

Safer On-Premise Medical AI: A Hybrid Framework for Hallucination Detection in Local LLMs

Sanne Wielinga¹, Clara Maathuis¹, Natasha Alechina^{1,2}

¹Open University of the Netherlands

²Utrecht University

sannwielinga@gmail.com, clara.maathuis@ou.nl, natasha.alechina@ou.nl

Abstract

LLMs (Large Language Models) are increasingly integrated into medical workflows to provide decision-making support, for example as Question Answering (QA) systems. However, in medical context, privacy requirements and resource constraints require the use of local and small (1B-parameter) LLMs. 1B-parameter LLMs are highly susceptible to hallucinations, and existing hallucination detection frameworks often suffer from poor calibration in these low-resource settings. To address this problem, we introduce a lightweight explainable hybrid detection method that integrates semantic entropy to quantify the uncertainty at the meaning-level in diverse response samples with an internal signal probe trained on the hidden states of the final layer of the model. The framework operates without external retrieval and consistently improves the detection of factual errors across biomedical benchmarks, outperforming single-signal baselines most clearly in calibration stability, the property that enables safe abstention in clinical workflows. These findings contribute to efforts dedicated to building transparent, safe, and privacy-preserving AI systems in medical decision-making support.

Code — <https://github.com/sannwielinga/hybrid-hallucination-detection>

1 Introduction

The integration of Large Language Models (LLMs) in the healthcare domain is surrounded by the tension between computational capability and deployable safety (Amadi and Ojo 2025). On the one hand, recent LLMs showed valuable results for clinically relevant tasks such as medical question answering (QA), documentation support, and information synthesis (Pal et al. 2025; Lv et al. 2025; Zeng et al. 2025). On the other hand, their real-world adoption is constrained by the governance demands applicable to this domain in which errors might have direct implications for patient safety (Aljohani et al. 2025). This tension is further amplified by regulation and institutional requirements related to the sensitivity of health data. To this end, in the EU the GDPR (European Parliament and Council of the European Union 2016) treats health data as a special category of personal data subject to heightened safeguards and in the US

the HIPAA Privacy Rule (U.S. Department of Health and Human Services 1996) contains standards for protecting individually identifiable health information. In parallel, the EU AI Act introduces a risk-based regulatory framework for AI and imposes stringent obligations on high-risk systems, and by this establishing requirements related to transparency, risk management, and accountability (European Parliament and Council 2024). Hence, the adoption of LLMs in healthcare needs to be evaluated both in terms of performance as well as in relation to privacy preservation, regulatory compliance, and the broader conditions required for safe, responsible, and trustworthy clinical deployment.

Furthermore, these constraints also render cloud-based API calls to proprietary models impractical for many clinical workflows. Consequently, healthcare institutions increasingly require on-premise solutions: smaller, resource-efficient LLMs that operate entirely on local infrastructure (Garg et al. 2025). Specifically, while 7B+ parameter models often require 24–48GB of VRAM, 1B-parameter architectures can run on standard clinical workstations, ensuring broad accessibility. This shift toward local deployment has an equity dimension: institutions that lack the budget for cloud API costs or the governance infrastructure for data-sharing agreements can nevertheless participate in AI-assisted healthcare if smaller models are designed, developed, and deployed in a responsible, safe, and trustworthy manner. Achieving this vision requires a human-centered AI approach (Shneiderman 2022) that prioritizes clinician oversight, interpretable safety signals, and harm-aware design.

Nevertheless, smaller models are more prone to hallucinations: they can generate fluent medical text that is factually incorrect or unsupported by the context (Ji et al. 2023). In critical medical contexts, a model that confidently fabricates a contraindication or misinterprets a diagnostic criterion poses an unacceptable safety risk.

Therefore, the viability of local medical LLMs hinges on both their generation quality and on the ability to reliably detect and abstain from answering when they are likely incorrect. The ethical imperative is clear: deploying an AI system that confidently provides incorrect medical information without adequate safeguards risks direct patient harm, making the detection and mitigation of such harms a critical requirement for responsible deployment.

Current approaches to hallucination detection present

a trade-off between accuracy and computational feasibility that is ill-suited for this resource-constrained setting. Retrieval- or evidence-grounded verification can reduce hallucinations, but introduces external dependencies, latency, and governance overhead that complicates on-premise deployment (Wada et al. 2025; Alu and Oluwadare 2026). At the same time, intrinsic uncertainty methods are of interest because they require only model-accessible signals, but the strongest results are often reported for larger models or in computationally unconstrained settings. Semantic self-consistency methods can be effective but require repeated inference passes (Farquhar et al. 2024) and internal state probes can be cheap at inference time, but are less studied for compact medical QA models and may produce uncalibrated scores if trained under noisy supervision. As a result, it remains unclear which uncertainty signals remain informative for smaller models such as $\sim 1\text{B}$ medical QA LLMs and how to combine them into a detector that can support abstaining from answering and requesting human review.

We address this gap by proposing a hybrid hallucination detection framework targeting complementary uncertainty signals in small, on-premise medical LLMs. The first signal, *Semantic Entropy* (SE), estimates meaning-level uncertainty by sampling a small set of diverse responses and measuring how probability mass spreads over distinct semantic answers. SE is sensitive to epistemic ambiguity and semantic instability, but it can fail when a model is confidently wrong and repeatedly samples the same incorrect answer. The second signal, an *Internal Signal* (IS) probe, trains a lightweight classifier on the model’s final-layer hidden state from a single low-temperature generation to predict incorrectness. IS can detect high-confidence failures that SE misses, however it relies on a training signal. In this context, correctness supervision is necessarily approximated at scale. Accordingly, we use automated overlap- and similarity-based correctness labels as *proxy* targets and interpret absolute metrics with appropriate caution, and further support key qualitative findings through expert annotation of hallucination subtypes. We combine SE and IS into a single $P(\text{Incorrect})$ score using simple averaging and a learned meta-combination by prioritizing calibration and deployability under on-prem constraints. Moreover, the framework is evaluated making use of a dual strategy: large-scale automated metrics using proxy labels, and a granular, expert-verified analysis of specific hallucination subtypes.

The contributions of this research can be summarized as follows:

- We formalize the challenge of hallucination detection specifically for local, $\sim 1\text{B}$ parameter medical LLMs.
- We propose a hybrid detector that integrates semantic entropy with an internal signal probe to produce a single $P(\text{Incorrect})$ score, improving discrimination (AUROC) over either signal alone in multiple configurations.
- We analyze probabilistic reliability by showing that internal signal and meta-combined scores can be substantially better calibrated than common self-evaluation baselines, enabling risk-aware filtering and review.
- We provide subtype-level analysis using expert annotations

and show a consistent, clinically salient asymmetry: factual hallucinations regarding indications/contraindications are easier to detect than conceptual errors in diagnostic definition/criteria, exposing a remaining blind spot for smaller-model medical QA reliability.

2 Related Work

Hallucinations represent outputs that are plausible yet factually incorrect, unfaithful to the input, or ungrounded in reliable sources, which makes them a significant barrier to reliable LLM deployment, particularly in domains such as healthcare (Ji et al. 2023; Huang et al. 2025). While both mitigation and detection strategies have been explored, this study focuses primarily on detection. Mitigation approaches, including Chain-of-Thought prompting (CoT) (Wei et al. 2022), Retrieval Augmented Generation (RAG) (Lewis et al. 2020), and iterative self-correction methods such as Self-Refine (Madaan et al. 2023) and CRITIC (Gou et al. 2023), aim to reduce hallucination occurrence during generation. Detection methods, by contrast, seek to identify hallucinated outputs post-generation, enabling abstention from answering or requesting human review. We focus on post-generation detection, as it provides a safety layer independent of the generation strategy and is compatible with on-premise deployment constraints.

2.1 LLMs in Healthcare and the Hallucination Challenge

The application of LLMs in biomedicine has expanded rapidly given their potential for clinical question answering, diagnostic support, and medical documentation (Lu et al. 2024). Specialized models such as Med-PaLM 2 demonstrated impressive performance on medical licensing exams and QA benchmarks (Singhal et al. 2025), while evaluations of general-purpose models such as GPT-4 showed strong results on medical challenge problems (Nori et al. 2023). Despite this potential, various considerations need to be taken into account considering the reliability and safety requirements imposed in the healthcare domain. Benchmarks such as Med-HALT (Pal, Umapathi, and Sankarasubbu 2023) and MedHallu (Pandit et al. 2025) were specifically designed to quantify hallucination rates in medical contexts, revealing non-trivial error rates even across highly capable models. Hallucinated medical outputs are particularly challenging because they are often articulated with a high degree of apparent confidence and may intertwine correct and incorrect information, making reliable detection difficult in the absence of domain expertise.

At the same time, given challenges such as the complex terminology, the need for nuanced clinical reasoning, and strict data privacy regulations, on-premise deployment is necessary using smaller and efficient LLMs (Rodríguez 2025). However, evidence suggests that smaller models may be inherently more susceptible to knowledge gaps and hallucinations, particularly on specialized topics (Lauscher, Glavaš et al. 2025; Moor et al. 2023). This motivates the development of robust detection mechanisms specifically

for the 1B-parameter class of models feasible for resource-constrained clinical environments.

2.2 Hallucination Detection Approaches

Detection methods can be broadly categorized by the degree of model access they require. Black-box detectors treat the generator as an input-output system, deriving reliability signals from token probabilities or self-evaluation. Token-level likelihood heuristics such as average log probability, perplexity, and naive token entropy, are known to be weak proxies for factual correctness, as models can assign high probability to fluent but incorrect text (Farquhar et al. 2024). Self-assessment methods like P(True), prompt the model to judge whether its own answer is correct (Kadavath et al. 2022). Nevertheless, these methods can be prompt-sensitive and poorly calibrated. Recent studies have moved beyond surface self-report. For instance, IRIS prompts models to verify statement truthfulness and uses verification uncertainty as a soft signal for hallucination detection (Srey, Wu, and Tuan 2025).

Consistency-based detectors use multiple sampled generations, treating disagreement as evidence of uncertainty. SelfCheckGPT considers this principle for zero-resource faithfulness evaluation by comparing sampled outputs without external references (Manakul, Liusie, and Gales 2023). SE addresses the limitation that lexical variation obscures semantic agreement by clustering sampled answers using bidirectional entailment and computing Shannon entropy over semantic clusters (Farquhar et al. 2024; Kuhn, Gal, and Farquhar 2023). Moreover, SE can distinguish the same meaning expressed in different ways from genuine semantic disagreement, but it incurs additional inference passes and can fail when a model is consistently wrong. (Vipulanandan, Premaratne, and Sarkar 2026) shows that robust semantic-level uncertainty can remain informative even under deployment constraints such as quantization and short generation lengths.

White-box methods use internal signals such as hidden states, attention patterns, or neuron activations. Unsupervised methods like MIND establish baseline activation patterns and identify deviations as potential hallucinations (Su et al. 2024). At the same time, supervised probes train lightweight classifiers on hidden representations to predict correctness labels (Orgad et al. 2025; Azaria and Mitchell 2023). Because hidden states can encode latent uncertainty not reflected in the final token distribution, probes can complement sampling-based methods. More computationally efficient alternatives have also been proposed. For example, RAUQ identifies uncertainty-sensitive attention heads to enable single-pass uncertainty estimation (Vazhentsev et al. 2026), whereas HALT introduces lightweight probes over intermediate hidden representations and supports selective routing to more powerful verification mechanisms when warranted (Bhatnagar et al. 2026).

Nevertheless, (Song et al. 2025) stresses that different types of hallucination can defeat various detection paradigms. This motivates the choice of considering a hybrid approach that combines complementary signals. This is especially relevant in the healthcare domain, where (Pandit

et al. 2025) shows that even strong LLMs struggle to reliably detect clinically relevant hallucinations, particularly when hallucinated answers are semantically close to correct ones.

The evaluation of response correctness remains a methodological challenge as well. To this end, computationally inexpensive token-overlap measures such as SQuAD F1 (Rajpurkar et al. 2016) provide a practical basis for large-scale assessment but are inherently limited in their ability to capture semantic equivalence beyond exact lexical correspondence. By contrast, semantic similarity metrics like BERTScore (Zhang et al. 2019) provide a richer representation of meaning while being indirect proxies for factual correctness. As a result, automated labeling procedures inevitably introduce a degree of noise into both detector evaluation and the training of supervised components which calls for complementary expert-verified evaluation to support more robust and clinically meaningful assessment.

2.3 Explainability for LLM Reliability

The opacity of LLMs creates a direct need for transparency about reliability, particularly in clinical settings where trust is derived not from perfect accuracy but from understanding a model's reasoning process and its limitations (Tonekaboni et al. 2019). Despite the expansion of XAI methods, their integration into real-time patient care remains limited, largely due to the lack of agreed-upon evaluation frameworks (Jung et al. 2023). Iterative co-design approaches that involve clinicians as active participants in the development of explanation interfaces have shown promise for improving both trust and usability (Panigutti et al. 2023), yet such human-centered evaluation remains rare in the context of LLM-based safety systems.

In the context of hallucination detection, the most relevant paradigm shifts the focus of explanation from the LLM's internal generative process to the auxiliary safety systems built around it. Wu et al. (2024) advocate for this approach under the concept of Usable XAI, arguing that explaining the decisions of a safety classifier is often more practical than attempting to explain the full generative process. In this way, model-agnostic methods such as SHAP (Lundberg and Lee 2017) can attribute importance to each input signal for a given detection decision assuring both global transparency (i.e., which signals matter most across instances) and local interpretability (i.e., why this specific output was flagged). This user-centric philosophy implies that an explanation's value is an empirical quality to be measured rather than a technical property to be asserted (Meske et al. 2022; Ali et al. 2023). Recent practitioner studies reinforce this point, finding that generic application of explainability methods without domain-specific evaluation and stakeholder involvement fails to meet user needs (Dhaini et al. 2025), which motivates our approach of grounding SHAP-based explanations in a clinician-evaluated, domain-specific safety context.

In general, while numerous studies tackle LLM hallucinations, a gap emerges at the junction of model size, application domain, and detection methodology: techniques such as SE (Farquhar et al. 2024) and IS probing (Orgad et al. 2025) have predominantly been evaluated on large-

scale models in general-domain settings, whereas privacy-sensitive healthcare deployment relies on smaller, institutionally hosted models that may be comparatively more prone to hallucination (Lauscher, Glavaš et al. 2025). Existing hybrid strategies typically couple retrieval-based verification with uncertainty signals (Lewis et al. 2020), introducing external dependencies that complicate on-premise deployment, while the two intrinsic paradigms exhibit a theoretically motivated complementarity: SE can produce low entropy when a model repeatedly generates the same incorrect response, whereas IS remains sensitive to such high-confidence failures through the model’s internal representations (Orgad et al. 2025; Su et al. 2024).

We are explicit about what is novel here versus reused: SE (Farquhar et al. 2024) and internal state probing (Orgad et al. 2025) are established techniques; the contribution of this work lies in their calibrated fusion, the domain-specific hallucination taxonomy, and the systematic evaluation of both in the underexplored $\sim 1\text{B}$ on-premise regime. Hence, this research addresses this gap by proposing a hybrid hallucination detection system applied in the healthcare domain by examining the behavior of these signals in $\sim 1\text{B}$ -parameter models in two medical question-answering benchmarks, combining them within a calibrated hybrid detection framework, and evaluating the framework with respect to both detection performance and transparency for clinical end users.

3 Methodology

To address the hallucination detection problem for small, on-premise LLMs, this research proposes a hybrid hallucination detection framework by combining complementary uncertainty signals that capture different failure modes of hallucination using SE with an IS probe based on a lightweight classifier trained on the model’s final-layer hidden state from a single low-temperature generation to predict incorrectness.

For this, we formalize the problem of medical hallucination detection as follows. Let M be a language model that produces a response y given a medical query x . The objective is to learn a risk scoring function

$$S(x, y) \in [0, 1] \quad \text{interpreted as } P(\text{Incorrect} \mid x, y),$$

which can be used to abstain from answering when risk exceeds a safety threshold.

Since collecting expert human annotations at the scale required for training is challenging in practice, a *proxy* target variable $z \in \{0, 1\}$ derived from automated metrics (e.g., SQuAD F1 or BERTScore against a reference) is used. Here, $z = 1$ denotes an “incorrect” answer. We acknowledge that z is a noisy label. However, we hypothesize that training on these proxies allows the model to learn generalizable patterns of error, which we subsequently validate against expert human annotations.

3.1 Semantic Entropy (SE)

Semantic entropy estimates meaning-level uncertainty from the distribution of sampled answers. Unlike “naive” entropy, which penalizes lexical variation, SE clusters responses by semantic equivalence.

Aligned with (Farquhar et al. 2024), we generate a set of N stochastic samples $\mathcal{Y} = \{y^{(1)}, \dots, y^{(N)}\}$ using high-temperature sampling. We define a semantic equivalence relation $y^{(i)} \sim y^{(j)}$ using a bidirectional entailment check with a domain-specific Natural Language Inference (NLI) model. In this context, two responses are equivalent if and only if they mutually entail each other. This relation partitions $\mathcal{Y}$ into K semantic clusters $\{C_1, \dots, C_K\}$.

The probability mass of a cluster C_K is estimated by marginalizing over its constituent sequences:

$$p(C_k) = \frac{\sum_{y \in C_k} p_\theta(y \mid x)}{\sum_{y' \in \mathcal{Y}} p_\theta(y' \mid x)}.$$

The *Semantic Entropy* score is then computed as the Shannon entropy over this cluster distribution:

$$\text{SE}(x) = - \sum_{k=1}^K p(C_k) \log p(C_k),$$

(where $\log$ is base-2 or natural; the choice only rescales). When $K = 1$, all samples agree semantically and $\text{SE}(x) = 0$.

For comparability across varying K , we also use normalized semantic entropy $\text{NSE}(x) = \text{SE}(x) / \log K$ when $K > 1$, and 0 otherwise. Here, a higher SE/NSE indicates greater dispersion over distinct semantic answers and thus higher likelihood of hallucination.

3.2 Internal Signal Probe (IS)

Semantic uncertainty can remain low even when the model produces a confident but incorrect response and repeatedly samples the same wrong answer. The IS probe explicitly targets this confidently wrong failure mode, which SE cannot capture by design. To capture complementary signals associated with such failures, we train a probe on the internal representations of the primary generation.

We define the *Internal Signal* as a function of the model’s hidden states during the generation of the primary answer $\hat{y}$. Let $\mathbf{h}_T \in \mathbb{R}^d$ denote the activation vector of the final transformer layer at the last generated token T . Following (Orgad et al. 2025), we hypothesize that $\mathbf{h}_T$ contains linearly separable information regarding the correctness of the sequence.

For this, we train a supervised probe $f_\phi : \mathbb{R}^d \rightarrow [0, 1]$ to predict the proxy correctness label z :

$$\text{IS}(x) = f_\phi(\mathbf{h}_T) = \sigma(\mathbf{w}^\top \mathbf{h}_T + b),$$

where σ is the sigmoid function.

The probe is a logistic regression classifier trained per model–dataset configuration on the proxy labels z , using only final-layer activations and no access to reference answers at inference time; the complete training configuration and pipeline are available in the released code repository. Because z is itself a noisy proxy, the probe may partly fit label artefacts rather than incorrectness as such; the BERTScore-based sensitivity analysis (Section 4.2) mitigates, but does not eliminate, this concern.

To prevent data leakage during evaluation and when the outputs are used as features in a downstream model, we compute IS scores using stratified K -fold cross-validation and retain out-of-fold probabilities for each instance.

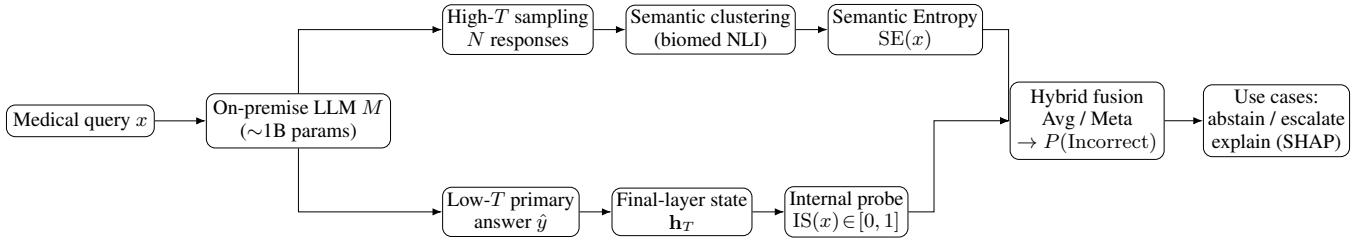


Figure 1: Hybrid hallucination detection framework architecture.

3.3 Hybrid Framework

The two modules capture complementary dimensions of hallucination risk: SE reflects semantic instability across generated responses, whereas IS estimates latent patterns associated with incorrectness. Their integration is not straightforward since SE is an unbounded information-theoretic measure, while IS is expressed as a probability in $[0, 1]$. To address this difference in scale and interpretation, we introduce a fusion strategy that normalizes both signals into a unified risk metric.

Feature Normalization. Within each cross-validation fold, we apply Min-Max scaling $\text{norm}(u; x) = (u(x) - u_{\min}) / (u_{\max} - u_{\min})$ to map raw features (SE, NSE, IS) to $[0, 1]$, with $(u_{\min}, u_{\max})$ computed on the fold’s training split only.

Hybrid (Avg). The simplest combination assumes equal contribution after normalization: $H_{\text{avg}}(x) = \frac{1}{2} \text{norm}(\text{SE}; x) + \frac{1}{2} \text{norm}(\text{IS}; x)$. This yields an interpretable baseline that requires no additional learned parameters beyond the IS probe. We emphasize that H_{avg} is treated as a discriminative risk score rather than a calibrated probability: Min-Max scaling serves only to place the unbounded SE score and the $[0, 1]$ IS score on a common scale for fusion. All calibration claims and abstention thresholds pertain to the probabilistic outputs of the IS probe and Hybrid (Meta), for which we report ECE.

Hybrid (Meta). To learn an adaptive combination, we train a meta-classifier m_ψ on a compact feature vector $v(x) = [\text{norm}(\text{SE}; x), \text{norm}(\text{IS}; x), \text{norm}(\text{RL}; x)]$, where RL is the response length (token count) as an auxiliary heuristic, based on the observation that hallucinations in smaller models often correlate with verbosity. The hybrid score $H_{\text{meta}}(x) = m_\psi(v(x)) \in [0, 1]$ is trained on the same proxy labels and evaluated with out-of-fold predictions. In this context, we explore both linear (Logistic Regression) and non-linear (Random Forest) meta-learners. Compared to averaging, m_ψ can attenuate the influence of potentially misleading signals in cases of disagreement, such as low SE and high IS, by learning dataset-specific interactions between the underlying uncertainty signals. Note that response length enters only this secondary meta-learned variant. The primary Hybrid (Avg) detector and all single-signal results use no length feature, so detection performance does not depend on response verbosity.

We retain both variants because they serve distinct purposes. Hybrid (Avg) is the primary discriminative method: it achieves the best AUROC in the majority of configurations (Section 4.2) and requires no additional learned parameters beyond the IS probe. It is also the variant reported as “Hybrid” in Table 5. Hybrid (Meta) is reported alongside in Table 3 to quantify the trade-off between adaptive combination and calibration, and its learned coefficients enable the global feature importance analysis in Section 4.5.

3.4 Hallucination Taxonomy for Medical QA

To analyze clinical risk beyond a single correctness label, we employ a domain-specific taxonomy developed in consultation with a medical expert with 15+ years of clinical experience. It focuses on factuality hallucinations, in which the model contradicts established medical knowledge, as these are the primary safety concern in the clinical QA setting studied here.

We distinguish five factuality subtypes (Table 1). The A-codes group hallucinations that contradict specific, verifiable medical information: medication dosage or administration errors (A1) involve incorrect details about dose, frequency, or route of administration; indication or contraindication errors (A2) occur when the model incorrectly states that a treatment is safe when it is contraindicated, or vice versa; and diagnostic criteria or definition errors (A3) capture misrepresentations of established diagnostic criteria, disease definitions, or symptom timelines. The B-codes group quantitative or fabrication-based error: statistical or epidemiological errors (B1) involve incorrect prevalence, incidence, or risk estimates; and fabricated entities, guidelines, or studies (B2) occur when the model invents non-existent sources, authorities, or clinical protocols. A1 is evaluated separately on the rule-based synthetic dosage task (Section 4.4), where ground truth is unambiguous; A2, A3, B1, and B2 are evaluated on the expert-annotated subset.

An example of a Type A2 error with direct patient safety implications is as follows. When asked whether Ibuprofen can be taken with Lisinopril, a model responded affirmatively, failing to recognize that NSAIDs (Nonsteroidal Anti-Inflammatory Drugs) can blunt the effect of ACE (Angiotensin-Converting Enzyme) inhibitors and harm kidney function. An example of a Type A3 error is the following: when asked to list DSM (Diagnostic and Statistical Manual of Mental Disorders) criteria for major depressive disorder, a model stated a minimum mood duration of two months rather than the correct two weeks answer. A Type

Code	Definition
A1	Medication dosage/administration error
A2	Indication/contraindication error
A3	Diagnostic criteria/definition error
B1	Statistical/epidemiological error
B2	Fabricated entity, guideline, or study

Table 1: Domain-specific taxonomy of medical factuality hallucinations used for expert annotation. A-codes group contradictions of specific factual medical information; B-codes group quantitative or fabrication-based errors.

A3 error is subtler and harder to detect through uncertainty signals since the response is structurally coherent and semantically close to the correct answer.

This taxonomy is used exclusively for the expert-labeled evaluation subset and does not affect the training process. Intuitively, A2 errors are anchored to concrete, verifiable facts and tend to produce distinct uncertainty signals, whereas A3 errors can be semantically close to correct statements, making them harder to detect. This distinction drives the clinically significant detectability gap we report in our results.

3.5 Explainability Methods

To enhance the transparency and operational usefulness of the proposed hybrid framework (Figure 1), a lightweight XAI component is added to understand how the underlying signals contribute to each risk prediction (Das and Rad 2020). This layer provides post-hoc explanations that attribute detection results to the combined effects of SE, IS, and response length. At the global level, feature-importance analysis is used to investigate how the meta-learner adjusts its reliance on these signals according to dataset characteristics reflecting task-specific patterns of hallucination. At the local level, SHAP-based explanations are used to audit individual predictions made by showing how particular features increase or decrease the final risk estimate. Hence, this module allows for better understanding both successful detections and failure cases, including instances in which the model is confidently but incorrectly factual, which contributes to strengthening the safety and accountability of the framework.

4 Experiments and Results

4.1 Experimental Setup

In this research, ~ 1 B-parameter models for on-premise deployment are considered. We evaluate four architectures: Llama-3.2-1B-Instruct, TinyLlama-1.1B-Chat-v1.0, its domain-adapted variant tiny-llama-1.1b-chat-medical, and Granite-3.1-1B.

These architectures were selected so that findings generalize across the 1B-parameter class rather than reflecting a single model type: they span large-scale pre-training with Grouped-Query Attention (TinyLlama), domain specialization (its medical variant), a newer architectural generation with an improved tokenizer (Llama-3.2-1B), and a Mixture-

	BioASQ	MedQuAD
<i>Dataset Properties</i>		
Validation questions	819	11,750
Question types	4	37
Mean ref. length (chars)	16.6	95.1
<i>Model Error Rates (%), SQuAD F1 < 0.5</i>		
Llama-3.2-1B	58.2	82.4
TinyLlama v1.1	71.3	86.9
TinyLlama-med	73.5	91.2
Granite-3.1-1B	62.4	89.7
<i>Expert Annotation</i>		
Annotated subset (N)	480	480

Table 2: Dataset characteristics and model baseline error rates. The high error rates of ~ 1 B models on both benchmarks underscore the need for reliable detection: the majority of generated answers are incorrect under automated labeling.

of-Experts design that increases effective capacity without proportional compute cost (Granite-3.1-1B).

Furthermore, the evaluation is performed on two medical QA benchmarks: (i) **BioASQ** which contains expert-level biomedical QA using a validation set of $n = 819$, and (ii) **MedQuAD** which contains patient-facing questions using a validation set of $n = 11,750$. Since the reference answers of MedQuAD can be long, we use concise summaries of the reference answers to support automated evaluation at scale. These benchmarks capture two distinct clinical communication modalities: expert-to-expert and patient-to-system. Insights into key dataset characteristics and model error rates are summarized in Table 2.

The primary responses are generated using low-temperature greedy decoding ($T = 0.1$) to obtain a stable baseline output for downstream analysis. For SE estimation, we sample ($N = 10$) candidate responses at a higher temperature ($T = 0.8$) to capture semantic variability across plausible generations. Semantic clustering is then performed using BiomedNLP-PubMedBERT, fine-tuned for NLI (Natural Language Inference) which serves as the entailment model for assessing semantic equivalence among the sampled responses.

In the evaluation process, a multi-tiered evaluation strategy is considered. First, we rely on **automated correctness labels** for scale, using thresholds on **SQuAD F1** (SQuAD F1 < 0.5) and **BERTScore F1** (BERTScore F1 < 0.85). Primary results use SQuAD F1-derived labels, while sensitivity uses BERTScore F1-derived labels.

Automated proxy labels enable scale but cannot fully characterize clinically meaningful hallucination modes. We therefore conduct an *initial exploratory* expert-driven evaluation designed as the first step of a human-centric validation program. Specifically, we sample a subset of model outputs from the pooled runs and annotate them with the taxonomy of Section 3.4. The annotations were performed in consultation with a medical expert with 15+ years of clinical experience. The results are treated as qualitative evidence and used to test whether clinically high risk factual hallucinations (e.g., contraindications) are systematically more de-

tectable than subtle conceptual errors (e.g., diagnostic criteria). Scaling to multi-expert panels and formal user study protocols is part of future research.

In this process, we employ a rule-based synthetic dosage dataset to test detection performance on purely numerical hallucinations where ground truth is unambiguous. The dataset consists of weight-based pediatric dosage calculations with explicit formulae provided in the prompt, each presenting a drug, a target dose in mg/kg/day, and a patient weight, requiring the model to compute the correct daily dose. This task is designed to evaluate the detector under unambiguous ground truth, not the generator. The weak arithmetic performance of $\sim 1B$ models is expected and orthogonal to the question of whether their errors can be detected.

We report discrimination via the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC), calibration via Expected Calibration Error (ECE) for detectors that output probabilities, and a practical abstention proxy via the Area Under the Rejection Accuracy Curve (AURAC). All experiments were conducted on a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB VRAM), demonstrating that the complete pipeline, including generation and detection, is feasible on standard clinical workstation hardware. In this context, the effectiveness of the framework is evaluated through three lenses: discriminative performance, safety alignment, and interpretability.

4.2 Overall Detection Performance

Table 3 summarizes model performance under SQuAD-F1 proxy labeling. Across 6 of the 8 model–dataset configurations, the **Hybrid (Avg)** method achieves the highest discriminative performance, as measured by AUROC. This result provides empirical support for the proposed framework reinforcing that semantic output inconsistency and internal-state-based signals capture complementary aspects of hallucination risk and, when combined, improve detection effectiveness. The AUROC gains over the best single signal are modest in absolute terms; the central contribution of the fusion lies in its consistency across configurations, combined with the calibration stability of the probabilistic components, which is what enables safe abstention thresholds in clinical workflows.

For example, using TinyLlama v1.1 on the MedQuAD dataset, the hybrid method reaches 0.930 AUROC, outperforming both SE (0.911) and IS (0.917) individually. While the IS probe alone is a strong discriminator, often matching SE, its primary advantage lies in calibration. The IS probe consistently achieves low ECE (< 0.06), significantly outperforming the uncalibrated $P(\text{True})$ baseline. This indicates that the probability scores produced by our internal probe are reliable proxies for actual correctness which is a prerequisite for setting safe abstention thresholds in clinical workflows. At the same time, this shows that the fusion of these components captures genuinely complementary information.

Despite the comparatively stronger discrimination performance observed, the practical utility of all methods, as reflected by AURAC, remains limited, particularly on the MedQuAD dataset (generally 0.28–0.34), compared

to BioASQ dataset (0.54–0.63). This shows an important boundary condition of the proposed method: even an effective detector can deliver only modest gains through abstention when the underlying model exhibits low baseline accuracy because the number of correct responses that can be preserved after filtering is intrinsically limited. Since AURAC is highly sensitive to this base-rate effect, the elevated error rates observed in 1B models (58–91%; Table 2) constrain the practical benefit attainable through rejection-based filtering alone. Accordingly, these results support framing the proposed framework as a risk indicator to assist human review, rather than as a standalone autonomous filtering mechanism.

Sensitivity analysis using BERTScore-derived labels (BERTScore F1 > 0.85) confirmed the overall trends on BioASQ, with AUROC scores remaining comparable. On MedQuAD, a consistent performance reduction was observed across all detectors, reflecting the dataset’s characteristics of containing diverse but valid paraphrased answers that BERTScore correctly identifies as correct, thereby making the discrimination task more difficult. Key qualitative findings, including strong IS calibration (ECE < 0.06) and low AURAC, persisted across both labeling schemes, suggesting the framework’s probabilistic estimates are robust to the choice of correctness proxy.

4.3 Subtype Analysis

From an ethical and safety perspective, the potential risk and harm associated with hallucinations varies across error types and application contexts (Granstedt et al. 2025; Artsi et al. 2025). For instance, a fabricated contraindication is life-threatening, and a verbose definition is merely inefficient. To understand this risk profile, we analyzed the aggregated Z-normalized uncertainty scores for the expert-annotated subset ($N = 960$ total, $N = 480$ per dataset).

Table 4 reports the mean Z-score normalized uncertainty per factuality subtype. Across both datasets and across SE, IS, and the hybrid detector, A2 (indication/contraindication) errors receive substantially more extreme uncertainty scores than any other factuality subtype, while A3 (diagnostic criteria/definition) errors sit close to the overall mean. B1 and B2 errors fall between these extremes.

This detectability gap is also statistically significant. Mann–Whitney U tests on the A2 vs. A3 comparison yield $p < 10^{-4}$ across all detectors, with medium-to-large effect sizes (rank-biserial $r = -0.47$ for SE on BioASQ to $r = -0.70$ for IS on MedQuAD), confirming that the separation is not merely significant but practically meaningful. Given the small size of some subtype cells, we report rank-biserial effect sizes alongside exact p -values. The nonparametric tests remain valid at these sample sizes, but subtype-level results should be read as qualitative evidence consistent with the exploratory scope of the expert evaluation.

The MedQuAD numbers are particularly striking: the IS probe assigns A2 errors a mean Z-score of -3.22 , more than three standard deviations below the population mean, whereas A3 errors sit within $+0.14$ of it. This suggests that when smaller models hallucinate a specific, verifiable medical fact, the internal and semantic signals are strong and

Dataset	Model	SE	IS	Hybrid(Avg)	Hybrid(Meta)	ECE(IS)	ECE(HybridMeta)
BioASQ	Llama-3.2-1B	0.860	0.826	0.867	0.873	0.033	0.072
BioASQ	TinyLlama v1.1	0.887	0.876	0.907	0.885	0.042	0.063
BioASQ	TinyLlama-med	0.893	0.915	0.922	0.900	0.035	0.069
BioASQ	Granite-3.1-1B	0.869	0.811	0.877	0.840	0.057	0.075
MedQuAD	Llama-3.2-1B	0.893	0.862	0.899	0.862	0.022	0.056
MedQuAD	TinyLlama v1.1	0.911	0.917	0.930	0.901	0.023	0.036
MedQuAD	TinyLlama-med	0.893	0.929	0.903	0.879	0.022	0.035
MedQuAD	Granite-3.1-1B	0.911	0.915	0.924	0.898	0.024	0.065

Table 3: AUROC (higher is better) and calibration (ECE; lower is better) using automated SQuAD-F1 proxy labels.

Dataset	Subtype	N	SE	IS	Hybrid (Avg)	Hybrid (Meta)
BioASQ	A2 Indication/Contraindication	54	-1.07	-1.32	-1.25	-1.25
	A3 Diagnostic Definition	294	+0.10	+0.09	+0.10	+0.10
	B1 Statistical/Epidemiological	26	+0.34	+0.46	+0.41	+0.38
	B2 Fabricated Entity/Guideline	19	+0.31	+0.49	+0.41	+0.36
MedQuAD	A2 Indication/Contraindication	16	-2.21	-3.22	-2.94	-1.92
	A3 Diagnostic Definition	321	+0.12	+0.14	+0.15	+0.10
	B1 Statistical/Epidemiological	19	+0.03	+0.09	+0.06	+0.06
	B2 Fabricated Entity/Guideline	8	+0.55	+0.20	+0.45	+0.78

Table 4: Mean Z-score normalized uncertainty by factuality subtype on the expert-annotated subset. Lower (more negative) scores indicate that the detector assigns higher uncertainty to that subtype and therefore detects it more reliably. Across both hybrid variants, A2 errors receive markedly more extreme scores than A3, B1, and B2 on both datasets. The two hybrid variants agree closely except on small subsets (notably MedQuAD A2 with $N = 16$ and MedQuAD B2 with $N = 8$), where the meta-learner’s adaptive weights deviate from simple averaging.

distinct; when they hallucinate conceptual definitions (A3), the error signal is noisier. This asymmetry is clinically fortunate: the framework acts as a guardrail for the most acute safety risks, even if it struggles with nuance.

Model (Acc.)	SE	IS	Hybrid	ECE _{IS}
TinyLlama (20.4%)	0.625	0.981	0.973	0.018
TinyLlama-med (13.3%)	0.773	0.995	0.975	0.017
Llama-3.2-1B (0.6%)	0.686	0.997	0.996	0.008
Granite-3.1-1B (13.3%)	0.515	0.965	0.943	0.038

4.4 Synthetic Dosage Task

To evaluate detection on numerical hallucinations with unambiguous ground truth, we employed a rule-based synthetic dosage calculation task. Baseline model performance differed markedly across the models: TinyLlama achieved 20.4%, while Llama-3.2-1B solved only 0.6% correctly, consistent with known limitations of small LLMs in precise numerical reasoning (Mirzadeh et al. 2024).

The IS probe was the dominant signal for this task, achieving AUROC scores above 0.96 across all models, substantially outperforming SE (AUROC 0.52–0.77). This indicates that for errors involving misapplication of explicit numerical rules, final-layer activations encode confidence cues that semantic consistency across samples cannot capture. SE’s weakness here is expected: a model that consistently applies an incorrect calculation will produce semantically consistent but factually wrong outputs, yielding low entropy. Calibration remained excellent for the IS probe (ECE < 0.04), while AURAC remained low due to the models’ poor baseline accuracy on the task itself. These results are summarized in Table 5 and reflect again the complementary nature of the two signals: SE excels at detecting semantic divergence in open-ended generation and IS captures confidence anomalies in rule-based reasoning.

Table 5: AUROC on the synthetic dosage task. The IS probe dominates, confirming that internal representations capture confidence signals invisible to sampling-based SE. Percentages indicate model baseline accuracy; ECE is reported for IS only.

4.5 Explainability

For on-premise deployment, a single risk probability is often insufficient. Clinicians need to understand *which signals* drove the warning to decide whether to override, abstain, or escalate to human review. We therefore accompany the hybrid detector with lightweight post-hoc explanations that attribute each decision to the underlying intrinsic signals.

Detection Signals Interpretation. The analysis conducted on global feature importance shows that the meta-learner adapts to the dataset. On BioASQ (expert-level queries), the logistic regression meta-learner assigns its highest coefficient to SE (2.16), followed by IS (1.48), and response length (0.78), relying heavily on semantic consistency. On MedQuAD (consumer queries), the random forest meta-learner distributes Gini importance more evenly: IS (0.38), SE (0.35), and response length (0.27), reflecting a more balanced integration of both signals.

At the local level, SHAP waterfall plots enable detailed

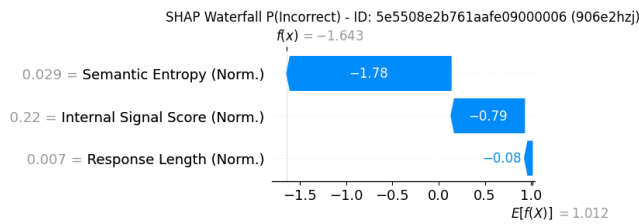


Figure 2: SHAP waterfall plot for a false negative detection (missed A2 hallucination). The model’s high confidence, reflected in low SE and IS scores, pushed $P(\text{Incorrect})$ to 0.162 despite the output being factually incorrect. This illustrates the “confidently wrong” failure mode where parametric errors evade both uncertainty signals.

auditing of individual failure cases. For example, in a **True Positive** detection of a fabricated syndrome definition, SHAP revealed that both high SE and high IS pushed the risk score to 1.0. Conversely, we identified **False Negatives** where the model was “confidently wrong” by generating a concise, logically consistent, but factually incorrect “Yes” to a contraindication query. In these cases, the low SE and IS scores explain *why* the detector failed: the LLM’s error was parametric, not stochastic (Figure 2). A systematic, clinician evaluated assessment of the reliability of these explanations across case types is a next step of the human-centered validation program outlined in Section 6.

Initial Expert-Based Evaluation. To validate the utility of these explanations, we conducted a qualitative study with a medical expert. In this process, each scenario was evaluated on three metrics: clarity (understanding of the detector’s reasoning), utility (actionability for clinical decisions), and trustworthiness (appropriate calibration of confidence in the system). The expert reviewed the model outputs alongside SHAP explanations for correct, hallucinated, and false positive scenarios. We stress the qualitative scope of this evaluation: with a single expert it can surface usability insights and candidate hypotheses, but it supports no quantitative reliability or generalizability claims. Multi-specialty panels with pre-registered protocols are part of future work.

The expert rated the clarity of the explanations highly (4/5), but the primary finding was the shift in trust calibration.

“[The explanation] helps me understand the system has weaknesses. Seeing a missed error reduces my trust somewhat, but it is a good reminder to remain skeptical.” – Medical Expert

The expert noted that while a binary flag is sufficient for obvious errors, the breakdown is vital for False Positives. When the detector flagged a correct answer, the expert could use the explanation to see that the flag was driven by “semantic noise” rather than internal uncertainty, allowing them to override the system with confidence. This confirms that explainable detection supports a “human-in-the-loop” workflow where AI has a supportive role and not a replacing one.

5 Discussion

In the healthcare domain, LLM deployment is often framed as a trade-off between proprietary API-based models, which may offer stronger performance but raise security, privacy, and sovereignty concerns, and locally deployed open models, which provide greater control and compliance but historically weaker clinical performance. Our findings show this trade-off can be mitigated through methods such as the proposed hybrid detection framework, with several implications for responsible AI deployment in resource-constrained environments, aligned with the principles of human-centered AI (Shneiderman 2022).

5.1 Implications for Responsible On-Premise Deployment

Data protection frameworks such as GDPR and HIPAA constrain the adoption of state-of-the-art cloud-based LLMs in many healthcare settings. As a result, the broader democratization of medical AI increasingly depends on the responsible, safe, and effective deployment of smaller and locally hosted models. Our findings indicate that although 1B-parameter models are prone to error, their failures are not random. By combining meaning-level uncertainty with internal representations, we achieved discrimination scores that support safer deployment of these smaller models. This suggests that post-hoc reliability layers can serve as a viable alternative to scaling model size, enabling hospitals to maintain data sovereignty without exposing patient data. A concrete deployment scenario is a clinical workstation providing decision-support QA under mandatory human review, in settings that lack local retrieval infrastructure or the governance arrangements required for external data flows.

This finding has implications beyond individual institutions. As regulatory frameworks such as the EU AI Act (European Parliament and Council 2024) increasingly classify medical AI systems as high-risk, the ability to demonstrate transparent safety mechanisms has become a compliance requirement. The hybrid framework’s capacity to produce calibrated risk scores accompanied by interpretable explanations directly addresses the transparency and accountability requirements that such regulations impose. In resource-limited settings, on-premise deployment of smaller models with robust detection may represent the only viable pathway to AI-assisted healthcare that respects both patient privacy and institutional data governance.

Furthermore, the subtype-level analysis has direct implications for deployment design. The consistent finding that factual indication/contraindication errors (A2) are substantially easier to detect than diagnostic definition/criteria errors (A3) suggests that a responsible deployment strategy should not rely on a single binary flag. Instead, a tiered alerting approach could leverage the distinct uncertainty signals. This implies that a high SE score could trigger a conceptual uncertainty warning, prompting the clinician to review the response for semantic inconsistencies, while a high IS score could trigger a more critical factual hallucination alert, recommending verification against established sources. In these lines, the explainability component is essential to making

such a tiered system usable, as it allows a clinician to understand the evidence behind any given alert and to decide whether to override, accept, or escalate the system’s judgment. Recent work on accountability in healthcare AI advocates that such decision-level transparency should function as a mechanism for joint accountability between developers and clinical users, rather than placing the burden on any single actor (Bagave et al. 2025).

5.2 Calibrated Trust and Human-AI Collaboration

The qualitative feedback from the medical expert emphasizes that the value of a system lies in its transparency. By visualizing why a response was flagged, the framework shifts the clinician’s role from a passive recipient of information to an active auditor. This reduces the risk of automation bias, where users over-rely on system outputs by explicitly surfacing the model’s own uncertainty.

As shown in the expert evaluation, the signal-level breakdown proved most valuable in ambiguous cases such as false positives, supporting a human-in-the-loop workflow in which AI supports rather than replaces clinical judgment. From an ethical standpoint, this transparency serves a dual purpose: it protects patients by ensuring that flagged outputs receive human review, and it protects clinicians by providing an auditable rationale for their decision to accept or override an AI recommendation (Bagave et al. 2025).

5.3 Limitations

Despite these promising results, several limitations exist and are further acknowledged.

First, the IS probe and meta-learner are trained on automated proxy labels derived from SQuAD F1 and BERTScore thresholds. These labels conflate factual incorrectness, incompleteness, and acceptable paraphrase, and may under-penalize semantically close but clinically important errors. Importantly, the supervised components are both trained and evaluated on these proxy labels, so systematic label biases can shape the learned decision boundary itself. The synthetic dosage task partially addresses this by providing rule-based ground truth independent of the proxies. Sensitivity analysis with BERTScore-derived labels confirmed the overall trends, but a larger-scale, human-verified dataset is necessary to fully characterize clinical utility across diverse patient populations.

Second, the expert subtype analysis and explainability review are explicitly initial and involve one clinician which does not permit inter-rater reliability estimation. While this provides valuable qualitative evidence, the findings cannot be generalized without multi-specialty panels and pre-registered study protocols.

Third, the SE component relies on biomedical NLI for equivalence clustering. Systematic NLI errors such as domain shift and negation sensitivity, propagate directly into SE scores and could bias detection toward or against certain error types. Quantifying NLI reliability on the target domain and testing robustness to alternative entailment models is planned future work.

Fourth, all four evaluated models fall within the $\sim 1\text{B}$ -parameter class. This scope is deliberate since the proposed framework targets the hardware constraints of on-premise clinical workstations, but it leaves open whether the hybrid’s advantages persist at larger scales (e.g., 3B–7B parameters), where individual detectors may already be stronger and the complementarity between SE and IS may diminish. A related open question is whether larger but quantized models, which fit similar hardware, shift this trade-off. Evaluating the hybrid on quantized models is a natural next step.

Fifth, the evaluation is limited to English-language, short-form medical QA. Other impactful clinical tasks, including summarization, discharge instructions, and guideline drafting, involve longer contexts and different hallucination modalities. The framework has not been evaluated in conjunction with extrinsic approaches such as RAG (Retrieval Augmented Generation) which may complement intrinsic detection when local retrieval infrastructure is available. Extending the evaluation to established benchmarks for clinical summarization and medical text simplification is a concrete next step. Establishing generalizability across languages, tasks, and model scales is necessary before any clinical integration.

6 Conclusion

This research proposed a hybrid framework for hallucination detection specifically designed for privacy-constrained, on-premise medical LLMs. By combining semantic entropy with internal signal probing, we obtained modest but consistent improvements in detection performance and, critically, in calibration on 1B-parameter models across two medical QA benchmarks and four architectures. The framework is most effective at detecting high-risk factual errors such as fabricated contraindications and provides explainable signals that support calibrated human trust. The subtype-level analysis revealed a clinically important asymmetry: indication and contraindication errors are identified with relatively high reliability, while more subtle conceptual hallucinations remain harder to flag. These results demonstrate that post-hoc reliability monitoring of smaller, local models can substantially improve their safety profile, supporting the broader goal of equitable and responsible medical AI.

Future work follows directly from the limitations outlined in Section 5.3: scaling expert annotation to multi-specialty panels with NLI-based correctness assessment, and extending evaluation to larger models, additional languages, and longer-form clinical tasks such as summarization and discharge-instruction generation. Beyond these, three complementary directions warrant dedicated investigation: (i) multi-layer probing strategies and alternative semantic clustering approaches to capture hallucination patterns our current signals miss; (ii) human-centered co-design studies with practitioners to empirically assess how alerting strategies, explanation formats, and feedback mechanisms affect clinician trust, decision quality, and workflow integration; (iii) fairness monitoring across patient demographic groups to ensure the framework performs equitably across populations.

Ethics Statement

This research aims to enhance patient safety by improving hallucination detection in medical AI systems. To this end, several ethical considerations are further discussed.

First, the supervised components of the framework are trained on automated correctness labels, which may systematically misclassify certain answer types. If these biases correlate with patient demographics or clinical subdomains, the detector could be less effective for underrepresented populations, potentially exacerbating health disparities. Continuous monitoring, regular retraining with diverse data, and mechanisms for clinicians to report and override detector flags are therefore core components of a responsible deployment strategy.

Second, the hallucination taxonomy and subtype annotations were developed and verified by a single medical expert. While this provided clinically grounded labels, it does not capture the full diversity of clinical perspectives. To address this, multi-specialty validation is needed before deployment.

Third, a well-calibrated detection system could paradoxically increase clinician trust to a degree that reduces vigilance for error types the system handles poorly. Hence, deployment needs to include clear communication of the system's known blind spots.

And fourth, while designed for safety purposes, the hallucination taxonomy and detection signals could theoretically be repurposed in order to identify which errors evade detection, enabling adversarial manipulation of medical LLM outputs. We believe the safety benefits substantially outweigh this risk, particularly given the focus on open evaluation methods.

This research did not involve human subjects beyond the expert annotator, who participated voluntarily. No patient data was used. All datasets are established, de-identified benchmarks.

Acknowledgments

We thank the AIES reviewers for their constructive and thoughtful feedback, and the consulting medical expert for the annotation and evaluation sessions.

References

- Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, 99: 101805.
- Aljohani, M.; Hou, J.; Kommu, S.; and Wang, X. 2025. A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 6720–6748. Association for Computational Linguistics.
- Alu, F. F.; and Oluwadare, S. 2026. An auditable and source-verified framework for clinical AI decision support: integrating retrieval-augmented generation with data provenance. *Frontiers in Artificial Intelligence*, 9: 1737532.
- Amadi, C.; and Ojo, A. 2025. Building Trustworthy AI in Healthcare. *IEEE Access*, 14: 1182–1212.
- Artzi, Y.; Sorin, V.; Glicksberg, B. S.; Korfiatis, P.; Freeman, R.; Nadkarni, G. N.; and Klang, E. 2025. Challenges of Implementing LLMs in Clinical Practice: Perspectives. *Journal of Clinical Medicine*, 14(17): 6169.
- Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 967–976.
- Bagave, P.; Westberg, M.; Janssen, M.; and Ding, A. Y. 2025. Accountability Framework for Healthcare AI Systems: Towards Joint Accountability in Decision Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 279–291.
- Bhatnagar, R.; Sun, Y.; Zhang, C. A.; Wen, Y.; and Yang, H. 2026. HALT: Hallucination Assessment via Latent Testing. *arXiv preprint arXiv:2601.14210*.
- Das, A.; and Rad, P. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv preprint arXiv:2006.11371*.
- Dhaini, M.; Müller, T.; Rabets, R.; and Kasneci, G. 2025. Adoption of Explainable Natural Language Processing: Perspectives from Industry and Academia on Practices and Challenges. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 719–730.
- European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council on harmonised rules on Artificial Intelligence. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>. Official Journal of the European Union L 202/1–60.
- European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
- Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.
- Garg, M.; Raza, S.; Rayana, S.; Liu, X.; and Sohn, S. 2025. The Rise of Small Language Models in Healthcare: A Comprehensive Survey. *arXiv preprint arXiv:2504.17119*.
- Gou, Z.; Shao, Z.; Gong, Y.; Shen, Y.; Yang, Y.; Duan, N.; and Chen, W. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.
- Granstedt, J.; Kc, P.; Deshpande, R.; Garcia, V.; and Badano, A. 2025. Hallucinations in medical devices. *Artificial Intelligence in the Life Sciences*, 8: 100145.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.
- Jung, J.; Lee, H.; Jung, H.; and Kim, H. 2023. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9(5).
- Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.
- Lauscher, A.; Glavaš, G.; et al. 2025. How Much Do LLMs Hallucinate across Languages? On Multilingual Estimation of LLM Hallucination in the Wild. *arXiv preprint arXiv:2502.12769*.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- Lu, Z.; Peng, Y.; Cohen, T.; Ghassemi, M.; Weng, C.; and Tian, S. 2024. Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9): 1801–1811.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lv, B.; Tang, C.; Liu, N.; Liu, X.; Zhang, R.; and Yu, Y. 2025. MIRE: A Medical Information Enhanced Framework for Long-Tail Medical Dialogue Synthesis. *Expert Systems with Applications*, 261: 130905.
- Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegrefe, S.; Alon, U.; Dziri, N.; Prabhunoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.
- Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheck-GPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Meske, C.; Bunde, E.; Schneider, J.; and Gersch, M. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information systems management*, 39(1): 53–63.
- Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *arXiv preprint arXiv:2410.05229*.
- Moor, M.; Banerjee, O.; Abad, Z. S. H.; Krumholz, H. M.; Leskovec, J.; Topol, E. J.; and Rajpurkar, P. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956): 259–265.
- Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Koteck, H.; and Belinkov, Y. 2025. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *International Conference on Learning Representations (ICLR)*. Poster Presentation.
- Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2023. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 314–334.
- Pal, A.; Wangmo, T.; Bharadia, T.; Ahmed-Richards, M.; Bhanderi, M. B.; Kachhadiya, R.; et al. 2025. Generative AI/LLMs for plain language medical information for patients, caregivers and general public: Opportunities, risks and ethics. *Patient Preference and Adherence*, 19: 2227–2249.
- Pandit, S.; Xu, J.; Hong, J.; Wang, Z.; Chen, T.; Xu, K.; and Ding, Y. 2025. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2858–2873.
- Panigutti, C.; Beretta, A.; Fadda, D.; Giannotti, F.; Pedreschi, D.; Perotti, A.; and Rinzivillo, S. 2023. Co-design of human-centered, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems*, 13(4): 1–35.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.
- Rodríguez, A. 2025. Privacy-Focused LLM for Local Data Processing: Implementing OLLAMA and RAG to Securely Query Personal Files in Closed Environments. *arXiv preprint*.
- Shneiderman, B. 2022. *Human-centered AI*. Oxford University Press.
- Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 1–8.
- Song, Y.; Qiu, L.; Zhang, X.; and Tang, Z. 2025. Hallucination Detection via Internal States and Structured Reasoning Consistency in Large Language Models. *arXiv preprint arXiv:2510.11529*.
- Srey, P.; Wu, X.; and Tuan, L. A. 2025. Unsupervised Hallucination Detection by Inspecting Reasoning Processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 22128–22140.

- Su, W.; Wang, C.; Ai, Q.; Hu, Y.; Wu, Z.; Zhou, Y.; and Liu, Y. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 14379–14391.
- Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, 359–380. PMLR.
- U.S. Department of Health and Human Services. 1996. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Pub. L. 104-191, 110 Stat. 1936.
- Vazhentsev, A.; Rvanova, L.; Kuzmin, G.; Fadeeva, E.; Lazichny, I.; Panchenko, A.; Panov, M.; Baldwin, T.; Sachan, M.; Nakov, P.; et al. 2026. Efficient Hallucination Detection for LLMs Using Uncertainty-Aware Attention Heads.
- Vipulanandan, P.; Premaratne, K.; and Sarkar, D. 2026. Semantic Uncertainty Quantification of Hallucinations in LLMs: A Quantum Tensor Network Based Method. *arXiv preprint arXiv:2601.20026*.
- Wada, A.; Tanaka, Y.; Nishizawa, M.; Yamamoto, A.; Akashi, T.; Hagiwara, A.; Hayakawa, Y.; Kikuta, J.; Shimoji, K.; Sano, K.; et al. 2025. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. *npj Digital Medicine*, 8(1): 395.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wu, X.; Zhao, H.; Zhu, Y.; Shi, Y.; Yang, F.; Liu, T.; Zhai, X.; Yao, W.; Li, J.; Du, M.; et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint arXiv:2403.08946*.
- Zeng, Z.; Cheng, Q.; Hu, X.; Zhuang, Y.; Liu, X.; He, K.; and Liu, Z. 2025. KoSEL: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems*, 309: 112837.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.