EDGE: Efficient Data Selection for LLM Agents via Guideline Effectiveness

Yunxiao Zhang[1] , **Guanming Xiong**[1] , **Haochen Li**[2] and **Wen Zhao**[1*]
[1]Peking University
[2]01.AI
yunxiao.zhang@stu.pku.edu.cn, gm_xiong@pku.edu.cn, lihaochen@01.ai, zhaowen@pku.edu.cn

Abstract

Large Language Models (LLMs) have shown remarkable capabilities as AI agents. However, existing methods for enhancing LLM-agent abilities often lack a focus on data quality, leading to inefficiencies and suboptimal results in both fine-tuning and prompt engineering. To address this issue, we introduce EDGE, a novel approach for identifying informative samples without needing golden answers. We propose the Guideline Effectiveness (GE) metric, which selects challenging samples by measuring the impact of human-provided guidelines in multi-turn interaction tasks. A low GE score indicates that the human expertise required for a sample is missing from the guideline, making the sample more informative. By selecting samples with low GE scores, we can improve the efficiency and outcomes of both prompt engineering and fine-tuning processes for LLMs. Extensive experiments validate the performance of our method. Our method achieves competitive results on the HotpotQA and WebShop and datasets, requiring 75% and 50% less data, respectively, while outperforming existing methods. We also provide a fresh perspective on the data quality of LLM-agent fine-tuning.

1 Introduction

In recent years, Large Language Models (LLMs) [Ouyang *et al.*, 2022; OpenAI, 2023] have demonstrated remarkable few-shot learning and reasoning capabilities. An increasing number of studies have begun exploring how to leverage LLMs as agents that can accomplish various tasks through multiple interactions with the environment [Deng *et al.*, 2023; Liu *et al.*, 2024b; Wang *et al.*, 2024]. For example, WebShop [Yao *et al.*, 2022] provides a simulated shopping environment where agents must select products that best match user requirements.

During interactions, LLMs frequently encounter complex or previously unseen scenarios, which places substantial demands on their generalization capabilities. Numerous studies have been dedicated to mitigating this challenge.

Prior work has demonstrated the importance of guidelines (or insights) in prompt-based multi-turn interaction methods. Guidelines are natural language prompts summarized from data that contain more information and cover more scenarios than exemplars, while typically consuming less context space. Existing approaches autonomously gather experiences from training tasks through trial and error to generate these guidelines [Zhao *et al.*, 2024].

Another line of research focuses on Supervised Fine-Tuning (SFT) of open-source LLMs to enhance their instruction-following capabilities. Prior work has shown that the effectiveness of SFT depends more on dataset quality than quantity [Wang *et al.*, 2023; Zhou *et al.*, 2023]. Current data filtering approaches, including GPT-4-based scoring [Chen *et al.*, 2024], instruction difficulty assessment [Li *et al.*, 2024], and semantic diversity metrics [Lu *et al.*, 2024], have demonstrated varying degrees of success.

Despite these advancements, current LLM-agent approaches still face several pressing challenges. In prompt-based methods, existing approaches for obtaining guidelines **do not consider data quality control**, instead randomly selecting samples from annotated data, which not only requires substantial and costly annotation efforts but also suffers from noisy data problems. Meanwhile, in SFT-based methods, current approaches heavily **rely on golden answer feedback** and primarily focus on single-turn instruction tuning, lacking necessary exploration of more complex multi-turn interaction scenarios that are essential for real-world applications.

To address these challenges, we propose Efficient Data selection for LLM agents via Guideline Effectiveness, a novel framework centered around a new metric called Guideline Effectiveness (GE) to select the most informative subset of samples from a vast unlabeled data (query) pool. These selected samples can be utilized for both prompt engineering and SFT.

Guidelines represent human understanding of tasks and serve as prior knowledge for agents, encompassing tool usage patterns and comprehension of complex scenarios [Zhao *et al.*, 2024; Fu *et al.*, 2024]. The GE score essentially quantifies the impact of guidelines on each data sample, enabling us to identify which samples are most challenging for the model and thus select more informative ones. Beginning with an initial guideline, we select a small number of samples with the

[*] Corresponding author.

lowest GE scores. These samples are then analyzed to summarize error causes and update the guideline. Next, we use the updated guideline and advanced API-based LLM to annotate more low-GE-score samples instead of relying on human annotators. Notably, the updated guideline incorporates solutions for challenging samples and deeper insights into the task and tools, ensuring that the annotated data is of high quality. Finally, we can use these informative and high-quality annotated samples to fine-tune open-source LLMs.

The main contributions of this work are summarized as follows:

- Propose a novel Guideline Effectiveness metric to identify informative samples using guidelines without requiring golden answers. This metric enables efficient sample selection for both prompt engineering and model fine-tuning.
- Derive effective guidelines and obtain high-quality data for challenging multi-turn interaction tasks without the need for manual annotation, by leveraging the GE score.
- Demonstrate the effectiveness of our approach through extensive experiments on HotpotQA and WebShop benchmarks, achieving state-of-the-art performance with only 75% and 50% data requirements compared to existing methods.

2 Related Work

This study investigates how to effectively utilize guidelines in the context of data selection for supervised fine-tuning (SFT).

Data Selection for SFT aims to select a high-quality subset of data. [Zhou *et al.*, 2023] demonstrates that only 1,000 carefully curated prompts and responses can achieve remarkably strong performance. [Chen *et al.*, 2024] proposes using GPT-4 for direct quality scoring, successfully identifying 9k high-quality samples from a dataset of 52k instances. [Li *et al.*, 2024] introduces the Instruction-Following Difficulty (IFD) metric to identify discrepancies between a model's expected responses and its intrinsic generation capabilities. [Liu *et al.*, 2024a] curates 6K training samples by evaluating them along three dimensions: complexity, quality, and diversity. [Bhatt *et al.*, 2024] conducts a comprehensive evaluation of existing data selection methods that aim to maximize uncertainty and/or diversity measures. However, these evaluation metrics inherently depend on golden answers as feedback. Furthermore, they primarily focus on single-turn interactions, neglecting the complexities of multi-turn interaction scenarios. AgentTuning [Zeng *et al.*, 2024] and FiReAct [Chen *et al.*, 2023] investigate fine-tuning LLMs with multi-turn interaction trajectories generated by GPT-4, further examining the effects of multi-task learning and prompt design methods, respectively. However, both methods randomly select samples for annotation, and assume that perfectly correct trajectories (reward = 1) represent high quality. This approach may result in the inclusion of simpler problems in fine-tuning datasets, leading to low quality of fine-tuning data.

Deep Active Learning aims to identify the most informative samples for annotation, thereby reducing labeling costs. The methods are typically categorized into uncertainty-based [Settles, 2011; Kremer *et al.*, 2014], diversity-based [Sener and Savarese, 2018; Bukharin *et al.*, 2024; Bhatt *et al.*, 2024], or hybrid approaches [Azeemi *et al.*, 2025]. In the era of large language models (LLMs), some studies have attempted to integrate active learning with LLMs to achieve efficient SFT. [Azeemi *et al.*, 2025] investigates active learning for improving label efficiency in natural language generation but reports inconsistent results. [Kung *et al.*, 2023] proposed a task-level active learning framework to explore the most effective SFT tasks. However, it makes the simplifying assumption that all instances are of equal value within a task. [Bhatt *et al.*, 2024] is most similar to ours. It is the first to utilize experimental design for SFT and formulates active learning as a facility location problem. This method focuses on selecting semantically diverse and representative samples, effectively improving the generative capabilities of LLMs. However, it does not focus on addressing agent tasks that require more reasoning and decision-making capabilities.

Guideline-based Prompting aims to summarize historical interaction experiences from datasets into natural language prompts that can guide future interactions. [Zhao *et al.*, 2024] introduces Experiential Learning, which autonomously gathers experiences from training tasks through trial and error to generate instructive guidelines. [Fu *et al.*, 2024] advances this approach by automatically generating context-aware guidelines and implementing a retrieval system that selects guidelines relevant to the agent's current state. However, these approaches rely on random sampling without quality consideration and their automated summarization lacks the depth and nuance of expert knowledge.

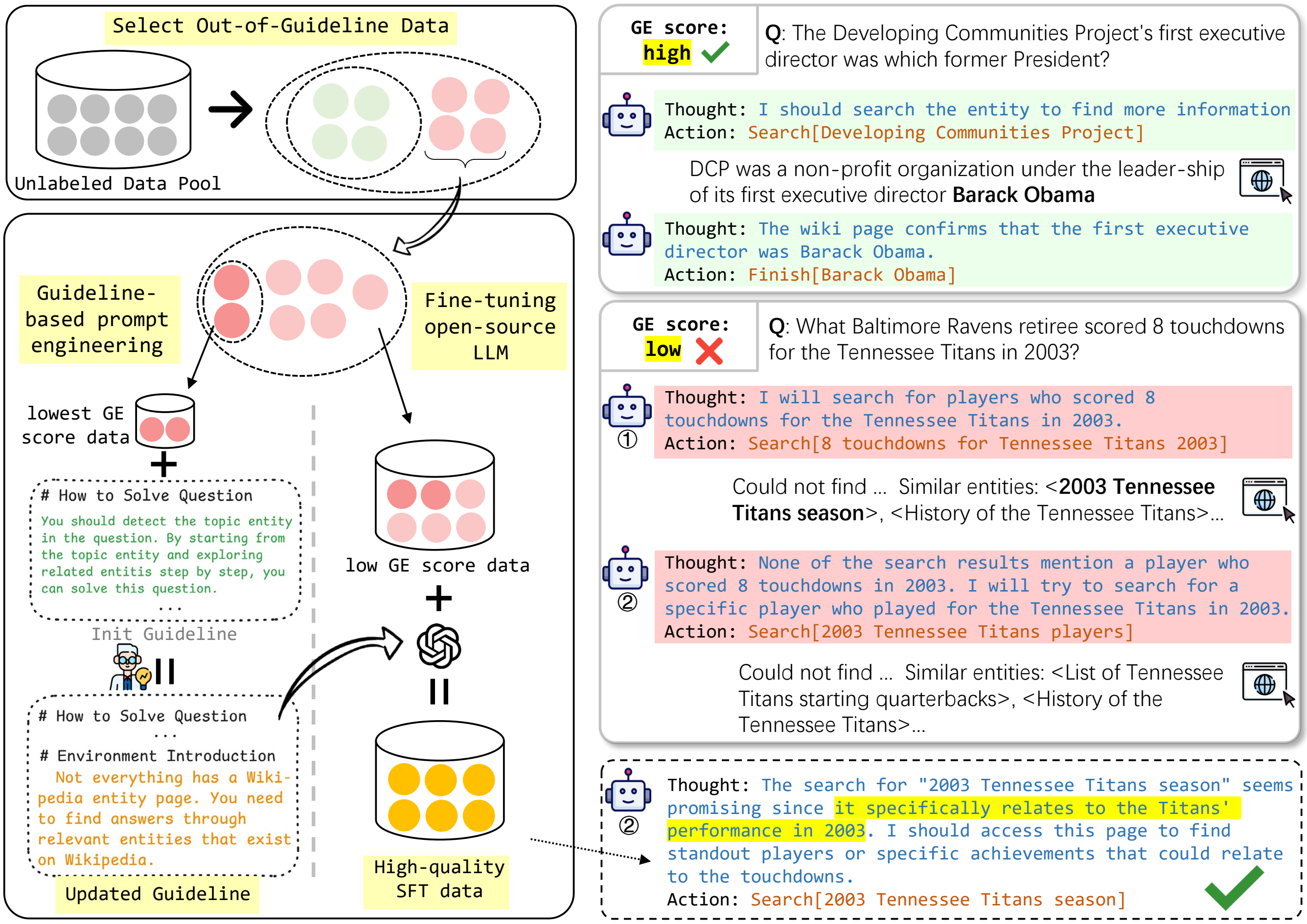
3 Methodology

3.1 Overview

Our core insight is to identify informative samples from an unlabeled data pool by leveraging guidelines, as illustrated in Figure 1. Given an unlabeled data pool and initial guidelines, we first compute the Guideline Effectiveness (GE) score for each sample using the initial guidelines. Samples with lower scores are selected for manual annotation, resulting in updated guidelines that can be directly applied to prompt-based methods. To fully utilize the unlabeled data, we incorporate the new guidelines into the prompt text and employ GPT-4 for annotation, generating question-interaction trajectory pairs as high-quality SFT data. Notably, this entire process does not require golden answers.

3.2 Preliminary

The guideline is defined as $\mathcal{G} = \text{CONCAT}(\text{g}_1, \text{g}_2, \ldots, \text{g}_n)$, where each g_i is a natural language instruction reflecting human understanding of the task. Given a set of questions $\mathcal{Q} = \{q_1, \ldots, q_n\}$, a language model LLM, and an initial guideline $\mathcal{G}^{init}$, we generate interaction trajectories $\mathcal{T} = \text{LLM}(\mathcal{Q}, \mathcal{G}^{init})$, where each trajectory $\mathcal{T} = (q, a_1, o_1, \ldots, a_T, o_T)$ consists of question-action-observation sequences with length T. Here a_i represents the action taken by the LLM at step i, and o_i denotes the observation or feedback received from the environment after taking



How to Solve Question
You should detect the topic entity in the question. By starting from the topic entity and exploring related entitis step by step, you can solve this question.
...

How to Solve Question
...
Environment Introduction
Not everything has a Wikipedia entity page. You need to find answers through relevant entities that exist on Wikipedia.

Q: The Developing Communities Project's first executive director was which former President?

Thought: I should search the entity to find more information
Action: Search[Developing Communities Project]

DCP was a non-profit organization under the leader-ship of its first executive director **Barack Obama**

Thought: The wiki page confirms that the first executive director was Barack Obama.
Action: Finish[Barack Obama]

Q: What Baltimore Ravens retiree scored 8 touchdowns for the Tennessee Titans in 2003?

Thought: I will search for players who scored 8 touchdowns for the Tennessee Titans in 2003.
Action: Search[8 touchdowns for Tennessee Titans 2003]

Could not find ... Similar entities: <**2003 Tennessee Titans season**>, <History of the Tennessee Titans>...

Thought: None of the search results mention a player who scored 8 touchdowns in 2003. I will try to search for a specific player who played for the Tennessee Titans in 2003.
Action: Search[2003 Tennessee Titans players]

Could not find ... Similar entities: <List of Tennessee Titans starting quarterbacks>, <History of the Tennessee Titans>...

Thought: The search for "2003 Tennessee Titans season" seems promising since it specifically relates to the Titans' performance in 2003. I should access this page to find standout players or specific achievements that could relate to the touchdowns.
Action: Search[2003 Tennessee Titans season]

Figure 1: The overall process of our method and an example of how guidelines correct LLM's behavior.

action a_i. Our objective is to select an informative subset $\mathcal{Q}'$ such that:

$$\mathcal{Q}' = \arg\max_{\mathcal{Q}' \subset \mathcal{Q}} (\text{Reward}(\text{LLM}')) \tag{1}$$

where Reward is the metrics of the task and LLM′ is fine-tuned by labeled $\mathcal{Q}'$. While Eq. (1) is not directly optimized due to the complexity of the reward landscape and non-convex behavior of LLMs, we approximate it using a heuristic, active learning approach in practice.

3.3 Guideline Effectiveness

Guidelines are natural language prompts enriched with expert knowledge that can cover more scenarios while consuming less context space compared to detailed exemplars. However, since humans may not initially recognize all potential challenging samples, the initial guidelines may be inadequate. Therefore, we propose a metric called Guideline Effectiveness to quantify the contribution of guidelines in solving a given question in order to identify questions that are challenging for the initial guidelines.

Given an question q, the prompt text is constructed as:

$$\text{Prompt} = \text{CONCAT}(\text{I}, \mathcal{G}^{init}, \text{E}) \tag{2}$$

where I represents the instruction text, and E denotes a set of interaction examplars. We measure the uncertainty of LLM's output action a_t at step t using the average cross-entropy loss of each token.

$$d_\theta^{\mathcal{G}}(a_t|\text{Prompt}) = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|\text{I}, \mathcal{G}^{init}, \text{E}, \mathcal{T}, w_{<i} : \theta) \tag{3}$$

where N is the number of tokens in a_t, w_i is the i-th token in a_t and θ denotes the LLM parameters. A lower $d_\theta^{\mathcal{G}}$ implies lower uncertainty in generating an action.

To evaluate guideline effectiveness, we construct $\text{Prompt}^{-\mathcal{G}} = \text{CONCAT}(\text{I}, \text{E})$ by excluding $\mathcal{G}^{init}$ from the context. The difficulty of generating action without guidelines is:

$$d_\theta^{I}(a_t|\text{Prompt}^{-\mathcal{G}}) = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|\text{I}, \text{E}, \mathcal{T}, w_{<i} : \theta) \tag{4}$$

The score d_θ^I reflects the uncertainty to generate a using only the LLM's intrinsic knowledge.

Based on $d_\theta^{\mathcal{G}}$ and d_θ^I, the GE score is defined as:

$$GE(q) = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{d_\theta^I(a_t|\text{Prompt}^{-\mathcal{G}})}{d_\theta^{\mathcal{G}}(a_t|\text{Prompt})} \tag{5}$$

GE quantifies the influence of guideline on generating each a_t by comparing the uncertainty of generating actions with and without guidelines. The intuitive interpretation of GE values is as follows: A lower value of d indicates the generation process is easier. The difficulty scores d_θ^I and $d_\theta^{\mathcal{G}}$ represent generation difficulty without and with guidelines respectively. When $GE > 0$, we have $d_\theta^I > d_\theta^{\mathcal{G}}$, indicating that guidelines facilitate action generation. A larger positive GE score suggests guidelines have a stronger positive impact on generation. As the score approaches zero, the similar magnitudes of d_θ^I and $d_\theta^{\mathcal{G}}$ indicate guidelines provide limited benefit, which are samples of particular interest. When $GE < 0$, meaning $d_\theta^I < d_\theta^{\mathcal{G}}$, guidelines appear to impede generation. This reveals cases where the LLM's inherent knowledge leads it to generate actions that conflict with the guidelines, another important category of samples to identify.

3.4 Efficiently Incorporating Human Expertise

This section describes how we utilize the GE score to incorporate human expertise into LLMs efficiently, as shown in Figure 1. Given the data pool $\mathcal{Q}$ and historical interaction trajectory $\mathcal{T}$, and initial guideline $\mathcal{G}^{init}$, we can calculate the GE score for each q_i. Lower GE scores indicate challenging questions which require additional learning by the LLM. We accomplish this in two stages: Guideline Update and High-Quality Data Generation.

Update the Guideline. We select m questions with the lowest GE scores. By observing these questions' interaction trajectories, we summarize the task-specific knowledge that the LLM lacks and update $\mathcal{G}^{init}$ to $\mathcal{G}^{new}$. Analyzing samples with the lowest GE scores allows $\mathcal{G}^{new}$ to integrate further human expertise necessary for addressing challenging samples, such as deeper insights into the task and tools. The value of m can be small (e.g., 30), which is manageable by humans within a reasonably short time. Human experts can either refine existing guidelines or introduce new ones as needed. We denote the ReAct framework augmented with the updated guideline as EDGE_{UG}.

High-Quality Data Generation. Similarly, we select k questions with the lowest GE scores and employ GPT-4 to generate interaction trajectories guided by $\mathcal{G}^{new}$. The incorporation of human expertise within $\mathcal{G}^{new}$ ensures that the trajectories maintain a high standard of quality. Utilizing these high-quality data for fine-tuning, open-sourced LLMs will implicitly learn human expertise from the annotated data. Fine-tuning the open-sourced LLMs is often crucial because: 1) As guidelines become more complex, it gets harder for open-sourced LLMs to follow; 2) Some tasks may be too intricate to distill into guidelines, making annotating data a simpler option.

4 Experiments

4.1 Baselines

To evaluate our approach, we have selected a range of state-of-the-art (SOTA) methods as baselines and conducted model comparisons along two dimensions:

EDGE vs. Other Agent Methods. We compare our method against several state-of-the-art agent methods: **ReAct** [Yao *et al.*, 2023] integrates reasoning and acting capabilities for sequential decision-making tasks; **Reflexion** [Shinn *et al.*, 2023] reinforces language agents through linguistic feedback; **AMOR** [Guan *et al.*, 2024] constructs reasoning logic over finite state machines for automated problem-solving across modules; **ExpeL** [Zhao *et al.*, 2024] leverages GPT-4 to extract guidelines from failed trajectories.

GE vs. Other Data Selection Strategies. For comparison with other label-efficient data selection strategies, we evaluate GE against several baseline approaches: **Random** selects data randomly for annotation; **Mean Entropy** [Settles, 2011; Kremer *et al.*, 2014] measures uncertainty through token-wise negative entropy of softmax probabilities; **FL** [Bhatt *et al.*, 2024] selects semantically representative samples based on diversity; **High Score** [Chen *et al.*, 2023; Zeng *et al.*, 2024] retains only fully correct interaction trajectories from annotated data.

4.2 Experiment Setup

Datasets. HotpotQA [Yang *et al.*, 2018] is a multi-hop question-answering benchmark that challenges an agent to retrieve Wikipedia passages to perform reasoning and question-answering. This involves utilizing API calls and LLM's knowledge to search for and retrieve information in order to find answers. Following [Yao *et al.*, 2023], we use three types of actions to support interactive information retrieval in HotpotQA: search[entity], lookup[query] and finish[answer]. **WebShop** [Yao *et al.*, 2022] is a simulated online shopping environment composed of a website with 1.18M real-world products. The agent's goal is to purchase a product that meets specific requirements based on a text instruction. This task requires the agent to query the website's search engine, select products with required features, and click the necessary options. Following [Liu *et al.*, 2024b], the system implements two valid actions: search[query] and click[button].

For HotpotQA, we use the first 10,000 training questions as the data pool and randomly select 500 dev questions. For WebShop, we use 8,500 instructions as the data pool and another 500 instructions for evaluation. For each dataset, we selected 30 samples with the lowest GE score for guideline updating, and then annotated 800 samples for fine-tuning. The statistical details of the test datasets are presented in Table 1.

Dataset	**Data Pool**	**EDGE Used (m / k)**	**Raw (Train / Dev)**
HotpotQA	10,000	30 / 800	90,564 / 7,405
WebShop	8,500	30 / 800	12,087 / -

Table 1: Statistics of the datasets.

Method	Base Model	HotpotQA		WebShop	
		EM	F1	Reward	SR
API-based LLM					
ReAct	GPT-4o	42.0	55.17	58.63	33.2
ExpeL	GPT-4o	47.8	60.92	64.16	42.0
Reflexion	GPT-4o	49.2	61.30	63.28	39.8
AMOR[†]	GPT-4-Turbo	55.2	65.20	-	-
EDGE$_{\mathrm{UG}}$ (Ours)	GPT-4o	**63.7**	**72.88**	**73.11**	**47.8**
Fine-tuned Open-source LLM					
ReAct	M-7B	22.6	38.31	30.77	14.2
w/ Random	M-7B	34.4	46.11	59.05	39.0
w/ ME	M-7B	35.8	47.00	58.79	38.8
w/ HS	M-7B	37.2	49.19	59.32	39.2
w/ FL	M-7B	32.8	46.06	59.00	39.0
w/ GE (Ours)	M-7B	**41.8**	**55.47**	**62.07**	**41.2**
ReAct	L-8B	35.4	45.96	37.42	18.0
w/ Random	L-8B	44.2	56.02	66.73	42.8
w/ ME	L-8B	46.0	56.13	64.3	42.4
w/ HS	L-8B	46.6	57.66	66.21	43.6
w/ FL	L-8B	40.6	52.38	66.09	43.2
w/ GE (Ours)	L-8B	**52.4**	**66.15**	**69.14**	**46.0**

Table 2: Main results. The best results are marked in **bold** and the second-best results are marked with underline. Results marked with [†] are reported in the original paper. ME denotes Mean Entropy, and HS denotes High Score.

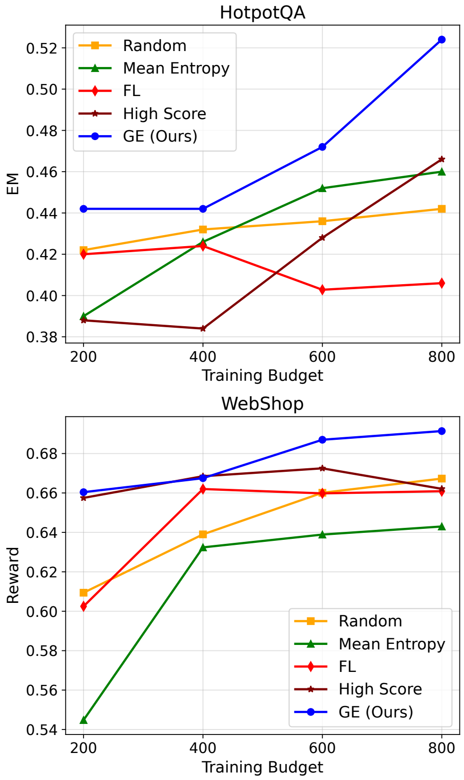


Figure 2: Comparison of different data selection strategies on various training budgets.

Evaluation Metrics. For HotpotQA, we employ two metrics: F1 and exact match (EM). F1 measures the token-level overlap between the prediction and ground truth, while EM calculates the proportion of items whose F1=1. For WebShop, reward $\in [0, 1]$ measures how well the purchased item matches the text instruction, and success rate (SR) measures the proportion of items that get reward=1.

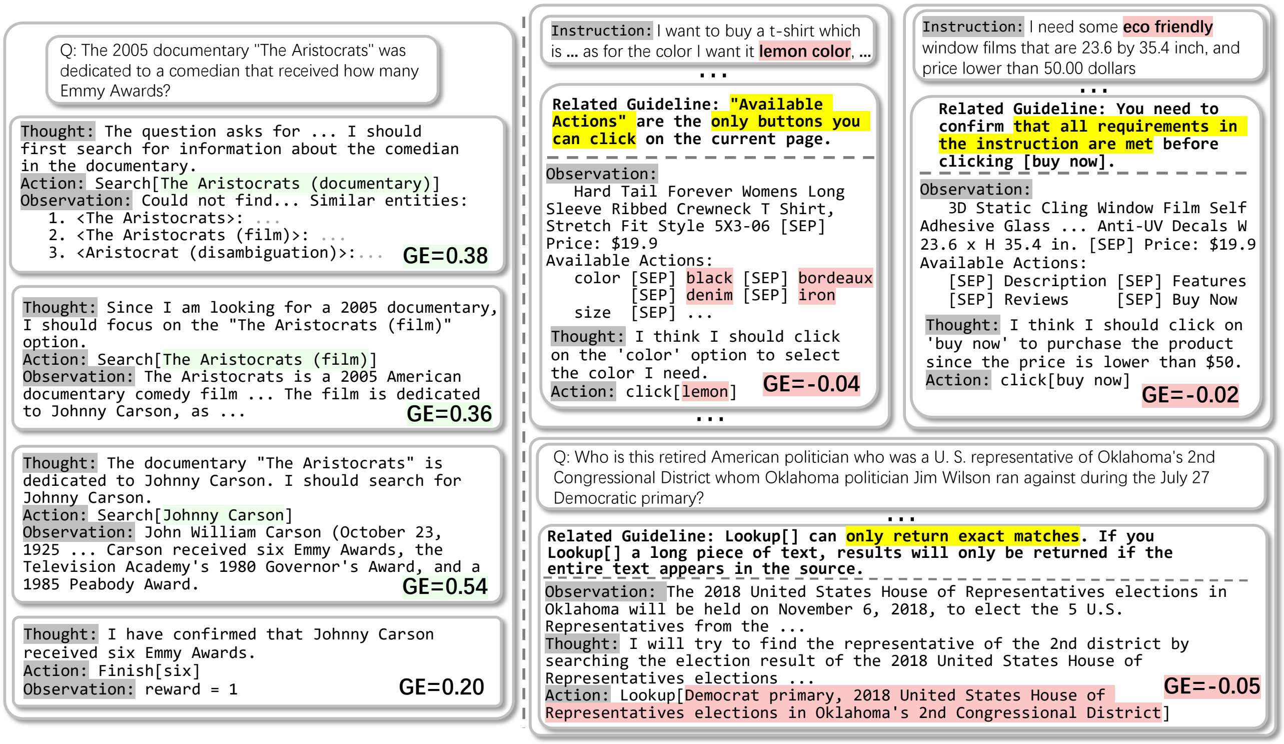
Implementation details. We invoke the OpenAI GPT4-4o (gpt-4o-2024-08-06) API. For all inference, we set temperature=0.7, top_p=0.95, max_length=512. For fine-tuning, we choose `LLAMA-3.1-8B-Instruct` (L-8B) and `Mistral-7B-Instruct-v0.3` (M-7B), training for 4 epochs with a learning rate of 5e-6 using 8 NVIDIA 80GB A100 GPUs. We use L-8B for the computation of GE score. Two master's students with domain expertise were responsible for annotation.

4.3 Main Result

EDGE yield effective guidelines. EDGE$_{\mathrm{UG}}$ surpasses baselines in both HotpotQA and WebShop, as shown in Table 2, achieving improvements of 13.3% and 13.9%, respectively. The ExpeL autonomously summarizes guidelines, but its effectiveness is constrained by the LLM's limited environmental understanding. This shortcoming hinders ExpeL's ability to generate guidelines that demand a deeper comprehension of the environment or tools. For example, WebShop displays candidate products in a semantic similarity ranking, making the top-ranked products more likely to be target products. Due to ExpeL's lack of understanding of the search engine, it is unable to summarize related guidelines.

GE-selected fine-tuning data outperform others. We use same prompt to generate fine-tuning dataset for other data selection methods. Results in Table 2 show that GE outperforms them across the two datasets using L-8B and M-7B. Notably, ReAct w/ GE (L-8B) even surpassed baselines that used GPT-4o. These findings indicate that the samples selected by GE are more challenging, and their solving trajectories integrate a greater depth of human expertise. FL focuses on selecting samples that are more semantically representative. Although it performs well on generation tasks, results indicate that it struggles with complex tasks that involve multi-turn interactions requiring reasoning and decision-making. High Score, which filters and selects entirely accurate samples from labeled data, performs relatively well. Notably, not all samples selected by GE are labeled totally correctly. This means that despite having higher rewards, the data selected by the High Score still yields inferior results compared to GE.

GE efficiently enhances fine-tuning. We compared the performance of different data selection methods with training budget k=[200, 400, 600, 800], as shown in Figure 2. GE consistently outperforms the baselines across all training budgets, achieving more efficient integration of human expertise. Moreover, our method reduces training data usage by 50% on WebShop and 25% on HotpotQA, respectively, while still achieving better performance compared to the baselines.



Q: The 2005 documentary "The Aristocrats" was dedicated to a comedian that received how many Emmy Awards?

```
Thought: The question asks for ... I should
first search for information about the comedian
in the documentary.
Action: Search[The Aristocrats (documentary)]
Observation: Could not find... Similar entities:
   1. <The Aristocrats>: ...
   2. <The Aristocrats (film)>: ...
   3. <Aristocrat (disambiguation)>:...        GE=0.38
```

```
Thought: Since I am looking for a 2005 documentary,
I should focus on the "The Aristocrats (film)"
option.
Action: Search[The Aristocrats (film)]
Observation: The Aristocrats is a 2005 American
documentary comedy film ... The film is dedicated
to Johnny Carson, as ...                        GE=0.36
```

```
Thought: The documentary "The Aristocrats" is
dedicated to Johnny Carson. I should search for
Johnny Carson.
Action: Search[Johnny Carson]
Observation: John William Carson (October 23,
1925 ... Carson received six Emmy Awards, the
Television Academy's 1980 Governor's Award, and a
1985 Peabody Award.                             GE=0.54
```

```
Thought: I have confirmed that Johnny Carson
received six Emmy Awards.
Action: Finish[six]
Observation: reward = 1                         GE=0.20
```

Instruction: I want to buy a t-shirt which is ... as for the color I want it **lemon color**, ...

...

```
Related Guideline: "Available
Actions" are the only buttons you
can click on the current page.
- - - - - - - - - - - - - - - - - -
Observation:
   Hard Tail Forever Womens Long
Sleeve Ribbed Crewneck T Shirt,
Stretch Fit Style 5X3-06 [SEP]
Price: $19.9
Available Actions:
   color [SEP] black [SEP] bordeaux
         [SEP] denim [SEP] iron
   size  [SEP] ...
Thought: I think I should click
on the 'color' option to select
the color I need.
Action: click[lemon]          GE=-0.04
```

...

Instruction: I need some **eco friendly** window films that are 23.6 by 35.4 inch, and price lower than 50.00 dollars

...

```
Related Guideline: You need to
confirm that all requirements in
the instruction are met before
clicking [buy now].
- - - - - - - - - - - - - - - - - -
Observation:
   3D Static Cling Window Film Self
Adhesive Glass ... Anti-UV Decals W
23.6 x H 35.4 in. [SEP] Price: $19.9
Available Actions:
   [SEP] Description [SEP] Features
   [SEP] Reviews     [SEP] Buy Now
Thought: I think I should click on
'buy now' to purchase the product
since the price is lower than $50.
Action: click[buy now]
                              GE=-0.02
```

Q: Who is this retired American politician who was a U. S. representative of Oklahoma's 2nd Congressional District whom Oklahoma politician Jim Wilson ran against during the July 27 Democratic primary?

...

```
Related Guideline: Lookup[] can only return exact matches. If you
Lookup[] a long piece of text, results will only be returned if the
entire text appears in the source.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Observation: The 2018 United States House of Representatives elections in
Oklahoma will be held on November 6, 2018, to elect the 5 U.S.
Representatives from the ...
Thought: I will try to find the representative of the 2nd district by
searching the election result of the 2018 United States House of
Representatives elections ...                                    GE=-0.05
Action: Lookup[Democrat primary, 2018 United States House of
Representatives elections in Oklahoma's 2nd Congressional District]
```

Figure 3: Examples of various actions' GE scores. The left side shows an example with a high GE score. The right side shows examples with lower GE scores on WebShop and HotpotQA. The red highlight indicates the reason for the lower GE value of the action.

Method	**Avg. Interaction Turns**				**Reward**				**LLM Reward**
	k=200	k=400	k=600	k=800	k=200	k=400	k=600	k=800	
FL	4.50	4.53	4.45	4.43	80.85	79.78	77.86	77.68	46.06
Random	4.97	4.84	4.75	4.87	77.56	79.21	79.81	77.69	46.11
Mean Entropy	4.82	4.92	4.89	4.96	72.17	72.97	73.18	71.10	47.00
High Score	4.77	4.60	4.65	4.64	100.00	100.00	100.00	100.00	49.19
GE (Ours)	6.40	6.38	6.43	6.39	68.06	70.84	71.09	70.27	55.47

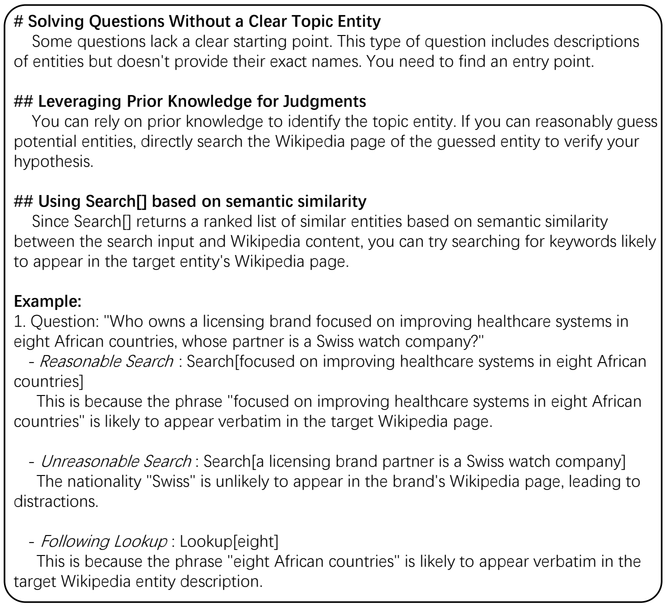
Table 3: Statistics of the annotated data by different data selection methods.

4.4 Analysis of EDGE

What kinds of samples have high/low GE score? Through manual observation of the samples, we found that GE values tend to be higher when the environment is simple, and the guideline incorporates relevant human expertise. Conversely, GE values are lower in complex environments or when the guideline lacks relevant expertise. These observations align with our hypothesis. We selected a few samples with high/low GE score for illustration. On HotpotQA, Figure 3 (Left) is a sample with high GE score. The LLM followed the guideline, progressively searching for related entities based on the topic entity until the answer was found. Figure 3 (Right) demonstrates actions with low GE score. On WebShop, samples with complex environment caused the LLM to violate the guidelines, resulting in low GE scores. In the top left example, the extensive list of available actions on the product page, combined with relevant information, misled the LLM into click[lemon], which was a non-existent button. The example below is a HotpotQA case, where this sample made the LLM use overly long search keywords.

What guidelines have we summarized? Based on the observation of low GE score samples, we have listed the following representative guidelines in order of their necessity within the samples. Table 4 shows the format of our guideline. For WebShop:

- *Click high-ranked products first.* A high ranking indicates that the semantics of the product's options or features are more similar to the search content. Even if the titles of high-ranked products sometimes do not fully meet the required criteria, it is still necessary to click on them.
- *What is a product title.* This guideline introduces that some attributes like color, flavor, etc., will not appear in the product title. LLMs often hope to find products



Solving Questions Without a Clear Topic Entity
Some questions lack a clear starting point. This type of question includes descriptions of entities but doesn't provide their exact names. You need to find an entry point.

## Leveraging Prior Knowledge for Judgments
You can rely on prior knowledge to identify the topic entity. If you can reasonably guess potential entities, directly search the Wikipedia page of the guessed entity to verify your hypothesis.

## Using Search[] based on semantic similarity
Since Search[] returns a ranked list of similar entities based on semantic similarity between the search input and Wikipedia content, you can try searching for keywords likely to appear in the target entity's Wikipedia page.

Example:
1. Question: "Who owns a licensing brand focused on improving healthcare systems in eight African countries, whose partner is a Swiss watch company?"
- *Reasonable Search* : Search[focused on improving healthcare systems in eight African countries]
This is because the phrase "focused on improving healthcare systems in eight African countries" is likely to appear verbatim in the target Wikipedia page.

- *Unreasonable Search* : Search[a licensing brand partner is a Swiss watch company]
The nationality "Swiss" is unlikely to appear in the brand's Wikipedia page, leading to distractions.

- *Following Lookup* : Lookup[eight]
This is because the phrase "eight African countries" is likely to appear verbatim in the target Wikipedia entity description.

Table 4: Example of a peice of our guideline on HotpotQA.

Method	Easy	Medium	Hard
Random	0%	0%	0%
Mean Entropy	+1.08%	+1.13%	-2.21%
High Score	-0.92%	+4.13%	-3.21%
FL	+0.71%	+0.00%	-0.71%
GE (Ours)	-1.67%	-1.50%	+3.17%

Table 5: The proportion of different difficulty levels across various methods. The percentages indicate the change in proportion compared to the original distribution of difficulty levels.

Method	Easy	Medium	Hard
ReAct	-	-	-
w/ Random	65.29	62.92	56.02
w/ Mean Entropy	62.12	62.29	56.13
w/ High Score	66.68	65.11	57.66
w/ FL	63.19	61.82	52.38
w/ GE (Ours)	68.64	66.98	66.15

Table 6: Performance comparison on different difficulty levels.

whose titles perfectly match the instructions.

- *Buying a similar product is better than buying nothing.* Before the rounds are exhausted, the LLM needs to balance the trade-off and compromise to buy a product that is not perfectly aligned when necessary.

For HotpotQA:

- *How to use 'Lookup'.* Despite being informed that "the Lookup[] only supports exact matching", the LLM still tends to search for longer keywords. This guideline details that the LLM should search for the shortest possible word that are likely to appear in the original text to match more search results.
- *Solving questions without a clear topic entity.* When the topic entity cannot be found, LLM can try leverage its prior knowledge to make a reasonable inference or use Search[] based on semantic similarity.
- *Understanding the Question.* In complex scenarios, if certain representative requirements are met (e.g., the 20th President of the United States), the LLM can respond directly without confirming other constraints.

Does filtering reward=1 trajectories truly lead to high quality? Intuitively, annotated samples with higher rewards suggest higher quality. Thus, existing methods often filter fully correct samples (reward = 1) from many annotated examples, known as High Score. However, does low reward always indicate low quality? We analyzed the annotated samples selected by different methods, as shown in Table 3. Compared to Random, High Score has lower average interaction turns, which typically indicates that the samples are simpler and easier. Combined with the observation from Table 5 that High Score is least likely to select hard questions, we can conclude that High Score includes more simple samples during filtering. This explains why High Score fails to achieve the best performance despite all data reward=1. Notably, despite the lowest annotated data rewards of GE, it achieves the best performance. This is because more challenging samples can better leverage the advantages of human experience from the guidelines. This suggests that data containing more "attempts at challenging problems" represents higher quality for fine-tuning, even if it is not fully correct.

4.5 Effective Analysis in HotpotQA

To investigate whether EDGE expanded the range of solvable problems, we analyzed the distribution of question difficulty levels in HotpotQA (easy, medium and hard). Table 5 presents the proportion of each difficulty level within the subsets selected by different methods. Among the various approaches, only GE selected a higher proportion of hard questions. As shown in Table 6, GE achieved slight advantages on easy and medium questions, and outperformed the baselines on hard questions. Consequently, GE effectively broadened the scope of solvable problems by focusing on out-of-guideline questions.

5 Conclusion

We propose GE metric, which effectively identifies the most informative samples without relying on golden answers. Selecting samples with low GE scores enhances the efficiency and outcomes of prompt engineering and fine-tuning processes for LLMs. Extensive experiments demonstrate the effectiveness of our method, and we provide a fresh perspective on the data quality of LLM-agent fine-tuning.

Limitations and Future Work. The involvement of human effort is required during guideline updates, which may hinder scalability. In the future, we aim to integrate automatic guideline refinement and explore self-improving agents to reduce human involvement.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant No. 2023YFC3304404.

References

[Azeemi *et al.*, 2025] Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. To label or not to label: Hybrid active learning for neural machine translation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3071–3082, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Bhatt *et al.*, 2024] Gantavya Bhatt, Yifang Chen, Arnav Das, Jifan Zhang, Sang Truong, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Du, Kevin Jamieson, Jordan Ash, and Robert Nowak. An experimental design framework for label-efficient supervised finetuning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6549–6560, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Bukharin *et al.*, 2024] Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3411–3425, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Chen *et al.*, 2023] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning, 2023.

[Chen *et al.*, 2024] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Deng *et al.*, 2023] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[Fu *et al.*, 2024] Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Guan *et al.*, 2024] Jian Guan, Wei Wu, zujie wen, Peng Xu, Hongning Wang, and Minlie Huang. AMOR: A recipe for building adaptable modular knowledge agents through process feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Kremer *et al.*, 2014] Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. Active learning with support vector machines. *Wiley Int. Rev. Data Min. and Knowl. Disc.*, 4(4):313–326, July 2014.

[Kung *et al.*, 2023] Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829, Singapore, December 2023. Association for Computational Linguistics.

[Li *et al.*, 2024] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[Liu *et al.*, 2024a] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Liu *et al.*, 2024b] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024.

[Lu *et al.*, 2024] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[Sener and Savarese, 2018] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[Settles, 2011] Burr Settles. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy, 16 May 2011. PMLR.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[Wang *et al.*, 2023] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Wang *et al.*, 2024] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Yao *et al.*, 2022] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc., 2022.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[Zeng *et al.*, 2024] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3053–3077, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Zhao *et al.*, 2024] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[Zhou *et al.*, 2023] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.