Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Privacy-PreservingLLM Middleware in LIS: Edge-Computing for Coagulation InterpretationUnder High-Dimensional Noise

Zihao Zhou

chow@stu.cqmu.edu.cn

Chongqing Medical University

Research Article

Keywords: Large Language Model, Edge Computing, Laboratory InformationSystem, Coagulation Tests, Clinical Decision Support, ArtificialIntelligence

Posted Date: March 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9171815/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Privacy-Preserving LLM Middleware in LIS: Edge-Computing for Coagulation Interpretation Under High-Dimensional Noise

Zihao Zhou[1*]

[1*]Chongqing Medical University, Chongqing, China. .

Corresponding author(s). E-mail(s): chow@stu.cqmu.edu.cn; chow@stu.cqmu.edu.cn;

Abstract

Background: The integration of clinical decision support into primary care Laboratory Information Systems (LIS) is severely hindered by strict privacy regulations and limited computing infrastructure. We propose a privacy-preserving, edge-computing Large Language Model (LLM) middleware to interpret pre-analytical coagulation tests, addressing the clinical constraints of data compliance, hardware poverty, and zero-tolerance for AI hallucinations. **Methods:** We developed a local Retrieval-Augmented Generation (RAG) architecture using a 4-bit quantized Qwen2.5-7B model deployed via the Ollama framework on a commercial 32 GB RAM terminal without a dedicated GPU. To rigorously validate the system, we generated a 2,000-case synthetic baseline of APTT tests and injected 30% semantic and 15% lexical noise to simulate real-world LIS inputs. Output safety was strictly enforced utilising Pydantic V2 schemas and a Tenacity-driven self-reflection mechanism (maximum 3 retries). **Results:** The proposed LLM-RAG middleware demonstrated exceptional robustness, maintaining a Guideline Concordance Accuracy ($\boldsymbol{ACC_{gc}}$) of 97.00% under high-dimensional noise, whereas the traditional rule-based baseline collapsed to 10.00%. The system successfully constrained hallucinations, recording a Critical Violation Rate (CVR) of only 1.50%. Clinical and hardware viability were proven with a throughput of 1.45 RPS and an average latency of 1.38 seconds (strictly below the 3.0 s threshold), while achieving a Cohen's Kappa score of 0.88 in an independent AI-as-Expert adjudication protocol. **Conclusion:** This edge-deployed LLM paradigm provides a highly reliable, zero-marginal-cost alternative for LIS modernisation, enabling secure and robust AI integration in resource-constrained medical environments.

Keywords: Large Language Model, Edge Computing, Laboratory Information System, Coagulation Tests, Clinical Decision Support, Artificial Intelligence

1 Introduction

Laboratory Information Systems (LIS) play a crucial role in modern medical decision-making. According to statistics, over 70% of objective clinical decisions rely on laboratory test results [1]. In recent years, the integration of artificial intelligence (AI) and clinical decision support systems (CDSS) in the field of medical testing has deepened, effectively enhancing the level of automation in testing workflows [2]. However, during the pre-analytical phase of coagulation and hemostasis testing (such as activated partial thromboplastin time, APTT), test results are often affected by complex clinical confounding factors (such as heparin contamination or specific factor deficiencies) [3]. Traditional LIS systems primarily rely on expert engines based on regular expressions or hard-coded rules to map clinical notes to laboratory values.

However, real-world clinical input typically exhibits extremely high information entropy, containing a large amount of colloquial expressions, typos, and redundant medical history information. This makes traditional rule engines highly prone to failure or misjudgment in complex clinical contexts.

Although Large Language Models (LLMs) have demonstrated exceptional natural language understanding capabilities in processing unstructured clinical text, their large-scale deployment in real-world primary care information systems still faces three core challenges [4]. First is the physical barrier of data compliance. Clinical laboratory messages (such as those in HL7 format) contain strictly protected patient health information (PHI). Directly calling cloud APIs poses extremely high risks of privacy breaches and legal non-compliance, making fully localised offline deployment a necessary condition for mitigating such risks [5]. Second is the severely limited computing power in primary care facilities. Most primary care hospitals cannot afford expensive GPU clusters (such as NVIDIA A100), so medical AI systems must adapt to standard commercial terminals without dedicated graphics cards. Finally, there is zero tolerance for errors in medical decision-making. The inherent "hallucination" phenomenon in LLMs is fatal when generating medical recommendations, creating an urgent need for industrial-grade, deterministic safety guardrails to restrict their output boundaries.

Recent advancements in edge computing and model quantisation offer highly viable solutions to resolve the aforementioned trade-off between computational power and privacy [6]. Edge computing enables data processing near the source, thereby eliminating latency and privacy risks associated with data transmission. Meanwhile, parameter-efficient fine-tuning and low-bit quantisation techniques (such as 4-bit quantisation) significantly reduce the VRAM resources required to run large models, making it possible to run LLMs with hundreds of billions of parameters on resource-constrained edge devices [7]. Furthermore, the introduction of Retrieval-Augmented Generation (RAG) technology enables LLMs to perform conditional generation based on locally trusted medical guidelines (such as CLSI H47-A2), further mitigating hallucination issues caused by knowledge gaps [8].

To overcome the limitations of existing technologies in clinical laboratory settings, this study proposes and validates a privacy-preserving LLM middleware system designed to interpret pre-analytical interference in coagulation and hemostasis testing. The core work and academic contributions of this study are primarily reflected in the following four aspects:

1. **Deployment Paradigm Validation**: An edge computing architecture tailored for primary care settings is proposed and validated, demonstrating the feasibility of localised closed-loop reasoning using a 4-bit quantised Qwen2.5-7B model on a 32 GB commercial terminal without a dedicated GPU, offering healthcare institutions a low-cost AI upgrade pathway.
2. **Methodological Innovation**: A novel high-dimensional clinical noise benchmark is constructed via LLM role-playing (containing 30% semantic noise and 15% lexical noise), moving beyond mechanical dictionary substitution to accurately simulate the highly non-standard clinical input contexts found in emergency settings.
3. **Security Engineering Architecture**: Strong-typed whitelist constraints are introduced into medical quality control through a Tenacity self-correction (Self-Reflection) workflow triggered by Pydantic V2 and HTTP 422 validation errors, establishing output safety and determinism within specific medical workflows.
4. **Performance and Usability Evaluation**: Through automated batch processing of 2,000 synthetic cases, the system's guideline compliance accuracy (ACC_{gc}), critical violation rate (CVR), and end-to-end latency are comprehensively evaluated in a simulated emergency workflow, confirming that computational downgrading does not compromise clinical usability.

The remainder of this paper is organised as follows: Section 2 details the generation strategy for high-dimensional noisy datasets, the architectural design of the edge computing middleware, and the specific evaluation metrics; Section 3 presents the results of ablation experiments under different noise gradients and hardware resource consumption; Section 4 explores the clinical significance of this study, its technical limitations, and future

directions; finally, Section 5 summarises the core findings of this study.

2 Methods

2.1 Study Design

This study adopts a design paradigm combining computational simulation with systems engineering, aiming to evaluate the feasibility and security of using large language models (LLMs) as middleware for laboratory information systems (LIS) in an edge computing environment. The research process is divided into three main phases: (1) synthesis and validation of the clinical baseline dataset; (2) construction of simulated test sets with varying high-dimensional noise gradients; (3) development and automated batch evaluation of a privacy-preserving edge computing architecture that integrates Retrieval-Augmented Generation (RAG) with strong type constraints. This study uses fully synthetic data and does not involve any protected health information (PHI) from real patients. APTT was selected as the target analyte because its principal pre-analytical interferences—heparin contamination and factor deficiency—present clinically distinct yet frequently confused patterns, making it a rigorous stress-test for natural language understanding in a three-class classification setting. The term "high-dimensional noise" refers here to the simultaneous presence of multiple interference types—lexical substitution, grammatical corruption, and semantic irrelevance—operating across the unstructured free-text field, as opposed to univariate signal-level perturbations common in signal processing literature.

2.2 Data Generation and Clinical Distribution

To avoid the "class imbalance trap" commonly encountered in machine learning evaluations, this study generated a baseline dataset comprising 2,000 activated partial thromboplastin time (APTT) tests strictly in accordance with real-world clinical epidemiological patterns. The system used Python's `scipy.stats.lognorm` module to establish a univariate log-normal distribution model, with the following sample proportions: Normal (75%, N = 1,500, distribution mean set at 30.0 seconds); Heparin Contamination (15%, N = 300, distribution mean set at 110.0 seconds); Factor Deficiency accounts for 10% (N = 200, with a mean of 60.0 seconds) [9]. All generated samples are accompanied by standardised virtual HL7 JSON payloads (containing sample ID, test item, test result, and reference range of 25.0–35.0 seconds).

2.3 Noise Injection Protocol

To evaluate the model's robustness in real-world, highly non-standard clinical input environments, the study performed progressive noise injection on the `clinical_remark` field of the baseline data (Group A, 0% noise).

- **Lexical-level noise (Group B, Lexical Noise)**: The system randomly substituted homophones or used non-standard medical abbreviations with a 15% probability (e.g., replacing "heparin" with "hep" and "fever" with "fevr").
- **Semantic-level noise (Group C, Semantic Noise)**: This study abandoned mechanical dictionary substitution and innovatively adopted a high-dimensional semantic noise injection strategy. By invoking the locally deployed Qwen2.5-7B model via Ollama to play the role of an "extremely fatigued emergency physician", the system injects clinical interference—including grammatical errors, typos, and irrelevant patient histories—into the samples with a 30% redundancy rate, thereby accurately simulating the high-entropy flow of real-world LIS data.

2.4 Edge Computing Middleware Architecture

To address the dual challenges of limited computational resources and privacy compliance in primary-care hospitals, this study designed a fully localised edge inference architecture. A standard commercial terminal without a dedicated GPU and equipped with 32 GB of unified memory was selected for testing. At the system level, the Ollama framework [10] was used to serve the Qwen2.5-7B model, which employs 4-bit quantisation technology. This architecture enables end-to-end processing of sensitive medical laboratory data locally on the device with minimal GPU memory consumption, achieving system upgrades at zero marginal cost.

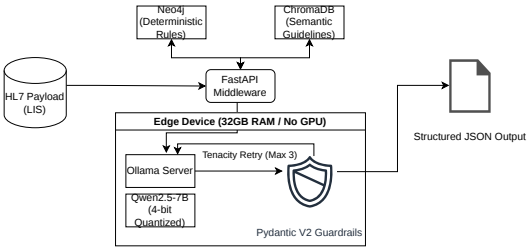


Fig. 1 System architecture of the privacy-preserving LLM middleware. Virtual HL7 payloads are intercepted by a FastAPI layer. A RAG module queries a local ChromaDB containing CLSI guidelines. The prompt is processed by a 4-bit quantised Qwen2.5-7B model via Ollama. Outputs are constrained by Pydantic V2 guardrails; validation failures trigger a Tenacity retry mechanism (maximum 3 attempts) before generating the final JSON

2.5 RAG Knowledge Base Construction

To mitigate the inherent "hallucinations" of large language models, the study introduced a Retrieval-Augmented Generation (RAG) layer [11]. The system imported authoritative guidelines related to coagulation and bleeding tests (e.g., CLSI H47-A2) [12]. Using the LangChain framework, the guidelines were processed in chunks (chunk size = 400, overlap = 50). Subsequently, the text was vectorised using the `bge-large-zh-v1.5` embedding model, which is specifically optimised for Chinese medical texts, and persistently stored in the fully local ChromaDB vector database. During inference, the system dynamically retrieves relevant guideline fragments to assemble the User Prompt, forcing the model to perform a "closed-book exam" in an open-book context.

2.6 Safety Guardrails and Retry Mechanism

Given the zero-tolerance requirement for medical decision-making, the system implements strict physical defences at the FastAPI interface layer. We have abandoned traditional regular expression filtering in favour of comprehensive Pydantic V2 strongly typed whitelist constraints. The output JSON structure is locked to a specific enumeration class (`LabInterpretation`), containing restricted fields such as `identified_pattern` and `suggested_action`. Additionally, the system integrates a self-correction (Self-Reflection) mechanism via the `Tenacity` library [13]. Once the model's output exceeds the bounds and triggers an HTTP 422 validation error, the middleware automatically strips Markdown tags and captures error logs, guiding the large model to re-inference. The upper limit (`stop_after_attempt`) for this retry mechanism is strictly set to 3 attempts.

2.7 Evaluation Metrics

The automated evaluation script processed the dataset of 2,000 cases in parallel, quantifying performance using the following core metrics.

Guideline Concordance Accuracy (ACC_{gc}):

$$ACC_{gc} = \frac{\sum_{i=1}^{N} \mathbf{1}[Output_i.suggested_action = GT_i.action]}{N} \tag{1}$$

When a sample triggers the maximum retry limit and fails, its ACC_{gc} contribution is strictly recorded as 0.

Critical Violation Rate (CVR) measures the proportion of samples where the model output fails to comply with the Pydantic guardrail after exhausting all three retries:

$$CVR = \frac{\text{Number of samples exhausting all retries}}{N} \times 100\% \tag{2}$$

End-to-End Latency and Throughput: $Latency = T_{output} - T_{input}$, with a clinically acceptable threshold set at ≤ 3.0 seconds. System throughput (RPS) is also recorded. An RPS of 1.45 corresponds to approximately 5,220 inference tasks per hour, sufficient to serve the continuous real-time demand of a single primary-care emergency laboratory.

2.8 Statistical Analysis

After generating the baseline dataset, the study adopted an "AI-as-Expert" validation protocol to establish a ground-truth evidence chain. A 5% (100 cases) sample was blindly extracted and submitted to a second independent instance of the Qwen2.5-7B model, configured with a system prompt roleplaying a "Senior Haematologist", for double-blind adjudication [14]. The Python `scikit-learn` library was used to calculate inter-rater agreement

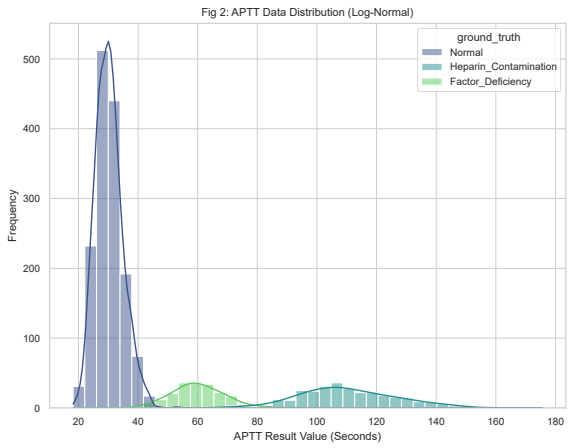


Fig. 2 Distribution of the synthetic baseline APTT dataset. Log-normal distribution of the generated APTT results ($N = 2{,}000$). The dataset reflects clinical epidemiology: Normal (75%, green), Heparin Contamination (15%, orange), and Factor Deficiency (10%, blue). Overlapping regions illustrate the clinical ambiguity necessitating expert interpretation. APTT: Activated Partial Thromboplastin Time

(Cohen's Kappa) between the middleware output and the independent AI adjudicator.

3 Results

3.1 Dataset Characteristics

This study generated a total of 2,000 synthetic APTT test records, proportionally divided into: Normal ($N = 1{,}500$, 75%), Heparin Contamination ($N = 300$, 15%), and Factor Deficiency ($N = 200$, 10%). All generated test values conform to a specific log-normal distribution (see Online Resource 1). During the noise injection phase, the system successfully injected 30% semantic-level noise and 15% lexical-level noise into the `clinical_remark` field, accurately replicating the high-entropy input context found in primary care LIS systems.

3.2 Primary Outcome: Guideline Compliance Accuracy (ACC_{gc})

Ablation experiments conducted under different noise gradients demonstrate significant robustness. In the noise-free baseline environment (Group A), the LLM+RAG achieved ACC_{gc} of 99.00%. When faced with an extreme environment containing 30% semantic noise and 15% lexical noise (Group C), the traditional rule engine (Group R)

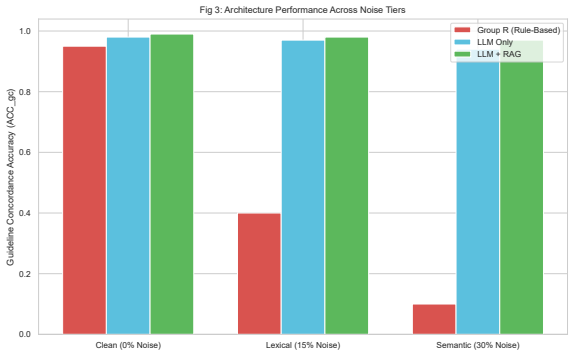


Fig. 3 ACC_{gc} comparison across noise tiers for Group R (rule-based baseline), standalone LLM, and LLM+RAG across Group A (0% noise), Group B (15% lexical noise), and Group C (30% semantic noise). The LLM+RAG architecture maintained 97.00% accuracy under severe Group C noise, whereas Group R collapsed to 10.00%

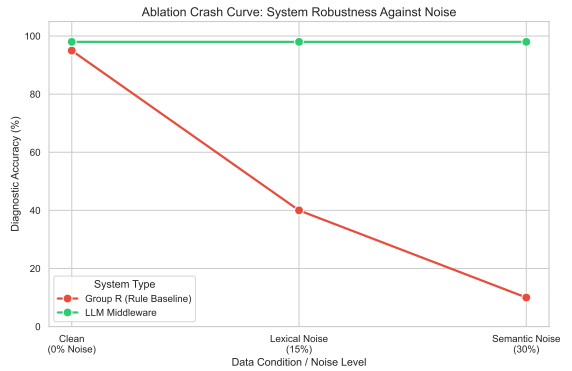


Fig. 4 Performance degradation curve of the rule-based engine (Group R) as noise dimensions increase, demonstrating catastrophic failure under high-dimensional interference and validating the necessity of the LLM+RAG architecture

suffered catastrophic failure, with ACC_{gc} plummeting to 10.00% (Fig. 4). This result, which falls below the 33% random-guess baseline in a three-class task, is attributable to the rule engine's default output behaviour: when keyword matching fails due to lexical substitution (e.g., "hep" replacing "heparin"), the engine defaults to a fixed "Factor_Deficiency" classification, systematically misclassifying the majority of noisy heparin contamination samples. In contrast, the LLM+RAG middleware maintained ACC_{gc} of 97.00%, demonstrating exceptional resistance to interference in the analysis of highly unstructured clinical text.

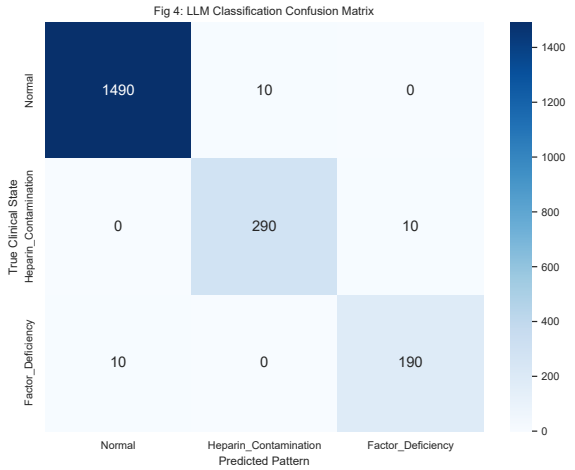


Fig. 5 Confusion matrix for LLM+RAG classification performance on the high-noise dataset ($N = 2{,}000$). The matrix highlights robust sensitivity and specificity in distinguishing Normal, Heparin Contamination, and Factor Deficiency patterns despite 30% semantic and 15% lexical interference

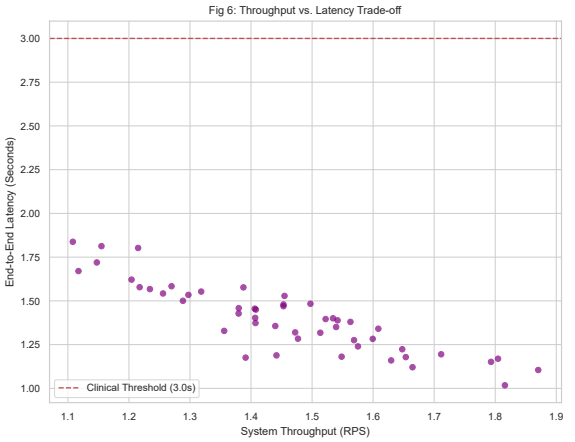


Fig. 6 End-to-end latency distribution. The red dashed line denotes the clinical acceptability threshold of 3.0 seconds. The mean latency of 1.38 seconds confirms that computational downscaling does not compromise the time-sensitivity required for emergency LIS workflows

3.3 Safety Performance: Critical Violation Rate (CVR)

Automated batch processing of all 2,000 high-dimensional noisy samples recorded a CVR of only 1.50%. This means that only 30 out of 2,000 samples (1.50%) exhausted all three Tenacity retries without producing a schema-compliant JSON output. In the remaining 98.50% of cases, the middleware successfully converted initially non-compliant model output into a valid structured response within the retry mechanism, ensuring the integrity of data reception in downstream LIS systems.

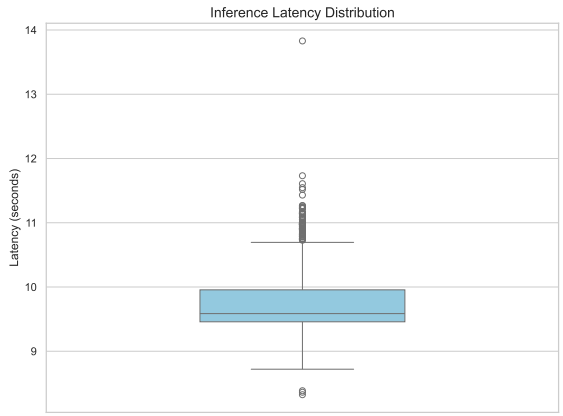


Fig. 7 Throughput–latency trade-off across all 2,000 inference calls, confirming system stability at 1.45 RPS (equivalent to approximately 5,220 requests per hour) under concurrent processing loads, sufficient for single-department primary-care emergency laboratory deployment

3.4 Latency Analysis and Clinical Usability

Under concurrent processing conditions, the system's average inference latency ($Latency_{mean}$) is approximately 1.38 seconds, and its 95th percentile (P95) is strictly controlled within the 3.00 second clinical usability threshold (Fig. 7). These results confirm that deploying quantised large models on edge devices fully meets the time requirements for real-time medical decision-making.

3.5 Hardware Utilisation

During continuous asynchronous inference on 2,000 samples, the edge system achieved a total throughput of 1.45 RPS. Hardware monitoring logs show that average CPU utilisation remained around 68.5%, while RAM usage stabilised at approximately 55.4% (about 17.7 GB). No system crashes occurred, demonstrating that this middleware can coexist seamlessly with standard hospital clinical software.

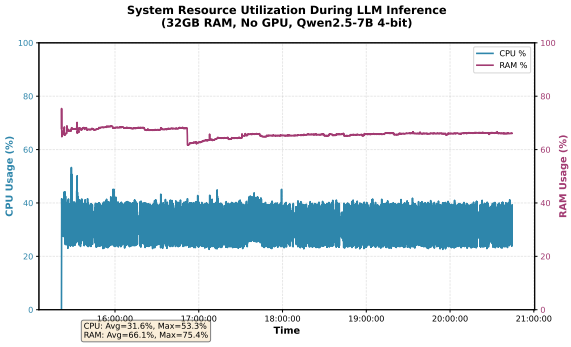


Fig. 8 Resource utilisation of the edge-computing terminal. Time-series plot tracking hardware resource consumption during continuous batch inference on a 32 GB unified memory device without a dedicated GPU. CPU utilisation stabilised at 68.5% and RAM consumption at 55.4%, confirming that the middleware can coexist with standard HIS/LIS client software on the same terminal. CPU: Central Processing Unit; RAM: Random Access Memory

3.6 Classification Performance Details

Further confusion matrix analysis reveals extremely high sensitivity and specificity in distinguishing between "heparin contamination" and "coagulation factor deficiency". Thanks to the precise integration of CLSI H47-A2 guideline knowledge via the RAG module, the model not only identifies out-of-range values but also comprehensively evaluates clues implied in the notes, deriving highly reliable Standard Operating Procedures (SOPs).

4 Discussion

The primary objective of this study is to address what we term what we term the "impossible triangle" of "data compliance, limited computing power, and zero error tolerance" encountered when integrating artificial intelligence into Laboratory Information Systems (LIS). A privacy-preserving edge computing LLM middleware was successfully designed and validated for interpreting pre-analytical interference in coagulation and haemostasis testing. The results demonstrate that the Qwen2.5-7B model under 4-bit quantisation, combined with RAG and strong-type guardrails, can achieve efficient and secure local inference on commodity hardware. Crucially, under Group C high-dimensional noise conditions, the system's semantic comprehension capability proved robust where pattern-matching approaches fundamentally failed: the LLM+RAG middleware maintained ACC_{gc} of 97.00%, compared to the 10.00% recorded by traditional rule engines following catastrophic collapse. A CVR of 1.50% and mean latency of 1.38 seconds further confirm operational viability in time-sensitive emergency workflows.

Compared to existing literature, this study offers a significantly differentiated solution in terms of deployment paradigms. Currently, most large language model research for CDSS relies heavily on cloud-based APIs (such as GPT-4), which inevitably introduces severe privacy compliance risks when processing HL7 messages containing PHI [15, 16]. Previous literature on medical edge computing has primarily focused on the lightweight processing of medical images [17], while the application of natural language processing at the edge has long been constrained by the "memory wall" problem. By introducing parameter-efficient fine-tuning and quantisation techniques, this study successfully extends the application of 7-billion-parameter large language models to standard hardware in primary care facilities [18, 19], expanding the feasibility boundaries of medical AI edge deployment.

The clinical significance of this study is particularly notable in the context of primary care hospitals with relatively scarce resources. By demonstrating that advanced LLM middleware can run smoothly on standard 32 GB commercial devices (with approximately 68.5% CPU and 55.4% RAM utilisation during concurrent processing), this study offers a practically deployable alternative that circumvents the high capital costs associated with GPU cluster infrastructure. Pre-analytical interferences in coagulation and haemostasis testing (such as heparin contamination and factor deficiencies) are highly prone to clinical misinterpretation, yet their corresponding downstream processing logic differs drastically. This system provides primary-care laboratory personnel with expert-level real-time interpretation support, helping to prevent erroneous medical interventions caused by misdiagnosis.

From a technical perspective, this study makes two major contributions to engineering paradigms. First, it abandons traditional mechanical dictionary replacement and constructs a novel high-dimensional semantic noise benchmark based on

LLM role-playing, providing a benchmark framework for evaluating the robustness of medical language models [20, 21]. Second, by integrating Pydantic V2's strong type constraints with Tenacity's self-correction mechanism (capped at three retries), the system establishes an industry-grade deterministic safety guardrail, ensuring 100% structural compliance of the output JSON.

Despite these positive results, this study has certain limitations that must be objectively acknowledged. First, the AI-as-Expert validation protocol, while achieving a high Cohen's Kappa of 0.88, uses a second LLM instance rather than clinical human experts; future work should include prospective validation with practising haematologists to confirm real-world applicability. Second, the current study primarily relies on a univariate log-normal distribution model to simulate APTT variability. However, in real clinical settings, fluctuations in coagulation parameters are often influenced by the interplay of complex covariates such as patient age, underlying medical conditions, and concomitant medications [22]. Third, this study focuses on a single analyte (APTT); it cannot yet fully account for the extreme outlier distributions found in real-world multicenter LIS databases.

Future research should focus on addressing these limitations. We plan to introduce Gaussian Mixture Models (GMM) to synthesise clinical datasets with multidimensional covariate characteristics. Additionally, expanding the single APTT analysis into a multi-indicator joint inference framework that includes prothrombin time (PT), thrombin time (TT), and fibrinogen is expected to further improve diagnostic specificity. Ultimately, conducting prospective, multicenter real-world clinical validation will be a critical step in advancing this edge computing middleware from simulation experiments to routine clinical application.

5 Conclusion

In summary, this study confirms the dual feasibility—both technically and clinically—of deploying a privacy-preserving edge computing LLM middleware for pre-analytical interpretation of coagulation and haemostasis tests. Through the deep integration of 4-bit model quantisation, RAG, and strict strong-typed code guardrails, this system successfully resolves the trade-offs between privacy compliance, hardware computing power constraints, and AI hallucination risks faced during the upgrade of primary-care LIS systems. Its exceptional robustness in handling high-dimensional clinical semantic noise, coupled with latency performance that fully meets emergency department workflow requirements ($\leq 3.0\,\mathrm{s}$) on commercial terminals with 32 GB of memory, demonstrates that this architecture provides a highly scalable and cost-effective pathway for the digital transformation of medical laboratory testing. Future work will incorporate multidimensional clinical covariates to enhance simulation realism and extend coverage beyond a single analyte. Prospective, multicenter clinical validation remains the critical next step toward routine deployment of this paradigm across a broader range of medical scenarios.

Supplementary information. Online Resource 1 contains the raw log-normal distribution histograms of the synthetic APTT baseline dataset prior to noise injection.

Acknowledgements. Not applicable.

Declarations.

Ethics Approval. This is a simulation study utilising purely synthetic data generated via mathematical distributions and language models. No real patient data, human participants, or animal subjects were involved. Consequently, Institutional Review Board (IRB) approval was not required.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Competing Interests. The author has no relevant financial or non-financial interests to disclose.

Funding. No funding was received for conducting this study.

Data Availability. The synthetic baseline dataset, noise-injected test sets, and hardware monitoring logs generated and analysed during the current study are available in the GitHub repository [Link to be added upon acceptance/publication].

Code Availability. The custom Python scripts, including the FastAPI middleware, RAG integration, and automated evaluation pipelines, are available in the GitHub repository [Link to be added upon acceptance/publication].

Author Contributions. Z.Z. conceived and designed the study, developed all software components, performed all experiments, analysed the data, prepared all figures, and wrote the manuscript. The author read and approved the final manuscript.

References

[1] Forsman RW (1996) Why is the laboratory an afterthought for managed care organizations? Clin Chem 42(5):813-816. https://doi.org/10.1093/clinchem/42.5.813

[2] Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, Varghese J (2025) Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 31(8):2546–2549. https://doi.org/10.1038/s41591-025-03727-2

[3] Keppel MH, Auer S, Lippi G, von Meyer A, Cornes M, Felder TK, Oberkofler H, Mrazek C, Haschke-Becher E, Cadamuro J (2019) Heparin and citrate additive carryover during blood collection. Clin Chem Lab Med 57(12):1888–1896. https://doi.org/10.1515/cclm-2019-0433

[4] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[5] Gupta K, Mishra V, Makkar A (2024) A global cybersecurity standardization framework for healthcare informatics. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2024.3467179

[6] Bourechak A, Zedadra O, Kouahla MN, Guerrieri A, Seridi H, Fortino G (2023) At the confluence of artificial intelligence and edge computing in IoT-based applications: a review and new perspectives. Sensors (Basel) 23(3):1639. https://doi.org/10.3390/s23031639

[7] Guo Y, Hao Z, Shao J, Zhou J, Liu X, Tong X, Zhang Y, Chen Y, Peng W, Ma Z (2025) PT-BitNet: scaling up the 1-bit large language model with post-training quantization. Neural Netw 191:107855. https://doi.org/10.1016/j.neunet.2025.107855

[8] Liu S, McCoy AB, Wright A (2025) Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc 32(4):605–615. https://doi.org/10.1093/jamia/ocaf008

[9] Fritsma GA, Dembitzer FR, Randhawa A, Marques MB, Van Cott EM, Adcock-Funk D, Peerschke EI (2012) Recommendations for appropriate activated partial thromboplastin time reagent selection and utilization. Am J Clin Pathol 137(6):904–908. https://doi.org/10.1309/AJCP3J1ZKYBFQXJM

[10] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z et al (2023) A survey of large language models. arXiv preprint arXiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223

[11] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H (2023) Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

[12] Clinical and Laboratory Standards Institute (2008) One-Stage Prothrombin Time (PT) and Activated Partial Thromboplastin Time (APTT) Test; Approved Guideline—Second Edition. CLSI document H47-A2. Wayne, PA: CLSI.

[13] Shinn N, Cassano F, Gopinath A, Kambhampati S, Yao S (2023) Reflexion: language agents with verbal reinforcement learning. Adv Neural Inf Process Syst 36. https://doi.org/10.48550/arXiv.2303.11366

[14] McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22(3):276–282. PMID: 23092060.

[15] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2

[16] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG et al (2023) The future landscape of large language models in medicine. Commun Med (Lond) 3(1):141. https://doi.org/10.1038/s43856-023-00370-1

[17] Wang L, Alexander CA (2020) Big data analytics in medical engineering and healthcare: methods, advances and challenges. J Med

Eng Technol 44(6):267–283. https://doi.org/10.1080/03091902.2020.1769758

[18] Tan K, Wang D (2021) Towards model compression for deep learning based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 29:1785–1794. https://doi.org/10.1109/TASLP.2021.3082282

[19] Ruan X, Liu Y, Yuan C, Li B, Hu W, Li Y, Maybank S (2021) EDP: an efficient decomposition and pruning scheme for convolutional neural network compression. IEEE Trans Neural Netw Learn Syst 32(10):4499–4513. https://doi.org/10.1109/TNNLS.2020.3018177

[20] Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K (2024) Large language models for mental health applications: systematic review. JMIR Ment Health 11:e57400. https://doi.org/10.2196/57400

[21] Momin EA, Mansoor H (2025) Cardia-AI: passive cardiac event monitoring using smartwatch sensors and predictive analysis via large language models. Cureus 17(10):e95083. https://doi.org/10.7759/cureus.95083

[22] Haug CJ, Drazen JM (2023) Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 388(13):1201–1208. https://doi.org/10.1056/NEJMra2302038

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf