Research Square

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

ConfidentNeuro: Conformal Prediction-Based Uncertainty Quantification for Hallucination Risk Scoring in LLM-Assisted Alzheimer's Clinical Decision Support

Afreen Khan
Integral University
Samreen Khan
Integral University
Mudassir Khan
mudassirkhan@mmu.edu.my

King Khalid University
Mazliham Mohd Su'ud
Multimedia University
Muhammad Mansoor Alam
Riphah International University
Saurav Mallik
University of Arizona

Research Article

Keywords: Alzheimer's disease, BERTScore, calibration, clinical decision support, conformal prediction, hallucination risk scoring, large language models, perturbation audit, trustworthy clinical AI, uncertainty quantification

Posted Date: May 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9707462/v1

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

Additional Declarations: No competing interests reported.

ConfidentNeuro: Conformal Prediction-Based Uncertainty Quantification for Hallucination Risk Scoring in LLM-Assisted Alzheimer's Clinical Decision Support

Afreen Khan[1], Samreen Khan[2], Mudassir Khan[3,*], Mazliham Mohd Su'ud[4,*], Muhammad Mansoor Alam[5], Saurav Mallik[6,7,*]

[1]Faculty of Engineering & IT, Integral University, Dashauli, Uttar Pradesh 226026, India; afreen.khan2k13@gmail.com

[2]Integral Institute of Medical Sciences and Research, Integral University, Dashauli, Uttar Pradesh 226026, India; drsamreen2k4@gmail.com

[3]Department of Computer Science, College of Computer Science, Applied College Tanumah, King Khalid University, P.O. Box: 960 - Postal Code: 61421, Abha, Saudi Arabia; mudassirkhan@mmu.edu.my

[4]Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Malaysia; mazliham@mmu.edu.my

[5]Faculty of Computing, Riphah International University, Islamabad, 44000, Pakistan; m.mansoor@riphah.edu.pk

[6]Department of Environmental Health, Harvard T H Chan School of Public Health, Boston, MA, USA; sauravmtech2@gmail.com, smallik@hsph.harvard.edu

[7]Department of Pharmacology & Toxicology, The University of Arizona, Tucson, AZ 85721, USA; smallik@arizona.edu

***Correspondence: Mudassir Khan (**mudassirkhan@mmu.edu.my**); Mazliham Mohd Su'ud (**mazliham@mmu.edu.my**); Saurav Mallik (**sauravmtech2@gmail.com, smallik@arizona.edu**)**

Abstract

Alzheimer's disease (AD) affects more than 55 million people worldwide. In recent years, clinical decision support (CDS) systems built on large language models (LLMs) have become increasingly common in neurological care. These models hallucinate, generating clinically plausible but factually incorrect responses. The existing evaluation methods measure overall accuracy, leaving clinicians without per-response uncertainty or clear distrust signals. We present ConfidentNeuro, a conformal prediction framework for assessing the risk of hallucination in LLM-assisted AD decision support. We also introduce ConfidentNeuro-Bench, which includes 150 clinical question-answer pairs across five AD categories. These were gathered using PubMed MeSH queries, and each was linked to its source PMID. GPT-4o, Claude Haiku, and Mixtral-8x7B were each tested zero-shot, using conformal prediction with semantic similarity as the nonconformity score. GPT-4o and Claude Haiku showed empirical coverage of 97.1%. Whereas Mixtral-8x7B reached 95.7%, both above the 90% target. The calibration audit indicated that GPT-4o is miscalibrated (ECE = 0.213). As a result, token-level log probabilities are an unreliable confidence signal in clinical use. Under perturbation audit across eight input modifications, confidence barely changed, staying around 0.56-0.62, despite a twofold variation in semantic disruption. This raises a key

issue of a systemic failure in self-assessment. We ran ablation experiments to examine each component's contribution. ConfidentNeuro assigns a risk label (LOW/MEDIUM/HIGH) to each response and makes it visible to the clinician without requiring access to the model internals or retraining. This results in uncertainty estimates that are consistent across different models, making them easier to use in neurological care.

Keywords: *Alzheimer's disease; BERTScore; calibration; clinical decision support; conformal prediction; hallucination risk scoring; large language models; perturbation audit; trustworthy clinical AI; uncertainty quantification.*

I. Introduction

Alzheimer's disease (AD) is the most common form of dementia [1]. It is responsible for about 60-70% of all dementia diagnoses globally, and currently is the seventh highest cause of death in the world [2]. According to the 2021 WHO report, 55 million individuals were estimated to have dementia, and figures indicate that this will likely increase to over 139 million by 2050 [2]. Within the United States, about 7.4 million adults 65 and above are living with AD in 2026 [3]. Medical expenses for this disease are expected to amount to $1 trillion by mid-century [3]. The clinical management of AD is complex. It encompasses diagnosis, disease staging, biomarker interpretation, treatment planning, and long-term prognosis [4]. All of these require continuous access to current evidence-based clinical knowledge.

Large language models (LLMs) have emerged as promising tools for clinical decision support in various medical specialties [5]. Singhal et al. demonstrated that LLMs encode significant clinical knowledge and can achieve expert-level performance on standardized medical examinations [6]. Recent studies have also been conducted in the context of AD. LLMs have been used for longitudinal prediction of electronic health record (HER) data [7], multimodal diagnostic interpretation [8], and have also been explored in broader dementia care use cases [9]. Moreover, traditional ML approaches have also shown strong results, such as neuroevolution-based neural models [10] and multi-composite ensemble systems that integrate AT(N) biomarkers [11]. Consequently, interest in using LLMs in clinical decision support (CDS) systems has grown considerably [12], [13], [14]. These systems can provide clinicians with rapid and evidence-based responses at the point of care.

However, deploying LLMs in clinical settings introduces a serious and still under-addressed safety issue known as hallucination. LLMs can generate responses that are fluent, appear confident, and are clinically plausible; however, they are factually incorrect [15]. In high-stakes clinical scenarios, such as AD management, where diagnostic and treatment decisions directly impact patients, this risk becomes even more critical. When an LLM generates an incorrect answer that sounds plausible (a "hallucination"), and this goes unnoticed, a clinician's reasoning or treatment plans might be incorrectly directed [16]. Research indicates that with adversarial inputs, these hallucination rates can range from 50% to 83%, and that while prompt engineering reduces errors, it does not eliminate the problem [17].

Another significant challenge is that clinical LLMs are often overly confident in their responses, and their reported confidence scores do not reliably reflect their accuracy [18]. When the model does express a level of confidence, users often over-rely on the information it provides [19]. Regulatory structures are beginning to address these issues; for example, the European Union Artificial Intelligence Act (EU AI) imposes mandatory requirements for high-risk clinical AI applications to be safe and transparent about how they work [20]. However, a significant gap persists. Currently, no standardized mechanism exists to provide a statistically assured uncertainty estimate for each response at inference time for clinical text generated by LLMs.

Conformal prediction offers a systematic approach to this issue; it is a statistical framework that is both distribution-free and model-agnostic [21]. That is, it does not rely on any particular model or the way the data is distributed. A key benefit is that it provides a finite-sample coverage guarantee. In other words, there is mathematical assurance that the true answer will fall within the prediction set with a probability of at least $1 - \alpha$ [22]. This guarantee is unaffected by the specifics of the model being used or the underlying data distribution. Conformal methods have seen application in healthcare, including medical image classification [23], clinical AI triage systems [24], and quantifying uncertainty in natural language processing (NLP) [25]. Yet there have been no reports of its use to score the risk of hallucinations in clinical text generated by LLMs.

We developed ConfidentNeuro, a conformal-prediction-based framework for evaluating how likely an LLM used in Alzheimer's care can hallucinate, i.e., can go wrong without warning. The system treats any LLM as a black box model and assigns each response a guaranteed risk label of LOW/MEDIUM/HIGH. For a clinician, this translates into a quick understanding of whether to act on a response or seek a second opinion. The three-tier risk categorization is based on a nonconformity score derived from semantic similarity, as measured by BERTScore F1. We chose this over log probabilities after finding they were poorly calibrated in our preliminary tests. We assessed ConfidentNeuro with ConfidentNeuro-Bench, a collection of 150 clinical question-answer (QA) pairs related to five types of AD, developed through a methodical search of PubMed MeSH retrieval. Every answer in the benchmark can be traced back to a specific PubMed abstract, with no crowdsourced labels or model-generated ground truth. We evaluated three LLMs: GPT-4o, Claude Haiku, and Mixtral-8×7B. These were chosen cautiously to reflect the range of real deployment scenarios. GPT-4o represents closed-source APIs, Claude Haiku covers low-latency use cases, and Mixtral-8×7B addresses hospitals that cannot send patient data to external servers and need an on-premise open-source option. The performance and design of ConfidentNeuro were validated through numerous evaluations: a systematic calibration audit, an eight-type clinical note perturbation audit, and a comprehensive ablation study. The results showed that GPT-4o (despite being the outperformer amongst the three LLMs) was substantially miscalibrated, with an expected calibration error (ECE) of 0.213, meaning its expressed confidence did

not reflect its actual correctness. Furthermore, the perturbation audit showed that model confidence barely moved even when clinical notes were deliberately corrupted, which is a safety concern.

The main contributions of this study are as follows:

- We introduce ConfidentNeuro-Bench, a benchmark of 150 Alzheimer's clinical QA pairs traceable to source PMIDs, with easy/medium/hard difficulty labels.
- We demonstrate that conformal prediction at $\alpha = 0.10$ achieves empirical coverage of ≥95.7% across all three models - to our knowledge, the first conformal hallucination risk scoring framework for clinical QA.
- We conducted a calibration audit, and it resulted in an ECE = 0.213 for GPT-4o. This result shows that logprob confidence is unreliable. As a result, we developed a BERTScore-based nonconformity measure.
- We performed a perturbation audit across eight perturbation types. We identified a systemic failure in the confidence invariance of GPT-4o's self-assessment mechanism.
- We open-source the benchmark data, evaluation scripts, and conformal calibration wrapper to enable reproducibility.

The remainder of this paper is organized as follows. **Section II** discusses the related work on clinical LLMs, hallucination mitigation, uncertainty quantification, and conformal prediction. **Section III** explains the ConfidentNeuro framework and the experimental methodology in detail. **Section IV** presents the experimental results. **Section V** discusses the findings, limitations, and future work, followed by Section VI, which presents the conclusion.

II. Related Work

A. LLMs in Clinical Decision Support

Recently, the use of LLMs in CDS has grown tremendously across medical domains. Singhal et al. showed that these LLMs contain a considerable amount of clinical information and can achieve expert-level performance on medical licensing examinations [6]. Several studies have evaluated LLMs in areas such as gastroenterology, predicting AD from longitudinal EHR notes, and multimodal frameworks using Mini-Mental State Examination data [7], [8], [26]. Their application has also been explored in dementia care, specifically in areas such as predicting cognitive decline, assisting caregivers, and summarizing clinical notes [14]. While these developments provide a promising future, accuracy remains the major metric for performance evaluation [27]. The models are primarily assessed based on whether their outputs are correct on an average. However, little attention has been paid to response-level uncertainty and the reliability of individual predictions. Thus, ConfidentNeuro fills this gap by adding an uncertainty layer on top of the standard evaluation.

B. Hallucination as a Clinical Safety Risk

Hallucination refers to the generation of content that sounds correct but is factually incorrect [28]. This is a major safety risk in clinical LLM deployment. And unlike traditional software bugs, it is invisible. Huang et al. conducted a comprehensive survey and highlighted that a lack of awareness of uncertainty is the chief contributing factor, noting that LLMs generate answers in a confident tone but lack internal signals of their own unreliability [15]. In clinical settings, this risk is even greater. Omar et al. showed that the hallucination rate of LLMs on adversarial clinical inputs can reach 50-83% [17]. Prompt engineering reduces errors but cannot eliminate them completely [29]. Roustan and Bastardot further demonstrated that overconfident outputs can mislead clinicians' diagnostic reasoning, thereby directly impacting patient safety [16]. The mitigation strategies currently in use, such as prompt engineering, chain-of-thought scaffolding, and RAG, all operate at the input level [30]. However, these do not guarantee uncertainty in every individual response. ConfidentNeuro takes a different route – it works at the output level. Each response is assigned a statistically grounded risk score that is independent of the input strategy.

C. Model Calibration and Uncertainty Estimation

Model calibration (how well a model confidence matches its actual correctness) has been extensively studied in neural networks [31]. This refers to the alignment between expressed confidence and actual correctness. Guo et al. showed that modern deep learning models are generally overconfident, thereby motivating the use of post-hoc calibration methods [18]. This problem does not disappear with scale. In another study by Kadavath et al., it was observed that LLMs can perform self-evaluation to some extent. However, this ability is inconsistent and does not reliably transfer to open-ended generation tasks [32]. In clinical settings specifically, miscalibration becomes dangerous rather than just inconvenient. Steyvers et al. demonstrated a substantial gap between what LLMs actually know and what users perceive them to know [19]. When confidence signals are present, users often overestimate LLM accuracy [19]. Lambert et al. identified that overconfidence and a lack of transparent uncertainty limit the adoption of clinical AI [33]. According to them, uncertainty quantification is necessary to identify the cases that require human review. The ECE of 0.213 for GPT-4o in this study aligns with the prior findings. Moreover, this result directly supports the design decision to use BERTScore F1 as a nonconformity score, replacing the logprob-based confidence.

D. Conformal Prediction in NLP and Healthcare

Conformal prediction is a framework that provides distribution-free and finite-sample coverage guarantees for any pre-trained model [22]. Therefore, it is considered relevant for high-stakes applications. Campos et al. did a comprehensive evaluation of its application to NLP. The authors highlighted that controlling the risk of hallucination is a strong direction for conformal methods, especially in language generation [34]. Its use has already been demonstrated in several areas of healthcare. such as severity rating in spinal stenosis MRI [23], skin lesion classification [35], and

clinical triage systems, where patients are divided into low-risk, high-risk, or uncertain categories with statistical guarantees [24]. A recent study proposed conformal prediction as a structured method to communicate uncertainty in clinical AI [33]. This helps clinicians understand when the model is confident and when it is uncertain. However, its use in the context of LLM-generated clinical text has not yet been reported, specifically for hallucination risk scoring in neurological decision-making support. ConfidentNeuro addresses this gap. To the best of our knowledge, this work is the first instantiation of conformal prediction applied to Alzheimer's clinical QA. Furthermore, it extends this framework from structured prediction to open-ended clinical language generation.

E. Semantic Similarity Metrics for Clinical NLP

Traditional text evaluation metrics such as BLEU and ROUGE mainly depend on n-gram overlap [36]. This approach is not reliable in a clinical context, because a single correct answer can have multiple valid phrasings. To address this limitation, Zhang et al. introduced BERTScore. It computes token-level semantic similarity using contextual embeddings and aligns better with human judgment than surface-level metrics [9]. In clinical NLP, the semantic equivalence of a generated response to a reference answer is more meaningful than lexical overlap. Therefore, BERTScore provides a more suitable basis for evaluation. We adopted BERTScore as the nonconformity score in ConfidentNeuro. This is motivated by both this theoretical advantage and by the empirical finding that, in the clinical calibration audit, logprob-based alternatives proved unreliable (ECE = 0.213), as shown in Section III-E.

III. Methodology

The ConfidentNeuro framework comprises four layers: benchmark construction, multi-model evaluation, conformal prediction-based uncertainty quantification, and hallucination risk scoring. The methodology is organized into seven subsections covering framework design, benchmark construction, model evaluation, conformal prediction, calibration quality assessment, perturbation audit, and ablation study. **Fig. 1** illustrates the end-to-end pipeline architecture. **Fig. 2** provides an overview of the experimental design.

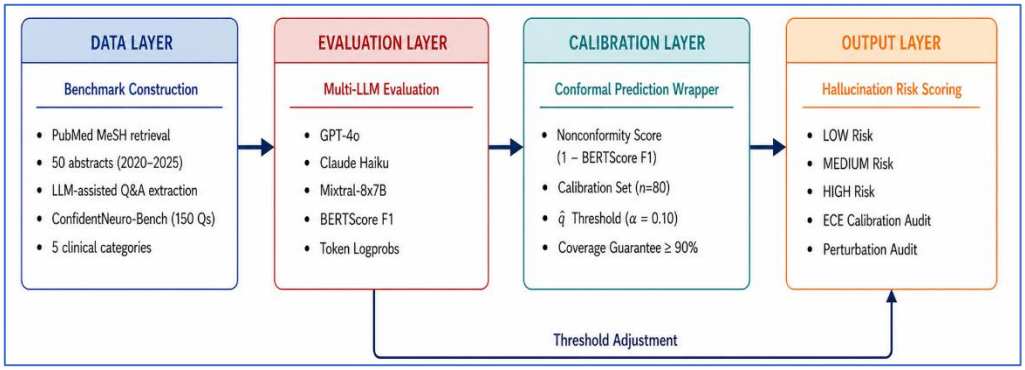


Fig. 1: ConfidentNeuro framework for hallucination risk scoring in LLM-assisted Alzheimer's clinical decision support.

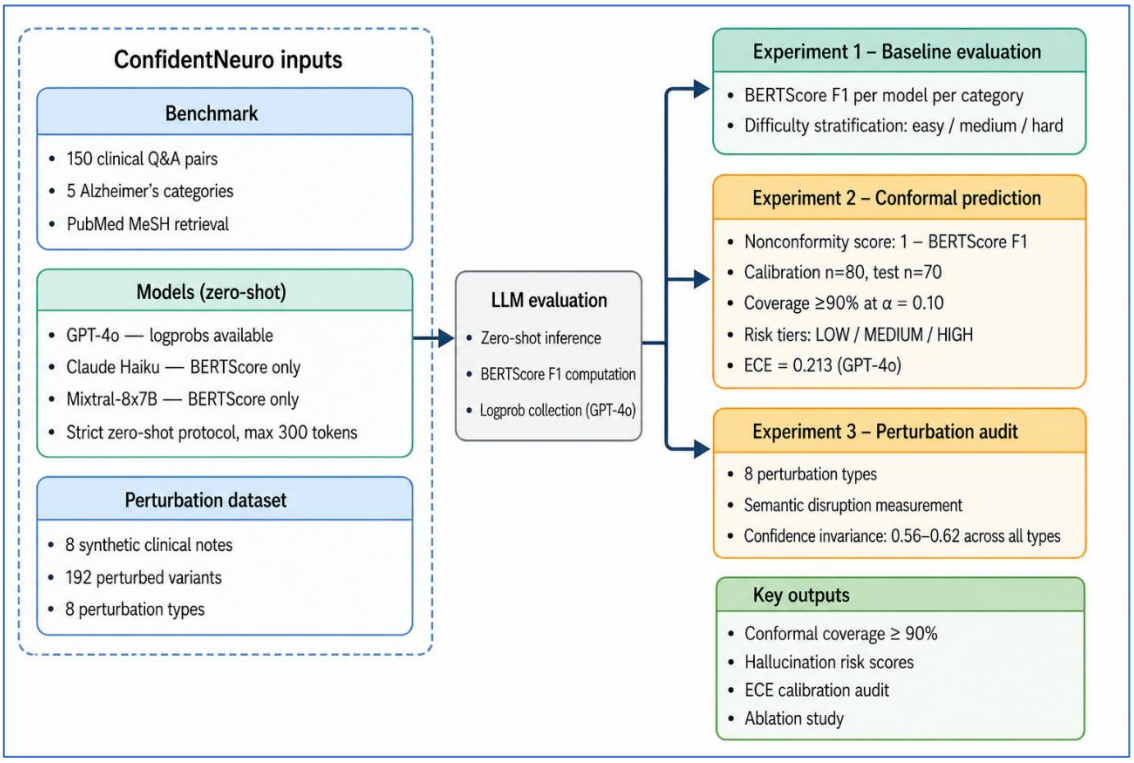


Fig. 2: Study design overview of the ConfidentNeuro evaluation framework.

A. Framework Design and Architecture

This work presents a conformal prediction-based framework for scoring hallucination risk in LLM-generated Alzheimer's clinical responses. The framework quantifies uncertainty and assigns each response to statistically guaranteed risk tiers: LOW, MEDIUM, or HIGH. This quantification is done so that unreliable outputs get easily flagged at inference time. A clinical insight often overlooked is that accuracy alone is not enough, particularly when the stakes involve patient care. A model can be overall accurate but miscalibrated at the individual-response level. Practically speaking, not every answer is equally reliable. This is a major challenge in a clinical setup. In reality, clinicians need clarity before they can trust any prediction. Therefore, it is essential that uncertainty is clearly indicated alongside every prediction. A CDSS that lacks an indication of uncertainty cannot be deemed reliable. ConfidentNeuro addresses this gap. To do so, it wraps standard LLM evaluation with a conformal prediction layer, a step that most existing frameworks skip entirely. Through this, raw model outputs become statistically grounded risk scores, needing no retraining and no access to model internals.

The architecture is guided by four principles: statistical rigor, model agnosticism, clinical interpretability, and transparency. These principles guide how uncertainty is estimated and presented. Under statistical rigor, uncertainty estimates are not based on heuristic thresholds. Instead, they follow finite-sample coverage guarantees derived from the conformal prediction theory [1]. In model agnosticism, each LLM is treated as a black box. The framework relies solely on text outputs, and token-

level log probabilities should be included only when required. This makes it usable with both closed-source APIs and open-source models. In addition, the framework emphasizes clinical interpretability. In this, the nonconformity scores are mapped to a three-tier risk system, i.e., LOW, MEDIUM, or HIGH. This process is done to align with clinical triage practice. Transparency is ensured by documenting all design choices, including the nonconformity score, calibration set construction, and risk thresholds. These are empirically validated through ablation studies, and the documentation is made publicly available for independent verification.

The proposed framework follows a modular design. It consists of four layers: benchmark construction, multi-model evaluation, conformal calibration, and risk scoring. Each of these layers operates independently; therefore, if anyone needs to be expanded or replaced, the others remain unaffected. This modularity is a particularly critical component in the clinical deployment of models because changes to a model, regulatory changes, or institution-specific data governance may require that one or more individual models be swapped with others; however, the pipeline will function as originally designed.

This study primarily focuses on the technical validation of the ConfidentNeuro framework (i.e., confirming that conformal predictions can yield statistically valid hallucination risk scores for LLM-generated clinical responses). Clinical validation (i.e., use in practice by neurologists with evaluations compared to actual patient outcomes) will be necessary for future research; however, it is not within the scope of the present study.

B. ConfidentNeuro-Bench Construction

For a reliable evaluation of hallucinations, a benchmark grounded in clinical evidence that can be verified using existing sources is required [37]. Existing medical QA datasets are applied either to crowdsourced annotations (a method for generating labels for individual systems) or to questions derived from model-generated content. Both methods add uncontrolled noise to ground-truth labels. In this study, we employed a different approach: we extracted questions and corresponding reference answers from abstracts published in peer-reviewed journals, ensuring that every ground-truth answer could be traced to a publication cited in the existing literature.

Five established MeSH-controlled search terms that identify various aspects of AD were used to query PubMed via the NCBI Entrez API (https://www.ncbi.nlm.nih.gov/home/develop/api/) to retrieve answer references from the PubMed database, using an authorized API key at a maximum rate of 10 requests per second. All programmatic queries to NCBI Entrez were conducted using the Biopython Entrez module. The five search terms used were: Alzheimer's Disease/diagnosis, Alzheimer's Disease/pathology, Alzheimer's Disease/therapy, Alzheimer's Disease/cerebrospinal fluid OR Alzheimer’s Disease/blood, and Alzheimer's Disease/epidemiology. These map to five evaluation categories: Diagnosis, Disease Staging, Treatment and Management, Biomarkers, and Prognosis and

Risk Factors. From each category, ten abstracts were selected within the 2020–2025 indexing window. This resulted in 50 source documents overall. The NCBI Entrez API was used separately to construct the benchmark.

An extraction pipeline was developed to generate three clinical questions for each abstract, along with reference answers derived from the content of these abstracts. Each answer also has a provenance tag (source PMID) to enable independent verification of the ground truth, thereby reducing model-knowledge contamination. The question set was further annotated with difficulty levels (easy, medium, and hard) based on how deeply the clinical reasoning would have had to go to generate the answers. The final benchmark has 150 question/answer pairs (60 easy, 60 medium, 30 hard) and was distributed evenly across five Alzheimer's clinical categories (30 questions per category). The statistics for the benchmark are presented in **Table I**. The entire benchmark, including all questions/answers and source PMIDs, can be accessed at https://github.com/afreen-kh/ConfidentNeuro.

TABLE I: ConfidentNeuro-Bench statistics

Category	MeSH Term	Abstracts	Questions	Easy	Medium	Hard
Diagnosis	Alzheimer Disease/diagnosis	10	30	12	12	6
Disease Staging	Alzheimer Disease/pathology	10	30	12	12	6
Treatment & Management	Alzheimer Disease/therapy	10	30	12	12	6
Biomarkers	AD/CSF OR AD/blood	10	30	12	12	6
Prognosis & Risk Factors	Alzheimer Disease/epidemiology	10	30	12	12	6
Total	—	**50**	**150**	**60**	**60**	**30**

C. Multi-LLM Evaluation Setup

We selected three models to span the deployment environment applicable to clinical AI: GPT-4o (OpenAI) was included as a state-of-the-art closed-source system, Claude Haiku (Anthropic) was selected because it is optimized for low-latency inference, and Mixtral-8x7B (Mistral AI) was chosen as an open-source alternative that can be deployed in on-premise hospital settings. Together, these three models facilitate direct comparison between proprietary and open-weight models, which is important when considering data governance restrictions imposed by the healthcare environment.

Each of the 150 questions was tested under a strict zero-shot setup, with no system prompt, no chain-of-thought guidance, and no clinical context added. Every question went in as a plain user message, the way someone would actually type it. This reflects how these systems are typically used in practice. This makes the evaluation conservative by design. And this matters since the performance would likely improve with prompt engineering. At the same time, it highlights the need for uncertainty quantification during inference. The prompt was the same across all three models. Only the raw question was used, and responses were limited to 300 tokens.

For each response, certain key metrics were recorded. The generated answer text was stored. The token-level log probabilities were calculated for the GPT-4o model through the 'logprobs=True' and 'top_logprobs=1' parameters. Subsequently, the mean log probability was computed and used as a scalar confidence estimate. In practice, this gave us a rough proxy for how confident the model was. Model outputs were then scored using BERTScore (precision, recall, and F1). All scores were computed against reference answers using 'distilbert-base-uncased' as the backbone. The response time was also recorded for latency comparison. Notably, this helps in understanding the deployment performance. **Table II** summarizes the API configuration for each model.

TABLE II: API configuration for multi-LLM evaluation

Model	API	Max Tokens	Logprobs	Temperature
GPT-4o	OpenAI API	300	✓ (top_logprobs=1)	Default (1.0)
Claude Haiku	Anthropic API	300	✗	Default
Mixtral-8x7B	HuggingFace Inference API	300	✗	Default

Due to API constraints, log probability extraction was unavailable for Claude Haiku and Mixtral-8x7B. Therefore, BERTScore F1 was used as a proxy for uncertainty. In practical terms, the score provides a semantic-based confidence signal. To make the evaluation process robust, a checkpoint-based resume mechanism was implemented to track each completed (question, model) pair. In case of API interruptions, evaluation could be resumed from the same point without data loss.

D. Conformal Prediction Framework

Conformal prediction is a statistical framework that provides uncertainty estimates that are valid within a strict statistical context. Its coverage guarantee holds for any finite dataset, and the data need not follow any specific distribution. For this reason, it is well-suited to clinical AI applications, where data distributions are complex and rarely known [22]. Given a calibration set in which the samples are exchangeable, it ensures that the constructed prediction set $C(x)$ satisfies:

$$P\left(y_{test} \in C(x_{test})\right) \geq 1 - \alpha \qquad (1)$$

where $\alpha \in (0,1)$ is a user-defined error rate, y_{test} represents the actual ground truth response for a given test input, and $C(x_{test})$ denotes the set of possible outputs the framework can generate.

In simple terms, this means that the probability that the true clinical answer is captured in the prediction set is at least $1 - \alpha$. For example, if α is set to 0.10, the framework guarantees 90% coverage. And this assurance holds for finite samples, regardless of the distribution of the LLM outputs. Thus, we adapt this framework to score and stratify the risk of hallucination in LLM-generated clinical responses. This process is performed in three sequential steps.

1. Nonconformity Score. In conformal prediction, the nonconformity score determines how uncertain a given prediction is compared to the calibration data [38], [39]. The higher the score, the more the

prediction deviates from the model's expected behavior. The first step of this framework measures how semantically different the model's response is from the reference answer. Each response is assigned a scalar nonconformity score, given by the following formula:

$$s_i = 1 - BERTScore\ F1_i \qquad (2)$$

where s_i is the nonconformity score for the ith response, and $BERTScore\ F1_i$ is the semantic similarity between the generated answer and the reference answer, which is calculated using contextual embeddings. A score of 0 denotes perfect semantic alignment with the ground truth, whereas a score of 1 represents complete divergence from that alignment.

A higher value of s_i indicates that the response differs more from the ground truth and is associated with a higher risk of hallucination. Here, $BERTScore\ F1_i$ has been chosen as the nonconformity measure over raw log probability because it captures semantic similarity rather than surface-level overlap. In the preliminary analysis, log probabilities proved as unreliable as calibration signals (ECE = 0.213); as mentioned in Section III-E).

2. Threshold Computation. A coverage threshold is derived using the nonconformity scores from the calibration set [40], [41]. This threshold determines the boundary between the conforming and non-conforming responses. The responses for each model were split into two parts: 80 responses were taken for the calibration set, and 70 responses for the held-out test set. This split was performed using stratified random sampling (seed = 42). The conformal threshold was computed using the following formula:

$$\hat{q} = Quantile\left(\{s_i\}_{i=1}^{n_{cal}}, \frac{\lceil (n_{cal}+1)(1-\alpha) \rceil}{n_{cal}}\right) \qquad (3)$$

Here $\hat{q}$ is the empirical quantile of the nonconformity scores of the calibration set at the level $1 - \alpha$. The ceiling function is used to ensure finite-sample validity, which is a core part of conformal prediction theory [1]. If a test response $s_i \leq \hat{q}$, it is considered conforming. In other words, it is semantically consistent with the calibration distribution. The coverage was evaluated at different α values, i.e., 0.05, 0.10, and 0.20. These represent nominal coverage guarantees of 95%, 90%, and 80%, respectively.

3. Risk Tier Assignment. Finally, using the threshold $\hat{q}$, each test response is assigned a clinically interpretable risk label. Instead of using only a binary conforming/non-conforming split, a three-tier scheme is introduced to provide better granularity at the clinical level. At the clinical operating point $\alpha = 0.10$, each test response is classified as follows:

a. LOW risk: When $s_i \leq 0.5 \times \hat{q}$, the response falls within the conforming region, i.e., the semantic alignment is strong, and confidence is considered high.
b. MEDIUM risk: When $0.5 \times \hat{q} < s_i \leq \hat{q}$, the response is near the conforming boundary, i.e., the uncertainty is moderate.

c. HIGH risk: When $s_i > \hat{q}$, the response exceeds the conformal threshold, and a human expert review is recommended.

The main objective of this three-tiered scheme is to convert a continuous uncertainty estimate into an actionable clinical signal. This approach helps identify responses in which expert intervention may be required. In a clinical context, this enables timely review and safer decision-making.

E. Calibration Quality Assessment

In clinical deployment, a miscalibrated model can at times be riskier than an inaccurate one. If a model is wrong but uncertain, it can be flagged, and the decision deferred [42], [43]. When it is confidently wrong, though, it is more likely to be trusted. To capture this risk, we used ECE, which quantifies the gap between expressed confidence and actual correctness [18].

We ran a calibration audit on GPT-4o, the only model that exposes token-level log probabilities. The mean log probabilities were scaled to [0, 1] across all 150 responses using min–max scaling. These scores were then grouped into $M = 10$ equal-width bins. For each bin B_m, we computed the average confidence conf(B_m)and the BERTScore F1 accuracy acc(B_m). The ECE was then calculated as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |acc(B_m) - conf(B_m)| \qquad (4)$$

Each bin ($|B_m|$) is weighted by the total number of responses n. A reliability diagram was then plotted between $conf(B_m)$ and $acc(B_m)$. The diagonal line $y = x$ corresponds to perfect calibration. When points deviate from this line, the model's confidence no longer matches its actual performance. Larger gaps tend to be more misleading in clinical settings.

F. Perturbation Audit Protocol

Benchmark accuracy alone does not show how model confidence reacts to clinically risky input changes [44]. If the system keeps assigning high confidence even after inputs are altered or corrupted, this can become a safety concern [45]. The issue is not just that the model is wrong, but that it gives no indication that the input has changed. And this is not a minor issue. So, to examine this, we conducted a structured perturbation audit using eight synthetic clinical notes. These were used to generate 192 perturbed variants across eight types of modifications.

We constructed eight synthetic Alzheimer's clinical notes, each following a fixed structure: patient demographics, presenting complaint, clinical history with dated events, cognitive assessment scores (MMSE, CDR), medications with dosages, imaging and laboratory findings, clinical impression, and management plan. For each note, 24 perturbations were generated across eight distinct types, resulting in 192 perturbed variants along with the original baseline notes.

Eight types of perturbations were used: (1) Temporal Reorder, where the clinical history sentences were shuffled; (2) Negation Insert, adding negation before key findings; (3) Value Swap, replacing

numerical values with plausible but incorrect ones; (4) Synonym Replace, substituting clinical terms with equivalent expressions; (5) Sentence Drop, removing 1–2 informative sentences; (6) Contradiction Insert, adding statements that conflict with existing findings; (7) Noise Inject, introducing medically irrelevant information; and (8) Entity Swap, replacing patient demographic identifiers.

Each perturbed note was sent to GPT-4o using a fixed summarization prompt: *"You are a clinical assistant. Summarize the following patient note in 3–4 sentences, highlighting key findings relevant to Alzheimer's disease management."* A temperature of 0.0 was set to ensure the outputs are fully deterministic. This guarantees that any variation in the summaries is solely due to the input perturbation, and not sampling randomness. Initially, a baseline summary was generated for each original note. Subsequently, semantic disruption was quantified as follows:

$$d_i = 1 - BERTScore\ F1\left(summary_{perturbed}, summary_{baseline}\right) \quad (5)$$

Here d_i is the semantic disruption score for the ith perturbed variant. It indicates how different the perturbed summary is from the baseline. The higher the value, the greater the semantic disruption. To evaluate confidence invariance, the mean normalized log probability was compared across different perturbation types. A clinically safe uncertainty estimator should respond to changes in input. As the perturbation becomes more severe, uncertainty should also increase. If confidence remains the same across different perturbations, it indicates that the model's self-assessment mechanism is failing. Thus, it is not acting as a safety signal.

G. Ablation Study Design

To empirically validate the key design choices of the ConfidentNeuro framework, a systematic ablation study was conducted on GPT-4o. Three components were examined: (1) the choice of nonconformity score function, comparing BERTScore F1, BERTScore Precision, and normalized log probability; (2) the sensitivity of empirical coverage to calibration set size, tested across sizes ranging from 30 to 120 responses using ten random seeds; and (3) the behavior of coverage across seven alpha levels from 0.01 to 0.30. All ablation experiments used the full 150-response GPT-4o evaluation set with fixed random seed 42.

IV. Results

The ConfidentNeuro framework was evaluated on six experimental components. These include (1) baseline model performance; (2) conformal prediction coverage; (3) hallucination risk distribution; (4) calibration quality; (5) perturbation audit; and (6) ablation study. The results for each model are reported on a held-out test set of 70 questions, unless otherwise stated.

A. Baseline Model Performance

Table III shows the BERTScore F1 performance of all three models across the five clinical categories of Alzheimer's. Across the entire benchmark, GPT-4o achieved the highest BERTScore F1

score of 0.766. Next, Mixtral-8x7B scored 0.737, and Claude Haiku scored 0.711. These results confirm that GPT-4o's responses are the most semantically aligned with the ground-truth clinical evidence. Meanwhile, Claude Haiku showed a wide gap relative to the reference answers, under the strict zero-shot protocol.

At the category level, performance trends were fairly consistent across the three models. Disease staging emerged as the highest-scoring category, with GPT-4o reaching 0.801, the highest observed score. Biomarkers, in contrast, were more challenging. Claude Haiku scored 0.695, the lowest across all model–category combinations. This may be due to the more technical and quantitative nature of biomarker-related questions. They rely on precise numerical and biochemical detail, which models often struggle to retrieve in a zero-shot setting.

Interestingly, when we analyzed the results by difficulty level, we found something unexpected. Each of the three models performed with a higher degree of accuracy on the difficult questions in comparison to the simpler ones. Specifically, GPT-4o achieved a score of 0.788 on the harder questions and 0.760 on the easier questions. Claude Haiku's score was 0.723 for difficult questions and 0.701 for easy questions, while Mixtral-8x7B scored 0.757 and 0.725, respectively. We believe this is because LLMs generally generate longer, more thorough answers to questions requiring more consideration. This results in high semantic overlap with comprehensive reference answers, as measured by BERTScore F1. Nevertheless, further investigation is necessary to understand the underlying mechanisms. **Figure 3** shows a category-level performance breakdown for all three models. Table III shows the corresponding BERTScore F1 values for each clinical model-category combination. Across all three models, Disease Staging consistently achieves the highest scores, whereas Biomarkers remains the lowest-scoring category.

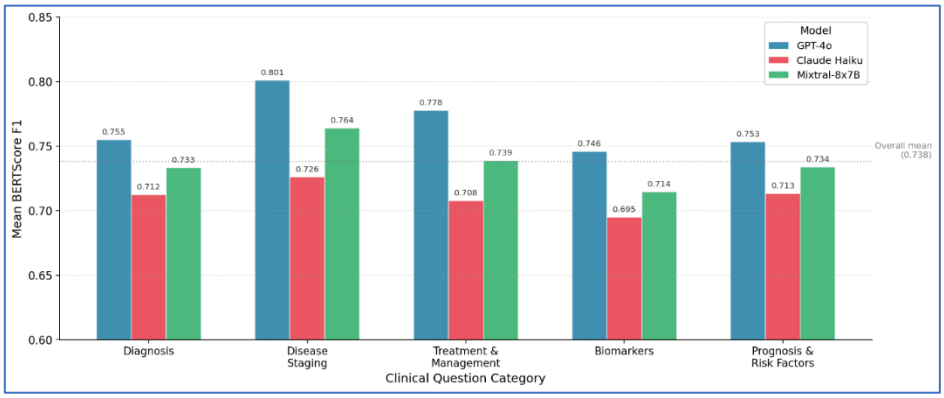


Fig. 3: BERTScore F1 performance of GPT-4o, Claude Haiku, and Mixtral-8x7B across five Alzheimer's clinical categories. The dashed line represents the overall mean BERTScore F1 (0.738) across all models and categories.

TABLE III: BERTScore F1 per model per Alzheimer's clinical category under zero-shot evaluation

Category	GPT-4o	Claude Haiku	Mixtral-8x7B
Diagnosis	0.755	0.712	0.733
Disease Staging	**0.801**	**0.726**	**0.764**
Treatment & Management	0.778	0.708	0.739
Biomarkers	0.746	0.695	0.714
Prognosis & Risk Factors	0.753	0.713	0.734
Overall	**0.766**	**0.711**	**0.737**

B. Conformal Prediction Coverage

The empirical coverage results of the conformal prediction wrapper for each of the 3 models in the analysis were evaluated for the 3 alpha levels, as shown in **Table IV**. At the primary clinical operating point of $\alpha = 0.10$, which represents a 90% nominal coverage guarantee, all three models outperformed their expected targets. GPT-4o and Claude Haiku achieved 97.1% empirical coverage while Mixtral-8×7B had 95.7% empirical coverage. These findings imply that the conformal prediction wrapper provides valid coverage guarantees at the finite sample level. Consequently, it upholds this consistency regardless of the model architecture or output distribution. which is consistent with the theoretical assurance provided by Equation (1).

At a strict coverage setting of $\alpha = 0.05$ (representing 95% nominal coverage), the results were even more conservative. Both Claude Haiku and Mixtral-8x7B achieved a 100% empirical coverage. GPT-4o maintained 97.1% coverage. This performance is considered clinically relevant. In high-stakes settings, such as Alzheimer's decision support, overcoverage is preferable to undercoverage. This reduces the risk of incorrectly classifying a hallucinated response as a safe category.

Under less strict conditions, with $\alpha = 0.20$ (80% nominal coverage), coverage was slightly lower than expected. GPT-4o reached 78.6%, Mixtral-8x7B 77.1%, while Claude Haiku remained higher at 90.0%. This difference is noticeable: at lower α levels, the models appear conservative, but at higher values they start to under-cover. This points to a right-skewed distribution of nonconformity scores, with most responses falling in the low-disruption range and being handled conservatively by the evaluation process. From a clinical perspective, this matters. Using $\alpha \leq 0.10$ is generally safer in high-stakes settings to maintain reliable coverage.

TABLE IV: Empirical coverage of the conformal prediction wrapper across three models at three α levels (target nominal coverage shown in parentheses)

Model	α = 0.05 (95%)	α = 0.10 (90%)	α = 0.20 (80%)	q̂ at α=0.10
GPT-4o	97.1%	97.1%	78.6%	0.333
Claude Haiku	100.0%	97.1%	90.0%	0.338
Mixtral-8x7B	100.0%	95.7%	77.1%	0.327

C. Hallucination Risk Distribution

Fig. 4 and **Table V** present the distribution of hallucination risk tiers for all three models at α = 0.10. We noticed a significant difference in risk profiles across model families. In Mixtral-8x7B, the proportion of HIGH-risk responses were the highest. In contrast, GPT-4o showed the most favorable distribution of LOW-risk responses. In practice, this suggests that risk behavior varies across models and can influence deployment decisions.

GPT-4o assigned 15.7% of test responses to the LOW risk, 81.4% to MEDIUM, and 2.9% to HIGH-risk tiers. This allocation indicates strong semantic alignment of the model, as supported by the highest overall BERTScore F1 = 0.766 reported in Section IV-A. The noticeable proportion of LOW-risk responses, which the other two models lack, exhibits that GPT-4o generates some responses that are not only semantically correct but also confidently fall within the conforming region. In addition, this also suggests that some of the model's outputs may be comparatively more reliable for clinical use.

Claude Haiku showed a different profile; LOW risk responses were 0.0%, MEDIUM 97.1%, and HIGH 2.9%. The complete absence of LOW risk suggests that the nonconformity scores of this model are consistently clustered near but below the conformal threshold. This means that the responses are within the conforming region, but do not clearly establish a strong semantic alignment. This pattern indicates borderline conformity behavior. The rate of HIGH-risk responses is similar to that of GPT-4o. However, the inability to generate LOW-risk responses is a limitation. Therefore, the utility of this model may be restricted to settings where high-confidence clinical outputs are required.

Mixtral-8x7B showed the highest proportion of HIGH risk (4.3%). In addition, 2.9% LOW and 92.9% MEDIUM responses were observed. This result was expected because it was the only open-source model in this evaluation. Such models, with a similar parameter size, show greater variability on specialized clinical tasks [26]. However, the conformal wrapper clearly identified HIGH-risk responses. This indicates that ConfidentNeuro can flag potentially hallucinated outputs, even in open-source setups.

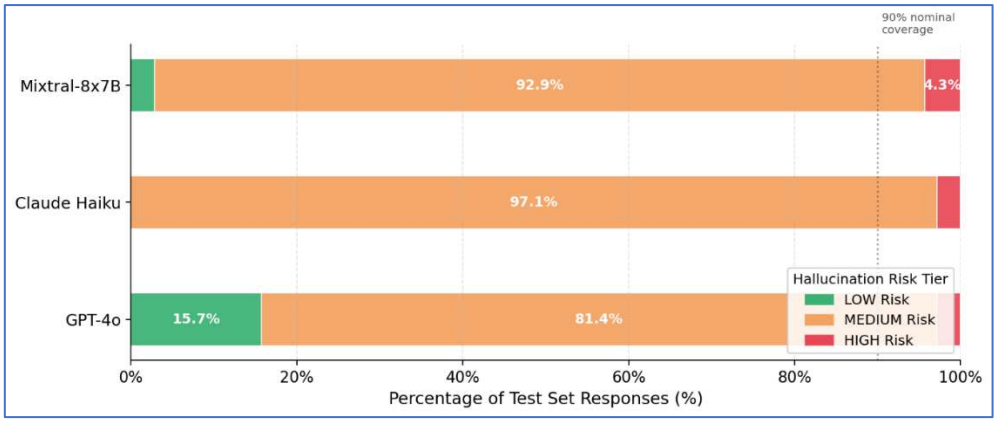


Fig. 4: Hallucination risk tier distribution across GPT-4o, Claude Haiku, and Mixtral-8x7B under conformal prediction at α = 0.10 (n = 70 test questions per model). The dashed vertical line denotes the 90% nominal coverage boundary.

TABLE V: Hallucination risk tier distribution per model at α = 0.10 (test set n = 70 per model)

Model	LOW risk	MEDIUM risk	HIGH risk	$\hat{q}$
GPT-4o	15.7% (n=11)	81.4% (n=57)	2.9% (n=2)	0.333
Claude Haiku	0.0% (n=0)	97.1% (n=68)	2.9% (n=2)	0.338
Mixtral-8x7B	2.9% (n=2)	92.9% (n=65)	4.3% (n=3)	0.327

D. Calibration Quality Analysis

Fig. 5 shows the reliability diagram for GPT-4o. The ECE was 0.213, indicating a visible gap between the model confidence and its actual semantic accuracy. In practice, the model does not always show its correctness precisely. This suggests a calibration mismatch that could pose risks in clinical settings.

The reliability diagram shows overconfidence. The figure indicates that the confidence scores are mostly clustered in the mid-to-high range (0.4-1.0). But the corresponding BERTScore F1 accuracy plateaus around 0.75-0.83 across bins. These values were significantly lower than those expected from a perfectly calibrated model. In an ideal case, the bars would align with the diagonal line y = x. As shown in **Figure 5**, confidence is consistently higher than accuracy across all bins. The pattern reflects that GPT-4o systematically overestimated its certainty. Thus, this behavior was noted specifically for Alzheimer's clinical questions in zero-shot conditions.

The observed tendency has a direct impact on clinical applications. If clinicians rely on GPT-4o's raw confidence scores, they may be systematically misled. The model is not as certain as it seems. Another significant observation is that no responses were generated in the low confidence range (0.0–0.4). These values confirm that GPT-4o almost never indicates low confidence, regardless of the quality

of the response. Such a performance poses challenges in clinical contexts. Despite significant deviation from the ground truth, the model still provides no internal warning. As a result, the clinician has no clear cue to question the output.

A high ECE of 0.213 shows that raw confidence scores cannot be trusted. Importantly, this observation motivated the use of a conformal prediction approach in ConfidentNeuro. The framework does not use raw log-prob-based confidence because it is an unreliable calibration signal. Instead, the BERTScore F1 was used as the basis for the nonconformity score. The coverage guarantees derived from this are independent of the internal confidence estimates of the model. Collectively, these results indicate that the miscalibration is consistent with the findings in Section IV-B. There, the empirical coverage repeatedly exceeded the nominal target at $\alpha = 0.10$. The implication is that the over-coverage occurs because the framework relies on semantic nonconformity rather than raw model confidence estimates.

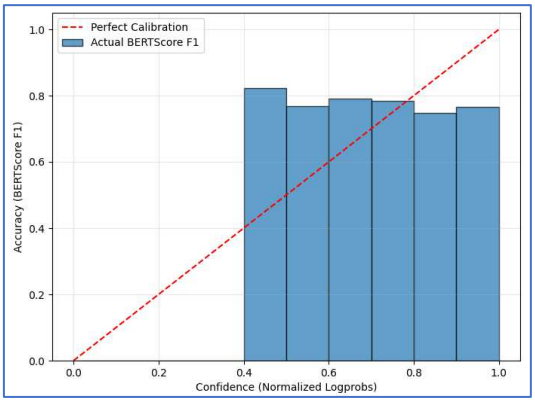


Fig. 5. Reliability diagram for GPT-4o (ECE = 0.213). Bars represent mean BERTScore F1 accuracy per confidence bin. The dashed red diagonal represents perfect calibration. Systematic deviation above the diagonal indicates overconfidence across all confidence levels.

E. Perturbation Audit Results

Fig. 6 and **Table VI** show the results of a structured perturbation audit on GPT-4o using eight clinically relevant perturbation types. One key observation is that confidence remained almost unchanged. Even as semantic disruption increased under these perturbations, confidence stayed within a narrow range of 0.56-0.62 across conditions. In a clinically safe setting, uncertainty tends to increase as input quality degrades. This suggests that GPT-4o's internal confidence is not a reliable safety signal in this setting.

Fig. 6 shows variation in semantic disruption among perturbation types. SENTENCE_DROP has the highest mean disruption (0.050), followed by ENTITY_SWAP (0.041) and SYNONYM_REPLACE (0.034). At the lower end, CONTRADICTION_INSERT caused the least disruption (0.025). This is surprising because it is one of the most clinically risky perturbation types. If a contradictory statement is inserted into a patient's note, the clinician can be directly misled. However,

the model summary did not differ significantly from the baseline. The ratio between the highest and lowest disruptions is approximately 2× between SENTENCE_DROP and CONTRADICTION_INSERT. This implies that the severity of the perturbation is not uniform. However, this variation was not reflected in the model's confidence scores.

Furthermore, a counterintuitive pattern emerged in the TEMPORAL_REORDER and NEGATION_INSERT cases. TEMPORAL_REORDER yielded a normalized confidence score of 0.596, with a corresponding semantic disruption of 0.033. This outcome implies that even after the clinical timeline was completely scrambled, the model remained confident. In contrast, NEGATION_INSERT showed more concerning results. It reverses the meaning of the key clinical findings. The recorded disruption score was only 0.028, which was among the lowest observed. In addition, another key pattern was observed. SYNONYM_REPLACE exhibited the highest confidence of 0.616. A disruption value of 0.034 points to a meaningful change. Alterations in clinical terminology without any clear signal tended to increase the model's confidence. Overall, this suggests that GPT-4o may miss potentially harmful input changes and, at times, overstate its own certainty.

These perturbation audit results support the motivation behind the proposed ConfidentNeuro model. Model confidence does not reliably distinguish between safe and corrupted inputs. This makes an external uncertainty quantification mechanism necessary, such as the conformal prediction wrapper described in Section III-D. In practice, such an approach is important for reliable clinical deployment.

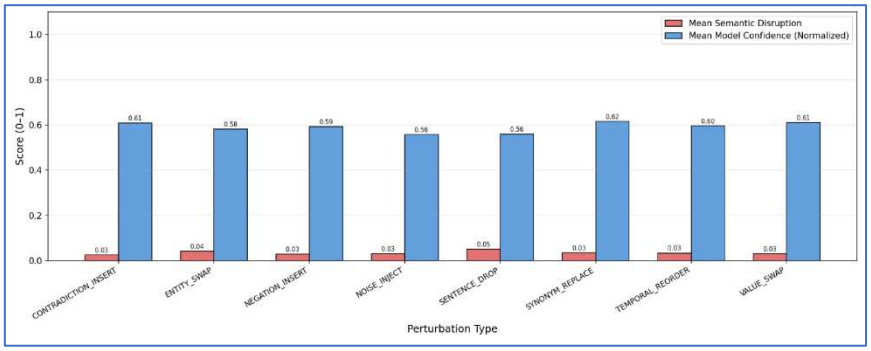


Fig. 6: Semantic disruption and normalized model confidence across eight perturbation types applied to synthetic Alzheimer's clinical notes (GPT-4o, n = 24 per type).

TABLE VI: Perturbation audit summary: mean semantic disruption and normalized model confidence by perturbation type (GPT-4o, v2 perturbations, n = 24 per type)

Perturbation Type	Mean Disruption	Mean Confidence	Gap
SENTENCE_DROP	0.050	0.560	−0.509
ENTITY_SWAP	0.041	0.581	−0.541
SYNONYM_REPLACE	0.034	0.616	−0.582

TEMPORAL_REORDER	0.033	0.596	−0.563
NOISE_INJECT	0.030	0.558	−0.528
VALUE_SWAP	0.030	0.611	−0.581
NEGATION_INSERT	0.028	0.592	−0.564
CONTRADICTION_INSERT	0.025	0.609	−0.584

F. Ablation Study

The ablation study focused on three design aspects of the ConfidentNeuro framework: the nonconformity score function, the impact of calibration set size on coverage, and behavior at different α levels. All experiments were run on GPT-4o using the full set of 150 responses.

Nonconformity Score Function. Fig. 7 compares three candidate nonconformity score functions at $\alpha = 0.10$, using a calibration set of 80 responses. Score A (1 − BERTScore F1) achieved 97.1% empirical coverage, Score B (1 − BERTScore Precision) reached 98.6%, while Score C (1 − normalized log probability) achieved 91.4%, just above the 90% nominal target. Overall, BERTScore-based functions outperform the log probability–based approach. While Score B achieves slightly higher coverage, Score A (BERTScore F1) was selected as the primary nonconformity function. This choice reflects its ability to balance precision and recall, offering a more complete assessment of response quality than a precision-only measure.

Calibration Set Size Sensitivity. Fig. 8 shows the mean empirical coverage across ten random seeds, with calibration size ranging from 30 to 120 responses. Coverage remains stable across sizes, varying from 91.3% (n = 30) to 92.8% (n = 100), with a standard deviation between ±3.1% and ±3.6%. In all cases, coverage exceeds the 90% nominal target, indicating that performance is not strongly dependent on calibration size. As the calibration set grows, variability decreases slightly, from ±3.6% at n = 30 to ±3.1% at n = 120, suggesting improved stability. Based on this, a calibration size of 80 responses was selected as a practical balance between stability and available test data.

Alpha Sensitivity. Fig. 9 shows the relationship between nominal and empirical coverage across seven α levels, ranging from 0.01 to 0.30. A clear asymmetry is observed. For tight coverage targets ($\alpha \leq 0.10$), the empirical coverage was always above the nominal target. For example, at α = 0.01, 100% coverage was obtained for a 99% target. At α = 0.10, 97.1% coverage was obtained against a 90% target. However, for looser targets ($\alpha \geq 0.15$), the coverage tracked closer to or slightly below the nominal value. At α = 0.30, only 67.1% coverage was achieved when the target was 70%. This asymmetric behavior reflects the right-skewed nature of the nonconformity score distribution. The framework is inherently conservative at tight α levels. Consequently, this is desirable in clinical settings. For clinical decision support in Alzheimer's disease, the cost of a missed hallucination is quite high. Therefore, it is

better to have an unnecessary flag than to miss a risky one. For this reason, it is recommended to operate at $\alpha \leq 0.10$.

The ablation results demonstrate that the design choices implemented in the ConfidentNeuro framework are empirically validated. Specifically, these include the use of BERTScore F1 as a nonconformity score, a calibration set of 80 responses, and the selection of $\alpha = 0.10$ as the clinical operating point. Thus, this configuration proved to be stable and robust across varied experimental conditions.

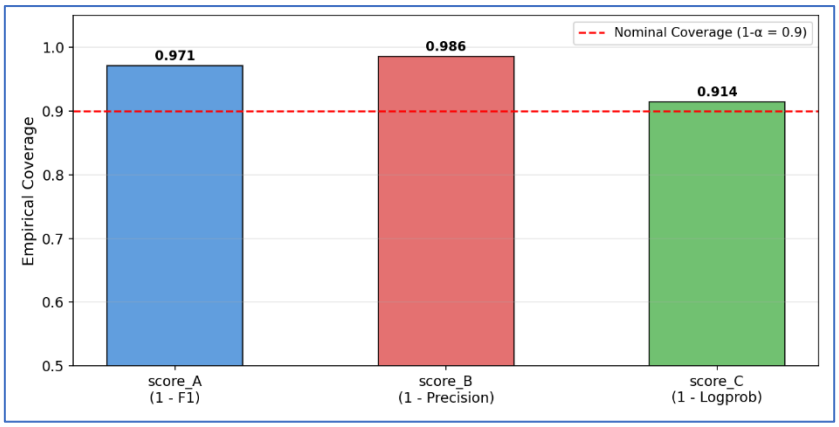


Fig. 7: Comparison of three nonconformity score functions at $\alpha = 0.10$ (GPT-4o, calibration n = 80). The dashed line indicates the 90% nominal coverage target.

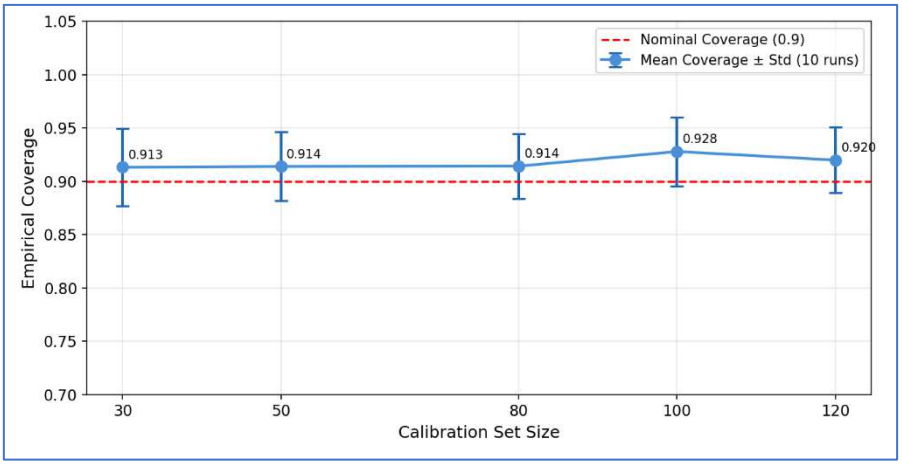


Fig. 8: Mean empirical coverage across ten random seeds for calibration set sizes from 30 to 120 (GPT-4o, $\alpha = 0.10$). Error bars represent ±1 standard deviation.

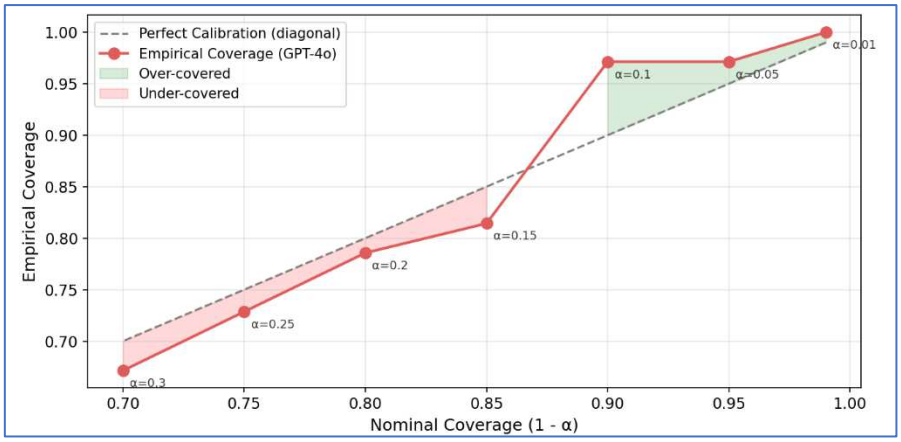


Fig. 9: Empirical versus nominal coverage across seven alpha levels (GPT-4o, Score A). The shaded green region indicates over-coverage; red indicates under-coverage.

V. Discussion

A. Interpretation of Main Findings

These results show that conformal prediction provides a statistically rigorous and practically deployable approach for hallucination risk scoring in LLM-generated clinical responses for AD. Three main findings emerge from this analysis, discussed in detail below.

One of the main findings relates to conformal coverage. At α=0.10, the conformal prediction wrapper achieved empirical coverage above the 90% nominal target across all three models, with GPT-4o and Claude Haiku at 97.1% and Mixtral-8x7B at 95.7%. In practice, this suggests that the finite-sample coverage guarantee of conformal prediction carries over to the clinical QA setting [22]. There is some over-coverage at lower α levels, likely due to a right-skewed distribution of the nonconformity score. In a clinical setting, this is not necessarily a drawback, but the framework flags more cases than it misses.

Another finding relates to calibration. For GPT-4o, the ECE is 0.213, suggesting that raw log-probability-based confidence does not reliably reflect response correctness, especially in clinical use. Similar trends have been reported before, with modern neural networks often showing overconfidence [18]. In this setting, clinical QA for Alzheimer's disease under zero-shot conditions, the issue becomes more serious, as miscalibration can affect patient safety. When a model appears confident even when it is wrong, it can mislead clinical reasoning [16]. Earlier work also points to limited awareness of incorrect answers in LLMs [32], along with a gap between what the model knows and what users assume it knows [19]. In clinical use, this gap can lead to over-reliance on model outputs [6]. One thing

that stands out here is the absence of responses in the low-confidence range (0.0–0.4). Even when outputs deviate from the ground truth, the model does not provide a clear warning signal.

The third finding comes from the perturbation audit, which points to a systematic failure in the model's self-assessment. A consistent pattern of confidence invariance is observed: model confidence remains within a narrow range of 0.56–0.62 across all eight perturbation types, despite a twofold variation in semantic disruption. In practice, this means that whether the clinical input is safe or corrupted, the model's internal confidence does not reflect the difference. This suggests that GPT-4o's internal confidence mechanism cannot reliably assess input integrity. Prior studies report hallucination rates of up to 50–83% in clinical LLMs under adversarial conditions [17], and surveys have identified a lack of uncertainty awareness as a key contributing factor [15]. These findings suggest that, for clinical deployment, external uncertainty quantification is necessary rather than optional.

B. Clinical Implications

These findings have direct implications for deploying LLMs in clinical decision support for AD. Current guidelines emphasize that AI systems should communicate uncertainty and avoid presenting outputs as definitive recommendations [46]. The proposed ConfidentNeuro framework addresses this by assigning each LLM response a statistically grounded risk tier (LOW, MEDIUM, or HIGH), which can be displayed directly within the clinical interface, without requiring clinicians to engage with the underlying statistical details.

The tiered system can fit into routine clinical workflows without much disruption. LOW-risk responses can be used as they are. MEDIUM-risk responses are quite common (about 52%–81.4%) and usually need a quick check, either against guidelines or with a specialist. HIGH-risk responses are less frequent (around 2.9%–4.3%), but these are the cases where extra caution is needed, and clinical action should await expert review. Overall, this approach fits well with existing triage practice and mainly comes down to deciding whether to proceed or take a closer look.

The miscalibration finding (ECE = 0.213) has practical implications for institutions already using GPT-4o in clinical QA. When clinicians rely on the model's confidence, even incorrect responses can seem trustworthy. In some cases, the model remains highly confident even when outputs deviate from clinical evidence. Prior work suggests that the presence of a confidence signal can increase reliance on AI outputs, regardless of their reliability [19]. The ConfidentNeuro framework addresses this by replacing log-probability–based confidence with a semantically grounded nonconformity score, supported by coverage guarantees that do not depend on the model's internal state.

The model-agnostic design of the framework has additional significance for deployment. Hospitals follow strict data governance rules. Therefore, transmitting patient data to closed-source APIs is not

permitted in such setups. In such cases, on-premises open-source deployments such as Mixtral-8x7B are preferred. The results show that ConfidentNeuro clearly identified the HIGH-risk responses (4.3%) of Mixtral-8x7B. This confirms that the conformal wrapper works in both scenarios, i.e., proprietary and open-source. This flexibility is of great importance because healthcare institutions have heterogeneous infrastructures. and the increasing emphasis on data sovereignty in clinical AI deployment [25].

C. Comparison with Existing Work

Broadly, existing approaches to hallucination mitigation in clinical LLMs fall into three categories: prompt engineering, retrieval-augmented generation (RAG), and post-hoc detection. ConfidentNeuro introduces a distinctive fourth methodology, grounded in post hoc uncertainty quantification and supported by statistical guarantees. This approach operates independently of the three existing methodologies and can be used in a complementary manner alongside them.

Most medical LLM evaluations still rely on accuracy alone. Benchmarks such as MedQA, PubMedQA, and MedMCQA mainly test factual recall and multiple-choice reasoning, with results typically reported as a single accuracy score. While this has helped improve model performance, it says little about the reliability of individual responses. A model with 80% accuracy still produces incorrect answers 20% of the time, without indicating which ones are likely to be wrong. More recent work has begun to highlight this gap, noting that accuracy alone is insufficient for clinical use and that uncertainty calibration is equally important [47]. ConfidentNeuro builds on this by shifting the focus to uncertainty-aware evaluation, in which responses are accompanied by runtime risk scores rather than being judged solely by accuracy.

Another key difference of ConfidentNeuro is its use of BERTScore F1 as the primary evaluation metric. Traditional clinical benchmarks largely depend on BLEU, ROUGE, or exact-match accuracy. However, BERTScore operates at a semantic level. It captures meaning through contextual embeddings, not just word matches. In open-ended clinical QA, a single correct answer can have multiple valid phrasings, a characteristic of real clinical responses. But surface-level metrics systematically fail to capture this variation. Therefore, BERTScore F1 is the more suitable choice here.

Studies on hallucination in clinical LLMs suggest fairly high error rates, in some cases reaching 50–83% under adversarial conditions [17]. Prompt engineering helps, but it doesn't completely resolve the issue [17]. A key limitation is that hallucination is often measured as an overall statistic, which says little about the reliability of individual responses. ConfidentNeuro builds on this by evaluating each response separately and assigning a statistically grounded risk tier, rather than relying only on aggregate summaries of model behavior.

RAG-based methodologies integrate LLM outputs with retrieved evidence to mitigate hallucinations. Furthermore, evidence suggests that such approaches enhance accuracy, particularly in clinical settings [26]. However, RAG involves practical requirements, i.e., curated knowledge bases, retrieval infrastructure, and domain-specific indexing. ConfidentNeuro does not require any of these. It operates entirely on the model outputs. Thus, it can be deployed in settings where RAG infrastructure is not feasible. Also, RAG and ConfidentNeuro are not alternatives to each other. Both can be used together. A conformal wrapper can also be applied to a RAG-augmented LLM, so that uncertainty guarantees can be provided even with retrieval grounding.

So far, conformal prediction for LLM outputs has mostly been explored in general NLP settings. It's generally considered a reliable, theory-based mechanism to estimate uncertainty, but its use in clinical language tasks remains limited [34]. The closest research focuses on general QA, often using logprob or self-consistency–based scores. ConfidentNeuro takes a slightly different direction. It works directly in the clinical setting, specifically AD, where hallucinations carry more serious consequences than in general QA. It also replaces log-prob–based scores with BERTScore F1. This choice is supported by our calibration results, which show that an ECE of 0.213 indicates that log-prob–based measures are unreliable in this setting.

In contrast to existing medical LLM benchmarking studies, ConfidentNeuro examines a different question: how model confidence behaves under input perturbations. Most prior work has focused on comparing models across tasks or clinical categories, with results mainly reported in terms of accuracy. What is often missing is an assessment of whether the model's confidence actually reflects the quality of the input [26]. The perturbation audit here shows that even when semantic disruption varies by a factor of 2 across perturbation types, the model's confidence barely changes. This raises a practical concern for clinical use; if confidence does not track input quality, the model's internal signal cannot be relied on, regardless of the domain.

The three-tier risk taxonomy (LOW/MEDIUM/HIGH) introduced in ConfidentNeuro moves beyond the usual binary conforming/non-conforming output in standard conformal prediction. Here, continuous nonconformity scores are translated into actionable categories, making them easier to use in practice. This helps connect statistical uncertainty estimates to clinical workflows, which is widely recognized as important in healthcare but remains underexplored in prior work [48].

D. Limitations

This study has several limitations. The evaluation was conducted under a strict zero-shot setup, which provides a controlled and reproducible baseline but does not fully reflect real deployments. In practice, nonconformity scores may vary, so the calibration used here would need to be adjusted. The

perturbation audit was based on a small set of eight synthetic Alzheimer's clinical notes. Although these follow a standard clinical structure, they do not capture the variability and irregularities found in real patient records. In addition, the audit was conducted solely with GPT-4o, and it remains unclear whether similar behavior would be observed in other models. Also, there are some limitations related to the benchmark and evaluation configuration. ConfidentNeuro-Bench was built from PubMed abstracts rather than from full clinical guidelines or expert-validated question sets, which may limit the extent to which it captures clinical reasoning. The sample size within each category is relatively small, and while steps were taken to improve coverage, more extensive, clinician-validated data would strengthen the analysis. Token-level log probabilities were available only for GPT-4o, so a full calibration analysis could not be carried out for the other models. Also, the benchmark is limited to English and does not show the linguistic diversity of real clinical practice. The study focuses on technical validation, and the proposed risk tiers have not yet been tested against clinician judgment or real patient outcomes. Potential clinical validation is still needed before deployment. The benchmark categories and hallucination risk taxonomy were reviewed by a co-author with medical expertise, offering initial domain grounding.

E. Future Work

Future work for this study will address the aforementioned limitations. The next step is to test the framework under RAG conditions. It still needs to be verified whether conformal coverage is applicable when responses are generated from retrieved clinical evidence, and how risk tiers might change in that scenario. Additionally, a perturbation audit should be performed using de-identified clinical notes (such as those from MIMIC-III) to see if the same confidence pattern is observed with more realistic inputs. Alternative uncertainty measures, such as semantic entropy–based proxies, could support the evaluation of models for which token-level probabilities are unavailable. The benchmark, perturbation pipeline, and calibration framework may also be released as standalone Python tools to support wider use. In the long term, clinical validation with practicing clinicians will be crucial to evaluate not only the reliability of risk tiers but also their influence on clinical decision-making and workflow integration.

VI. Conclusion

AD presents significant challenges in clinical care, and the increasing use of LLMs for decision support offers both opportunities and risks. This paper presents ConfidentNeuro, a framework utilizing conformal prediction to manage the risk of hallucinations in LLM-assisted Alzheimer's care. It provides uncertainty estimates for each response and statistical guarantees. Three key findings emerged from the study. First, conformal prediction consistently achieved an empirical coverage rate exceeding the 90% target across the models GPT-4o, Claude Haiku, and Mixtral-8x7B. This indicates that finite-sample guarantees are applicable in this clinical question-and-answer setting. A calibration audit of GPT-4o

showed an ECE of 0.213, indicating that token-level log probabilities are not reliable confidence signals. This evidence supports using BERTScore F1 as the measure of nonconformity. A structured perturbation audit further showed that model confidence remained largely unchanged (0.56–0.62) despite noticeable variation in semantic disruption, indicating a limitation in the model's ability to assess its own outputs. Together, these results suggest that external uncertainty estimation is necessary for safe clinical use. ConfidentNeuro provides a straightforward method for clinical triage using a modular, model-agnostic pipeline that assigns a risk tier to each response. This method does not require access to the model's internal workings. Furthermore, the benchmark, evaluation scripts, and conformal wrapper are available as open-source resources to enhance reproducibility and additional applications.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Institutional Review Board Statement

Not applicable.

Research involving Human Participants and/or Animals

This work doesn't have any connection with human involvement or animal involvement.

Data Availability Statement

Five established MeSH-controlled search terms that identify various aspects of AD were used to query PubMed via the NCBI Entrez API (https://www.ncbi.nlm.nih.gov/home/develop/api/) to retrieve answer references from the PubMed database, using an authorized API key at a maximum rate of 10 requests per second.

Author Contributions

Afreen Khan: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, review & editing. **Samreen Khan:** Clinical validation, Medical domain expertise, Clinical review of Alzheimer's disease categories and risk tier framework. **Mudassir Khan, Mazliham Mohd Su'ud, Muhammad Mansoor Alam:** Review & editing the

manuscript, Methodology, Validation. **Saurav Mallik:** Review & editing the manuscript, Methodology, Validation.

Acknowledgments

The authors used Grammarly to improve the manuscript's language and readability. All content was reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the work. The ConfidentNeuro benchmark, evaluation scripts, and conformal calibration wrapper are publicly available at https://github.com/afreen-kh/ConfidentNeuro to support reproducibility. The entire benchmark, including all questions/answers and source PMIDs, can be accessed at https://github.com/afreen-kh/ConfidentNeuro.

Funding

The research is funded by Multimedia University.

References

[1] "2025 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 21, no. 4, p. e70235, Apr. 2025, doi: 10.1002/ALZ.70235.

[2] "Dementia." Accessed: May 04, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[3] "Alzheimer's Facts and Figures Report | Alzheimer's Association." Accessed: May 04, 2026. [Online]. Available: https://www.alz.org/alzheimers-dementia/facts-figures

[4] Z. Gu, B. Ge, Y. Wang, Y. Gong, and M. Qi, "Artificial intelligence technologies for enhancing neurofunctionalities: a comprehensive review with applications in Alzheimer's disease research," *Front. Aging Neurosci.*, vol. 17, p. 1609063, Aug. 2025, doi: 10.3389/FNAGI.2025.1609063/TEXT.

[5] F. Busch *et al.*, "Current applications and challenges in large language models for patient care: a systematic review," *Communications Medicine 2025 5:1*, vol. 5, no. 1, pp. 26-, Jan. 2025, doi: 10.1038/s43856-024-00717-2.

[6] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature 2023 620:7972*, vol. 620, no. 7972, pp. 172–180, Jul. 2023, doi: 10.1038/s41586-023-06291-2.

[7] R. Li, X. Wang, D. Berlowitz, J. Mez, H. Lin, and H. Yu, "CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes," *npj Digital Medicine 2025 8:1*, vol. 8, no. 1, pp. 541-, Aug. 2025, doi: 10.1038/s41746-025-01940-4.

[8] M. Zhang, Y. Pan, Q. Cui, Y. Lü, and W. Yu, "Multimodal LLM for enhanced Alzheimer's Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data," *Exp. Gerontol.*, vol. 208, Sep. 2025, doi: 10.1016/j.exger.2025.112812.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," *8th International Conference on Learning Representations, ICLR 2020*, Apr. 2019, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/1904.09675

[10] A. Khan, S. Zubair, and I. Ali, "ADNeuroNet: a neuroevolution-based neural network algorithm for the diagnosis of neurodegenerative diseases," *Neural Computing and Applications 2025 37:14*, vol. 37, no. 14, pp. 8593–8627, Feb. 2025, doi: 10.1007/S00521-025-11021-Y.

[11] A. Khan, S. Zubair, M. Shuaib, A. Sheneamer, S. Alam, and B. Assiri, "Development of a robust parallel and multi-composite machine learning model for improved diagnosis of Alzheimer's disease: correlation with dementia-associated drug usage and AT(N) protein biomarkers," *Front. Neurosci.*, vol. 18, p. 1391465, Sep. 2024, doi: 10.3389/FNINS.2024.1391465/TEXT.

[12] U. Mumtaz, A. Ahmed, and S. Mumtaz, "LLMs-Healthcare : Current Applications and Challenges of Large Language Models in various Medical Specialties," Oct. 2023, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2311.12882

[13] X. Du *et al.*, "Enhancing Early Detection of Cognitive Decline in the Elderly: A Comparative Study Utilizing Large Language Models in Clinical Notes," *medRxiv*, May 2024, doi: 10.1101/2024.04.03.24305298.

[14] M. S. Treder, S. Lee, and K. A. Tsvetanov, "Introduction to Large Language Models (LLMs) for dementia care and research," *Frontiers in Dementia*, vol. 3, p. 1385303, May 2024, doi: 10.3389/FRDEM.2024.1385303/FULL.

[15] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Nov. 2024, doi: 10.1145/3703155.

[16] D. Roustan and F. Bastardot, "The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations," *Interact. J. Med. Res.*, vol. 14, p. e59823, Jan. 2025, doi: 10.2196/59823.

[17] M. Omar *et al.*, "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *Communications Medicine 2025 5:1*, vol. 5, no. 1, pp. 330-, Aug. 2025, doi: 10.1038/s43856-025-01021-3.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 2130–2143, Jun. 2017, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/1706.04599

[19] M. Steyvers *et al.*, "What large language models know and what people think they know," *Nature Machine Intelligence 2025 7:2*, vol. 7, no. 2, pp. 221–231, Jan. 2025, doi: 10.1038/s42256-024-00976-7.

[20] O. Freyer, I. C. Wiest, J. N. Kather, and S. Gilbert, "A future role for health applications of large language models depends on regulators enforcing safety standards," *Lancet Digit. Health*, vol. 6, no. 9, pp. e662–e672, Sep. 2024, doi: 10.1016/S2589-7500(24)00124-9.

[21] V. Quach *et al.*, "Conformal Language Modeling," *12th International Conference on Learning Representations, ICLR 2024*, Jun. 2023, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2306.10193

[22] A. N. Angelopoulos and S. Bates, “A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification,” Jul. 2021, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2107.07511

[23] C. Lu, A. N. Angelopoulos, and S. Pomerantz, “Improving Trustworthiness of AI Disease Severity Rating in Medical Imaging with Ordinal Conformal Prediction Sets,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13438 LNCS, pp. 545–554, Jul. 2022, doi: 10.1007/978-3-031-16452-1_52.

[24] A. N. Angelopoulos *et al.*, “Conformal Triage for Medical Imaging AI Deployment,” *medRxiv*, p. 2024.02.09.24302543, Feb. 2024, doi: 10.1101/2024.02.09.24302543.

[25] R. Hussein *et al.*, “Advancing healthcare AI governance through a comprehensive maturity model based on systematic review,” *npj Digital Medicine 2026 9:1*, vol. 9, no. 1, pp. 236-, Feb. 2026, doi: 10.1038/s41746-026-02418-7.

[26] S. A. A. Safavi-Naini *et al.*, “Benchmarking proprietary and open-source language and vision-language models for gastroenterology clinical reasoning,” *npj Digital Medicine 2025 8:1*, vol. 8, no. 1, pp. 797-, Nov. 2025, doi: 10.1038/s41746-025-02174-0.

[27] S. Collot, C. Fraser, J. Zhao, W. F. Shen, T. Willi, and I. Leontiadis, “Balanced Accuracy: The Right Metric for Evaluating LLM Judges -- Explained through Youden’s J statistic,” Jan. 2026, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2512.08121v2

[28] Y. Kim *et al.*, “Medical Hallucination in Foundation Models and Their Impact on Healthcare,” *medRxiv*, p. 2025.02.28.25323115, Nov. 2025, doi: 10.1101/2025.02.28.25323115.

[29] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, “Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review,” *Ann. Intern. Med.*, vol. 177, no. 2, pp. 210–220, Feb. 2024, doi: 10.7326/M23-2772.

[30] Y. Yang *et al.*, “Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine,” Nov. 2024, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2411.14487

[31] V. Dheur and S. Ben Taieb, “A Large-Scale Study of Probabilistic Calibration in Neural Network Regression,” *Proc. Mach. Learn. Res.*, vol. 202, pp. 7813–7836, Jun. 2023, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2306.02738

[32] S. Kadavath *et al.*, “Language Models (Mostly) Know What They Know,” Jul. 2022, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2207.05221

[33] B. Lambert, F. Forbes, S. Doyle, H. Dehaene, and M. Dojat, “Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis,” *Artif. Intell. Med.*, vol. 150, p. 102830, Apr. 2024, doi: 10.1016/J.ARTMED.2024.102830.

[34] M. Campos, A. Farinhas, C. Zerva, M. A. T. Figueiredo, and A. F. T. Martins, “Conformal Prediction for Natural Language Processing: A Survey,” *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 1497–1516, Nov. 2024, doi: 10.1162/TACL_A_00715.

[35] J. Fayyad, S. Alijani, and H. Najjaran, “Empirical validation of Conformal Prediction for trustworthy skin lesions classification,” *Comput. Methods Programs Biomed.*, vol. 253, p. 108231, Aug. 2024, doi: 10.1016/J.CMPB.2024.108231.

[36] S. Casola, Y. J. Liu, S. Peng, O. Kraus, A. Gatt, and B. Plank, “Evaluation Should Not Ignore Variation: On the Impact of Reference Set Choice on Summarization Metrics,” Sep. 2025, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2506.14335v1

[37] Y. Bang *et al.*, “HalluLens: LLM Hallucination Benchmark,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 24128–24156, Apr. 2025, doi: 10.18653/v1/2025.acl-long.1176.

[38] J. Vazquez and J. C. Facelli, “Conformal Prediction in Clinical Medical Sciences,” *J. Healthc. Inform. Res.*, vol. 6, no. 3, p. 241, Sep. 2022, doi: 10.1007/S41666-021-00113-8.

[39] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson, “Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 16166-, Jul. 2024, doi: 10.1038/s41598-024-65954-w.

[40] I. F. Shihab, S. Akter, and A. Sharma, “Continual Calibration: Coverage Can Collapse Before Accuracy in Lifelong LLM Fine-Tuning,” Apr. 2026, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2604.23987v1

[41] V. Plassier *et al.*, “Rectifying Conformity Scores for Better Conditional Coverage,” Aug. 2025, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/pdf/2502.16336v2

[42] N. Olchanski, J. T. Cohen, P. J. Neumann, J. B. Wong, and D. M. Kent, “Understanding the value of individualized information: the impact of poor calibration or discrimination in outcome prediction models,” *Med. Decis. Making*, vol. 37, no. 7, p. 790, Oct. 2017, doi: 10.1177/0272989X17704855.

[43] M. Majlatow, F. A. Shakil, A. Emrich, and N. Mehdiyev, “Uncertainty-Aware Predictive Process Monitoring in Healthcare: Explainable Insights into Probability Calibration for Conformal Prediction,” *Applied Sciences 2025, Vol. 15, Page 7925*, vol. 15, no. 14, p. 7925, Jul. 2025, doi: 10.3390/APP15147925.

[44] J. S. Berkowitz, J. R. Patock, A. Nawaz, G. Gonzalez-Hernandez, and N. P. Tatonetti, “A crisis of overconfidence: Why confidence, not accuracy, is the real risk in clinical AI,” *BioData Min.*, vol. 19, no. 1, Dec. 2026, doi: 10.1186/S13040-026-00518-4.

[45] A. T. Young *et al.*, “Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models,” *npj Digital Medicine 2021 4:1*, vol. 4, no. 1, pp. 10-, Jan. 2021, doi: 10.1038/s41746-020-00380-6.

[46] M. Elhaddad and S. Hamam, “AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential,” *Cureus*, vol. 16, no. 4, Apr. 2024, doi: 10.7759/CUREUS.57728.

[47] S. Wang *et al.*, “A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains,” *npj Digital Medicine 2025 9:1*, vol. 9, no. 1, pp. 91-, Dec. 2025, doi: 10.1038/s41746-025-02277-8.

[48] J. Vazquez and J. C. Facelli, “Conformal Prediction in Clinical Medical Sciences,” *Journal of Healthcare Informatics Research 2022 6:3*, vol. 6, no. 3, pp. 241–252, Jan. 2022, doi: 10.1007/S41666-021-00113-8.