Beyond Single Models: Unsupervised Ensemble Selection for Small Language Models in Medical QA

Nicolas Ventulett, Fabian Nicklas, Eric Gaida, Dieter Wallach, and Jan Conrad

University of Applied Sciences Kaiserslautern
{nicolas.ventulett, fabian.nicklas, eric.gaida, dieter.wallach, jan.conrad}@hs-kl.de

Abstract. Small Language Models (SLMs) provide efficient alternatives to large models for clinical open-ended question answering (QA) but often show variable performance. We propose two unsupervised answer selection strategies for SLM en-sembles: a confidence-based method using normalized perplexity and a consensus-based medoid method capturing semantic similarity among model outputs. Eval-uations on three clinical QA benchmarks show that both strategies outperform single-model and random selection baselines. The results show that unsupervised confidence and consensus mechanisms can enhance the performance of SLM en-sembles for medical QA without requiring additional training or increasing model size.

Keywords: Artificial Intelligence, Large Language Models, Small Language Mod-els, Medical Question Answering, Clinical NLP

1 Introduction

Recent advances in Large Language Models (LLMs) have shown impressive performance in question answering (QA) on clinical documents, enabling patients and clinicians to ef-ficiently extract information from medical records. However, deploying large-scale LLMs in healthcare scenarios remains challenging due to strict privacy regulations and com-putational cost. Consequently, small language models (SLMs) are gaining attention for on-premise or on-edge deployment. While SLMs tend to show greater variability and lower accuracy compared to LLMs, ensemble methods that aggregate the predictions of multiple SLMs offer a promising way to improve performance without increasing model size. This work presents two unsupervised answer selection strategies for SLM ensem-bles in medical QA: a confidence-based method using normalized model perplexity and a consensus-based medoid method capturing semantic similarity among model outputs. Both approaches are evaluated against random selection in ensembles and single-model answering as baselines.

2 Related Work

Ensemble methods have demonstrated improvements in language model performance over various tasks, however most approaches focus on large-scale models or require su-pervised training [1][2]. In the field of medical QA, recent work has achieved expert-level performance through domain-specific fine-tuning and ensemble refinement [3][4]. While ensemble approaches show promising improvement in model performance, unsupervised output selection strategies for SLM ensembles in a clinical context for open-ended QA remain largely unexplored.

3 Methodology

The ensemble consists of six state-of-the-art instruction-tuned SLMs with ≤ 7B param-eters: SmolLM3-3B [5], Qwen3-4B-Instruct [6], Phi-4-mini-instruct (4B) [7], DeepSeek-R1-5B [8], Gemma-3-4B-it [9] and Mistral-7B-Instruct-v0.3 [10]. Each model in the en-semble receives the same structured prompt including a patient discharge summary and a corresponding question to be answered based on the clinical note, resulting in six can-didate answers to select from. We then compare two unsupervised ensemble selection strategies: The first strategy selects the candidate with the highest model confidence, estimated through perplexity scores computed during generation. Since raw perplexity is not comparable across different model architectures, we employ bits-per-byte (BPB) as byte-normalized perplexity variant [11]. The answer with the lowest BPB value is selected as the most confident prediction and final output. The second strategy selects the answer that is semantically most similar to all other ensemble answers. For each candidate, we compute its sentence embedding [12] and measure its average cosine similarity to all other candidates in the ensemble. The answer with the highest average similarity is selected as the final output. This consensus-based strategy selects the most representative (medoid) answer from the ensemble, under the assumption that agreement among diverse models indicates higher reliability.

4 Evaluation

The proposed methods were evaluated on three different benchmark datasets: EHR-DS-QA (506 samples) [13], MeDiSumQA (416 samples) [14] and CliniQG4QA (1287 samples) [15]. These datasets contain human-verified and medically relevant QA-pairs on deiden-tified discharge summaries from MIMIC-III/-IV clinical notes [16][17]. All SLMs and ensemble methods were evaluated using BERTScore F1, capturing semantic similarity between prediction and ground-truth answer [18]. We compared against two baselines: random ensemble selection and individual model performance without ensembling. Across all three datasets, both ensemble methods consistently outperformed random answer se-lection and single-model inference across all six SLMs. The semantic medoid approach achieved higher mean BERTScore F1 across all samples than any individual model, while the BPB-based confidence method showed the best overall performance across all benchmarks. On the EHR-DS-QA dataset, for example, the confidence-based selection improved mean BERTScore F1 by 7.3 points (0.8441 vs. 0.7674) compared to DeepSeek-R1-5B single-model answering.

5 Conclusion

This work demonstrates that unsupervised ensemble selection strategies, confidence-based and consensus-based, can effectively enhance the performance of SLMs in clini-cal open-ended question answering, consistently outperforming single-model and random selection baselines across three benchmark datasets. The results highlight a practical approach for privacy-sensitive healthcare scenarios that require on-premise or on-edge deployment while maintaining a balance between computational efficiency and answer quality. Future research should investigate adaptive or dynamic ensemble selection meth-ods to further reduce computational overhead.

References

1. Ashiga, M., Jie, W., Wu, F., Voskanyan, V., Dinmohammadi, F., Brookes, P., Gong, J., Wang, Z.: Ensemble learning for large language models in text and code generation: A survey (2025)
2. Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., et al.: Meditron-70b: Scaling medical pretraining for large language models (2023)
3. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3) (March 2025) 943–950
4. Lucas, M.M., Yang, J., Pomeroy, J.K., Yang, C.C.: Reasoning with large language models for medical question answering. Journal of the American Medical Informatics Association **31**(9) (September 2024) 1964–1975
5. Bakouch, E., Ben Allal, L., Lozhkov, A., Tazi, N., Tunstall, L., et al.: SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3 (2025)
6. Team, Q.: Qwen3 technical report (2025)
7. Microsoft, :, Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., et al.: Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras (2025)
8. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025)
9. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., et al.: Gemma 3 technical report (2025)
10. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., et al.: Mistral 7b (2023)
11. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., et al.: The pile: An 800gb dataset of diverse text for language modeling (2020)
12. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro-cessing, Association for Computational Linguistics (11 2019)
13. Kotschenreuther, K.: EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Dis-charge Summaries for Enhanced Medical Information Retrieval Systems (2024)
14. Dada, A., Koras, O.A., Bauer, M., Butler, A., Smith, K.E., Kleesiek, J., Friedrich, J.: MeDiS-umQA: Patient-Oriented Question-Answer Generation from Discharge Letters (February 2025) arXiv:2502.03298 [cs].
15. Yue, X., Zhang, X.F., Yao, Z., Lin, S., Sun, H.: CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering (December 2021) arXiv:2010.16021 [cs].
16. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (2024)
17. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., et al.: Phys-ioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23) (2000) e215–e220 Publisher: Am Heart Assoc.
18. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert (2020)