CODES: A CONTEXT-EFFICIENT FRAMEWORK FOR ENHANCING SMALL LANGUAGE MODELS VIA DOMAIN-SPECIFIC ADAPTATION AND MODEL ENSEMBLING

LAN HU[1†*], YUTING XIN[2†], BINQI SHEN[3], HANYU CAI[3], LIER JIN[4]
[1]Carnegie Mellon University
[2]University of Minnesota
[3]Northwestern University
[4]Duke University
E-MAIL:lanh@alumni.cmu.edu, yuting.xin@outlook.com,
binqishen2021@u.northwestern.edu, hanyucai2022@u.northwestern.edu,
lierjin@alumni.duke.edu
[†]Equal contribution [*]Corresponding author

Abstract:

Adapting language models to specialized domains remains challenging under limited computational resources. We introduce CoDES (Context-efficient Domain Ensemble System), a framework that improves small language model performance through domain-specific fine-tuning and weighted parameter ensembling. CoDES combines parameter-efficient adaptation via Low-Rank Adaptation (LoRA) with completion-only supervision, and merges two fine-tuned models through weighted parameter averaging to improve robustness and accuracy.

We evaluate CoDES on two biomedical question answering benchmarks, MedMCQA and MedQA. On MedMCQA, the ensemble achieves 74.8% accuracy, approaching a 72B-parameter model (77.1%) while consuming 2.5 times less energy. Consistent improvements on MedQA further demonstrate the framework's generalizability across datasets and examination styles. Taken together, these results show that targeted domain adaptation combined with model ensembling provides a practical pathway for deploying competitive language model systems under realistic resource constraints.

Keywords:

small language models; parameter-efficient fine-tuning; low-rank adaptation (LoRA); model ensembling; domain adaptation

1 Introduction

Large language models (LLMs) have reshaped what is possible in natural language processing, achieving strong results across reasoning, knowledge retrieval, and language understanding [1]. Their reach has extended well beyond text, with large-scale models now deployed across diverse modalities and real-world applications [2, 3]. Yet these advances carry significant computational, environmental, and financial costs [4, 5], and selecting an appropriately sized model has become a practical necessity for many organizations [6, 7].

Smaller language models have emerged as a compelling alternative, offering faster training, lower inference latency, and broader accessibility [8]. Their efficiency makes them well suited for latency-sensitive and resource-constrained settings [9, 10], and they are seeing growing adoption in production systems where cost and scalability matter [11, 12]. Nevertheless, a meaningful performance gap persists between small and large models [13]. While scaling remains the dominant strategy for large models [14], smaller models must rely on more targeted approaches to narrow this gap [15, 16]. Prior work has explored architectural improvements [17, 18, 19] and advances in training strategies and knowledge representation [20, 21, 22], yet it remains unclear whether carefully optimized compact models can reliably match the reasoning performance of much larger counterparts.

To address this question, we propose CoDES (Context-efficient Domain Ensemble System), a framework that improves small language model performance through domain-specific adaptation and weighted parameter ensembling. Rather than relying on model scale, CoDES targets how compact models are trained and combined within specialized domains. We evaluate the framework on two established biomedical question answering benchmarks, MedMCQA [23] and MedQA [24], demonstrating that carefully adapted small models can approach the performance of much larger models at a fraction of

the computational cost.

2 Related Work

2.1 Domain Specialization

Existing work has begun to examine domain-specific modeling for high-stakes applications, showing that structured inputs and domain signals improve accuracy, interpretability, and output reliability [25, 26]. Recent work has further demonstrated that domain-specialized models can achieve strong performance even in low-resource settings through targeted training strategies [27]. Similar domain-aligned optimization strategies have also been explored in other high-stakes settings, such as legal and finance service systems [28, 29]. However, systematic studies that bring together compact model tuning, calibrated probabilistic outputs, and deployment constraints for the biomedical field remain limited [30].

2.2 Parameter- and Compute-efficient Adaptation

Adapting pretrained transformers under hardware and cost constraints typically updates only a small set of parameters or uses low-precision weight formats. Techniques such as low-rank adapters (LoRA) [31] and 4-bit/mixed-precision loading reduce memory and compute while keeping pretrained behavior, making them practical choices for fine-tuning compact models on domain data [32, 33].

In this work, we adopt LoRA as a representative and widely used PEFT method to isolate the effect of the proposed framework, particularly weighted parameter ensembling. Other PEFT approaches, such as QLoRA and adapter tuning, follow similar principles but differ in implementation details, including quantization and parameter injection [34, 35]. Retrieval-augmented generation (RAG) incorporates external knowledge at inference time [36], whereas this work focuses on internalizing domain knowledge within model parameters to reduce inference-time latency and infrastructure requirements.

Beyond parameter-efficient adaptation methods, complementary directions have explored incorporating structured domain knowledge into model representations, for example through latent or autoencoder-based approaches, which have been shown to be effective in modeling complex processes across domains [37, 38]. Related work has also investigated broader system-level, including agent-level architectures, highlighting alternative ways of organizing computation and knowledge beyond parameter updates [39, 40]. These directions are complementary to our focus on parameter-efficient adaptation and model ensembling for improving small language model performance under resource constraints.

2.3 Model Ensembling and Calibration

Combining complementary small models can improve accuracy and calibration without switching to much larger models. Selective processing reduces wasted work and prevents the introduction of misleading background details by routing computation only to the most relevant input regions or scales [41, 42]. Similar ideas have been explored in retrieval-augmented systems that selectively route and filter contextual information to improve the relevance of contextual inputs and final performance [43, 44]. Related work further shows that efficiency can be improved by pruning less useful intermediate reasoning or action paths in LLM systems [45]. Compact multistage fusion modules can integrate complementary feature streams before prediction, and similar benefits of combining feature-based and relational signals have been observed in other domains [46, 47]. Simple model integration techniques such as weighted parameter averaging and light ensembling effectively blend strengths of different fine-tuned checkpoints with little extra cost; these ideas motivate our weighted parameter-averaging approach.

These insights motivated our design: we target a domain-focused biomedical setting, use 4-bit loading with LoRA to maintain computational efficiency, format data as conversational templates and apply completion-only supervision so fine-tuning concentrates on answer tokens, and merge two fine-tuned checkpoints by weighted parameter averaging to combine their complementary strengths.

3 Methodology

We propose a context-efficient framework designed to enhance small language models through domain-specific adaptation and model ensembling, with the goal of achieving performance comparable to much larger LLMs while maintaining practical feasibility in resource-constrained environments. The framework is evaluated in a high-impact domain: medical question answering, and is designed to be extensible to other knowledge-intensive domains such as legal and financial services. This section describes the datasets, models, training pipeline, and evaluation metrics used in our experiments.

3.1 Data Selection

To evaluate both the effectiveness and generalizability of the proposed framework, experiments were conducted across two

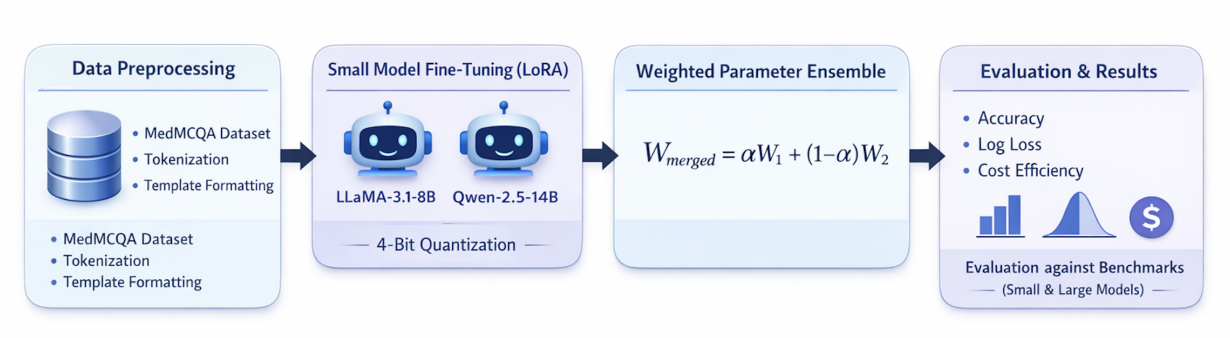


FIGURE 1. Overview of the Context-efficient Domain Ensemble System (CoDES) Framework

complementary biomedical question answering benchmarks spanning diverse medical knowledge domains, examination styles, and difficulty levels: MedMCQA and MedQA.

MedMCQA consists of multiple-choice questions drawn from standardized Indian medical entrance examinations (AIIMS and NEET-PG). MedQA consists of multiple-choice questions drawn from professional medical licensing examinations across multiple countries, including the United States (USMLE), Mainland China, and Taiwan. Together, these datasets provide the following complementary strengths:

- **Diverse Knowledge:** MedMCQA covers 21 medical subject areas and 2.4K distinct healthcare topics; MedQA spans a broad range of clinical specialties across multiple national licensing systems, collectively testing across a wide spectrum of medical reasoning tasks.
- **Domain Relevance:** Both datasets reflect real-world clinical reasoning demands in high-stakes healthcare environments, where domain-specific tuning must compensate for the limited capacity of general LLMs [48, 49, 50].
- **Benchmark Popularity:** Both are broadly adopted in the medical NLP community [51], enabling transparent and reproducible comparison with other work.

3.2 Model Selection

To study performance across model scales, we include both a high-capacity reference model and smaller, computationally economical alternatives.

Baseline Model: Qwen2.5-72B [52] is used as a high-capacity reference to establish a performance ceiling in the absence of specialized medical models. Although evaluated zero-shot, this model provides a robust upper-bound benchmark for the native reasoning capabilities of frontier-scale LLMs.

Candidate Models: Two smaller models were selected based on accessibility, community adoption, and suitability for controlled comparison:

- Qwen2.5-14B [52]: a reduced-scale counterpart to the baseline model, enabling direct examination of how parameter count influences performance under comparable architectural conditions.
- LLaMA3.1-8B [53]: a compact, open-source model with strong general-language proficiency and significantly lower computational requirements.

For both candidate models, we report zero-shot performance as well as performance after applying context-efficient fine-tuning, isolating the effect of the proposed framework from model size. We further examine whether combining multiple fine-tuned candidate models through ensembling yields additional improvements.

3.3 Training Method

The overall methodology is illustrated in Fig. 1. Our framework processes domain-specific data through targeted preprocessing, then fine-tunes small language models using parameter-efficient adaptation, which are further combined via weighted parameter ensembling. The resulting models are evaluated on downstream tasks using standard metrics such as accuracy and log loss, allowing assessment of both predictive quality and computational efficiency.

Data Preprocessing: Each sample was formatted into a conversational chat template with system and user messages, then tokenized and truncated to a maximum length. Labels were generated so that only answer tokens contributed to the loss, while all prompt tokens were masked.

Parameter-Efficient Fine-Tuning: The candidate models were loaded in 4-bit precision to reduce GPU memory overhead. We applied LoRA with rank $r = 16$ targeting all linear modules. This configuration was chosen to prevent overfitting to the training samples while preserving pretrained reasoning capabilities and enabling efficient domain adaptation with a reduced memory footprint.

Completion-Only Supervision: Parameter updates were restricted to answer tokens, reducing gradient noise and improving the model's ability to produce concise, discrete outputs.

Ensemble Strategy: The two best-performing fine-tuned models were combined via weighted parameter averaging [54, 55], with a scaling factor α controlling each model's contribution. Different weight configurations were tested to identify the optimal combination, allowing the ensemble to leverage complementary strengths of both models.

3.4 Evaluation Metrics

Model performance was evaluated using two metrics. **Accuracy** measures the proportion of questions for which the predicted choice matched the correct answer. **Log Loss** (cross-entropy loss) measures the quality of probabilistic predictions:

$$L(y, p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\{A,B,C,D\}} \mathbf{1}[y_i = c]\log(p_{i,c}),$$

where N is the total number of evaluated questions, y_i is the correct label, $p_{i,c}$ is the predicted probability for choice c, and $\mathbf{1}[\cdot]$ is the indicator function.

Models that assign high probability to incorrect answers result in a heavier penalty. Consequently, when two models have the same accuracy, the model with lower log loss exhibits more consistent and better-calibrated predictive behavior.

4 Experiments & Results

Building on the dataset preparation, model selection, and training techniques detailed in the Methodology section, we conducted a set of experiments aiming to improve performance of small language models.

4.1 Baseline Performance

To establish a baseline performance, we conducted zero-shot evaluation on one large language model and two smaller models on an identical dataset consisting of 10,000 randomly-selected examples from MedMCQA dataset. Specifically, we compare Qwen2.5-72B, LLaMA 3.1-8B, and Qwen2.5-14B under the same evaluation protocol. The results are reported in Table 1.

The Qwen2.5-72B model achieves the highest accuracy of 77.1%, whereas Llama 3.1-8B and Qwen2.5-14B achieve 63.5% and 64.0%, respectively. This performance gap is attributable to the increased parameter capacity of large LLMs, which enables more expressive representations and improved generalization. The 77.1% accuracy achieved by the large model serves as a baseline reference and will be used for comparison in subsequent sections.

TABLE 1. Baseline Model Performance Comparison

Baseline Model	Sample Size	Accuracy
Qwen2.5-72B	10k	77.1%
Llama3.1-8B	10k	63.5%
Qwen2.5-14B	10k	64.0%

4.2 Small Model Fine Tuning with LoRA

We fine-tuned Llama3.1-8B and Qwen2.5-14B using LoRA on 10,000 samples from the MedMCQA training set, with 3,000 examples used for validation.

To assess whether parameter-efficient tuning is sufficient, we compare LoRA with full fine-tuning under the same setup. As shown in Table 2, the performance difference is minimal, with accuracy differing by at most 0.2% and nearly identical log loss. This suggests that LoRA can match full fine-tuning performance while being significantly more efficient.

Table 3 reports results across different LoRA hyperparameter settings. In general, moderate learning rates (around 5×10^{-5}) perform best, while smaller learning rates tend to underperform. Increasing the number of epochs from one to two leads to consistent improvements, indicating that additional exposure to domain data helps the models adapt more effectively.

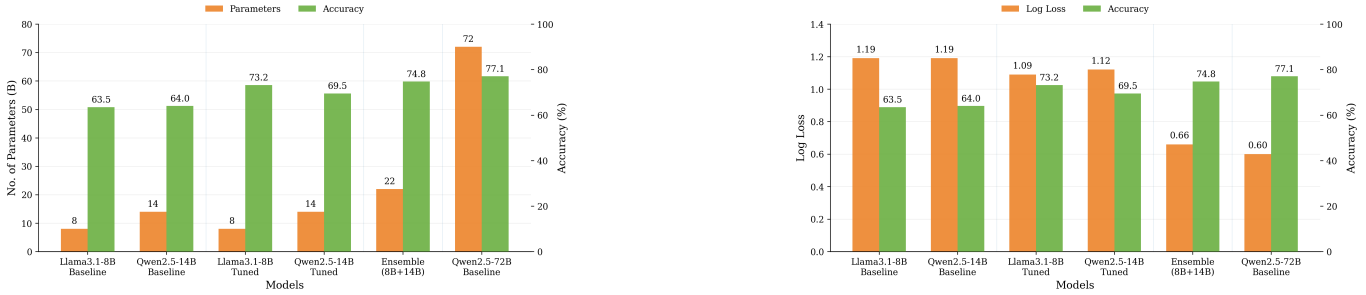
Compared to the baseline results in Table 1, fine-tuning leads to clear performance gains, with accuracy improving by 9.7% for Llama3.1-8B and 5.5% for Qwen2.5-14B. Overall, these results show that even relatively small models can benefit substantially from domain-specific adaptation.

4.3 Ensemble of Small Models v.s. Large model

We ensembled the two best-performing fine-tuned models (Llama3.1-8B[3] and Qwen2.5-14B[2]) via weighted parameter averaging. The formula for ensembling two models is as below,

TABLE 2. Comparison between full fine-tuning (SFT) and LoRA-based fine-tuning.

Model	Method	Learning Rate	Epoch	Accuracy	Log Loss
Llama3.1-8B	Full SFT	1×10^{-4}	1	72.7%	1.09
Llama3.1-8B	LoRA	1×10^{-4}	1	72.5%	1.09
Qwen2.5-14B	Full SFT	1×10^{-4}	1	63.8%	1.18
Qwen2.5-14B	LoRA	1×10^{-4}	1	63.7%	1.19



(a) Accuracy vs. Parameter Size

(b) Accuracy vs. Log Loss

FIGURE 2. Performance comparison across baseline, fine-tuned, and ensemble model configurations.

TABLE 3. Performance of Llama3.1-8B and Qwen2.5-14B under different LoRA configurations.

Model	Learning Rate	Epoch	Accuracy	Log Loss
Llama3.1-8B[1]	1×10^{-4}	1	72.5%	1.09
Llama3.1-8B[2]	5×10^{-5}	1	70.6%	1.10
Llama3.1-8B[3]	5×10^{-5}	2	73.2%	1.09
Llama3.1-8B[4]	2×10^{-5}	1	63.6%	1.19
Llama3.1-8B[5]	2×10^{-5}	2	68.5%	1.14
Qwen2.5-14B[1]	1×10^{-4}	1	63.7%	1.19
Qwen2.5-14B[2]	5×10^{-5}	2	69.5%	1.12
Qwen2.5-14B[3]	2×10^{-5}	1	63.5%	1.19

TABLE 4. Performance of weighted parameter ensembles combining fine-tuned Llama3.1-8B and Qwen2.5-14B models.

Llama Weight	Qwen Weight	Accuracy	Log Loss
0.90	0.10	73.1%	0.71
0.80	0.20	73.4%	0.70
0.70	0.30	74.1%	0.68
0.65	**0.35**	**74.8%**	**0.66**
0.60	0.40	74.3%	0.67
0.50	0.50	73.8%	0.68
0.40	0.60	73.5%	0.69
0.30	0.70	73.4%	0.68

where alpha is a scaling factor that determines the contribution of each model.

$$W_{\text{merged}} = \alpha W_1 + (1 - \alpha) W_2$$

In Table 4, we report the performance of ensemble model with weights range from 0.3 to 0.9. We focus on this range to ensure both models contribute meaningfully, while avoiding degenerate cases where the ensemble collapses to a single model. The results show that, although performance varies across configurations, all ensemble configurations show improved accuracy and lower log loss compared to a single model. Performance peaks when α is between 0.6 and 0.7, with the best observed result at $\alpha = 0.65$; beyond this point, increasing the Qwen weight leads to slight declines in both accuracy and calibration. The best-performing ensemble achieves an accuracy of 74.8%, closely approaching the 77.1% reached by the Qwen2.5-72B baseline. Although the 72B baseline is evaluated zero-shot, this comparison highlights a practical reality: for most organizations, the computational burden of fine-tuning massive models makes the specialized adaptation of smaller, ensembled architectures a more viable pathway to high-performance medical reasoning. This performance gain aligns with a bias–variance perspective: by selecting intermediate α values, the ensemble harmonizes complementary representations, effectively reducing variance without sacrificing beneficial and model-specific biases.

Notably, accuracy remains relatively stable ($\alpha \in [0.3, 0.9]$),

ranging only from 73.1% to 74.8%. This 1.7% variance suggests that the framework's effectiveness stems primarily from combining complementary model strengths rather than hypersensitivity to weight allocation. All ensemble weight configurations listed in Table 4 yield higher accuracy and lower log loss compared with the individual Llama3.1-8B and Qwen2.5-14B models. This consistent improvement suggests that the two fine-tuned models capture complementary patterns; by merging their parameters, the ensemble leverages their combined strengths for superior predictive performance. Beyond accuracy, the ensemble significantly improves calibration, reducing log loss to the 0.66–0.71 range compared to 1.09–1.19 for individual models. This reduction indicates better-calibrated predicted probabilities, ensuring the system produces more reliable confidence estimates alongside its improved accuracy.

4.4 Practical Implications of the Framework

The results demonstrate that combining context-specific tuning, parameter-efficient adaptation, and model ensembling enables small models to achieve competitive performance while retaining practical advantages for real-world deployment.

In terms of performance, fine-tuned small models show substantial improvements over their baselines, and the ensemble further enhances accuracy, reaching 74.8% compared to 77.1% for the 72B model. While a gap remains, this result suggests that targeted domain adaptation and model combination can significantly reduce the reliance on large-scale models.

From an efficiency perspective, the CoDES framework operates with substantially lower computational requirements. As shown in Table 5, the 72B model consumes considerably more energy than the smaller models, while the ensemble of small models remains efficient. This reduction in resource usage is particularly important for applications where cost and energy consumption are key constraints.

This framework also supports faster inference due to the smaller model sizes, which reduces computational overhead during deployment. This makes the approach more suitable for real-time systems or large-scale applications where latency is a critical factor.

Finally, the modular design allows for flexible domain customization. Through context-specific fine-tuning, models can be efficiently adapted to new domains without requiring retraining of large-scale models, making the approach more practical for scenarios where domain knowledge evolves over time.

TABLE 5. Energy consumption comparison for 10K questions.

Model	Energy (kWh)
Qwen2.5-72B	0.20
Llama3.1-8B	0.03
Qwen2.5-14B	0.05
Qwen2.5-14B + Llama3.1-8B	0.08

4.5 Generalization to Additional Dataset

To further validate the effectiveness of the proposed framework, we apply it to the MedQA dataset, a widely used benchmark for medical question. This experiment evaluates whether the performance improvements observed on the primary dataset generalize to another domain-relevant benchmark. The data size is consistent with the previous experiment, with 10k samples for zero-shot evaluation, 10k for training, and 3k for validation. For LoRA fine-tuning and ensembling, only the best-performing configurations are reported.

We applied the same training pipeline and model configurations to MedQA. As shown in Table 6, LoRA-based fine-tuning improves performance over the zero-shot baseline for both models. Llama3.1-8B increases from 60.8% to 68.8% accuracy, while Qwen2.5-14B improves from 61.4% to 66.7%. These gains are accompanied by reductions in log loss, indicating improved prediction calibration.

The ensemble model achieves the best overall performance, reaching 71.6% accuracy and reducing log loss to 0.78. Notably, it narrows the performance gap with the large-scale Qwen2.5-72B model (73.4%).

These results demonstrate that the proposed framework generalizes effectively across datasets. The consistent improvements in both accuracy and log loss indicate that combining parameter-efficient adaptation with model ensembling provides a robust and scalable approach for enhancing small language models under varying data conditions. This approach offers a practical pathway for deploying high-performing language model systems under realistic resource constraints, particularly in knowledge-intensive domains where models must be frequently updated.

5 Conclusion

This work examined whether small language models can achieve strong performance when adapted using domain-specific context in an efficient manner. We proposed CoDES, a framework that combines context-specific fine-tuning, parameter-efficient adaptation, and model ensembling

TABLE 6. Generalization performance of the proposed framework on the MedQA dataset.

Model / Method	Accuracy	Log Loss
Qwen2.5-72B (Baseline)	73.4%	1.09
Llama3.1-8B (Baseline)	60.8%	1.24
Qwen2.5-14B (Baseline)	61.4%	1.22
Llama3.1-8B + LoRA	68.8%	1.14
Qwen2.5-14B + LoRA	66.7%	1.16
Ensemble (Llama + Qwen)	**71.6%**	**0.78**

to improve the performance of compact models.

Experiments on biomedical question answering show that fine-tuned small models, especially when combined through ensembling, can approach the performance of a much larger 72B model while requiring significantly fewer computational resources. These results highlight the potential of targeted adaptation and model combination as an alternative to scaling model size.

While the current evaluation focuses on biomedical question answering, CoDES is designed to generalize to other knowledge-intensive domains. This model-centric approach remains compatible with external modules, such as RAG, which can be integrated into the framework in future work. Subsequent research will explore broader domain adaptation, additional model architectures, and integration with retrieval-augmented methods. Improving model robustness under distribution shift and real-world variability remains an important open challenge [56, 57], and addressing this will be essential for reliable deployment across diverse application settings.

References

[1] T. Shahzad, T. Mazhar, M. U. Tariq, W. Ahmad, K. Ouahada, and H. Hamam, "A comprehensive review of large language models: issues and solutions in learning environments," *Discov. Sustain.*, vol. 6, no. 1, Jan. 2025.

[2] Z. Xu, X. Zhang, Q. Huang, X. Zhou, and J. Zhang, "Avatarshield: Visual reinforcement learning for human-centric synthetic video detection," *arXiv preprint arXiv:2505.15173*, 2025.

[3] Z. Xu, X. Zhang, Y. Xu, Q. Huang, S. Chen, T. Yao, S. Ding, and J. Zhang, "Genshield: Unified detection and artifact correction for ai-generated images," in *International Conference on Machine Learning*. PMLR, 2026.

[4] M. Dauner and G. Socher, "Energy costs of communicating with ai," *Frontiers in Communication*, vol. Volume 10 - 2025, 2025. [Online]. Available: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1572947

[5] X. Li, Y. Ma, Y. Huang, X. Wang, Y. Lin, and C. Zhang, "Synergized data efficiency and compression (sec) optimization for large language models," in *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2024, pp. 586–591, doi: 10.1109/EIECS63941.2024.10800533.

[6] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov, "Small language models are the future of agentic ai," 2025. [Online]. Available: https://arxiv.org/abs/2506.02153

[7] C. Wu, B. Li, M. Gao, and Z. Wang, "From efficiency to adaptivity: A deeper look at adaptive reasoning in large language models," 2025. [Online]. Available: https://arxiv.org/abs/2511.10788

[8] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, and S. Wang, "A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 6, Nov. 2025. [Online]. Available: https://doi.org/10.1145/3768165

[9] I. Lamaakal, Y. Maleh, K. El Makkaoui, I. Ouahbi, P. Pławiak, O. Alfarraj, M. Almousa, and A. A. Abd El-Latif, "Tiny language models for automation and control: Overview, potential applications, and future research directions," *Sensors (Basel)*, vol. 25, no. 5, p. 1318, Feb. 2025.

[10] J. Jiang, P. Yang, R. Zhang, and F. Liu, "Towards efficient large language model serving: A survey on system-aware kv cache optimization," *Authorea Preprints*, 2025. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.176046306.66521015/v3

[11] H. Kim, H. Hwang, J. Lee, S. Park, D. Kim, T. Lee, C. Yoon, J. Sohn, J. Park, O. Reykhart, T. Fetherston, D. Choi, S. H. Kwak, Q. Chen, and J. Kang, "Small language models learn enhanced reasoning skills from medical textbooks," *NPJ Digit. Med.*, vol. 8, no. 1, p. 240, May 2025.

[12] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, and J. Hernández-Orallo, "Larger and

more instructable language models become less reliable," *Nature*, vol. 634, no. 8032, pp. 61–68, Oct. 2024.

[13] M. J. J. Bucher and M. Martini, "Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification," *arXiv preprint arXiv:2406.08660*, 2024.

[14] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[15] J. Xing, D. Luo, C. Xue, and R. Xing, "Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective," 2025. [Online]. Available: https://arxiv.org/abs/2411.14654

[16] C. Xiao, J. Cai, W. Zhao, B. Lin, G. Zeng, J. Zhou, Z. Zheng, X. Han, Z. Liu, and M. Sun, "Densing law of LLMs," *Nature Machine Intelligence*, vol. 7, no. 11, pp. 1823–1833, Nov. 2025.

[17] Y. Wang, R. Wu, X. Li, and D. Kutscher, "Pactrain: Pruning and adaptive sparse gradient compression for efficient collective communication in distributed deep learning," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, 2025, pp. 1–7, doi:10.1109/DAC63849.2025.11133419.

[18] Y. Zhou, Z. Zhao, D. Cheng, Z. Wu, J. Gui, Y. Yang, F. Wu, Y. Cheng, and H. Fan, "Dropping experts, recombining neurons: Retraining-free pruning for sparse mixture-of-experts llms," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 15 169–15 186, doi:10.48550/arXiv.2509.10377.

[19] S. Zeng, D. Qi, X. Chang, F. Xiong, S. Xie, X. Wu, S. Liang, M. Xu, X. Wei, and N. Guo, "Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation," 2026. [Online]. Available: https://arxiv.org/abs/2509.22548

[20] Z. Chen, M. Zhang, X. Yu, X. Luo, M. Sun, Z. Pan, Y. Feng, P. Pei, X. Cai, and R. Huang, "Think with 3d: Geometric imagination grounded spatial reasoning from limited views," 2025. [Online]. Available: https://arxiv.org/abs/2510.18632

[21] R. Luo, X. Chen, and Z. Ding, "Sequda-rec: Sequential user behavior enhanced recommendation via global unsupervised data augmentation for personalized content marketing," *arXiv preprint arXiv:2509.17361*, 2025.

[22] J. Tian, Z. Wang, J. Zhao, and Z. Ding, "Mmrec: Llm based multi-modal recommender system," in *2024 19th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*. IEEE, 2024, pp. 105–110, doi:10.48550/arXiv.2408.04211.

[23] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering," 2022. [Online]. Available: https://arxiv.org/abs/2203.14371

[24] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[25] S. Xu, Z. Ding, Z. Wei, C. P. Yang, Y. Li, X. Chen, and H. Wang, "Beyond keywords: modeling the semantic complexity of deceptive communication on instant messaging platforms," *Journal of Computational Social Science*, vol. 9, no. 2, p. 29, Feb. 2026.

[26] X. Qiu, L. Yu, Y. Zhang, A. Yang, N. Kokhlikyan, N. Cancedda, D. Garcia-Olano *et al.*, "Hallucination reduction with casal: Contrastive activation steering for amortized learning," *arXiv preprint arXiv:2510.02324*, 2025.

[27] M. Hu, Q. Zeng, and L. Luo, "Zero-shot vulnerability detection in low-resource smart contracts through solidity-only training," 2026. [Online]. Available: https://arxiv.org/abs/2603.21058

[28] Z. Chen, M. Hu, J. Yi, and W. Sun, "Reinforcement learning for option hedging: Static implied-volatility fit versus shortfall-aware performance," *arXiv preprint arXiv:2601.01709*, 2026.

[29] W. Sun, Q. Shen, Y. Gao, Q. Mao, T. Qi, and S. Xu, "Objective over architecture: Fraud detection under extreme imbalance in bank account opening," *Computation*, vol. 13, no. 12, p. 290, 2025.

[30] D. Liu, Q. Shen, and J. Liu, "The health-wealth gradient in labor markets: Integrating health, insurance, and social metrics to predict employment density," *Computation*, vol. 14, no. 1, p. 22, 2026, doi:10.3390/computation14010022.

[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[32] T. Zhang, K. Kasichainula, Y. Zhuo, B. Li, J.-S. Seo, and Y. Cao, "Transformer-based selective super-resolution for efficient image refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7305–7313, doi:10.48550/arXiv.2312.05803.

[33] M. Hu, J. Wang, W. Zhao, Q. Zeng, and L. Luo, "Flowmaltrans: Unsupervised binary code translation for malware detection using flow-adapter architecture," 2025. [Online]. Available: https://arxiv.org/abs/2508.20212

[34] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10 088–10 115, 2023.

[35] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[36] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[37] W. Yan, E. Wu, A. G. Schwing, and E. Rosenbaum, "Semantic autoencoder for modeling beol and mol dielectric lifetime distributions," in *2023 IEEE International Reliability Physics Symposium (IRPS)*, 2023, pp. 1–9.

[38] W. Yan, E. Wu, and E. Rosenbaum, "New loss function for learning dielectric thickness distributions and generative modeling of breakdown lifetime," in *2025 IEEE International Reliability Physics Symposium (IRPS)*, 2025, pp. 1–9.

[39] J. Liu, Y. Jiang, G. Zhang, Z. Zhang, H. Chang, Z. Yin, Q. Ren, and J. Yan, "Todoevolve: Learning to architect agent planning systems," 2026. [Online]. Available: https://arxiv.org/abs/2602.07839

[40] M. Long, J. Liu, Y. Li, H. Xiong, J. Yan, K. Wang, Y. Cao, and J. Ding, "Towards practical large-scale dynamical heterogeneous graph embedding: Cold-start resilient recommendation," 2025. [Online]. Available: https://arxiv.org/abs/2512.13120

[41] Q. Shen and J. Zhang, "Mftformer: Meteorological-frequency-temporal transformer with block-aligned fusion for traffic flow prediction," *Research Square*, 2026, preprint, doi:10.21203/rs.3.rs-8770196/v1.

[42] S. Jia and L. Li, "Adaptive masking enhances visual grounding," *arXiv preprint arXiv:2410.03161*, 2024.

[43] Z. Cheng, L. Lai, and Y. Liu, "Resolving the robustness-precision trade-off in financial rag through hybrid document-routed retrieval," 2026. [Online]. Available: https://arxiv.org/abs/2603.26815

[44] Z. Cheng, L. Lai, Y. Liu, K. Cheng, and X. Qi, "Enhancing financial report question-answering: A retrieval-augmented generation system with reranking analysis," 2026. [Online]. Available: https://arxiv.org/abs/2603.16877

[45] S. Yang, S. C. Han, Y. Ding, S. Wang, and E. Hoy, "Tooltree: Efficient llm agent tool planning via dual-feedback monte carlo tree search and bidirectional pruning," *arXiv preprint arXiv:2603.12740*, 2026.

[46] Q. Wang, F. Liu, B. Zhang, J. Liu, F. Xu, and Y. Wang, "Siamctca: Cross-temporal correlation aggregation siamese network for uav tracking," *Drones*, vol. 9, no. 4, p. 294, 2025, doi:10.3390/drones9040294.

[47] S. Zhao, Z. Xue, Y. Qi, X. Zeng, and Z. Yu, "Non-intrusive graph-based bot detection for e-commerce using inductive graph neural networks," 2026. [Online]. Available: https://arxiv.org/abs/2601.22579

[48] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges," *Artificial Intelligence Review*, 2024.

[49] Z. Chen, C. Zhao, M. Zhao, Q. Zhan, Y. Dong, S. Zhao, S. Yu, C. Yao, Y. Xie, Z. Dong, Q. Sun, Y. Liu, C.-H. Yu, H. Yang, and G. Wang, "Rethinking subjective trust in llm: Actualizing tangibility from uncertainty," Dec. 2025. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.176463545.53813864/v1

[50] J. Yang, T. Liu, Y. T. Luo, T. Niu, P. C.-I. Pang, A. Xiang, and Q. Yang, "Exploring the application boundaries of llms in mental health: A systematic scoping review," *Frontiers in Psychology*, vol. 16, p. 1715306, 2025, doi:10.1109/ACCESS.2024.3406469.

[51] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding," *arXiv preprint arXiv:2305.12031*, 2023.

[52] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan, "Qwen2 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2407.10671

[53] Meta AI, "Introducing llama 3.1: Open source ai models," https://ai.meta.com/blog/meta-llama-3-1/, 2024, accessed: 2026-03-07.

[54] E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao, "Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities," *ACM Computing Surveys*, vol. 58, no. 8, pp. 1–41, 2026.

[55] J. Tian and Z. Wang, "Dlrrec: Denoising latent representations via multi-modal knowledge fusion in deep recommender systems," *arXiv preprint arXiv:2512.00596*, 2025.

[56] X. Zhou, Z. Lee, W. Ye, and S. Zhang, "Dynamic knowledge transfer for mitigating spurious correlations in deep learning," *International Journal of Computer Vision*, vol. 134, no. 2, p. 64, 2026.

[57] X. Zhou, M. Zhang, Z. Lee, Y. Hua, W. Ye, F. D. Salim, S. Zhang *et al.*, "Boosting resilience of large language models through causality-driven robust optimization," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.