# Small Language Model Enhancement Strategies in Practice: A Signal-Oriented Taxonomy and Open Questions (Extended Abstract)

**YongKyung Oh**[1], **Jaesun Yeom**[2]

[1]Medical & Imaging Informatics (MII) Group, University of California, Los Angeles (UCLA)
[2]Department of Industrial and Management Engineering, Hanbat National University
yongkyungoh@mednet.ucla.edu, jsyeom@hanbat.ac.kr

## Abstract

Deploying small language models (SLMs) for strategic decision-making in regulated domains requires balancing accuracy, cost, and auditability. Fine-tuning, cascading, and prompt augmentation can improve performance, but these approaches are typically studied independently. This paper presents a taxonomy that organizes enhancement strategies by the type of signal delivered to the SLM. Comparing strategies along this dimension shows differences not only in training requirements but also in what information crosses the model boundary at inference time. These distinctions have practical consequences for cost, auditability, and robustness that remain underexamined. The paper concludes with open questions for enterprise deployment.

## Research Background

Small language models (SLMs; <10B parameters) (Subramanian, Elango, and Gungor 2025) are increasingly deployed where cost, latency, and regulatory constraints limit frontier API models. However, their accuracy on domain-specific tasks remains inconsistent. Instruction-tuned models improve on seen data but degrade on unseen inputs (Du et al. 2024; Fatemi, Hu, and Mousavi 2025), a pattern also reported in financial, medical, and legal text classification. Yet existing work compares performance, not signal type.

## Taxonomy of Enhancement Strategies

Table 1 classifies five strategies by *signal type*: what information, if any, reaches the SLM at inference time.

| Strategy | Signal to SLM | Training |
|---|---|---|
| Capacity Reallocation | None (weights modified) | Required |
| Model Replacement | None (input routed away) | Not required |
| Retrieval Augmentation | Retrieved documents | Not required |
| Prompt Engineering | Natural-language guidance | Optional |
| Symbolic Injection | Structured rules / ontologies | Optional |

Table 1: SLM enhancement strategies grouped by signal type. The Signal to SLM column indicates what information, if any, is added to the model input at inference time.



**Capacity Reallocation.** Fine-tuning reallocates representational capacity toward the target task at the cost of general capability (Fu et al. 2023; Wei et al. 2022a).

**Model Replacement.** Cascade architectures route low-confidence inputs to larger models, discarding the small model's probability distribution (Chen, Zaharia, and Zou 2024; Rabanser et al. 2025; Gupta et al. 2024).

**Retrieval Augmentation.** Retrieval-augmented generation (RAG) appends retrieved documents to the prompt (Lewis et al. 2020; Fan et al. 2024), adding context but not classification-level confidence signals.

**Prompt Engineering.** Chain-of-thought prompting (Wei et al. 2022b) and model-generated guidance (Juneja et al. 2023; Bi et al. 2025) fall into this category. These approaches introduce structured linguistic cues but require prompt design or additional training.

**Symbolic Injection.** Ontology-based methods encode domain constraints as formal rules, either by post-correcting neural outputs or aligning model representations with structured knowledge (Gueddes and Mahjoub 2026; Liu et al. 2025), though they depend on ontology construction.

## Gaps and Open Questions

**Cost–performance trade-offs.** Fine-tuning reduces general capability (Fu et al. 2023), and guidance methods increase inference cost (Bi et al. 2025). Whether zero-training signal injection can narrow the accuracy gap at lower overhead remains an open question, particularly in cost-sensitive decision-making settings (Kirtac and Germano 2024).

**Uncertainty as input rather than filter.** Neural networks exhibit systematic overconfidence (Guo et al. 2017; Geng et al. 2024), but confidence is usually reduced to a routing threshold (Geifman and El-Yaniv 2017). Whether a classifier's full probability distribution can serve as structured input to an SLM remains an open research question.

**Auditability of information flow.** Replacement cascades distribute the decision rationale across multiple models, which complicates bias tracing in algorithmic decision-making systems (Marabelli, Newell, and Handunge 2021). LLM-based financial analysis shows systematic bias linked to model selection and firm characteristics (Nakagawa, Hirano, and Fujimoto 2024; Lee et al. 2025). A single-path design that varies only the input signal remains underexplored.

## References


Bi, J.; Wu, Y.; Xing, W.; and Wei, Z. 2025. Enhancing the Reasoning Capabilities of Small Language Models via Solution Guidance Fine-Tuning. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 9074–9084. Abu Dhabi, UAE: Association for Computational Linguistics.

Chen, L.; Zaharia, M.; and Zou, J. 2024. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Trans. Mach. Learn. Res.*, 2024.

Du, K.; Xing, F.; Mao, R.; and Cambria, E. 2024. Financial Sentiment Analysis: Techniques and Applications. *ACM Comput. Surv.*, 56(9).

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 6491–6501. New York, NY, USA: Association for Computing Machinery.

Fatemi, S.; Hu, Y.; and Mousavi, M. 2025. A Comparative Analysis of Instruction Fine-Tuning Large Language Models for Financial Text Classification. *ACM Trans. Manage. Inf. Syst.*, 16(1): 6:1–6:30.

Fu, Y.; Peng, H.; Ou, L.; Sabharwal, A.; and Khot, T. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, 10421–10430. PMLR.

Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 6577–6595.

Gueddes, A.; and Mahjoub, M. A. 2026. BERT-OntoSent: combining BERT language model with sentiment ontology for enhanced sentiment analysis on social media. *Journal of Information and Telecommunication*, 10(1): 1–23.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1321–1330. Sydney, NSW, Australia: JMLR.org.

Gupta, N.; Narasimhan, H.; Jitkrittum, W.; Rawat, A. S.; Menon, A. K.; and Kumar, S. 2024. Language Model Cascades: Token-Level Uncertainty And Beyond. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Juneja, G.; Dutta, S.; Chakrabarti, S.; Manchanda, S.; and Chakraborty, T. 2023. Small Language Models Fine-tuned to Coordinate Larger Language Models improve Complex Reasoning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3675–3691. Singapore: Association for Computational Linguistics.

Kirtac, K.; and Germano, G. 2024. Sentiment trading with large language models. *Finance Research Letters*, 62: 105227.

Lee, H.; Seo, J.; Park, S.; Lee, J.; Ahn, W.; Choi, C.; Lopez-Lira, A.; and Lee, Y. 2025. Your AI, Not Your View: The Bias of LLMs in Investment Analysis. In *Proceedings of the 6th ACM International Conference on AI in Finance*, 150–158. New York, NY, USA: Association for Computing Machinery.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Liu, Z.; Gan, C.; Wang, J.; Zhang, Y.; Bo, Z.; Sun, M.; Chen, H.; and Zhang, W. 2025. OntoTune: Ontology-Driven Self-training for Aligning Large Language Models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, 119–133. New York, NY, USA: Association for Computing Machinery.

Marabelli, M.; Newell, S.; and Handunge, V. 2021. The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. *The Journal of Strategic Information Systems*, 30(3): 101683.

Nakagawa, K.; Hirano, M.; and Fujimoto, Y. 2024. Evaluating Company-specific Biases in Financial Sentiment Analysis using Large Language Models. In *2024 IEEE International Conference on Big Data (BigData)*, 6614–6623.

Rabanser, S.; Rauschmayr, N.; Kulshrestha, A.; Poklukar, P.; Jitkrittum, W.; Augenstein, S.; Wang, C.; and Tombari, F. 2025. Gatekeeper: Improving Model Cascades Through Confidence Tuning.

Subramanian, S.; Elango, V.; and Gungor, M. 2025. Small Language Models (SLMs) Can Still Pack a Punch: A survey.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.