Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 280 (2026) 518–525

Procedia Computer Science

www.elsevier.com/locate/procedia

The 17th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 14-16, 2026, Istanbul, Türkiye

An Ontology-Guided Verification Pipeline for Accurate Medical Specialist Classification Using LLMs: Benchmarking Batch and Quantization Effects

Aisvarya Adeseye[a,*], Jouni Isoaho[a]

[a]*Department of Computing, University of Turku, Vesilinnantie 5, Turku 20014, Finland*

Abstract

Patients usually lack the medical knowledge to choose an appropriate specialist to consult, resulting in delayed diagnoses, wasted clinical resources, and unnecessary consultations. While scalable automated specialist selection is possible with large language models (LLMs), instability, hallucination risk, and reliability degradation because of quantization are deployment challenges. Safe usage also requires sensitive data handling that complies with GDPR and HIPAA; locally hosted LLMs help meet the requirement than commercial cloud-based systems. This study proposes an ontology-guided verification pipeline, paired with batch processing and a self-evaluation approach to improve the accuracy of seven LLM families, including both general-purpose and medical-tuned open weight models. The pipeline integrates weighted clinical ontologies, automated ambiguity detection, ensemble voting with contrast validation, confidence calibration, and multi-prompt batch verification to enhance decision robustness. Also, prioritization of lighter-weight quantized local models facilitates real-time operational and efficient resource utilization within e-health ambient environments. Furthermore, these models demonstrate that high performance can be achieved without compromising privacy. The models were evaluated across 4,999 expert-labeled clinical cases. Batch processing improved accuracy by approximately 1–3% for full-precision models and produced substantially larger gains for lighter-weight quantized models, where stabilization effects played a significantly stronger role. Importantly, medical-tune models in the lower-billion parameter range, particularly MedGemma-4B and MediPhi-4B, consistently outperformed larger general-purpose LLMs under both full-precision and quantized inference settings. These findings indicate that domain-specialized models and structured verification can outweigh base model scaling for reliable clinical recommendation systems.

Keywords: Ontology-Guided Verification; Medical Specialist Classification; Large Language Models (LLMs); Edge AI in Healthcare; Batch Processing and Self-Evaluation; Model Quantization; Privacy-Preserving Clinical AI; Ambient Healthcare Systems

* Corresponding author
E-mail address: aisvarya.a.adeseye@utu.fi

1. Introduction

Patients usually lack the required medical knowledge needed to select the appropriate clinical specialist for their symptoms, leading to unnecessary consultations, delayed diagnosis, and inefficient use of healthcare resources. Referral inaccuracies contribute significantly to appointment bottlenecks and increased system burden, particularly within primary-care triage workflows [1, 2]. Large Language Models (LLMs) have demonstrated strong potential for automating textual data interpretation and recommendation tasks [4, 5]; however, their real-world adoption faces several challenges, such as response instability, hallucination risk, and consistency degradation caused by inference optimizations such as quantization [6, 7]. These effects have more impact on lighter-weight models intended for low-cost or real-time deployment [23, 12].

Privacy and regulatory compliance from GDPR and HIPAA further streamline deployment choices. Under these laws, sensitive clinical data cannot be processed in commercial cloud-hosted LLMs because it introduces legal complexity and data governance risks; even with anonymization, there is still a risk of re-identification [3]. Consequently, locally hosted solutions represent a more suitable, secure, and compliant alternative [4, 5] for clinical decision-support systems. However, despite these benefits, resource requirement is a major bottleneck. Larger models are usually quantized to make them lightweight so they can be deployed in resource constrained local environment. Quantization introduces concerns such as computational stability, accuracy loss due to compression, and reasoning reliability. Previous studies show that domain specialization and structured verification methods significantly improve model performance and mitigate hallucination risk compared to scaling model size alone [13].

This study addresses these challenges by proposing an Ontology-Guided Verification Pipeline, augmented with batch processing and self-evaluation mechanisms to systematically improve the accuracy of both general-purpose and medical-tuned LLMs. Particular emphasis is placed on the benchmarking of deployable, lighter-weight quantized models that support resource-efficient and privacy-preserving compliance requirements, while maintaining robust classification performance without sacrificing accuracy. The following are the two main objectives of this study:

1. To introduce and design an ontology-guided verification pipeline for achieving accurate medical specialist classification using LLMs.
2. To evaluate the effects of model types and quantization on classification accuracy, and to study how batch processing and LLM self-evaluation compensate for performance degradation and improved output accuracy.

2. Related Works

Recent studies have evaluated LLMs for general text classification. Trust and Minghim [14] showed that full-precision fine-tuning typically outperforms quantized variants, although gains do not scale linearly with model size. Also, their work is domain-agnostic and does not address medical safety or verification mechanisms. Ontology-guided clinical NLP has also been explored. Mavridis et al. [16] applied LLMs for ontology mapping and medical knowledge graph construction, focusing on RDF generation rather than real-time specialist classification. Also, Subramaniam et al. [17] demonstrated that ontology-guided ML can outperform zero-shot foundation models for cardiac report analysis, highlighting the importance of domain structure. Quantization studies mainly target efficiency and safety. Kharinaev et al. [18] evaluated low-bit quantization schemes and reported stability trade-offs, however, generic benchmarks without clinical verification workflows were used. These previous works treat general classification, ontology feature engineering, or quantization safety independently. Our study converges this by uniquely integrating ontology-guided verification, batch processing, and self-evaluation within a single pipeline to make quantized local LLMs clinically reliable for real-world medical specialist classification.

3. Methodology

Figure 1 presents the overall ontology-guided verification pipeline for medical specialist classification. The pipeline starts with a validated baseline dataset to build domain keyword sets mapped to the clinical ontology. Keywords are weighted by clinical relevance and passed through an ontology compatibility check to ensure semantic alignment. Parallel text-fusion paths with and without preassigned specialists enable comparative validation.

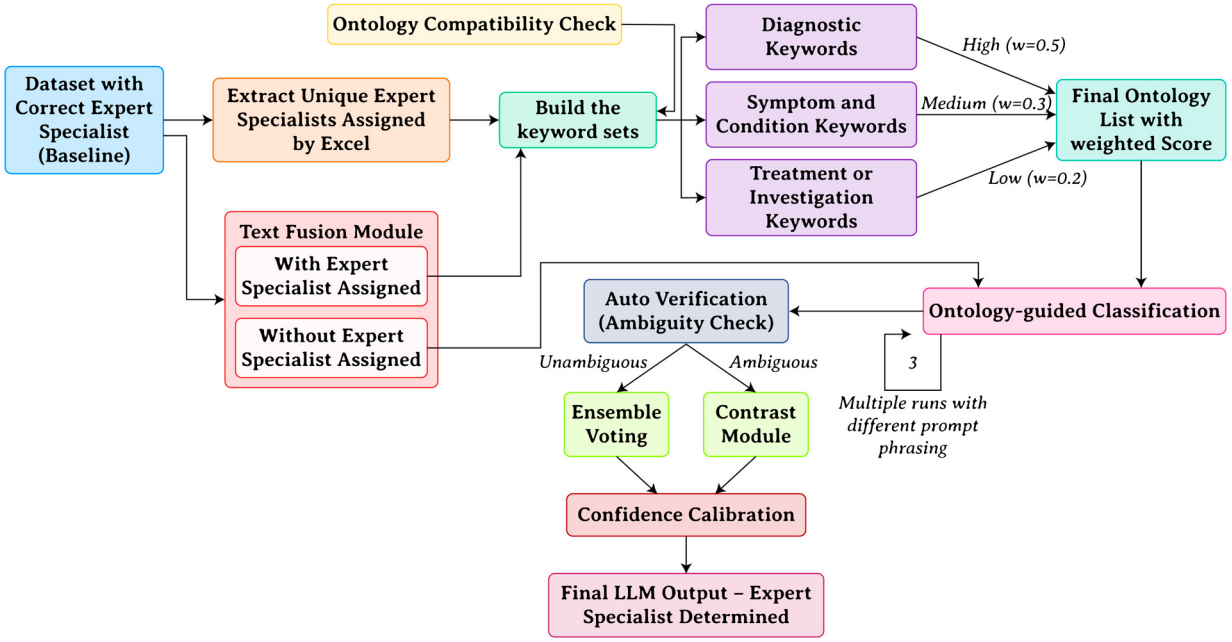

Fig. 1. Ontology-Guided Verification Pipeline for Reliable Medical Specialist Classification

An automated verification stage detects ambiguous outputs, while ensemble voting stabilizes clear cases. Ambiguous predictions undergo contrast prompting across multiple runs to reduce uncertainty. The final outputs are confidence-calibrated and filtered through ontology-guided consistency checks. In general, the pipeline enhances classification reliability, reduces hallucinations, and supports stable, clinically aligned decision-making.

3.1. Key Modules of the Pipeline

3.1.1. Text Fusion

It combines generated outputs with and without assigned specialist labels. This dual-path process helps ameliorate bias introduced by single prompts, capturing both guided and unguided reasoning patterns while increasing the semantic coverage of the input. Also, this module helps detect contradictory interpretations early and thus improves robustness before ontology filtering. In general, it enhances classification reliability and minimizes prompt-dependency errors.

3.1.2. Ontology List Generation

It uses domain knowledge to build a structured keyword list grouped as *diagnostic*, *symptom and condition*, and *treatment* categories. Clinical concept separation is enforced via its structure, while medically significant terms are prioritized via weights. This improves classification semantic clarity and reduces irrelevant or weak signal noise. Generally, this module creates a reliable foundation for ontology-guided reasoning.

3.1.3. Ontology Guided Classification

This module uses weighted category scoring to map extracted keywords to the medical ontology. Each category (grouped as *diagnostic*, *symptoms and conditions*, and *treatment*) is assigned a weight based on its diagnostic reliability. As a result of disease identity being directly linked to diagnostic indicators, the diagnostic category received the highest weight. Furthermore, they were the most reliable discriminators in disease classification and ontology reasoning [19]. Treatment reference was assigned moderate weight because there is a strong association between therapeutic procedures and disease entities. However, it often overlaps across conditions, reducing direct diagnos-

tic specificity [20]. Since symptoms are inherently non-specific and shared across multiple disorders [22], they have weaker discriminative value, so they were assigned the lowest weights.

The foundation of the weighting hierarchy was from clinical decision-making and ontology engineering literature [8]. In literature, emphasis is given on diagnostic evidence over symptomatic cues to improve precision and reduce classification ambiguity [9]. Previous work [23] indicates that medical text classification accuracy increases and semantic drift reduces when differential term weighting is applied. Also, clinical decision support systems widely use ontology-guided weighted reasoning to maintain reliability and logical consistency [11].

Generally, enforcement of weighted constraints reduces mis-classification as a result of vague or overlapping terminology. Moreover, it improves specialist assignment accuracy with both full precision and quantized models, ensuring that outputs clinically reflect validated reasoning and not just surface-level keyword matching.

This module was executed three times per sample using slightly rephrased variants of the prompt, and then outputs from the runs are checked for consistency agreement (self-consistency check). Consistent outputs are deemed reliable if the generated stochastic noise is weak. It also improves output stability, especially for quantized models, strengthening confidence when assigning a specialist.

3.1.4. Verification - Ambiguity Check

Ontology-guided classifications from the previous stage are checked here. Firstly, it checks if the output is Ambiguous or Unambiguous. An Ambiguous output means that the self-consistency check yielded different ontology classification, while an unambiguous output means the results yielded the same classification during the self-consistency check. Ambiguous cases are passed to the *Contrast Module* for deeper evaluation to prevent premature or unstable classification. Also, the unambiguous output is passed to the *Ensemble Voting* to check if more than one expert specialist is matched.

Contrast Module applies contrast prompting to resolve ambiguous cases by explicitly comparing competing specialist labels. It reasons over differences to select the best-aligned option in multiple runs to reduce residual uncertainty. It improves the resolution accuracy of borderline cases by going beyond surface-level matching to deeper analytical evaluation. However, **Ensemble Voting** uses a majority-vote strategy to determine the final label by aggregating predictions from multiple inference outputs. This helps eliminate single-run errors when diverse specialists are matched.

3.2. Techniques

Two complementary inference techniques were adopted by this study to improve classification reliability. They include *Batch Processing* and *Self-Evaluation*, with both used as a post-prompting control mechanism. While batch processing helps stabilize output via controlled repetition, self-evaluation uses boundary testing, contrast reasoning, negation checks, and step-wise validation to strengthen logical verification. Together, they help reduce hallucination and improve consistency with different model sizes and quantization levels.

3.2.1. Batch Processing

This means multiple prompts are run in parallel with the same input. Initially, the batch size was kept small and then progressively increased. During each increase, the accuracy and stability gains were evaluated until results degraded due to batch size changes. Afterwards, the best performing batch size was selected. This method was applied to each model to select an optimal batch size. Larger size models generally had bigger batch sizes than lighter ones. This approach balances robustness gains with computational cost.

3.2.2. Self-Evaluation

This involves the performance of explicit reasoning checks on model output. It uses slightly varied prompts, running it three times to compare the consistency of the output. Only stable agreements are retained as valid predictions. Consequently, this process helps reduce stochastic noise, filtering uncertain reasoning paths, improving stability, especially for lighter-weight quantized models. Generally, it serves as a logical verification layer that improves final decision confidence.

3.3. LLM Models

GPT-5o model was included as a leading large, commercial cloud-based LLM to observe its performance in comparison to local open-source models enhanced with self-evaluation and batch processing. Chosen local models includes LLaMA-8B, LLaMA-3B, LLaMA-1B, Gemma-2B, Phi-1.3B, MedGemma-4B, and MediPhi-4B. This model covers both general-purpose and domain-specialized variants with parameter size ranges of 1B-8B. Hence, this variation enables controlled analysis of model capacity effects on stability, quantization tolerance, and self-evaluation effectiveness. Quantization levels of INT8, INT4, and INT2 were chosen to reflect realistic deployment settings of moderate compression to extreme resource-saving configurations. However, INT2 was applied only to larger size models where stability could be evaluated without collapse, while lighter-weight models were limited to INT4 due to known reliability risks. This design enables evaluation across safe and aggressive compression thresholds. These diverse deployment settings enable robust conclusions about accuracy trade-offs, privacy, computational efficiency, and deployment feasibility across various LLMs.

3.4. Dataset

Medical data is extremely difficult to obtain because of strict patient privacy regulations. Therefore, to prevent privacy violations, the anonymized data used in this study was obtained from this public repository: Kaggle: Medical Transcriptions Dataset. The dataset contains 4,999 medical transcription records across multiple specialties. Each record contains a short patient description, the labeled medical specialty, a sample title, the full clinical transcription text, and extracted keywords. The transcription followed structured clinical formats such as history, medications, assessment, and plan, which reflects real-world documentation practice. The dataset framing helps evaluate the ontology-guided classification, self-evaluation, and quantized inference. Also, the dataset contains both general and specialized clinical domains, making it possible to test across various semantic complexities. Furthermore, the narrative style diversity and specialty overlaps introduce realistic ambiguity, making the dataset useful for assessing the model's verification mechanisms under domain-sensitive conditions, stability, and quantization tolerance.

4. Performance Validation

This section examines the effect of the proposed Ontology-Guided Verification Pipeline, self-evaluation, and batch processing on the performance of quantized LLMs. The baseline condition means the application of the pipeline without self-evaluation and batch processing. Quantization reduces model size and compute cost, but at the expense of accuracy degradation. This study evaluated whether reflective self-evaluation can help offset this loss. The results were compared across general and domain-specific quantized models (INT8 and INT4) with batch inference.

Table 1 shows that self-evaluation consistently improves prediction accuracy across all models, and this effect remains evident regardless of model scale. The gain is larger for lighter-weight models, as reflective checking helps compensate for weaker reasoning capacity, whereas larger models rely less on such support. Batch processing produces a small but stable additional improvement, which suggests a reduction in inference variance across repeated evaluations. Larger models benefit less from these enhancements because their baseline accuracy is already near saturation; hence, the relative margin for improvement is limited. Generally, combining self-evaluation with batch inference offers the best accuracy-to-compute trade-off, and this is particularly important for lighter-weight and domain-specific models, where even modest error reduction yields the greatest practical impact.

Also, table 2 shows that self-evaluation strongly offsets the accuracy loss caused by quantization in LLaMA 8B and 3B models, and this compensatory effect becomes more visible as model precision decreases. Gains increase as precision decreases, because the mechanism is most effective under larger model compression. Batch processing further stabilizes outputs and amplifies recovery, thereby improving result consistency. INT2 and INT4 benefit the most since their baseline errors were the highest. For INT8, improvements were smaller but still consistent. lighter-weight models relied more on self-evaluation for performance recovery, as their reasoning redundancy is limited under quantization. Generally, self-evaluation acts as an error-correction layer that enables aggressive quantization to remain practically viable without leading to severe accuracy collapse.

Table 1. Performance comparison of Full precision LLM predictions with and without self-evaluation (N = 4,999)

Setting	Metric	GPT-5o		LLaMA-8B		LLaMA-3B		LLaMA-3.2-1B		Gemma-3n-2B		MedGemma-4B		Phi-1.5 (1.3B)		MediPhi-4B	
		NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B
No Self-Evaluation	Correct	4,649	4,658	4,645	4,655	4,386	4,402	4,009	4,031	4,256	4,268	4,579	4,590	4,181	4,195	4,562	4,572
	Wrong	350	341	354	344	613	597	990	968	743	731	420	409	818	804	437	427
	Accuracy (%)	***93.00***	***93.18***	***92.92***	***93.11***	***87.74***	***88.08***	***80.18***	***80.62***	***85.13***	***85.37***	***91.59***	***91.81***	***83.64***	***83.91***	***91.25***	***91.45***
With Self-Evaluation	Correct	4,690	4,709	4,695	4,716	4,498	4,531	4,095	4,131	4,357	4,387	4,723	4,748	4,260	4,287	4,706	4,732
	Wrong	309	290	304	283	501	468	904	868	642	612	276	251	739	712	293	267
	Accuracy (%)	***93.82***	***94.20***	***93.92***	***94.34***	***90.00***	***90.63***	***81.92***	***82.63***	***87.16***	***87.76***	***94.45***	***94.98***	***85.20***	***85.75***	***94.16***	***94.66***

Note: NB means No Batch Processing was used; B means Batch Processing was used.

Table 2. Performance of quantized LLaMA models with and without self-evaluation (batch processing) (N = 4,999)

Setting	Metric	LLaMA-8B						LLaMA-3B						LLaMA-3.2-1B			
		INT8		INT4		INT2		INT8		INT4		INT2		INT8		INT4	
		NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B
No Self-Evaluation	Correct	4,219	4,238	3,962	4,021	3,561	3,652	3,889	3,924	3,441	3,512	3,014	3,096	3,739	3,772	3,468	3,522
	Wrong	780	761	1,037	978	1,438	1,347	1,110	1,075	1,558	1,487	1,985	1,903	1,260	1,227	1,531	1,477
	Accuracy (%)	***84.38***	***84.76***	***79.25***	***80.42***	***71.23***	***73.05***	***77.79***	***78.47***	***68.83***	***70.24***	***60.29***	***61.93***	***74.80***	***75.46***	***69.38***	***70.45***
With Self-Evaluation	Correct	4,432	4,457	4,286	4,331	4,011	4,071	4,106	4,183	3,783	3,899	3,382	3,541	3,902	3,978	3,676	3,739
	Wrong	567	542	713	668	988	928	893	816	1,216	1,100	1,617	1,458	1,097	1,021	1,323	1,260
	Accuracy (%)	***88.66***	***89.15***	***85.74***	***86.63***	***80.23***	***81.45***	***82.14***	***83.68***	***75.72***	***78.00***	***67.65***	***70.83***	***78.06***	***79.58***	***73.53***	***74.80***

Note: NB means No Batch Processing was used; B means Batch Processing was used.

Similarly, in table 3, we show that quantization-induced performance loss was substantially reduced with self-evaluation across all Gemma and Phi model families. This effect is consistently observed across all precision levels. The improvement for INT4 settings is the strongest because reflective reasoning compensates for heavily constrained numerical precision. Domain-specific models (MedGemma and MediPhi) recovered more effectively than general models, indicating that domain-specialized model enhances self-evaluation. Smaller gains and prediction stabilization that improved reliability were observed with batch processing. Quantized lighter-weight models like Phi 1.5 depended heavily on self-evaluation to remain useful because of limited baseline resilience. Generally, reliability is preserved after aggressive compression, especially for domain-specific models, due to self-evaluation, which acts as a robustness layer.

Table 3. Performance of quantized Gemma, MedGemma, Phi, and MediPhi models with and without self-evaluation via batch processing (N = 4,999)

Setting	Metric	Gemma-3n-2B FP				MedGemma-4B				Phi-1.5 (1.3B)				MediPhi-4B			
		INT8		INT4		INT8		INT4		INT8		INT4		INT8		INT4	
		NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B
Without Self-Evaluation	Correct	4,049	4,078	3,709	3,774	4,379	4,412	4,039	4,088	3,989	4,023	3,639	3,703	4,339	4,372	3,989	4,042
	Wrong	950	921	1,290	1,225	620	587	960	911	1,010	976	1,360	1,296	660	627	1,010	957
	Accuracy (%)	***80.98***	***81.55***	***74.18***	***75.50***	***87.59***	***88.24***	***80.79***	***81.78***	***79.79***	***80.46***	***72.79***	***74.06***	***86.79***	***87.44***	***79.80***	***80.84***
With Self-Evaluation	Correct	4,267	4,298	3,984	4,014	4,576	4,608	4,281	4,324	4,204	4,238	3,913	3,954	4,563	4,588	4,257	4,293
	Wrong	732	701	1,015	985	423	391	718	675	795	761	1,086	1,045	436	411	742	706
	Accuracy (%)	***85.35***	***85.97***	***79.67***	***80.29***	***91.53***	***92.18***	***85.63***	***86.50***	***84.11***	***84.77***	***78.27***	***79.11***	***91.29***	***91.78***	***85.16***	***85.87***

Note: NB means No Batch Processing was used; B means Batch Processing was used.

5. Discussion

All results in this study were obtained using the Ontology-Guided Verification Pipeline, which demonstrated that GPT-5o achieved very high baseline accuracy (with no self-evaluation or batch processing). However, local models were not far behind as they achieved comparable performance once self-evaluation and batch processing were applied. In several cases, domain-specialized local models even surpassed ChatGPT, demonstrating that targeted training combined with reflective inference can outperform large general-purpose commercial cloud systems. Hence, privacy, regulatory, and cost risks of using such commercial systems are reduced. Under regulations such as GDPR (EU) and HIPAA (US), transferring sensitive data to external servers is legally restricted or prohibited. This makes local deployment a safer and more compliant choice.

The results confirm that investing effort in optimizing local models is worthwhile. Fine-tuning significantly improves both accuracy and stability across model families as seen in the medically-tuned models. LLaMA models at 8B, 3B, and 1B showed strong base accuracy, but stability decreases as size shrinks and quantization increases. The 8B variant remains the most stable, while 3B and 1B became more sensitive under INT4 compression. Only the larger LLaMA models were evaluated with INT2, where heavy quantization produced sharp accuracy drops that self-evaluation only partially recovered. Also, the general Gemma and Phi models achieve strong accuracy and show clear gains after domain teaching. Consequently, the medical-domain models (MedGemma and MediPhi) outperformed their general counterparts, with Overall Recovery Value (OVR), the net post-quantization performance recovery reaching nearly twice that of general models (OVR = 2x). This highlights the importance of task-specific training and alignment over relying on larger, resource-intensive models, demonstrating that targeted optimization can be more effective than base model scaling to achieve high performance. From a resource-efficiency perspective, aggressive Compression Ratios (CR), representing the extent of quantization applied (for example, INT8 → INT4 → INT2), strongly influenced the benefit of self-evaluation. Higher CR decreases memory and computational accuracy. INT4 provides the best accuracy-efficiency trade-off, while INT2 maximizes efficiency at the cost of stability, making it suitable only for large, well-trained models. Smaller heavily quantized models tend to repeat systematic errors, limiting ensemble diversity and reducing batching gains.

Overall, lighter-weight local models finetuned to a specific domain have high potential usage in ambient systems. With careful prompt engineering, batch processing, targeted teaching, and a self-evaluative approach, they could reach competitive accuracy with far lower computational cost while ensuring privacy protection and regulatory compliance. This positions fine-tuned lighter-weight LLMs as the most practical choice for secure, scalable, and privacy-sensitive real-world deployment.

6. Conclusion

This study demonstrated that local lighter-weight LLMs, when supported by ontology-guided verification, self-evaluation, and controlled batch processing, can achieve performance comparable to, and in some cases exceeding, commercial cloud models such as GPT-5o. Results confirm that domain specialization outweighs base model size scaling, as medical-tuned models in the low-billion parameter range consistently outperformed larger general LLMs under both full-precision and quantized settings. Self-evaluation proved to be a critical robustness layer, particularly in mitigating quantization-induced accuracy loss, while batch processing provided additional stability gains with minimal computational overhead. The pipeline supports privacy-preserving deployments compliant with GDPR and HIPAA, enabling real-time, secure clinical integration without reliance on cloud inference.

Future work will extend this pipeline to real clinical triage workflows, incorporating prospective validation with live electronic health record data. Additional studies will explore adaptive quantization strategies, dynamic batch sizing, and reinforcement learning guided prompt optimization to further improve stability under extreme compression. Finally, extending the ontology across multi-specialty and cross-lingual datasets will support broader deployment in international and resource-constrained healthcare environments.

References

[1] Walderhaug, Karina Ellingsen, Marie Kaltenborn Nyquist, and Bente Prytz Mjølstad (2022) "GP Strategies to Avoid Imaging Overuse: A Qualitative Study in Norwegian General Practice," *Scandinavian Journal of Primary Health Care* **40**(1): 48–56.

[2] Marco-Ibáñez, Almudena, Isabel Aguilar-Palacio, and Carlos Aibar (2023) "Does virtual consultation between primary and specialised care improve healthcare quality? A scoping review of healthcare quality domains assessment," *BMJ Open Quality* **12**(4): e002388.

[3] Nunan, D., and M. Di Domenico (2016) "Exploring Reidentification Risk: Is anonymisation a promise we can keep?," *International Journal of Market Research* **58**(1): 19–34.

[4] Adeseye, Aisvarya, Jouni Isoaho, and Mohammad Tahir (2025) "Systematic Prompt Framework for Qualitative Data Analysis: Designing System and User Prompts," in *Proceedings of the 2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS)*, Abu Dhabi, UAE, pp. 229–234.

[5] Adeseye, Aisvarya, Jouni Isoaho, and Tahir Mohammad (2025) "LLM-Assisted Qualitative Data Analysis: Security and Privacy Concerns in Gamified Workforce Studies," *Procedia Computer Science* **257**: 60–67.

[6] Tripathi, Shivangi, Henry Griffith, and Heena Rathore (2024) "Assessing Hallucination in Large Language Models Under Adversarial Attacks," in *Proceedings of the 2024 Ninth International Conference on Mobile and Secure Services (MobiSecServ)*, pp. 1–6.

[7] Yu, Jiangyong, Sifan Zhou, Dawei Yang, Shuoyu Li, Shuo Wang, Xing Hu, Chen Xu, Zukang Xu, Changyong Shu, and Zhihang Yuan (2025) "MQuant: Unleashing the Inference Potential of Multimodal Large Language Models via Static Quantization," in *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, Dublin, Ireland, pp. 1783–1792.

[8] Tarus, J. K., Z. Niu, and G. Mustafa (2018) "Knowledge-based recommendation: A review of ontology-based recommender systems for e-learning," *Artificial Intelligence Review* **50**(1): 21–48.

[9] Meyer, A. N. D., T. D. Giardina, L. Khawaja, and H. Singh (2021) "Patient and clinician experiences of uncertainty in the diagnostic process: Current understanding and future directions," *Patient Education and Counseling* **104**(11): 2606–2615.

[10] Tavabi, Nazgol, James Pruneski, Shahriar Golchin, Mallika Singh, Ryan Sanborn, Benton Heyworth, Assaf Landschaft, Amir Kimia, and Ata Kiapour (2024) "Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline," *Artificial Intelligence in Medicine* **151**: 102847.

[11] Wu, D., J. An, S. Nan, Y. She, H. Duan, and N. Deng (2025) "A knowledge-based clinical decision support system for personalized health examination items in China: Design and evaluation," *BMC Medical Informatics and Decision Making* **25**(1): 183.

[12] Adeseye, Aisvarya, Jouni Isoaho, Seppo Virtanen, and Mohammad Tahir (2025) "Efficient Prompt Design for Resource-Constrained Deployment of Local LLMs," in *Proceedings of the 2025 IEEE Nordic Circuits and Systems Conference (NorCAS)*, Riga, Latvia, pp. 1–7.

[13] Jing, X., H. Min, Y. Gong, P. Biondich, D. Robinson, T. Law, C. Nohr, A. Faxvaag, L. Rennert, N. Hubig, and R. Gimbel (2023) "Ontologies Applied in Clinical Decision Support System Rules: Systematic Review," *JMIR Medical Informatics* **11**: e43053.

[14] Trust, Paul, and Rosane Minghim (2024) "A Study on Text Classification in the Age of Large Language Models," *Machine Learning & Knowledge Extraction* **6**(4): 2688.

[15] Garla, Vijay N., and Cynthia Brandt (2012) "Ontology-guided feature engineering for clinical text classification," *Journal of Biomedical Informatics* **45**(5): 992–998.

[16] Mavridis, Apostolos, Stergios Tegos, Christos Anastasiou, Maria Papoutsoglou, and Georgios Meditskos (2025) "Large language models for intelligent RDF knowledge graph construction: Results from medical ontology mapping," *Frontiers in Artificial Intelligence* **8**: 1546179.

[17] Subramaniam, Suganya, Sara Rizvi, Ramya Ramesh, Vibhor Sehgal, Brinda Gurusamy, Hikmatullah Arif, Jeffrey Tran, Ritu Thamman, Emeka C. Anyanwu, Ronald Mastouri, G. Burkhard Mackensen, and Rima Arnaout (2025) "Ontology-guided machine learning outperforms zero-shot foundation models for cardiac ultrasound text reports," *Scientific Reports* **15**(1): 5456.

[18] Kharinaev, Artyom, Viktor Moskvoretskii, Egor Shvetsov, Kseniia Studenikina, Mikhail Bykov, and Evgeny Burnaev (2025) "Investigating the Impact of Quantization Methods on the Safety and Reliability of Large Language Models," *arXiv preprint*.

[19] Cimino, James J. (1998) "Desiderata for controlled medical vocabularies in the twenty-first century," *Methods of Information in Medicine* **37**(04/05): 394–403.

[20] Witte, René, Thomas Kappler, and Christopher J. O. Baker (2007) "Ontology Design for Biomedical Text Mining," in Christopher J. O. Baker and Kei-Hoi Cheung (eds.), *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, Boston, MA: Springer US, pp. 281–313.

[21] Kesiku, Cyrille YetuYetu, Andrea Chaves-Villota, and Begonya Garcia-Zapirain (2022) "Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review," *Information* **13**(10): 499.

[22] Kesiku, Cyrille YetuYetu, Andrea Chaves-Villota, and Begonya Garcia-Zapirain (2022) "Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review," *Information* **13**(10): 499.

[23] Tavabi, Nazgol, James Pruneski, Shahriar Golchin, Mallika Singh, Ryan Sanborn, Benton Heyworth, Assaf Landschaft, Amir Kimia, and Ata Kiapour (2024) "Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline," *Artificial Intelligence in Medicine* **151**: 102847.