to explore ways to reduce latency, such as enhancing the

awareness of LLMs in tool utilization, enabling them to better assess when the use of tools is genuinely necessary. Additionally, maintaining the simplicity and responsiveness of tools is crucial. Overloading a single tool with too many features should be avoided to maintain efficiency and effectiveness.

6.2 Rigorous and comprehensive evaluation

Although current research demonstrates considerable advancements in tool learning with LLMs, evidenced by empirical studies across various applications, there remains a notable gap in establishing solid quantitative metrics to evaluate and understand how effectively LLMs utilize tools. Additionally, while numerous strategies have been suggested to enhance the tool learning capabilities of LLMs, a thorough comparative evaluation of these approaches is still missing. For instance, while human evaluation is capable of accurately reflecting human preferences, it is associated with significant costs and exhibits issues with repeatability, lacking in universal applicability. While the automated evaluation method, ToolEval [18], has enhanced the efficiency and reproducibility of assessments, it does not necessarily reflect the genuine preference of users. There is a need for a rigorous and comprehensive evaluation framework that considers efficiency, precision, cost, and practicality holistically. Specifically, this framework should provide independent assessments and attribution analysis for improvements at different stages, clearly delineating their specific contributions to the final response. This could involve defining new evaluation metrics and constructing assessment environments that simulate the complexity of the real world.

6.3 Comprehensive and accessible tools

While existing efforts have predominantly focused on leveraging tools to enhance the capabilities of LLMs, the quality of these tools critically impacts the performance of tool learning [46]. The majority of current tools are aggregated from existing datasets or public APIs, which imposes limitations on their accessibility and comprehensiveness. Moreover, existing datasets only contain a limited set of tools, which is unable to cover a diverse range of user queries [90]. Such constraints curtail the practical applicability and depth of tool learning. Additionally, current work acquires tools from different sources such as Public APIs [19], RESTful APIs [93], Rapid APIs [18], Hugging Face [53,89,188], or the OpenAI plugin list [31]. The diverse origins of these tools result in a variance in the format of descriptions, which hampers the development of a unified framework for tool learning. There is a pressing need to develop and compile a more comprehensive and easily accessible tool set. Given the substantial overhead associated with manually creating tools, a viable approach is to employ LLMs for the mass automatic construction of tool set [198,199]. Furthermore, the tool set should encompass a wider range of fields, offering a diverse array of functionalities to meet the specific needs of various domains. We posit that a comprehensive and accessible tool set will significantly accelerate the advancement of tool learning.

6.4 Safe and robust tool learning

Current research predominantly emphasizes the capabilities of LLMs in utilizing tools within well-structured environments, yet it overlooks the inevitable presence of noise and emerging safety considerations relevant to real-world applications. Upon deploying tool learning with LLMs into practical scenarios, safety issues and noise become unavoidable, necessitating a thoughtful approach to defend against these potential attacks. Ye et al. (2024) [149] introduce five levels of noise (Clean, Slight, Medium, Heavy, and Union) to evaluate the robustness of LLMs in tool learning. Their findings indicate a significant degradation in performance, revealing that even the most advanced model, GPT-4, exhibits poor resistance to interference. Furthermore, Ye et al. (2024) [150] devise six safety scenarios to evaluate the safety of LLMs in tool learning, uncovering a pronounced deficiency in safety awareness among LLMs, rendering them incapable of identifying safety issues within tool learning contexts. With the extensive deployment of tool learning systems across various industries, the imperative for robust security measures has significantly intensified. This necessitates a profound investigation and the introduction of innovative methodologies to fortify the security and resilience of these systems. Concurrently, in anticipation of emergent attack vectors, it is essential to perpetually refine and update security strategies to align with the rapidly evolving technological landscape.

6.5 Unified tool learning framework

As discussed in Section 4, the process of tool learning can be categorized into four distinct stages. However, prevailing research predominantly concentrates on only one of these stages for specific problems, leading to a fragmented approach and a lack of standardization. This poses significant challenges to scalability and generality in practical scenarios. It is imperative to explore and develop a comprehensive solution that encompasses task planning, tool selection, tool invocation, and response generation within a singular, unified tool learning framework.

6.6 Real-world benchmark for tool learning

Despite the substantial volume of work already conducted in the field of tool learning, the majority of queries in existing benchmarks are generated by LLMs rather than originating from real-world user queries. These synthesized queries may not accurately reflect genuine human interests and the manner in which users conduct searches. To date, there has been no publication of a tool learning dataset that encompasses authentic interactions between users and tool-augmented LLMs. The release of such a dataset, along with the establishment of a corresponding benchmark, is believed to significantly advance the development of tool learning.

6.7 Tool learning with multi-modal

While numerous studies have focused on bridging the LLMs with external tools to broaden the application scenarios, the majority of existing work on LLMs in tool learning has been confined to text-based queries. This limitation potentially leads to ambiguous interpretations of the true user intent. LLMs are poised to enhance understanding of user intent

through the integration of visual and auditory information. The increasing use of multi-modal data, such as images, audio, 3D, and video, opens up significant opportunities for further development. This encompasses exploring the capabilities of multi-modal LLMs in tool use and the combination of multi-modal tools to generate superior responses. Several pioneering research projects have explored this area. For example, Wang et al. (2024) [88] propose the MLLM-Tool, a system that incorporates open-source LLMs and multi-modal encoders, enabling the learned LLMs to be aware of multi-modal input instructions and subsequently select the correctly matched tool. Despite these initial efforts, the exploration of tool learning with multi-modal inputs has not been extensively studied. A comprehensive understanding of the capabilities of multi-modal LLMs in tool use is crucial for advancing the field.

7 Conclusion

In this paper, with reviewing more than 150 papers, we present a comprehensive survey of tool learning with LLMs. We begin the survey with a brief introduction to the concepts of “tool” and “tool learning”, providing beginners with a foundational overview and essential background knowledge. Then we elucidate the benefits of tool integration and tool learning paradigm, detailing six specific aspects to underscore why tool learning is crucial for LLMs. Moreover, to provide a more detailed introduction to how to conduct tool learning, we break down the tool learning process into four distinct phases: task planning, tool selection, tool calling, and response generation. Each phase is discussed in depth, integrating the latest research advancements to provide a thorough understanding of each step. Additionally, we summarize and categorize existing benchmarks and evaluation methods specific to these stages of tool learning, offering a structured overview of evaluation protocols. Finally, we highlight some potential challenges and identify future directions for research within this evolving field. We hope this survey can provide a comprehensive and invaluable resource for researchers and developers keen on navigating the burgeoning domain of tool learning with LLMs, thereby paving the way for future research endeavors.

Acknowledgements This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (Grant No. 62377044), Beijing Key Laboratory of Big Data Management and Analysis Methods, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, funds for building world-class universities (disciplines) of Renmin University of China, and PCC@RUC. The authors would like to extend their sincere gratitude to Yankai Lin for his constructive feedback throughout the development of this work.

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Washburn S L. Tools and human evolution. *Scientific American*, 1960, 203(3): 62–75
2. Gibson K R, Ingold T. *Tools, Language, and Cognition in Human Evolution*. Cambridge: Cambridge University Press, 1994
3. Von Eckardt B. *What Is Cognitive Science?* The MIT Press, 1995, ISBN: 9780262720236
4. Shumaker R W, Walkup K R, Beck B B. *Animal Tool Behavior: the Use and Manufacture of Tools by Animals*. Baltimore: Johns Hopkins University Press, 2011
5. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. 2024, arXiv preprint arXiv: 2303.08774
6. El-Kassab W S, Salama C R, Rafea A A, Mohamed H K. Automatic text summarization: a comprehensive survey. *Expert Systems with Applications*, 2021, 165: 113679
7. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto T B. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 2024, 12: 39–57
8. Zhang B, Haddow B, Birch A. Prompting large language model for machine translation: a case study. In: *Proceedings of the 40th International Conference on Machine Learning*. 2023, 41092–41110
9. Feng Z, Zhang Y, Li H, Liu W, Lang J, Feng Y, Wu J, Liu Z. Improving LLM-based machine translation with systematic self-correction. 2024, arXiv preprint arXiv: 2402.16379v2
10. Yang Z, Qi P, Zhang S, Bengio Y, Cohen W W, Salakhutdinov R, Manning C D. HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: *Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, 2369–2380
11. Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang M W, Dai A M, Uszkoreit J, Le Q, Petrov S. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019, 7: 452–466
12. Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2023, 9802–9822
13. Vu T, Iyyer M, Wang X, Constant N, Wei J, Wei J, Tar C, Sung Y H, Zhou D, Le Q, Luong T. FreshLLMs: refreshing large language models with search engine augmentation. In: *Proceedings of Findings of the Association for Computational Linguistics ACL 2024*. 2024, 13697–13720
14. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y J, Madotto A, Fung P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023, 55(12): 248
15. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A T, Bi W, Shi F, Shi S. Siren’s song in the AI ocean: a survey on hallucination in large language models. 2023, arXiv preprint arXiv: 2309.01219
16. Qin Y, Hu S, Lin Y, Chen W, Ding N, et al. Tool learning with foundation models. 2024, arXiv preprint arXiv: 2304.08354
17. Schick T, Dwivedi-Yu J, Dessi 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 Conference on Neural Information Processing Systems*. 2023, 36
18. 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
19. Tang Q, Deng Z, Lin H, Han X, Liang Q, Cao B, Sun L. ToolAlpaca: generalized tool learning for language models with 3000 simulated cases. 2023, arXiv preprint arXiv: 2306.05301
20. Wang H, Qin Y, Lin Y, Pan J Z, Wong K F. Empowering large language models: tool learning for real-world interaction. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024, 2983–2986

21. Yao S, Chen H, Yang J, Narasimhan K. WebShop: towards scalable real-world web interaction with grounded language agents. In: Proceedings of the 36th Conference on Neural Information Processing Systems. 2022, 20744–20757
22. Lazaridou A, Gribovskaya E, Stokowiec W, Grigorev N. Internet-augmented language models through few-shot prompting for open-domain question answering. 2022, arXiv preprint arXiv: 2203.05115
23. Lu Y, Yu H, Khashabi D. GEAR: augmenting language models with generalizable and efficient tool resolution. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024, 112–138
24. Pan L, Wu X, Lu X, Luu A T, Wang W Y, Kan M Y, Nakov P. Fact-checking complex claims with program-guided reasoning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 6981–7004
25. Wang X, Wang Z, Liu J, Chen Y, Yuan L, Peng H, Ji H. MINT: evaluating LLMs in multi-turn interaction with tools and language feedback. In: Proceedings of the 12th International Conference on Learning Representations. 2024
26. Parisi A, Zhao Y, Fiedel N. TALM: tool augmented language models. 2022, arXiv preprint arXiv: 2205.12255
27. Karpas E, Abend O, Belinkov Y, Lenz B, Lieber O, Ratner N, Shoham Y, Bata H, Levine Y, Leyton-Brown K, Muhlgaay D, Rozen N, Schwartz E, Shachaf G, Shalev-Shwartz S, Shashua A, Tenenholz M. MRKL systems: a modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. 2022, arXiv preprint arXiv: 2205.00445
28. Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, Hesse C, Jain S, Kosaraju V, Saunders W, Jiang X, Cobbe K, Eloundou T, Krueger G, Button K, Knight M, Chess B, Schulman J. WebGPT: Browser-assisted question-answering with human feedback. 2022, arXiv preprint arXiv: 2112.09332
29. Suris D, Menon S, Vondrick C. ViperGPT: visual inference via python execution for reasoning. In: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. 2023, 11854–11864
30. Li M, Zhao Y, Yu B, Song F, Li H, Yu H, Li Z, Huang F, Li Y. API-Bank: a comprehensive benchmark for tool-augmented LLMs. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 3102–3116
31. Huang Y, Shi J, Li Y, Fan C, Wu S, Zhang Q, Liu Y, Zhou P, Wan Y, Gong N Z, Sun L C. MetaTool benchmark for large language models: deciding whether to use tools and which to use. In: Proceedings of the 12th International Conference on Learning Representations. 2024
32. Chen Z, Du W, Zhang W, Liu K, Liu J, Zheng M, Zhuo J, Zhang S, Lin D, Chen K, Zhao F. T-Eval: evaluating the tool utilization capability of large language models step by step. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 9510–9529
33. Xu Q, Hong F, Li B, Hu C, Chen Z, Zhang J. On the tool manipulation capability of open-source large language models. 2023, arXiv preprint arXiv: 2305.16504
34. Gao S, Shi Z, Zhu M, Fang B, Xin X, Ren P, Chen Z, Ma J, Ren Z. Confucius: iterative tool learning from introspection feedback by easy-to-difficult curriculum. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. 2024, 18030–18038
35. Zhao Y, Wu J, Wang X, Tang W, Wang D, De Rijke M. Let me do it for you: towards LLM empowered recommendation via tool learning. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024, 1796–1806
36. Zhao W X, Zhou K, Li J, Tang T, Wang X, et al. A survey of large language models. 2024, arXiv preprint arXiv: 2303.18223
37. Huang X, Liu W, Chen X, Wang X, Wang H, Lian D, Wang Y, Tang R, Chen E. Understanding the planning of LLM agents: a survey. 2024, arXiv preprint arXiv: 2402.02716
38. Qiao S, Ou Y, Zhang N, Chen X, Yao Y, Deng S, Tan C, Huang F, Chen H. Reasoning with language model prompting: a survey. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 5368–5393
39. Sun J, Zheng C, Xie E, Liu Z, Chu R, et al. A survey of reasoning with foundation models. 2024, arXiv preprint arXiv: 2312.11562
40. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao W X, Wei Z, Wen J. A survey on large language model based autonomous agents. Frontiers of Computer Science, 2024, 18(6): 186345
41. Summers T R, Yao S, Narasimhan K, Griffiths T L. Cognitive architectures for language agents. Transactions on Machine Learning Research, 2024, ISSN: 2835-8856
42. Xi Z, Chen W, Guo X, He W, Ding Y, et al. The rise and potential of large language model based agents: a survey. 2023, arXiv preprint arXiv: 2309.07864
43. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H. Retrieval-augmented generation for large language models: a survey. 2024, arXiv preprint arXiv: 2312.10997
44. Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W, Jiang J, Cui B. Retrieval-augmented generation for AI-generated content: a survey. 2024, arXiv preprint arXiv: 2402.19473
45. Mialon G, Dessi R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Rozière B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T. Augmented language models: a survey. Transactions on Machine Learning Research, 2023, ISSN:2835-8856
46. Wang Z, Cheng Z, Zhu H, Fried D, Neubig G. What are tools anyway? A survey from the language model perspective. 2024, arXiv preprint arXiv: 2403.15452
47. Komeili M, Shuster K, Weston J. Internet-augmented dialogue generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 8460–8478
48. Zhang K, Zhang H, Li G, Li J, Li Z, Jin Z. Toolcoder: Teach code generation models to use api search tools. arXiv preprint arXiv:2305.04032, 2023
49. Shi W, Min S, Yasunaga M, Seo M, James R, Lewis M, Zettlemoyer L, Yih W T. REPLUG: retrieval-augmented black-box language models. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024, 8371–8384
50. Paranjape B, Lundberg S, Singh S, Hajishirzi H, Zettlemoyer L, Ribeiro M T. ART: automatic multi-step reasoning and tool-use for large language models. 2023, arXiv preprint arXiv: 2303.09014
51. Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Duan N, Chen W. CRITIC: large language models can self-correct with tool-interactive critiquing. In: Proceedings of the 12th International Conference on Learning Representations. 2024
52. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, et al. LaMDA: language models for dialog applications. 2022, arXiv preprint arXiv: 2201.08239
53. Patil S G, Zhang T, Wang X, Gonzalez J E. Gorilla: large language model connected with massive APIs. 2023, arXiv preprint arXiv: 2305.15334
54. Hao S, Liu T, Wang Z, Hu Z. ToolkenGPT: augmenting frozen language models with massive tools via tool embeddings. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023, 36
55. Zhuang Y, Yu Y, Wang K, Sun H, Zhang C. ToolQA: a dataset for LLM question answering with external tools. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2024, 36
56. Zhang K, Chen H, Li L, Wang W. Syntax error-free and generalizable tool use for LLMs via finite-state decoding. 2024, arXiv preprint arXiv: 2310.07075v1
57. Gu Y, Shu Y, Yu H, Liu X, Dong Y, Tang J, Srinivasa J, Latapie H, Su

- Y. Middleware for LLMs: tools are instrumental for language agents in complex environments. 2024, arXiv preprint arXiv: 2402.14672
58. Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J. Training verifiers to solve math word problems. 2021, arXiv preprint arXiv: 2110.14168
 59. Shao Z, Huang F, Huang M. Chaining simultaneous thoughts for numerical reasoning. In: Proceedings of Findings of the Association for Computational Linguistics. 2022, 2533–2547
 60. Kadlčik M, Štefánik M, Sotolar O, Martinek V. Calc-X and calcformers: empowering arithmetical chain-of-thought through interaction with symbolic systems. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 12101–12108
 61. He-Yueya J, Poesia G, Wang R E, Goodman N D. Solving math word problems by combining language models with symbolic solvers. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023
 62. Zhang B, Zhou K, Wei X, Zhao X, Sha J, Wang S, Wen J R. Evaluating and improving tool-augmented computation-intensive math reasoning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024, 1023
 63. Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Huang M, Duan N, Chen W. ToRA: a tool-integrated reasoning agent for mathematical problem solving. In: Proceedings of the 12th International Conference on Learning Representations. 2024
 64. 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. 2024, 942–966
 65. Veerendranath V, Shah V, Ghate K. Calc-CMU at semeval-2024 task 7: pre-calc – learning to use the calculator improves numeracy in language models. In: Proceedings of the 18th International Workshop on Semantic Evaluation. 2024, 1468–1475
 66. Bulusu A, Man B, Jagmohan A, Vempaty A, Mari-Wyka J, Akkil D. MathViz-E: A case-study in domain-specialized tool-using agents. 2024, arXiv preprint arXiv: 2407.17544
 67. 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
 68. Chen W, Ma X, Wang X, Cohen W W. Program of thoughts prompting: disentangling computation from reasoning for numerical reasoning tasks. Transactions on Machine Learning Research, 2023, ISSN: 2835-8856
 69. Lu P, Peng B, Cheng H, Galley M, Chang K W, Wu Y N, Zhu S C, Gao J. Chameleon: plug-and-play compositional reasoning with large language models. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023, 36
 70. Wang X, Peng H, Jabbarvand R, Ji H. LETI: learning to generate from textual interactions. In: Proceedings of Findings of the Association for Computational Linguistics. 2024, 223–239
 71. Wu J, Zhu R, Chen N, Sun Q, Li X, Gao M. Structure-aware fine-tuning for code pre-trained models. In: Proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024, 15362–15372
 72. Zhang K, Li J, Li G, Shi X, Jin Z. CodeAgent: enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 13643–13658
 73. Inaba T, Kiyomaru H, Cheng F, Kurohashi S. MultiTool-CoT: GPT-3 can use multiple external tools with chain of thought prompting. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 1522–1532
 74. Bran A M, Cox S, Schilter O, Baldassari C, White A D, Schwaller P. Augmenting large language models with chemistry tools. Nature Machine Intelligence, 2024, 6(5): 525–535
 75. Ramos M C, Collison C J, White A D. A review of large language models and autonomous agents in chemistry. 2024, arXiv preprint arXiv: 2407.01603
 76. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics, 2024, 40(2): btae075
 77. Theuma A, Shareghi E. Equipping language models with tool use capability for tabular data analysis in finance. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024, 90–103
 78. Gao S, Wen Y, Zhu M, Wei J, Cheng Y, Zhang Q, Shang S. Simulating financial market via large language model based agents. 2024, arXiv preprint arXiv: 2406.19966
 79. Zhang W, Zhao L, Xia H, Sun S, Sun J, Qin M, Li X, Zhao Y, Zhao Y, Cai X, Zheng L T, Wang X R, An B. A multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024, 4314–4325
 80. Jin Q, Wang Z, Yang Y, Zhu Q, Wright D, Huang T, Wilbur W J, He Z, Taylor A, Chen Q, Lu Z. AgentMD: empowering language agents for risk prediction with large-scale clinical tool learning. 2024, arXiv preprint arXiv: 2402.13225
 81. Li B, Yan T, Pan Y, Luo J, Ji R, Ding J, Xu Z, Liu S, Dong H, Lin Z, Wang Y. MmedAgent: learning to use medical tools with multi-modal agent. 2024, arXiv preprint arXiv: 2407.02483
 82. 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
 83. Liu Z, He Y, Wang W, Wang W, Wang Y, Chen S, Zhang Q, Yang Y, Li Q, Yu J, Li K, Chen Z, Yang X, Zhu X, Wang Y, Wang L, Luo P, Dai J, Qiao Y. InternChat: solving vision-centric tasks by interacting with chatbots beyond language. 2023, arXiv preprint arXiv: 2305.05662v2
 84. Gao D, Ji L, Zhou L, Lin K Q, Chen J, Fan Z, Shou M Z. AssistGPT: a general multi-modal assistant that can plan, execute, inspect, and learn. 2023, arXiv preprint arXiv: 2306.08640
 85. Gao Z, Du Y, Zhang X, Ma X, Han W, Zhu S C, Li Q. CLOVA: a closed-loop visual assistant with tool usage and update. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 13258–13268
 86. Zhao L, Yang Y, Zhang K, Shao W, Zhang Y, Qiao Y, Luo P, Ji R. DiffAgent: fast and accurate text-to-image API selection with large language model. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 6390–6399
 87. Ma Z, Huang W, Zhang J, Gupta T, Krishna R. *m&m*'s: a benchmark to evaluate tool-use for *multi-step multi-modal* tasks. 2024, arXiv preprint arXiv: 2403.11085
 88. Wang C, Luo W, Chen Q, Mai H, Guo J, Dong S, Xuan X, Li Z, Ma L, Gao S. Tool-LMM: a large multi-modal model for tool agent learning. 2023, arXiv preprint arXiv: 2401.10727v1
 89. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: solving AI tasks with chatGPT and its friends in hugging face. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2024, 36
 90. Lyu B, Cong X, Yu H, Yang P, Qin Y, Ye Y, Lu Y, Zhang Z, Yan Y, Lin Y, Liu Z, Sun M. GitAgent: facilitating autonomous agent with GitHub by tool extension. 2023, arXiv preprint arXiv: 2312.17294

91. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E H, Le Q V, 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, 24824–24837
92. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K R, Cao Y. ReAct: synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations. 2023
93. Song Y, Xiong W, Zhu D, Li C, Wang K, Tian Y, Li S. RestGPT: connecting large language models with real-world applications via RESTful APIs. 2023, arXiv preprint arXiv: 2306.06624v1
94. Ruan J, Chen Y, Zhang B, Xu Z, Bao T, Du G, Shi S, Mao H, Zeng X, Zhao R. TPTU: task planning and tool usage of large language model-based AI agents. 2023, arXiv preprint arXiv: 2308.03427v1
95. Zhuang Y, Chen X, Yu T, Mitra S, Bursztyjn V S, Rossi R A, Sarkhel S, Zhang C. ToolChain*: efficient action space navigation in large language models with a* search. In: Proceedings of the 12th International Conference on Learning Representations. 2024
96. Liu Z, Lai Z, Gao Z, Cui E, Li Z, Zhu X, Lu L, Chen Q, Qiao Y, Dai J, Wang W. ControlLLM: augment language models with tools by searching on graphs. 2023, arXiv preprint arXiv: 2310.17796
97. Chen Y, Lv A, Lin T E, Chen C, Wu Y, Huang F, Li Y, Yan R. Fortify the shortest stave in attention: enhancing context awareness of large language models for effective tool use. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 11160–11174
98. Huang T, Jung D, Kumar V, Kachuee M, Li X, Xu P, Chen M. Planning and editing what you retrieve for enhanced tool learning. In: Proceedings of Findings of the Association for Computational Linguistics. 2024, 975–988
99. Shi Z, Gao S, Chen X, Feng Y, Yan L, Shi H, Yin D, Chen Z, Verberne S, Ren Z. Chain of tools: large language model is an automatic multi-tool learner. 2024, arXiv preprint arXiv: 2405.16533v1
100. Wu X, Shen Y, Shan C, Song K, Wang S, Zhang B, Feng J, Cheng H, Chen W, Xiong Y, Li D. Can graph learning improve task planning? 2024, arXiv preprint arXiv: 2405.19119v1
101. Liu Y, Yuan Y, Wang C, Han J, Ma Y, Zhang L, Zheng N, Xu H. From summary to action: enhancing large language models for complex tasks with open world APIs. 2024, arXiv preprint arXiv: 2402.18157
102. Zheng Y, Li P, Yan M, Zhang J, Huang F, Liu Y. Budget-constrained tool learning with planning. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 9039–9052
103. Qu C, Dai S, Wei X, Cai H, Wang S, Yin D, Xu J, Wen J R. From exploration to mastery: enabling LLMs to master tools via self-driven interactions. 2024, arXiv preprint arXiv: 2410.08197
104. Liang Y, Wu C, Song T, Wu W, Xia Y, Liu Y, Ou Y, Lu S, Ji L, Mao S, Wang Y, Shou L, Gong M, Duan N. Taskmatrix. AI: Completing tasks by connecting foundation models with millions of APIs. *Intelligent Computing*, 2024, 3: 0063
105. Qian C, Xiong C, Liu Z, Liu Z. Toolink: linking toolkit creation and using through chain-of-solving on open-source model. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024, 831–854
106. Kong Y, Ruan J, Chen Y, Zhang B, Bao T, Shi S, Du G, Hu X, Mao H, Li Z, Zeng X, Zhao R. TPTU-v2: boosting task planning and tool usage of large language model-based agents in real-world systems. 2023, arXiv preprint arXiv: 2311.11315
107. Shen W, Li C, Chen H, Yan M, Quan X, Chen H, Zhang J, Huang F. Small LLMs are weak tool learners: a multi-LLM agent. 2024, arXiv preprint arXiv: 2401.07324
108. Gao S, Dwivedi-Yu J, Yu P, Tan X E, Pasunuru R, Golovneva O, Sinha K, Celikyilmaz A, Bosselut A, Wang T. Efficient tool use with chain-of-abstraction reasoning. 2024, arXiv preprint arXiv: 2401.17464
109. Gui A, Li J, Dai Y, Du N, Xiao H. Look before you leap: towards decision-aware and generalizable tool-usage for large language models. 2024, arXiv preprint arXiv: 2402.16696
110. Ge Y, Hua W, Mei K, Tan J, Xu S, Li Z, Zhang Y. OpenAGI: when LLM meets domain experts. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023, 36
111. Wang Y, Yu J, Yao Z, Zhang J, Xie Y, Tu S, Fu Y, Feng Y, Zhang J, Zhang J, Huang B, Li Y, Yuan H, Hou L, Li J, Tang J. A solution-based LLM API-using methodology for academic information seeking. 2024, arXiv preprint arXiv: 2405.15165
112. Chen S, Wang Y, Wu Y F, Chen Q G, Xu Z, Luo W, Zhang K, Zhang L. Advancing tool-augmented large language models: integrating insights from errors in inference trees. 2024, arXiv preprint arXiv: 2406.07115
113. Liu Z, Hoang T, Zhang J, Zhu M, Lan T, Kokane S, Tan J, Yao W, Liu Z, Feng Y, Murthy R, Yang L, Savarese S, Niebles J C, Wang H, Heinecke S, Xiong C. APIGen: automated pipeline for generating verifiable and diverse function-calling datasets. 2024, arXiv preprint arXiv: 2406.18518
114. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972, 28(1): 11–21
115. Robertson S, Zaragoza H. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 2009, 3(4): 333–389
116. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019, 3982–3992
117. Xiong L, Xiong C, Li Y, Tang K F, Liu J, Bennett P N, Ahmed J, Overwijk A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: Proceedings of the 9th International Conference on Learning Representations. 2021
118. Hofstätter S, Lin S C, Yang J H, Lin J, Hanbury A. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021, 113–122
119. Izacard G, Caron M, Hosseini L, Riedel S, Bojanowski P, Joulin A, Grave E. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022, ISSN: 2385-8856
120. Gao L, Callan J. Unsupervised corpus aware language model pre-training for dense passage retrieval. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 2843–2853
121. Yuan L, Chen Y, Wang X, Fung Y R, Peng H, Ji H. CRAFT: customizing LLMs by creating and retrieving from specialized toolsets. In: Proceedings of the 12th International Conference on Learning Representations. 2024
122. Anantha R, Bandyopadhyay B, Kashi A, Mahinder S, Hill A W, Chappidi S. ProTIP: progressive tool retrieval improves planning. 2023, arXiv preprint arXiv: 2312.10332
123. Zheng Y, Li P, Liu W, Liu Y, Luan J, Wang B. ToolRerank: adaptive and hierarchy-aware reranking for tool retrieval. In: Proceedings of 2024 Joint International Conference on Computational Linguistics,

- Language Resources and Evaluation. 2024, 16263–16273
124. Qu C, Dai S, Wei X, Cai H, Wang S, Yin D, Xu J, Wen J R. Towards completeness-oriented tool retrieval for large language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024, 1930–1940
 125. Chen Z, Zhou K, Zhang B, Gong Z, Zhao X, Wen J R. ChatCoT: tool-augmented chain-of-thought reasoning on chat-based large language models. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, 14777–14790
 126. Liu X, Peng Z, Yi X, Xie X, Xiang L, Liu Y, Xu D. ToolNet: connecting large language models with massive tools via tool graph. 2024, arXiv preprint arXiv: 2403.00839
 127. Mekala D, Weston J, Lanchantin J, Raileanu R, Lomeli M, Shang J, Dwivedi-Yu J. TOOLVERIFIER: generalization to new tools via self-verification. 2024, arXiv preprint arXiv: 2402.14158
 128. Qiao S, Gui H, Lv C, Jia Q, Chen H, Zhang N. Making language models better tool learners with execution feedback. In: Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics. 2024, 3550–3568
 129. Du Y, Wei F, Zhang H. AnyTool: self-reflective, hierarchical agents for large-scale API calls. In: Proceedings of the 41st International Conference on Machine Learning. 2024
 130. Fore M, Singh S, Stamoulis D. GeckOpt: LLM system efficiency via intent-based tool selection. In: Proceedings of Great Lakes Symposium on VLSI 2024. 2024, 353–354
 131. Zhang Y, Cai H, Song X, Chen Y, Sun R, Zheng J. Reverse chain: a generic-rule for LLMs to master multi-API planning. In: Proceedings of Findings of the Association for Computational Linguistics. 2024, 302–325
 132. Yuan S, Song K, Chen J, Tan X, Shen Y, Kan R, Li D, Yang D. EASYTOOL: enhancing LLM-based agents with concise tool instruction. 2024, arXiv preprint arXiv: 2401.06201
 133. Shi Z, Gao S, Chen X, Feng Y, Yan L, Shi H, Yin D, Ren P, Verberne S, Ren Z. Learning to use tools via cooperative and interactive agents. 2024, arXiv preprint arXiv: 2403.03031v4
 134. Yang R, Song L, Li Y, Zhao S, Ge Y, Li X, Shan Y. GPT4Tools: teaching large language model to use tools via self-instruction. In: Proceedings of the 37th Conference on Neural Information Processing Systems. 2023, 36
 135. Li L, Chai Y, Wang S, Sun Y, Tian H, Zhang N, Wu H. Tool-augmented reward modeling. In: Proceedings of the 12th International Conference on Learning Representations. 2024
 136. Wang B, Fang H, Eisner J, Van Durme B, Su Y. LLMs in the imaginarium: tool learning through simulated trial and error. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 10583–10604
 137. Xu F, Shi W, Choi E. RECOMP: improving retrieval-augmented LMS with compression and selective augmentation. In: Proceedings of the 12th International Conference on Learning Representations. 2024
 138. Ye J, Li G, Gao S, Huang C, Wu Y, Li S, Fan X, Dou S, Zhang Q, Gui T, Huang X. ToolEyes: fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. 2024, arXiv preprint arXiv: 2401.00741
 139. Huang S, Zhong W, Lu J, Zhu Q, Gao J, Liu W, Hou Y, Zeng X, Wang Y, Shang L, Jiang X, Xu R, Liu Q. Planning, creation, usage: benchmarking LLMs for comprehensive tool utilization in real-world complex scenarios. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 4363–4400
 140. Wu M, Zhu T, Han H, Tan C, Zhang X, Chen W. Seal-tools: self-instruct tool learning dataset for agent tuning and detailed benchmark. 2024, arXiv preprint arXiv: 2405.08355v1
 141. Basu K, Abdelaziz I, Chaudhury S, Dan S, Crouse M, Munawar A, Austel V, Kumaravel S, Muthusamy V, Kapanipathi P, Lastras L A. API-BLEND: a comprehensive corpora for training and benchmarking API LLMs. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 12859–12870
 142. Shen H, Li Y, Meng D, Cai D, Qi S, Zhang L, Xu M, Ma Y. ShortcutsBench: a large-scale real-world benchmark for API-based agents. 2024, arXiv preprint arXiv: 2407.00132v2
 143. Wang J, Ma Z, Li Y, Zhang S, Chen C, Chen K, Le X. GTA: a benchmark for general tool agents. 2024, arXiv preprint arXiv: 2407.08713
 144. Ning K, Su Y, Lv X, Zhang Y, Liu J, Liu K, Xu J. WTU-EVAL: a whether-or-not tool usage evaluation benchmark for large language models. 2024, arXiv preprint arXiv: 2407.12823
 145. Trivedi H, Khot T, Hartmann M, Manku R, Dong V, Li E, Gupta S, Sabharwal A, Balasubramanian N. AppWorld: a controllable world of apps and people for benchmarking interactive coding agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 16022–16076
 146. Ruan Y, Dong H, Wang A, Pitis S, Zhou Y, Ba J, Dubois Y, Maddison C J, Hashimoto T. Identifying the risks of lm agents with an lm-emulated sandbox. In: Proceedings of the 12th International Conference on Learning Representations. 2024
 147. Farn N, Shin R. ToolTalk: evaluating tool-usage in a conversational setting. 2023, arXiv preprint arXiv: 2311.10775
 148. Zhong Y, Qi M, Wang R, Qiu Y, Zhang Y, Ma H. VioTGPT: learning to schedule vision tools towards intelligent video internet of things. 2023, arXiv preprint arXiv: 2312.00401
 149. Ye J, Wu Y, Gao S, Huang C, Li S, Li G, Fan X, Zhang Q, Gui T, Huang X. RoTBench: a multi-level benchmark for evaluating the robustness of large language models in tool learning. 2024, arXiv preprint arXiv: 2401.08326
 150. Ye J, Li S, Li G, Huang C, Gao S, Wu Y, Zhang Q, Gui T, Huang X. ToolSword: unveiling safety issues of large language models in tool learning across three stages. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 2181–2211
 151. Ma Y, Gou Z, Hao J, Xu R, Wang S, Pan L, Yang Y, Cao Y, Sun A, Awadalla H, Chen W. SciAgent: Tool-augmented language models for scientific reasoning. 2024, arXiv preprint arXiv: 2402.11451
 152. Guo Z, Cheng S, Wang H, Liang S, Qin Y, Li P, Liu Z, Sun M, Liu Y. StableToolBench: towards stable large-scale benchmarking on tool learning of large language models. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 11143–11156
 153. Zhan Q, Liang Z, Ying Z, Kang D. InjecAgent: benchmarking indirect prompt injections in tool-integrated large language model agents. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 10471–10506
 154. Guo Z, Huang Y, Xiong D. CToolEval: a Chinese benchmark for LLM-powered agent evaluation in real-world API interactions. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 15711–15724
 155. Lu J, Holleis T, Zhang Y, Aumayer B, Nan F, Bai F, Ma S, Ma S, Li M, Yin G, Wang Z, Pang R. ToolSandbox: a stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. 2024, arXiv preprint arXiv: 2408.04682
 156. Zhu M. Recall, precision and average precision. University of Waterloo, Dissertation, 2004, 6
 157. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 422–446
 158. Papineni K, Roukos S, Ward T, Zhu W J. Bleu: a method for automatic

- evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002, 311–318
159. Lin C Y. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of Text Summarization Branches Out. 2004, 74–81
160. Blackwell M, Iacus S, King G, Porro G. CEM: coarsened exact matching in Stata. *The Stata Journal: Promoting communications on statistics and Stata*, 2009, 9(4): 524–546
161. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, 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
162. Dao X Q, Le N B. Investigating the effectiveness of chatGPT in mathematical reasoning and problem solving: evidence from the Vietnamese national high school graduation examination. 2023, arXiv preprint arXiv: 2306.06331
163. Wei T, Luan J, Liu W, Dong S, Wang B. CMATH: can your language model pass Chinese elementary school math test? 2023, arXiv preprint arXiv: 2306.16636
164. Chen M, Tworek J, Jun H, Yuan Q, de Oliveira Pinto H P, et al. Evaluating large language models trained on code. 2021, arXiv preprint arXiv: 2107.03374
165. Austin J, Odena A, Nye M, Bosma M, Michalewski H, Dohan D, Jiang E, Cai C, Terry M, Le Q, Sutton C. Program synthesis with large language models. 2021, arXiv preprint arXiv: 2108.07732
166. Singh S, Fore M, Stamoulis D. Evaluating tool-augmented agents in remote sensing platforms. 2024, arXiv preprint arXiv: 2405.00709v1
167. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*, 2021, 23(1): 18
168. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M. Explainability for large language models: a survey. *ACM Transactions on Intelligent Systems and Technology*, 2024, 15(2): 20
169. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, et al. Ethical and social risks of harm from language models. 2021, arXiv preprint arXiv: 2112.04359
170. Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6465–6488
171. Sun H, Cai H, Wang B, Hou Y, Wei X, Wang S, Zhang Y, Yin D. Towards verifiable text generation with evolving memory and self-reflection. 2024, arXiv preprint arXiv: 2312.09075
172. Wallace E, Feng S, Kandpal N, Gardner M, Singh S. Universal adversarial triggers for attacking and analyzing NLP. In: Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019, 2153–2162
173. Jin D, Jin Z, Zhou J T, Szolovits P. Is Bert really robust? A strong baseline for natural language attack on text classification and entailment. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020, 8018–8025
174. Wu F, Zhang N, Jha S, McDaniel P, Xiao C. A new era in LLM security: exploring security concerns in real-world LLM-based systems. 2024, arXiv preprint arXiv: 2402.18649
175. Zhang J. Graph-toolFormer: to empower LLMs with graph reasoning ability via prompt augmented by chatGPT. 2023, arXiv preprint arXiv: 2304.11116
176. Li C, Yang R, Li T, Bafarassat M, Sharifi K, Bergemann D, Yang Z. STRIDE: a tool-assisted LLM agent framework for strategic and interactive decision-making. 2024, arXiv preprint arXiv: 2405.16376
177. Huang W, Abbeel P, Pathak D, Mordatch I. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of the 39th International Conference on Machine Learning. 2022, 9118–9147
178. Chern I C, Chern S, Chen S, Yuan W, Feng K, Zhou C, He J, Neubig G, Liu P. FacTool: factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios. 2023, arXiv preprint arXiv: 2307.13528
179. Xu S, Pang L, Shen H, Cheng X, Chua T S. Search-in-the-chain: interactively enhancing large language models with search for knowledge-intensive tasks. In: Proceedings of ACM Web Conference 2024. 2024, 1362–1373
180. Kim G, Baldi P, McAleer S. Language models can solve computer tasks. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024, 1723
181. Liu Y, Peng X, Zhang Y, Cao J, Zhang X, Cheng S, Wang X, Yin J, Du T. Tool-planner: dynamic solution tree planning for large language model with tool clustering. 2024, arXiv preprint arXiv: 2406.03807v1
182. Erbacher P, Falissard L, Guigue V, Soulier L. Navigating uncertainty: optimizing API dependency for hallucination reduction in closed-book QA. In: Proceedings of the 46th European Conference on Information Retrieval. 2024, 393–402
183. Xu Q, Li Y, Xia H, Li W. Enhancing tool retrieval with iterative feedback from large language models. 2024, arXiv preprint arXiv: 2406.17465
184. Xiao S, Liu Z, Zhang P, Muennighoff N. C-pack: Pack-aged resources to advance general chinese embedding, 2023
185. Gemini Team Google. Gemini: a family of highly capable multimodal models. 2024, arXiv preprint arXiv: 2312.11805
186. Hsieh C Y, Chen S A, Li C L, Fujii Y, Ratner A, Lee C Y, Krishna R, Pfister T. Tool documentation enables zero-shot tool-usage with large language models. 2023, arXiv preprint arXiv: 2308.00675
187. Xu Y, Feng Y, Mu H, Hou Y, Li Y, Wang X, Zhong W, Li Z, Tu D, Zhu Q, Zhang M, Che W. Concise and precise context compression for tool-using language models. In: Proceedings of Findings of the Association for Computational Linguistics ACL 2024. 2024, 16430–16441
188. Shen Y, Song K, Tan X, Zhang W, Ren K, Yuan S, Lu W, Li D, Zhuang Y. TaskBench: benchmarking large language models for task automation. 2023, arXiv preprint arXiv: 2311.18760
189. Wang H, Wang H, Wang L, Hu M, Wang R, Xue B, Lu H, Mi F, Wong K F. TPE: towards better compositional reasoning over conceptual tools with multi-persona collaboration. 2023, arXiv preprint arXiv: 2309.16090
190. Qian C, Han C, Fung Y, Qin Y, Liu Z, Ji H. CREATOR: tool creation for disentangling abstract and concrete reasoning of large language models. In: Proceedings of Findings of Association for Computational Linguistics: EMNLP 2023. 2023, 6922–6939
191. Jacovi A, Caciularu A, Herzig J, Aharoni R, Bohnet B, Geva M. A comprehensive evaluation of tool-assisted generation strategies. In: Proceedings of Findings of Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 13856–13878
192. Nathani D, Wang D, Pan L, Wang W Y. MAF: multi-aspect feedback for improving reasoning in large language models. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6591–6616
193. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing*, 2024, 4: 100211
194. Cui T, Wang Y, Fu C, Xiao Y, Li S, Deng X, Liu Y, Zhang Q, Qiu Z, Li P, Tan Z, Xiong J, Kong X, Wen Z, Xu K, Li Q. Risk taxonomy, mitigation, and assessment benchmarks of large language model

- systems. 2024, arXiv preprint arXiv: 2401.05778
195. Das B C, Amini M H, Wu Y. Security and privacy challenges of large language models: a survey. 2024, arXiv preprint arXiv: 2402.00888
 196. Qin Y, Cai Z, Jin D, Yan L, Liang S, Zhu K, Lin Y, Han X, Ding N, Wang H, Xie R, Qi F, Liu Z, Sun M, Zhou J. WebCPM: interactive web search for Chinese long-form question answering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 8968–8988
 197. Miao X, Oliaro G, Zhang Z, Cheng X, Jin H, Chen T, Jia Z. Towards efficient generative large language model serving: a survey from algorithms to systems. 2023, arXiv preprint arXiv: 2312.15234
 198. Cai T, Wang X, Ma T, Chen X, Zhou D. Large language models as tool makers. In: Proceedings of the 12th International Conference on Learning Representations. 2024
 199. Wang Z, Neubig G, Fried D. TroVE: inducing verifiable and efficient toolboxes for solving programmatic tasks. In: Proceedings of the 41st International Conference on Machine Learning. 2024, 51177–51191



Changle Qu is currently pursuing the PhD degree at Gaoling School of Artificial Intelligence, Renmin University of China, China. His current research interests mainly include tool learning with large language models and information retrieval.



Sunhao Dai is a PhD candidate at Gaoling School of Artificial Intelligence, Renmin University of China, China. His current research interests lie in recommender systems and information retrieval. He has published several papers in top-tier conferences such as KDD, SIGIR, ICDE, CIKM, and RecSys.



Xiaochi Wei received PhD degree from Beijing Institute of Technology, China in 2018, under the supervision of Prof. Heyan Huang. He visited National University of Singapore, Singapore from 2015 to 2016, under the supervision of Prof. Tat-Seng Chua. He is a Senior R&D Engineer in Baidu Inc.. His research interests include question answering, multi-media information retrieval, and recommender systems. He has served as PC member in several conferences, e.g., AACL, IJCAI, ACL, and EMNLP.



Hengyi Cai received PhD degree from Institute of Computing Technology, Chinese Academy of Sciences (Outstanding Graduate), China in 2021. He joined JD's doctoral management trainee program in the summer of 2021. Previously, he was a research intern at Baidu's Search Science Team in 2020, under the supervision of Dr. Dawei Yin. His research interests include dialogue system, question

answering, and information retrieval. He served or is serving as PC member for top-tier conference including ACL, EMNLP, KDD, NeurIPS, and SIGIR.



Shuaiqiang Wang received the BSc and PhD degrees in computer science from Shandong University, China in 2004 and 2009, respectively. He is currently a principle algorithm engineer with Baidu Inc.. Previously, he was a research scientist with JD.com. Before that, he was an Assistant Professor with the University of Manchester, UK and the University of Jyväskylä, Finland. served as Senior PC Member of IJCAI, and PC Member of WWW, SIGIR, and WSDM in recent years. He is broadly interested in several research areas including information retrieval, recommender systems, and data mining.



Dawei Yin received PhD degree from Lehigh University, USA in 2013. He is senior director of engineering with Baidu inc.. He is managing the search science team with Baidu. Previously, he was senior director, managing the recommendation engineering team with JD.com between 2016 and 2019. Prior to JD.com, he was senior research manager with Yahoo Labs, leading relevance science team and in charge of Core Search Relevance of Yahoo Search. His research interests include data mining, applied machine learning, information retrieval and recommender system. He published more than 100 research papers in premium conferences and journals, and was the recipients of WSDM 2016 Best Paper Award, KDD 2016 Best Paper Award, WSDM 2018 Best Student Paper Award.



Jun Xu is a professor with the Gaoling School of Artificial Intelligence, Renmin University of China, China. His research interests focus on learning to rank and semantic matching in web search. He served or is serving as SPC for SIGIR, WWW, and AAAI, editorial board member for JASIST, and associate editor for ACM TOIS. He has won the Test of Time Award Honorable Mention in SIGIR (2019), Best Paper Award in AIRS (2010) and Best Paper Runner-up in CIKM (2017).



Ji-rong Wen is a professor of the Renmin University of China (RUC), China. He is also the dean of the School of Information and executive dean of the Gaoling School of Artificial Intelligence with RUC. His main research interests include information retrieval, data mining, and machine learning. He was a senior researcher and group manager of the Web Search and Mining Group with Microsoft Research Asia (MSRA).