Instruction Tuning and CoT Prompting for Contextual Medical QA with LLMs

Chenqian Le *, Ziheng Gong , Chihang Wang , Haowei Ni , Panfeng Li , Xupeng Chen

Posted Date: 4 June 2025

doi: 10.20944/preprints202506.0254.v1

Keywords: Large Language Model; Quantized Low-RankAdaptation; Instruction Fine Tuning; Chain of Thought; MedicalQuestion Answering

 doi:10.20944/preprints202506.0254.v1

Article

Instruction Tuning and CoT Prompting for Contextual Medical QA with LLMs

Chenqian Le [1,*], Ziheng Gong [1], Chihang Wang [1], Haowei Ni [2], Panfeng Li [3] and Xupeng Chen [1,3]

[1] New York University, New York, USA
[2] Columbia University, New York, USA
[3] University of Michigan, Ann Arbor, USA
* Correspondence: cl6707@nyu.ed

Abstract: Large language models (LLMs) have shown great potential in medical question answering (MedQA), yet adapting them to biomedical reasoning remains challenging due to domain-specific complexity and limited supervision. In this work, we study how prompt design and lightweight fine-tuning affect the performance of open-source LLMs on PubMedQA, a benchmark for multiple-choice biomedical questions. We focus on two widely used prompting strategies—standard instruction prompts and Chain-of-Thought (CoT) prompts—and apply QLoRA for parameter-efficient instruction tuning. Across multiple model families and sizes, our experiments show that CoT prompting alone can improve reasoning in zero-shot settings, while instruction tuning significantly boosts accuracy. However, fine-tuning on CoT prompts does not universally enhance performance and may even degrade it for certain larger models. These findings suggest that reasoning-aware prompts are useful, but their benefits are model- and scale-dependent. Our study offers practical insights into combining prompt engineering with efficient finetuning for medical QA applications.

Keywords: large language model; quantized low-rank adaptation; instruction fine tuning; chain of thought; medical question answering

1. Introduction

In the modern healthcare landscape, clinicians, researchers, and patients are inundated with an ever-growing volume of biomedical literature, clinical guidelines, and patient records [1,2]. Making timely and accurate decisions often hinges on the ability to extract, interpret, and synthesize complex medical knowledge from this unstructured textual data. Traditional search engines and keyword-based tools struggle to meet the demands of such high-stakes environments, where nuances in language, context, and reasoning can significantly alter clinical outcomes. Moreover, the steep learning curve associated with biomedical jargon makes it challenging for non-experts to engage with medical content effectively.

Recent progress in deep learning has significantly advanced natural language processing (NLP), enabling systems to interpret structured and unstructured data more effectively [3–7]. While early biomedical NLP systems, such as domain-specific transformers like BioBERT [8], brought substantial improvements, they often required large-scale supervised training and lacked flexibility in reasoning and adaptation. In parallel, advances in multimodal deep learning and vision-language alignment—such as image-to-image generation from textual prompts [9] and visual question answering systems [10]—demonstrate the broader potential of neural networks in context-rich, data-driven reasoning environments. Methods for detecting deceptive patterns in human communication [11] and predictive modeling in time series data [2,12] further illustrate how diverse deep learning strategies are being tailored to complex real-world applications, including those in healthcare.

The development of large language models (LLMs) marks a paradigm shift. Trained on massive corpora, LLMs exhibit strong capabilities in language understanding, abstraction, and contextual reasoning [13,14]. These models can perform complex tasks in zero- or few-shot settings and have

 doi:10.20944/preprints202506.0254.v1

shown promise in domains such as financial reporting analysis [15,16], image processing [17–19]. However, the use of LLMs in biomedical contexts raises serious concerns: many are trained on general-domain text, resulting in hallucinations and factual inaccuracies when applied to clinical reasoning. Their lack of transparency further complicates their integration into safety-critical workflows like diagnosis and patient care.

As machine learning techniques evolve, integration with LLMs has fueled next-generation systems that support intelligent reasoning across modalities and domains. Innovations in federated learning for cross-cloud health data privacy [20,21], and confidence-aware real-time detection systems [22] are redefining how models interact with complex data ecosystems. Likewise, work in data augmentation [23–25], ensemble forecasting [**?**], and large-scale scene reconstruction [26–28] underscore the breadth of technical advancements that, while not exclusive to healthcare, offer transferable methodologies applicable to biomedical AI. Within this broader landscape, medical question answering (MedQA) stands out as a high-impact application that demands precise factual grounding, contextual understanding, and explainability—particularly in clinical decision support and patient communication.

Traditional approaches to MedQA relied on smaller, domain-specific models such as BioBERT [8] and BioMedLM [29]. These models offer strong lexical understanding but are often limited in their capacity for reasoning or generalization beyond narrow datasets. The emergence of general-purpose LLMs like GPT-4[30] and open-source alternatives such as LLaMA3[31] and Qwen2.5[32] has changed the paradigm, making it feasible to leverage broad language understanding for specialized biomedical tasks.

Despite their potential, challenges remain in deploying LLMs for high-stakes tasks such as MedQA. First, most models are not aligned to the medical domain by default, leading to hallucinations or vague responses. Second, biomedical datasets are often limited in scale, making full fine-tuning costly and potentially unstable. Finally, the reasoning paths taken by LLMs remain opaque, hindering explainability and user trust.

To address these challenges, two complementary strategies have emerged. First, instruction fine-tuning aligns LLM behavior with task-specific expectations using lightweight supervised learning. Second, Chain-of-Thought (CoT) prompting[33] encourages models to reason explicitly by generating intermediate rationales before answering. Prior work (e.g., Med-PaLM [34]) has shown the promise of such techniques in improving factuality and safety in medical QA, though primarily in closed or proprietary models.

In this paper, we extend these ideas to open-source LLMs and ask:

- Can CoT prompting alone improve performance in biomedical QA under zero-shot conditions?
- How does instruction fine-tuning interact with CoT prompts?
- Are gains from CoT consistent across different model families and sizes?

We explore these questions on the PubMedQA dataset using multiple open-source LLMs, including Llama-3.1-8B, Llama-3.3-70B, and the Qwen2.5 series. Fine-tuning is performed using QLoRA [35] for efficiency. Our findings reveal that while CoT prompts are helpful in some base settings, their benefits after fine-tuning are model-dependent—suggesting the need for more nuanced prompt-model alignment in clinical reasoning tasks.

2. Data Collection and Preprocessing

We use the PubMedQA dataset [36], a widely used benchmark in biomedical question answering. Each instance consists of a medical research question, a context (abstract from PubMed), and four possible answers: `A/B/C/D`.

We created two versions of the input prompt. The standard prompt asks the model to choose one of the options directly, while the Chain-of-Thought prompt encourages intermediate reasoning. Examples are provided below.

 doi:10.20944/preprints202506.0254.v1

Standard Prompt
You are a medical expert. Based on the context provided, answer the question by selecting one option. Provide your final answer as one of the options A, B, or C. Your final response should only be the letter corresponding to your chosen option.
Context: {Context}
Question: Is anorectal endosonography valuable in dyschesia?
Options: A: yes B: no C: maybe
Answer: A

Chain-of-Thought Prompt
You are a medical expert. Based on the context provided, answer the question by selecting one option. First, think step by step about the reasoning for your answer. Then provide your final answer as one of the options A, B, or C. Your final response should only be the letter corresponding to your chosen option.
Context: {Context}
Question: Is anorectal endosonography valuable in dyschesia?
Options: A: yes B: no C: maybe
Think: Let me think through this problem step by step to find the correct answer. {Think}
Answer: A

3. Method

3.1. Framework

Our framework, shown in Fig. 1 supports both standard and CoT prompting. For each model, we perform inference in both base (zero-shot) and instruction fine-tuned (SFT) settings. We train the models using QLoRA for parameter-efficient finetuning and evaluate them using Accuracy and F1 scores on the PubMedQA test set.

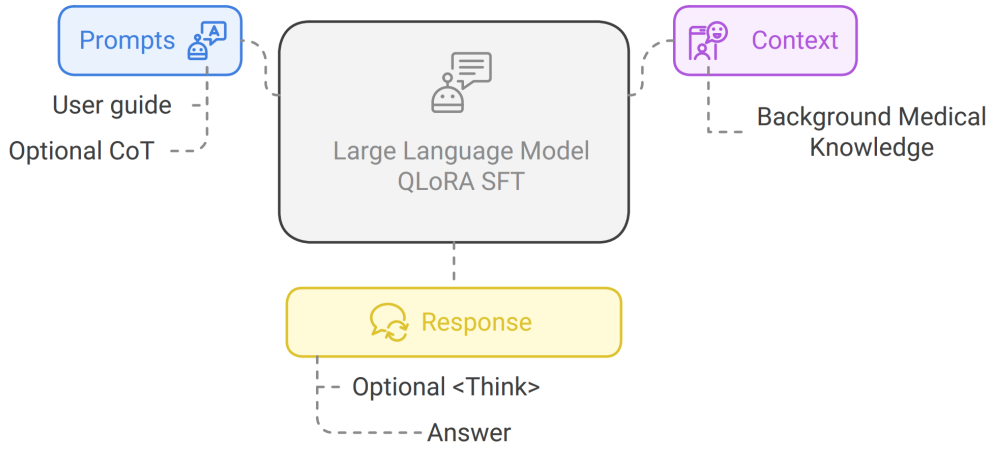

Figure 1. Overview of the framework for medical question answering using LLMs with standard vs. Chain-of-Thought prompting.

3.2. Instruction Fine Tuning

Instruction fine-tuning [37] is a widely adopted strategy for adapting large language models (LLMs) to domain-specific tasks using relatively small but high-quality supervised datasets. It trains LLMs to follow structured instructions and generate consistent, context-aware responses—particularly beneficial in specialized fields like biomedical QA, where accuracy and reasoning alignment are critical.

Our fine-tuning process consists of three main stages. First, we construct an instruction-following dataset by curating and formatting samples from the PubMedQA training set. Each sample is trans-

 doi:10.20944/preprints202506.0254.v1

formed into a structured prompt-response pair, supporting both standard and Chain-of-Thought (CoT) formats. The CoT variant includes intermediate reasoning steps (prefixed by Think:) as part of the supervision signal.

Second, we apply parameter-efficient fine-tuning using QLoRA [35], a method that quantizes model weights to 4-bit precision while learning low-rank adaptation matrices. This allows us to efficiently update large models without incurring the prohibitive memory and compute costs of full fine-tuning—enabling training on a single A100 GPU with long input sequences.

Finally, during inference, the fine-tuned model leverages its learned instruction-following capabilities to generate responses that are better aligned with the task requirements. We evaluate the impact of fine-tuning on both accuracy and reasoning quality, especially when combined with CoT-style prompting. This modular fine-tuning pipeline provides a scalable path to domain alignment without sacrificing efficiency.

3.3. Chain of Thought Prompting

Prompt design is pivotal in steering large language models (LLMs) toward more structured, interpretable, and context-aware reasoning—an essential feature in high-stakes domains such as biomedicine. Among various strategies, Chain-of-Thought (CoT) prompting has proven effective in guiding models to articulate intermediate reasoning steps before producing a final decision, offering both performance improvements and increased transparency.

In this work, we evaluate the utility of CoT prompting by comparing it with standard instruction-style prompts on multiple-choice biomedical question-answering tasks, where the model must select the most appropriate answer from three or four options: `A/B/C/D`. Standard prompts instruct the model to directly output a final choice, while CoT prompts add an explicit reasoning phase via a Think: prefix, encouraging the model to reason through the problem step by step before committing to an answer.

While our evaluation focuses on the final selected option for accuracy and F1 scoring, the generated reasoning chains offer additional insight into the model's internal decision-making process—critical for domains where explainability is valued alongside correctness. Notably, our setup does not include supervised reasoning labels; rather, we leverage the structure of the CoT prompts to induce reasoning behavior in the models.

By systematically comparing both prompt types across various model families and fine-tuning settings, we aim to better understand the benefits and limitations of reasoning-inductive prompting in medical multiple-choice QA.

3.4. QLoRA

Quantized Low-Rank Adaptation (QLoRA) [35] represents a state-of-the-art method for efficiently fine-tuning large models by applying quantization to reduce memory and compute usage. We employ 4-bit QLoRA to fine-tune all models, enabling training on a single A100 with long-context sequences.

3.5. Models and Settings

We evaluate four open-source LLMs: Llama-3.1-8B, Llama-3.3-70B, Qwen2.5-7B, and Qwen2.5-14B. For each model, we perform inference in two settings:

- **Base (Zero-shot)**: Direct inference using the pre-trained checkpoint.
- **Instruction Finetuned (SFT)**: Models fine-tuned on formatted PubMedQA data using QLoRA.

3.6. Prompting Strategies

Two types of prompts are designed:

- **Default Prompt**: A concise instruction asking the model to select one of the three options.
- **CoT Prompt**: Same as Default but with an explicit `Think:` section encouraging step-by-step reasoning.

 doi:10.20944/preprints202506.0254.v1

During finetuning, we supervise both the reasoning and final answer for CoT. During inference, only the final letter answer is used for evaluation.

4. Experiments

4.1. Training Configuration

To adapt open-source LLMs to the biomedical domain efficiently, we employed a parameter-efficient fine-tuning strategy using 4-bit Quantized Low-Rank Adaptation (QLoRA) [35]. This approach significantly reduces memory requirements while maintaining model performance, enabling the fine-tuning of large-scale models on a single A100 GPU.

All models were fine-tuned for one epoch on the PubMedQA training set using a consistent configuration to ensure comparability across experiments. Specifically, the optimization process utilized the 8-bit variant of the AdamW optimizer [38], which balances memory efficiency with convergence stability. The initial learning rate was set to 2×10^{-4}, with a linear scheduler to gradually decay the learning rate over the course of training.

Each training run supported a maximum input sequence length of 25,000 tokens, which is critical for handling the long context passages typical of biomedical abstracts. Due to the high memory demand of long-context fine-tuning, we dynamically adjusted the effective batch size through gradient accumulation. This strategy allowed for stable optimization without exceeding GPU memory limits while still enabling the models to benefit from large-scale sequence input.

While no weight decay or warm-up steps were used in this configuration, we observed that a single epoch of fine-tuning with QLoRA was sufficient to adapt the models effectively to the biomedical QA task. Our pipeline demonstrated that fine-tuning large models with long input contexts is both feasible and effective on limited hardware resources.

4.2. Metrics

We report performance using the following evaluation metrics:

- **Accuracy:** The proportion of correct predictions overall test instances.
- **Weighted F1 Score:** An F1 score weighted by class distribution is particularly important for PubMedQA due to its imbalanced answer types.

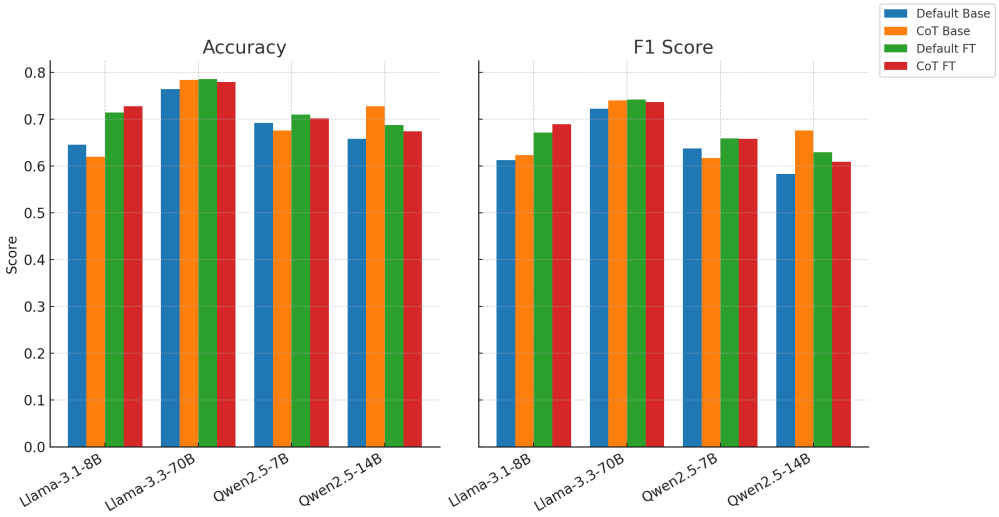

Figure 2. Model performance (Accuracy and F1) across different settings: Default vs. CoT prompt; Base vs. Fine-tuned.

 doi:10.20944/preprints202506.0254.v1

Table 1. Accuracy and F1 scores for each model under Default and CoT prompting, before and after instruction fine-tuning.

Model	Base		Finetune	
	Acc	F1	Acc	F1
Llama-3.1-8B (Default)	0.6460	0.6130	0.7140	0.6716
Llama-3.1-8B (CoT)	0.6200	0.6236	0.7280	0.6891
Llama-3.3-70B (Default)	0.7640	0.7224	0.7860	0.7420
Llama-3.3-70B (CoT)	0.7840	0.7399	0.7800	0.7366
Qwen2.5-7B (Default)	0.6920	0.6372	0.7100	0.6589
Qwen2.5-7B (CoT)	0.6760	0.6173	0.7020	0.6577
Qwen2.5-14B (Default)	0.6580	0.5834	0.6880	0.6298
Qwen2.5-14B (CoT)	0.7280	0.6760	0.6740	0.6087

4.3. Comparative Analysis of Model Performance

Table 1 and Figure 2 summarize the performance of all models under four settings: Base vs. Fine-tuned and Default vs. Chain-of-Thought (CoT) prompting. Three consistent patterns emerge across model families:

1. **Instruction fine-tuning leads to consistent improvements**. All models show performance gains after SFT, with accuracy increases ranging from +1.0% (Qwen2.5-7B CoT) to +8.0% (Llama-3.1-8B Default). This trend is especially notable under default prompting, where fine-tuning provides clearer gains than CoT.
2. **CoT prompting improves F1 in base models**. In 3 out of 4 models (Llama-3.1-8B, Llama-3.3-70B, Qwen2.5-14B), adding CoT to base models increases F1, suggesting that CoT can help models handle ambiguous or nuanced biomedical questions. For instance, Llama-3.3-70B shows an F1 increase from 0.7224 (Default) to 0.7399 (CoT) in the base setting.
3. **CoT + SFT is not always beneficial**. While CoT SFT gives the highest F1 in some settings (e.g., 0.6891 for Llama-3.1-8B), it can hurt performance in others. Notably, Qwen2.5-14B shows a drop in F1 from 0.6760 (Base CoT) to 0.6087 (CoT + SFT), indicating potential misalignment between CoT-style reasoning and instruction-tuned representations for larger models.

4.4. Model-Specific Insights

Llama-3.1-8B.

This mid-sized model exhibits strong gains from both CoT and instruction tuning. Notably, CoT + SFT reaches the highest F1 score among all configurations (0.6891), representing a +7.6% improvement over the base default (0.6130). This suggests that Llama-3.1-8B benefits from explicit reasoning signals when combined with domain adaptation.

Llama-3.3-70B.

The largest model in our study already performs well under default prompting (0.7224 F1) and shows a modest improvement with CoT (0.7399). However, post-finetuning, the gains become negligible or slightly negative: CoT + SFT yields 0.7366, slightly lower than the default SFT at 0.7420. This indicates that the model may already internalize reasoning processes, making additional CoT-style supervision unnecessary or even restrictive.

Qwen2.5-7B.

This model shows stable behavior under both prompting styles, but CoT prompts do not outperform default prompts, neither in base nor in fine-tuned settings. The highest F1 (0.6589) is achieved under default + SFT, while CoT + SFT lags slightly behind (0.6577), suggesting that the added reasoning structure provides marginal utility here.

 doi:10.20944/preprints202506.0254.v1

Qwen2.5-14B.

Interestingly, this model benefits the most from CoT prompting in the base setting (F1 from 0.5834 to 0.6760, a +9.3% jump), but fails to maintain this advantage after instruction tuning (F1 drops to 0.6087). This sharp decline highlights a potential misalignment between the model's internal reasoning and externally imposed CoT formats during SFT.

4.5. CoT: Help or Hurdle?

Our findings suggest that Chain-of-Thought prompting is most helpful in zero-shot or base model scenarios, which scaffolds latent reasoning without altering model weights. However, once models are fine-tuned—particularly under instruction-aligned datasets—the benefit of CoT becomes less predictable. For smaller models like Llama-3.1-8B, CoT and instruction tuning appear synergistic. For larger models, the interaction can be neutral or even detrimental.

These results highlight the importance of aligning reasoning style (e.g., CoT) with model scale and tuning objectives. Over-constraining a model with verbose or unnatural reasoning patterns during fine-tuning may hinder its generalization ability.

5. Conclusion and Future Work

In this study, we explored the impact of prompt design and instruction fine-tuning on the performance of large language models for biomedical question answering. We find that Chain-of-Thought prompting improves zero-shot performance, while instruction fine-tuning further enhances model quality under default prompts. However, the benefit of CoT fine-tuning is not universal and can occasionally degrade performance, especially in large-scale models.

In future work, we aim to explore:

- Multi-stage training with CoT pretraining before full instruction tuning;
- Faithfulness and reasoning quality evaluation for Think: outputs;
- Expansion to real-world clinical tasks with explainability constraints;
- Combining retrieval-based methods (RAG) with CoT prompting for grounded biomedical QA.

We hope this work serves as a foundation for robust and explainable medical LLM systems in high-stakes settings.

References

1. Bian, W.; et al. A Review of Electromagnetic Elimination Methods for low-field portable MRI scanner. In Proceedings of the 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA). IEEE, 2024, pp. 614–618. https://doi.org/10.1109/ICMLCA63499.2024.10753737.
2. Ni, H.; Meng, S.; et al. Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers. In Proceedings of the 2024 6th International Conference on Electronic Engineering and Informatics (EEI). IEEE, 2024, pp. 584–589. https://doi.org/10.1109/EEI63073.2024.10695966.
3. Yang, Z.; Jin, Y.; Zhang, Y.; Liu, J.; Xu, X. Research on Large Language Model Cross-Cloud Privacy Protection and Collaborative Training based on Federated Learning. *arXiv preprint arXiv:2503.12226* **2025**.
4. Zhang, Z.; Qin, W.; Plummer, B.A. Machine-generated text localization. *arXiv preprint arXiv:2402.11744* **2024**.
5. Yang, Z.; Jin, Y.; Xu, X. HADES: Hardware Accelerated Decoding for Efficient Speculation in Large Language Models. *arXiv preprint arXiv:2412.19925* **2024**.
6. Zhang, Z.; Zheng, J.; Fang, Z.; Plummer, B.A. Text-to-image editing by image information removal. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2024, pp. 5232–5241.
7. Jin, Y.; Yang, Z.; Xu, X.; Zhang, Y.; Ji, S. Adaptive Fault Tolerance Mechanisms of Large Language Models in Cloud Computing Environments. *arXiv preprint arXiv:2503.12228* **2025**.
8. Lee, J.; et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**.
9. Ding, Z.; et al. Enhance Image-to-Image Generation with LLaVA-generated Prompts. In Proceedings of the 2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS). IEEE, 2024, pp. 77–81. https://doi.org/10.1109/ISPDS62779.2024.10667513.

 doi:10.20944/preprints202506.0254.v1

10. Li, P.; Yang, Q.; Geng, X.; et al. Exploring Diverse Methods in Visual Question Answering. In Proceedings of the 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). IEEE, 2024, pp. 681–685. https://doi.org/10.1109/ICECAI62591.2024.10674838.
11. Li, P.; Abouelenien, M.; Mihalcea, R.; et al. Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks. In Proceedings of the 2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS). IEEE, 2024, pp. 263–267. https://doi.org/10.1109/ISPDS62779.2024.10667569.
12. Qiu, S.; Wang, Y.; Ke, Z.; Shen, Q.; Li, Z.; Zhang, R.; Ouyang, K. A Generative Adversarial Network-Based Investor Sentiment Indicator: Superior Predictability for the Stock Market. *Mathematics* **2025**, *13*, 1476.
13. Li, P.; Lin, Y.; Schultz-Fellenz, E. Contextual Hourglass Network for Semantic Segmentation of High Resolution Aerial Imagery. In Proceedings of the 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). IEEE, 2024, pp. 15–18. https://doi.org/10.1109/ICECAI62591.2024.10674750.
14. Zhao, P.; Fan, R.; Wang, S.; Shen, L.; Zhang, Q.; Ke, Z.; Zheng, T. Contextual Bandits for Unbounded Context Distributions. *arXiv preprint arXiv:2408.09655* **2024**.
15. Ni, H.; Meng, S.; Chen, X.; Zhao, Z.; et al. Harnessing Earnings Reports for Stock Predictions: A QLoRA-Enhanced LLM Approach. In Proceedings of the 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS). IEEE, 2024, pp. 909–915. https://doi.org/10.1109/DOCS63458.2024.10704454.
16. Meng, S.; Chen, A.; Wang, C.; Zheng, M.; Wu, F.; Chen, X.; Ni, H.; Li, P. Enhancing Exchange Rate Forecasting with Explainable Deep Learning Models. In Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS). IEEE, 2024, pp. 892–896. https://doi.org/10.1109/EIECS63941.2024.10800545.
17. Ding, Z.; et al. Regional Style and Color Transfer. In Proceedings of the 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL). IEEE, 2024, pp. 593–597. https://doi.org/10.1109/CVIDL62147.2024.10604182.
18. Zhang, Z.; He, H.; Plummer, B.A.; Liao, Z.; Wang, H. Complex scene image editing by scene graph comprehension. *arXiv preprint arXiv:2203.12849* **2022**.
19. Shree, A.; Jia, K.; Xiong, Z.; Chow, S.F.; Phan, R.; Li, P.; Curro, D. Image analysis. *US Patent App. 17/638,773* **2022**.
20. Luo, H.; Ji, C. Cross-Cloud Data Privacy Protection: Optimizing Collaborative Mechanisms of AI Systems by Integrating Federated Learning and LLMs. *arXiv preprint arXiv:2505.13292* **2025**.
21. Ji, C.; Luo, H. Cloud-Based AI Systems: Leveraging Large Language Models for Intelligent Fault Detection and Autonomous Self-Healing. *arXiv preprint arXiv:2505.11743* **2025**.
22. Ding, Z.; Lai, Z.; Li, S.; et al. Confidence Trigger Detection: Accelerating Real-Time Tracking-by-Detection Systems. In Proceedings of the 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). IEEE, 2024, pp. 587–592. https://doi.org/10.1109/ICECAI62591.2024.10674884.
23. Yang, Q.; Ji, C.; Luo, H.; et al. Data Augmentation Through Random Style Replacement. In Proceedings of the 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL). IEEE, 2025.
24. Xu, X.; Wang, Z.; Zhang, Y.; Liu, Y.; Wang, Z.; Xu, Z.; Zhao, M.; Luo, H. Style Transfer: From Stitching to Neural Networks. In Proceedings of the 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2024, pp. 526–530.
25. Yang, Q.; et al. A comparative study on enhancing prediction in social network advertisement through data augmentation. In Proceedings of the 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE). IEEE, 2024, pp. 214–218. https://doi.org/10.1109/MLISE62164.2024.10674203.
26. Zhou, Y.; Zeng, Z.; Chen, A.; Zhou, X.; Ni, H.; Zhang, S.; et al. Evaluating Modern Approaches in 3D Scene Reconstruction: NeRF vs Gaussian-Based Methods. In Proceedings of the 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS). IEEE, 2024, pp. 926–931. https://doi.org/10.1109/DOCS63458.2024.10704527.
27. Qiu, S.; Wang, H.; Zhang, Y.; Ke, Z.; Li, Z. Convex Optimization of Markov Decision Processes Based on Z Transform: A Theoretical Framework for Two-Space Decomposition and Linear Programming Reconstruction. *Mathematics* **2025**, *13*. https://doi.org/10.3390/math13111765.
28. Langerman, J.M.; Endres, I.; Rethage, D.; Li, P. Three-dimensional building model generation based on classification of image elements. *US Patent App. 18/703,608* **2025**.

 doi:10.20944/preprints202506.0254.v1

29. Bolton, E.; et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv* **2024**.
30. Achiam, J.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
31. Grattafiori, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
32. Yang, A.; et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* **2024**.
33. Wei, J.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**.
34. Singhal, K.; et al. Toward expert-level medical question answering with large language models. *Nature Medicine* **2025**.
35. Dettmers, T.; et al. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314* **2023**.
36. Jin, Q.; et al. Pubmedqa: A dataset for biomedical research question answering. *arXiv:1909.06146* **2019**.
37. Zhang, S.; et al. Instruction tuning for large language models: A survey. *arXiv:2308.10792* **2023**.
38. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**.