Citation Behavior in AI vs Human-Written Scientific Texts: A Controlled Comparison

Raffaello Mastromarino

Research Question

How do the citation patterns of generative AI (LLM)-augmented scientific texts differ from those of human-written texts on the same scientific topics? In particular:

- What do LLMs cite (journal prestige, recency, diversity)?
- How do they cite (number and types of references)?
- Are there detectable biases (geographic, gender, centrality)?

Motivation

Recent advances in generative AI have introduced large-scale language models (LLMs) into scientific writing workflows. This project explores whether these models reproduce or distort existing norms in citation behavior. Understanding these dynamics is crucial to:

- Detect potential biases introduced by LLMs
- Assess how AI may alter knowledge diffusion
- Inform policy on the responsible use of LLMs in science

Data Sources

- **Human-written papers:** We collected abstracts and metadata from OpenAlex for papers published after the Qwen 3 32B knowledge cutoff (end of 2024), ensuring that the LLM had no access to their citations. In particular, we selected only those papers that included a bibliography. In total, we sampled 3,662 papers.
- **AI-generated citations:** For each human-written abstract, Qwen 3 32B was prompted to generate a plausible bibliography, effectively producing a reference list that could accompany the paper.
- **Metadata extraction:** OpenAlex API was used to collect citation metadata (author affiliations, countries, gender estimates, prestige via journal/source, etc.)

Methodology

Proposed variant to Option A: Instead of asking the LLM to write full papers with citations, I inverted the direction: I used real scientific abstracts (post-knowledge-cutoff) and asked the LLM to generate citations for them. This approach:

- Ensures the scientific content is held constant between human and AI allowing for pairwise comparisons.
- Allows controlled comparisons focusing purely on citation behavior.
- Allow for comparisons between different models.

Analysis Pipeline:

1. Select a set of real titles and abstracts from recent AI papers on OpenAlex (in our case 100 out of the starting 3662).
2. Generate citation lists using Qwen 3 32B.
3. Extract metadata for cited papers via OpenAlex.
4. Detect hallucinations using a fuzzy matching approach between generated citations and OpenAlex records.
5. Compare:
    - Number of citations
    - Prestige (citation count, venue)
    - Recency (publication year distributions)
    - Diversity (author geography, gender)
    - Overlap with real human-written references

Prompt Design and Rationale

To simulate citation behavior, we asked Qwen 3 32B to generate bibliographies based solely on the *title* and *abstract* of scientific papers. This setup allows us to study how the model extrapolates plausible citations based on its internalized structure of scientific knowledge and citation norms.

A key design decision was to use a clear and minimal prompt. This simplicity is intentional and crucial: overly detailed or leading prompts risk introducing bias or constraining the model toward specific citation styles, topics, or author identities. Instead, our prompt asks the model to construct a bibliography that is plausible and diverse, encouraging it to rely on its general knowledge rather than memorized associations or heuristic shortcuts.

The prompt used is as follows:

```
Generate a list of references for a paper having as title and abstract:

Title: {title}

Abstract: {abstract}

Bibliography:

Please format the bibliography as a numbered list, where each reference follows this pattern:
Author(s): <authors>; Title: "<title>"; Year: <year>; Venue: <journal/conference>
```

This neutral and standardized phrasing helps reduce the risk of steering the model and ensures consistency across outputs, thereby supporting fair and interpretable comparisons between LLM-generated and human-written citation patterns.

Preliminary Insights

- Qwen 3 32B displays a clear bias toward citing papers with higher citation counts, indicating a prestige bias.
- The publication years of cited papers in the AI-generated bibliographies tend to be more recent than those in human-written references.
- No straightforward gender bias was detected in the authorship of cited works.
- There is a small bias toward journals with higher SJR rankings in AI-generated citations.
- A possible geographic bias in AI citations is observed but requires further in-depth analysis to confirm and characterize.

- Citation overlap between AI-generated and real references remains low, as we can deduce by Jaccard score.

Conclusion and Future Work

This methodology allows a controlled investigation of citation behavior without relying on speculative stylometry or full LLM-generated articles. Next steps include:

- Expanding to more LLMs (e.g., Claude, Gemini, Mistral) to compare citation behaviors.
- Studying field-specific citation patterns (e.g., AI vs biology).
- Refining bias metrics and deepening the geographic bias analysis.
- Trying other methods.

GitHub: `https://github.com/huygenssteiner/citation-diversity-ai-vs-human` **Notebook:** Curated Jupyter notebook included in the repository