Data-Efficient Meta-Models for Evaluation of Context-based Questions and Answers in LLMs

Julia Belikova[1,2] Konstantin Polev[1] Rauf Parchiev[1] Dmitry Simakov[1]

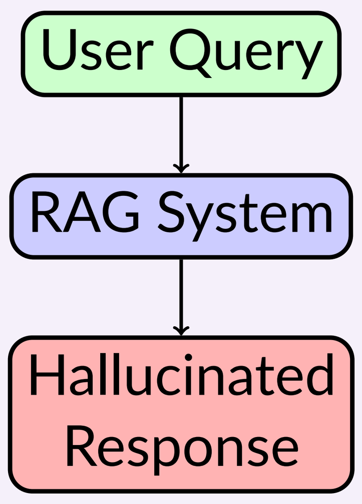
[1]Sber AI Lab [2]Moscow Institute of Physics and Technology

Problem

Background: RAG systems are widely deployed in production & **contextual hallucinations** in such systems undermine user trust

Industrial Deployment Constraints:

1. Limited Annotated Data
   - Domain-specific annotation is expensive ($$$)
   - Time-consuming manual labeling process
   - Expertise requirements vary by domain
2. Computational Efficiency
   - Proprietary LLMs → prohibitive latency & costs
   - Real-time deployment requirements
   - Scalability concerns for high-volume applications
3. Privacy & Data Sovereignty
   - Sensitive enterprise data cannot leave premises
   - Regulatory compliance (GDPR, HIPAA)
   - Need for local, open-source solutions



Current SoTA local hallucination detectors, such as attention-based methods [1] and internal probing techniques [2], have proved to be effective on academic benchmarks with **extensive annotated training sets** that are infeasible for industrial settings.

Research Gap: Bridge between academic benchmarks and feasible industrial solutions in hallucination detection

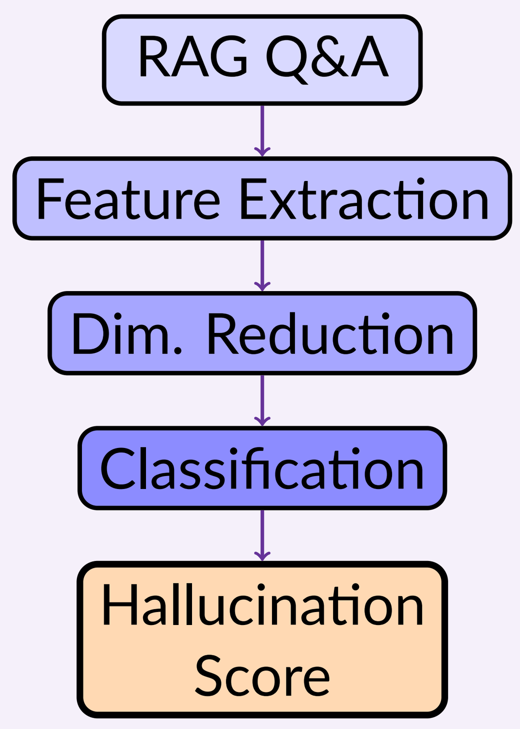
Contributions

- **A data-efficient meta-modeling framework for hallucination detection** that achieves competitive performance with SoTA baselines using as few as 250 training samples, significantly lowering annotation costs.
- **The first rigorous empirical validation of the SoTA tabular classifier** TabPFNv2 [3] for hallucination detection that demonstrates superiority in data-scarce scenarios across multiple RAG benchmarks.
- **Demonstration of the industrial viability** of using probing with smaller, open-source LLMs as feature extractors, offering a private, cost-effective, and scalable alternative to proprietary model-based evaluators.

Framework

- Small proxy LLMs as internals extractors
- Multi-strategy feature extraction
- Dimensionality reduction
- Efficient tabular classifiers

Goal: Combine efficient classification with effective feature extraction to minimize train size while preserve high performance



Methodology

Let D = dataset, C = context, R = LLM response, S = model states

(1) Hidden States Feature Extraction:

$$p_{mean} = \frac{1}{|R|}\sum_{i=1}^{|R|} h_i, \quad p_{max} = \max(h_1, \dots, h_{|R|}), \quad p_{last} = h_{|R|}$$

where $h_i \in \mathbb{R}^{d_{model}}$ are hidden states of response tokens R.

(2) Attention-based Feature Extraction (Lookback Lens):

$$p_{lookback}^{(l,h)} = \frac{1}{|R|}\sum_{r \in R} \frac{\sum_{c \in C} A^{(l,h)}[r,c]}{\sum_{t=1}^{|T|} A^{(l,h)}[r,t]}$$

where $A^{(l,h)} \in \mathbb{R}^{|T|\times|T|}$ is attention matrix for layer l and head h.

(3) Dimensionality Reduction:

$$p^{red} = \text{PCA}(p^{orig}, d = 30) \quad \text{or} \quad p^{red} = \text{UMAP}(p^{orig}, d = 30)$$

(4) Meta-Classification:

$$f_{hall}(C, R, S; \phi) = g_\phi(p(C, R, S))$$

where $g_\phi \in \{\text{LogReg}, \text{CatBoost}, \textbf{TabPFNv2}\}$

Experimental Setup

Datasets:

- **EManual:** Enterprise manual Q&A[4]
- **ExpertQA:** Expert-level questions[5]
- **RAGTruth (QA):** RAG benchmark[6]

Baselines:

- LLM-as-Judge (GPT-4o)
- RAGAS-Faithfulness[7] (GPT-4o)
- Attention-Pooling Probe[2]

Models:

- **Generator:** GPT-3.5-turbo
- **Extractors:** Gemma-2-9b-it, Llama-3.1-8B, Qwen2.5-7B

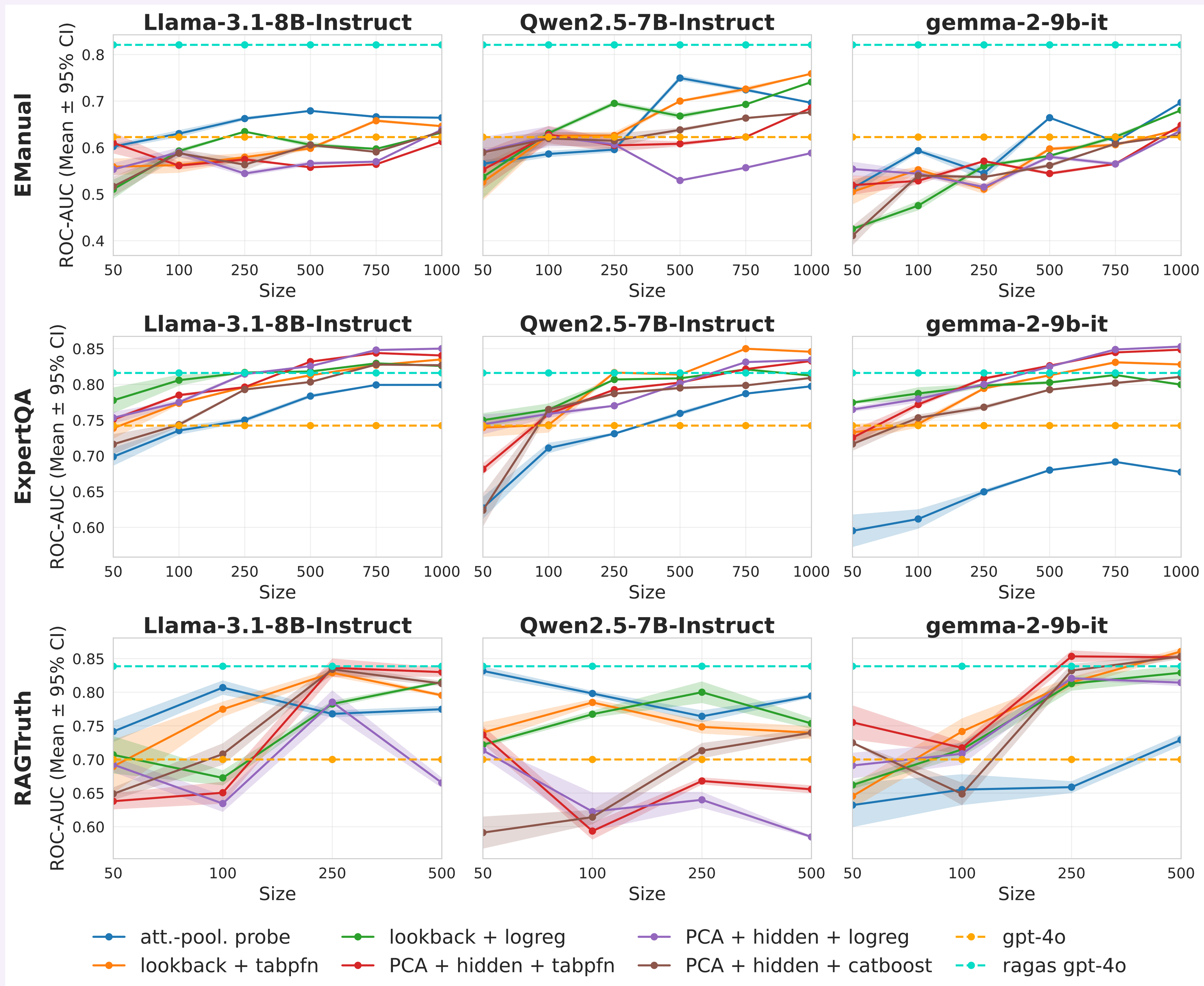
Evaluation:

- Data scarcity: **50–1000 samples**
- Metric: **ROC-AUC**
- 5-fold cross-validation

*to obtain LLM internals, we use local proxy models – extractors

Results

Method Performance with Limited Data



• Competitiveness to SOTA • Data Efficiency • Model Flexibility

Classifier Performance with Limited Data

Classifier	EManual	ExpertQA	RAGTruth	Average
TabPFNv2	0.7161	0.8204	0.8139	0.7834
LogReg	0.6896	0.8218	0.8087	0.7734
CatBoost	0.6832	0.7932	0.8176	0.7646
Att.-pool. probe	0.6776	0.7611	0.8002	0.7463

• Best performance across a variety of size and feature combinations

Key Insights

Technical Insights

- **TabPFNv2 Effectiveness:** Consistent #1 ranking across datasets and extractors
- **Small LLMs as extractors:** Match/outperform proprietary evaluators at lower cost
- **Data Efficiency Threshold:** Performance plateau at **250 samples**

Practical Benefits:

- **90% reduction** in annotation requirements
- Local deployment feasible
- Real-time processing capability
- Cost-effective scaling

Deployment Recommendation: Start with 250 training samples with TabPFNv2

References

[1] Yung-Sung Chuang et al. "Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps". In: *arXiv preprint arXiv:2407.07071* (2024).
[2] S. CH-Wang et al. "Do Androids Know They're Only Dreaming of Electric Sheep?" In: *arXiv preprint arXiv:2312.17249* (2024).
[3] N. Hollmann et al. "Accurate Predictions on Small Data with a Tabular Foundation Model". In: *Nature* 637.8045 (2025), pp. 319–326.
[4] A. Nandy et al. "Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework". In: *arXiv preprint arXiv:2109.05897* (2021).
[5] C. Malaviya et al. "ExpertQA: Expert-Curated Questions and Attributed Answers". In: *arXiv preprint arXiv:2309.07852* (2024).
[6] C. Niu et al. "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models". In: *arXiv preprint arXiv:2401.00396* (2024).
[7] Explodinggradients. *RAGAS: Evaluation Framework for Retrieval Augmented Generation*. https://github.com/explodinggradients/ragas. 2024.