Semi-supervised Fine-tuning for Large Language Models

Junyu Luo$^\heartsuit$, **Xiao Luo**$^{\spadesuit\dagger}$, **Xiusi Chen**$^\diamond$, **Zhiping Xiao**$^{\clubsuit\dagger}$, **Wei Ju**$^\heartsuit$, **Ming Zhang**$^{\heartsuit\dagger}$

$^\heartsuit$ State Key Laboratory for Multimedia Information Processing,
School of Computer Science, PKU-Anker LLM Lab, Peking University
$^\spadesuit$ University of California, Los Angeles $^\diamond$ University of Illinois Urbana-Champaign
$^\clubsuit$ University of Washington
Github Repository: https://github.com/luo-junyu/SemiEvol.

Abstract

Supervised fine-tuning (SFT) is crucial in adapting large language models (LLMs) to a specific domain or task. However, only a limited amount of labeled data is available in practical applications, which poses a severe challenge for SFT in yielding satisfactory results. Therefore, a data-efficient framework that can fully exploit labeled and unlabeled data for LLM fine-tuning is highly anticipated. Towards this end, we introduce a semi-supervised fine-tuning (SemiFT) task and a framework named SEMIEVOL for LLM alignment from a *propagate-and-select* manner. For knowledge propagation, SEMIEVOL adopts a bi-level approach, propagating knowledge from labeled data to unlabeled data through both in-weight and in-context methods. For knowledge selection, SEMIEVOL incorporates a collaborative learning mechanism, selecting higher-quality *pseudo-response* samples. We conducted experiments using GPT-4o-mini and Llama-3.1 on seven general or domain-specific datasets, demonstrating significant improvements in model performance on target data. Furthermore, we compared SEMIEVOL with SFT and self-evolution methods, highlighting its practicality in hybrid data scenarios.

1 Introduction

Supervised fine-tuning (SFT) is a crucial method for enhancing large language models' (LLMs) performance on instructional or domain-specific tasks (Raffel et al., 2020; Chung et al., 2024), playing a vital role in adapting LLMs for specific scenarios. However, SFT relies on a substantial amount of annotated labeled data, which can be increasingly costly in real-world applications (Honovich et al., 2023; Kung et al., 2023). While existing LLMs often employ unsupervised pretraining methods (Devlin, 2018; Radford et al., 2019;
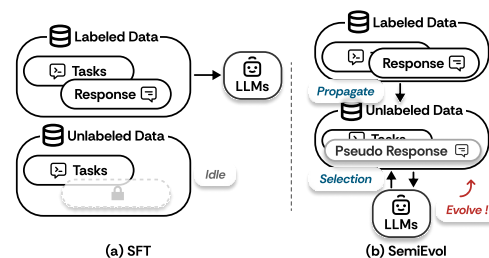

Figure 1: Comparison of SEMIEVOL with previous SFT methods. SEMIEVOL enables interaction between diverse data types for superior performance evolution.

Brown, 2020) to improve their capabilities, this approach typically requires vast datasets and substantial computational resources, making it impractical for scenarios with limited accessible samples.

In practice, however, it often presents a hybrid situation, where a small amount of labeled data coexists with a relatively larger volume of unlabeled data. On the one hand, when deploying LLMs to new target tasks, a limited amount of task-specific annotations can be valuable without incurring excessive costs (Perlitz et al., 2023; Kung et al., 2023). On the other hand, during the continuous inference process of LLMs, a substantial amount of unlabeled data accumulates (Tao et al., 2024; Honovich et al., 2023; Wang et al., 2023). Effectively leveraging the labeled data to enhance model performance on unlabeled data, while simultaneously selecting high-quality unlabeled samples, can improve LLMs' performance in target scenarios, offering substantial practical utility. Therefore, we aim to address the following question:

> *Can LLMs evolve in a real-world scenario of limited labeled data and abundant unlabeled data?*

Designing an evolution framework for hybrid-data scenarios is non-trivial due to the following reasons: First, semi-supervised learning (Kipf and Welling, 2016; Shi et al., 2023), which has been widely studied in machine learning, primarily fo-

† Corresponding authors.

cuses on classification tasks. When considering *generative* tasks, the previous techniques such as pseudo-labeling (Sohn et al., 2020) and contrastive learning (He et al., 2020), cannot be directly applied to LLM use cases, like reasoning and planning (Chen et al., 2022; Hendrycks et al., 2020). Second, previous SFT and unsupervised pretraining methods typically deal with a single type of data (either labeled or unlabeled) (Zhang et al., 2023). Under hybrid-data circumstances, effectively maximizing their combined potential for model improvement becomes challenging.

In this work, we introduce SEMIEVOL for improving LLM reasoning in hybrid-data scenarios, as illustrated in Figure 1. SEMIEVOL employs a bi-level strategy for knowledge *propagation-and-selection*. For knowledge propagation, SEMIEVOL enhances LLMs' inference performance using labeled data through both in-weight and in-context scopes. During in-weight propagation, SEMIEVOL uses labeled data to adapt the model. During in-context propagation, SEMIEVOL employs k-nearest neighbor retrieval in latent space to assist prediction. Moreover, SEMIEVOL introduces a bi-level approach for data selection and generating *pseudo-responses*. First, it introduces a collaborative learning framework, utilizing multiple LLMs with different configurations for inference and self-justification of responses, yielding more accurate predictions. Second, SEMIEVOL adaptively selects unlabeled data by confidence based on response entropy. By mining on unlabeled data leveraging labeled data, we obtain high-quality *pseudo-responses*. Using these *pseudo-response* data, the model enhances its performance on target tasks. We conducted tests on seven general or domain-specific datasets (*e.g.*, MMLU, MMLU-Pro and ConvFinQA), covering tasks such as question-answering, reasoning, and numerical computation. We compared SEMIEVOL with popular methods like retrieval augmented generation, self-evolution and SFT, demonstrating SEMIEVOL's consistent effectiveness across various scenarios.

We summarize the contributions as follows:

- To the best of our knowledge, we are the first to study a practical problem of **semi-supervised fine-tuning (SemiFT)**, aiming to adapt LLMs into different domains data-efficiently.

- We introduce SEMIEVOL, a unified framework for knowledge *propagation-and-selection* that effectively combines labeled and unlabeled data for model evolution.

- We demonstrate the consistent effectiveness of SEMIEVOL across seven widely used general or domain-specific generative tasks in comparison to extensive baseline models.

2 Challenges for Real-world LLM Fine-tuning

2.1 Supervised Fine-tuning

Supervised fine-tuning (SFT) aims to adapt Large Language Models (LLMs) to domain-specific scenarios. Given an LLM $\mathcal{M}$ and a dataset $\mathcal{D}_{\text{labeled}} = \{T_i, Y_i\}_{i=1}^{N}$, where T_i represents the input task or context and Y_i denotes the corresponding expected response. The model minimizes the loss function for each token of the anticipated output during the fine-tuning process FT.

Challenge: Annotation Cost. Despite the effectiveness of supervised fine-tuning, it would require expensive labeling costs to access abundant labeled data. An economic solution is to utilize easily accessible unlabeled data without feedback as a supplement for fine-tuning.

2.2 Background and Problem Definition: Semi-supervised Fine-tuning (SemiFT)

In real-world scenarios, it is more common to have access to both a small amount of labeled data $\mathcal{D}_{\text{labeled}}$ and a larger volume of unlabeled data $\mathcal{D}_{\text{unlabeled}} = \{T_i\}_{i=1}^{M}$. Labeled data offers higher confidence, while unlabeled data represents a broader sample distribution. In this paper, we propose SEMIEVOL approach, which primarily focuses on how to leverage both types of data $\mathcal{D}_{\text{semi}} = \mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}}$ to optimize the LLM $\mathcal{M}$. Our SEMIEVOL not only improves model performance but also offers greater practical value.

Challenge: Generative Task. In fact, developing a semi-supervised fine-tuning framework is highly challenging. Tradition semi-supervised approaches usually focus on classification problems solved by pseudo-labeling while our problem is a generative task, which requires us to generate expected responses instead.

3 Methodology

3.1 Overview

In this paper, we develop SEMIEVOL to integrate labeled and unlabeled data for improving LLM performance in reasoning. The core idea of SEMIEVOL is to leverage labeled data through a
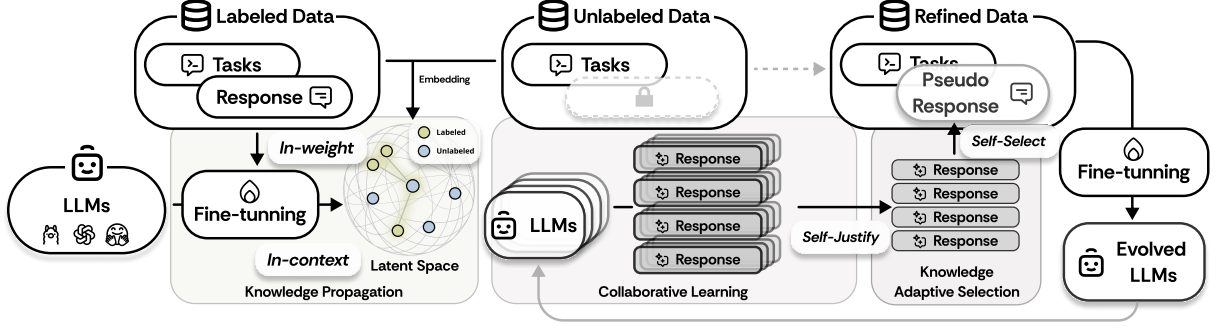
Figure 2: Overview of SEMIEVOL. It maximizes the utility of labeled data through a bi-level knowledge *propagation-and-selection* framework, while leveraging collaborative learning among multiple LLMs to exploit unlabeled data, thereby unleashing the full data potential.

bi-level *propagation-and-select* process. As illustrated in Figure 2, SEMIEVOL is featured by three key components: *(1) Knowledge Propagation*: We utilize labeled data to enhance model $\mathcal{M}$'s performance on unlabeled data. This process focuses on two aspects, *i.e.*, model weights and context. The propagation process involves model adaptation using labeled data and providing the most relevant references from the latent space to assist model inference. *(2) Collaborative Learning*: We employ multiple LLMs with different configurations as mutual teachers to infer unlabeled data. We pay particular attention to inconsistent responses, using the models to self-justify these discrepancies. *(3) Knowledge Self-selection*: We design the adaptive selection for unlabeled data and pseudo-responses. Using labeled data as a guide, we identify the most valuable unlabeled data for learning. By optimizing LLMs on these selected data samples, the model achieves superior evolution performance.

In summary, SEMIEVOL addresses the prevalent real-world scenario where both labeled and unlabeled data coexist. By leveraging the labeled data and the capabilities of LLMs themselves, we perform knowledge propagation, mining, and selection on unlabeled data. This strategy improves model performance in the target scenarios.

3.2 Knowledge Propagation

Labeled data contain expected target responses, while unlabeled data represents a broader task distribution. To leverage this, we aim to propagate knowledge from labeled to unlabeled data, enabling the model to effectively utilize and learn from unlabeled instances. We design a bi-level knowledge propagation framework that operates simultaneously on two fronts: *in-weight* and *in-context*.

For in-weight propagation, we initially warm

up the base model $\mathcal{M}_{base}$ on labeled data $\mathcal{D}_{\text{labeled}}$ to enhance its predictive capabilities for the target task. Specifically, we fine-tune the model, leveraging task data and target responses to obtain a preliminary adapted model ($\mathcal{M}_{warm}$). This process is formulated as:

$$\mathcal{M}_{warm} = FT\left(\mathcal{M}_{base}, \mathcal{D}_{\text{labeled}}\right), \quad (1)$$

where FT is the fine-tuning process.

For in-context propagation, we first embed labeled dataset into latent space using an embedding function $\epsilon(\cdot)$:

$$E_{\text{labeled}} = \{\epsilon\left(t_i\right) \mid \left(t_i, y_i\right) \in \mathcal{D}_{\text{labeled}}\}. \quad (2)$$

During inference on unlabeled data, for each task $t_j \in \mathcal{D}_{\text{unlabeled}}$, we retrieve the k nearest labeled instances in the embedding space:

$$\mathcal{N}\left(t_j\right) = NN\left(E_{\text{labeled}}, \epsilon\left(t_j\right), k\right), \quad (3)$$

where k is set to 3, NN is the nearest neighbors search. We use $\mathcal{N}\left(t_j\right)$ as context to improve the inference on the unlabeled data.

In summary, labeled data facilitates knowledge propagation to unlabeled data through both in-weight and in-context manners. In practice, we first adapt the model to obtain the warm-up LLM $\mathcal{M}_{warm}$, then utilize labeled data as context to enhance inference on unlabeled instances.

3.3 Collaborative Learning

To further exploit unlabeled data, we designed a collaborative learning framework tailored for LLMs. This framework utilizes the inherent capabilities of LLMs for *self-justify* to obtain high-confidence *pseudo-responses* from unlabeled data. Some concurrent works also attempt to use LLMs for similar

functionality (Wang et al., 2024b), while their focus differs from ours.

Initially, we employ a set of n LLMs, denoted as $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_n$ to perform inference on the unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$, where n is 4 by default and will be discussed in Section 4.3.2. Each model is configured with different inference contexts and settings, providing diverse perspectives and yielding more comprehensive results. For each unlabeled sample $t_j \in \mathcal{D}_{\text{unlabeled}}$, we obtain multiple predictions:

$$\{y_j^m\} = \{\mathcal{M}_m(t_j)\}_{m=1}^n . \qquad (4)$$

Subsequently, we implement a *self-justification* process using LLMs. This step synthesizes the inferences from various models to select and summarize the most accurate response $\hat{y}_j$:

$$\tilde{y}_j = \text{Self-Justify}\left(\{y_j^m\}_{m=1}^n\right) . \qquad (5)$$

where the *Self-Justify* operator is implemented via prompting $\mathcal{M}_{warm}$ by natural language instructions. In summary, our LLM-specific collaborative learning framework harnesses multiple differently configured LLMs for multi-perspective inference. By utilizing the LLMs' inherent abilities to *self-justify*, we effectively mine unlabeled data, and generate high-confident pseudo-responses.

3.4 Knowledge Adaptive Selection

While the *pseudo-responses* $\tilde{y}_j$ generated through the collaborative learning framework enrich the training data, they may still contain noise or low-quality information that could misguide the model's learning. To address this issue, we design an adaptive data selection approach within the SEMIEVOL framework. Specifically, we measure the confidence of the responses $\tilde{y}_j$ for the unlabeled data selection.

We use the entropy of the LLM's responses to measure the model's confidence in the answers. Since LLMs generate responses token by token, we calculate the per-token negative log-likelihood, which serves as an approximation of the entropy. For each data sample $t_j \in \mathcal{D}_{\text{unlabeled}}$, the entropy $H(\tilde{y}_j)$ is computed on pseudo-response $\tilde{y}_j$ after Eq. 5 as:

$$H(\tilde{y}_j) = -\frac{1}{L_j} \sum_{k=1}^{L_j} \log P\left(r_j^k \mid t_j, r_j^{<k}\right) , \quad (6)$$

where L_j is the length of the response r_j generated by $\mathcal{M}_{warm}$, r_j^k is the k-th token in the response,

$r_j^{<k} = \left\{r_j^1, r_j^2, \cdots, r_j^k\right\}$ are the preceding tokens of $\tilde{y}_j$, and $P\left(r_j^k \mid t_j, r_j^{<k}\right)$ is $\mathcal{M}_{warm}$'s predicted probability of token r_j^k at position k.

For the unlabeled data, we compute the entropy $H(\tilde{y}_j)$ for each pseudo-response $\tilde{y}_j$ corresponding to task $t_j \in \mathcal{D}_{\text{unlabeled}}$. We then use the θ percentile of the entropy values from the labeled data to establish a dynamic threshold τ:

$$\tau = \text{Percentile}_\theta\left(\{H(\tilde{y}_j)\}_{j=1}^M\right) , \qquad (7)$$

where M is the amount of unlabeled samples, and θ is default to 50% and will be investigated in Section 4.3.2.

Using this dynamic threshold, we select confident samples from the unlabeled data. In formula,

$$\mathcal{D}_{\text{selected}} = \{(t_j, \tilde{y}_j) \mid H(\tilde{y}_j) \leq \tau\} . \qquad (8)$$

We filter the pseudo-responses obtained previously, resulting in the refined dataset $\mathcal{D}_{\text{selected}}$.

Finally, we combine the selected pseudo-labeled data with the original labeled data to fine-tune the base model, which can enhance its performance and adaptability on the target task:

$$\mathcal{M}_{\text{evol}} = FT\left(\mathcal{M}_{\text{base}}, \mathcal{D}_{\text{selected}} \cup \mathcal{D}_{\text{labeled}}\right) , \quad (9)$$

where $\mathcal{M}_{\text{base}}$ is the pre-trained LLM, and FT denotes the fine-tuning process.

By leveraging both high-quality pseudo-labeled data and original labeled data, we enhance the model's performance and adaptability on the target task while reducing the influence of noisy or erroneous information.

3.5 Summary

SEMIEVOL enhances the performance and adaptability of LLMs in target tasks through a two-stage knowledge mining process, combining labeled and unlabeled data for model evolution. Firstly, we leverage a small amount of labeled data to enhance knowledge propagation across unlabeled data. Secondly, we employ knowledge mining and adaptive selection. This strategy effectively integrates both labeled and unlabeled data, culminating in the evolved model $\mathcal{M}_{\text{evol}}$.

4 Experiment

4.1 Experiment Setup

4.1.1 Datasets

We employed both general-purpose and domain-specific evaluation datasets to provide a comprehensive assessment. These datasets encompass a

Model and Strategy	MMLU	MMLU Pro	ARC	FPB	USMLE	PubMedQA	ConvFinQA
GPT-4o-mini *Vanilla*	77.4	57.8	91.5	93.4	73.8	77.5	63.9
GPT-4o-mini *SFT*	77.8	58.8	90.3	98.0	75.0	77.5	88.8
GPT-4o-mini **SEMIEVOL**	**79.9**	**60.8**	**92.7**	**98.9**	**77.2**	**79.5**	**89.2**
Llama3.1-8B *Vanilla*	66.4	47.1	81.1	81.7	70.2	73.5	51.1
Llama3.1-8B *SFT*	67.9	49.8	81.8	96.2	70.8	75.0	81.3
AdaptLLM	–	–	–	49.7	31.5	27.6	30.9
InstructPT	–	–	–	76.1	47.4	44.5	55.2
MemoryLLM	56.4	31.8	56.3	57.7	37.8	55.5	37.2
RAG (BM25)	66.6	37.4	80.8	83.7	69.3	69.0	63.4
RAG (FAISS)	66.5	38.8	81.3	82.5	69.1	71.5	64.6
Hermes-3	63.6	37.9	74.9	73.9	54.5	68.5	54.9
Reflection-Llama	65.5	37.5	82.2	80.8	67.4	77.5	40.8
Llama3.1-8B **SEMIEVOL**	**68.8**	**50.3**	**83.4**	**96.2**	**71.6**	**76.0**	**82.4**

Table 1: **Performance comparison** across different models on various datasets.

variety of tasks, including multiple-choice questions, reasoning, numerical computations, *etc.*. Specifically, our general evaluation datasets include MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024d), and ARC (Clark et al., 2018), while domain-specific datasets comprise FPB (Malo et al., 2014), USMLE (Jin et al., 2021), PubMedQA (Jin et al., 2019), and ConvFinQA (Chen et al., 2022), covering various fields such as finance and healthcare. This diverse selection enables a thorough evaluation of the model's performance across different task types and knowledge domains.

4.1.2 Backbones and Baselines

Base Models. To demonstrate the generalization capability of SEMIEVOL, we employed a diverse range of leading models, encompassing both commercial and open-source and LLMs, including GPT-4o-mini and Llama-3.1-8B (Dubey et al., 2024).

Baselines. We evaluated our method against baselines from several categories: *(1) Vanilla*, which involves testing solely through API calls or using the original model; *(2) Supervised Fine-tuning* (SFT) (Hu et al., 2021; Wei et al., 2021), which adapts the model to the target task using the labeled data; *(3) Self-Evolution Methods* (Self-Evol), which enhance LLM capabilities using additional unlabeled data. We compare with Reflection-

Llama (Li et al., 2024)[1] and Hermes-3 (Teknium et al., 2024)[2], both of which evolve from the Llama-3.1-8B model; *(4) Domain Adaptation Methods*, including AdaptLLM (Cheng et al., 2024b) and InstructPT (Cheng et al., 2024a), utilize domain-specific data (*e.g.*, finance and medical). We select models adapted to corresponding domains for testing, all with comparable parameter counts of 8B; *(5) Inference-time enhancement methods*, such as Retrieval Augmented Generation (RAG) (Lewis et al., 2020), including BM25 (Jones et al., 2000) and FAISS (Douze et al., 2024) algorithms. We also compare with MemoryLLM (Wang et al., 2024c), with the nearest labeled sample as memory;

This comprehensive comparison allows us to assess the effectiveness of our proposed method across various state-of-the-art approaches in LLM fine-tuning and adaptation.

4.1.3 Implementation Details

For the setting of semi-supervised fine-tuning of LLMs, we have $\mathcal{D}_{\text{labeled}}$, $\mathcal{D}_{\text{unlabeled}}$ and $\mathcal{D}_{\text{test}}$. The data proportion in our experiments is $labeled : unlabeled : test = 2 : 6 : 2$ and will be further discussed in Section 4.3.6. The answer information for $\mathcal{D}_{\text{unlabeled}}$ is inaccessible in our setting. We fine-tuned Llama-3.1-8B using Low-Rank Adap-

---
[1]https://huggingface.co/Solshine/reflection-llama-3.1-8B
[2]https://huggingface.co/NousResearch/Hermes-3-Llama-3.1-8B

tation (LoRA) ([Hu et al., 2021](#)) and applied fine-tuning with the official API for GPT-4o-mini[3]. All fine-tuning processes take 2 epochs. n is set to 4 and θ is set to 50%, with further investigation planned in subsequent experiments. Our dataset is publicly available at Hugging Face[4].

We evaluated all methods using the test sets $\mathcal{D}_{\text{test}}$. Model inference followed default settings for each approach. Codes are available in our GitHub repository[5].

4.2 Main Result

We present the main results of SEMIEVOL in Table 1. We can draw the following insights. **Firstly, the tasks are generally challenging.** Off-the-shelf LLMs perform poorly on these tasks, highlighting the necessity of leveraging scenario data to enhance model performance. **Secondly, SEMIEVOL consistently improves both commercial and open-source models.** Notably, SEMIEVOL is one of the few approaches that demonstrably enhances state-of-the-art commercial models, underscoring its practical value. **Thirdly, SFT yield modest improvements,** demonstrating the effectiveness of labeled data. Given the high cost of data labeling, SEMIEVOL effectively utilizes unlabeled data to complement this approach. **Fourthly, the self-evolution method fails to achieve consistent improvements,** showing limited improvement or even adverse effects on most datasets. **Fifthly, adaptive fine-tuning methods can enhance performance only on specific tasks** (*e.g.*, ConvFinQA). Also, these methods may compromise the model's instruction-following ability, leading to significant performance drops in some tasks (*e.g.*, USMLE and PubMedQA). **Lastly, SEMIEVOL consistently outperforms SFT methods**, which demonstrates the effectiveness of incorporating unsupervised data and leveraging labeled data to fully utilize unsupervised data. Even when base models perform poorly (*e.g.*, MMLU-Pro and ConvFinQA), SEMIEVOL can still achieve substantial improvements in model performance.

4.3 Analysis and Discussions

4.3.1 Ablation Study

To evaluate the effectiveness of different components, we conducted an ablation analysis on

---

Variant	MMLU	MMLU-Pro	ARC
Llama3.1-8B SEMIEVOL	68.8	50.3	83.4
w/o IWP	67.8	48.8	82.4
w/o ICP	68.0	49.4	83.0
w/o CL	67.0	49.2	82.4
w/o AS	67.9	49.0	82.1

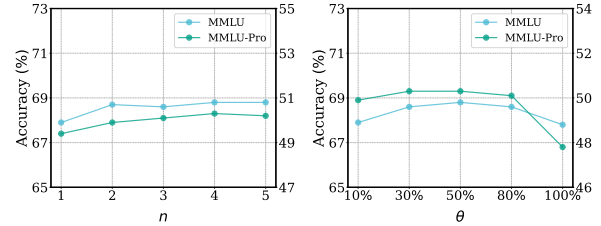Table 2: **Ablation study** via performance comparison of different variants on SEMIEVOL.

Figure 3: **Sensitivity analysis** of SEMIEVOL's performance under different n and θ on variant datasets.

SEMIEVOL, with results presented in Table 2. The findings reveal several key insights: **(1)** The full model consistently outperforms all other configurations across the three datasets, demonstrating its comprehensive effectiveness. **(2)** In terms of knowledge propagation, both In-weight Propagation (IWP) and In-context Propagation (ICP) contribute significantly to the transfer of knowledge from labeled to unlabeled data and subsequent model evolution. In-weight Propagation, in particular, shows a more pronounced impact. **(3)** Removing Collaborative Learning (CL) negatively affects model performance. This suggests that Collaborative Learning effectively leverages predictions from multiple LLMs to autonomously identify more accurate answers, thereby enhancing the prediction quality on unlabeled data. **(4)** The absence of Adaptive Selection (AS) also leads to decreased model performance. This indicates that AS successfully selects more confident samples, thus improving the accuracy of unlabeled data and enhancing the model's evolutionary process.

4.3.2 Sensitivity Analysis

We analyze the number of collaborating models (n) and the data selection ratio (θ), with results illustrated in Figure 3. From the results, we have the following observations. **(1)** Our method demonstrates robust performance across various settings, indicating low sensitivity to these parameters. **(2)** Model accuracy generally increases with n, as
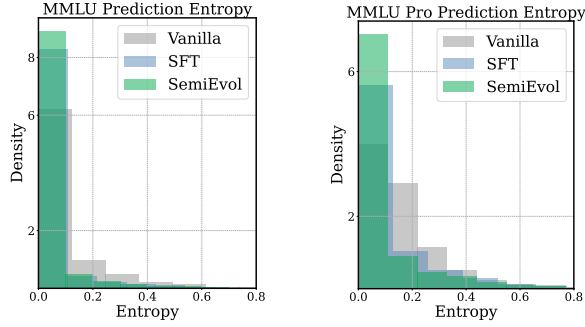
Figure 4: **Entropy distribution** indicates SEMIEVOL can enhanced response confidence. Lower entropy values indicate more confident predictions.
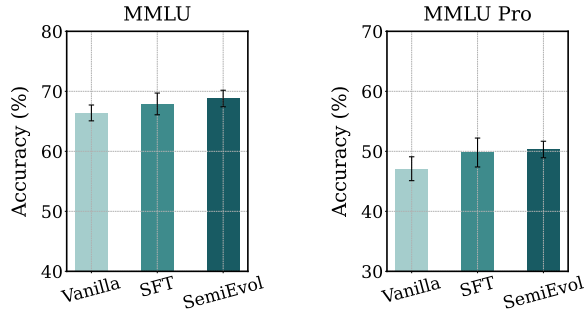

Figure 5: **Stability analysis** via mean performance and standard deviation across multiple inference prompts.

more collaborating LLMs enhance prediction accuracy. However, this also introduces additional computational overhead. We chose $n = 4$ as the default. **(3)** Accuracy initially increases with θ but subsequently decreases, suggesting that introducing excessively noisy data is detrimental to model evolution. Consequently, we empirically set $\theta = 50\%$ as the default value. It is noteworthy that we did not conduct extensive hyperparameter searches, as our primary focus was on validating the overall framework's effectiveness.

4.3.3 Response Entropy Analysis

We present the entropy distribution of different methods on the test set, as illustrated in Figure 4. Lower entropy indicates more confident responses. Compared to the Vanilla and SFT model, SEMIEVOL demonstrates a significant improvement in response confidence. This observation substantiates the effectiveness of SEMIEVOL in producing more decisive and assured outputs. This signifies that SEMIEVOL not only improves accuracy but also enhances the model's ability to generate more confident and reliable responses.
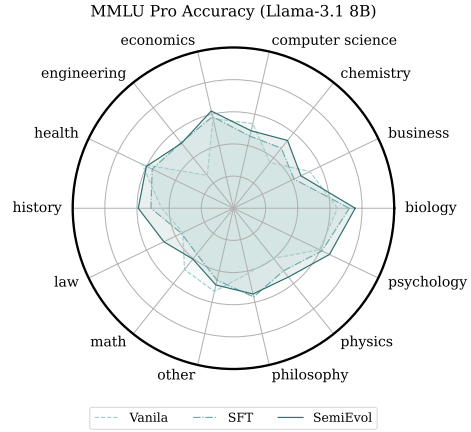

Figure 6: **Category-wise performance** of SEMIEVOL.

4.3.4 Category-wise Performance Analysis

We conducted an in-depth investigation into the differential impact of SEMIEVOL across various categories in MMLU-Pro, as illustrated in Figure 6. We find that **(1)** SEMIEVOL demonstrates enhanced performance across the majority of domains compared to both SFT and Vanilla approaches. This broad-spectrum improvement underscores the method's versatility and effectiveness across diverse subject areas. **(2)** SEMIEVOL achieves substantial gains in specific fields such as Law, Engineering, and Philosophy. This notable improvement suggests that knowledge in these domains is underrepresented in common knowledge bases, highlighting the necessity for targeted adaptation.

4.3.5 Stability Analysis

We evaluate the inference stability of different models by utilizing diverse prompts. Specifically, we employed GPT-4o to rephrase the instructions and conducted 5 tests on each model, reporting the average performance and standard deviation. As illustrated in Figure 5, changing the inference prompts had minimal impact on the various models. Notably, SEMIEVOL even demonstrated a slight improvement in model stability.

4.3.6 Discussion on Continuous Evolution

In real-world scenarios, unlabeled data often accumulates continuously, altering the ratio between labeled and unlabeled data. Table 3 illustrates the impact of various data proportions on SEMIEVOL's performance. As illustrated, model performance consistently improves with an increase in unsupervised data across different base models. This validates SEMIEVOL's effectiveness in addressing real-world scenarios, where model performance in

Base Model	MMLU ($\mathcal{D}_{unlabeled}$ / $\mathcal{D}_{labled}$)				MMLU-Pro ($\mathcal{D}_{unlabeled}$ / $\mathcal{D}_{labled}$)			
	50%	100%	200%	300%	50%	100%	200%	300%
GPT-4o mini	78.2	78.6	79.3	79.9	58.9	59.5	60.1	60.8
Llama3.1-8B	67.6	67.9	68.6	68.8	49.8	49.8	50.0	50.3

Table 3: **Performance of continuous evolution** with varying amounts of unlabeled data.
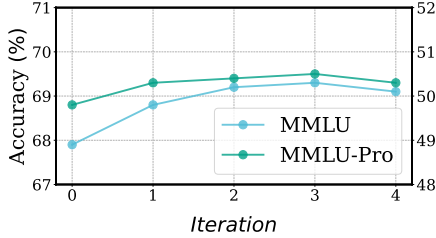

Figure 7: **Iterative evolution performance**, each iteration means perform a round of SEMIEVOL.

specific domains can be progressively enhanced as more unsupervised data accumulates.

4.3.7 Discussion on Iterative Evolution

We verify the model's iterative evolution capability, as illustrated in Figure 7. After applying SEMIEVOL, we utilized the labeled data and pseudo-response data as new labeled data, initiating a fresh round of SEMIEVOL on the previously filtered unlabeled data. By the fourth iteration, we had utilized most of the unlabeled data, resulting in further performance improvements in the target scenario. This iterative evolution capability further demonstrates the practicality of SEMIEVOL.

5 Related Work

5.1 Data Engineering for SFT

With the rapid advancement of Large Language Models (LLMs) (Zhao et al., 2023), researchers have discovered that employing suitable data for Supervised Fine-Tuning (SFT) can enhance model performance on downstream tasks (Taori et al., 2023; Longpre et al., 2023; Hou et al., 2024; Luo et al., 2024b). Some researchers focus on data selection (Bhatt et al., 2024; Parkar et al., 2024; Xia et al., 2024; Bukharin and Zhao, 2023), aiming to improve data quality to boost model effectiveness within limited training budgets. Others concentrate on data synthesis (Mukherjee et al., 2023; Chung et al., 2024; Honovich et al., 2022; Cheng et al., 2023), attempting to enhance models' instruction-following capabilities through synthesized instruction data. Researchers also shifted their focus to

model self-evolution (Tao et al., 2024; Madsen et al., 2024). These include self-instruction (Wang et al., 2022) and self-play (Chen et al., 2024) techniques that enable models to acquire task-specific capabilities without extensive annotated data. Complementary to these approaches, SEMIEVOL focuses on LLMs' ability to continuously evolve in real-world **semi-supervised fine-tuning (SemiFT)** scenarios, relying solely on their inherent capabilities. It effectively utilizes small amounts of labeled data to improve model evolution performance.

5.2 Semi-supervised Learning

Semi-supervised learning aims to reduce the annotation cost during model training (Zhu, 2005; Tarvainen and Valpola, 2017; Ju et al., 2024; Yang et al., 2024; Feng et al., 2024), which has received increasing attention in various fields such as text classification (Duarte and Berton, 2023; Thangaraj and Sivakami, 2018; Linmei et al., 2019) and neural machine translation (Cheng et al., 2016; Pham et al., 2023). Current semi-supervised learning approaches can be mainly divided into two types, *i.e.*, pseudo-labeling (Lee et al., 2013) and consistency regularization (Sohn et al., 2020; Berthelot et al., 2019). Pseudo-labeling approaches usually add extra unlabeled data into the labeled dataset by leveraging the labels predicted by the model. Recent studies attempt different techniques to enhance pseudo-labeling such as considering adaptive thresholds (Zhang et al., 2024; Rhee and Cho, 2019) and class imbalance (Luo et al., 2024a; Wang et al., 2024a). In contrast, consistency regularization aims to encourage the consistency of predictions under different perturbations. However, these approaches focus on classification problems (Shi et al., 2023), which cannot be applied to LLM fine-tuning. To tackle this issue, we propose a new framework SEMIEVOL in a propagate-and-select manner for LLM adaptation.

6 Conclusion

We for the first time investigate the practical challenge of utilizing hybird-data (*i.e.*, both labeled and

unlabeled data) to enhance LLMs performance in specific scenarios. We designed a bi-level framework SEMIEVOL for knowledge *propagation-and-selection*. This framework leverages in-weight and in-context knowledge propagation from labeled data, while employing collaborative learning and adaptive selection to generate high-quality pseudo-responses. We validated SEMIEVOL's efficacy on both general and domain-specific datasets, conducting a detailed analysis of the improvements it yields. Furthermore, we demonstrated SEMIEVOL's capability for continuous iterative evolution, which plays a crucial role in enhancing LLMs' effectiveness in real-world applications.

Limitations

One limitation of our work is that due to the limit of computational resources, we do not evaluate our framework on more LLMs such as GPT-4o and Llama3.1 70B. In future work, we will attempt to incorporate our framework into these LLMs. Moreover, although our framework is evaluated on various benchmark datasets, we do not involve more complicated domains which require more scientific knowledge. To solve this, we will extend our framework to more advanced scientific domains such as genomics analysis.

Acknowledgement

References

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Gantavya Bhatt, Yifang Chen, Arnav M Das, Jifan Zhang, Sang T Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S Du, Kevin Jamieson, et al. 2024. An experimental design framework for label-efficient supervised finetuning of large language models. *arXiv preprint arXiv:2401.06692*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Alexander Bukharin and Tuo Zhao. 2023. Data diversity matters for robust instruction tuning. *arXiv preprint arXiv:2311.14736*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024a. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024b. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

José Marcio Duarte and Lilian Berton. 2023. A review of semi-supervised learning for text classification. *Artificial intelligence review*, 56(9):9401–9469.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srbuhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. 2024. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14409–14428.

Zhichao Hou, Weizhi Gao, Yuchen Shen, Feiyi Wang, and Xiaorui Liu. 2024. Protransformer: Robustify transformers via plug-and-play paradigm. *arXiv preprint arXiv:2410.23182*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

Wei Ju, Siyu Yi, Yifan Wang, Qingqing Long, Junyu Luo, Zhiping Xiao, and Ming Zhang. 2024. A survey of data-efficient graph learning. *arXiv preprint arXiv:2402.00447*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2, page 896. Atlanta.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024. Selective reflection-tuning: Student-selected data recycling for LLM instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16189–16211, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4821–4830.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Junyu Luo, Yiyang Gu, Xiao Luo, Wei Ju, Zhiping Xiao, Yusheng Zhao, Jingyang Yuan, and Ming Zhang. 2024a. Gala: Graph diffusion-based alignment with jigsaw for source-free domain adaptation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1–14.

Junyu Luo, Xiao Luo, Kaize Ding, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. 2024b. Robustft: Robust supervised fine-tuning for large language models under noisy response. *arXiv preprint arXiv:2412.14922*.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 295–337.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Ritik Sachin Parkar, Jaehyung Kim, Jong Inn Park, and Dongyeop Kang. 2024. Selectllm: Can llms select important instructions to annotate? *arXiv preprint arXiv:2401.16553*.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein Dor. 2023. Active learning for natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877.

Viet H Pham, Thang M Pham, Giang Nguyen, Long Nguyen, and Dien Dinh. 2023. Semi-supervised neural machine translation with consistency regularization for low-resource languages. *arXiv preprint arXiv:2304.00557*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hochang Rhee and Nam Ik Cho. 2019. Efficient and robust pseudo-labeling for unsupervised domain adaptation. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 980–985. IEEE.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking semi-supervised learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5614–5634.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report.

Muthuraman Thangaraj and Muthusamy Sivakami. 2018. Text classification techniques: A literature review. *Interdisciplinary journal of information, knowledge, and management*, 13:117.

Pengyun Wang, Yadi Cao, Chris Russell, Siyu Heng, Junyu Luo, Yanxin Shen, and Xiao Luo. 2024a. Delta: Dual consistency delving with topological uncertainty for active graph domain adaptation. *arXiv preprint arXiv:2409.08946*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024b. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. 2024c. Memoryllm: Towards self-updatable large language models.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024d. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Junwei Yang, Hanwen Xu, Srbuhi Mirzoyan, Tong Chen, Zixuan Liu, Zequn Liu, Wei Ju, Luchen Liu, Zhiping Xiao, Ming Zhang, et al. 2024. Poisoning medical knowledge using large language models. *Nature Machine Intelligence*, 6(10):1156–1168.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xuerong Zhang, Li Huang, Jing Lv, and Ming Yang. 2024. Self adaptive threshold pseudo-labeling and unreliable sample contrastive loss for semi-supervised image classification. In *International Conference on Artificial Neural Networks*, pages 61–75. Springer.

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

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey.

A   Algorithm

In Algorithm 1, we present the comprehensive algorithmic process of SEMIEVOL, which incorporates a bi-level framework for knowledge propagation and selection. This process ultimately yields the evolved model, $\mathcal{M}_{evol}$.

Algorithm 1 Algorithm of SEMIEVOL

Require: Labeled data $\mathcal{D}_{labeled}$, Unlabeled data $\mathcal{D}_{unlabeled}$, LLM $\mathcal{M}$;
Ensure: Evolved LLM $\mathcal{M}_{evol}$;

1: *// In-Weight Knowledge Propagation*
2: Fine-tune $\mathcal{M}$ on $\mathcal{D}_{labeled}$, obtain $\mathcal{M}_{warm}$;
3: *// Collaborative Learning*
4: **for** m = 1, $\cdots$, n **do**
5:    *// In-Context Propagation*
6:    Get the prediction $\left\{ y_j^m \right\}$ as Eq. 4;
7: **end for**
8: Self-Justify to generate pseudo-responses;
9: *// Adaptive Selection*
10: Select pseudo-responses with entropy as Eq. 8, obtain $\mathcal{D}_{selected}$;
11: Fine-tune on $\mathcal{D}_{selected}$, obtain $\mathcal{M}_{evol}$;

B   Experimental Settings

We will detail the experimental process, including parameter settings, prompt configurations, and computational resource consumption.

B.1   Parameter Settings

Inference Process For commercial methods (*e.g.*, GPT4o-mini), we utilize API calls. For open-source models (*e.g.*, Llama3.1 8B), we employ the vLLM (Kwon et al., 2023) framework locally for inference. The inference parameters are as Table 4.

Parameter	Value
Model Name	Llama3.1 8B Instruct
Computing Arch.	NVIDIA A100-80GB
Max Sequence Length	4096

Table 4: Parameters configuration during inference.

In the collaborative learning process, we employ n LLMs with diverse configurations for mutual learning, setting their temperature to 1. These configurations are sampled from the following options: (1) Utilization of $\mathcal{M}_warm$: $\{0, 1\}$; (2) Number of

labeled samples referenced for in-context knowledge propagation: $\{0, 1, 2, 3\}$.

Fine-tuning Process. We use platform APIs for commercial methods fine-tuning and LlamaFactory (Zheng et al., 2024) for open-source models fine-tuning. Both of them are fine-tuned for 2 epochs. Commercial models use default settings (adaptively configured by OpenAI based on the task). For open-source models, our hyperparameter settings are as Table 5, all of which are the default parameters from LlamaFactory.

Parameter	Value
Model Name	Llama3.1 8B Instruct
Computing Arch.	NVIDIA A100-80GB
Fine-tuning Type	LoRA
Gradient Accumulation Steps	8
Learning Rate	1e-4
θ in Adaptive Selection	50%

Table 5: Hyperparameter settings for fine-tuning open-source models.

B.2 Instruction Settings

We employ concise instructions for inference as shown in Table 6. During this process, we present questions to the LLM and elicit responses.

We extract answers using regular expression matching. For character-based answers, we check for exact matches. For numerical answers, we assess whether they fall within an acceptable error margin (maximum error of 1e-2).

During the SEMIEVOL process, we require additional instructions for tasks such as self-justification and in-context knowledge propagation. For these supplementary commands, we provide instruction templates in a table for reference.

Instruction Template

Multiple Choice
Answer the multiple-choice question.
Your response should be in the format: 'Answer: LETTER' (without quotes).

Question:
{question}

Options:
{options}

Free-from QA
Answer the following question.
Output the value in the format: 'Answer: VALUE' (without quotes).

Question:
{question}

Options:
{options}

Provide your answer on a new line after 'Answer:', without using a \boxed command.

Table 6: Instruction templates for different types of questions.

Instruction Template

Self-Justify
Here are the multiple answers to the question.
Please consider them thoroughly and give me
the correct answer. Your response should be in
the following format:
'Answer: LETTER' (without quotes).

Question:
{question}

Options: {options}

Multiple Answers:
{answers}

Now, please give me the final correct answer:

In-context Knowledge Propagation
You are an expert in the question answering.
Below are some examples of questions and their
corresponding answer.

{reference}

Table 7: Instruction templates for SEMIEVOL process.