

Uncertainty Guardrails for LLM-Based Medical Diagnostic Systems

Gleb Kuzmin^{1,7,9*} Lyudmila Rvanova^{1*} Iliia Kopanichuk^{2,4} Vladimir Makharev^{2,5}
Ivan Bakulin² Vladimir Shaposhnikov^{2,8} Artem Oskin² Petr Anokhin^{2,6}
Aleksandr Nesterov² Artem Vazhentsev³ Alexander Panchenko^{8,2}
Elena Tutubalina^{2,10} Iaroslav Radionovich Beshpalov² Dmitry V. Dylov^{2,8}
Artem Shelmanov^{3†} Ivan Oseledets^{2,6†}

¹FusionBrain Lab ²AXXX ³MBZUAI ⁴Moscow Center for Advanced Studies
⁵Innopolis University ⁶Lomonosov Moscow State University ⁷RUDN ⁸Applied AI Institute
⁹ISP RAS Research Center for Trusted Artificial Intelligence ¹⁰HSE University
kuzmin.gyu@gmail.com artem.shelmanov@mbzuai.ac.ae

Abstract

Large language models (LLMs) are increasingly being adopted in clinical practice, e.g., in medical diagnostic systems. However, their reliability is still a concern, as LLMs are prone to hallucinations and could lead to misdiagnosis, inappropriate treatment decisions, and potential harm to patients. To mitigate these risks, LLM-based medical diagnostic systems should incorporate safety guardrails capable of detecting suspicious predictions and flagging them for human review or further clinical assessment. In this work, we employ uncertainty quantification (UQ) as a safety mechanism within an enterprise diagnostic system. We evaluate 16 UQ methods and propose two task-specific modifications tailored to the medical diagnostic workflow. First, we show that methods based on semantic similarity between generated samples substantially benefit from replacing a general-purpose NLI model with a domain-specific model. Second, we adapt a supervised UQ method that relies on LLM internal states for the detection of erroneous generation from a black-box enterprise LLM accessible only through a text-generation API. Our approach leverages the internal states of an open-source, lightweight proxy white-box model. We demonstrate that this approach outperforms all other UQ methods in the considered setting.¹

1 Introduction

Large language models (LLMs) are increasingly being integrated into medical question answering and clinical decision-support systems (Tang et al., 2024; Bao et al., 2023). Despite their strong linguistic fluency, these systems remain prone to generating responses that appear confident even when they are

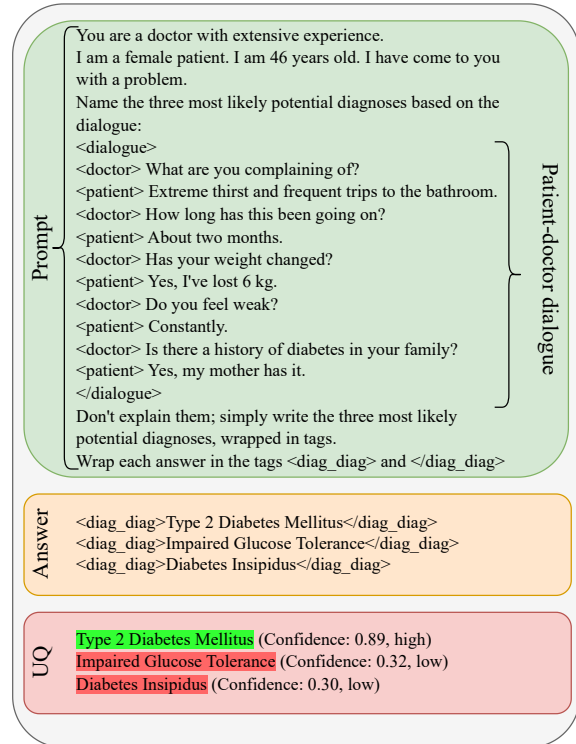


Figure 1: Example of a dialog between a patient and LLM in diagnostic system with the proposed safety mechanism. Candidate diagnoses from the generated response are assigned confidence scores, highlighting uncertain outputs for physician review.

factually incorrect or unsupported by medical evidence – a phenomenon commonly referred to as hallucination (Xu et al., 2024; Huang et al., 2025). Such errors typically arise from two main sources: (1) limitations in the model’s domain knowledge, particularly in specialized or rapidly evolving medical fields, and (2) ambiguity or complexity in clinical queries that require multi-step diagnostic reasoning or additional patient-specific evidence. In clinical settings, these failure modes pose substantial safety risks, as fluent medical language can obscure incorrect or unsupported diagnostic claims while providing no explicit indication of

*Equal contribution

†Equal supervision

¹<https://github.com/glkuzi/uncertainty-guardrails-medical-llms>

uncertainty. This misalignment between expressed confidence and true predictive reliability may cause clinicians to over-trust model outputs, thereby influencing clinical reasoning and decision-making. As a result, hallucinated or unreliable responses may lead to misdiagnosis, inappropriate treatment decisions, delayed interventions, and potentially irreversible harm to patients. Consequently, LLM-based medical systems require dedicated safety mechanisms capable of detecting unreliable predictions and explicitly communicating uncertainty to human experts (Wang et al., 2025).

One approach to constructing such guardrails is Uncertainty Quantification (UQ) (Malinin and Gales, 2021; Kuhn et al., 2023; Farquhar et al., 2024; Vashurin et al., 2025). Uncertainty is a fundamental concept in machine learning that captures the inherent incompleteness of our knowledge about model parameters, data distributions, and predictions (Houlsby et al., 2011; Gal and Ghahramani, 2016; Hüllermeier and Waegeman, 2021). UQ represents a principled framework for estimating this missing information, building on the idea that ML models can emit signals reflecting the reliability of their predictions alongside the predictions themselves (Zhang et al., 2019; He et al., 2020; Xin et al., 2021; Duan et al., 2024). Many UQ methods can be deployed without substantial architectural changes or fine-tuning of the underlying LLMs. This is particularly important in enterprise AI systems, where deployed models are often closed-source black boxes that cannot be modified (Lin et al., 2024; Manakul et al., 2023).

In this work, we develop a UQ-based safety mechanism for an enterprise medical AI diagnostic system that produces diagnosis-level confidence scores and highlights potentially unreliable statements for physician review (see Figure 1). The mechanism operates as a wrapper around an existing diagnostic system without modifying the underlying model and augments only the visualization of its outputs. We first experiment with various UQ methods and adapt consistency-based UQ techniques to the clinical domain by using a specialized NLI model to compute similarity between generated diagnoses. While this adaptation proves useful, we observe that supervised UQ (Shelmanov et al., 2025) for white-box LLMs still provides a substantial advantage. To support proprietary black-box LLMs accessible only through text-generation APIs, we further propose a supervised UQ method based on a lightweight proxy white-box model.

The proxy model exposes internal representations that can be used to learn uncertainty signals aligned with the black-box model’s outputs, enabling reliable confidence estimation even when the proprietary model’s internal states are unavailable. This makes the proposed safety mechanism practical for real-world enterprise systems where direct access to proprietary models is often restricted.

Our main contributions are as follows:

- We conduct a comprehensive empirical study of UQ methods for building safety guardrails in enterprise medical AI diagnostic systems.
- We show that consistency-based UQ methods and other techniques relying on semantic similarity between generations benefit substantially from domain-specific NLI models.
- We adapt the supervised UQ method **UHead** to the black-box setting using a proxy white-box LLM. The resulting method achieves the best performance on the real-world task of detecting incorrectly generated diagnoses in an enterprise AI diagnostic system.

2 Related Work

AI-based diagnostic systems have advanced substantially with the emergence of LLMs. Systems such as DISC-MedLLM (Bao et al., 2023), AMIE (Tu et al., 2025), MEDAGENTS (Tang et al., 2024), and CLARITY (Shaposhnikov et al., 2025) combine multi-turn dialog, natural language understanding, and broad medical knowledge (Mukherjee et al., 2024; Sun et al., 2024; Yu et al., 2024) to support interactive diagnostic consultations. Despite this progress, LLM-based diagnostic systems can produce inconsistent diagnostic suggestions across dialog turns and often lack well-calibrated confidence scores, making it difficult for clinicians to assess when their outputs can be trusted. In this work, we address this safety-critical gap.

Uncertainty quantification methods can be categorized along two main dimensions. First, they can be unsupervised or supervised, depending on whether they need any labeled training data (Fomicheva et al., 2020; Fadeeva et al., 2024). Second, they can be classified as white-box or black-box, depending on the level of access to the underlying model (e.g., internal states and logits versus only generated text).

In the black-box setting, only the generated text is accessible. Here, uncertainty is commonly quantified using consistency-based approaches,

which measure agreement across multiple generations sampled from the model. Representative methods include lexical-similarity-based uncertainty (Fomicheva et al., 2020), the Number of Semantic Sets, and similarity-graph-based measures such as the Sum of Eigenvalues of the Graph Laplacian, Degree Matrix, and Eccentricity (Lin et al., 2024). In the majority of methods, semantic similarity is computed using an NLI model. The consistency-based methods are among the strongest techniques for UQ in general but introduce a huge computational overhead.

In the white-box setting, full access to the LLM is available, including its internal states and token-level probability distributions. Many unsupervised approaches rely on information-theoretic signals derived from token-level probabilities: Maximum Sequence Probability (MSP) (Fadeeva et al., 2023), Perplexity, Maximum Token Entropy (TE) (Fomicheva et al., 2020), PMI (Takayama and Arase, 2019), SelfCertainty (Kang et al., 2025), CCP (Fadeeva et al., 2024), TokenSAR (Duan et al., 2024), Frequency Scoring (FS) (Mohri and Hashimoto, 2024), PTrue (Kadavath et al., 2022). Few methods use the internal states of the model, such as Attention Score (Sriramanan et al., 2024).

Supervised UQ methods learn a classifier or regressor to predict the correctness of LLM generations using its internal states, e.g., hidden states (Azaria and Mitchell, 2023; CH-Wang et al., 2024; Vazhentsev et al., 2025b) or attention maps (Chuang et al., 2024; Vazhentsev et al., 2025a; Shelmanov et al., 2025). They typically achieve substantially better performance than unsupervised methods, particularly in in-domain settings. However, because they rely on access to internal representations of the LLM, they are generally difficult to apply in black-box scenarios.

In this work, we develop guardrails for an enterprise medical diagnostic system with the goal of improving patient safety. First, we systematically investigate and enhance unsupervised UQ techniques. Second, we propose a method for adapting supervised UQ approaches to the black-box setting, enabling them to achieve strong performance while remaining compatible with the constraints of the enterprise system, where access to LLM internal representations is unavailable.

3 Diagnostic Task and Clinical Data

An AI diagnostic assistant is intended to support physicians during initial consultations. Given a multi-turn chat that records the collection of a patient’s medical history and anamnesis (symptoms, their dynamics, relevant background, and answers to follow-up questions), the system must produce a ranked list of plausible diagnoses. Concretely, for each conversation, the target output is a set of up to three candidate diagnoses that best explain the reported clinical picture.

We created a dataset with 360 chat-based consultations in Russian that simulate the information-gathering stage of a primary-care visit, where a clinician asks clarifying questions and the patient responds². The first 180 chats are human-human dialogs between two licensed physicians: one physician plays the patient with a known underlying condition and presents corresponding symptoms, while the other plays the clinician and conducts the interview to identify the disease. The remaining 180 chats are real patient-LLM dialogs, where an LLM conducts the same type of diagnostic interview. To determine reference diagnoses for evaluation, seven physicians independently reviewed every conversation and provided up to three diagnoses per chat. In addition to these expert labels, we also collected model-generated diagnoses by prompting GigaChat (Mamedov et al., 2025) to output up to three diagnoses for each conversation, following the protocol of prior work (Kopanichuk et al., 2026). This post-hoc annotation setup infers realistic diagnoses from the conversation alone, as the chats simulate an early diagnostic interview rather than a complete clinical workup.

4 Confidence-Based Safety Mechanism

The confidence-based safety mechanism wraps a medical AI diagnostic system with a UQ-based method. More precisely, we extract meaningful diagnoses from the answer of the AI medical system and assign each of them a confidence score, as shown in Figure 1. These scores quantify the predictive confidence of the model and enable fine-grained selective question answering (Xin et al., 2021; Vashurin et al., 2025). When uncertainty exceeds a predefined rejection threshold h , the system

²The anonymized evaluation data is available at <https://huggingface.co/datasets/kopan/med-eval-360>

transfers it to a human expert:

$$\text{Diagnosis} = \begin{cases} \text{generation,} & \text{if } U(x, y) < h, \\ \text{expert answer} & \text{if } U(x, y) \geq h, \end{cases} \quad (1)$$

where $U(x, y)$ corresponds to the uncertainty score for the dialog x and the LLM prediction y . The choice of threshold h is task-specific and is described for our system in Section 6.2. To design an effective safety mechanism, we evaluate a broad range of state-of-the-art UQ methods described in Section 2.

Consistency-based black-box methods.

Among unsupervised approaches, we focus on consistency-based techniques due to their strong empirical performance and compatibility with black-box deployment settings. We observe that the effectiveness of these methods depends on the NLI model used to assess semantic similarity between generated outputs. We hypothesize that general-purpose NLI models struggle with clinical diagnoses because distinct conditions may have similar surface forms, while equivalent diagnoses may be expressed using different terminology (e.g., rhinitis vs. allergic rhinitis). To mitigate this issue, we train our small encoder-based matching model based on ModernBERT to assess the similarity between diagnoses (see Appendix D) and use it in consistency-based methods.

Supervised black-box methods. We hypothesize that in specialized domains such as medicine, supervised UQ methods can outperform unsupervised approaches. We used the matching model together with the ground-truth diagnoses to construct training labels. We focus on two supervised methods – the Text-Based UQ (TB UQ) baseline and UHead (Shelmanov et al., 2025).

TB UQ trains a lightweight encoder-based transformer to classify the correctness of generated textual diagnoses. It uses the LLM prompt as the premise and a generated diagnosis as the hypothesis, with the ground-truth diagnosis providing the supervision signal. Since the method operates purely on text, it is directly compatible with both white-box and black-box LLM deployments.

UHead (Shelmanov et al., 2025) leverages the internal states of the LLM, such as hidden states, attention maps, and token probabilities, as features for an auxiliary transformer-based UQ head that predicts the correctness of the generated diagnosis. The original UHead is compatible only with white-box models. To use UHead with black-box models,

we propose using a *proxy model*. Specifically, we reuse the prompts and generated output from the black-box model but compute hidden states, attention maps, and token probabilities using a smaller open-source white-box LLM. This approach enables the training and inference of UHeads on top of a white-box proxy LLM while approximating the reliability of the underlying black-box LLM. As we show in experiments, the best results can be achieved when the proxy model and the black-box model are from the same LLM family.

The motivation for using a proxy is twofold. First, proxy-based UHead can approximate the aleatoric uncertainty associated with the inherent complexity and ambiguity of diagnostic tasks, as this type of uncertainty is largely model-independent. Second, if the proxy belongs to the same LLM family as the enterprise model, it is likely derived through distillation and, therefore, may partially inherit the epistemic uncertainty of the enterprise model, reflecting similar knowledge limitations and failure modes.

5 Experimental Setup

Datasets. We conducted experiments on the medical question-answering generation task, as described in Section 3. Examples represent multi-turn conversations; the full dialog is used in the input prompt. An example of a dataset instance is provided in Table 11 in Appendix G.

Expert-derived targets. The medical dataset initially consisted of 360 examples with annotations from medical professionals. Because some dialogs exceeded our context length limitations, we filtered the dataset and retained 323 examples for further analysis. Each sample may contain several diagnostic hypotheses, reflecting natural disagreement and uncertainty in clinical reasoning. We used 259 samples for evaluation and 64 as a holdout split for additional control and threshold selection. These sets contain approximately 780 and 180 diagnoses, respectively, as each sample includes approximately three diagnoses.

Model-derived targets. For training, we use a larger set of 1500 dialogs (approximately 4500 diagnoses), where target diagnoses were generated using a strong reference model – DeepSeek-V3.2 (DeepSeek-AI et al., 2025). To validate the quality of DeepSeek annotation, we compared its performance with the average performance of expert physicians. Table 9 in Appendix F shows that

DeepSeek has comparable performance and, thus, can be used to generate training labels. These synthetic labels are used only to train supervised UQ methods; all test-set evaluation and metric computation use expert-derived reference diagnoses.

Models. We use two LLMs for generating diagnoses: GigaChat-2-Max and GigaChat-2-Lite (Mamedov et al., 2025).³ These models achieve strong performance in Russian and satisfy the deployment constraints of the enterprise diagnostic system. *GigaChat-2-Max* is a large-scale instruction-tuned LLM used as the primary diagnostic generator. The enterprise medical diagnostics system is built on top of this model. GigaChat-2-Max is a black-box model deployed via API, and we observe only the generated text. This setting is a common real-world deployment scenario in which access to model weights and internal states is restricted due to proprietary constraints, security considerations, or API-based usage. *GigaChat-2-Lite* is an open-weight instruction-following LLM with 7 billion parameters. We use it in the white-box regime: in addition to the generated text, we have access to the internal states of the model, such as hidden states, attention maps, logits, etc.

UQ baselines. We conduct experiments with unsupervised and supervised methods discussed in Sections 2 and 4. Several UQ baselines use an NLI model for calculating the semantic similarity between samples. We evaluated CCP, FS, and TokenSAR with multilingual NLI DeBERTa⁴ for CCP and FS, and RoBERTa CrossEncoder⁵ for TokenSAR; this setting is denoted as (base). We then evaluated these methods with the NLI model adapted for the medical domain (ModernBERT (Warner et al., 2025)) (med). The details of the adaptation are provided in Appendix D; details of all UQ methods are presented in Appendix A.

Both supervised methods, UHead and TB UQ, were trained on 1,500 dialog samples, with 260 samples left for evaluation. The training setup and hyperparameters are presented in Appendix C. The TB UQ method was used with the best backbone model (see Appendix E). We also compared the supervised methods in the black-box setting with the consistency-based baselines described above.

Metrics. The performance of UQ methods is evaluated using the Prediction Rejection Rate

(PRR) (Malinin et al., 2017). PRR is calculated as the ratio of the area between the rejection curves of the UQ method and the random UQ baseline and the same area between the ideal UQ method and the random UQ baseline. In this work, we report PRR normalized at a risk threshold of 0.5 (PRR@0.5; Vashurin et al. (2025)), which measures how effectively an uncertainty score enables the rejection of unreliable outputs while preserving high-quality predictions.

Answer quality is evaluated using a single metric, Relaxed Exact Match (Relaxed EM), calculated with a dedicated medical diagnosis matching model distilled from an ensemble (see Appendix D). This matcher is used consistently across all experiments. Evaluation is performed at the diagnosis level. Suppose that for the prompt x_i we have the LLM’s generation g_i , which contains up to $k = 3$ diagnoses $g_i = \{g_i^j\}_{j=0..k}$. For this prompt, we also have the ground truth diagnosis, denoted as $y_i = \{y_i^j\}_{j=0..h}$, which has up to h diagnoses. Then our metric for one diagnosis j for the sample i is defined as:

$$\text{Relaxed EM}_{i,j} = \max_{k=1..h} \text{EM}(g_i^j, y_i^k). \quad (2)$$

See Appendix B for more details.

6 Results and Discussion

6.1 Main Results

White-box setting. We first evaluate existing UQ methods with a white-box model, GigaChat-2-Lite. The results are presented in Table 1.

We divide the table into four groups: (1) base methods, (2) methods using an additional model together with the base model, (3) methods using an additional model with a medically adapted NLI model, and (4) supervised methods. Among the base methods, MSP achieves the strongest performance, reaching 0.597 PRR@0.5. Methods that rely on an additional base NLI model show comparable results, with TokenSAR (base) achieving 0.577 PRR@0.5. However, all such approaches benefit substantially from medical-domain adaptation, which improves performance for two out of three methods and raises the score to 0.631 PRR@0.5 for the best-performing FS (med) variant. Finally, the supervised UHead method achieves the best overall performance, outperforming the second-best method, FS (med).

Black-box setting. The results for the black-box setting are presented in Table 2. Among the unsupervised approaches, the best performance

³<https://huggingface.co/ai-sage>

⁴<https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>

⁵<https://huggingface.co/cross-encoder/stsb-roberta-large>

UQ method	PRR@0.5 $\uparrow$
MSP	0.597
Perplexity	0.382
TE	0.522
SelfCertainty	0.100
PMI	-0.314
PTrue	0.111
AttentionScore	0.189
Focus	0.517
CCP (base)	0.462
FS (base)	0.366
TokenSAR (base)	0.577
CCP (med)	0.534
FS (med)	<u>0.631</u>
TokenSAR (med)	0.565
UHead	0.653
TB UQ (ModernBERT-large)	0.310

Table 1: Performance of UQ methods for *white-box* GigaChat-2-Lite on the *test* set. The best result is highlighted in **bold**, and the second best is underlined. The table blocks correspond to base UQ methods, NLI-based UQ methods using base and medical NLI models, and supervised methods, respectively.

is achieved by Eccentricity with the medical NLI model and Jaccard similarity, reaching 0.561 PRR@0.5. Overall, most black-box baselines perform similarly, except LexicalSimilarity, which shows slightly lower performance.

The supervised TB UQ baseline, which predicts the correctness of diagnoses using only the generated text, demonstrates relatively weak performance. In contrast, UHead with a proxy model achieves the best results, reaching 0.714 PRR@0.5. We conducted a paired bootstrap test comparing UHead with the best baseline, Eccentricity with the medical NLI model and Jaccard similarity, and obtained a p-value of 0.006, indicating a statistically significant improvement. This substantial gain over the TB UQ baseline highlights the value of leveraging internal model representations, even from a lightweight proxy LLM, as a signal for estimating the reliability of LLM-generated diagnoses. Table 7 in Appendix E further shows that UHead remains effective when using a smaller proxy model from a different model family and still outperforms the other UQ methods.

6.2 Selecting Abstention Threshold

For deployment in a real enterprise diagnostic system, the uncertainty score must be mapped to actionable thresholds for abstention or other interventions. In practical use, ranking quality alone is not sufficient; the uncertainty scores should also

UQ method	PRR@0.5 $\uparrow$
NumSemSets	0.552
EigValLaplacian (med, NLI entail)	0.523
EigValLaplacian (med, NLI contra)	0.523
EigValLaplacian (med, Jaccard score)	0.494
DegMat (med, NLI entail)	0.518
DegMat (med, NLI contra)	0.518
DegMat (med, Jaccard score)	0.510
Eccentricity (med, NLI entail)	0.554
Eccentricity (med, NLI contra)	0.554
Eccentricity (med, Jaccard score)	<u>0.561</u>
KernelLanguageEntropy	0.547
LexicalSimilarity (rougeL)	-0.039
LUQ	0.517
UHead (approx. via GigaChat-2-lite)	0.714
TB UQ (ModernBERT-large)	0.499

Table 2: Performance of UQ methods for *black-box* GigaChat-2-Max on the *test* set. The best results are highlighted with **bold**, the second best is underlined.

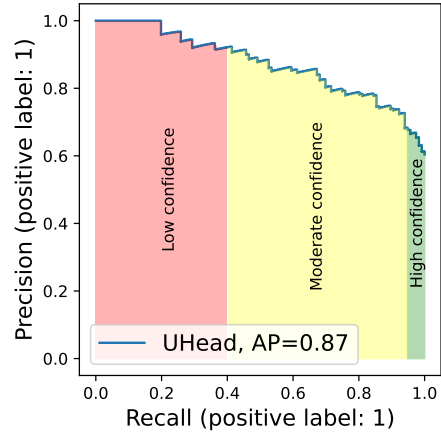


Figure 2: Precision–recall curve for diagnosis errors (positive class) detection on the *holdout* set with UHead based on proxy for black-box GigaChat-2-Max.

be converted into discrete confidence tiers that can be presented to physicians. We propose discretizing confidence estimates into three levels: high, moderate, and low confidence, and selecting the corresponding thresholds on a hold-out validation set. UHead with a proxy LLM maintains strong performance on the holdout set, achieving 0.704 PRR@0.5, consistent with the results on the test set. We construct the precision–recall curve shown in Figure 2. Based on this curve, we set the thresholds for low and moderate confidence to 0.37 and 0.76, respectively, which correspond to (error detection precision = 0.92, recall = 0.40) and (precision = 0.68, recall = 0.95). Overall, these calibrated thresholds translate UHead scores into interpretable confidence tiers that can directly support safer decision-making in production. In

particular, they enable conservative abstention at low confidence while preserving higher precision when the model is confident. In the industrial setting, physicians are shown a separate confidence tier for each generated candidate diagnosis and can use this information when reviewing the model output. The expert physician review provides error categorization and shows that most errors occur under low or moderate confidence for all types of error (Appendix H).

Since inference latency is also an important requirement for an industrial system, we estimated the latency of the full pipeline. In our experiments, UHead using GigaChat-2-Lite as a proxy model (7B parameters) adds less than 10% to the inference time of the main black-box LLM and has only a minor impact on the overall application latency (approximately 2%).

7 Conclusion

We present a study of uncertainty-based guardrails in an enterprise medical AI system that provides confidence estimates for generated medical diagnoses. The specificity of our case is the lack of direct access to a large proprietary backbone LLM, GigaChat-2-Max. For this black-box setup, we adapted a supervised white-box method, UHead (Shelmanov et al., 2025), introducing a small white-box proxy model to estimate the uncertainty of the larger black-box model. We showed that the proposed solution demonstrates the best performance for the given model. Hence, it was integrated as a safety mechanism in the considered medical AI diagnostic system. We also discovered that the performance of UQ methods that assess semantic similarity between generated samples (CCP, FS, and others) can be substantially improved using a specialized medical NLI model, which may be a useful insight for white-box industrial setups.

Limitations

The data used in the project described in this paper are closed-source and cannot be shared for ethical and privacy reasons. However, in the paper, we presented several synthetic examples of dialogs that are similar to the real data. The raw clinical data remain closed for privacy and ethical reasons. To support reproducibility, we release an anonymized evaluation dataset, following the required ethical and legal procedures

All experiments were conducted exclusively on

Russian-language data. However, the UQ methods considered in this work are largely language-agnostic, and we expect the main findings to generalize to other languages, as can be confirmed, e.g., by comparable experiments of Vashurin et al. (2025) and Fadeeva et al. (2024) with UQ in English, Russian, Chinese, and Arabic, inter alia.

The scope of the experiments (LLMs, data, language, etc.) is limited by the constraints of the considered enterprise solution. The presented proxy-based UHead method supposing that the proxy model belongs to the same LLMs family as the main model, which are not always possible; however, our experiments show that this method could be used even with the cross-family proxy. In future work, we aim to demonstrate the effectiveness of using proxy LLMs for UQ with other open-source models and datasets.

Ethical Considerations

The study discussed in the paper aims to improve the safety of AI medical diagnostic systems.

During data collection, all potential participants were informed about the purpose of the study, the potential risks and benefits, and the source of funding, after which they gave their voluntary informed consent. To ensure confidentiality, no personally identifiable information was collected. Access to raw data is granted only to a limited number of individuals who have completed mandatory certification in the protection of human research participants. Nevertheless, the dataset used in the study was made closed-source to avoid unintentional harm to the involved patients.

To support reproducibility while preserving participant confidentiality, we plan to release a carefully anonymized version of the test set, subject to all required ethical, legal, and institutional approvals.

Acknowledgments

We thank anonymous reviewers for their insightful feedback towards improving this paper. The work of Gleb Kuzmin was supported by a grant, provided by the MED of Russia (agreement identifier 000000C313925P4G0002) and the agreement with the ISP RAS dated June 20, 2025 No. 139-15-2025-011.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. [Optuna: A next-generation hyperparameter optimization framework](#). In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.
- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [DISC-MedLLM: Bridging general large language models and real-world medical consultation](#). *Preprint*, arXiv:2308.14346.
- Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. [Do androids know they’re only dreaming of electric sheep?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.
- Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. [Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [DeepSeek-V3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. [Shifting uncertainty to relevance: Towards the predictive uncertainty quantification of free-form large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.
- Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. [AutoGluon-Tabular: Robust and accurate autotml for structured data](#). *Preprint*, arXiv:2003.06505.
- Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. [Fact-checking the output of large language models via token-level uncertainty quantification](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand. Association for Computational Linguistics.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. [LM-Polygraph: Uncertainty estimation for language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. [Unsupervised quality estimation for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a Bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. [Towards more accurate uncertainty estimation in text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372, Online. Association for Computational Linguistics.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. [Bayesian active learning for classification and preference learning](#). *Preprint*, arXiv:1112.5745.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and](#)

- open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable Best-of-N selection for large language models via self-certainty. *Advances in Neural Information Processing Systems*, 38:19720–19745.
- Iliia Kopanichuk, Petr Anokhin, Vladimir Shaposhnikov, Vladimir Makharev, Ekaterina Tsapieva, Iaroslav Beshpalov, Victor Gomboleviskiy, Dmitry Kurapeev, Dmitry V. Dylov, and Ivan Oseledets. 2026. Beyond majority voting: a new framework for evaluating AI diagnostic systems against expert consensus. *Scientific Reports*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.
- Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–50, Vancouver, Canada. Association for Computational Linguistics.
- Valentin Mamedov, Evgenii Kosarev, Gregory Leleytner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, Ainur Israfilova, Ivan Baskov, Artem Chervyakov, Emil Shakirov, Mikhail Kolesov, Daria Khomich, Daria Latortseva, Sergei Porkhun, Yury Fedorov, and 14 others. 2025. GigaChat family: Efficient Russian language modeling through mixture of experts architecture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 93–106, Vienna, Austria. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR.
- Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, and 7 others. 2024. Polaris: A safety-focused LLM constellation architecture for healthcare. *Preprint*, arXiv:2403.13313.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.
- Vladimir Shaposhnikov, Alexandr Nesterov, Iliia Kopanichuk, Ivan Bakulin, Zhelvakov Egor, Ruslan Abramov, Tsapieva Ekaterina Olegovna, Iaroslav Radionovich Beshpalov, Dmitry V. Dylov, and Ivan Oseledets. 2025. CLARITY: Clinical assistant for routing, inference, and triage. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1805–1821, Suzhou (China). Association for Computational Linguistics.
- Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35712–35731, Suzhou, China. Association for Computational Linguistics.
- Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

- Zhoujian Sun, Cheng Luo, Ziyi Liu, and Zhengxing Huang. 2024. [Conversational disease diagnosis via external planner-controlled large language models](#). *Preprint*, arXiv:2404.04292.
- Junya Takayama and Yuki Arase. 2019. [Relevant and informative response generation using pointwise mutual information](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, Florence, Italy. Association for Computational Linguistics.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. [MedAgents: Large language models as collaborators for zero-shot medical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. [Benchmarking uncertainty quantification methods for large language models with LM-Polygraph](#). *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2025a. [Unconditional Truthfulness: Learning Unconditional Uncertainty of Large Language Models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35673–35694, Suzhou, China. Association for Computational Linguistics.
- Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025b. [Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. 2025. [Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond](#). *Eng. Appl. Artif. Intell.*, 139:109553.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. [The art of abstention: Selective prediction and error regularization for natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. [Hallucination is inevitable: An innate limitation of large language models](#). *Preprint*, arXiv:2401.11817.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qinkai Yu, Mingyu Jin, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, and Yongfeng Zhang. 2024. [Health-LLM: Personalized Retrieval-Augmented Disease Prediction System](#). *Preprint*, arXiv:2402.00746.
- Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. [LUQ: Long-text uncertainty quantification for LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.
- Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. [Enhancing uncertainty-based hallucination detection with stronger focus](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.
- Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. [Mitigating uncertainty in document classification](#). In *Proceedings of the 2019 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

A UQ Methods

This appendix contains a high-level description for uncertainty quantification used in our experiments.

Information-based methods. These methods treat the generation process itself as a source of uncertainty evidence.

MSP. Maximum sequence probability. Estimates uncertainty using the probability assigned by the model to the generated sequence. The uncertainty score is computed as one minus the probability of the most likely generated sequence:

$$\text{MSP}(y \mid x, \theta) = 1 - P(y \mid x, \theta).$$

Perplexity (Fomicheva et al., 2020). The perplexity of the answer reflects how unlikely the generated response is according to the language model.

$$\text{PPL}(y \mid x, \theta) = \exp \left\{ -\frac{1}{L} \log P(y \mid x, \theta) \right\},$$

where L is a length of generation.

Token entropy (Fomicheva et al., 2020). Mean entropy of probability distribution.

$$\mathcal{H}_T(\mathbf{y} \mid \mathbf{x}; \theta) = \frac{1}{L} \sum_{l=1}^L \mathcal{H}(y_l \mid y_{<l}, \mathbf{x}; \theta),$$

where $\mathcal{H}(y_l \mid y_{<l}, \mathbf{x}; \theta)$ is the entropy of the token probability distribution at generation step l . Higher entropy indicates higher uncertainty.

PMI (Takayama and Arase, 2019). Pointwise mutual information between answer and question measures how strongly the response depends on the prompt.

$$\text{PMI}(x, y) = \log \frac{P(y \mid x)}{P(y)}.$$

Lower PMI indicates higher uncertainty.

AttentionScore (Sriramanan et al., 2024). Estimates uncertainty from internal attention scores using token self-attention across heads and layers. Concretely, self-attention values corresponding to the diagonal entries of the attention matrices are averaged across transformer layers and attention heads. Resulting scores were used as an uncertainty estimate.

Focus (Zhang et al., 2023). Uses token importance scores estimated by TF-IDF or other weighting methods and uses them to aggregate token-level uncertainty. Instead of averaging uncertainty across

all generated tokens, it emphasizes uncertainty associated with informative tokens, under the assumption that errors in content-bearing words are more indicative of response quality.

Self-evaluation methods. These methods obtain an uncertainty signal from the model’s own assessment of the generated answer.

SelfCertainty (Kang et al., 2025). Model’s self-reported confidence. After the generation of the response, the model is asked to estimate the confidence of its own answer.

PTrue (Kadavath et al., 2022). Reflexive method in which the model estimates the probability of a correct answer. The model is asked to estimate the probability after answer generation.

Meaning-diversity methods. These methods estimate uncertainty from semantic variability, either from likely token-level alternatives or from disagreement across multiple sampled generations.

CCP (Fadeeva et al., 2024). Claim-Conditioned Probability. Decomposes the answer into individual claims and estimates the conditional generation probability of each claim. For each word in the claim, CCP considers top-K token alternatives and groups them into entailment and contradiction sets using an NLI model. Word-level scores are then multiplied into a claim-level uncertainty estimate.

TokenSAR (Duan et al., 2024). Aggregates token-level uncertainty scores into claim uncertainty. It combines token likelihood with token relevance, which reflects how much the response meaning changes when the token is removed. For each extracted claim, the scores of tokens aligned to this claim are aggregated into a claim-level uncertainty estimate.

Frequency Scoring (FS) (Mohri and Hashimoto, 2024). Estimates uncertainty scores based on the semantic consistency of extracted claims across multiple sampled generations. For each extracted claim, the method compares it with sampled generations. Each comparison is assigned an entailment, neutral, or contradiction judgment, and the claim-level score is computed as the number of contradiction judgments minus the number of entailment judgments:

$$\text{FS}(c) = N_{\text{contra}}(c) - N_{\text{entail}}(c),$$

where higher scores indicate higher uncertainty.

NumSemSets (Lin et al., 2024). Measures uncertainty as a number of semantic clusters among generation samples. Sampled responses are grouped

Model	UQ method	Learning Rate	Num. of Epochs	Attn. Window Size
GigaChat-2-Lite	UHead	0.1	15	10
GigaChat-2-Max	UHead	0.1	10	5
GigaChat-2-Lite	TB UQ (ModernBERT-large)	1e-5	10	-
GigaChat-2-Max	TB UQ (ModernBERT-large)	1e-5	10	-

Table 3: Hyperparameters for supervised UQ methods.

into clusters of semantically equivalent meanings. More clusters mean higher uncertainty.

EigValLaplacian (Lin et al., 2024). Estimates uncertainty using the eigenvalues of the semantic similarity graph Laplacian. The graph is constructed from sampled responses, with edge weights representing semantic similarity between pairs of responses.

DegMat (Lin et al., 2024). Uses the degree matrix of the semantic similarity graph to estimate uncertainty. The graph is constructed from response samples using pairwise semantic similarity scores. On the diagonal elements, the degree matrix total similarity of each response to all other responses is situated. Lower overall similarity means higher uncertainty.

Eccentricity (Lin et al., 2024). Estimates uncertainty based on how dispersed the sampled generations are in the spectral embedding space. A semantic similarity graph is built over sampled answers.

Kernel Language Entropy (Nikitin et al., 2024). Estimates uncertainty from the entropy of a semantic similarity kernel constructed from sampled generations. The kernel captures the overall semantic agreement between generated responses, where more consistent generations result in lower entropy and uncertainty

Lexical Similarity (Fomicheva et al., 2020). Uses lexical similarity between generated samples. Estimates uncertainty as the negative mean pairwise lexical similarity. More lexically diverse responses indicate higher uncertainty.

Long-text Uncertainty Quantification (LUQ) (Zhang et al., 2024). Measures how consistently multiple sampled generations support each other semantically using an NLI model. This is computed by averaging pairwise entailment scores between sampled generations. Lower average entailment corresponds to higher uncertainty.

B UQ Metrics

For methods comparison, we used the Prediction Rejection Ratio (PRR). For each input $\mathbf{x}_j$, the language model generates a prediction $\mathbf{y}_j$, which is assigned an uncertainty score $u_j = U(\mathbf{x}_j)$. The rejection curve characterizes the average prediction quality $Q(\mathbf{y}_j, \hat{\mathbf{y}}_j)$ as increasingly uncertain predictions are rejected according to a threshold u_j ; $\hat{\mathbf{y}}_j$ corresponds to the ground-truth answer for $\mathbf{x}_j$. PRR measures how much the resulting curve improves over random selection, normalized by the improvement achieved by an oracle ranking:

$$PRR = \frac{AUC_{\text{unc}} - AUC_{\text{rnd}}}{AUC_{\text{oracle}} - AUC_{\text{rnd}}}. \quad (3)$$

Higher PRR values indicate better uncertainty scoring. We report PRR@0.5, which computes the prediction-rejection area up to rejecting at most 50% of the test instances.

PRR was chosen because of its ability to directly evaluate uncertainty estimation by ranking unreliable predictions so that they can be rejected. Since PRR depends entirely on the ordering of uncertainty scores, it enables a fair comparison between methods that produce heterogeneous uncertainty estimates.

To estimate the quality of the predictions, we used the Relaxed EM metric, introduced in Equation (2). This metric is based on the diagnosis match regarding the specialized matching function, which is denoted EM in Equation (2) (see the Appendix D for more details on the matching function).

C Training Setup for Supervised UQ Methods

We trained a linear UHead model for both GigaChat-2-Lite and GigaChat-2-Max, using GigaChat-2-Lite as an approximation model and token probabilities, attentions, and hidden states from all layers as features. The UHead model was trained with the hyperparameters from Table 3. The TB UQ method was trained on texts only, with the hyperparameters as in Table 3.

Approach	Model / Pipeline	Accuracy	Precision	Recall	F_1
LLM-as-a-judge	GPT-4o	0.914	0.575	0.905	0.703
LLM-as-a-judge	Qwen2.5-72B	0.919	0.595	0.894	0.714
LLM-as-a-judge	Qwen2.5-7B	0.949	0.751	0.823	0.785
LLM-as-a-judge	Llama-3.1-70B	0.948	0.751	0.805	0.777
Meta-model	AutoGluon bagged LightGBM ensemble	0.980	0.910	0.900	0.910
Distillation	ModernBERT-base (thr.=0.76)	<u>0.970</u>	<u>0.880</u>	<u>0.890</u>	<u>0.882</u>
Distillation	bert-base-multilingual-cased (thr.=0.77)	0.969	0.880	0.870	0.877
Distillation	ModernBERT-base (Optuna, thr.=0.48)	0.963	0.840	0.860	0.853

Table 4: Comparison of diagnosis match implementations on the test set ($n = 1469$). Precision/Recall/ F_1 are reported for the positive class. Best overall values per column are in **bold**; best distillation run is underlined.

D Diagnosis Matching

Determining whether two diagnostic text strings refer to the same underlying medical condition is a challenging problem, as equivalent diagnoses may be expressed in different forms. For instance, "acute pancreatitis" and "chronic pancreatitis exacerbation" represent distinct conditions, while "rotavirus gastroenteritis" and "acute intestinal infection gastroenteritis" may be considered a match depending on the clinical context. To address this challenge, we adopted a supervised meta-modeling approach from (Kopanichuk et al., 2026).

The match function $M : \mathfrak{F} \mapsto \mathbb{B}$ determines whether a pair of diagnosis strings (d^A, d^E) should be considered equivalent. This function was trained on a dataset of 4,833 labeled diagnosis pairs (with 1,469 held out for testing), where labels were determined by majority vote among three medical experts using criteria specifically developed for diagnosis similarity assessment. The meta-model pipeline combines 17 engineered features across four categories: (i) direct LLM prompting using five language models (Llama-3.1-70B-Instruct, Qwen2.5-7B-Instruct, GigaChat-Max, Vikhr-Nemo-12B-Instruct, Phi-4) selected for the highest F_1 scores, (ii) RAG-based ICD-10 code matching using BERGAMOT embeddings and FAISS vector search over 20,000+ ICD entries, (iii) text embedding similarity computed with three models (Giga, BERGAMOT, Arctic-Embed), and (iv) linguistic similarity ratios including n-gram overlap, stemming, lemmatization, POS tags, and normalized Levenshtein distance.

Model training was performed using AutoGluon AutoML (Erickson et al., 2020) with optimization targeting average precision on cross-validation folds, resulting in a bagged ensemble of LightGBM classifiers as the best-performing model. This meta-

model achieved 0.98 accuracy, 0.91 precision, 0.90 recall, and 0.91 F_1 -score on the test set, substantially outperforming LLM-as-a-judge baselines, including GPT-4o (0.70 F_1) and Qwen2.5-72B (0.71 F_1), as shown in Table 4.

To reduce inference complexity and deployment costs while maintaining high performance, we applied knowledge distillation to compress the meta-model into smaller transformer-based student models. The distillation process used 56,412 training samples, where each diagnosis pair was encoded as a sequence-pair classification input. Students were trained using a hybrid objective that combined hard expert labels and soft teacher probabilities with temperature scaling to preserve uncertainty information. The training objective enables the student to learn the teacher’s decision boundary while remaining consistent with expert annotations.

The best distilled model (ModernBERT-base with a batch size of 32, 5 epochs, $\alpha=0.3$, temperature=3.0, and a decision threshold of 0.76) achieved 0.97 accuracy, 0.88 precision, 0.89 recall, and 0.88 F_1 -score—retaining 96.7% of the meta-model’s F_1 performance while requiring significantly fewer computational resources at inference time. Additional distillation experiments with bert-base-multilingual-cased and Optuna-based hyperparameter optimization (Akiba et al., 2019) yielded comparable results (Table 4).

E Supervised UQ Method Ablation

To evaluate the best encoder-based model as a TB UQ method for the white-box GigaChat-2-Lite and black-box GigaChat-2-Max models, we compared several encoder-based models on this task and assessed their performance. The results are provided in Tables 5 and 6. As shown in the tables, the ModernBERT-large achieves the best performance

for the white-box and black-box models and is thus used in our experiments.

To further investigate the effect of the proxy model used in UHead under the black-box setting, we trained multiple UHead variants using different proxy LLMs. The results are presented in Table 7 and indicate that the proposed approach is robust across a range of proxy models, including the models from the different model families (e.g., Gemma-3 (Gemma Team et al., 2025), Qwen-3 (Yang et al., 2025), and GigaChat-2).

UQ method	PRR@0.5 $\uparrow$
ModernBERT-large	0.310
RuBioRoBERTa	0.030
RuRoBERTa	0.083
ClinicalBERT	0.026

Table 5: Results for various encoder-based models as TB UQ method for GigaChat-2-Lite model on *test* set, diagnosis-level.

UQ method	PRR@0.5 $\uparrow$
ModernBERT-large	0.499
RuBioRoBERTa	0.461
RuRoBERTa	0.387
ClinicalBERT	-0.179

Table 6: Results for various encoder-based models as TB UQ method for GigaChat-2-Max model on *test* set, diagnosis-level.

UQ method	PRR@0.5 $\uparrow$
Eccentricity (med, Jaccard score)	0.561
UHead (approx. via GigaChat-2-lite)	0.714
UHead (approx. via GigaChat-20B-A3B)	0.670
UHead (approx. via Gemma-3-4B-it)	0.650
UHead (approx. via Qwen-3-4B-Instruct-2507)	0.609

Table 7: Results for various proxy models with UHead for GigaChat-2-Max model on *test* set, diagnosis-level.

F Data Statistics and Labeling

We provide additional statistics for the dataset in Tables 8 and 10. Table 8 summarizes the key statistics in all data splits while Table 10 presents the distribution of diagnoses in the test set, demonstrating that the dataset covers a broad range of medical conditions and clinical scenarios with respect to ICD-10.

To further validate the use of DeepSeek-V3.2 for labeling the training data used by supervised UQ methods, we compared its labeling performance

against that of a typical expert on the test set, for which both expert annotations and ground-truth labels are available. Specifically, we computed the proportion of samples for which at least one of the predictions matches the ground-truth targets. For DeepSeek-V3.2, we used the prediction of the model, while for human annotators we used the diagnoses provided by individual experts. The ground-truth labels were obtained by aggregating the annotations of all seven experts. The results are presented in Table 9 and show that the labeling performance of the average expert is comparable to that of DeepSeek-V3.2, making it feasible for automatic train set labeling. We also want to explicitly note that this approach was used only for labeling additional training data for supervised UQ methods and hence does not affect the evaluation procedure.

Split	Avg. Turns	Avg. Words	Avg. Words in Sentence	Approx. Number of Samples	Approx. Number of Diagnoses
Train	4.4	64.4	7.3	1500	4500
Test	6.8	98.0	7.2	259	780
Holdout	6.6	97.7	7.4	64	192

Table 8: Dataset statistics. We present an average number of turns in multi-turn dialogues, average number of words per sample and per sentence for each split, as well as the numbers of samples and diagnoses. Note that training and evaluation were conducted at the diagnosis level.

Labeling Scheme	Performance
Experts	65.6
DeepSeek-V3.2	63.7

Table 9: Comparison of expert labels and DeepSeek-V3.2 labels on test set.

G Data Examples

We provide an illustrative example of the dataset structure. To protect patient privacy and comply with ethical and legal requirements, we present a synthetic example instead of a real clinical conversation. The synthetic sample preserves the structure and key characteristics of the original data, allowing readers to understand the methodology without exposing sensitive medical information. The example is shown in Table 11 and demonstrates the full dialog. We also want to explicitly state that we used the full doctor-patient dialogs, without any summarization or reduction.

ICD-10 Category	A	B	D	E	F	G	H	I	J	K	L	M	N	R	T
Percentage of category in test set	1.0	1.8	1.3	5.6	1.6	5.0	4.7	9.6	10.9	17.8	5.6	19.5	13.8	0.8	1.0

Table 10: Distribution of ICD-10 codes for test set samples; the dataset contains a diverse distribution of medical scenarios.

Sample fields	Text
Conversation with patient (model’s prompt)	You are a doctor with extensive experience. I am a male patient. I am 52 years old. I came to you with a problem. Name the 3 most likely potential diagnoses based on the dialogue: <dialogue> <doctor> What’s bothering you? <patient> Pain behind the sternum, squeezing, radiating to the left arm. <doctor> When does pain occur? <patient> Usually when walking or if I get nervous. <doctor> How long does the attack last? <patient> 10-15 minutes, if I sit down, he lets go. <doctor> Are you taking any medications? <patient> Sometimes nitroglycerin. </dialogue> Don’t explain them, just write the 3 most likely potential diagnoses wrapped in tags Wrap each answer in <diag_diag> and </diag_diag> tags
Prediction of the model	<diag_diag>Stable angina</diag_diag> <diag_diag>Ischemic heart disease</diag_diag> <diag_diag>Cardialgia</diag_diag>

Table 11: Example of conversation with patient. The provided data sample is synthetic; however, it is close to the used train and test samples. The example is translated to English.

H Holdout Case Study: Diagnosis-Level Mismatches

To complement the aggregate UQ evaluation, we manually inspect the holdout split at the diagnosis level. The split contains 192 generated diagnoses from GigaChat-2-Max; 116 are negative under Relaxed EM. We group these formal mismatches into categories that reflect why a prediction failed to match the expert reference set.

Each mismatch was assigned one primary category based on the predicted diagnosis, the reference diagnoses, and the evidence available in the dialogue. We then aggregate the categories by the confidence tiers introduced above. Table 12 summarizes the resulting taxonomy.

Category	#	%	Low	Low+M	H/M/L
Taxonomy/granularity	40	34.5	13 (32.5)	38 (95.0)	2/25/13
Competing differential	36	31.0	15 (41.7)	35 (97.2)	1/20/15
Sparse evidence	20	17.2	9 (45.0)	17 (85.0)	3/8/9
Unsupported extrapolation	20	17.2	9 (45.0)	20 (100.0)	0/11/9
Total	116	100.0	46 (39.7)	110 (94.8)	6/64/46

Table 12: Manual taxonomy of diagnosis-level Relaxed-EM mismatches on the holdout split. “Low” counts mismatches assigned low confidence; “Low+M” counts mismatches assigned either low or moderate confidence. Values in parentheses are row-wise percentages. H/M/L gives high-, moderate-, and low-confidence counts.

The taxonomy highlights four qualitative UQ patterns. The most reliable pattern appears for unsupported specificity or extrapolation. This class includes predictions that add an etiology, complication, comorbidity, rare condition, or overly specific conclusion not supported by the dialogue. All 20 mismatches in this category fall into the low-or moderate-confidence region, and none receives high confidence. Thus, when the diagnostic generator moves beyond the available clinical evidence, UQ tends to downweight the prediction.

Taxonomy and label-granularity mismatches form the opposite pattern. They are the largest category, but only 13 of 40 are assigned low confidence. Many of these predictions are clinically close to the reference label, so UQ often assigns moderate confidence rather than rejecting them strongly.

Competing-differential mismatches occupy an intermediate regime. They correspond to plausible alternative diagnoses supported by the observed anamnesis but absent from the expert reference set. UQ assigns only one such mismatch high confidence and captures 35 of 36 under low-or-moderate confidence, suggesting that these outputs are treated as clinically ambiguous rather than clearly unsupported.

Sparse-evidence cases are qualitatively different.

They account for only 20 of 116 mismatches but produce 3 of the 6 high-confidence mismatches. These cases typically lack the discriminating clinical information needed to separate several plausible diagnoses. As a result, the formal mismatch may reflect underspecification of the input rather than an unsupported or clinically implausible generation. We hypothesize that such cases are primarily associated with aleatoric uncertainty and could potentially be detected by a separate guardrail before the diagnostic model attempts to generate a diagnosis. We leave the investigation of this hypothesis for future work.

We further audit extreme disagreements between the UQ score and Relaxed EM: high-confidence mismatches and low-confidence matched diagnoses. This targeted failure analysis identifies 8 cases with 10 diagnosis-level predictions. The cases cluster into taxonomy, synonymy, or subtype mismatches (3 cases), sparse or non-specific anamnesis (4 cases), and competing etiologies or treatment-related confounding (1 case). Representative examples include headache cases without sufficient criteria to distinguish migraine from tension-type headache, anorectal complaints without visual examination, and dermatological cases where procedure-related irritation, allergic reaction, and infection remain competing explanations.

Overall, the case study suggests that UQ is most informative for model-driven diagnostic overreach: extrapolations beyond the dialogue, unsupported specificity, and rare or comorbid conditions not grounded in the evidence. UQ is less decisive when uncertainty comes from the clinical input or from the evaluation setup itself, such as sparse anamnesis, non-specific symptoms, multiple plausible differentials, or close-but-not-identical labels. This explains why many formally incorrect predictions receive moderate, rather than low, confidence: the confidence signal separates clearly unsupported predictions from ambiguous or boundary cases more effectively than it resolves the boundary cases themselves.