

Small Models are Valuable Plug-ins for Large Language Models (/pdf?id=SMv1SmITF8G)

Anonymous

16 Dec 2023 ACL ARR 2023 December Blind Submission Readers: December, Paper541

Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers, Paper541 Authors Show

Revisions (/revisions?id=SMv1SmITF8G)

Abstract: Large language models (LLMs) such as GPT-3 and GPT-4 are powerful but their weights are often publicly unavailable and their immense sizes make the models difficult to be tuned with common hardware. As a result, effectively tuning these models with large-scale supervised data can be challenging. As an alternative, In-Context Learning (ICL) can only use a small number of supervised examples due to context length limits. In this paper, we propose Super In-Context Learning (SuperICL) which allows black-box LLMs to work with locally fine-tuned smaller models, resulting in superior performance on supervised tasks. Our experiments demonstrate that SuperICL can improve performance beyond state-of-the-art fine-tuned models while addressing the instability problem of in-context learning.

Paper Type: long

Research Area: NLP Applications


Contribution Types: Model analysis & interpretability, NLP engineering experiment

Languages Studied: EN, AR, BG, DE, EL, ES, FR, HI, RU, SW, TH, TR, UR, VI, ZH


Revealed to Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, Julian McAuley

14 Dec 2023 ACL ARR 2023 December Submission

Authors: *Canwen Xu (/profile?id=~Canwen_Xu1), Yichong Xu (/profile?id=~Yichong_Xu1), Shuohang Wang (/profile?id=~Shuohang_Wang1), Yang Liu (/profile?id=~Yang_Liu50), Chenguang Zhu (/profile?id=~Chenguang_Zhu1), Julian McAuley (/profile?id=~Julian_McAuley1)*

Responsible NLP Research:  pdf (/attachment?id=ycDNvCaKARi&name=responsible_NLP_research)

Previous URL: /forum?id=b-8ijb9EE0X (/forum?id=b-8ijb9EE0X)

Previous PDF:  pdf (/attachment?id=ycDNvCaKARi&name=previous_PDF)

Reassignment Request Action Editor: Yes, I want a different action editor for our submission

Reassignment Request Reviewers: Yes, I want a different set of reviewers

Justification For Not Keeping Action Editor Or Reviewers: The reviewers fail to understand the importance of our work in LLM planning.

Preprint: no

Existing Preprints: <https://arxiv.org/abs/2305.08848> (<https://arxiv.org/abs/2305.08848>)

Preferred Venue: NAACL 2024

Consent To Share Data: no

Consent To Review: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

Add

Author-Editors Confidential Comment

Withdraw

Reply Type: Author:

9 Replies

Visible To: Hidden From:

[−] Meta Review of Paper541 by Area Chair 4LJ3

ACL ARR 2023 December Paper541 Area Chair 4LJ3

01 Feb 2024, 13:58 ACL ARR 2023 December Paper541 Meta Review Readers:

Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Authors, Paper541 Reviewers

Submitted, Program Chairs [Show Revisions \(/revisions?id=ji4LpP7HmA\)](#)

Metareview:

The paper presents a novel approach, Super In-Context Learning (SuperICL), to overcome the challenges associated with fine-tuning large language models (LLMs) such as GPT-3 and GPT-4. The authors propose a method that combines the capabilities of inaccessible, large-scale LLMs with locally fine-tuned smaller models, enhancing performance on supervised learning tasks. The paper is well-structured and clear, making its content accessible and valuable to the academic community. The authors have conducted extensive experiments to validate their approach, demonstrating that SuperICL surpasses the performance of fine-tuned models in terms of accuracy and instability. The paper would fit well in sessions related to language model fine-tuning, in-context learning, and the application of small models in large language model frameworks.

Summary Of Reasons To Publish:

The paper presents a novel and significant contribution to the field of natural language processing, specifically in the area of large language model fine-tuning. The authors propose a unique approach, SuperICL, that combines the capabilities of large and small models, addressing the challenges associated with fine-tuning large language models. The paper is well-structured and clear, making its content accessible and valuable to the academic community. The authors have conducted extensive experiments to validate their approach, demonstrating that SuperICL surpasses the performance of fine-tuned models in terms of accuracy and instability. The paper's findings could be of interest to both broad and narrow sub-communities within the field of natural language processing.

Summary Of Suggested Revisions:

The reviewers have suggested several revisions that could improve the paper. Firstly, the authors should provide a more compelling motivation for their work, possibly focusing on the high cost of fine-tuning large language models. Secondly, the authors should consider evaluating their approach on more recent benchmarks, such as MMLU and GSM8K. Thirdly, the authors should include fine-tuned, moderately-sized LLMs, such as Llama-7b and Mistral-7b, as baselines for comparison. Additionally, the authors should explore the impact of the plug-in model's effectiveness on the performance of LLMs and investigate why the inclusion of additional demonstration examples does not enhance generative performance. Lastly, the authors should consider comparing their work with similar existing works and provide a deeper insight into in-context learning.

Overall Assessment: 3 = There are major points that may be revised

Best Paper Ae: No

Ethical Concerns:

None

Needs Ethics Review: No

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add

[−] Official Review of Paper541 by Reviewer 7oAL

ACL ARR 2023 December Paper541 Reviewer 7oAL

16 Jan 2024, 19:42 (modified: 27 Jan 2024, 03:48) ACL ARR 2023 December Paper541

Official Review Readers: Program Chairs, Paper541 Senior Area Chairs, Paper541 Area

Paper Summary:

This paper introduces Super In-Context Learning (SuperICL), designed to overcome the challenges associated with fine-tuning large language models (LLMs) such as GPT-3 and GPT-4. SuperICL combines the capabilities of inaccessible, large-scale LLMs with locally fine-tuned smaller models, enhancing performance on supervised learning tasks. The experimental results demonstrate that SuperICL surpasses the performance fine-tuned models in terms of accuracy and instability

Summary Of Strengths:

1. The collaborative use of small and large models, balancing computational efficiency and performance, is a strategic and commendable approach.
2. Investigating the influence of small model predictions on large model generativity is a novel and significant contribution, previously underexplored in the field.
3. The paper is clear and well-structured, making its content accessible and valuable to the academic community.

Summary Of Weaknesses:

1. The results indicate that the performance of LLMs is significantly dependent on the plug-in model's effectiveness. Further experiments using a weaker model might be necessary to explore this aspect comprehensively. Moreover, the improvements that SuperICL provides over Fine-Tuned-RoBERTa are relatively minor.
2. The experimental data indicate that adding more supervised examples (n-shots) does not necessarily improve performance.
3. A lack of testing on generative tasks, such as multiple-choice questions and completion-style question answering. This omission leaves a gap in understanding the full scope of SuperICL's applicability and effectiveness in diverse task formats.

Comments, Suggestions And Typos:

1. The proposed solution presumes complete access to the training set. However, it's known that the performance of ICL is highly contingent on the choice of exemplars. Therefore, it is curious why the selection of these examples is random. It seems that incorporating the plug-in model in the selection process could potentially enhance this aspect.
2. Regarding the ablation study in Table 5, the significant impact of the Ref component on final performance is noted. It would be insightful to explore how the overall performance of SuperICL is affected when the plug-in model's accuracy is suboptimal. Understanding this relationship could provide valuable insights into the robustness of SuperICL under varying conditions of plug-in model performance.
3. An investigation into why the inclusion of additional demonstration examples does not enhance generative performance would be beneficial. This inquiry should also revisit the issue of whether the plug-in model's output predominantly influences the final prediction in SuperICL. Clarifying this aspect could lead to a deeper understanding of the interaction between the plug-in model and the large language model within the SuperICL framework.
4. Figure 2 falls short in clearly illustrating the relationship between overridden behavior and the confidence scores from the plug-in model. A comparative analysis examining the differences in the distribution of overridden instances in SuperICL, with and without confidence scores, is needed. Additionally, if there is a strong correlation between overridden behavior and plug-in model confidence, it may be worth considering allowing LLM to predict primarily in cases where the plug-in model exhibits low confidence. This approach could optimize the synergy between the two models in the proposed SuperICL.

Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Confidence: 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.

Best Paper: No

Ethical Concerns:

None

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add **Author-Editors Confidential Comment**

[-] Response to Reviewer 7oAL

ACL ARR 2023 December Paper541 Authors Canwen Xu (/profile?id=~Canwen_Xu1) (privately revealed to you)

25 Jan 2024, 20:09 ACL ARR 2023 December Paper541 Official

Comment Readers: Program Chairs, Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers Submitted, Paper541 Authors Show Revisions (/revisions?id=sDWGwj-EXG)

Comment:

We would like to thank the reviewer for your insightful comments.

The proposed solution presumes complete access to the training set. However, it's known that the performance of ICL is highly contingent on the choice of exemplars. Therefore, it is curious why the selection of these examples is random. It seems that incorporating the plug-in model in the selection process could potentially enhance this aspect.

Unlike ICL, we found that the performance of SuperICL is not sensitive to the exemplars as shown in Table 7. We also experimented with example-retrieval approach and the improvement is marginal.

Regarding the ablation study in Table 5, the significant impact of the Ref component on final performance is noted. It would be insightful to explore how the overall performance of SuperICL is affected when the plug-in model's accuracy is suboptimal. Understanding this relationship could provide valuable insights into the robustness of SuperICL under varying conditions of plug-in model performance.

This is a great point. In fact, we do have this analysis in Appendix B, where the reference (i.e., the plug-in model's output) is compromised by adversarial attack. This has a direct impact on the final performance although the large models can correct some predictions successfully.

An investigation into why the inclusion of additional demonstration examples does not enhance generative performance would be beneficial.

In Figure 3, we observe a diminishing effect on MNLI and SST-2 when we add more demonstration examples to the context. This is consistent with previous findings on ICL. An interesting observation is if we do not have enough examples to demonstrate all available classes, for example, in MNLI, when we have 1 or 2 examples, the performance is significantly lower.

A comparative analysis examining the differences in the distribution of overridden instances in SuperICL, with and without confidence scores, is needed. Additionally, if there is a strong correlation between overridden behavior and plug-in model confidence, it may be worth considering allowing LLM to predict primarily in cases where the plug-in model exhibits low confidence.

Thanks for the insightful suggestion. We will add this in the revision.

Add **Author-Editors Confidential Comment**

[-] Response to authors' rebuttal

ACL ARR 2023 December Paper541 Reviewer 7oAL

27 Jan 2024, 03:47 ACL ARR 2023 December Paper541 Official
Comment Readers: Program Chairs, Paper541 Senior Area Chairs, Paper541
Area Chairs, Paper541 Reviewers Submitted, Paper541 Authors Show
Revisions (/revisions?id=MV3gvXKtNk)

Comment:

Thank you for your response. My primary concern lies in how the final performance is significantly influenced by the plug-in model's performance. This might explain why varying the context examples doesn't notably impact the results. The approach's integration of predictions and confidences directly into the prompt format is a reasonable for the reported results, as these elements are likely to be treated as crucial keywords by LM. (please correct me if I am wrong).

However, I am still interested in observing how these large models perform when the plug-in model's accuracy is low. This could be a crucial test to see if the larger models can effectively correct errors made by the plug-in model.

In light of some of the responses to my concerns, I am inclined to revise my evaluation score.

Add

Author-Editors Confidential Comment

[-] Official Review of Paper541 by Reviewer Uoju

ACL ARR 2023 December Paper541 Reviewer Uoju

15 Jan 2024, 18:06 ACL ARR 2023 December Paper541 Official Review Readers:
Program Chairs, Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers
Submitted, Paper541 Authors Show Revisions (/revisions?id=-VbKRVDADI)

Paper Summary:

This paper proposed SuperICL which enables the usage of a large number of supervised examples by integrating finetuned small models with large language models. Specifically, the in-context samples are fed to the large language models after they are concatenated with the ground-truth label, predicted label and confidence of the fine-tuned small language model. Then, the test sample is fed to the large language model after it is concatenated with the prediction and confidence of the fine-tuned small language model. Experiments in the paper validates that high performance of small language models can help in-context learning of large language model.

Summary Of Strengths:

- The main strength of this paper is that their proposed method is validated through extensive experiments such as different small/large language models.

Summary Of Weaknesses:

- Though this paper claims that their proposed SuperICL is a new paradigm for utilizing large language models(79-80), there are lines of works aiming at similar goals and this paper lacks the comparison to such works. For example, the most similar one is [1] which integrates fine-tuned small language models into the black-box large language model. Also, many works aim to adapt large language models for large supervised datasets such as adapter.
- The proposed method of concatenating the test input with the plug-in model's prediction attached is straightforward and lacks deep insights into in-context learning.

[1] Welleck, Sean, et al. "Generating Sequences by Learning to Self-Correct." The Eleventh International Conference on Learning Representations. 2022.

Comments, Suggestions And Typos:

I hope Section 4. and Appendix to be more organized. Though the experiments in Appendix section are interesting, they are not fully explained in the context of the whole paper.

Soundness: 3.5

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:

None

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Author Identity Guess: 4 = From an allowed pre-existing preprint or workshop paper, I know/can guess at least one author's name.

Add **Author-Editors Confidential Comment**

[−] Response to Reviewer UoJu

ACL ARR 2023 December Paper541 Authors Canwen Xu (/profile?id=~Canwen_Xu1) (privately revealed to you)

25 Jan 2024, 20:08 ACL ARR 2023 December Paper541 Official

Comment Readers: Program Chairs, Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers Submitted, Paper541 Authors Show Revisions (/revisions?id=BgLtGHGO9tj)

Comment:

Thank you for your insightful comments.

I hope Section 4. and Appendix to be more organized. Though the experiments in Appendix section are interesting, they are not fully explained in the context of the whole paper.

Thanks for the advice. Since we have conducted many experiments, we had to make hard decisions about what to put in the main content and what goes into the appendix. We will review the paper to better incorporate the appendix into the whole paper.

Add **Author-Editors Confidential Comment**

[−] Official Review of Paper541 by Reviewer HQVy

ACL ARR 2023 December Paper541 Reviewer HQVy

13 Jan 2024, 22:52 ACL ARR 2023 December Paper541 Official Review Readers:

Program Chairs, Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers Submitted, Paper541 Authors Show Revisions (/revisions?id=c0RcBjId6T)

Paper Summary:

The authors propose Super In-Context Learning (SuperICL), a method that utilizes smaller models as plug-ins. These plug-ins provide predictions with confidence scores, which are then concatenated with the input text and ground-truth labels as context. Extensive experiments conducted by the authors demonstrate that SuperICL significantly improves performance on supervised tasks and addresses the instability of In-Context Learning (ICL).

Summary Of Strengths:

1. The paper is well-written and easy to follow.
2. The authors conduct extensive experiments to support their findings.

Summary Of Weaknesses:

1. The motivation for this work is not clearly articulated. A more compelling motivation might be the high cost of fine-tuning large language models, suggesting the use of smaller models to assist in prediction, rather than the

inaccessibility of large language models (LLMs).

2. The evaluation benchmarks used in this study are outdated. Evaluating the approach on more recent benchmarks, such as MMLU and GSM8K, would be beneficial.
3. There is a lack of baseline comparisons. In addition to in-context learning methods, it would be useful if the authors included fine-tuned, moderately-sized LLMs, such as Llama-7b and Mistral-7b, as baselines.

Comments, Suggestions And Typos:

1. The content between lines 020 and 046 requires comprehensive revision.

Soundness: 3.5

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:

None

Needs Ethics Review: No

Reproducibility: 5 = They could easily reproduce the results.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Author Identity Guess: 2 = From social media/a talk/other informal communication, I know/can guess at least one author's name.

Add **Author-Editors Confidential Comment**

[-] Response to Reviewer HQVy

ACL ARR 2023 December Paper541 Authors [Canwen Xu \(/profile?id=~Canwen_Xu1\)](/profile?id=~Canwen_Xu1) (privately revealed to you)

25 Jan 2024, 20:08 [ACL ARR 2023 December Paper541 Official](#)

Comment Readers: Program Chairs, Paper541 Senior Area Chairs, Paper541 Area Chairs, Paper541 Reviewers Submitted, Paper541 Authors [Show Revisions \(/revisions?id=YxxJHpK-OSI\)](#)

Comment:

We would like to thank you for your insights.

When we first kicked-off this project, the LLaMA family did not exist and our main focus was to allow light-weight training for black-box LLMs. Indeed, as our field has been changed drastically, we will adopt the motivation you suggested in the revision. We are working on adding more recent benchmarks and baselines as you suggested.

Add **Author-Editors Confidential Comment**

[-] Supplementary Materials by Program Chairs

ACL ARR 2023 December Program Chairs

16 Dec 2023, 06:06 [ACL ARR 2023 December Paper541 Supplementary](#)

Materials Readers: Program Chairs, Paper541 Reviewers, Paper541 Authors, Paper541 Area Chairs, Paper541 Senior Area Chairs [Show Revisions \(/revisions?id=2CH7GnfDTf3\)](#)

Responsible NLP Research: [pdf \(/attachment?id=2CH7GnfDTf3&name=responsible_NLP_research\)](#)

Previous URL: </forum?id=b-8jlb9EE0X> (</forum?id=b-8jlb9EE0X>)

Previous PDF: [pdf \(/attachment?id=2CH7GnfDTf3&name=previous_PDF\)](#)

Reassignment Request Action Editor: Yes, I want a different action editor for our submission

Reassignment Request Reviewers: Yes, I want a different set of reviewers

Justification For Not Keeping Action Editor Or Reviewers: The reviewers fail to understand the importance of our work in LLM planning.

Note From EiCs: These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.

Add

Author-Editors Confidential Comment

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[Frequently Asked Questions](#)

[\(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

[Feedback](#)

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2024 OpenReview