Improving In-Context Learning with Small Language Model Ensembles

M. Mehdi Mojarradi, Lingyi Yang, Robert McCraith, Adam Mahdi
Correspondence email: m.mehdi.mojarradi@gmail.com

In-Context Learning

In-context learning (ICL) is an effective method for adapting large language models (LLMs) to perform specific tasks *without the need of updating model parameters through fine-tuning*. It involves prompting an LLM with *few-shot training examples* and a test input, allowing the LLM to infer the correct output from the provided context. While ICL is *time and cost-efficient*, it *lacks the accuracy* when compared with fine-tuning and retrieval-augmented generation.

Last year, Xu et. al. and Yang et. al. proposed SuperICL [1] and SuperContext [2], which **enrich the in-context demonstrations with the predictions of a small language model (SLM) and its confidence scores**.

Ensemble Super In-Context Learning

We propose *Ensemble SuperICL*, enabling an LLM to utilise several SLMs through ICL. The stages of this method (illustrated in Figure 1) are:

1. In-context examples are sampled from a dataset, where each example is a pair of input and true label.
2. Two or more fine-tuned SLMs produce *ensemble super context*: one demonstration in Ensemble SuperICL consists of an input, the predicted labels and confidence scores of two to five SLMs on this in-context example, and the true label.
3. A test question is concatenated with its predicted labels and confidence scores from SLMs and fed to the LLM.

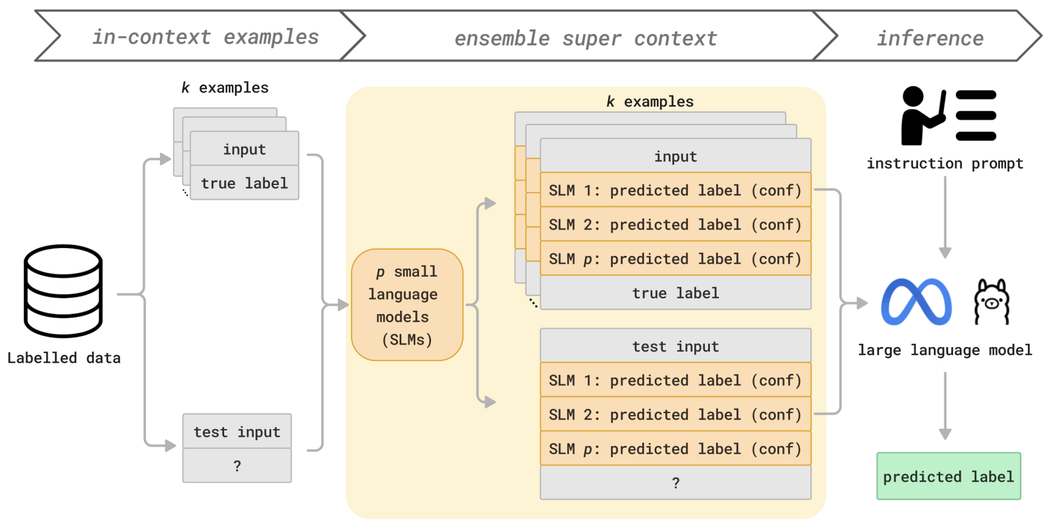

Figure 1 – Stages of ensemble super-ICL.

Figure 2 gives an example of Ensemble SuperICL applied to the SST-2 dataset.

a. choose *k* demonstrations	**input:** sam mendes has become valedictorian at the school for soft landings and easy ways out. **true label:** negative ... *k*
b. construct ensemble super context	**input:** sam mendes has become valedictorian at the school for soft landings and easy ways out. **BART-Large Prediction:** positive (Confidence: 0.93) **ELECTRA-Large Prediction:** negative (Confidence: 0.88) **true label:** negative ... *k*
c. construct test input	**input:** unflinchingly bleak and desperate **BART-Large Prediction:** negative (Confidence: 0.98) **ELECTRA-Large Prediction:** negative (Confidence: 1.0) **label:** ?
d. LLM is instructed to use ensemble super context and test input to predict the correct label	**instruction:** You are tasked with predicting the sentiment of a given sentence (positive or negative). BART-Large, ELECTRA-Large are language models fine-tuned on this task, and you may use their output as an aid to your judgement. Fill in your answer after 'Label: ' at the end of the prompt. *+ ensemble super context + test input*

Figure 2 – Example of Ensemble SuperICL procedure on the SST-2 dataset.

Data

We used five datasets for our experiments: four natural language understanding (NLU) benchmarks and one domain-specific dataset.

The General Language Understanding Evaluation (GLUE) benchmark. We used four of the eleven GLUE datasets to evaluate a range of NLU abilities: the Multi-Genre Natural Language Inference corpus (MNLI), the Stanford Sentiment Treebank (SST-2), the Microsoft Research Paraphrase Corpus (MRPC), and the Corpus of Linguistic Acceptability (CoLA)

The Medical Multiple-Choice Question Answering dataset (MedMCQA). This contains over 183k medical entrance exam questions. Each question is assigned one of 21 medical subjects such as surgery, dental, and pathology. We task our models with inferring the subject of a given question.

Small language models

We used *Llama3-8b-Instruct as the LLM*, and considered seven SLMs detailed in the table below along with their accuracy on the five datasets. For the GLUE datasets, a fine-tuned version of the SLM was used. For MedMCQA, all SLMs used were fine-tuned on MNLI. Dashes indicate where fine-tuned SLMs were unavailable for a dataset or not considered.

SLM	Size	SST-2	MRPC	MNLI	CoLA	MedMCQA
MobileBERT	25M	-	-	-	52.78	-
flan-t5-base	248M	-	-	88.68	-	70.43
ELECTRA-large	335M	96.56	89.95	90.28	67.43	29.86
DeBERTa-large	350M	94.95	89.71	90.39	64.06	71.43
RoBERTa-large	356M	96.44	89.71	88.68	65.65	61.57
BART-large	407M	95.30	87.50	88.85	-	68.71
T5-large	770M	-	-	-	53.51	-

Results and Discussion

We present accuracies of the best performing versions of Ensemble SuperICL (with 2, 3, 4, and 5 SLMs) and show that:

- **Ensemble SuperICL (E-SuperICL) outperforms all baselines on three natural language understanding benchmarks and a domain-specific labelling task (MedMCQA).**
- **Ensemble SuperICL boosted ICL performance by 3 to 20 percentage points, with greater gains on more challenging tasks.**

	SST-2	MRPC	MNLI	CoLA	MedMCQA
ICL (Llama3-8b-Instruct)	94.15	75.25	71.24	55.43	79.43
SuperICL	96.56	89.22	87.45	67.21	82.71
SLM Majority vote	96.67	90.69	**91.39**	65.64	68.14
E-SuperICL 2	**97.13**	90.69	88.41	**70.36**	**84.29**
E-SuperICL 3	97.02	**91.42**	90.25	**70.36**	83.71
E-SuperICL 4	96.79	**91.42**	90.47	70.32	82.43
E-SuperICL 5	-	-	91.27	68.17	80.57

We investigated the effects of removing three components:

- (a) the SLM predictions for the in-context examples (Ctx);
- (b) the confidence scores of the SLMs in both the in-context examples and test input (Con);
- (c) the SLM predictions for the test input (Tst).

We see that **all components are essential for optimal performance** – removing of the SLM predictions for in-context examples has the greatest impact.

	Components							
	(a) Ctx	(b) Con	(c) Tst	SST-2	MRPC	MNLI	CoLA	Med
(1)	×	×	✓	96.90	89.71	84.76	67.94	78.57
(2)	✓	×	✓	96.79	89.71	89.41	67.25	83.00
(3)	×	✓	✓	97.02	**91.42**	86.94	69.85	78.86
(4)	✓	✓	✓	**97.13**	**91.42**	**91.14**	**70.36**	**84.29**

Summary

Ensemble SuperICL is an ICL method that enables an LLM to leverage the predictions and confidence of off-the-shelf SLMs. This **improves classification accuracy while preserving low time, compute, and data requirements**.

We show that:

1. Ensemble SuperICL outperforms ICL, SLM, and SuperICL baselines on natural language understanding tasks;
2. Knowledge can be transferred as it labels large-scale domain-specific data more accurately than all baselines.

Check out our paper!

References

[1] C. Xu, Y. Xu, S. Wang, Y. Liu, C. Zhu, and J. McAuley, "Small models are valuable plug-ins for large language models," *arXiv preprint arXiv:2305.08848*, 2023.
[2] L. Yang, S. Zhang, Z. Yu, G. Bao, Y. Wang, J. Wang, R. Xu, W. Ye, X. Xie, W. Chen *et al.*, "Supervised knowledge makes large language models better in-context learners," *arXiv preprint arXiv:2312.15918*, 2023.