# Task 2 – Fact-Checking & Reliability Test

## Objective

Check factual accuracy and hallucination tendencies of different Large Language Models (LLMs).

## Instructions (Summary)

* Each LLM was asked 3 factual questions where the answers are publicly verifiable.
* Correctness was checked using reliable sources such as Google and Wikipedia.
* Special attention was paid to whether the tool gave confident but incorrect answers (hallucinations).
* Deliverable: A table showing the question, tool response, correctness, and confidence level.

## Comparison Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | LLM Model | Tool Response | Correctness | Confidence Level |
| Who is the current Prime Minister of India (2025)? | ChatGPT | Narendra Modi | Correct | High |
| Who is the current Prime Minister of India (2025)? | Claude | Narendra Modi | Correct | High |
| Who is the current Prime Minister of India (2025)? | Gemini | Narendra Modi | Correct | High |
| Who is the current Prime Minister of India (2025)? | Llama 3 | Narendra Modi | Correct | Medium |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | LLM Model | Tool Response | Correctness | Confidence level |
| What is the capital city of Canada? | ChatGPT | Ottawa | Correct | High |
| What is the capital city of Canada? | Claude | Ottawa | Correct | High |
| What is the capital city of Canada? | Gemini | Ottawa | Correct | High |
| What is the capital city of Canada? | Llama 3 | Toronto | Incorrect (Hallucination) | High (confident but wrong) |
|  |  |  |  |  |
| Question | LLM Model | Tool Response | Correctness | Confidence Level |
| When did World War II end? | ChatGPT | 1945 | Correct | High |
| When did World War II end? | Claude | 1945 | Correct | High |
| When did World War II end? | Gemini | 1945 | Correct | High |
| When did World War II end? | Llama 3 | 1946 | Incorrect | Medium |

## Observations

1. **ChatGPT**  – Consistently gave correct answers across all tested questions. It not only provided the factual answer but also explained reasoning in some cases. Its confidence level aligned well with accuracy, meaning that when it answered confidently, the response was indeed correct. This makes it highly reliable for fact-checking tasks.

2. **Claude** – Similar to ChatGPT, Claude performed very well on factual accuracy. It was slightly more verbose in its responses, often providing additional historical or contextual details, which is useful for deeper understanding. Confidence was appropriate, and no major hallucinations were observed.

3. **Gemini** – Accurate in most responses, especially strong in factual queries due to its integration with Google’s knowledge base. Its strength lies in being up-to-date, which makes it particularly good for dynamic or current event questions. However, answers were sometimes shorter and less elaborated compared to ChatGPT or Claude.

4. **Llama 3** – Showed weaknesses in factual consistency. For example, it incorrectly stated that Toronto is the capital of Canada and gave 1946 as the end year of World War II. Importantly, these mistakes were presented with high confidence, which can be misleading. This overconfidence in incorrect answers indicates higher hallucination risk compared to the other tools.

**Overall Insight:**

* For reliable fact-checking, ChatGPT (GPT-5) and Claude are the most trustworthy.
* Gemini is excellent for current facts and live updates, but less detailed.
* Llama 3 is best used for exploratory or open-source contexts but requires manual verification due to its hallucination tendency.