Polysemantic Dropout: Conformal OOD Detection for Specialized LLMs

Ayush Gupta[1,2], Ramneet Kaur[1], Anirban Roy[1], Adam D. Cobb[1], Rama Chellappa[2], Susmit Jha[1]

[1]Computer Science Lab, SRI, [2]Johns Hopkins University

Correspondence: agupt120@jh.edu

Abstract

We propose a novel inference-time out-of-domain (OOD) detection algorithm for specialized large language models (LLMs). Despite achieving state-of-the-art performance on in-domain tasks through fine-tuning, specialized LLMs remain vulnerable to incorrect or unreliable outputs when presented with OOD inputs, posing risks in critical applications. Our method leverages the Inductive Conformal Anomaly Detection (ICAD) framework, using a new non-conformity measure based on the model's dropout tolerance. Motivated by recent findings on polysemanticity and redundancy in LLMs, we hypothesize that in-domain inputs exhibit higher dropout tolerance than OOD inputs. We aggregate dropout tolerance across multiple layers via a valid ensemble approach, improving detection while maintaining theoretical false alarm bounds from ICAD. Experiments with medical-specialized LLMs show that our approach detects OOD inputs better than baseline methods, with AUROC improvements of 2% to 37% when treating OOD datapoints as positives and in-domain test datapoints as negatives.

1 Introduction

LLMs' ability to generate coherent, contextually relevant and often human-level language has led to their rapid adoption in industry and research, powering applications such as recommendation systems (Sun et al., 2019), legal analysis (Chalkidis et al., 2020), literature review (He et al., 2024), drug discovery (Guan and Wang, 2024), and clinical decision support (Thirunavukarasu et al., 2023). When these models are fine-tuned for specialized tasks, they achieve state-of-the-art performance by leveraging the domain-specific knowledge (Parthasarathy et al., 2024). However, these fine-tuned LLMs remain susceptible to errors when confronted with data that falls outside the scope of their domain. Fig. 1 shows incorrect responses

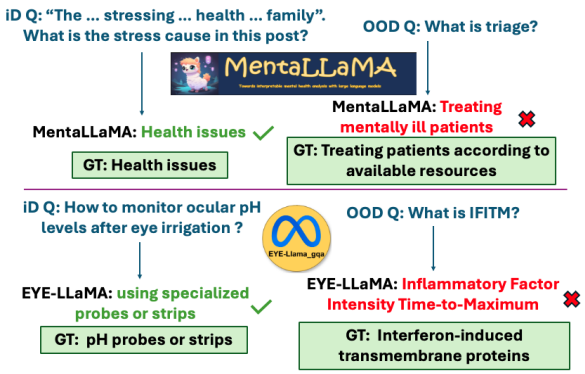

Figure 1: LLMs specialized in Medical Domain work well on in-domain queries, but are prone to make mistakes on out-of-domain (OOD) queries: MentaLLaMA associates its responses to OOD queries with mental health, and EYE-LLaMA hallucinates on those. Here 'GT' stands for the Ground Truth Answer.

by two medical LLMs when prompted for out-of-domain (OOD) questions. According to our analysis, MentaLLaMA (Yang et al., 2024b): an LLM specialized in mental health analysis, associates most of its responses on OOD inputs with mental health. On the other hand, EYE-LLaMA (Haghighi et al., 2024): an LLM specialized in ophthalmology, mostly hallucinates on these OOD queries.

In this paper, we address the challenge of detecting OOD inputs for specialized LLMs, aiming to enhance their reliability and safety in real-world applications. We leverage the Inductive Conformal Anomaly Detection (ICAD) framework (Laxhammar and Falkman, 2015) for OOD detection with bounded false alarms. Central to this framework is the non-conformity measure (NCM), a real-valued function that quantifies the non-conformity of an input to the training distribution. The success of ICAD depends on the choice of NCM: a good measure that can distinguish between in and out-of-domain inputs. We propose to utilize the dropout tolerance of these specialized LLMs as the NCM in the ICAD framework for OOD detection. We define the dropout tolerance of an LLM on an in-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11757–11770
November 4-9, 2025

put query x as the minimum fraction of neurons required to be dropped from a layer of the model to change its original prediction on x. Fig. 2 gives an overview of the proposed approach.

LLMs are expected to be robust to perturbations such as dropout due to redundancy or distribution of concepts across neurons. Polysemantic code learned by these networks favors redundancy (Marshall and Kirchner, 2024), and we hypothesize this redundancy to be higher for in-domain than OOD inputs for LLMs specialized in a particular domain. Polysemanticity (Huben et al., 2023) is a widely investigated topic in the mechanistic interpretability research community. It refers to the phenomenon where neurons activate on multiple concepts to maximize the model's capacity, thereby making these models challenging to interpret. The contributions of this paper can be summarized as follows:

1. Novel NCM: We propose a novel NCM based on dropout tolerance of LLMs in the ICAD framework for detection of out-of-domain or out-of-distribution detection for specialized LLMs[1].

2. Ensemble Approach: Instead of relying on a single layer for the dropout, we propose an ensemble approach where detections from different layers can be combined by a valid merging function while preserving the false alarm rate guarantees of the ICAD framework.

3. OOD Detection Algorithm: We propose an inference-time OOD detection algorithm based on the ensemble approach with the proposed NCM.

4. Experimental Evaluation: We perform extensive experiments on LLMs specialized in the medical domain: MentaLLaMA and EYE-LLaMA, on multiple OOD datasets. We compare AUROC and ROC results on OOD detection with three baselines, and empirically evaluate false alarm guarantees along with several ablation studies for the proposed algorithm.

2 Background

2.1 Polysemanticity and Dropout

Polysemanticity refers to the phenomenon where individual neurons within LLMs activate on multiple, often disparate concepts or features. The hypothesized cause of polysemanticity is *superposition*, where these models encode far more features than neurons. This is accomplished by distributing features to an over-complete set of directions in the activation space rather than to individual neurons to maximize the model's capacity, making it difficult to interpret LLMs (Huben et al., 2023).

Marshall and Kirchner (2024) connect information theory with polysemanticity where polysemantic code learned by LLMs is not only efficient but also favors *redundancy* for robustness (Fig. 2 of their paper). Redundancy refers to the distribution of features across neurons, therefore, discouraging monosemantic code [2] and making the model robust to noise or perturbations such as dropout. Dropout is a technique that drops a fraction of neurons from the neural network while making predictions on an input. It was introduced as a regularization technique to avoid overfitting during the training phase where a neuron would be dropped with the probability p (dropout rate) at each training iteration, and weights of all neurons would be scaled down by the dropout rate during inference (Srivastava et al., 2014). The idea is to prevent the network from becoming too dependent on certain nodes, or in other words, promote redundancy of features among nodes for generalizability and robustness.

2.2 Conformal Prediction

Conformal prediction (Balasubramanian et al., 2014) is a statistical framework used to assess the degree to which a new input conforms to the training distribution. Central to this framework is the non-conformity measure (NCM), a real-valued function that quantifies the non-conformity of an input (x) with respect to the training distribution by assigning it a non-conformity score α_x. Given a training dataset $X = \{x_1, x_2, \ldots, x_l\}$, the NCM evaluates how atypical an input is with respect to X, with larger scores indicating a higher degree of non-conformity. A variety of NCMs have been proposed in literature, employing methods such as k-nearest neighbors (Vovk et al., 2005), support vector machines (Vovk et al., 2005), random forests (Devetyarov and Nouretdinov, 2010), variational autoencoders (Cai and Koutsoukos, 2020a), memory prototypes (Yang et al., 2024c), transformation equivariance (Kaur et al., 2022, 2024).

Conformal anomaly detection (CAD) (Laxhammar and Falkman, 2011) makes use of the NCM to flag inputs anomalous to the training distribution from its p-value. The p-value of an input x is computed by comparing its non-conformity score α_x

[1] We use 'OOD' for out-of-domain or out-of-distribution in the paper.

[2] Monosemanticity refers to the mapping of a concept or feature to a single neuron.

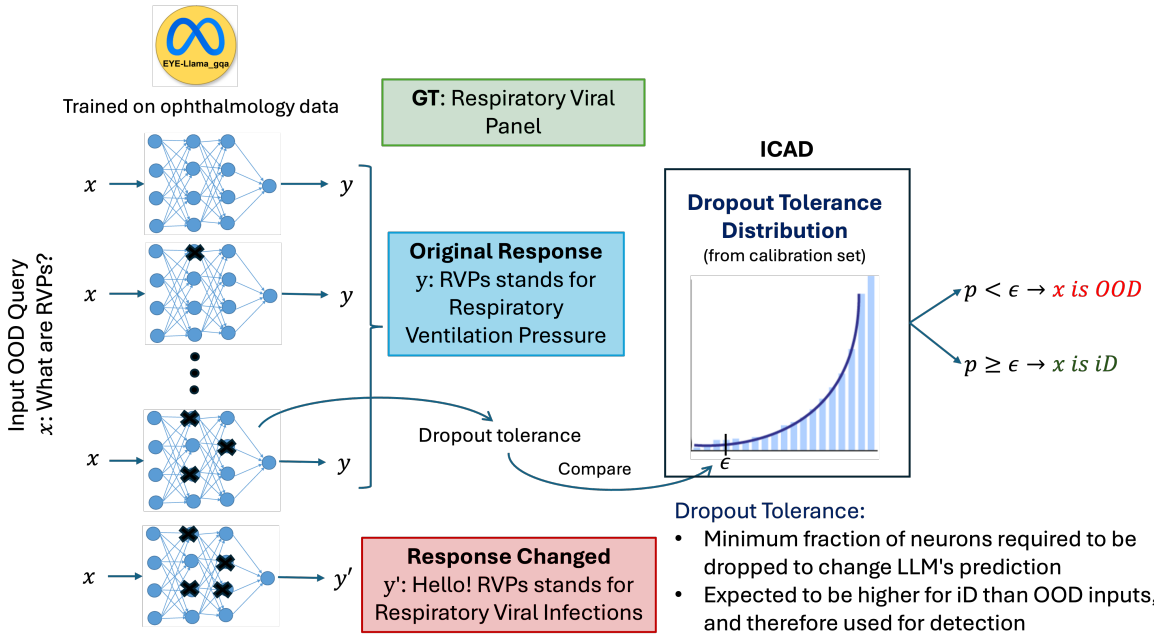

Figure 2: An overview of the proposed approach for OOD detection for specialized LLMs: We propose to leverage *dropout tolerance* for non-conformity measure in the Inductive Conformal Anomaly Detection (ICAD) framework for detection with bounded false alarm rate.

to these scores of the training samples:

$$p\text{-}value = \frac{|i = 1, ..., l : \alpha_x \leq \alpha_i| + 1}{l + 1},$$

where α_i denotes the non-conformity scores of the training data, calculated using the NCM defined on the new dataset constructed from the training dataset of size l and the new input x. If x follows the same distribution as the training data, its score is expected to lie within the range of these scores for the training data, resulting in a higher p-value. Conversely, x is flagged as anomalous if its p-value falls below a chosen detection threshold $\epsilon \in (0, 1)$.

In scenarios where the NCM is computationally expensive, recalculating scores for the entire dataset upon the arrival of each new input imposes a substantial computational overhead. To address this limitation of CAD, inductive conformal anomaly detection (ICAD) was introduced (Laxhammar and Falkman, 2015). In ICAD, the training set is partitioned into a proper training set $X_{\text{tr}} = \{x_1, \ldots, x_m\}$ and a calibration set $X_{\text{cal}} = \{x_{m+1}, \ldots, x_l\}$. NCM is defined on X_{tr}, and the p-value for a new input x is calculated by comparing its non-conformity score α_x to those computed for the calibration set:

$$p\text{-}value = \frac{|i = m + 1, ..., l : \ \alpha_x \leq \alpha_i| + 1}{l - m + 1}. \quad (1)$$

Scores for the calibration set are precomputed offline and used during inference, improving the computational efficiency of the CAD framework. Again, an input is considered anomalous if its p-value is less than the detection threshold ϵ.

Lemma 1. *(Balasubramanian et al., 2014) If the test input x and the calibration points are independent and identically distributed (i.i.d.), the p-value computed in (1) is uniformly distributed in $(0, 1)$. Therefore, the probability of a false alarm (i.e., erroneously labeling a non-anomalous input as anomalous) is upper bounded by the detection threshold ϵ.*

The success of ICAD depends on the choice of NCM used in the framework. We propose to leverage dropout tolerance, i.e. the minimum fraction of neurons that must be deactivated to alter LLMs's prediction. Our hypothesis is that these polysemantic models exhibit greater robustness (i.e., can tolerate higher levels of dropout) on in-distribution (iD) inputs compared to OOD inputs.

2.3 Combining hypothesis

Same hypothesis of "an input drawn from the training distribution" can be tested using multiple conformal anomaly detectors, and merging their results with an ensemble approach can lead to better performance than individual detectors. Multiple p-values $(p_1, \ldots, p_K)$ of an input from (1) by K conformal anomaly detectors can be combined using the fol-

lowing averaging function $M_{r,K}$:

$$M_{r,K}(p_1, \ldots, p_K) = \left(\frac{p_1^r + \ldots + p_K^r}{K} \right)^{1/r},$$

with the special cases of $r = 0$, ∞, and $-\infty$ defined as follows:

$$M_{0,K}(p_1, \ldots, p_K) = \exp\left(\frac{\ln p_1 + \ldots + \ln p_K}{K} \right) = \left(\prod_{k=1}^{K} p_k \right)^{1/K}, \quad (2)$$

$$M_{\infty,K}(p_1, \ldots, p_K) = \max(p_1, \ldots, p_K).$$

$$M_{-\infty,K}(p_1, \ldots, p_K) = \min(p_1, \ldots, p_K), \quad (3)$$

Vovk and Wang (2020) make use of $M_{r,K}$ for defining *valid* merging functions on K p-values:

$$a_{r,K} M_{r,K}(p_1, \ldots, p_K),\ r \in [-\infty, \infty], K \geq 2. \quad (4)$$

Here $a_{r,K}$ is a constant required for making the merging function valid, i.e. preserving the false alarm rate guarantees of ICAD (Lemma 2). We consider the following four merging functions for combining K p-values computed from dropout in K layers of the LLM. These functions vary in the value of r (Vovk and Wang, 2020):

1. **Harmonic Mean (HM)**: With $r = -1$, $a_{-1,K} = (\ln K)$, and $M_{-1,K}(p_1, \ldots, p_K) = \left(\frac{p_1^{-1} + \ldots + p_K^{-1}}{K} \right)^{-1}$.
2. **Arithmetic Mean (AM)**: With $r = 1$, $a_{1,K} = (1 + r)^{1/r} = 2$, and $M_{1,K}(p_1, \ldots, p_K) = \left(\frac{p_1 + \ldots + p_K}{K} \right)$.
3. **Geometric Mean (GM)**: With $r = 0$, $a_{0,K} = e$, and $M_{0,K}(p_1, \ldots, p_K)$ is as defined in (2).
4. **Bonferroni Method (BM)**: With $r = -\infty$, $a_{-\infty,K} = K$, and $M_{-\infty,K}(p_1, \ldots, p_K)$ is as defined in (3).

Lemma 2 (Vovk and Wang, 2020). *If $p_1, \ldots, p_K$ are uniformly distributed in $[0, 1]$, then the value obtained by applying a merging function from (4) is a valid p-value—meaning it is uniformly distributed in $[0, 1]$. Thus, for any $\epsilon \in (0, 1)$, we have $Pr(\text{merged } p\text{-value} < \epsilon) \leq \epsilon$. This validity holds without any assumptions about the dependence among the K p-values.*

3 OOD Detection for Specialized LLMs

Proposed NCM: NCM assigns a score which measures non-conformity of an input with respect to the training distribution. So, it is expected to be higher for OOD and lower for in-distribution (iD) inputs. We, therefore, propose to use $1-$*dropout tolerance* of an LLM's layer as the NCM. Here, dropout tolerance for a layer L is defined as the minimum fraction of neurons in L that must be dropped to change the LLM's original prediction. By original prediction, we mean the prediction made by the model without any dropout.

The intuition behind this score is that an LLM specialized in a particular domain is expected to have higher dropout tolerance for iD inputs than OOD inputs, resulting in higher non-conformity score for OOD than iD inputs. We propose to use this score in the ICAD framework (1) for OOD detection with bounded false alarms.

Ensemble Approach: Instead of relying on a single layer for OOD detection, we propose an ensemble approach where p-values (1) from different layers can be combined by valid merging functions (Vovk and Wang, 2020). We consider the four merging functions defined in Section 2.3 on the K p-values computed from K layers by using the proposed NCM for each layer; thereby preserving the theoretical guarantees from the ICAD framework.

Proposed Algorithm for OOD Detection: With different ways of choosing the neurons to be dropped, comparing semantics of the original response with the one after dropout etc., there can be different ways of implementing an OOD detection algorithm with the merged p-value from the proposed NCM. We describe the specifics of different steps in the proposed Algorithm 1 as follows.

1. Selection of neurons to be dropped: Based on the activations of a layer while generating the last token on an input query x, we construct the list L of the m most activated neurons in the layer [3]. We choose the last token as it captures context for the entire response.

2. Iterative dropout: We query the LLM for x multiple times, each time dropping n additional neurons from L, and checking if the generated response after the dropout is semantically similar to the pre-dropout response from the model.

3. Checking for change in the response: After each dropout iteration, we prompt GPT-4o (Achiam

[3] L contains m maximally activated neurons stored in ascending order of activation.

Algorithm 1 OOD Detection for Specialized LLMs

1: **Input:** Specialized LLM M, Input query x, Maximum number m of neurons to be dropped from a layer, K layers selected for dropout for the ensemble approach, Merging function $M_{r,K}$, K sets of calibration set alphas $\{\alpha_j^k : 1 \leq k \leq K, m+1 \leq j \leq l\}$, Evaluation LLM E, detection threshold ϵ
2: **Output:** "1" if x is detected as OOD; "0" otherwise
3: $y_{orig} = M(x)$ {Original response}
4: Initialize $k = 1$ {For iteration over layers}
5: **while** $k \leq K$ **do**
6: $L =$ list of m maximally activated neurons in layer k on last token of y_{orig}
7: Initialize $i = n$ {For iterative dropout}
8: **while** $i < m$ **do**
9: Drop the first i neurons from L
10: $y_{dropout} = M_{dropout}(x)$
11: **if** $E(y_{current}, y_{dropout})$ == "different" **then**
12: Goto line 16
13: **end if**
14: $i = i + n$
15: **end while**
16: $\alpha_x^k = 1 - \frac{i}{\text{total \#neurons in } k}$
17: $p_k = \frac{|j=m+1,\ldots,l:\ \alpha_x^k \leq \alpha_j^k| + 1}{l-m+1}$
18: **end while**
19: $p_{merged} = M_{r,K}(p_1, \ldots, p_K)$
20: **If** $p_{merged} < \epsilon$ **then return** 1 **else return** 0

et al., 2023) to evaluate whether the generated response is semantically similar to the pre-dropout response. If the response is the same, we continue to the next iteration. Otherwise, we compute the *dropout tolerance* on x as the fraction of neurons dropped to change the pre-dropout response on x.

4. OOD Detection: We compute the non-conformity score α_x as $1-$*dropout tolerance* of the LLM on x. We compare α_x to these scores of all queries in the calibration set (computed offline), to obtain the p-value for x. We do this for (K) layers in the model for getting K p-values, calculate the merged p-value p_{merged} from (4), and compare it with the detection threshold ϵ. x is detected as OOD if $p_{merged} < \epsilon$, and iD otherwise.

4 Experiments

4.1 Specialized LLMs and iD datasets

We consider two LLMs specialized in particular domains of healthcare: EYE-LLaMA and MentaLLaMA. Details of these models and their in-distribution (iD) datasets are as follows.

EYE-LLaMA (Haghighi et al., 2024) is a specialized LLM developed to enhance natural language understanding and QA capabilities within the field of ophthalmology. Built upon LLaMA 2 (Touvron et al., 2023), EYE-LLaMA addresses the unique linguistic and informational needs of ophthalmic practitioners, researchers, and educators. EYE-LLaMA was trained in two phases: pre-training on 766K ophthalmology documents and fine-tuning on the EyeQA dataset.

EyeQA (iD Dataset for EYE-LLaMA) (Haghighi et al., 2024) amalgamates approximately 744,000 unsupervised text samples sourced from PubMed abstracts, 22,000 samples from nearly 570 textbooks, and articles from EyeWiki and Wikipedia's ophthalmology category as the pretraining dataset. For supervised fine-tuning, the dataset includes around 18,000 QA pairs from medical datasets, 1,500 QA pairs from medical forums and is further enriched by 15,000 QA pairs generated by GPT-3.5. This dataset contains a mix of multiple-choice and descriptive queries and was used to fine-tune EYE-LLaMA - to make it specialized for the ophthalmology domain.

MentaLLaMA (Yang et al., 2024b) is an open-source instruction-following LLM developed for interpretable mental health analysis on social media data. Again built upon the LLaMA 2 architecture, MentaLLaMA is designed to perform mental health classification tasks while providing human-readable explanations for its predictions. Multiple MentaLLaMA versions – 7B, 13B and 33B – are available. We utilize the MentaLLaMA-chat-7B variant in this work. This model's training dataset, IMHI, is described below.

IMHI (iD Dataset for MentaLLaMA) (Yang et al., 2024b) curate a dataset called the Interpretable Mental Health Instruction (IMHI) dataset to train MentaLLaMA. IMHI is a multi-task, and multi-source corpus designed to facilitate instruction-tuning of LLMs for interpretable mental health analysis. Comprising approximately 105,000 instruction-response pairs, the dataset has distinct mental health tasks—including depression

detection, stress cause identification, and wellness classification—sourced from platforms such as Reddit and Twitter. The dataset contains long descriptive questions, with answers and detailed reasoning behind the answers generated by ChatGPT (Yang et al., 2024b).

4.2 OOD datasets

We use COVID-QA and MedMCQA as the OOD datasets in our work and evaluate our approach on both LLMs using both of these OOD datasets.

COVID-QA (Möller et al., 2020) is a specialized dataset comprising 2,019 QA pairs annotated by 15 biomedical experts. These annotations are derived from 147 scientific articles focusing on COVID-19-related content. Each entry includes a question, a contextual passage from the source article, and a corresponding answer, formatted in the SQuAD style (Rajpurkar et al., 2016). The dataset features contexts averaging around 6,000 tokens and answers averaging 14 words.

MedMCQA (Pal et al., 2022) is a large-scale MCQ dataset, comprising over 194,000 high-quality questions sourced from AIIMS and NEET PG exams across 21 medical subjects such as Anatomy, Pathology, Pharmacology, and Surgery. Each question includes four answer options and an explanation. Notably, this dataset contains ophthalmology questions, which we filter when using it as OOD for EYE-LLaMA.

We also report results with EYE-QA and IMHI datasets as OOD datasets for MentaLLaMA and EYE-LLaMA respectively, in the Appendix.

4.3 Metrics

With OOD inputs as positives and iD test datapoints as negatives, we compare the Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUROC) for both LLMs on both OOD datasets against the baselines. We also report the false alarm rate guarantee curves by Algorithm 1 for both models and OOD datasets.

4.4 Baselines

We compare our method with three baselines involving the iterative dropout procedure. These baselines are described below.

1. Base Score Method: Score α_x from the proposed NCM can also be directly used for OOD detection without the ICAD framework (or computing p-value from α_x). We use these non-conformity scores for detection and refer to this baseline as the *base score* method. This method, however, does not provide any guarantees on false alarms.

2. Single p-value Method: In this baseline, we use the traditional ICAD approach with just one p-value. This p-value is calculated from the proposed NCM with dropout in a single layer.

3. Ensemble Approach with Majority Voting: Here, we use majority voting instead of a valid merging function on the individual p-values from different layers. Specifically, we run the single p-value method on different layers and use majority voting on those detections. This baseline also does not provide any false alarm rate guarantees.

4.5 Results

With the number of layers $K = 3$ – specifically, layers 7, 15, and 22 – we compare AUROC results with baselines for both the models in Table 1. We choose layers 7, 15, and 22 as each of these layers lies at a different stage of inference: starting, middle, and towards the end, and hence the model has a different understanding of the query at each of these individual layers (Lad et al., 2024). Our approach consistently outperforms all baselines across all test cases. We also compare ROC curves with the single p-value method in Figure 3. Again, the proposed approach achieves the best results with comparable performance from the single p-value method with Layer 7.

Plots on the bounded false alarm rate guarantees by Algorithm 1 are shown in Figure 4 with the range of detection threshold $\epsilon = \{0, 0.05, 0.1, \ldots, 0.5\}$. These plots show that the false alarm rate is upper bounded by the detection threshold for all values of ϵ for EYE-LLaMA. For MentaLLaMA, the guarantees are satisfied for most of the cases except for the higher range of ϵ. This can be attributed to the statistical insignificance of the empirical calibration data representing the training distribution.

All reported results are averaged over five runs with random $80\% - 20\%$ splits of the iD test and calibration sets, and we observe low standard deviation in all cases, typically as low as 0.001.

4.6 Ablation Studies

We perform the following studies relevant to the proposed Algorithm 1.

1. Choice of the Valid Merging Function: As mentioned in Section 2.3, different valid merging functions can be used to combine the K p-values.

Model	EYE-LLaMA		MentaLLaMA	
OOD Dataset	**CovidQA**	**MedMCQA**	**CovidQA**	**MedMCQA**
Base Score Method with Layer 7	0.53	0.54	0.73	0.72
Base Score Method with Layer 15	0.48	0.58	0.71	0.69
Base Score Method with Layer 22	0.48	0.57	0.70	0.70
Single p-value Method with Layer 7	0.77	0.83	0.93	0.94
Single p-value Method with Layer 15	0.61	0.79	0.78	0.78
Single p-value Method with Layer 22	0.56	0.68	0.74	0.73
Ensemble Approach with Majority Voting	0.75	0.81	0.55	0.55
Ours with K=3 (Layers 7, 15, and 22)	**0.83**	**0.91**	**0.95**	**0.96**

Table 1: AUROC (↑) results for EYE-LLaMA and MentaLLaMA on both OOD datasets. Best results are in bold.

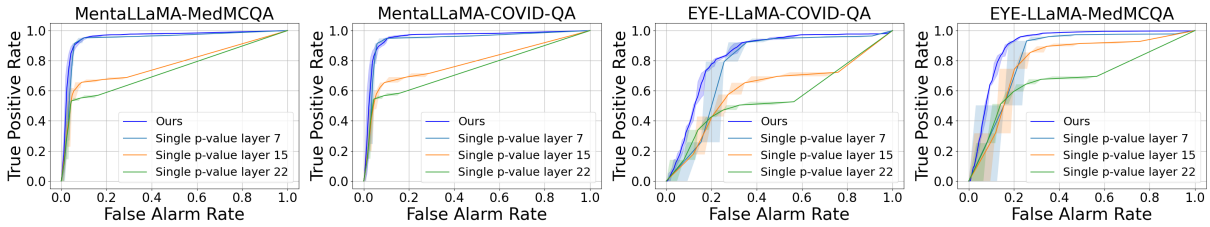

Figure 3: Comparison of ROC curves for OOD Detection with the single p-value baselines across layers.

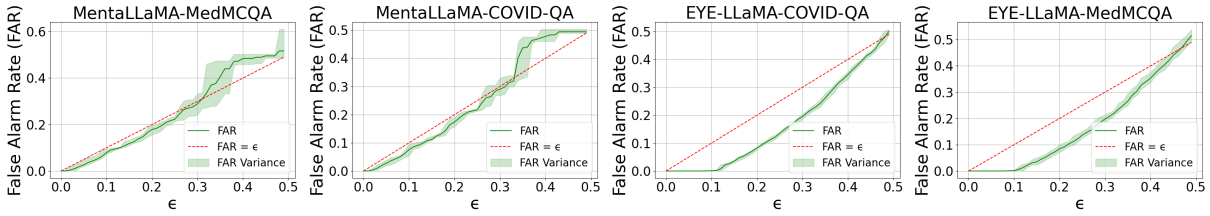

Figure 4: False Alarm Rate Guarantees Plots: False alarm is upper bounded by ϵ on average in most test cases.

Model	EYE-LLaMA		MentaLLaMA	
OOD Dataset	**CovidQA**	**MedMCQA**	**CovidQA**	**MedMCQA**
Bonferroni Method	0.67	0.79	0.93	0.93
Harmonic Mean	0.76	0.85	0.94	0.94
Geometric Mean	0.80	0.89	**0.95**	0.95
Arithmetic Mean	**0.83**	**0.91**	**0.95**	**0.96**

Table 2: AUROC by the proposed approach with different valid p-value merging functions.

We use Arithmetic Mean on p-values from the three layers (7, 15, and 22) for reporting the results in Section 4.5. We also experiment with the other three functions and report the AUROC results in Table 2. We observe that Arithmetic Mean performs the best with comparable performance by Geometric Mean. Performance by all the functions is comparable in the case of MentaLLaMA. ROC curves for the other three valid merging functions (with similar performance as Arithmetic Mean) are included in the Appendix.

2. Unchanged Responses: We set an upper limit of $m = 30$ on the maximum number of neurons to be dropped in Algorithm 1. It is possible that the response to a particular query might not be changed even after dropping all the m maximally activated neurons. For example, while running dropout on layer 7 of EYE-LLaMA on the EyeQA calibration set, we observe that approximately 91% of responses changed with the number of dropped neurons $i \leq m$. However, this percentage varies by layer: 74% for layer 15 and 56% for layer 22. Across our experiments, dropout in earlier layers more often changes responses than in the later layers. We hypothesize this is because early layers are crucial for understanding the query, as also suggested by Lad et al. (2024), resulting in the model being more sensitive to dropout in earlier layers.

For the proposed algorithm, the non-conformity score, and hence the p-value is undefined if a response is not changed within this upper limit of m. When aggregating the p-values from multiple layers, we only consider those layers where the p-value is well defined. If no p-value is defined for any of the layers, we resort to a default prediction of in-distribution for that particular query. The idea being if the response does not change even after dropping all the m maximally activated neurons in all the K layers, then the model is highly robust to the input query: a property more likely to be for iD queries than OOD. This default prediction, however, occurs very rarely: only 3% of the total

queries for Eye-LLaMA. Due to the ensemble approach, our method is able to make a prediction through one of the layers most of the time.

3. Ablation on m: In our experiments, we iteratively drop up to a maximum of $m = 30$ neurons. We analyze the effect of varying this limit on a subset of the EYEQA dataset. Specifically, we fix layer 15 (middle layer) as the dropout layer, and run Algorithm 1 on this subset with $m = 10, 30,$ and 50. We observe that responses for $59\%, 78\%,$ and 81% of total queries are changed within these limits, respectively. As expected, we observe that a higher m is more likely to change a response. It should be noted that increasing m increases the number of iterations in the algorithm. We choose $m = 30$, offering a balance between the fraction of changed responses and computational efficiency.

4. Difficulty Level of Queries: Based on the number of dropped neurons required to change the response, we also try to categorize queries. A venn diagram categorizing queries in the sets based on when they change is shown in Figure 6 of the Appendix. We analyze two extreme 'sets' of queries: 'Set A' with queries whose responses were changed at all $m = 10, 30,$ and 50 (500 queries) and 'Set B' with queries whose responses changed only at $m = 50$ (54 queries). We observe that in Set A, roughly 41% of the queries are MCQs, requiring a choice from given options. On the other hand, Set B has only 10% of such choice-based questions. We further check the proportion of choice-based questions in 'Set C' - containing responses altered at $m = 30$ and 50 but not at $m = 10$ - which is 17%: in between that of Sets A and B.

Thus, we observe that MCQs are more easily altered, requiring fewer dropped neurons compared to subjective queries. This is an interesting observation, which also seems to be supported by our results in Table 1, where we perform better when the OOD dataset is MedMCQA: a set containing entirely MCQs, as compared to COVID-QA, containing entirely subjective queries.

5. Size of Calibration Set: In our experiments, we use 20% of the iD dataset for calibration and the remaining 80% for testing. In this section, we experiment with different proportions of the calibration-test set split. The results are plotted in Figure 5. We observe that a smaller size of calibration set reduces the AUROC. This is because we require a statistically significant amount of calibration data to compute the p-value p_k for a given sample.

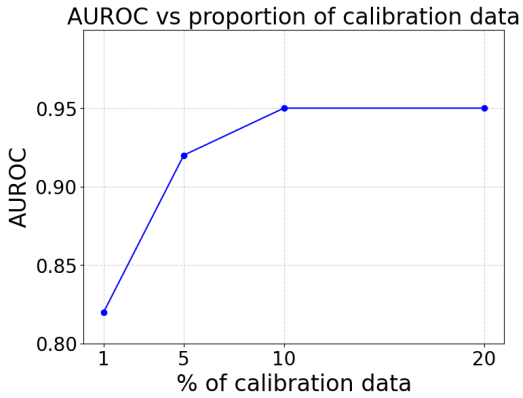

Figure 5: Plot of AUROC against the proportion of the calibration set. Here, the model used is MentaLLaMA. The OOD dataset is CovidQA.

6. Different Entailment Checker Models: In our experiments, we check for semantic entailment between two model responses using GPT-4o. In this section, we also experiment with the DeBERTa (He et al., 2021) and the LLaMA3-8B (AI@Meta, 2024) models to check for semantic entailment.

We observe that the DeBERTa model is not able to perform well on medical domain queries due to the complex terminologies which might not have been part of its training. Manual analysis of the entailment outputs shows that the performance of this model is not satisfactory on the medical queries we use.

However, the LLaMA3-8B model performs decently well on our queries, just like GPT-4o. Using the LLaMA3 model for semantic entailment on the EYE-LLaMA model, and using CovidQA as the OOD dataset, we obtain an AUROC of 0.79 on the OOD detection task (only Layer 7 used for dropout). The GPT-4o equivalent of this achieves a similar AUROC of 0.77.

Thus, we conclude that different LLMs can be used as entailment checkers in our algorithm.

4.7 Compute cost

The proposed Algorithm 1, if run sequentially, requires two nested loops per query. Thus, the number of forward passes are $O(K * \frac{m}{n})$ where K is the number of layers, m is the maximum number of neurons dropped and n is the incremental number of neurons dropped in each iteration. It should be noted that each of the forward passes can be run in parallel since they do not depend on each other. This approach of running multiple inferences on the same model to provide guarantees has been explored in other works (Samplawski et al., 2025;

Wang et al., 2024; Yang et al., 2024a; Kuhn et al., 2023a; Lin et al., 2023a; Kaur et al.; Wang et al., 2022; Padhi et al., 2025).

We use our 7B LLMs in 8bit inference mode, which need 7GB of GPU memory per forward pass. Running multiple forward passes together will increase the memory but reduce the time, so our algorithm gives the user the option to balance time and performance based on the end requirement.

5 Related Work

OOD detection has been of significant research focus for reliable deployment of traditional deep learning as standalone models (Hendrycks and Gimpel, 2016; Lee et al., 2018; Tack et al., 2020; Kaur et al., 2021; Macêdo et al., 2021; Kaur et al., 2023a) or as learning enabled components in closed-loop systems (Cai and Koutsoukos, 2020b; Yang et al., 2022; Ramakrishna et al., 2022; Sundar et al., 2020; Yang et al., 2024c). Some of these approaches are built on ICAD for providing false alarm guarantees (Kaur et al., 2022; Cai and Koutsoukos, 2020b; Kaur et al., 2023b, 2024; Yang et al., 2024c). MC-dropout has also been used as the Bayesian inference approach for quantifying uncertainty in traditional deep learning models (Gal and Ghahramani, 2016; Ryu et al., 2019).

For LLMs, main focus has been on quantifying uncertainty in the models' responses (Kadavath et al., 2022; Lin et al., 2023b; Kuhn et al., 2023b; Shorinwa et al., 2024; Kaur et al.; Padhi et al., 2025). Conformal prediction has been also used to generate sets instead of a single prediction to account for uncertainty in LLM's predictions (Quach et al., 2023; Wang et al., 2025).

OOD detection for LLMs has started to emerge for specialized LLMs such as conditional language models (Ren et al., 2022), models fine-tuned with LoRA adapters (Salimbeni et al., 2024), and for specific tasks like sentiment analysis (Ouyang et al., 2025). In contrast, our method is model- and application-agnostic, applicable to any domain-specialized LLM. Unlike existing works like Ouyang et al. (2025), it requires no learning and operates directly at inference. To the best of our our knowledge, this is the first OOD detection method for LLMs with theoretical guarantees.

6 Conclusion

In this paper, we introduce a novel, model-agnostic conformal OOD detection method for specialized LLMs, leveraging dropout tolerance as a non-conformity measure within the ICAD framework. Our inference-time OOD detection approach aggregates dropout tolerance across multiple layers using valid ensemble merging functions, preserving theoretical false alarm guarantees. Extensive experiments with medical-specialized LLMs demonstrate that our method consistently outperforms baseline approaches—achieving AUROC gains of up to 37%—and adapts to a variety of OOD dataset types and query complexities. Furthermore, our ablation studies provide additional insights into the effects of various design choices, informing the method's applicability, generalizability, and practical utility. Moving forward, this work opens avenues for extending conformal OOD detection to multi-modal LLMs and broader deployment in safety-critical applications.

Limitations

The proposed non-conformity measure, and hence the p-value is undefined if the original response to the query does not change even after dropping the maximum number of neurons. Although this situation was observed only rarely in our case studies and can be mitigated by increasing the number of layers used in the ensemble approach, it may still occur and cause the algorithm to incorrectly classify an OOD input as in-distribution. Further, increasing the number of layers improves the OOD detection performance but it also increases computational cost at inference time. Lastly, our experiments show that the approach is dependent on a representative and large-enough calibration data set to calculate a p-value for an unseen sample.

Acknowledgments

This material is based upon work supported by the United States Air Force and DARPA under Contract No. FA8750-23-C-0519 and HR0011-24-9-0424, and the U.S. Army Research Laboratory under Cooperative Research Agreement W911NF-17-2-0196 and Defense Logistics Agency (DLA) and the Advanced Research Projects Agency for Health (ARPA-H) under Contract Number SP4701-23-C-0073. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force, DARPA, the U.S. Army Research Laboratory, ARPA-H or the United States Government.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. 2014. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Feiyang Cai and Xenofon Koutsoukos. 2020a. Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 174–183. IEEE.

Feiyang Cai and Xenofon Koutsoukos. 2020b. Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 174–183. IEEE.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Dmitry Devetyarov and Ilia Nouretdinov. 2010. Prediction with confidence based on a random forest classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 37–44. Springer.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Shenghui Guan and Guanyu Wang. 2024. Drug discovery and development in the era of artificial intelligence: From machine learning to large language models. *Artificial Intelligence Chemistry*, 2(1):100070.

Tania Haghighi, Sina Gholami, Jared Todd Sokol, Enaika Kishnani, Adnan Ahsaniyan, Holakou Rahmanian, Fares Hedayati, Theodore Leng, and Minhaj Nur Alam. 2024. Eye-llama, an in-domain large language model for ophthalmology. *bioRxiv*.

Junda He, Christoph Treude, and David Lo. 2024. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. *ACM Transactions on Software Engineering and Methodology*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park13, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. 2022. idecode: In-distribution equivariance for conformal out-of-distribution detection.

Ramneet Kaur, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. 2021. Are all outliers alike? on understanding the diversity of outliers for detecting oods. *arXiv preprint arXiv:2103.12628*.

Ramneet Kaur, Xiayan Ji, Souradeep Dutta, Michele Caprio, Yahan Yang, Elena Bernardis, Oleg Sokolsky, and Insup Lee. 2023a. Using semantic information for defining and detecting ood inputs. *arXiv preprint arXiv:2302.11019*.

Ramneet Kaur, Colin Samplawski, Adam D Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander Michael Berenbeim, John A Pavlik, Nathaniel D Bastian, and 1 others. Addressing uncertainty in llms to enhance reliability in generative ai. In *Neurips Safe Generative AI Workshop 2024*.

Ramneet Kaur, Kaustubh Sridhar, Sangon Park, Yahan Yang, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. 2023b. Codit: Conformal out-of-distribution detection in time-series data for cyber-physical systems.

Ramneet Kaur, Yahan Yang, Oleg Sokolsky, and Insup Lee. 2024. Out-of-distribution detection in dependent data for cyber-physical systems with conformal guarantees. *ACM Transactions on Cyber-Physical Systems*, 8(4):1–27.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023a. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023b. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Vedang Lad, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of llms: Stages of inference? *Preprint*, arXiv:2406.19384.

Rikard Laxhammar and Göran Falkman. 2011. Sequential conformal anomaly detection in trajectories based on hausdorff distance. In *14th International Conference on Information Fusion*, pages 1–8. IEEE.

Rikard Laxhammar and Göran Falkman. 2015. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74(1):67–94.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023a. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023b. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano LI Oliveira, and Teresa Ludermir. 2021. Entropic out-of-distribution detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Simon C Marshall and Jan H Kirchner. 2024. Understanding polysemanticity in neural networks through coding theory. *arXiv preprint arXiv:2401.17975*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Tinghui Ouyang, Yoshiki Seo, and Isao Echizen. 2025. Textual out-of-distribution (ood) detection for llm quality assurance. *Knowledge-Based Systems*, 310:112975.

Trilok Padhi, Ramneet Kaur, Adam D Cobb, Manoj Acharya, Anirban Roy, Colin Samplawski, Brian Matejek, Alexander M Berenbeim, Nathaniel D Bastian, and Susmit Jha. 2025. Calibrating uncertainty quantification of multi-modal llms using grounding. *arXiv preprint arXiv:2505.03788*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Shreyas Ramakrishna, Zahra Rahiminasab, Gabor Karsai, Arvind Easwaran, and Abhishek Dubey. 2022. Efficient out-of-distribution detection using latent space of β-vae for cyber-physical systems. *ACM Transactions on Cyber-Physical Systems (TCPS)*, 6(2):1–34.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*.

Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. 2019. Uncertainty quantification of molecular property prediction with bayesian neural networks. *arXiv preprint arXiv:1903.08375*.

Etienne Salimbeni, Francesco Craighero, Renata Khasanova, Milos Vasic, and Pierre Vandergheynst. 2024. Beyond fine-tuning: Lora modules boost near-ood detection and llm security. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Colin Samplawski, Adam D Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. 2025. Scalable bayesian low-rank adaptation of large language models via stochastic variational subspace inference. In *The 41st Conference on Uncertainty in Artificial Intelligence*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Vijaya Kumar Sundar, Shreyas Ramakrishna, Zahra Rahiminasab, Arvind Easwaran, and Abhishek Dubey. 2020. Out-of-distribution detection in multi-label datasets using latent space of β-vae. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 250–255. IEEE.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Vladimir Vovk and Ruodu Wang. 2020. Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Sean Wang, Yicheng Jiang, Yuxin Tang, Lu Cheng, and Hanjie Chen. 2025. Copu: Conformal prediction for uncertainty quantification in natural language generation. *arXiv preprint arXiv:2502.12601*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. 2024. Blob: Bayesian low-rank adaptation by backpropagation for large language models. In *Conference on Neural Information Processing Systems*.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. 2024a. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Yahan Yang, Ramneet Kaur, Souradeep Dutta, and Insup Lee. 2024c. Memory-based distribution shift detection for learning enabled cyber-physical systems with statistical guarantees. *ACM Transactions on Cyber-Physical Systems*, 8(2):1–28.

Yahan Yang, Ramneet Kaur, Suradeep Dutta, and Insup Lee. 2022. Interpretable detection of distribution shifts in learning enabled cyber-physical systems. In *ACMIEEE International Conference on CyberPhysical Systems*.

A Appendix

A.1 Implementation details

We set the limit on the maximum number of neurons to be dropped, $m = 30$ in our main experiments. We set the step size $n = 5$, the incremental number of neurons dropped in each successive iteration of dropout. We used GPT-4o-mini to evaluate whether the pre-dropout response is semantically similar to the post-dropout response. Inspired by (Lad et al., 2024), we choose layers 7, 15, and 22 to perform the dropout - each of these layers lies in a different stage of inference, and hence the model has a different understanding of the query at each of these individual layers. We use a random $20\% - 80\%$ calibration-test split on the in-domain data for each run. We load the LLMs in 8-bit precision and perform inference on A100 GPUs. We use PyTorch hooks to get internal activations of the LLMs and trigger dropout.

Also, both EYE-LLaMA and MentaLLaMA are open-source and publicly available to be used for research purposes under MIT license.

A.2 More evaluation results

We have already evaluated with CovidQA and MedMCQA as the OOD datasets. However, the EYE-QA and IMHI dataset are also OOD for MentaLLaMA and EYE-LLaMA respectively. In this section, we evaluate the AUROC for these model-dataset combinations as well.

The ROC and false alarm curves for these combinations are shown in Figure 7. Our method achieves an AUROC of 0.93 for the MentaLLaMA - EyeQA combination, and 0.82 for EYE-LLaMA - IMHI combination. This shows that our method is not specific to particular OOD datasets - even the datasets that are in-domain for one model can be used as OOD for another model, and our method can still detect them as OOD.

A.3 Analysis of different merging functions

In Table 2, we provide the average AUROC values for when we use different merging functions in our method. Here, we plot the ROC and False alarm curves for each of these valid merging functions.

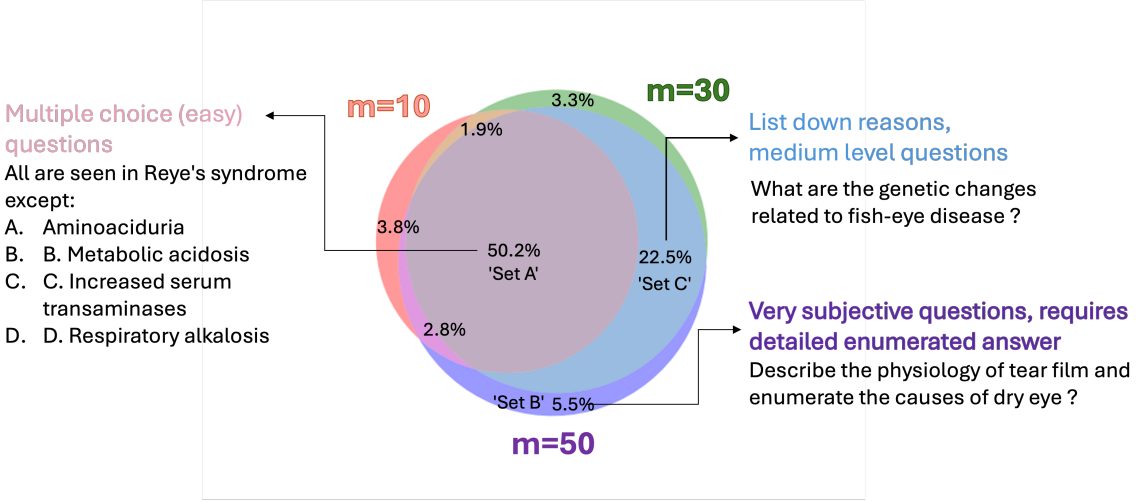

Figure 6: Ablation on the difficulty level of queries measured by varying the maximum number m of dropped neurons in the Layer 15 of EyeLLaMA model. We observe that responses to MCQs are more easily altered, requiring fewer neurons to be dropped compared to subjective queries. As the query becomes more subjective, it requires more neurons to change the response.

Figure 8 shows the plots for Geometric mean, Figure 9 for Harmonic mean and Figure 10 for the Bonferroni merging function. We note that while there is not a significant difference in the AUROC regardless of the merging function used, the false alarm curves have a distinct shapes depending on the merging function. Specifically, the Bonferroni method is seen to have a larger false alarm rate for a given ϵ than other methods, potentially because it is more sensitive to extremely low p-values than other methods.

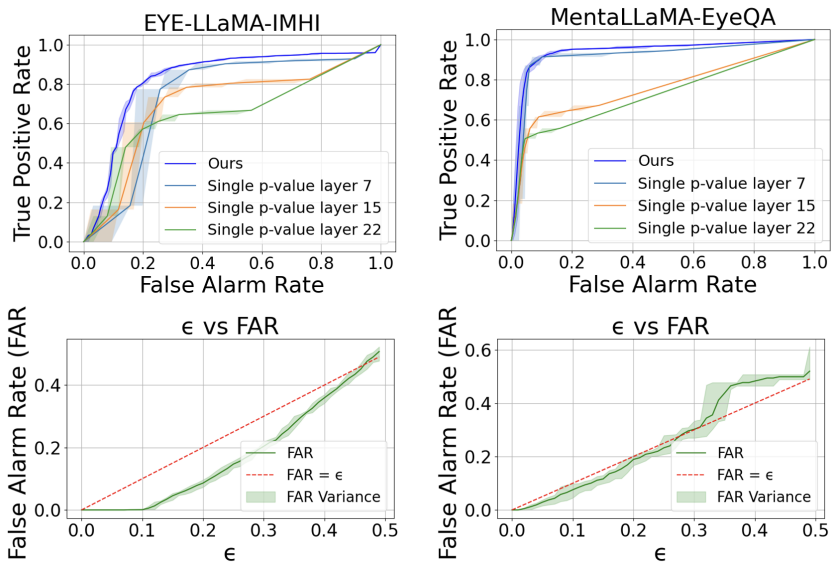

Figure 7: ROC and False Alarm Guarantee curves on more datasets, using Arithmetic Mean as the valid merging function.

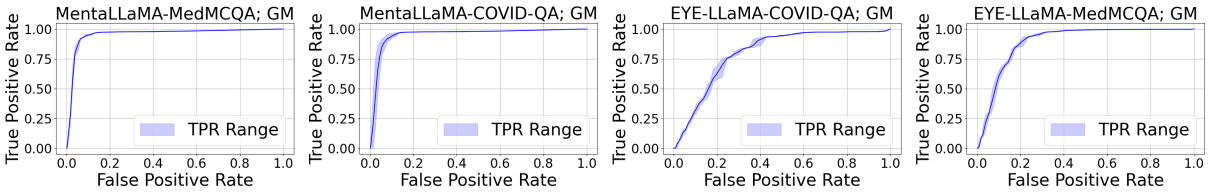

Figure 8: ROC curves with Geometric Mean as the valid merging function.

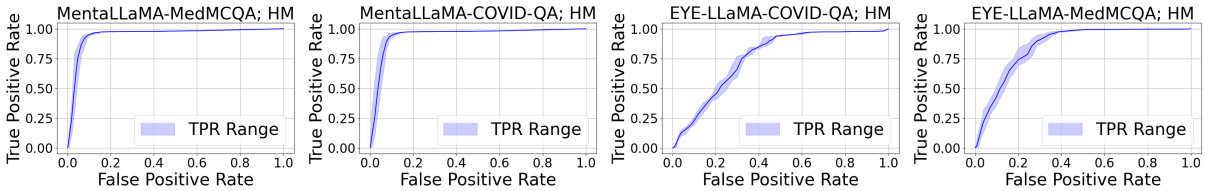

Figure 9: ROC curves with Harmonic Mean as the valid merging function.

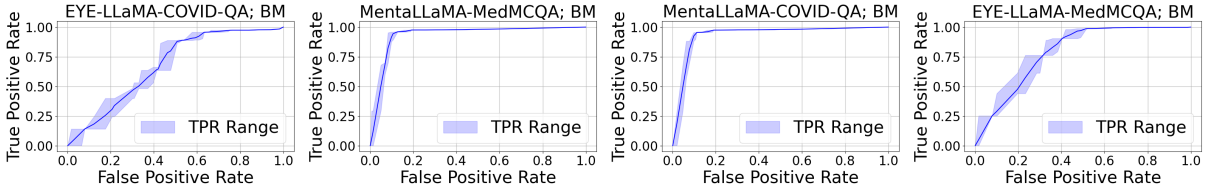

Figure 10: ROC curves with Bonferroni Method as the valid merging function.