Real or *Hallucinated*? Improving the Truthfulness of LLMs

Authors: Salil Fernandes, Shruti Pareshbhai Gandhi, Xenus Gonsalves,
Vedant Hareshbhai Patel, Reuben Suju Varghese

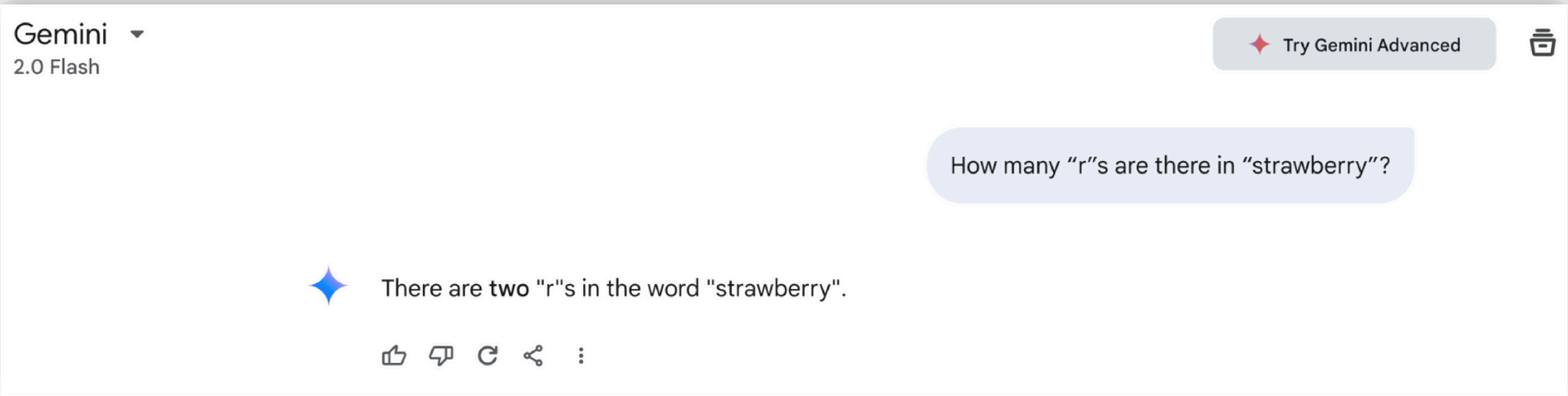
~ Group 40

USC Viterbi
School of Engineering

University of Southern California

Introduction

As of April 14, 2025,



So,

- **How can we accurately detect hallucinations in LLM outputs?**
- **Can we reduce hallucinations consistently across different models?**

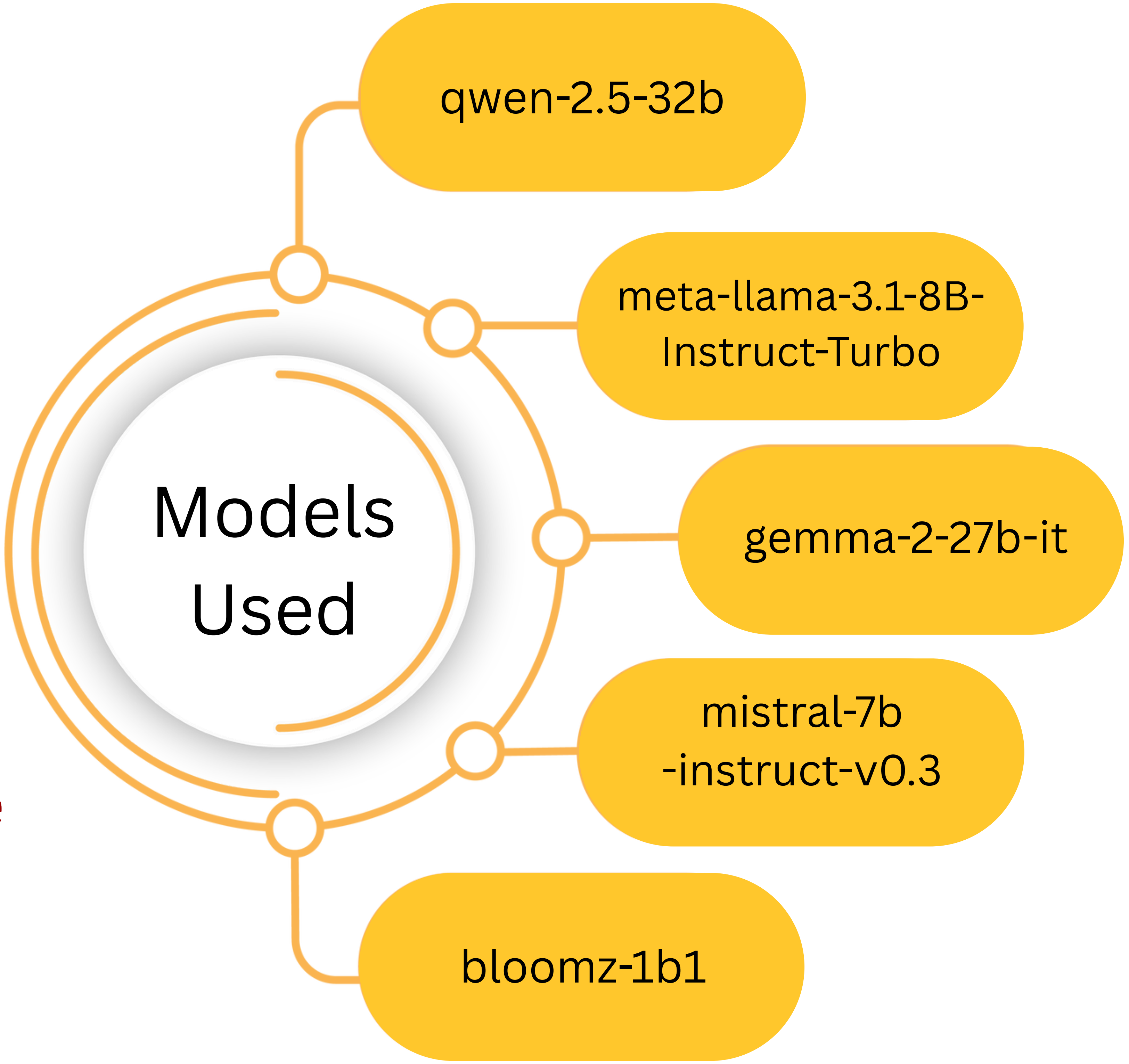
Dataset Preparation

question

right_answer

Responses from 5 LLMs

Binary is_hallucinated labels for each response



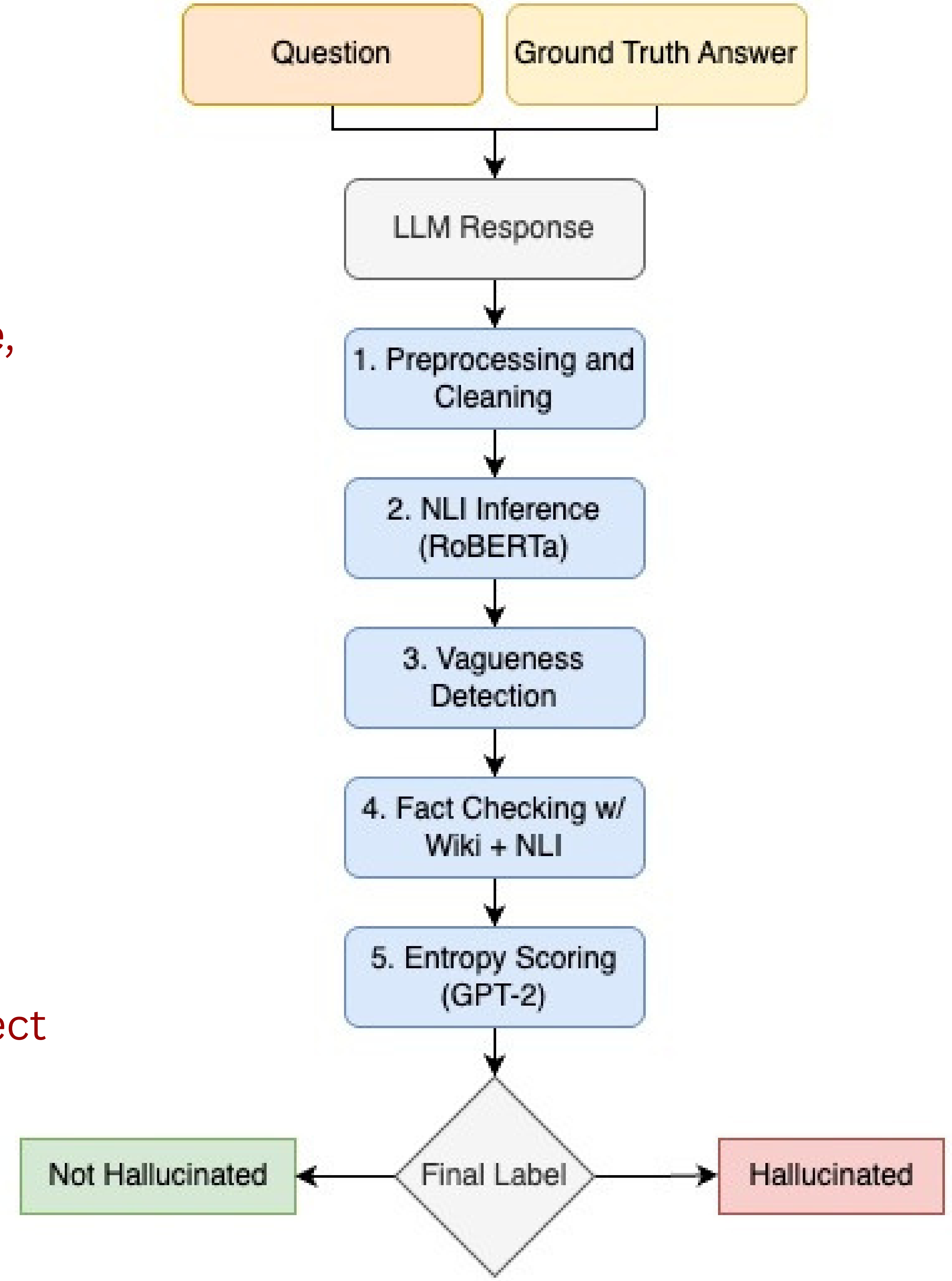
Hallucination Detection: Core Pipeline

Stage 1: Preprocessing
LLM responses are cleaned to remove markdown, disclaimers, and boilerplate, ensuring meaningful input for NLI.

Stage 2: NLI (RoBERTa-large-MNLI)
We form evidence-claim pairs and classify them using NLI. High-confidence contradictions are flagged as hallucinations.

Stage 3: Vagueness Detection
Responses with evasive phrases (e.g., “I don’t know”) are marked as hallucinated, overriding NLI if needed.

Stage 4: NLI + Wikipedia Fact-Check
Misclassified cases are rechecked using Wikipedia summaries and NLI to correct misclassified labels (FP + FN).



Stage 5: Entropy-Based Uncertainty Estimation

LLM responses through GPT-2, computing token-level probability distributions. A higher average entropy indicates the model was less confident in its response — a strong signal of hallucination.

Impact of Fact Checking & Entropy

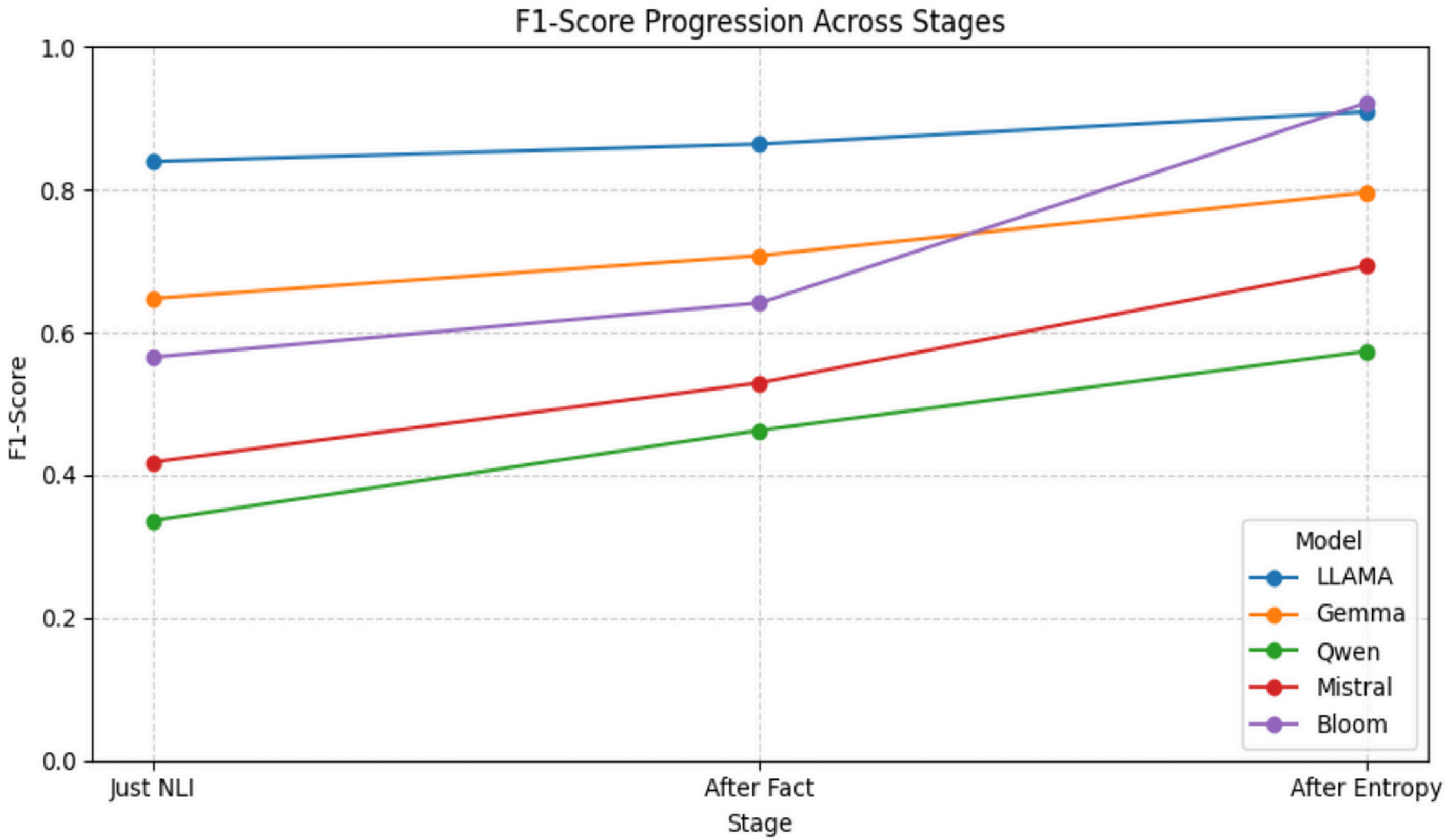
Model	False Positives	False Negatives
LLAMA	1055	1078
Gemma	449	2250
Qwen	1168	1940
Mistral	1825	1561
Bloom	1658	3004

Post NLI

Model	False Positives	False Negatives
LLAMA	534	665
Gemma	98	1537
Qwen	143	1573
Mistral	271	1161
Bloom	918	93

Post FC & Entropy

Combining NLI, heuristics, and entropy provides a multi -dimensional safety net for hallucination detection. This hybrid approach generalizes well across all 5 models tested showing an average improvement of 20-50%.

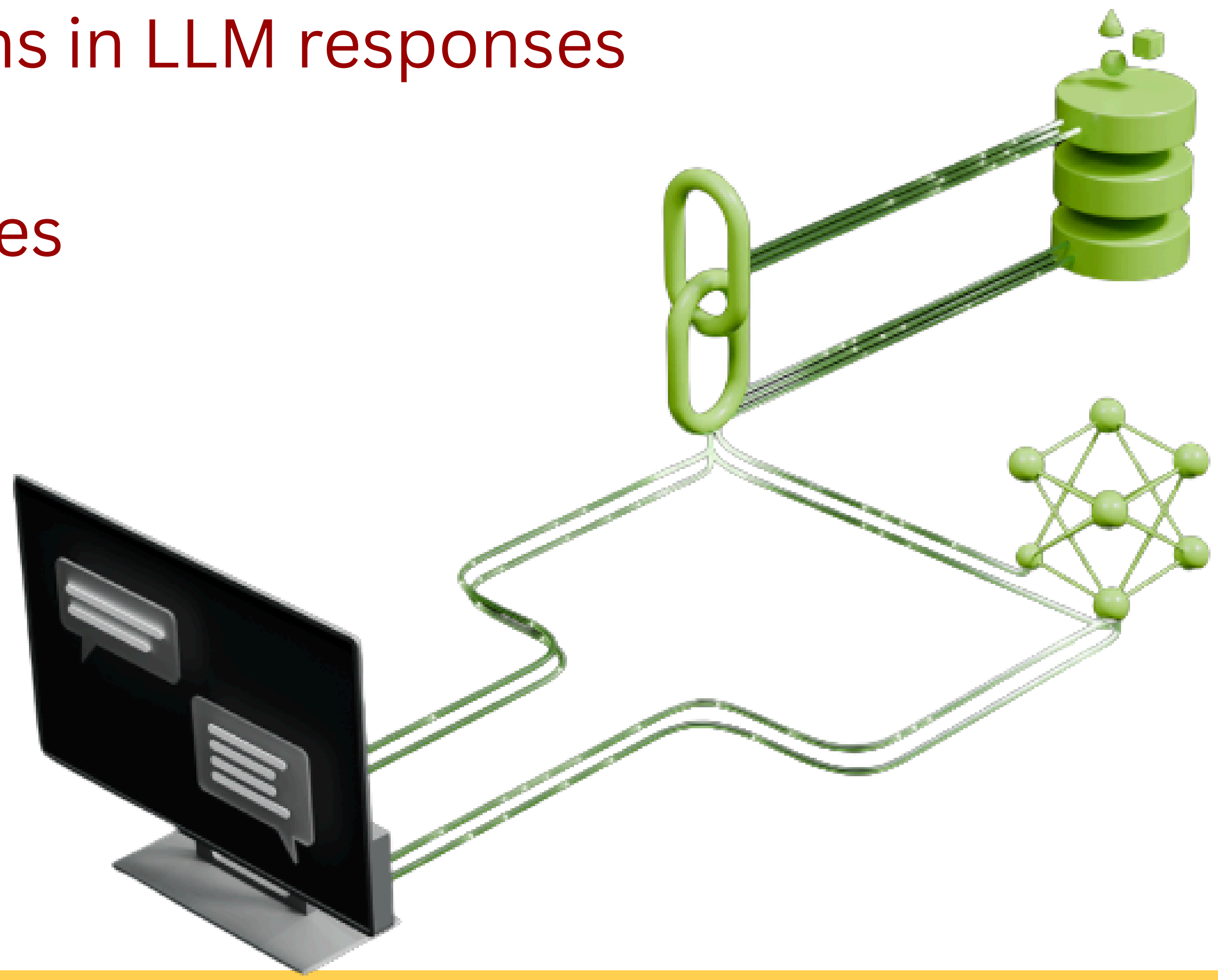


All models show consistent F1-score improvement after applying fact-checking and entropy-based filtering

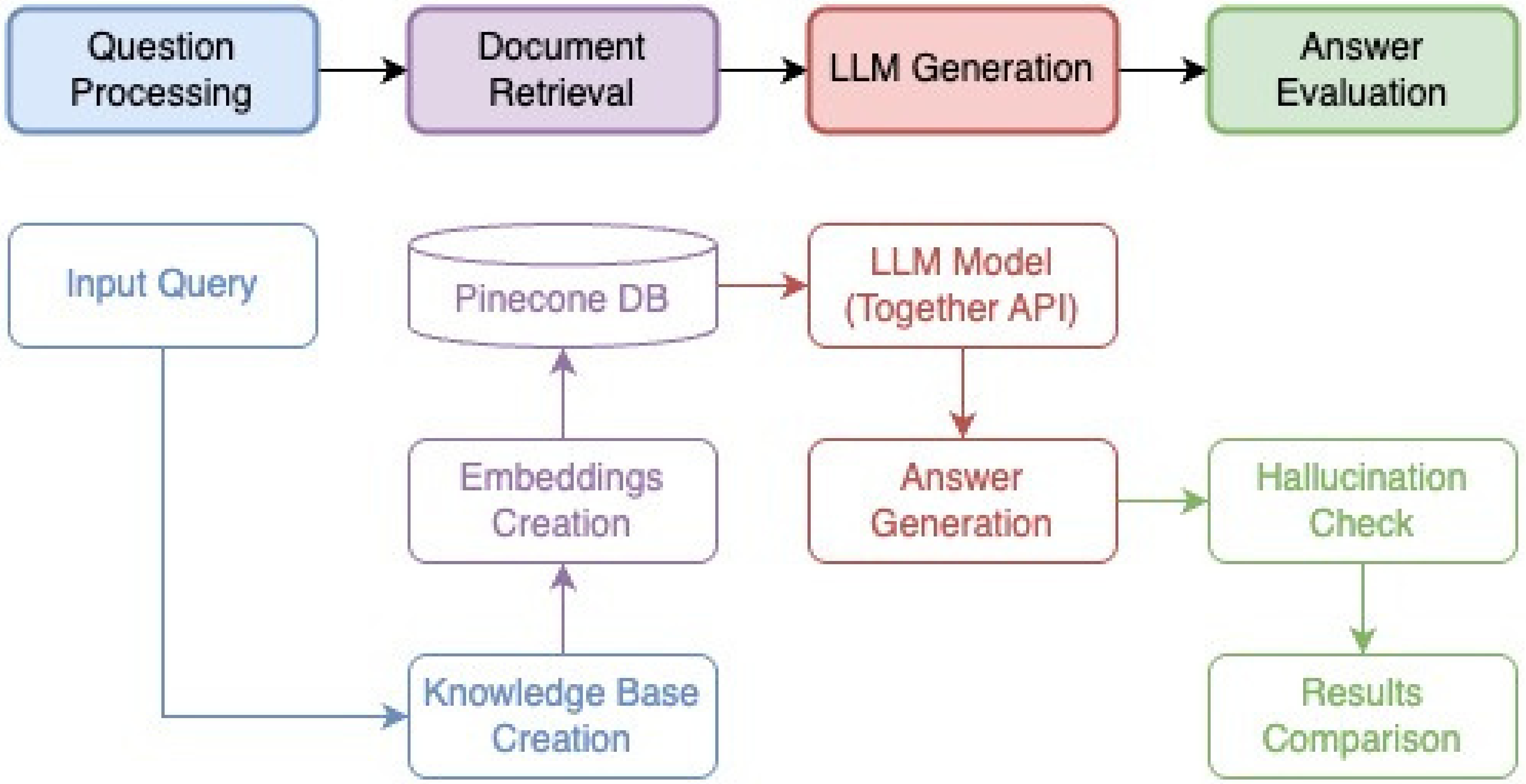
Bloom, Mistral, and Qwen show the most dramatic gains — up to 30–40% increase in F1.

Hallucination Reduction: RAG

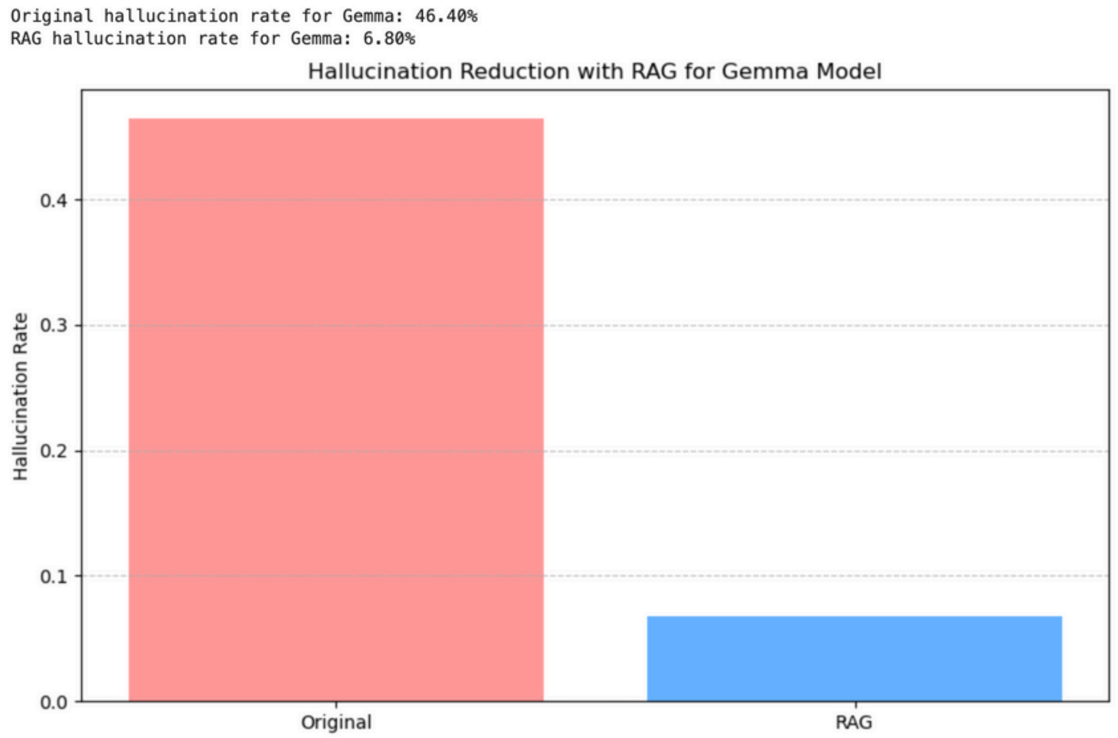
- **Objective:** Evaluate improvement in factual correctness when generation is context-aware
- **Retrieval-Augmented Generation (RAG):** Improves answers by incorporating external factual information which in turn helps reducing hallucinations in LLM responses
- Applied RAG pipeline to LLAMA, Gemma, and Mistral responses



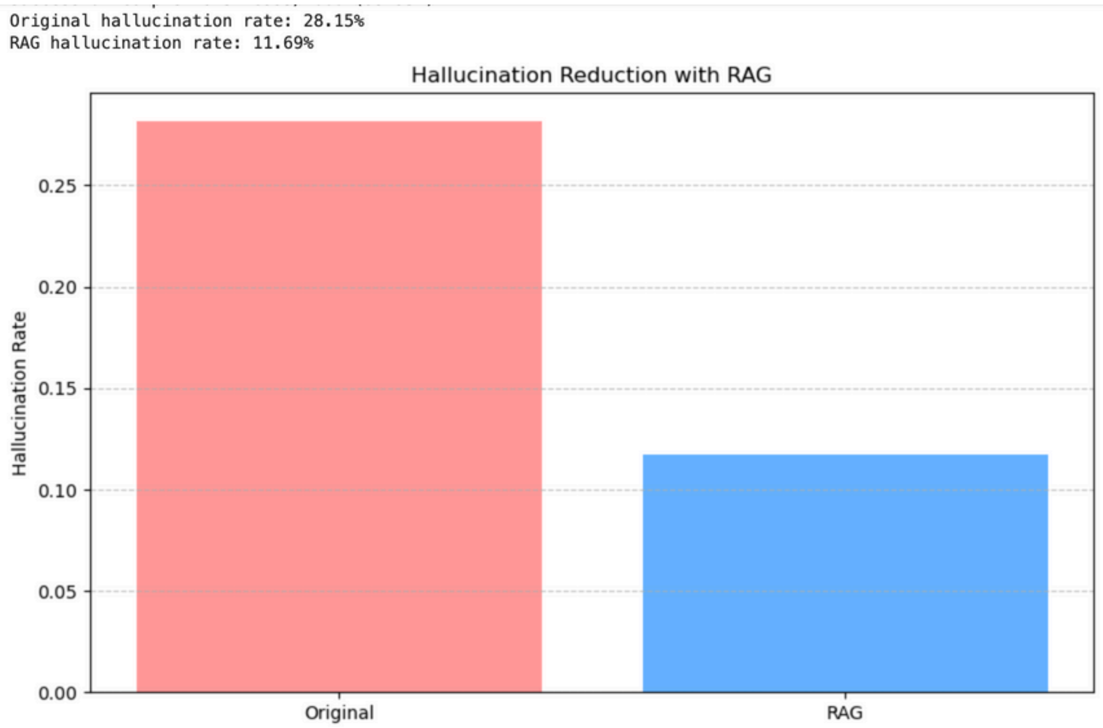
Hallucination Reduction: RAG Pipeline



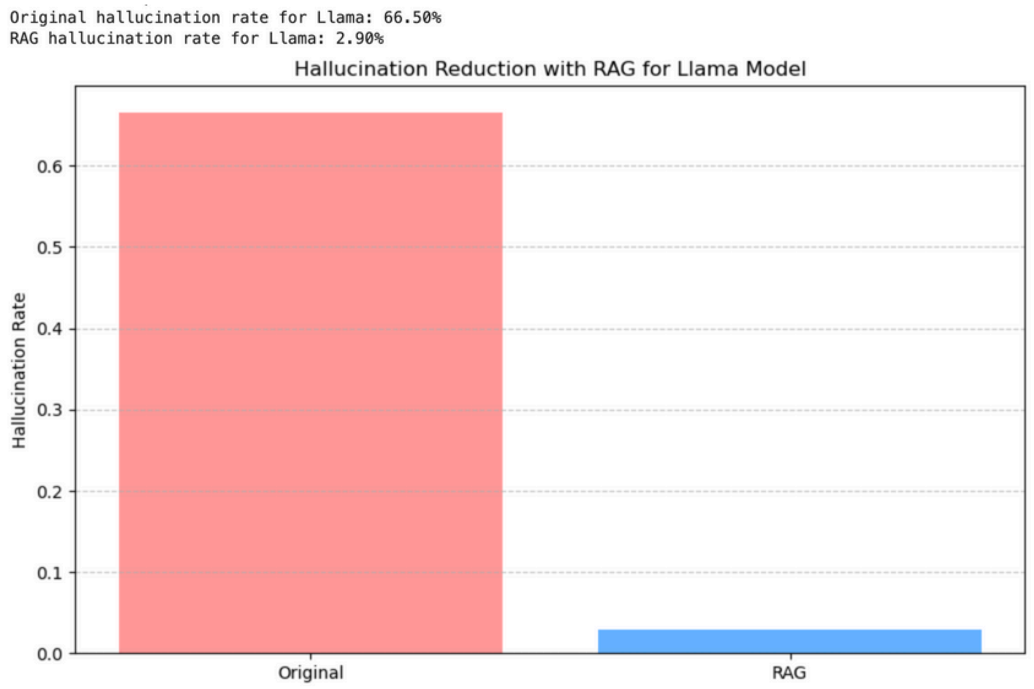
Hallucination Reduction: RAG Results



Gemma Model



Mistral Model



Llama Model

Challenges

- **Inference Misclassifications:** RoBERTa NLI model struggles with domain-specific or nuanced inputs, causing false positives and negatives.
- **API Rate Limits and Cost Constraints:** Together & Groq APIs (e.g., 1,000 responses/key for Groq) require batch processing & scheduling to avoid throttling & ensure consistent data collection.
- **Vagueness Detection Limitations:** Heuristic-based detection may miss vague responses not covered by predefined phrases.
- **Fact Source Constraints:** Sole reliance on Wikipedia limits fact-checking for obscure or dynamic topics.

Q&A