Position: The field of small language models needs greater attention and a more systematic approach from the CS research community.

Ivan Samoylenko [1 2]

Abstract

Recent advancements in artificial intelligence have predominantly focused on scaling large language models (LLMs), with major corporations like OpenAI, Google, and Meta developing models comprising hundreds of billions of parameters. This emphasis has fostered a well-organized research community dedicated to optimizing these expansive "black-box" models, extensively discussing both their successes and limitations. Conversely, the systematic exploration of how to utilize these black-box models—through approaches such as prompt engineering and ensemble techniques, which leverage LLMs without retraining—has received comparatively less attention, and the attention it has received is often poorly organized.

In this position-paper, we aim to demonstrate why the field of small language models deserves significantly more focus. We outline the unique challenges facing research in this area, pose key questions related to the development and application of small LLMs, and highlight gaps that persist in the current literature.

1. Introduction

The release of OpenAI's GPT-3.5 in November 2022 marked a significant milestone in the field of Large language model (LLM). With 175 billion parameters, GPT-3.5 demonstrated unprecedented capabilities in generating coherent and contextually relevant text, showcasing the potential of large-scale transformer-based models. The release of GPT-3 marked a turning point in the field of language models, triggering explosive growth in interest from both the research community and industry. The model's ability to solve tasks across diverse domains sparked a global "arms race" in language model development, pushing organizations to pursue increasingly powerful systems.

It is well-established in the literature that increasing training data, model size, and training time often leads to improved model performance (Fedus et al., 2022). As a result, major corporations have embraced scaling as a central strategy. Today, several large-scale compute facilities are either operational or under construction, including *Stargate* (a joint project between Microsoft and OpenAI), xAI's planned datacenter, and the Colossus cluster. Google, Meta, and leading Chinese firms are also rapidly expanding their compute infrastructure.

However, as model size and compute requirements grow, so do the costs of both training and inference, limiting the accessibility and deployment of such models in many real-world applications. To address this, recent industrial research has focused on two primary directions: (i) optimizing compute through better software-hardware co-design, compilation techniques, and parallelism strategies; and (ii) employing ensemble-based architectures such as *Mixture-of-Experts* (MoE).

While such approaches are primarily relevant for the largest models— for "giants of industry" it is often more important in these contexts to showcase the strength of enterprise-scale solutions than to seek excellence with smaller ones. Nevertheless, there are several compelling reasons why research on small language models is not only justified, but potentially crucial:

1. **Accessibility and Customization.** Small language models are far more accessible: they can be trained, fine-tuned, and deployed on local hardware by smaller organizations—including academic labs or independent groups—that may have the resources to develop models in the OLMo-7B or 13B (Groeneveld et al., 2024) range, but are unlikely to operate at the scale required for state-of-the-art "heavyweights." This accessibility translates into greater control, customizability, and security. For many applications, local deployment can offer stronger data protection compared to reliance

*Equal contribution [1]Moscow Institute of Physics and Technology, Institutsky lane 9, Dolgoprudny, Moscow region, 141700 [2]HSE University, Russian Federation. Correspondence to: Ivan Samoylenko <uvan63@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

on commercial API-based models.

2. **Broader User Base and Societal Impact.** As a direct consequence of this accessibility, small language models are used by a much broader (and potentially less technically sophisticated) audience. For instance, in HR tech companies and startups(Xu et al., 2024), (TTMS, 2024), LLM technologies are already being adopted for resume screening, while small models are increasingly embedded in search, recommendation, and other applied systems. As a result, understanding how such models operate in practice becomes essential from the perspective of safety, fairness, and ethical use.

3. **Environmental Impact.** Smaller language models consume significantly less energy than their larger counterparts. Beyond reducing CO_2 emissions(Authors, 2024), this also yields direct economic benefits. Moreover, in the context of ensemble approaches, the combination of small and large LLMs may unlock considerable potential for both performance and efficiency.

4. **Efficiency.** Large language models are not always necessary for handling simple queries. In fact, there is active development of approaches in which a router system first determines whether a query should be handled by a large or a small model, and then dispatches simple queries to a smaller model, thereby saving inference costs. Such hybrid and routing-based methods are already being explored and show promising results in reducing resource consumption without compromising overall system performance (Team, 2024; Ding et al., 2024).

To summarize, we have formulated a set of key statements regarding the importance and challenges associated with small language models. In the following sections, we aim to systematically justify and illustrate why we believe these statements to be true:

- In Section 2, we provide an illustrative example showing how even fundamental prompt-engineering ideas can yield dramatically different results depending on the size of the underlying model.
- In Section 3, we demonstrate why the current research landscape is insufficiently cohesive, highlighting gaps and fragmentation in the study of prompt engineering and ensemble approaches.
- In Section 4, we discuss a promising alternative direction: ensemble methods and the Mixture-of-Experts (MoE) paradigm, including recent developments on prompt ensembles and routing strategies. We argue that these approaches, particularly in the context of small language models, may offer untapped potential in both efficiency and practical utility.

Finally, in the Conclusion, we summarize our main findings and outline promising directions for future research in the domain of small and ensemble-based language models.

2. Black-box utilization Across Model Sizes: Prompt-Engineering Insights Are Not Universal for small and large models

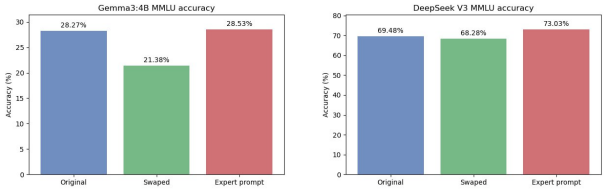
It is a generally acknowledged—but surprisingly poorly documented—fact that large and small language models (LLMs) often behave very differently, especially as the proliferation of models at both ends of the scale accelerates. But how important is this difference in practice? Can we assume that guidelines and recommendations for using and prompting large LLMs will reliably transfer to smaller models, or vice versa?

To illustrate why such differences may be critical, we present results from a simple experiment comparing the performance of one large and one small model (specifically, `deepseekV3-0324` as a large model and `gemma3:4b` as a small model). We do not claim that the observed effects are universal for all (or even most) model pairs, but at the very least, for this pair the issue is demonstrably relevant.

The experimental setup was as follows: we evaluated model effectiveness using the MMLU-pro dataset and zero-shot prompting. In the baseline, the prompt was of the standard form, `"you are a knowledge expert"`. Building on prior findings regarding positional bias—the tendency of models to select the first options in a list—and noting that small LLMs (not just gemma, but also llama 3.2 3b and falcon 3b in our additional tests) tended to output predominantly A/B/C answers, we performed a swap experiment: the correct answer was randomly placed in the middle positions of the list (positions 4 through 6), and model accuracy was re-evaluated (denoted *swapped* in figures).

We also tested whether expert prompting, known to improve performance in larger models, would generalize to small models. For this, we added to each zero-shot prompt both `"you are a knowledge expert"` and `"PhD level in ..."` followed by the question category, creating an *expert prompt* variant.

The results were striking: positional bias dramatically harmed performance for small models—including gemma, which was the best among those tested—with accuracy drops of up to 5–7%. In contrast, large model performance was only mildly affected. Conversely, the addition of expert-level phrasing failed to benefit small LLMs, while improving large model accuracy by 3.5%.



(a) gemma3:4b (small) (b) deepseekV3-0324 (large)

Figure 1. Comparison of model performance on the MMLU-pro dataset across different zero-shot prompting strategies: (i) standard zero-shot prompt without modifications, (ii) with the correct answer shuffled to a middle position in the list (*swapped*), and (iii) with an expert-level prompt incorporating explicit expertise cues (*expert prompt*). Results are shown for both a small model (gemma3:4b) and a large model (deepseekV3-0324).

These findings suggest that prompt-engineering advice and heuristics do not always transfer across model scales, and highlight the importance of tailored approaches. They also point to the potential of hybrid or ensemble models: for instance, a combination of a small LLM and a model with strong probabilistic ranking abilities (such as RoBerta) could mitigate some of these issues. However, we were unable to find substantial literature exploring such ensemble approaches, and our broader search further confirmed the lack of systematic investigation in this area.

3. Fragmentation in the Prompt-Ensemble Research Landscape

In this section, we aim to highlight the factors that currently hinder prompt-engineering-based ensemble methods from becoming a viable alternative (or complement) to research that depends heavily on large-scale compute.

3.1. Sparse Citation Network

The first and most striking observation we made is the lack of citation links between key papers in this area, even though many of them were published at top-tier (A*) conferences or in technical blogs of leading companies. For instance, works such as *AlphaEvolve*, *EvoPrompt*, and others (**????**)—despite operating under strikingly similar paradigms—do not reference each other. This suggests that the research community in this space is highly fragmented and largely unaware of parallel efforts.

To support this hypothesis, we constructed a citation graph based on data retrieved from the OpenAlex API, focusing on literature related to prompt ensembles. Despite using a broad range of relevant keywords and obtaining a reasonably large corpus of nearly 700 papers, the resulting graph was extremely sparse as shown in Figure 2. (To support our analysis, detailed statistics are provided in Table 1.)

Table 1. Citation networks: Prompt-Ensemble vs. MoE

METRIC	PROMPT-ENSEMBLES	MOE
NUMBER OF NODES	642	382
NUMBER OF EDGES	111	725
SIZE OF GIANT COMPONENT	31	162

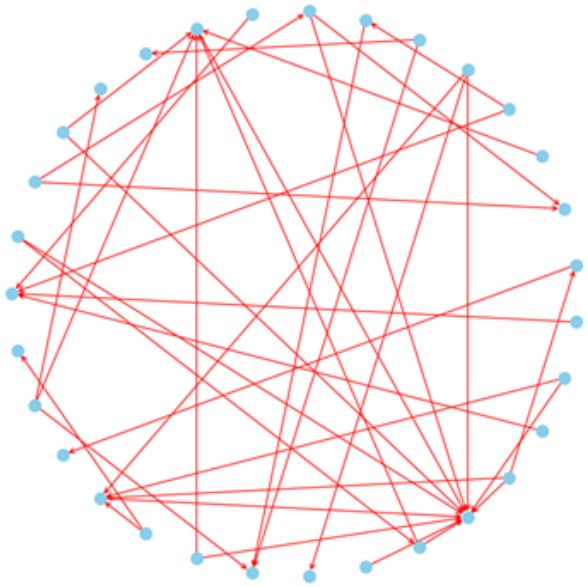


Figure 2. Visualization of the giant component in the citation network related to prompt ensemble methods. Although a large number of papers have been published, the main connected component remains unexpectedly sparse and limited in size.

As shown in Figure 3, for comparison, we examined the citation graph in the field of Mixture-of-Experts (MoE)—a domain that shares a philosophically similar approach through ensemble-based methods, though it focuses specifically on model training. Even with a narrower query scope, we observed significantly more cross-references and a giant component that constitutes a substantial portion of the overall paper set (see Appendix 5 for the list of keywords used to construct the graph).

3.2. Lack of Reproducibility

A notable challenge in the field of adaptive prompting is the lack of reproducibility across studies. Many works cite or rely on outdated language models that are no longer publicly available or relevant to the current generation of LLMs. In other cases, studies rarely converge on a common

set of models, making direct comparisons difficult. Some researchers evaluate adaptive prompting techniques on relatively small, open-source LLMs such as LLaMA or Alpaca, while others conduct experiments exclusively on proprietary models like ChatGPT or GPT-4. Although these lines of research share a philosophical affinity—particularly in their use of ensemble-based approaches and multi-agent prompting—the specific models employed can vary widely: for instance, recent studies have evaluated methods on models such as GPT-4, GPT-3.5, Claude-v1, Alpaca-13B, LLaMA-13B e.t.c (

Moreover, there is currently no widely accepted or regularly used benchmark for evaluating adaptive prompting or prompt-ensemble methods across multiple language models. Most studies introduce their own evaluation protocols and report results on a single model or a limited set of models, which hinders the development of standardized practices and makes it difficult to assess the generality of proposed methods. As a result, the field lacks a robust framework for systematically testing and comparing adaptive prompting techniques across diverse LLM architectures.

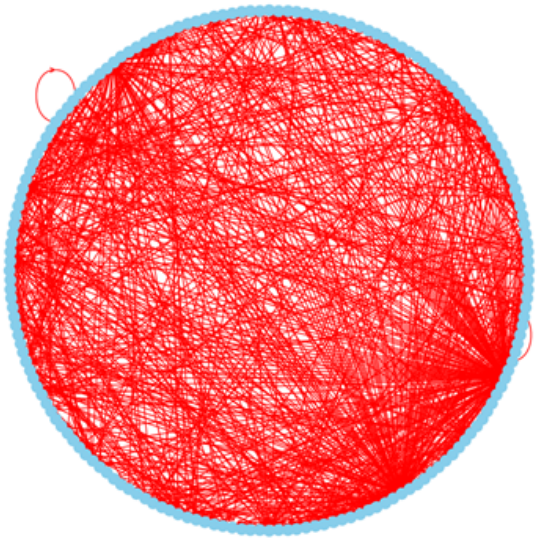


Figure 3. Visualization of the giant component in the citation network related to MoE. Giant component is sufficiently large and dence.

4. Smart Prompt Engineering as an Accessible Alternative to Other Ensemble Methods

As discussed earlier, Mixture-of-Experts (MoE) has become one of the most prevalent methods for achieving high performance in large language models while attempting to keep compute costs manageable. MoE enables efficient inference by activating only a fraction of the total model parameters for each input, often resulting in sublinear compute growth. However, despite these advantages, the development and training of large-scale MoE models remain extremely resource-intensive and costly. For example, recent state-of-the-art models such as DeepSeek R1 reportedly required upwards of $6M in compute, making such approaches inaccessible to most academic and smaller industrial teams.

Given these constraints, it is increasingly important to explore alternative methods that are more accessible for both research and real-world deployment. In this context, the choice of an appropriate model—including black-box framework that do not require any internal modifications—as well as the use of adaptive or ensemble prompt engineering techniques, can be particularly impactful. Recent internal evaluations from OpenAI, for instance, have shown that careful prompt design alone can improve performance by up to 20% on certain benchmarks, further underscoring the significance of both model selection and prompt-based strategies.

This has naturally led to the idea of using ensembles of prompts or automatically generating diverse prompt variations to further enhance the performance of black-box models. Several recent works have explored this avenue. For example, *EvoPrompt* (Microsoft + Tsinghua) employs evolutionary algorithms to optimize prompts, demonstrating significant improvements. Specifically, EvoPrompt achieved up to a 25% performance gain on the BIG-Bench Hard (BBH) benchmark and over 3-point increases in SARI scores for text simplification tasks across both Alpaca and GPT-3.5 models (Guo et al., 2024). Similarly, *AlphaEvolve* (Novikov et al., 2025) (Google DeepMind) utilizes an ensemble of Gemini models within an evolutionary framework to discover and optimize algorithms. Notably, AlphaEvolve improved data center scheduling efficiency by recovering approximately 0.7% of compute resources, enhanced Gemini's training performance by reducing kernel execution time by 23%, and discovered a novel matrix multiplication algorithm that surpassed a 56-year-old record.

Prompt-level analogues of the MoE approach are also beginning to emerge. In (Hu et al., 2024), the authors demonstrate that a well-trained router can simultaneously reduce compute costs and improve accuracy by over 15%. Similarly, (Long et al., 2024) proposes generating several synergistic system prompts to collectively cover different aspects of a task, achieving 5–10% improvements over both zero-shot and few-shot baselines.

In another study, (Wang et al., 2024), an architecture is introduced that learns to route queries among a set of specialized prompts, achieving comparable gains across multiple benchmarks.

Importantly, the latter work draws a direct parallel to the MoE framework, observing that the selection of prompts can be viewed as expert selection:

> 1. Prompt optimization can be viewed as model selection: an LLM pre-trained on next-token prediction can be seen as a collection of conditional probabilistic models, each induced by a specific prompt. Designing diverse prompts is therefore equivalent to selecting different submodels, making automatic prompt routing analogous to expert model routing in MoE.
>
> 2. Theoretical properties of MoE apply to MoP: this conceptual framework allows direct application of MoE theory to the task of optimizing mixtures of prompt experts.

The authors emphasize that, despite the simplicity of prompt-level ensembles, such approaches may provide an accessible and scalable path to performance gains—without the prohibitive costs of training large MoE models.

5. Conclusion

In this position paper, we have highlighted the critical importance of small language models, as well as the need for systematic research and the development of techniques for operating with black-box models. We believe that advancing this line of inquiry will help make LLM technology more accessible and, potentially, more secure for a broader range of users. As a possible solution to the performance limitations of small models, we see significant promise in the development of ensemble-based prompt engineering methods.

Nevertheless, the current state of the field is highly fragmented, with research efforts often proceeding in isolation and lacking reproducibility. Even the performance figures cited throughout this paper, drawn from existing studies, may be somewhat optimistic and difficult to generalize.

To move the field forward, we advocate for the following concrete actions:

- Conducting transparent, comparative panel studies that systematically evaluate major prompt engineering techniques across a diverse set of language models, including not only the most powerful but also the most accessible (for both commercial and personal use). Initiatives such as (Liang et al., 2023) represent an important step in this direction, but such benchmarks need to be actively adopted and continuously developed by the research community.
- Investigating effective ensemble constructions in which language models can be utilized as black-boxes, with a focus on practical applicability and robust evaluation. Transparent benchmarks for such ensembles and AI-agent systems are also needed to facilitate systematic comparison and reproducibility.

We hope these directions will help foster a more cohesive and impactful research landscape for the future of small language models and their real-world deployment.

Acknowledgements

This work was supported by the Ministry of Economic Development of the Russian Federation (code 25-139-66879-1-0003).

Impact Statement

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

References

Authors. Reconciling the contrasting narratives on the environmental impact of ai. *Nature*, 2024. URL https://www.nature.com/articles/s41598-024-76682-6.

Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. V. S., and Awadallah, A. H. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024. URL https://arxiv.org/abs/2404.14618.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (1):1–39, 2022. URL https://arxiv.org/abs/2101.03961.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N., and Hajishirzi, H. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., and Yang, Y. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.08532.

Hu, Q. J., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ranganath, G., Keutzer, K., and Upadhyay, S. K. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Promptbench: Towards reproducible benchmarks for prompt engineering methods. *arXiv preprint arXiv:2306.04528*, 2023. URL https://arxiv.org/abs/2306.04528.

Long, D. X., Yen, D. N., Luu, A. T., Kawaguchi, K., Kan, M.-Y., and Chen, N. F. Multi-expert prompting improves reliability, safety, and usefulness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.1135/.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J. R., Mehrabian, A., Kumar, M. P., See, A., Chaudhuri, S., Holland, G., Davies, A., Nowozin, S., Kohli, P., and Balog, M. Alphaevolve: A coding agent for scientific and algorithmic discovery. Technical report, Google DeepMind, May 2025. URL https://deepmind.google/discover/blog/alphaevolve.

Team, L. Routellm: An open-source framework for cost-effective llm routing. https://lmsys.org/blog/2024-07-01-routellm/, 2024. Accessed: 2025-05-27.

TTMS. Small language models: Key features and business applications, 2024. URL https://ttms.com/small-language-models-key-features-and-business-applications/.

Wang, R., An, S., Cheng, M., Zhou, T., Hwang, S. J., and Hsieh, C.-J. One prompt is not enough: Boosting llms via ensemble of diverse prompts. *arXiv preprint arXiv:2407.00256*, 2024. URL https://arxiv.org/abs/2407.00256.

Xu, W., Desai, J., Wu, F., Valvoda, J., and Sengamedu, S. H. Hr-agent: A task-oriented dialogue (tod) llm agent tailored for hr applications. *arXiv preprint arXiv:2410.11239*, 2024. URL https://arxiv.org/abs/2410.11239.

Algorithm 1 Build Seed Citation Graph

Input: List of seed articles W
$G \leftarrow$ new directed graph
$S \leftarrow$ set of IDs from W
for each $id \in S$ **do**
  Add node id to G
end for
for each article w in W **do**
  $src \leftarrow w.\text{id}$
  for each $ref \in w.\text{referenced\_works}$ **do**
    if $ref \in S$ **then**
      Add edge from src to ref in G
    end if
  end for
end for
Output: Graph G

A. Citation Graph Construction Details

To analyze the structure and fragmentation of research in the areas of prompt-ensemble methods and Mixture-of-Experts (MoE), we constructed two separate citation networks using the OpenAlex API. The overall procedure consisted of two main steps: collecting seed articles and building the citation graphs.

A.1. Seed Article Collection

For each research area, we queried the OpenAlex database using a targeted set of keywords. For the **prompt-ensemble** domain, the following search terms were used: *"Meta Prompting", "EvoPrompt", "Mixture of Prompts", "Mixture-of-Expert Prompts", "Automatic Prompt Optimization", "Prompt Expert Routing", "FunSearch", "AI agents", "Prompt ensemble", "Adaptive prompting", "LLM router"*. For the **Mixture-of-Experts (MoE)** domain, we used: *"Mixture of experts", "Mixture-of-Experts", "MoE"*.

All articles returned by these queries were considered as "seed works" for their respective domains.

A.2. Citation Graph Construction

We then constructed a directed citation graph for each set of seed articles using the following approach:

- **Nodes:** Each node represents a seed article.
- **Edges:** A directed edge from article A to article B is included if and only if A cites B and B is also among the seed articles.

To ensure the inclusion of isolated articles (i.e., those not citing or cited by other seeds), we added all seed nodes to the graph even if they had no connections.

The pseudocode for constructing the citation graph is provided above