ACSAC 2025

Anti-Knowledge Corruption: Document Augmentation Defense

Yizhu Wen[1], Xun Chen[2], Igor Molybog[1], Hanqing Guo[1]
University of Hawaii at Manoa[1]

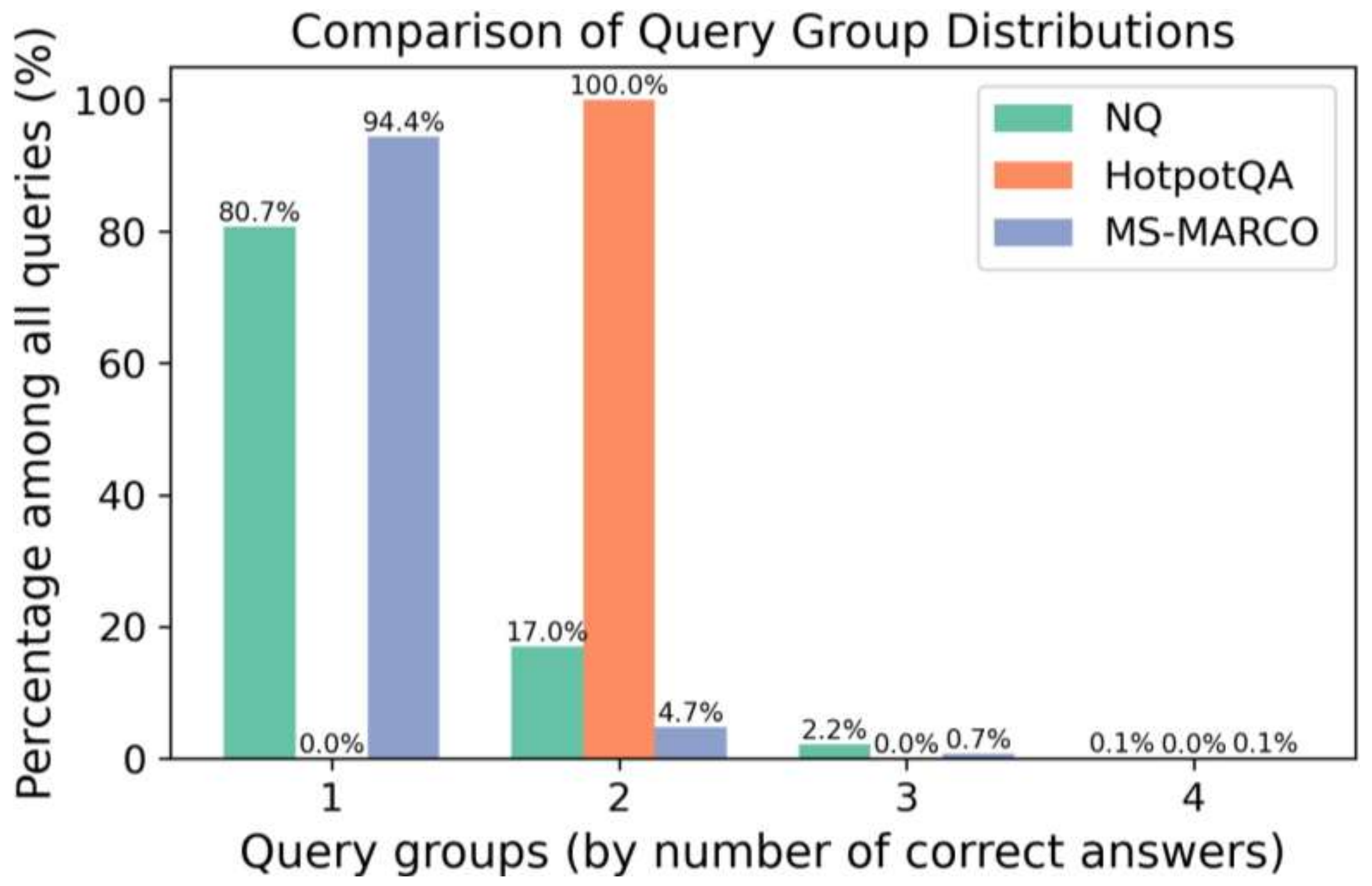
Contact Email: {yizhuw, guohanqi, molybog}@hawaii.edu[1], xunchen@outlook.com[2]

Problem:

RAG: LLMs that retrieve external documents for grounded answers.

Knowledge Corruption Attack: attacker inject fake texts into databases mislead RAG outputs.

Our Finding: Past attacks succeed mainly because the **fake data overwhelms real data**!



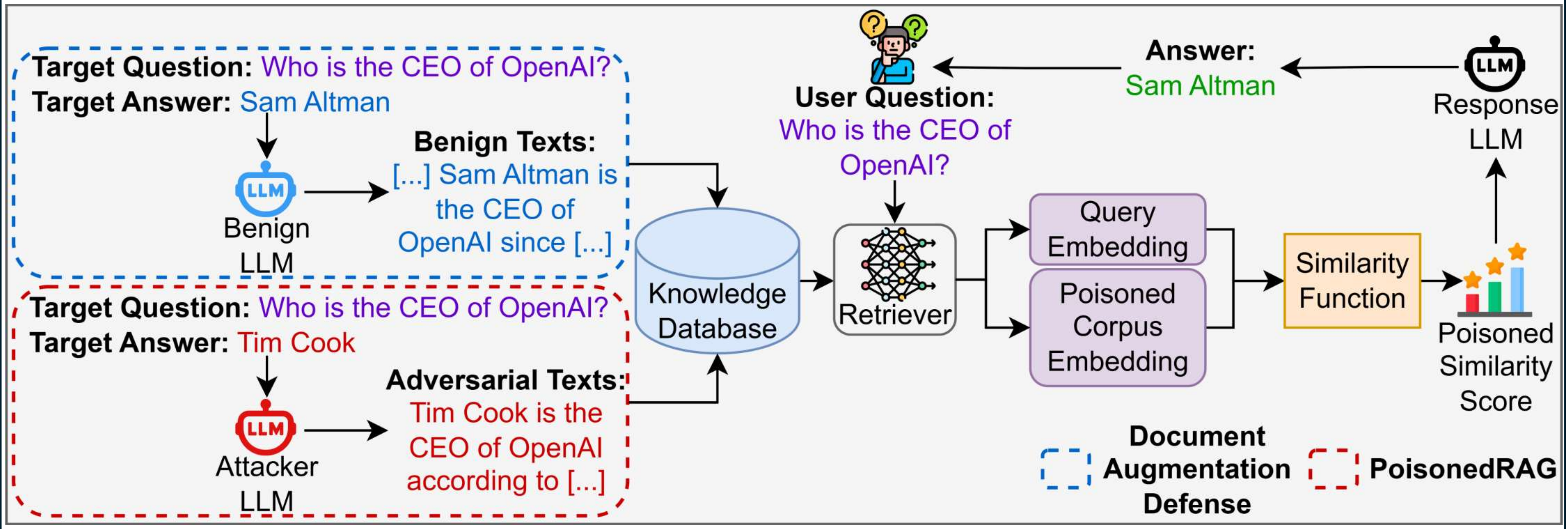
Only **1** or **2** correct data per document!

But attacker inject 5 fake data to run the attack!

Hypothesis:

H0: The **attack success rate (ASR) is the same** regardless of **benign-to-fake ratio.**

Framework Overview:


Target Question: Who is the CEO of OpenAI?
Target Answer: Sam Altman
Benign LLM
Benign Texts:
[...] Sam Altman is the CEO of OpenAI since [...]
Target Question: Who is the CEO of OpenAI?
Target Answer: Tim Cook
Attacker LLM
Adversarial Texts:
Tim Cook is the CEO of OpenAI according to [...]
Knowledge Database
User Question:
Who is the CEO of OpenAI?
Retriever
Query Embedding
Poisoned Corpus Embedding
Similarity Function
Poisoned Similarity Score
Response LLM
Answer:
Sam Altman
Document Augmentation Defense
PoisonedRAG

- Knowledge corruption attack in the red box inject LLM-crafted query relevant fake corpus into the database;
- Document augmentation defense in the blue box injects LLM-crafted query relevant benign corpus into the database;
- The RAG system gives correct answers under the knowledge corruption attack!
- It can **generalize to any existing RAG architecture** with minimal effort.

Experiment:

With 5 adversarial corpus per query, we add 5 query relevant benign corpus and track ASR. On Natural Questions dataset, **ASR drops by half** from 98% and 83% to 31% and 42%, indicating attack dilution.

Top k	Adv / query	Benign / query	ASR mean	Precision Mean	Recall Mean
5	1	0	65.00%	20.00%	99.00%
5	3	0	84.00%	59.00%	98.00%
5	5	0	98.00%	96.00%	96.00%
5	**5**	**5**	**31.00%**	**42.00%**	**42.00%**
10	5	0	83.00%	50.00%	99.00%
10	**5**	**5**	**42.00%**	**49.00%**	**98.00%**

Conclusion:

The results reject H0. More benign, query relevant data sharply reduces ASR. **The attack is much weaker in practice**. This highlights document augmentation defense as a simple, effective way to protect RAG systems.