Online Training of Large Language Models: Learn while Chatting

JUHAO LIANG*, ZIWEI WANG*, and ZHUOHENG MA*, The Chinese University of Hong Kong, Shenzhen, China

JIANQUAN LI, ZHIYI ZHANG, and XIANGBO WU, The Chinese University of Hong Kong, Shenzhen, China

BENYOU WANG, The Chinese University of Hong Kong, Shenzhen & Shenzhen Research Institute of Big Data, China

Large Language Models (LLMs) have dramatically revolutionized the field of Natural Language Processing (NLP), offering remarkable capabilities that have garnered widespread usage. However, existing interaction paradigms between LLMs and users are constrained by either inflexibility, limitations in customization, or a lack of persistent learning. This inflexibility is particularly evident as users, especially those without programming skills, have restricted avenues to enhance or personalize the model. Existing frameworks further complicate the model training and deployment process due to their computational inefficiencies and lack of user-friendly interfaces. To overcome these challenges, this paper introduces a novel interaction paradigm-'Online Training using External Interactions'-that merges the benefits of persistent, real-time model updates with the flexibility for individual customization through external interactions such as AI agents or online/offline knowledge bases ¹.

 $\label{eq:ccs} \mbox{CCS Concepts: \bullet Human-centered computing \rightarrow Human computer interaction (HCI); User interface toolkits; \bullet Computing methodologies \rightarrow Natural language processing; }$

Additional Key Words and Phrases: Large Language Model, User Interaction, Natural Language Processing

ACM Reference Format:

1 INTRODUCTION

Large language models (LLMs) [26, 62, 69, 78] has witnessed remarkable advancements in recent years, revolutionizing various natural language processing (NLP) tasks [2, 7, 14]. In a world where knowledge and user requirements are constantly shifting, it's critical for these models to engage in incremental learning [85] . Existing work [29, 63] tried to improve language models by self-consistency or self-reflection; however such improvement is limited to be in the scenarios where

Authors' addresses: Juhao Liang, juhaoliang1997@gmail.com; Ziwei Wang, ziweiwang2@link.cuhk.edu.cn; Zhuoheng Ma, zhuohengma@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, China; Jianquan Li; Zhiyi Zhang; Xiangbo Wu, The Chinese University of Hong Kong, Shenzhen, China; Benyou Wang, The Chinese University of Hong Kong, Shenzhen & Shenzhen Research Institute of Big Data, China, wangbenyou@cuhk.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

XXXX-XXXX/2023/3-ART \$15.00

https://doi.org/10.1145/nnnnnnn.nnnnnnn

^{*}Both authors contributed equally to this research.

¹https://github.com/FreedomIntelligence/Online-Training.git

deterministic rule based check could be used like coding and numerical answers, which might be used in open-world problems. We argue that using LLMs to improve themselves is limited; interactions between LLMs and environment are essential incremental learning.

There are two typical ways for incremental learning: offline incremental training and online incontext learning. Offline incremental training in language models involves training the system with new sets of annotated data, allowing for some level of adaptability, such as supervised fine-tuning [46] and reinforcement learning with human feedback [11, 88]. However, this approach is fraught with several limitations. First, it suffers from a lagging nature, meaning there is a delay between the emergence of new information and the model's update to include it, rendering the model less reliable for immediate or evolving scenarios. Second, the method is generally non-personalized; it updates the model based on broad data sets rather than tailoring to individual user preferences or specific contextual needs. This method's high computational cost and static nature post-update make it inflexible for adapting to real-time changes or diverse user requirements.

Another approach in the realm of incremental learning is known as 'online in-context learning'. In this setup, a LLM typically interacts with an AI agent or connects to either offline or online knowledge sources. For instance, the model might retrieve information from a given knowledge base or from web content, a.k.a, Retrieval-Augmented Generation (RAG) [30, 36, 74]. Such methods often occur within the paradigm of in-context learning (ICL) [8, 40]. However, a significant limitation of this approach is its lack of knowledge persistency. When the session changes, the learned information is not retained, leading to a loss of any updates or learning that took place.

To address the shortcomings of both 'Offline Incremental Training' and 'Online In-Context Learning,' we introduce a cutting-edge paradigm, 'Online Training using External Interactions.' This new approach offers the benefits of persistent model updates and real-time learning. Unlike traditional methods, it necessitates external interactions during the learning process, such as interfacing with AI agents or connecting to offline or online knowledge sources. These capabilities allow the model to continually adapt and stay updated, effectively addressing the limitations seen in previous frameworks. Furthermore, this innovative human-computer interaction paradigm presents a unique opportunity for ordinary users to modify LLMs themselves, thereby shifting from a developer-owned model to a more user-centric approach. This paradigm could mark the beginning of a broader practical application of LLMs in the everyday lives of an increasing number of people.

The key contributions of this work are as follows:

- We propose a novel paradigm for human interaction with language models, shifting from static
 to adaptable models. Instead of humans adapting to fixed model parameters, we introduce
 models that evolve in response to human input, fostering a dynamic and sometimes reciprocal
 relationship.
- The classification of interaction paradigm and proposal of a novel paradigm: Online Training using External Interactions, which is a user-friendly incremental learning methodology.

2 RELATED WORK AND MOTIVATION

From the users' perspective, there are currently two well-known interaction paradigms with LLMs: offline parameter-variant and online parameter-invariant. In this section, we first introduce these two prevalent user interaction paradigms and their applications, along with their applications, and detail their respective advantages and disadvantages in Sec. 2.1. Following this, in comparison to existing works, we outline the research objectives and motivation behind the proposed novel interaction paradigm in Sec. 2.2.

Table 1. Comparison of interaction paradigms. Online and offline refer to whether the model is serving, while training and parameter-invariant and parameter-variant indicate whether the parameter of the model changed. To assess and compare various interaction approaches, we focus on five key attributes: 1) Knowledge Persistency, which indicates if updated information remains accessible across different sessions; 2) Flexibility, evaluating if LLMs become static post-training; 3) Efficient Update, gauging the time and computational costs involved in model updates; 4) Knowledge Timeliness, assessing if the model's information is current; and 5) Knowledge Quality, which verifies the accuracy and reliability of the model's information. In the context of traditional training, LLMs are fine-tuned using a specific set of annotated data.

Paradigm	Methodology	Knowledge Persistency	Flexibility	Efficient Update	Knowledge Timeliness	Knowledge Quality
Offline Parameter-Variant	Traditional Training [8, 44, 61]	✓				✓
Online Parameter-Invariant	Retrieval-based methods [30, 36, 74]		✓	✓		
	Prompt-based methods [8, 40]		\checkmark	✓		
	Tool-based methods [19, 23, 35, 54, 70, 73, 80]				\checkmark	
Online Parameter-Variant	Online Training using External Interactions	✓	✓	✓	✓	✓

2.1 Related Work

Offline parameter-variant paradigm. The offline parameter-variant paradigm is the most commonly used interaction paradigm, wherein models are updated during periods of non-service. This paradigm comprises methods that are trained on a given labeled dataset. The dataset can be compiled solely by humans, as suggested by Ouyang et al. (2022) [46], which represents a conventional method of model training, specifically referred to as supervised fine-tuning (SFT). Alternatively, the process can be assisted by a retriever, as explored by [28, 42, 43]. Interaction with external knowledge during training can enhance the model's representation by integrating a larger volume of factual knowledge. For developers, offline parameter-variant methods are the most reliable and effective options for training a language model from scratch or for model updates due to their 'once and for all' characteristic, leading to high knowledge persistency and quality. However, from the user's perspective, the entire model training process can be complex, inflexible and time-consuming, ranging from data collection to computing resource configuration and model training.

Online parameter-invariant paradigm. For online parameter-invariant paradigm, there are three kinds of techniques, retrieval-based methods [30, 36, 74], prompt-based methods [8, 40], and tool-based methods [19, 23, 35, 54, 70, 73, 80]. RAG [36] is a representation of retrieval-based methods, which emphasizes the use of external knowledge sources to augment language models during inference time. Interaction with the knowledge base during inference can aid the language model in generating more precise, contextually relevant, and informed responses by dynamically leveraging external knowledge sources based on the specific input or query at hand. RAG amalgamates pretrained parametric and non-parametric memory for language generation. Designed to enhance the factual accuracy of dialogue agents, it aims to mitigate the issue of knowledge hallucination. Whereas, its performance heavily relies not only on the quality of the knowledge but also on the effectiveness of the retrieval method.

The primary objective of the *prompt-based method* is to sustain real-time, continuous engagement, making it ideal for application scenarios such as dialogue systems, real-time translations, and multi-round question answering. This iterative interaction process allows the model's output to incrementally adjust to meet user demands. Typically, this interaction paradigm does not modify the

model's parameters during the interaction, instead necessitating users to incessantly input or revise prompts to draw more meaningful responses from the language model. In-context learning [8] is a method of prompt-based, which operates without access to any external memory or knowledge beyond its pre-training phase. Consequently, while generating responses, it primarily relies on the immediate context of the conversation or task for information. Consequently, conversations can become rigid and labor-intensive due to the necessity for prompt engineering or dialogue engineering. The drawbacks of this interaction pattern are the inefficiency of proper prompts construction and failures in non-textual tasks.

By fragmenting a downstream task into multiple steps, *tool-based methods* can assign specific stages to external tools or APIs, such as those specializing in mathematical computations, web searches, image generation, and etc. Typically, tasks that emphasize fidelity and accuracy, such as real facts, complex mathematical operations, and tasks that transcend the LLM training corpus including up-to-date knowledge, low-resource languages, and image generation, are more effectively resolved using external tools than LLMs. ToolLLM [54] is a well-known tool-enhanced method. It constitutes a comprehensive framework for tool-use, offering tangible tools and components for LLMs. It is specifically engineered to empower LLMs to execute higher-level tasks, such as adhering to human instructions for utilizing external tools (APIs). However, it suffers from challenges related to invoking tools at the appropriate time and determining the most suitable tool to utilize.

Pros and Cons of Existing Paradigms. The two existing User-LLM interaction paradigms are extensively utilized in both research and practical applications. Offline parameter-variant approaches excel in knowledge persistency and quality, a result of the substantial engineering effort required before model training. This includes thorough data collection and meticulous training configuration. However, such a workload leads to inflexibility and high costs in model updates. Moreover, these methods inherently lack timeliness in knowledge updates due to their demanding workload. On the other hand, online parameter-invariant methods enable the enhancement of trained LLMs without extra training costs through prompting. However, as observed in previous research [41], the efficacy of prompting is not consistently positive. It may significantly degrade, especially if the relevant information's position varies in long context, even in models designed for long-context scenarios. Additionally, online parameter-invariant methods, like retrieval-based and tool-based approaches, often require extensive systems support, such as external databases, posing a burden for system development and sharing. Furthermore, the integrated knowledge in these methods has a short effective lifespan, expiring at the session's end or upon context removal. Overall, the pros and cons of existing paradigms are listed in Table 1.

2.2 Motivation

Motivation. In the context of knowledge persistence, the offline parameter-variant paradigm can be likened to the human brain's long-term memory, which necessitates extensive training and preparation for embedding specific knowledge. Conversely, the online parameter-invariant paradigm resembles short-term memory, where knowledge or skills can be rapidly acquired through quick learning but are not retained in the model for an extended period. This analogy highlights the strengths and weaknesses of these existing interaction paradigms. Motivated by this, we propose an intermediate interaction paradigm that amalgamates the benefits of both: the 'Online Parameter-Variant' method. This novel approach aims to reduce training costs compared to offline parameter-variant methods while offering more robust improvements than the online parameter-invariant paradigm. The 'Online Parameter-Variant' method, which is grounded in model training, focuses on several key metrics: knowledge persistency, flexibility, efficient updating, knowledge timeliness, and superior knowledge quality relative to the previous paradigms.

3 USER INTERFACE: ONLINE TRAINING USING EXTERNAL INTERACTIONS

We will first introduce the overall design in Sec. 3.1 which consists of three interactions. These three interactions are detailed in Sec. 3.2, Sec. 3.3 and Sec. 3.4.

3.1 Overall design

3.1.1 Philosophy. To address these challenges, we introduce a new interactive interface that facilitates user engagement with LLMs through conversational interactions while concurrently enabling fine-tuning through natural language instructions. The proposed system allows users to engage in conversations with a LLMs while providing specific instructions to trigger immediate fine-tuning. Users can seamlessly trigger the training process by employing natural language prompts preceded by "[Learn]," like "[Learn] I wish you could fetch more news on environmental pollution," within an interface resembling a chat, thereby commencing training grounded on network-sourced information.

Upon receiving the triggering signal, our system will comprehend the user's intended meaning and initiate distinct learning processes accordingly. After the training is completed, the newly enhanced model, enriched with incremental knowledge, will immediately replace the preceding model and seamlessly resume the ongoing conversation with users.

3.1.2 The three interactions. Online Training using External Interactions introduces three unique learning functionalities that form a comprehensive and versatile toolkit for interactive model training: Instruction-Guided Learning, Document-Driven Learning, and Web Search-Enabled Learning.

Instruction-Guided Learning serves as a soft knowledge source, leveraging conversational interfaces like ChatGPT ² to facilitate human-like, adaptive responses. This functionality is particularly powerful for nuanced or subjective queries where human-like interpretation and flexibility are required.

On the other hand, **Document-Driven Learning** and **Web Search-Enabled Learning** act as hard knowledge sources. **Document-Driven Learning** relies on offline sources, allowing for quality-controlled, curated information to be used in model training. This is particularly advantageous for tasks that require authoritative or highly reliable information. **Web Search-Enabled Learning** utilizes online data, offering the advantage of real-time information retrieval. While this allows the model to stay current, it can sometimes introduce bias or less reliable data into the training set.

By using self-instruct [75], instruction backtranslation [38], and online search augmentation, these functionalities allow for a high degree of customizability and adaptability. They empower users to shape their models according to specific instructions, documents, or real-time web data, thus bridging the gap between static, pre-trained models and dynamic, personalized user needs. Together, these functionalities make our framework not only versatile but also user-centric, enabling continuous improvement and adaptability across various application domains. This effectively highlights the complementary nature of the three functionalities in offering different sources and reliability of knowledge, serving to create a balanced, comprehensive approach for online learning with interaction.

3.1.3 Content Moderation Control. Content moderation is crucial for maintaining the integrity of LLMs. To ensure effective moderation and reduce the risk of generating biased, toxic, or unethical content, the proposed interface utilizes two primary strategies: Prevention and Feedback. First, we employ an external interface specifically designed to monitor and address content moderation issues

²https://chat.openai.com

during training. Second, we integrate a user feedback mechanism, enabling users to contribute to the moderation process through their interactions and observations. For the prevention aspect, all data used for model updates will undergo rigorous scrutiny, filtering out any inappropriate content to ensure that sources, whether local documents or Internet-based, are suitable for LLM moderation. As for the feedback component, a 'feedback' button is available to users, allowing them to report biases or any unsatisfactory elements of the response. When used, the model initiates an updating process, making corrections through a pre-defined mechanism. Interestingly, this feedback mechanism can be viewed as a form of online version reinforcement learning with human feedback (online RLHF) [46], aimed at aligning LLMs with actual user values and accentuating the personalized aspect of the proposed method.

3.1.4 Benefits and potential.

Lifelong Learning. The concept of lifelong learning [5] refers to the ability of a system, in this case, a language model, to continuously acquire and integrate new knowledge and skills throughout its existence. Unlike traditional training methods that rely on static datasets, our approach leverages continuous user interactions to adapt and evolve the model over time. This enables lifelong learning, where the model can continuously improve and stay relevant to the user's changing needs and interests.

Personalization [10]. Through our interactive training mode, users have the power to customize the language model to their liking. This customization extends beyond simple prompt-based instructions and allows users to fine-tune the model's behavior to suit their unique requirements. This level of personalization results in models that are highly specialized and context-aware, enhancing their utility for specific tasks and domains.

Accessibility. The heart of our approach lies in allowing users to engage in natural language conversations with the language model. This familiarity with conversational interactions makes it accessible to a wide range of users, regardless of their technical background. Users don't need to learn complex programming or command syntax; they simply converse as they would with another person.

User Empowerment. Users are in control of the training process. They can decide when and how to fine-tune the model based on their needs. This sense of empowerment fosters a feeling of ownership over the model, enhancing user engagement and motivation to participate in the training process. Additionally users can track the progress of their customized model over time. They can see how their interactions and commands have shaped the model's behavior, providing a sense of achievement and transparency in the training process.

3.2 Instruction-Guided Learning

Instruction-Guided Learning constitutes a fundamental component of our interactive language model fine-tuning system, enabling users to impart specific directives to the model regarding information retention and contextual understanding during ongoing conversational interactions. This method is particularly instrumental in customizing the model's responses to align with the user's unique requirements and preferences.

Within the Instruction-Guided Learning method, users are granted the capability to issue explicit instructions to the language model. These instructions may pertain to what facts or details the model should remember or any specific information that should be considered during the discourse. Users can convey their instructions in a natural language format, making it an intuitive and user-friendly

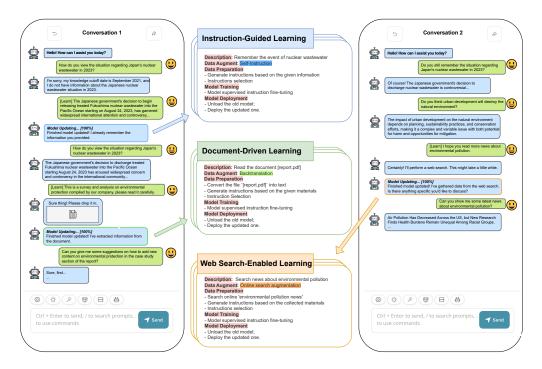


Fig. 1. The figure depicts the manner in which dialogues are conducted between LLM and user within our interactive mode. Notably, users issue distinct directives, each leading to the trigger of three distinct training processes. Furthermore, the figure underscores the model's ability to retain knowledge acquired during prior conversational session, even when transitioning across different conversation sessions.

process. These directives will be transmuted into trainable data via the self-instruct approach [75], after which we will proceed to iteratively enhance our model using the data thus generated.

After receiving user instructions, the model promptly incorporates the provided information into its understanding of the ongoing conversation. This entails the identification and retention of salient details and context specified by the user. The model's responses are then guided by the personalized context, resulting in responses that align closely with the user's directives.

As the conversation unfolds, the model continuously adapts its responses based on the instructions and context provided by the user. This iterative adaptation process allows the model to tailor its responses, ensuring that it adheres to the user's preferences and maintains a coherent and contextually relevant dialogue. Consequently, the user experiences a personalized and highly responsive conversational interface.

Instruction-Guided Learning offers users a powerful means to personalize the language model's behavior and responses in a conversational setting. By issuing explicit directives, users can shape the model's understanding and context, thereby tailoring its responses to their unique requirements. This method enhances the utility of language models across various applications, including personal virtual assistants, domain-specific chatbots, and tailored information retrieval systems, making them versatile tools for a diverse range of user needs.

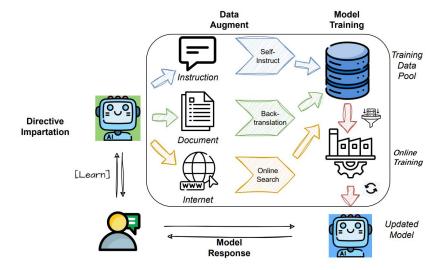


Fig. 2. This figure delineates our comprehensive workflow of chat-based online training. During the interaction between the user and the model, the user issues learning instructions to trigger the learning process. Three different learning methods correspond to three data augmentation techniques with the generated data as input to train new model. Then new model replace the old one seamlessly, allowing the user to continue the conversation.

3.3 Document-Driven Learning

Document-Driven Learning constitutes a pivotal facet of our interactive language model fine-tuning system, offering users the capability to enrich the model's knowledge base with structured and specialized information. This mode is particularly suited for users who seek to imbue the model with domain-specific expertise or train it on authoritative documents, academic texts, or specialized knowledge sources.

Users initiate the Document-Driven Learning method by selecting and uploading documents relevant to their specific area of interest or domain. These documents may encompass scholarly articles, technical manuals, legal documents, or any textual resources germane to the subject matter. The system accepts a variety of file formats, including PDFs, text documents, and web links.

Upon document submission, the system undertakes a comprehensive preprocessing and transformation procedure. Document-derived data produced through the utilization of Instruction Backtranslation [38] will be meticulously curated for high-quality training purposes. Through iterative fine-tuning, the model adapts to the new information derived from the uploaded documents. It learns to contextualize the content, recognize domain-specific terminology, and develop a deeper understanding of the subject matter. Consequently, the model's responses become more nuanced and contextually relevant when engaging in discussions related to the uploaded documents or the associated domain.

Document-Driven Learning represents a potent mechanism for users to imbue language models with domain-specific knowledge and expertise. By leveraging structured textual resources, users can enhance the model's contextual understanding and its capacity to provide informed responses within specialized domains. This method extends the utility of language models across a wide array of professional and academic applications, enabling them to serve as versatile and knowledgeable conversational partners.

3.4 Web Search-Enabled Learning

The integration of Web Search-Enabled Learning constitutes the third facet within the framework of our interactive language model fine-tuning system, affording users the capability to harness the vast knowledge repository of the internet to augment the model's understanding and responsiveness. This method is particularly valuable for users seeking real-time information, staying updated on current events, or training the model on a dynamic and ever-evolving knowledge landscape.

Upon receiving user search instructions, the system promptly conducts web searches using well-established search engines and APIs. The retrieved web content, which may include news articles, blog posts, research papers, and other relevant sources, is then subjected to information extraction and summarization processes to distill the key insights and facts. The extracted information from web searches serves as a valuable source of training data for the model [75]. During the subsequent fine-tuning phase, the model is exposed to the insights obtained from web searches.

An inherent advantage of Web Search-Enabled Learning is the model's adaptability to real-time information. As the web content evolves, the model continuously adapts to the dynamic knowledge landscape, ensuring that its responses remain up-to-date and accurate in the context of the ongoing conversation. This real-time adaptation is particularly advantageous for users seeking the latest information and insights.

The knowledge derived from web searches becomes an integral part of the model's memory, enriching its understanding of contemporary topics and factual information. This knowledge integration ensures that the model remains a reliable source of current events, trending topics, and dynamic knowledge domains over time.

Web Search-Enabled Learning empowers users to leverage the extensive resources of the internet to enhance the language model's knowledge and responsiveness. By instructing the system to retrieve real-time information, users ensure that the model remains current and up-to-date, making it a valuable resource for information retrieval, news updates, and dynamic knowledge domains. This method extends the utility of language models to domains requiring real-time knowledge integration and adaptation, making them versatile tools for a wide array of applications, including news summarization, trend analysis, and current event discussion.

4 APPLICATION: A CASE STUDY ON TOOL LEARNING

This section is dedicated to evaluating the effectiveness and efficiency of the proposed novel interaction paradigm, termed *Online Training using External Interactions*, abbreviated as *Online Training (OT)* in this section.

4.1 Problem setting

In this task, we assume that the user's objective is to train a LLM to effectively utilize external tools [54, 67]. To achieve this goal, we adopt the tool invocation data format outlined in Sun et al. (2023) [67], as demonstrated in Appendix A. We assess the model's accuracy in invoking the correct plugin and its corresponding inputs when presented with multiple APIs for various questions.

As illustrated in Figure 3, we employ two baseline methods: the prompt-based method (abbreviated as *Prompt*) and the full-parameter training method (abbreviated as *Full-SFT*). In the *Prompt* method, we use the base model (Llama2-7b-chat [72]) to generate answers with few-shot prompts [76] listed in the context. Conversely, the *Full-SFT* method involves leveraging external annotated training data to train the model, subsequently using the same few-shot prompts as in the *Prompt* method for the test set. And, the proposed approach, online-training (OT), involves generating corresponding training data based on user training instructions, then filtering out low-quality, including toxic or biased, data to train the original model. Subsequently, a prompt-based approach

is employed to generate answers for the test set. We select three tools from previous research [53], randomly choosing 300 data points for the test set and an additional 6k data points for the training dataset for the Full-SFT method

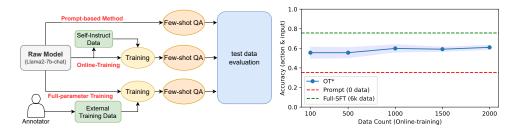


Fig. 3. Overview of the experimental design

Fig. 4. The results of the experiment, where the symbol * refers to the average of three experiments with random seed.

4.2 Result Analysis

The result are shown in Figure 4, where the x-coordinate indicates the data utilized by the *online-training* methods, and the y-coordinate shows the accuracy of each method using external tools. Accuracy is measured by counting instances of correctness in both the action (correct tool selection) and input (accurate tool parameters generation). The *Full-SFT* method utilizes 6,000 labeled indomain data points for model training. In comparison, the performance of the OT method is evaluated using between 100 and 2,000 model-generated data points for training. It is evident that using a single round of model generation ³ can achieve almost double the improvement over the vanilla model in the tool learning task, increasing from nearly 30% to 50%. However, the performance of online-training trained on 100 data points compared to 2000 data points appears similar. This is attributed to data distribution misalignment, which is still considered acceptable.

Method	Prompt	OT (0.1k)	SFT (6k)
Accuracy	0.35	0.56	0.76
Train (time)	/	2mins	40mins
Inference (time)	2mins	2mins	2mins

Table 2. Analysis of experiment duration: the time expended during the training process and the inference time for the test set are detailed.

Table 2 presents the analysis of experiment duration, detailing the time expenditures for both the training and inference processes. It becomes evident that the online-training method effectively amalgamates the benefits of both online parameter-invariant and offline parameter-variant methods, as demonstrated by its reduced training cost and heightened effectiveness on the test set.

5 DISCUSSION

This section, drawing upon the experiments and related works, delves into some concerned issues and potential challenges associated with the proposed method. Also, future possibilities in the development of User-LLM interaction paradigms are introduced.

 $^{^3\}mathrm{Approximately}$ 100 valid data points can be obtained from a single GPT-4 API call.

5.1 In-context Learning or Fine-Tuning?

There is always a question regarding language model downstream adaptation: should we opt for in-context learning or model fine-tuning for the continuous learning of trained LLMs? Both approaches have garnered considerable interest from researchers and users alike. As previously discussed, each paradigm has its strengths and weaknesses, depending on the scenario, and they are, in fact, not mutually exclusive. Moreover, increasing research [15, 47] is focusing on the relationship between ICL and Fine-Tuning. Dai et al. (2023) [15] found that ICL behaves similarly to explicit fine-tuning at the prediction level, representation level, and attention behavior level. In light of these findings, we propose a novel interaction paradigm that bridges the gap between these two existing approaches. This method involves injecting knowledge directly into the parameters, rather than solely in the context, thereby enhancing its persistency and robustness.

Scalability: One of the prominent advantages of the proposed method over ICL is its superior transferability and compositionality. For example, one could prepare specific training data for each learning job and then later decide which training data to be combined for final application. Such a combination could be done without extra inference cost as the increase of learning jobs does not affect the inference cost. Also thanks to the efficient compositionality, our approach could have a better capacity to deal with a larger-scale training, which benefits the scalability.

Inference Efficiency Our method stands out for its superior inference efficiency when compared to Retrieve and Generate (RAG) or In-Context Learning (ICL) strategies. Both RAG and ICL often result in significantly longer input prompts, which in turn leads to an increase in computational cost - a cost that grows quadratically with the length of the input. Although the training cost of our approach might be higher than that of RAG or ICL, it's important to note that this is a one-time expense related to the training phase, and remains constant regardless of the number of requests. In contrast, the cost of inference increases linearly with the number of requests, making our method more efficient in the long run. Moreover, unlike approaches that require large-scale databases or additional plugins, our method incorporates knowledge directly into the model through online training, thereby eliminating the need for external dependencies. This not only simplifies the process, but also enhances deployment readiness and operation efficiency.

5.2 Challenges

Despite the potential of the online training method, it faces several challenges:

- Knowledge Injection and Overfitting: Ovadia et al. (2023) [47] noted that LLMs often struggle to assimilate new factual information through fine-tuning. A key challenge is effectively injecting necessary knowledge into LLMs within a user-acceptable timeframe to enhance user experience. Rather than relying on a high number of training epochs, which may cause models to overfit by repeatedly training on the same data, our approach increases data diversity. This aligns with user requirements and ensures model generalizability and knowledge acquisition.
- Knowledge Persistency: Maintaining the knowledge persistency in LLMs is a crucial aspect of our proposed system. Unlike ICL-type knowledge persistency, which may involve storing information on disk, our parameter-variant methods embed knowledge directly into the LLMs' parameters. This approach ensures long-lasting knowledge retention, akin to pre-trained knowledge.
- Concurrency in LLMs Deployment: The online-training interaction paradigm we propose
 has similar deployment costs to conventional methods. It can be trained into a specific set of
 parameters, such as using LoRA, which offers the flexibility to load or unload specific training

for users. This ensures privacy and scalability without significantly increasing deployment demands.

6 CONCLUSION

In this paper, we introduce a novel interaction paradigm, online parameter-variant, and a new method, online learning using external interactions. This approach focuses on explicit model fine-tuning and instant responses to natural language instructions via a user-friendly interface. As for the future direction, we aim to break down more restrictions on users' utilization of LLMs and bring more engaging and beneficial model human-computer interactions to users, leveraging the 'Online Training using External Interactions' paradigm.

ACKNOWLEDGEMENT

This work is supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), and Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608).

REFERENCES

- [1] Albers, S. Online algorithms: a survey. Mathematical Programming 97 (2003), 3-26.
- [2] Bahl, L. R., Brown, P. F., DE SOUZA, P. V., AND MERCER, R. L. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37, 7 (1989), 1001–1008.
- [3] BENGIO, Y., AND LECUN, Y. Scaling learning algorithms towards AI. In Large Scale Kernel Machines. MIT Press, 2007.
- [4] BIAN, Z., LIU, H., WANG, B., HUANG, H., LI, Y., WANG, C., CUI, F., AND YOU, Y. Colossal-ai: A unified deep learning system for large-scale parallel training. arXiv preprint arXiv:2110.14883 (2021).
- [5] BIESIALSKA, M., BIESIALSKA, K., AND COSTA-JUSSA, M. R. Continual lifelong learning in natural language processing: A survey. In Proceedings of the 28th International Conference on Computational Linguistics (2020), International Committee on Computational Linguistics.
- [6] Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745 (2022).
- [7] Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. Large language models in machine translation.
- [8] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.
- [9] CHASE, H. LangChain, Oct. 2022.
- [10] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Lian, D., and Chen, E. When large language models meet personalization: Perspectives of challenges and opportunities, 2023.
- [11] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems 30* (2017).
- [12] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022).
- [13] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- [14] COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K., AND KUKSA, P. Natural language processing (almost) from scratch. *Journal of machine learning research 12*, ARTICLE (2011), 2493–2537.
- [15] DAI, D., SUN, Y., DONG, L., HAO, Y., MA, S., SUI, Z., AND WEI, F. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers, 2023.
- [16] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [17] Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. arXiv preprint arXiv:2301.00234 (2022).
- [18] FAIRSCALE AUTHORS. Fairscale: A general purpose modular pytorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale, 2021.
- [19] FASTCHAT AUTHORS. Lm-sys: Fastchat (vicuna: An open-source chatbot). https://github.com/lm-sys/FastChat, 2023.
- [20] GOODFELLOW, I., BENGIO, Y., COURVILLE, A., AND BENGIO, Y. Deep learning, vol. 1. MIT Press, 2016.

- [21] GUU, K., LEE, K., TUNG, Z., PASUPAT, P., AND CHANG, M. Retrieval augmented language model pre-training. In *International conference on machine learning* (2020), PMLR, pp. 3929–3938.
- [22] HAO, S., LIU, T., WANG, Z., AND HU, Z. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. arXiv preprint arXiv:2305.11554 (2023).
- [23] HAYSTACK AUTHORS. Haystack. https://github.com/deepset-ai/haystack, 2023.
- [24] HENDRYCKS, D., BURNS, C., BASART, S., ZOU, A., MAZEIKA, M., SONG, D., AND STEINHARDT, J. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).
- [25] HINTON, G. E., OSINDERO, S., AND TEH, Y. W. A fast learning algorithm for deep belief nets. *Neural Computation 18* (2006), 1527–1554.
- [26] HOFFMANN, J., BORGEAUD, S., MENSCH, A., BUCHATSKAYA, E., CAI, T., RUTHERFORD, E., CASAS, D. D. L., HENDRICKS, L. A., WELBL, J., CLARK, A., ET AL. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022).
- [27] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021).
- [28] Hu, L., Liu, Z., Zhao, Z., Hou, L., Nie, L., and Li, J. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [29] HUANG, J., GU, S. S., HOU, L., WU, Y., WANG, X., YU, H., AND HAN, J. Large language models can self-improve. arXiv preprint arXiv:2210.11610 (2022).
- [30] IZACARD, G., LEWIS, P., LOMELI, M., HOSSEINI, L., PETRONI, F., SCHICK, T., DWIVEDI-YU, J., JOULIN, A., RIEDEL, S., AND GRAVE, E. Few-shot learning with retrieval augmented language models. arXiv preprint arXiv:2208.03299 (2022).
- [31] JAISWAL, A., BABU, A. R., ZADEH, M. Z., BANERJEE, D., AND MAKEDON, F. A survey on contrastive self-supervised learning. *Technologies 9*, 1 (2020), 2.
- [32] JIANG, Z., XU, F. F., GAO, L., SUN, Z., LIU, Q., DWIVEDI-YU, J., YANG, Y., CALLAN, J., AND NEUBIG, G. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983* (2023).
- [33] JIN, D., PAN, E., OUFATTOLE, N., WENG, W.-H., FANG, H., AND SZOLOVITS, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.
- [34] KÖPF, A., KILCHER, Y., VON RÜTTE, D., ANAGNOSTIDIS, S., TAM, Z.-R., STEVENS, K., BARHOUM, A., DUC, N. M., STANLEY, O., NAGYFI, R., ET AL. Openassistant conversations—democratizing large language model alignment. *arXiv* preprint *arXiv*:2304.07327 (2023).
- [35] LANGCHAIN AUTHORS. LangChain. https://github.com/langchain-ai/langchain, 2023.
- [36] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W.-T., ROCK-TÄSCHEL, T., ET AL. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020), 9459–9474.
- [37] LI, S., FANG, J., BIAN, Z., LIU, H., LIU, Y., HUANG, H., WANG, B., AND YOU, Y. Colossal-ai: A unified deep learning system for large-scale parallel training. arXiv preprint arXiv:2110.14883 (2021).
- [38] LI, X., Yu, P., Zhou, C., Schick, T., Zettlemoyer, L., Levy, O., Weston, J., and Lewis, M. Self-alignment with instruction backtranslation. *arXiv* preprint arXiv:2308.06259 (2023).
- [39] LI, X., YU, P., ZHOU, C., SCHICK, T., ZETTLEMOYER, L., LEVY, O., WESTON, J., AND LEWIS, M. Self-alignment with instruction backtranslation, 2023.
- [40] Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? arXiv preprint arXiv:2101.06804 (2021).
- [41] LIU, N. F., LIN, K., HEWITT, J., PARANJAPE, A., BEVILACQUA, M., PETRONI, F., AND LIANG, P. Lost in the middle: How language models use long contexts, 2023.
- [42] LIU, Q., YOGATAMA, D., AND BLUNSOM, P. Relational memory-augmented language models. Transactions of the Association for Computational Linguistics 10 (2022), 555–572.
- [43] Lu, Y., Lu, H., Fu, G., AND Liu, Q. Kelm: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. *arXiv* preprint arXiv:2109.04223 (2021).
- [44] Malmi, E., Dong, Y., Mallinson, J., Chuklin, A., Adamek, J., Mirylenka, D., Stahlberg, F., Krause, S., Kumar, S., and Severyn, A. Text generation with text-editing models. *arXiv preprint arXiv:2206.07043* (2022).
- [45] Megatron-DeepSpeed authors. Megatron-DeepSpeed. https://github.com/microsoft/Megatron-DeepSpeed, 2023.
- [46] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35 (2022), 27730–27744.
- [47] OVADIA, O., BRIEF, M., MISHAELI, M., AND ELISHA, O. Fine-tuning or retrieval? comparing knowledge injection in llms, 2023
- [48] Parisi, A., Zhao, Y., and Fiedel, N. Talm: Tool augmented language models. arXiv preprint arXiv:2205.12255 (2022).
- [49] PARK, J. S., O'BRIEN, J., CAI, C. J., MORRIS, M. R., LIANG, P., AND BERNSTEIN, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and*

- Technology (2023), pp. 1-22.
- [50] PATIL, S. G., ZHANG, T., WANG, X., AND GONZALEZ, J. E. Gorilla: Large language model connected with massive apis. arXiv preprint arXiv:2305.15334 (2023).
- [51] Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813* (2023).
- [52] PETERS, M. E., NEUMANN, M., LOGAN IV, R. L., SCHWARTZ, R., JOSHI, V., SINGH, S., AND SMITH, N. A. Knowledge enhanced contextual word representations. arXiv preprint arXiv:1909.04164 (2019).
- [53] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., Fung, Y. R., Su, Y., Wang, H., Qian, C., Tian, R., Zhu, K., Liang, S., Shen, X., Xu, B., Zhang, Z., Ye, Y., Li, B., Tang, Z., Yi, J., Zhu, Y., Dai, Z., Yan, L., Cong, X., Lu, Y., Zhao, W., Huang, Y., Yan, J., Han, X., Sun, X., Li, D., Phang, J., Yang, C., Wu, T., Ji, H., Liu, Z., and Sun, M. Tool learning with foundation models, 2023.
- [54] QIN, Y., LIANG, S., YE, Y., ZHU, K., YAN, L., LU, Y., LIN, Y., CONG, X., TANG, X., QIAN, B., ET AL. Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv preprint arXiv:2307.16789 (2023).
- [55] RADFORD, A., NARASIMHAN, K., SALIMANS, T., SUTSKEVER, I., ET AL. Improving language understanding by generative pre-training.
- [56] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I., ET AL. Language models are unsupervised multitask learners. OpenAI blog 1, 8 (2019), 9.
- [57] RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W., AND LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research 21*, 1 (2020), 5485–5551.
- [58] RASLEY, J., RAJBHANDARI, S., RUWASE, O., AND HE, Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020), pp. 3505–3506.
- [59] SANH, V., WEBSON, A., RAFFEL, C., BACH, S. H., SUTAWIKA, L., ALYAFEAI, Z., CHAFFIN, A., STIEGLER, A., SCAO, T. L., RAJA, A., ET AL. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207 (2021).
- [60] SCHICK, T., DWIVEDI-YU, J., DESSÌ, R., RAILEANU, R., LOMELI, M., ZETTLEMOYER, L., CANCEDDA, N., AND SCIALOM, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* (2023).
- [61] SCHICK, T., DWIVEDI-YU, J., JIANG, Z., PETRONI, F., LEWIS, P., IZACARD, G., YOU, Q., NALMPANTIS, C., GRAVE, E., AND RIEDEL, S. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663* (2022).
- [62] Shanahan, M. Talking about large language models. arXiv preprint arXiv:2212.03551 (2022).
- [63] Shinn, N., Labash, B., and Gopinath, A. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366* (2023).
- [64] SHOEYBI, M., PATWARY, M., PURI, R., LEGRESLEY, P., CASPER, J., AND CATANZARO, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053 (2019).
- [65] Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mass: Masked sequence to sequence pre-training for language generation, 2019.
- [66] SORENSEN, T., ROBINSON, J., RYTTING, C. M., SHAW, A. G., ROGERS, K. J., DELOREY, A. P., KHALIL, M., FULDA, N., AND WINGATE, D. An information-theoretic approach to prompt engineering without ground truth labels. arXiv preprint arXiv:2203.11364 (2022).
- [67] Sun, T., Zhang, X., He, Z., Li, P., Cheng, Q., Yan, H., Liu, X., Shao, Y., Tang, Q., Zhao, X., Chen, K., Zheng, Y., Zhou, Z., Li, R., Zhan, J., Zhou, Y., Li, L., Yang, X., Wu, L., Yin, Z., Huang, X., and Qiu, X. Moss: Training conversational language models from synthetic data.
- [68] Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 (2021).
- [69] TAYLOR, R., KARDAS, M., CUCURULL, G., SCIALOM, T., HARTSHORN, A., SARAVIA, E., POULTON, A., KERKEZ, V., AND STOJNIC, R. Galactica: A large language model for science. arXiv preprint arXiv:2211.09085 (2022).
- [70] TEXT GENERATION INFERENCE AUTHORS. Text Generation Inference. https://github.com/huggingface/text-generation-inference, 2023.
- [71] TIRUMALA, K., MARKOSYAN, A., ZETTLEMOYER, L., AND AGHAJANYAN, A. Memorization without overfitting: Analyzing the training dynamics of large language models. Advances in Neural Information Processing Systems 35 (2022), 38274– 38290.
- [72] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., RODRIGUEZ, A., JOULIN, A., GRAVE, E., AND LAMPLE, G. Llama: Open and efficient foundation language models, 2023.
- [73] vLLM authors. vllm. https://github.com/vllm-project/vllm, 2023.
- [74] Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoeybi, M., Dong, Y., Kuchaiev, O., Li, B., Xiao, C., et al. Shall we

- pretrain autoregressive language models with retrieval? a comprehensive study. arXiv preprint arXiv:2304.06762 (2023).
- [75] WANG, Y., KORDI, Y., MISHRA, S., LIU, A., SMITH, N. A., KHASHABI, D., AND HAJISHIRZI, H. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 (2022).
- [76] WANG, Y., YAO, Q., KWOK, J. T., AND NI, L. M. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) 53, 3 (2020), 1–34.
- [77] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021).
- [78] WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., XIA, F., CHI, E., LE, Q. V., ZHOU, D., ET AL. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022), 24824–24837.
- [79] XI, Z., CHEN, W., GUO, X., HE, W., DING, Y., HONG, B., ZHANG, M., WANG, J., JIN, S., ZHOU, E., ET AL. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023).
- [80] YANG, Z., WU, Z., LUO, M., CHIANG, W.-L., BHARDWAJ, R., KWON, W., ZHUANG, S., LUAN, F. S., MITTAL, G., SHENKER, S., ET AL. {SkyPilot}: An intercloud broker for sky computing. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23) (2023), pp. 437–455.
- [81] YAO, Z., AMINABADI, R. Y., RUWASE, O., RAJBHANDARI, S., WU, X., AWAN, A. A., RASLEY, J., ZHANG, M., LI, C., HOLMES, C., ET AL. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. arXiv preprint arXiv:2308.01320 (2023).
- [82] YAO, Z., AMINABADI, R. Y., RUWASE, O., RAJBHANDARI, S., WU, X., AWAN, A. A., RASLEY, J., ZHANG, M., LI, C., HOLMES, C., ZHOU, Z., WYATT, M., SMITH, M., KURILENKO, L., QIN, H., TANAKA, M., CHE, S., SONG, S. L., AND HE, Y. DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales. arXiv preprint arXiv:2308.01320 (2023).
- [83] ZENG, G., HAN, X., ZHANG, Z., LIU, Z., LIN, Y., AND SUN, M. Openbmb: Big model systems for large-scale representation learning. In *Representation Learning for Natural Language Processing*. Springer, 2023, pp. 463–489.
- [84] ZENG, H. Measuring massive multitask chinese understanding. arXiv preprint arXiv:2304.12986 (2023).
- [85] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. arXiv preprint arXiv:2303.18223 (2023).
- [86] ZHENG, L., CHIANG, W.-L., SHENG, Y., ZHUANG, S., WU, Z., ZHUANG, Y., LIN, Z., LI, Z., LI, D., XING, E. P., ZHANG, H., GONZALEZ, J. E., AND STOICA, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [87] Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. Lima: Less is more for alignment, 2023.
- [88] ZIEGLER, D. M., STIENNON, N., Wu, J., BROWN, T. B., RADFORD, A., AMODEI, D., CHRISTIANO, P., AND IRVING, G. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019).

A EXPERIMENT DETAILS

For all experiments conducted in this study, we utilized four A100 GPUs, each equipped with 80GB of memory. The learning rate for SFT was set to 2e-06, while for online training, it was established at 2e-5. Following the guidelines in Tirumala et al. (2022) [71], we set the training batch epoch at 10 for OT and at 2 for SFT.

Additionally, the data format for tool invocation is exemplified as follows:

```
    Human: Can you provide a weather forecast for Rio de Janeiro, Brazil for the upcoming weekend?
    GPT: Thought: I need to use the forecast_weather API to get the weather forecast for Rio de Janeiro, Brazil for the upcoming weekend.
    Action: weather.forecast_weather
    Action Input: {"location": "Rio de Janeiro, Brazil", "days": 2}
```