

Accuracy of large language models for scientific reference attribution

1

David McMinn,<sup>a</sup> Steven Grant,<sup>b</sup> and Chirlene Hall<sup>c</sup>

<sup>a</sup>McMinn Medical Communications, East Kilbride, Scotland  
<sup>b</sup>GHIJK, Hairini, New Zealand  
<sup>c</sup>Napier University, Edinburgh, Scotland

Background & aims

- Large language models (LLMs) have the potential to provide efficiencies across a range of tasks in the field of medical communications
- One such task is attributing published references to scientific statements
- To date, however, there have been several examples of government,<sup>1</sup> business,<sup>2</sup> and academic<sup>3</sup> reports that contain inaccurate (or fabricated) references, possibly as a result of LLM use
- The purpose of this study was to assess the feasibility and accuracy of using LLMs for identifying real and relevant published references to support scientific statements

Methods

1

**Procedures**

- The following two LLMs were tested:
  - Chat GPT 4.0
  - Gemini 2.5-Pro-preview-05-06
- Each LLM was provided with a total of thirty true and thirty fabricated scientific statements using three different prompts

2

**Prompts**

- **Prompt 1:** asked the LLM to provide a published reference to support the scientific statement
- **Prompt 2:** Prompt 1 with an additional clause to search for the reference only within PubMed
- **Prompt 3:** Prompt 2 with the addition of an anti-hallucination clause, whereby the model was not to provide a reference if none could be found to support the scientific statement

3

**Outcomes**

- Outcomes for the true scientific statements included
  1. was a reference provided (yes/no)
  2. was the provided reference real (yes/no)
  3. for real references: was the associated published content relevant to the scientific statement (yes/no)
- These three outcomes were used to calculate a composite overall accuracy score (%) for each prompt for each model
- The outcome for the fabricated scientific statements was the accuracy (%) for correctly not providing a reference

Results

Accuracy of reference attribution for true scientific statements

- For GPT 4.0, overall reference attribution accuracy was 78%, 89%, and 76% for Prompts 1, 2, and 3, respectively. Prompt refinement (Prompt 1 to 2) improved accuracy (**Figure 1**)
- For Gemini, overall reference attribution accuracy was 74%, 67%, and 57% for Prompts 1, 2, and 3, respectively. Prompt refinement (Prompt 1 to 2) did not improve accuracy (**Figure 2**)
- When comparing performance between models, overall accuracy was higher for Chat GPT 4.0 irrespective of prompt, driven largely by better performance in providing real references. Where real references were provided, Gemini performed better with respect to content relevance

Figure 1. Chat GPT accuracy for attributing references to true scientific statements

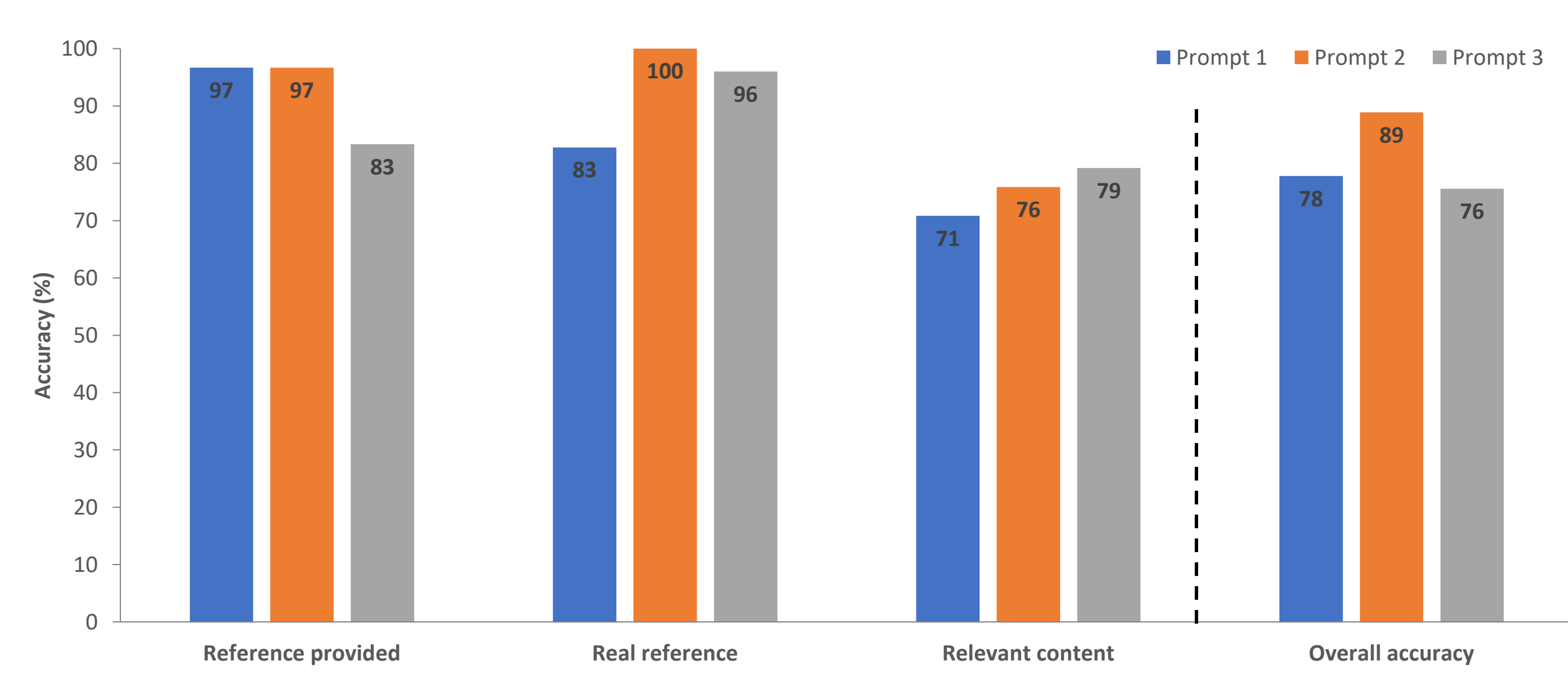
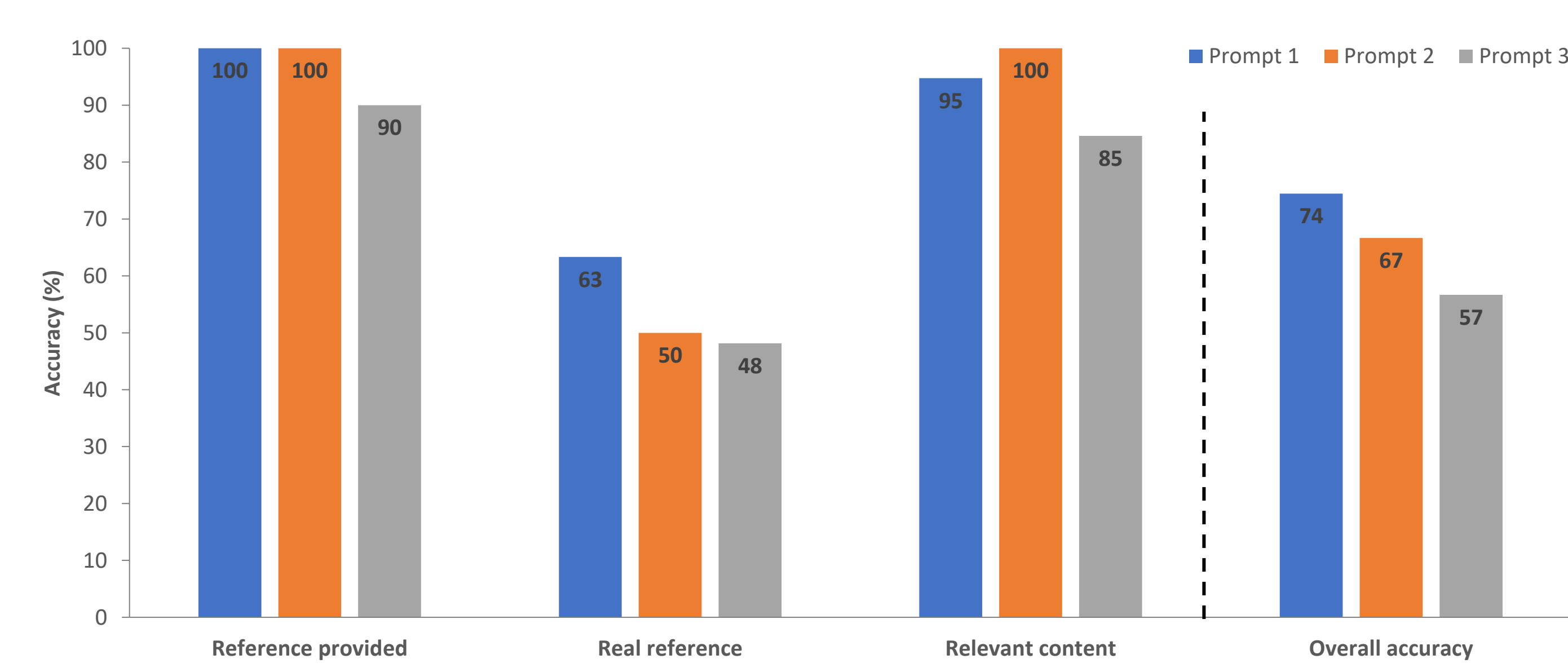


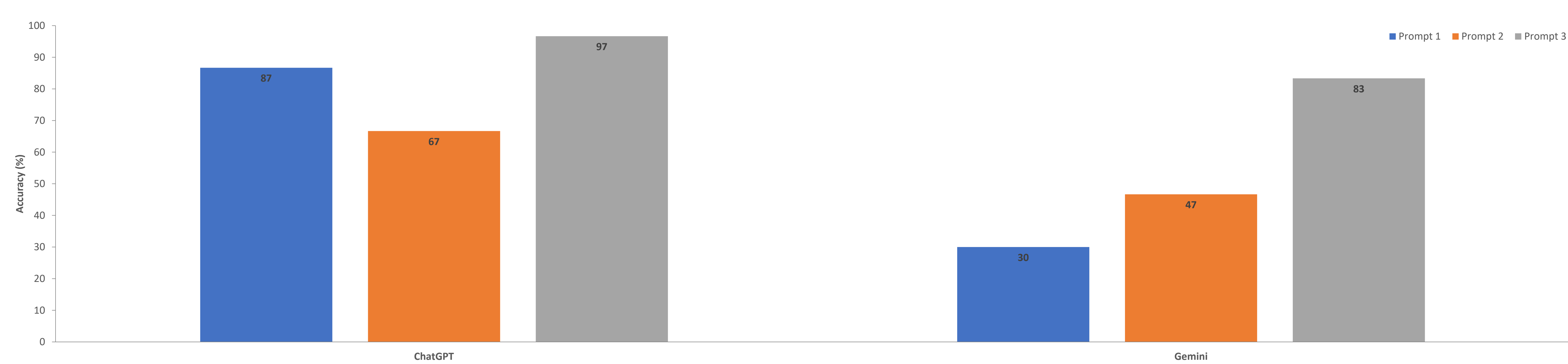
Figure 2. Gemini accuracy for attributing references to true scientific statements



Accuracy for correctly attributing no reference to fabricated scientific statements

- For GPT 4.0, accuracy was 87%, 67%, and 97% for Prompts 1, 2, and 3, respectively. Inclusion of the anti-hallucination clause (Prompt 3) improved performance
- For Gemini, accuracy was 30%, 47%, and 83% for Prompts 1, 2, and 3, respectively. Inclusion of the anti-hallucination clause (Prompt 3) improved performance
- When comparing performance between models, accuracy was higher for Chat GPT 4.0 irrespective of prompt

Figure 3. Accuracy for correctly attributing no reference to fabricated scientific statements



Conclusions

- There is variability between LLMs and prompts for accurately attributing references to scientific statements, with Chat GPT 4.0 performing better than Gemini
- Prompt refinement does not necessarily improve model performance, though inclusion of an anti-hallucination clause helps improve accuracy with respect to not providing references for fabricated scientific statements
- These findings support the potential utility of LLMs for use in scientific reference attribution, though human input is still required to ensure accuracy

References

<sup>1</sup>Dyer, O. *BMJ* 2025;389:r1150. doi: <https://doi.org/10.1136/bmj.r1150>  
<sup>2</sup>Pandey, R. *SSRN* 2025. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=5652350](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5652350)  
<sup>3</sup>Sharifi, A. *Nature* 2025;644:<https://www.nature.com/articles/d41586-025-02482-1>

Disclosures

- David McMinn is a shareholder of McMinn Medical Communications Ltd
- Steven Grant is a shareholder of GHIJK
- Chirlene Hall has nothing to disclose