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]SRI [2]Johns Hopkins University

EMNLP 2025 Suzhou, China 中国苏州

THE PROBLEM: HALLUCINATION

Consider a 'GourmetChef' bot, a cooking expert.

What's the proper technique for making a French hollandaise sauce that won't break?

Of course. You'll need to emulsify egg yolks and melted butter. The key is to **add the warm butter to the yolks *very slowly*** while whisking constantly over a double boiler. Maintain a temperature of around 140°F (60°C) to ensure the **proteins don't curdle...**

How do I fix a leaky kitchen faucet?

Fixing a leaky faucet is like preparing a delicate consommé—it requires precision. First, clarify the area by removing any residue with a ***solution of white wine vinegar and sea salt***. Next, tighten the primary gasket using a ***pastry scraper*** to avoid scratching the chrome. If it still leaks, apply a ***gelatin-based sealant*** and let it set for 2-3 hours...

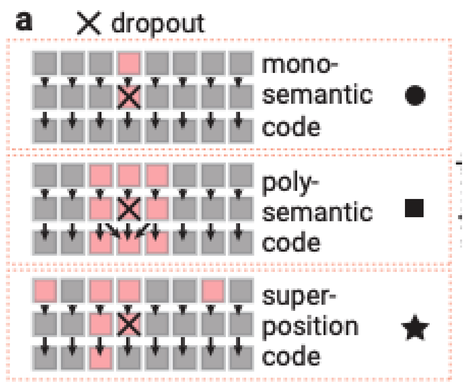
It hallucinates on out of domain data

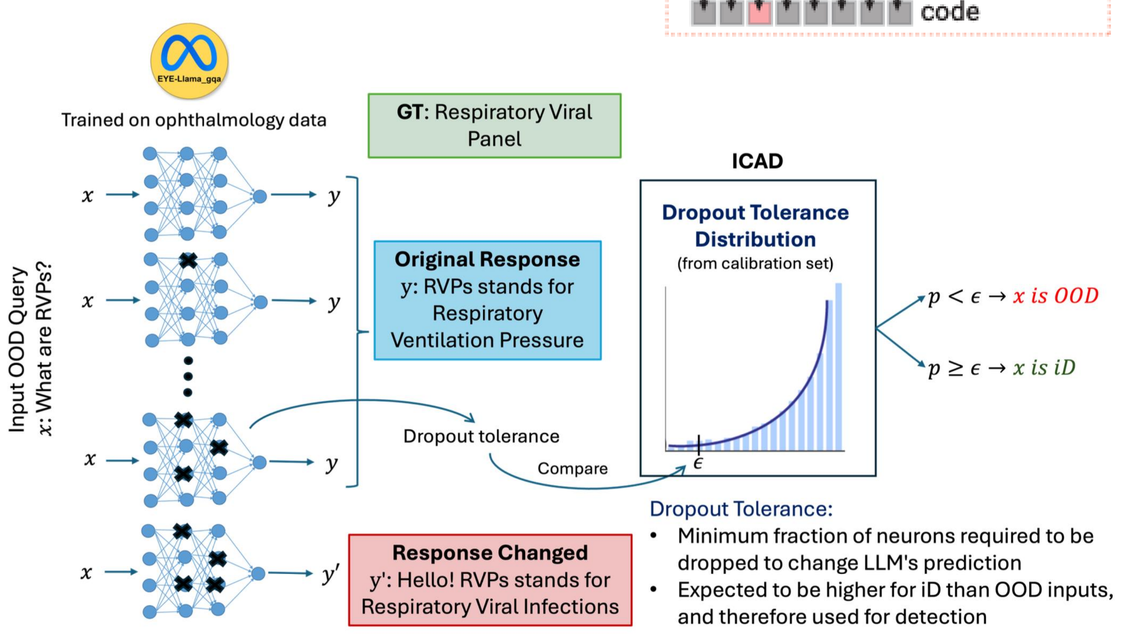
Dangerous in critical applications like healthcare!

OUR IDEA: Detecting OOD inputs

Background: LLMs are *Polysemantic*, thus, somewhat robust to dropout

- Is this 'level of robustness' an indication of confidence?
- **Can we exploit this?**





Measure LLM confidence through 'iterative dropout'

Ask same question repeatedly, each time introducing more dropout.

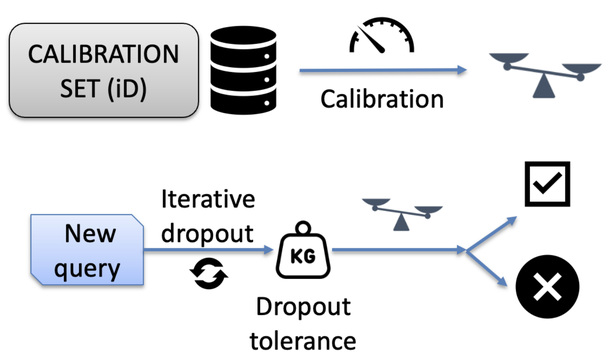
- **If response changes easily, LLM is not confident. Probably OOD!**
- **If response is robust, LLM is confident. Probably iD!**

ICAD: INDUCTIVE CONFORMAL ANOMALY DETECTION

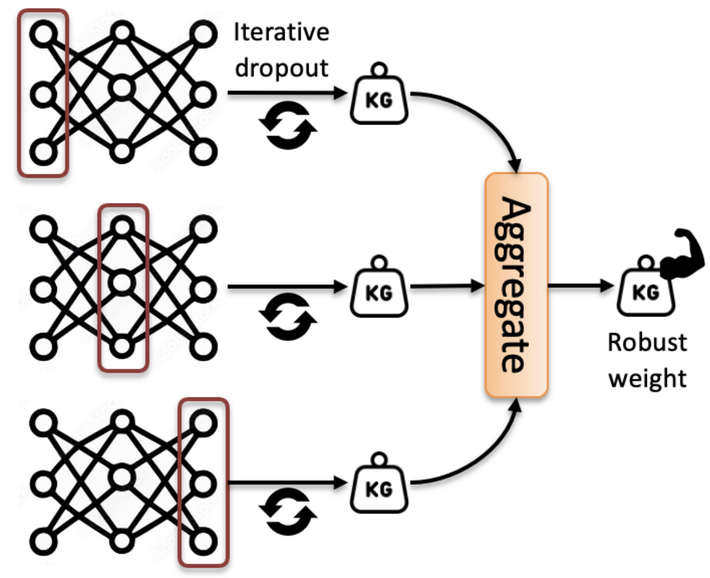
Mathematical framework for anomaly detection, theoretical bounds on false alarm rates

1. Use an in-domain query set to calibrate the scales of ICAD
2. Use calibrated ICAD to detect anomalies (OOD queries)

NO TRAINING NEEDED



ENSEMBLING



Iterative dropout across one layer gives *one* measurement for the non-conformity of the query

To reduce noise and get a more robust prediction, we aggregate the measurements across different layers

Aggregation method should be 'valid' to maintain theoretical bounds on False Alarms!

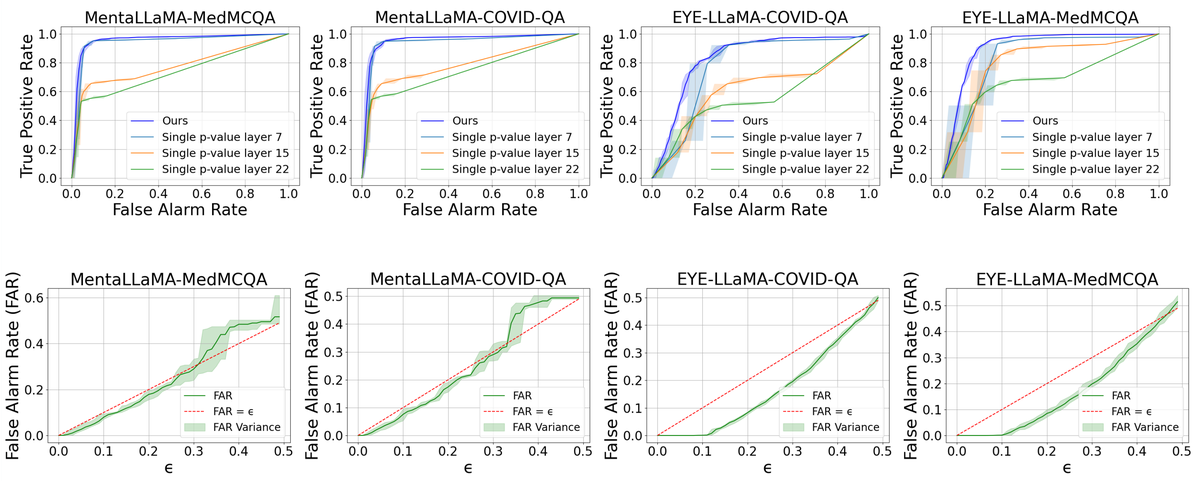
Use *valid merging functions* (Vovk and Wang, 2020) → Arithmetic Mean, Geometric Mean etc.

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

RESULTS

Across multiple LLMs and datasets, our method can detect OOD queries quite well.
All while having a user-defined bounded false alarm rate!

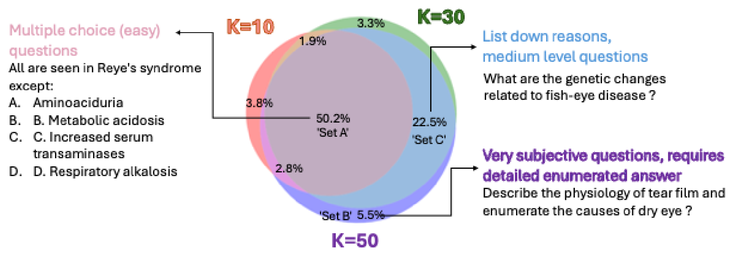
Model	EYE-LLaMA		MentaLLaMA	
OOD Dataset	**CovidQA**	**MedMCQA**	**CovidQA**	**MedMCQA**
Base Score Method with Layer 7	0.53	0.54	0.73	0.72
Base Score Method with Layer 15	0.48	0.58	0.71	0.69
Base Score Method with Layer 22	0.48	0.57	0.70	0.70
Single p-value Method with Layer 7	0.77	0.83	0.93	0.94
Single 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**



ANALYSIS AND ABLATIONS

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**

Different *valid* merging functions



Difficulty of queries is related to the ease of breaking the response!

CONTACT

Ayush Gupta
Final Year PhD student
Johns Hopkins University

Open for

- **Spring 2026 internships**
- **Full-time roles starting Summer 2026**