Research Seminar

Feb 29th, 2024

Supervised Knowledge Makes Large Language Models Better In-context Learners

Linyi Yang

School of Engineering

Hangzhou, China

yanglinyi@westlake.edu.cn

WESTLAKE UNIVERSITY

Microsoft Research 微软亚洲研究院

Motivation: On the OOD generalization of LLMs

Performance Decay: Human< InstructGPT/ChatGPT< Fine-tuned Model

Model	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	Avg	Avg Δ↓
Humans (OOD)	*92.36*	*84.13*	*81.10*	*83.47*	*84.70*	*85.43*	*80.28*	*58.98*	*80.14*	*7.82*
GPT-3 (ID)	93.68	69.27	79.20	80.20	79.21	72.15	88.10	50.13	76.49	-
GPT-3.5 (ID)	95.75	72.25	82.78	82.71	73.36	75.69	89.55	54.99	78.39	-
GPT-3 (OOD)	92.33	61.50	79.00	71.03	59.55	55.41	73.74	27.31	64.98	11.51
GPT-3.5 (OOD)	**95.92**	66.01	75.84	66.15	58.43	67.96	74.01	30.77	66.90	**11.49**
ELECTRA-large (OOD)	95.14	**76.94**	**80.44**	**78.74**	**69.96**	**77.24**	**81.14**	**37.85**	**69.68**	21.87

Table 7: OOD performance of GPT-3 and GPT3.5 using in-context learning compared with human performance and ELECTRA-large. We randomly select a single instance for each label. GPT-3 refers to text-davinci-003, and GPT-3.5 denotes the gpt-3.5-turbo.

The OOD performance of GPT-like models still underperforms the humans.

[Yang et al., 2023] GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-distribution Generalization Perspective. In Proceedings of the ACL 2023 Findings.

Huggingface Daily Selection (More than 25K views in one day)

AK @_akhaliq

Subscribe ···

Supervised Knowledge Makes Large Language Models Better In-context Learners

paper page: huggingface.co/papers/2312.15...

Large Language Models (LLMs) exhibit emerging in-context learning abilities through prompt engineering. The recent progress in large-scale generative models has further expanded their use in real-world language applications. However, the critical challenge of improving the generalizability and factuality of LLMs in natural language understanding and question answering remains under-explored. While previous in-context learning research has focused on enhancing models to adhere to users' specific instructions and quality expectations, and to avoid undesired outputs, little to no work has explored the use of task-Specific fine-tuned Language Models (SLMs) to improve LLMs' in-context learning during the inference stage. Our primary contribution is the establishment of a simple yet effective framework that enhances the reliability of LLMs as it: 1) generalizes out-of-distribution data, 2) elucidates how LLMs benefit from discriminative models, and 3) minimizes hallucinations in generative tasks. Using our proposed plug-in method, enhanced versions of Llama 2 and ChatGPT surpass their original versions regarding generalizability and factuality. We offer a comprehensive suite of resources, including 16 curated datasets, prompts, model checkpoints, and LLM outputs across 9 distinct tasks. Our empirical analysis sheds light on the advantages of incorporating discriminative models into LLMs and highlights the potential of our methodology in fostering more reliable LLMs.

SUPERVISED KNOWLEDGE MAKES LARGE LANGUAGE MODELS BETTER IN-CONTEXT LEARNERS

Linyi Yang[1,2]*, **Shuibai Zhang**[1]*, **Zhuohao Yu**[3]*, **Guangsheng Bao**[1], **Yidong Wang**[3], **Jindong Wang**[4], **Ruochen Xu**[4], **Wei Ye**[3], **Xing Xie**[4], **Weizhu Chen**[4], **Yue Zhang**[1,2]†

[1]School of Engineering, Westlake University, [2]Westlake Institute for Advanced Study
[3]Peking University, [4]Microsoft

ABSTRACT

Large Language Models (LLMs) exhibit emerging in-context learning abilities through prompt engineering. The recent progress in large-scale generative models has further expanded their use in real-world language applications. However, the critical challenge of improving the generalizability and factuality of LLMs in natural language understanding and question answering remains under-explored. While previous in-context learning research has focused on enhancing models to adhere to users' specific instructions and quality expectations, and to avoid undesired outputs, little to no work has explored the use of task-Specific fine-tuned Language Models (SLMs) to improve LLMs' in-context learning during the inference stage. Our primary contribution is the establishment of a simple yet effective framework that enhances the reliability of LLMs as it: 1) generalizes out-of-distribution data, 2) elucidates how LLMs benefit from discriminative models, and 3) minimizes hallucinations in generative tasks. Using our proposed plug-in method, enhanced versions of Llama 2 and ChatGPT surpass their original versions regarding generalizability and factuality. We offer a comprehensive suite of resources, including 16 curated datasets, prompts, model checkpoints, and LLM outputs across 9 distinct tasks. Our empirical analysis sheds light on the advantages of incorporating discriminative models into LLMs and highlights the potential of our methodology in fostering more reliable LLMs.

11:19 AM · Dec 27, 2023 · **25.7K** Views

Small Models Help LLMs Efficiently

Challenges: LLMs generally **underperform SLMs** in such natural language understanding tasks, with an increased tendency for **hallucination** when completing classification tasks.

Tasks: We introduce SuperContext, a versatile and straightforward in-context learning strategy to harness the strength of small models to augment LLMs, particularly focusing on Task 1: **OOD generalization** and Task 2: **Factuality**.

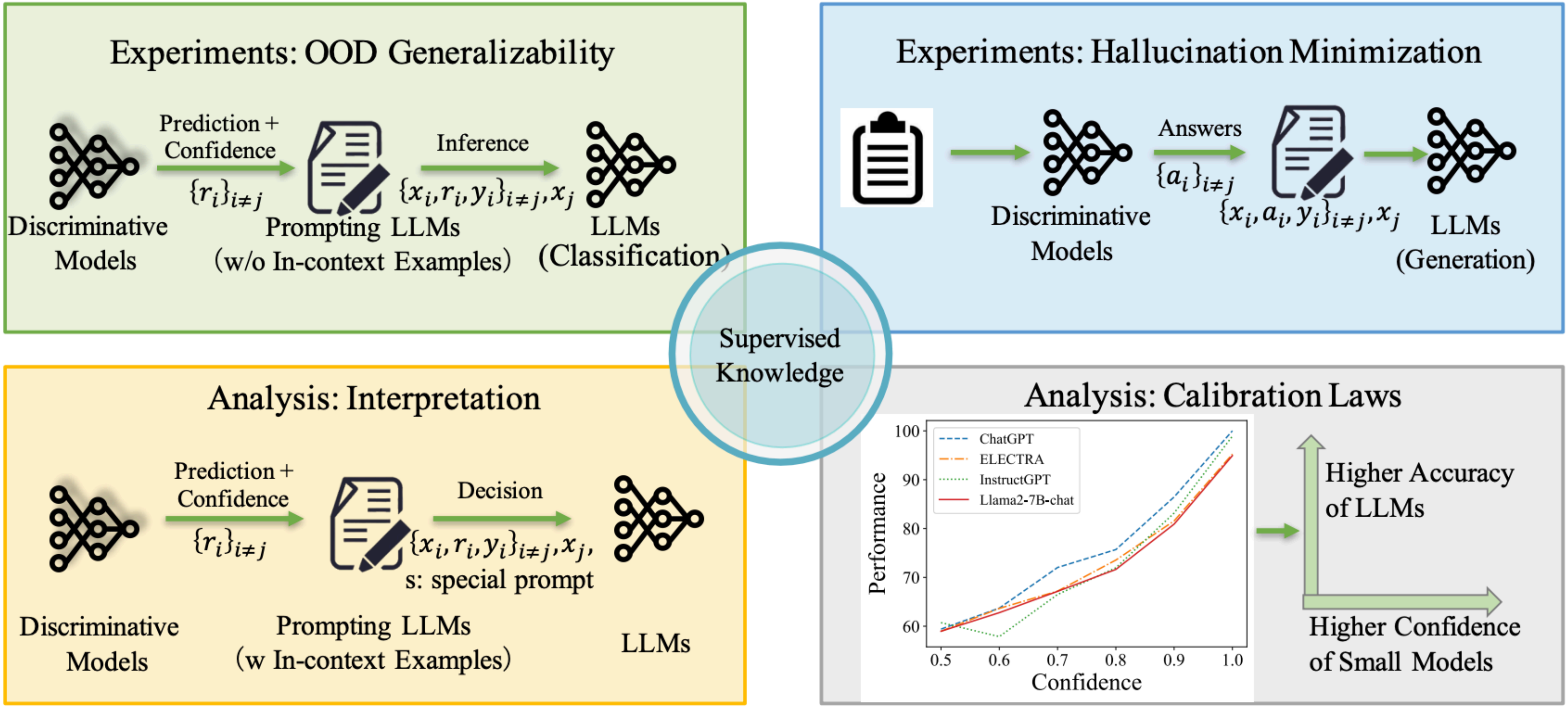
Methods: At the heart of SuperContext is the integration of SLM outputs representing the supervised knowledge into LLM prompts, exemplified by **incorporating the predictive results and confidence of a discriminative model** with LLMs during the inference stage.

In-context Learning:

$$p_{LLM}\left(y_j \mid \{x_i, y_i\}_{i \neq j}, x_j\right) \approx p_{LLM}\left(y_j \mid \{x_i, y_i\}_{i \in S_j}, x_j\right), \quad \forall S_j \subset [1, N] \backslash \{j\}.$$

SLM: refers to cost-efficient, task-specific, fine-tuned language models
LLM: Large language models
OOD: Out-Of-Distribution

SuperContext Contains Two Experiments and Two Analysis Tasks



We denote (x_i, y_i) as a question-answer pair and our receipt r_i is inserted between the question-answer pair.

SuperContext Are Evaluated by 8 NLU Tasks and 1 Generation Task

r_i: the supervised knowledge provided by the discriminative model
s_i: the optional interpretation prompt
NLU: Natural Language Understanding

Prompt Designs

r_i — Auxiliary Information

Use the prediction from the pre-trained model (334M Parameters) fine-tuned on a sentiment analysis dataset as a reference to aid your judgment.
Sentence: "Rich veins of funny stuff in this movie."
Model's Prediction: Positive
Model's Confidence: 0.967

Interpretation — s_i

Please provide your analysis using the format below and then give your final prediction: 1. Influence Degree: On a scale of 0 to 1 (in increments of 0.1), how much did the fine-tuned model's prediction influence your judgment? 2. Critical Features: Identify the specific word or phrase in the test case that played a pivotal role in your prediction of sentiment analysis.

Task 1: 8 NLU tasks and Task 2: 1 generation task

ID	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	SQuAD 2.0
OOD	IMDB Yelp Amazon Flipkart	MNLI-mis SNLI	NewsQA	SciTail HANS	QQP Twitter	MRPC Twitter	SICK	Textbook	Train: 130,319 Dev:11,873

Zero-shot SuperContext Outperforms 16-shot In-context Learning

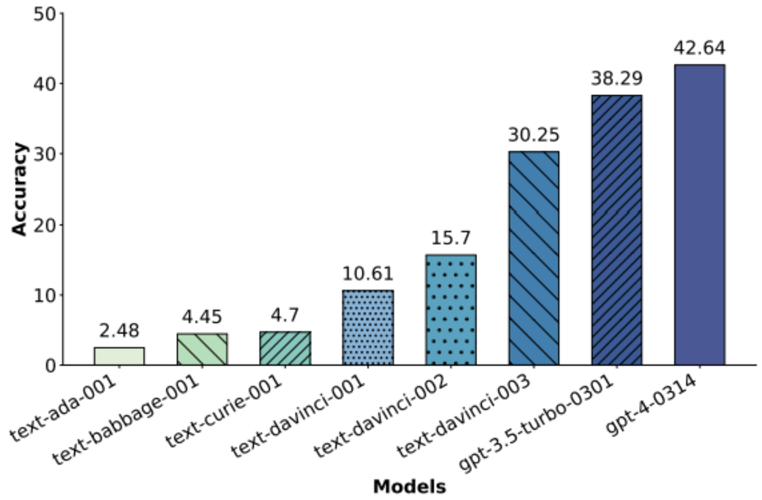
The NLU experiments consist with eight tasks across 15 unique OOD datasets. 'AVG' denotes the average results across 15 datasets. The ChatGPT equipped with SuperContext can even surpass the performance of using ChatGPT with 16-shot in-context examples.

Model	SST-2 OOD	MNLI OOD	QNLI OOD	RTE OOD	MRPC OOD	QQP OOD	STS-B OOD	CoLA OOD	Avg OOD
Human Performance	97.69	91.80	92.33	91.12	83.50	79.13	92.62	66.47	86.83
ELECTRA-large Different Datasets	94.84	87.30	82.66	78.45	63.60	78.08	80.74	40.29	79.86
ChatGPT	94.83	41.54	81.82	68.56	60.23	43.23	72.61	39.05	66.67
ChatGPT (+16-shot)	94.72	64.24	74.14	68.34	60.91	74.24	64.60	**47.15**	72.28
ChatGPT (+BM25)	94.84	64.19	74.00	60.31	**64.29**	68.35	65.22	42.50	71.69
SuperContext (w/o confidence)	94.84	77.21	82.66	78.45	63.60	78.08	80.74	40.29	78.43
SuperContext (+interpreter)	94.84	80.73	**83.81**	78.60	64.26	77.80	76.15	39.47	78.77
SuperContext (zero-shot)	**95.19**	**87.24**	82.91	**78.71**	63.87	**78.65**	**78.75**	41.47	**80.05**
ELECTRA-large Different Datasets	95.42	87.29	82.69	78.84	37.59	77.18	80.74	45.73	76.84
Llama2-chat	90.56	34.30	66.85	60.77	36.20	51.57	37.12	6.94	55.92
Llama2-chat (+16-shot)	94.72	48.20	67.70	61.62	35.72	59.15	18.01	11.52	58.54
Llama2-chat (+BM25)	92.87	48.14	68.48	59.40	37.08	58.24	39.19	10.57	59.69
SuperContext (zero-shot)	94.95	85.45	81.60	78.39	36.70	61.79	45.67	40.84	73.89
SuperContext (w/o confidence)	94.29	76.68	**82.66**	78.46	43.41	**78.17**	80.74	40.26	75.68
SuperContext (16-shot)	**95.45**	**87.14**	82.17	**79.07**	**54.63**	77.18	**80.74**	**45.47**	**79.08**

SuperContext Enhances the Performance of LLMs

"True wisdom is knowing what you don't know."

– Confucius



Comparison between the davinci series and human self-knowledge in instruction input form.

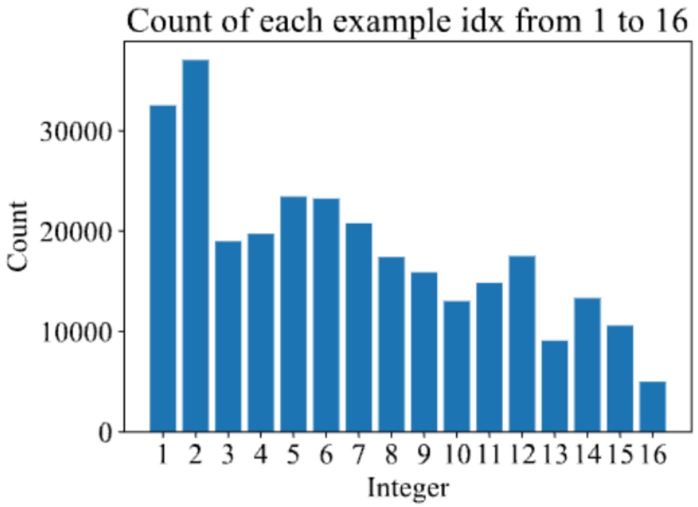
Source: Do Large Language Models Know What They Don't Know?

Table 3: Results of ChatGPT and Llama2-7B-chat, and their variants on SQuAD 2.0. the exact match and valid EM only accounts for the exact match of valid JSON. ACC the accuracy for no-answer questions and ACC accounts for the accuracy of has-answ

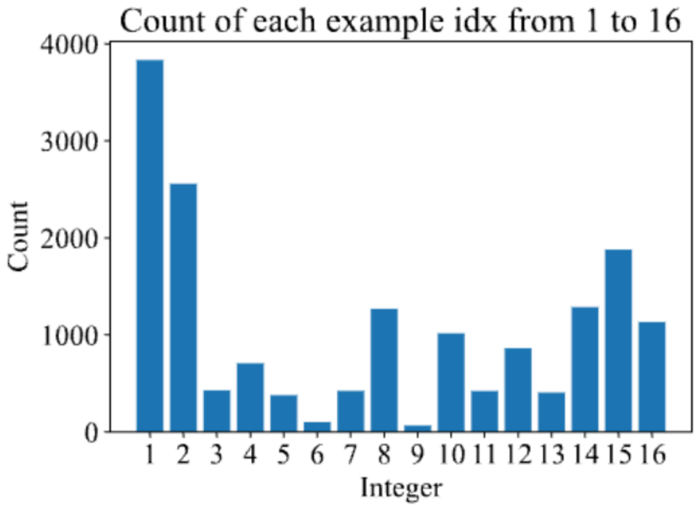
Model	Valid JSON	EM	Valid EM	ACC. No.	ACC. Has.
SuperContext (zero-shot)	85.18	**57.68**	**57.81**	**54.65**	60.71
ChatGPT (cluster+filter)	94.47	49.31	48.81	24.22	74.48
ChatGPT (16-shot)	99.49	44.69	44.52	13.22	76.25
ChatGPT	96.97	55.82	54.76	32.35	**79.35**
SuperContext (16-shot)	41.73	**47.91**	43.27	**63.65**	32.12
Fine-tuned multi-turn	96.40	25.70	26.66	10.47	40.16
Fine-tuned single-turn	97.17	47.22	**48.60**	39.44	**55.02**
Llama2-7B-chat (16-shot)	28.50	37.56	5.32	58.99	6.08
Llama2-7B-chat	40.09	46.48	40.13	3.72	31.87

Lost-in-middle Phenomenon and Calibration Laws

x-axis: order of in-context examples
y-axis: total number of times selected as influential examples



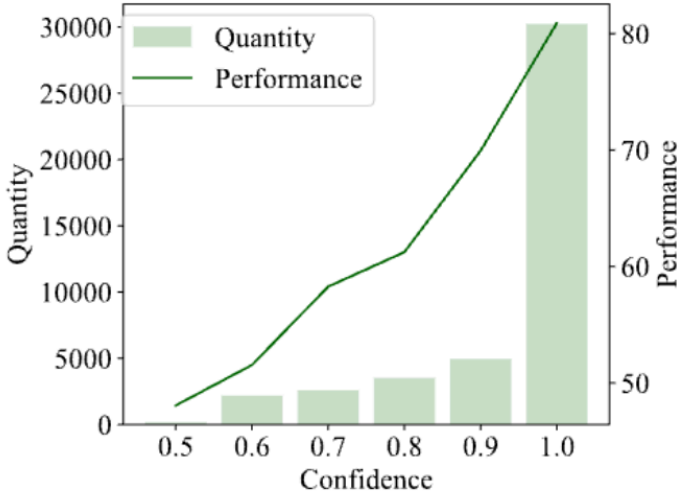
(a) Interpretation results of ChatGPT.



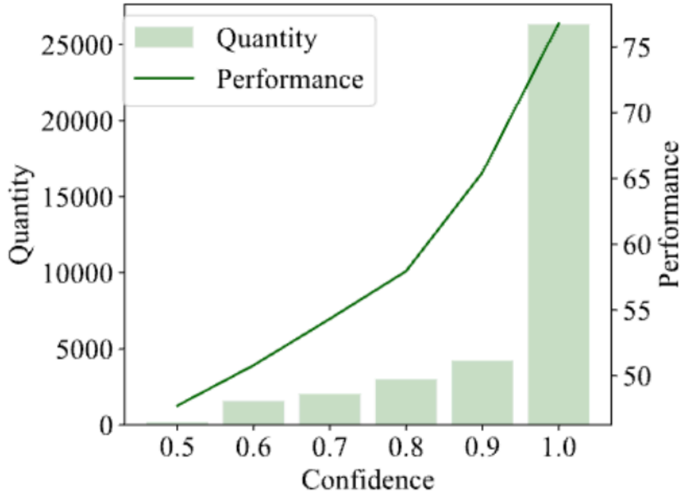
(b) Interpretation results of InstructGPT.

We find that LLMs tend to pay attention to the beginning or the end of the input context, and lost in the middle.

x-axis: confidence output of discriminative models
y-axis: quantity / performance of large language models



(a) The calibration laws of ChatGPT.



(b) The calibration laws of Llama2-7B-chat.

Both ChatGPT and Llama2-7B-chat demonstrate a positive correlation between SLMs' confidence and LLM' performance.

SuperCotext Contributes 15% Gain in Real-world Scenarios

Congratulations from the Epic's CEO and Dr Weizhu Chen

SuperContext contributes to 15% gain in the real-world scenarios.

(1) SuperContext can bring decent performance benefits compared to **original LLMs.**

(2) We highlight the potential in using the knowledge fusion between fine-tuned models and LLMs in practical applications in the future, such as **AI in Finance**.

Limitation: Beyond Learning from In-context Examples

SLM-LLM Interaction:

Tuning Language Models by Proxy

Alisa Liu[♡] **Xiaochuang Han**[♡] **Yizhong Wang**[♡♣] **Yulia Tsvetkov**[♡]
Yejin Choi[♡♣] **Noah A. Smith**[♡♣]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Allen Institute for AI
alisaliu@cs.washington.edu

https://arxiv.org/pdf/2401.08565.pdf

Our Open-source Projects have Received Widespread Attention

PandaLM: ReProducible and Automated Language Model Assessment, ICLR 2024, **Citation: 51, Stars: 815**

USB: A Unified Semi-supervised learning Benchmark for CV, NLP, and Audio, NeurIPS 2022, **Citation: 60, Stars: 1.1k**

A Survey on Evaluation of Large Language Models, TIST 2024, **Citation: 320, Stars: 1.1K**

PromptBench: A Unified Library for Evaluating and Understanding LLMs, arxiv 2024, **Citation: 75, Stars: 1.8k**

Thanks for your attention!
Q&A