Explainable Fact Checking with LLMs:
How do different LLMs compare in their rationales?

Matei Bordea
BSc Computer Science and Engineering
TU Delft – CSE3000 Research Project
m.bordea@student.tudelft.nl

1. Problem Description

In today's digital world, fact-checking is **more important than ever**. Large Language Models are becoming increasingly capable, and they can generate explanations for claim verifications. These explanations are **often more important than the final result** or the assigned label.

The only thing **stopping** the wide adoption of the LLMs is **the distrust that users have** for the thinking process.

There are a lot of LLMs openly available with very **different** training strategies and training datasets. In this research project we will compare **four of them** and validate if LLMs still **need more time or fine-tuning** before **mainstream adoption.**

2. Research Questions

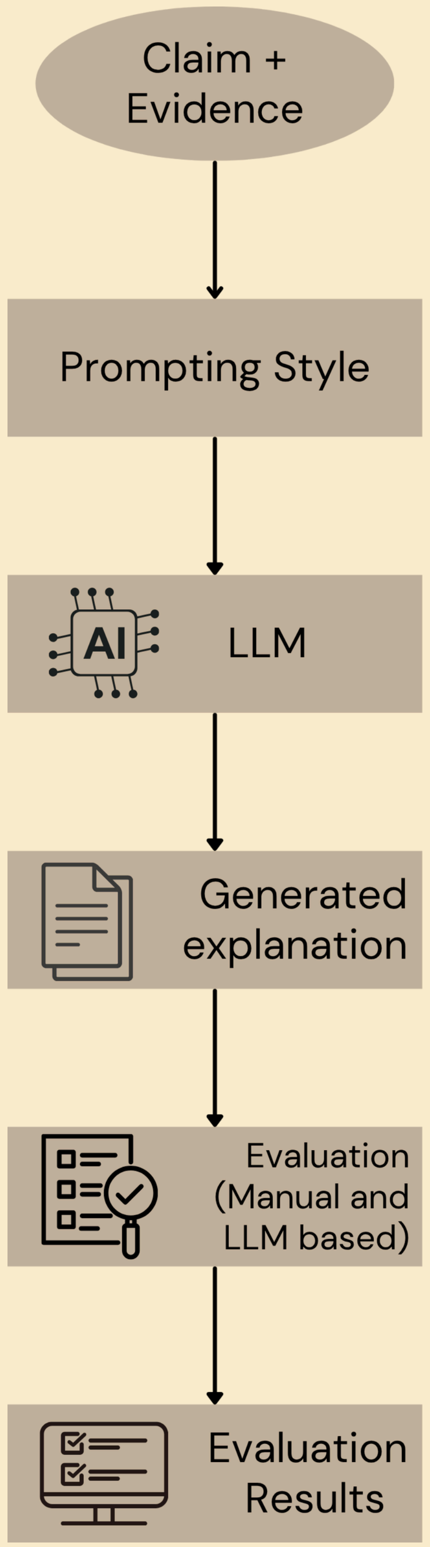
1) To what **extent** do different LLMs maintain **factual consistency** between the **provided evidence** and their generated **explanations**?

2) How do **different LLMs** treat **different types** of evidence?

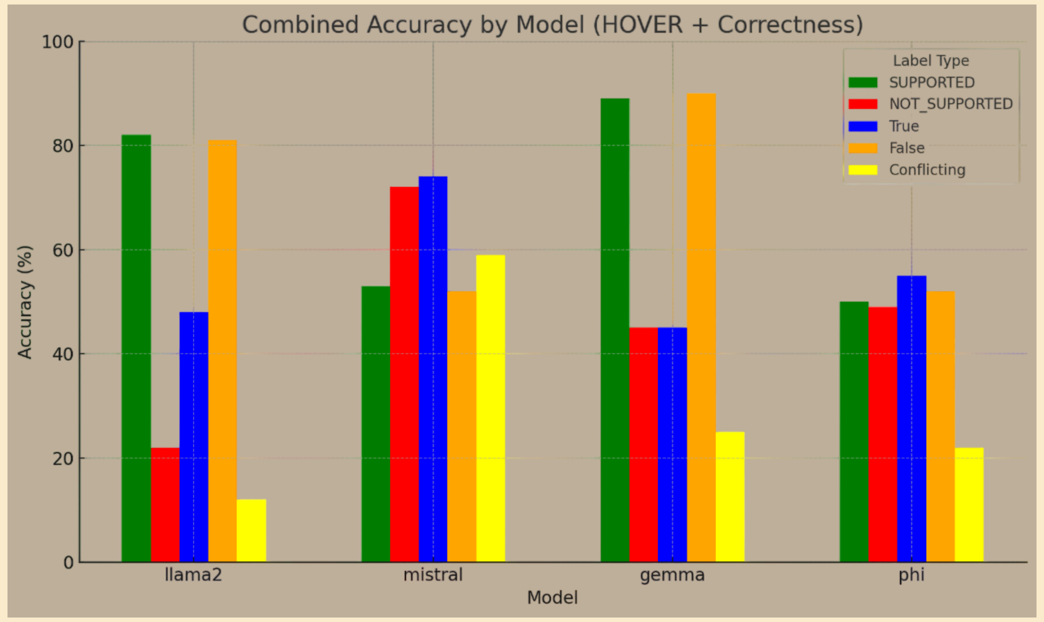
3) Can **automatic evaluation** using an LLM **correlate with human judgment** of faithfulness for LLM explanations?

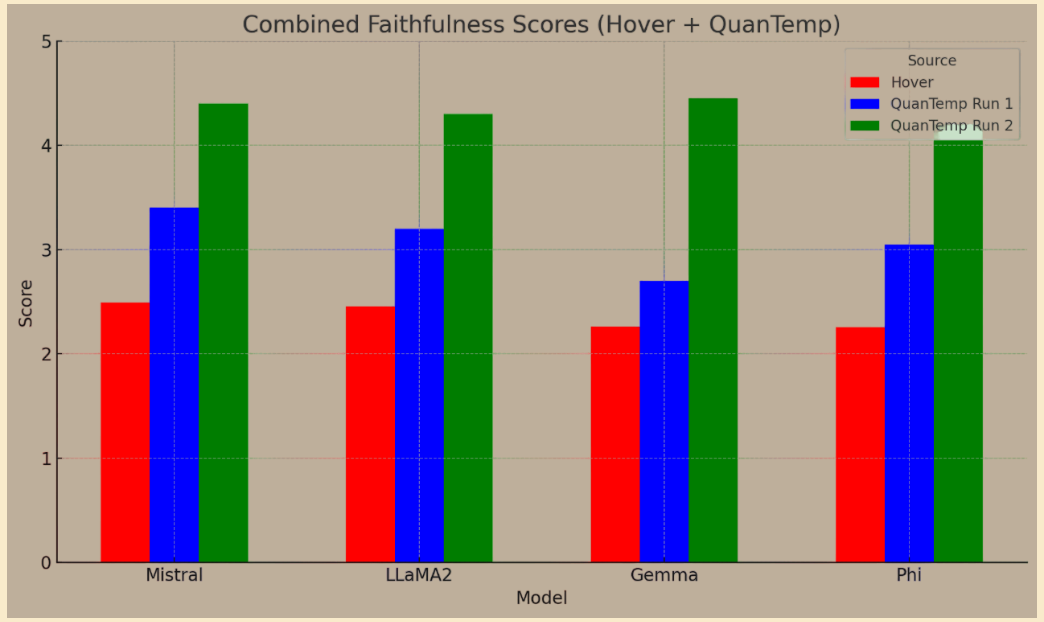
4) Are there **systematic patterns** in the **hallucinations or inconsistencies** produced by different LLMs?

3. Experimental Setup



4. Research results





5. Discussion

Claim Complexity Matters

- **Interval** and **multi-hop** claims caused the **most difficulty** for all models.
- **Statistical** claims were the most **reliably** handled
- **Small modifiers** (like dates or seasons) were **frequently overlooked**, even when important.

Label Accuracy Patterns

- LLMs performed **best** on **false** and **supporting** claims while **conflicting** claims had the **lowest** accuracy.

Faithfulness and Prompting

- Giving the **correct label in the prompt significantly improved** explanation quality.
- LLMs **sometimes justified incorrect labels** when instructed, showing they're **prone to agree with users** even if the evidence doesn't necessarily support it.

LLMs as Evaluators

- Evaluation styles varied:
  - **Phi** focused on precision and penalized unsupported claims.
  - **Gemma** valued fluency and detail.
  - **Mistral** and **LLaMA2** offered balanced, cautious reviews.

Hallucination Trends

- **Hallucinations** were more **common** when the label **wasn't provided**.
- **Gemma** and **LLaMA2** hallucinated by **adding unsupported** but fluent reasoning.
- **Mistral** hallucinated **the least** but sometimes missed subtle implications.
- Phi had **little hallucinations** as it would sometimes **abandon** harder tasks

6. Conclusion + Future Work

In this research the **current limitations of LLMs** are tested and recommendations for the future are made. As observed in this study LLMs **perform moderately well** on most claim types.

However, the models exhibit **inconsistent behavior** across tasks like justification generation. To fix this future research should explore **optimal training strategies** for each type of claim in order **to reduce hallucinations** and **improve evidence faithfulness.**

From an **evaluation perspective** LLMs already have **good language skills** so their current limitation is **computational capacity** and **breaking down the problem** in multiple parts to solve.