LLM-Based Automated Hallucination Detection in Multilingual Customer Service RAG Applications

Nikilkumar Patel – ex Research Assistant at IADS, University Essex

Haralambos Mouratidis – Director of IADS and Professor at School of CSEE, University of Essex

Kenneth Ng Kai Zhi – Data Scientist at Algomo

#217 - Session 7 - AIAI 2025 - 21st International Conference on Artificial Intelligence Applications and Innovations
Cyprus University of Technology, Limassol, Cyprus
Hybrid @ Limassol - Cyprus & Web

AIAI ARTIFICIAL INTELLIGENCE APPLICATIONS & INNOVATIONS

IADS | The Institute for Analytics and Data Science

University of Essex

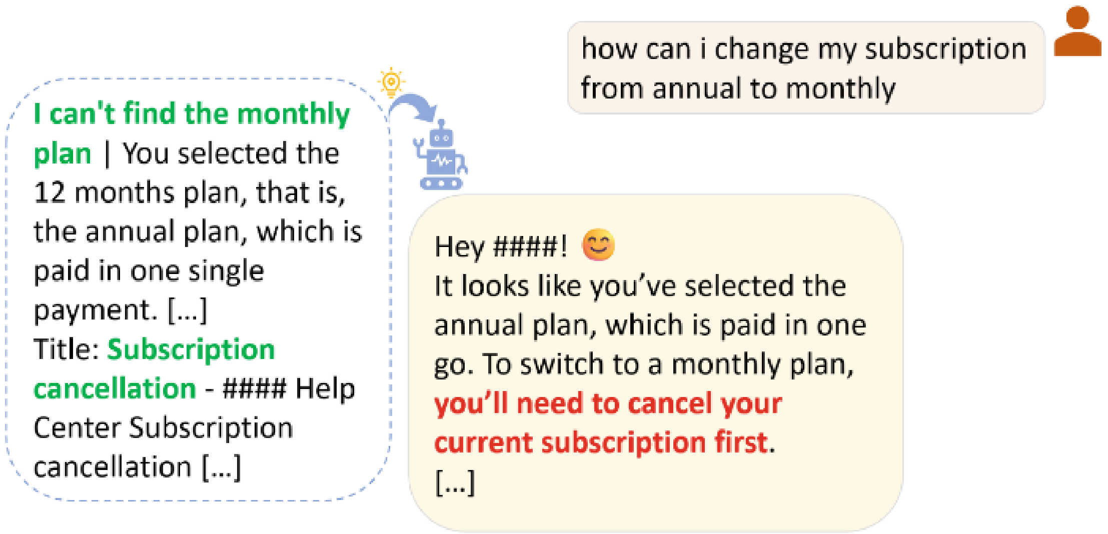
UKRI Innovate UK

Project Background

- A collaboration between academic institution University of Essex and a company Algomo limited
- A Part of Innovate UK sponsored project- “Addressing Hallucinations in Generative AI for customer service applications”
- Project timeline: Mar2024 - Feb2025 (Pre-reasoning, Pre-Agantic)
- Aim: Investigating & addressing challenges of LLM hallucinations involving:
  - ➢RAG
  - ➢Closed Large Language Models(LLM)
  - ➢Multilingual Customer Service Q&A
- Approach: To increase trustworthiness in LLMs in production by automatic, economic and real-time hallucination detection

Examples

Hallucination in Algomo's Customer Service chatbot on production*



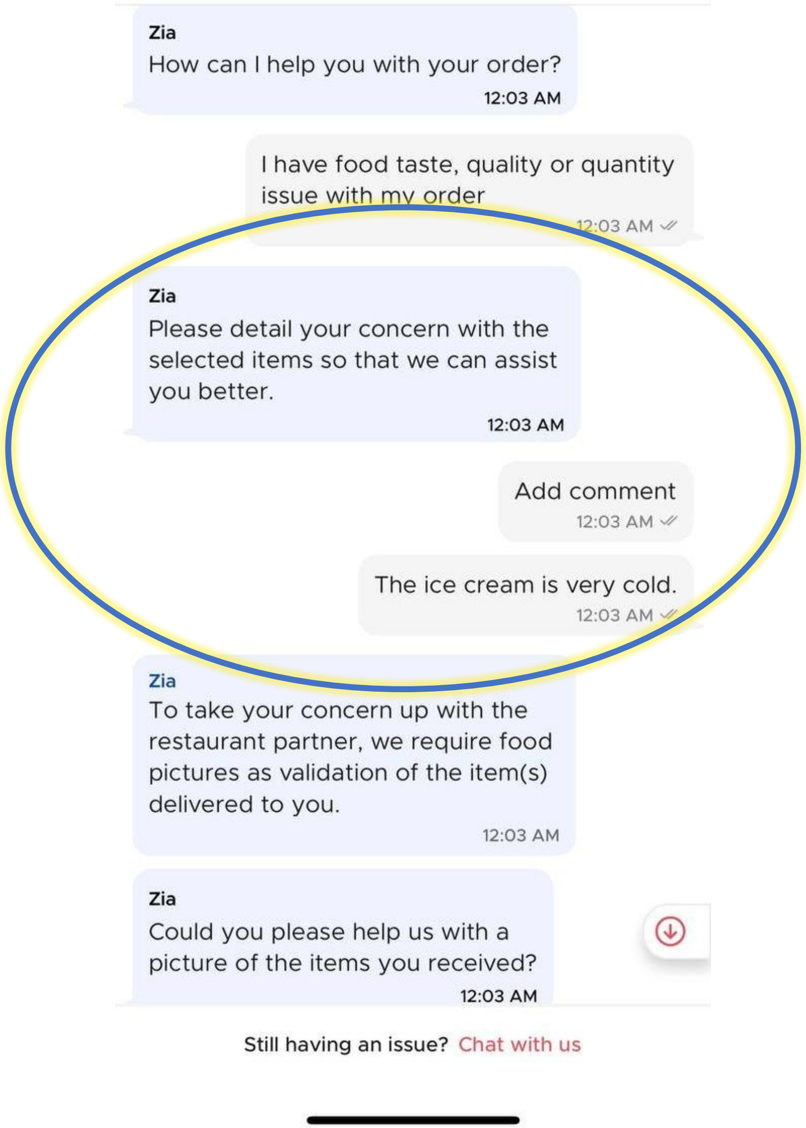
* #### = masked sensitive information due to Client and Customer Data Privacy

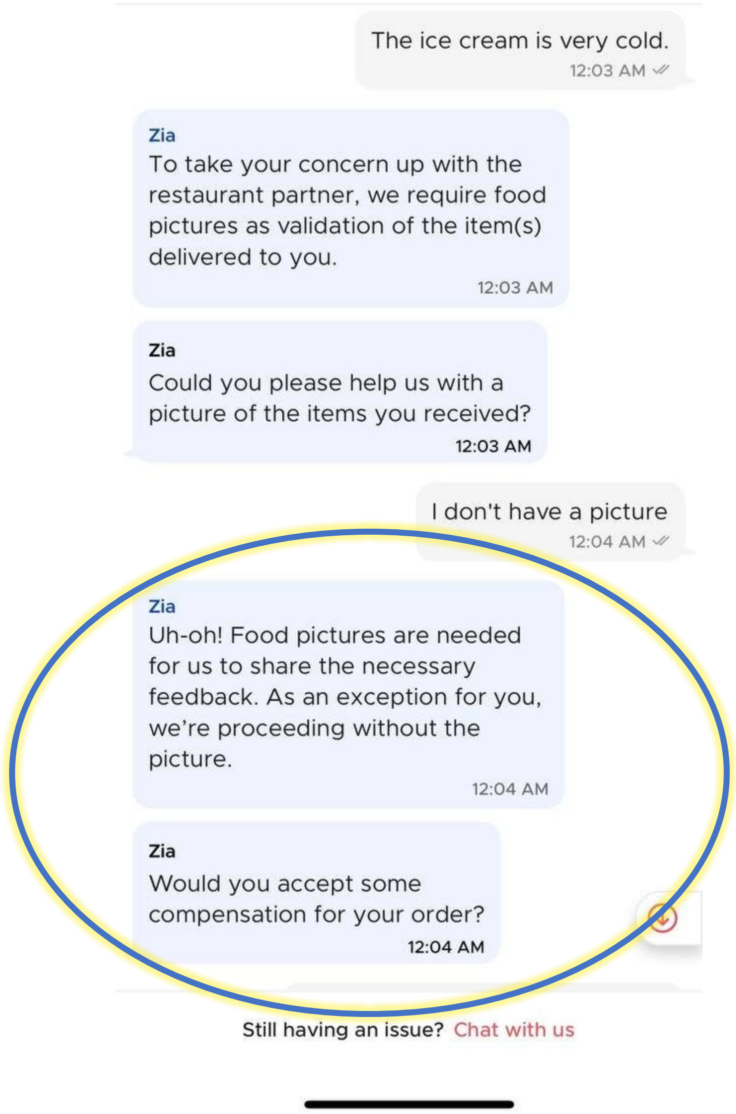
Examples…

Hallucinations in someone else's production*

Source: https://x.com/ri5hitripathi/status/1932724310969626741

*Not included in our work





Literature Review and Key Observations

Conversational Hallucinations are mainly attributed to following reasons:

- <u>Situated Unfaithfulness</u> - Overreliance on contextual information in RAG
- <u>Context bias</u> - Susceptibility to noise/faults in context; for more than 50% cases the answer reflects the wrong contextual knowledge despite knowing the correct information without the context
- <u>Exposure bias</u> - General-Purpose model (non- finetuned) on uncommon languages/slangs and deep, long-tailed data
- Prompt sensitivity, similarity-based retrieval and software design constraints introduce inconsistency in LLM Pipelines – Little/No reproducibility for similar type of queries
- <u>Emerging capabilities and unsafe behaviour</u> - In-context scheming and user deception

Existing Automatic Detection Methods/Metrics

- All are using LLM as a judge – Binary Classification – Hallucination +ve or -ve

- Types:

1. Natural Language Inference(NLI) Based
2. Prompt Based

- Selection for Evaluation Criteria:

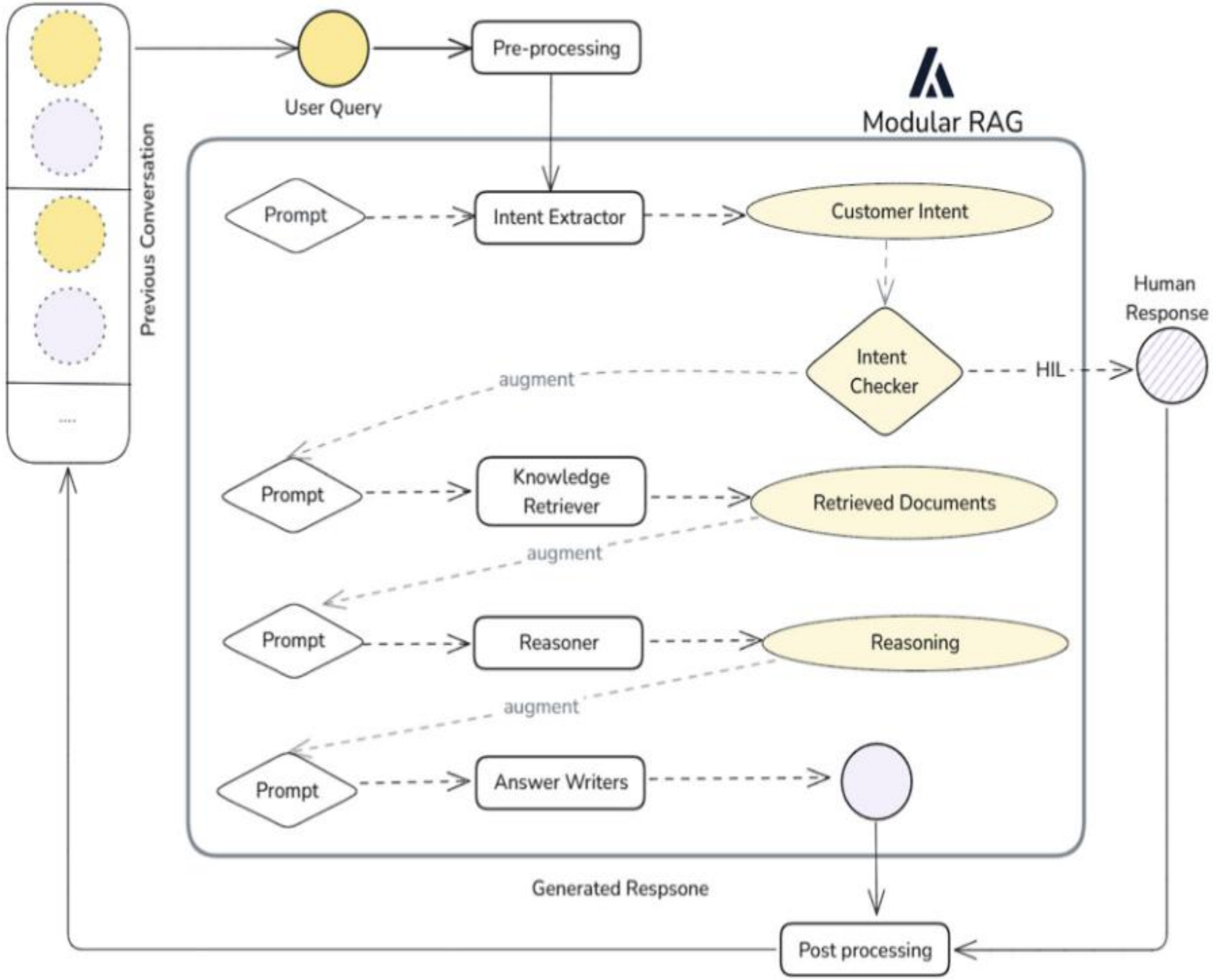
1. Use case compatibility
2. Economic
3. Real-time

Method	RAG-support	Multilingual	Type
RefChecker	Yes	Yes	NLI
~~CLUE~~	~~No~~	~~No~~	~~NLI~~
ChainPoll	Yes	Yes	Prompt
~~SelfCheckGPT~~	~~No~~	~~Yes~~	~~Prompt~~
Answer Faithfulness	Yes	Yes	Prompt
~~G-Eval~~	~~No~~	~~Yes~~	~~Prompt~~
~~ARES~~	~~Yes~~	~~No~~	~~Prompt~~
~~AutoHall~~	~~Yes~~	~~Yes~~	~~Prompt~~
~~SAC3~~	~~No~~	~~Yes~~	~~Prompt~~
~~ReID~~	~~Yes~~	~~No~~	~~Prompt~~
DeepEval Hallucination	Yes	Yes	Prompt

Experimental Setup

Algomo's AI AutomationWorkflow in Prod.:

- Modular RAG:
  1. Intent extractor
  2. Knowledge Retriever (Vector DB)
  3. Reasoner (Planning)
  4. Answer Writers
- Condition based Human escalation as "safe exit" for unknown/unfamiliar/custom topic
- Conversation level memory



Experimental Setup…

LLMs used in AI Automation Workflow:

- High cognitive tasks(Planner and Intent Checker): OpenAI’s gpt4o
- Less cognitive tasks(Answer Writer): OpenAI’s gpt4o-mini
- Embedding: OpenAI’s text-embedding-ada-002
- Hallucination Detection Methods/Metrics: Llama3-8b(Lynx), OpenAI’s gpt4o-mini, Anthropic’s Claude Haiku3.5

Experiment Methodology:

1. Real time LLM monitoring and data collection → 2. Label data with pre-defined hallucination criteria → 3. Run evaluation experiments with selected methods → 4. Data Analysis and Reflection

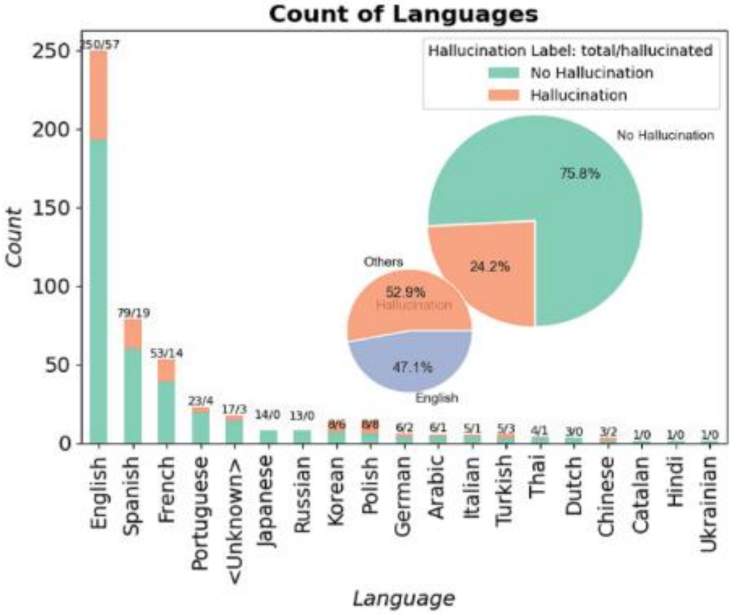
Experimental Setup…

Labelling conditions for Hallucination positive label for a <u>conversation</u> if any of its generated responses:

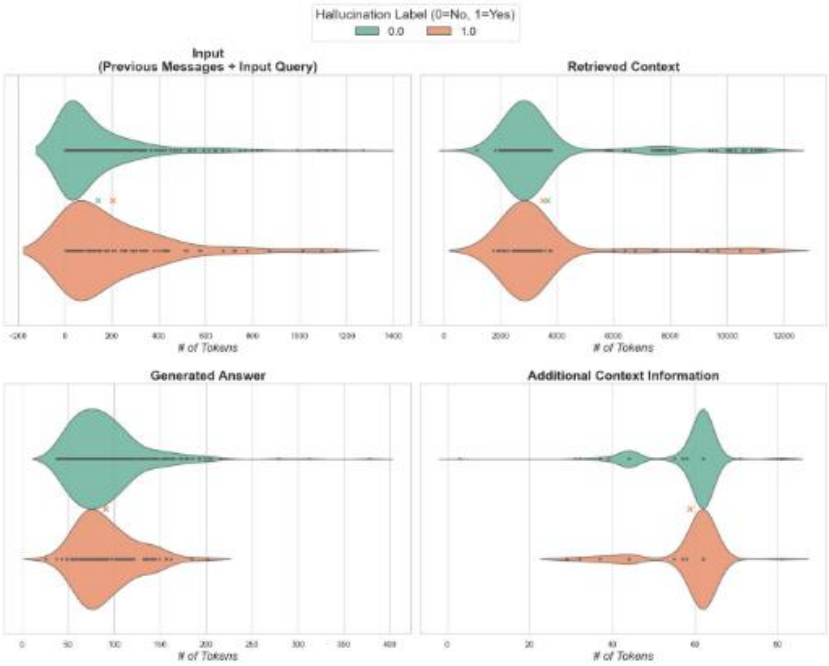
1. does not follow the prompt instructions - *Common*
2. fails to understand the user's intention – *Intent module*
3. contains repeated answer referring to previous messages multiple times. – *Reasoner module*
4. has claims that are not supported by retrieved contexts. - *Reasoner & Writer module*
5. contains any invented entities such as URLs, numbers, currencies not present in the retrieved context – *Writer module*
6. is in different language than input. – *Writer module*

Evaluation Data…

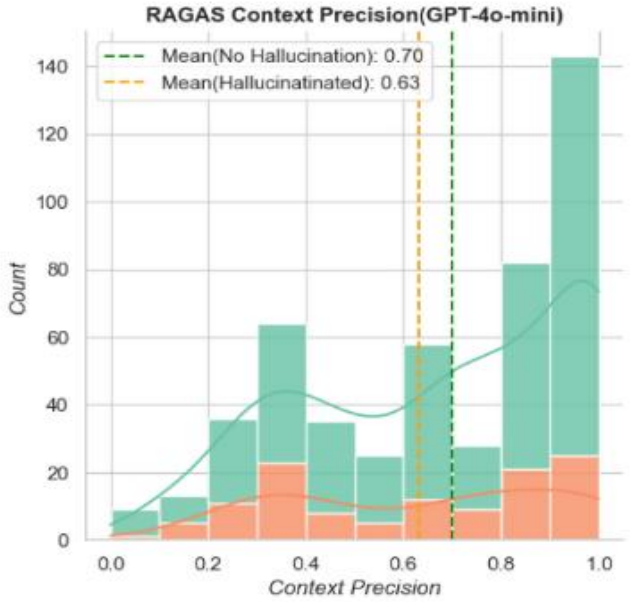
- Labelled Conversations : 500
- 250 English
- 250 all other languages
- Total Hallucination : ~**24**%



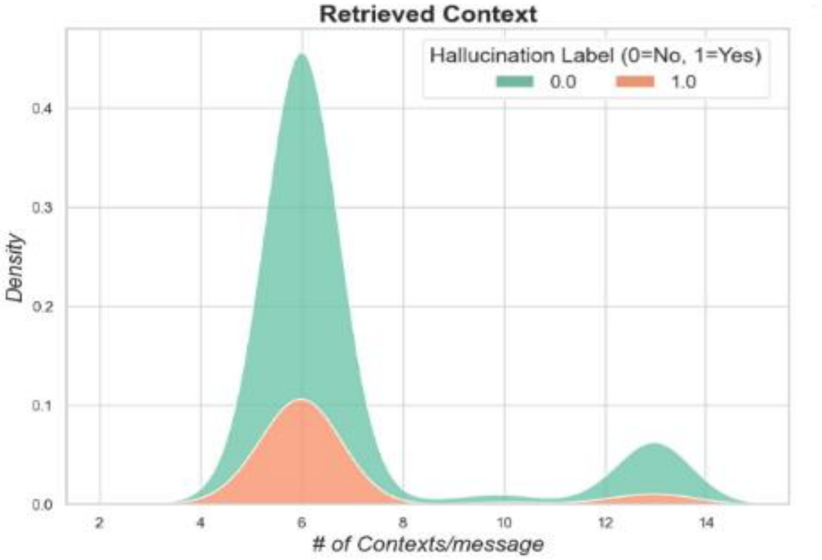
(a) Languages and Hallucination%



(b) Tiktoken Token distributions



(c) Mean Context Precision



(d) KDE of Retrieved context/message

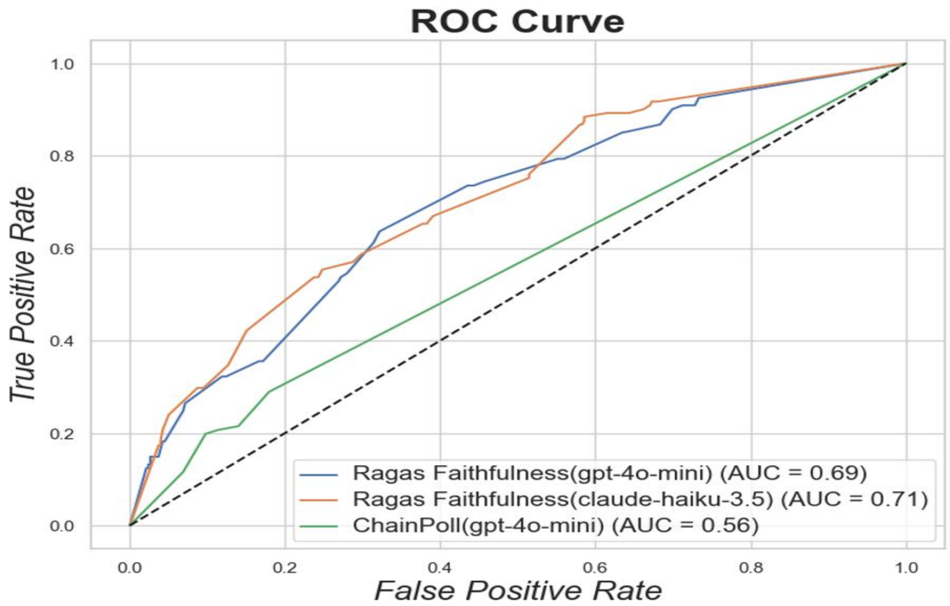
No of Messages/Conversation

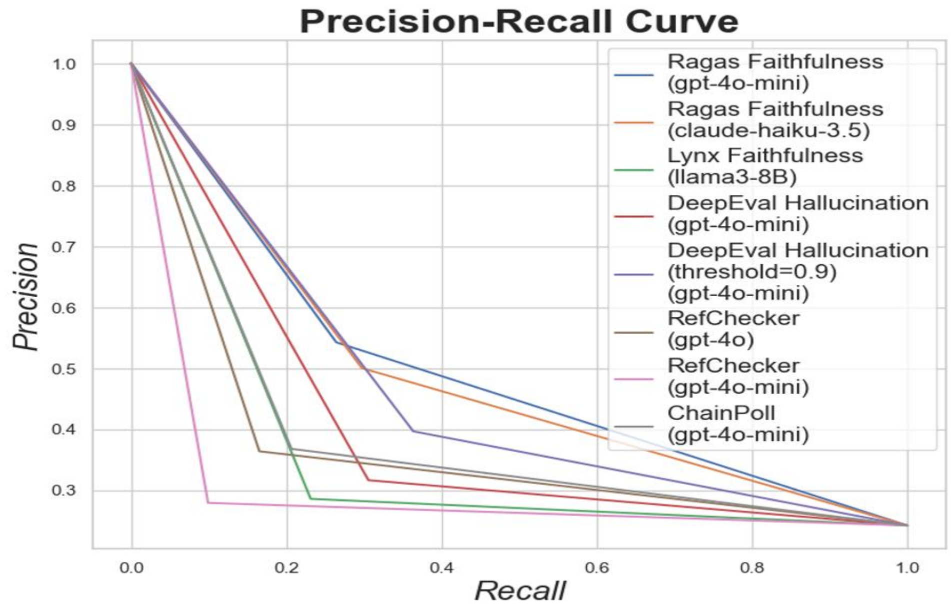
- Mean: ~3
- Min: 0 | Max: 28
- 25%:0 | 50%: 2 |75%: 4

Evaluation Results

Method	Judge-LLM	θ	Accuracy	Precision	Recall	F1
★ RefChecker	gpt-4o	-	0.73	0.36	0.17	0.23
RefChecker	gpt-4o-mini	-	0.72	0.28	0.10	0.15
Lynx Faithfulness	llama3-8B	-	0.67	0.29	0.23	0.26
ChainPoll	gpt-4o-mini	0.56	0.72	0.37	0.21	0.26
★ RAGAs Faithfulness	gpt-4o-mini	0.69	**0.77**	**0.54**	0.26	0.36
RAGAs Faithfulness	claude-haiku3.5	0.71	0.76	0.50	0.30	0.37
DeepEval Hallucination	gpt-4o-mini	0.50	0.67	0.32	0.31	0.31
DeepEval Hallucination	gpt-4o-mini	0.90	0.71	0.40	**0.36**	**0.38**

★ Best Ensemble
F1 Score : **0.439**



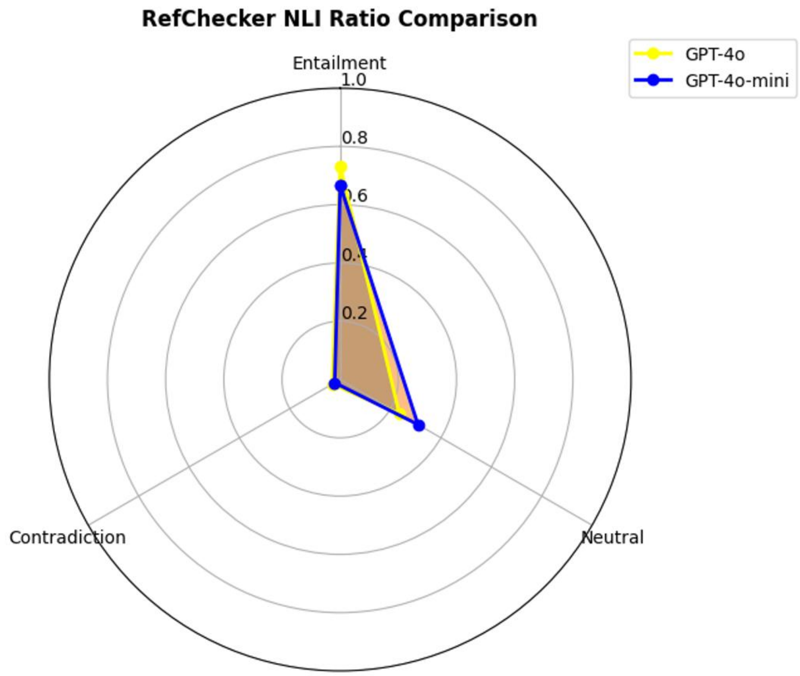


Discussion

- *67.77%* hallucinations occurred in planning (Reasoner) stage
- Problems noticed with using direct prompts with LLM as judge:
  - ➢Prone to self-hallucinations
  - ➢Prompt sensitivity of differentt models
- Problems noticed with Nature Language Inference(NLI):
  - ➢NLI LMs lack multilingual capabilities
  - ➢Inherent bias in NLI methods

Mean NLI Label Ratio assigned by Checker LLMs in RefChecker

Checker	Entailment%	Neutral%	Contradiction%
gpt-4o-mini	0.667	0.308	**0.022**
gpt-4o	0.732	0.235	**0.031**



Discussion…

Faithfulnesss(RAGAs) vs DeepEval Hallucination

$$\text{Faithfulness Score} = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}}$$

- Focuses on extracting claims from generated answers that are supported by context
- Problems noticed :
  - Derivative Compliance- Inferred vs Explicit mentions-causing False Positives
  - Superficial Compliance- Referring to the wrong context- causing False Negatives.

$$\text{Hallucination} = \frac{\text{Number of Contradicted Contexts}}{\text{Total Number of Contexts}}$$

- Focuses on finding contradicted context only
- Favourable attributes over Faithfulness and NLI:
  - Comes with tuning parameter to control the strictness of judging
  - Doesn’t extract “claims” and may prevents itself from self hallucinations

Limitations

1. Impact is unknown for advanced RAG methodologies: Graph RAG, CAG
2. Limited evaluation data due to resource constraints
3. Other open and proprietary models as judges

Key Takeaways

1. Continuous and granular LLM monitoring and evaluations (AI Observability in general) are required for operating trustworthy real-world AI applications.
2. Companies are advised to have their own evaluation data in addition to the standard evaluations sets
3. Due to complexity of real-world use cases, we require more robust evaluation methods and datasets.
4. Hallucination is a common and still an open-ended problem but its mitigation may be specific to the model, nature of problem and the use case.

Thank you for your time

Questions?