Hide or ***Highlight***: Understanding the Impact of Factuality Expression on User Trust

Hyo Jin (Gina) Do and Werner Geyer

IBM Research

- Large language models (LLMs) often produce convincing but non-factual information, known as "**hallucinations**".
- **Factuality estimates** can help users identify incorrect information in AI-generated answers.
- A common method for presenting factuality estimates is using visual **highlights** on content that is estimated to be either more or less factual, which assumes that the original AI-generated answer is always disclosed to the user.
- Can we build user trust by **hiding** less factual content, either by removing it or making it vague?
- In this work, we explore four strategies for presenting an AI-generated response with factuality estimates and a baseline:

Tell me a bio of Gonzalo Fonseca

Gonzalo Fonseca was a Uruguayan sculptor and painter born on July 2, 1922, in Montevideo, Uruguay. He studied architecture at the University of Montevideo before moving to Paris in 1950, where he became increasingly interested in sculpture.

BASELINE (No factuality estimates)

Gonzalo Fonseca was a Uruguayan sculptor and painter born on July 2, 1922, in Montevideo, Uruguay. He studied architecture at the University of Montevideo before moving to Paris in 1950, where he became increasingly interested in sculpture.

TRANSPARENT (Highlights less factual content)

Gonzalo Fonseca was a Uruguayan sculptor and painter born on July 2, 1922, in Montevideo, Uruguay. He studied architecture at the University of Montevideo [..], where he became increasingly interested in sculpture.

OPAQUE (Removes less factual content)

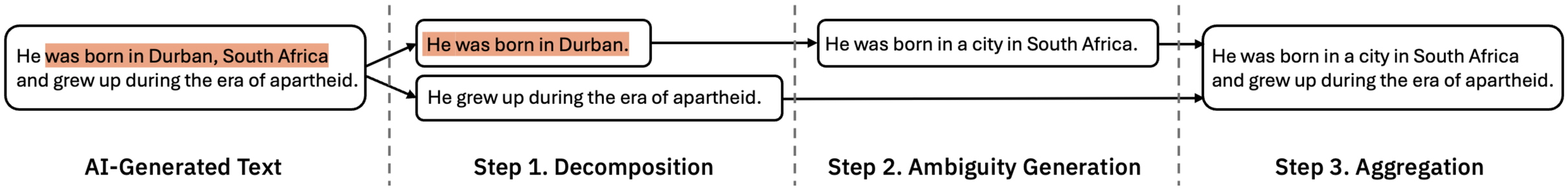
Gonzalo Fonseca was a Uruguayan sculptor and painter born on July 2, 1922, in Montevideo, Uruguay. He studied architecture at the University of Montevideo before moving to Paris in 1950, where he became increasingly interested in sculpture.

ATTENTION (Highlights factual content)

Gonzalo Fonseca was a Uruguayan sculptor and painter born on July 2, 1922, in Montevideo, Uruguay. He studied architecture at the University of Montevideo before moving to another country in the 1950s, where he became increasingly interested in sculpture.

AMBIGUITY (Makes less factual content vague)

AMBIGUITY STRATEGY IMPLEMENTATION



Used an LLM to generate atomic facts for each sentence and annotated factuality (Min et al. 2023)

Used an LLM to convert each incorrect fact to an ambiguous statement and aggregate them into a single sentence

HUMAN SUBJECT EXPERIMENT

An online survey experiment with 148 participants, each completing 4 tasks.

RQ1. How do varying strategies affect user trust and reliance in AI?

- Trust Belief: Transparent, Attention, Baseline < Opaque, Ambiguity ($p<.05$)
- Appropriate Compliance: Transparent < Opaque, Ambiguity ($p<.01$)
- Perceived Transparency: No significant difference

RQ2. How do varying strategies affect the perceived answer quality?

- Correctness: Transparent, Attention, Baseline < Opaque, Ambiguity ($p<.05$)
- Relevance, Conciseness, Completeness, Coherence: No significant difference

<AI-generated answer>
Michael Valpy is a Canadian journalist, author, and academic. He was born in 1944 in Montreal, Quebec, and studied at the University of Toronto, where he earned a Bachelor of Arts degree in 1965 and a Master of Arts degree in 1968. Valpy began his journalism career as a reporter for The Globe and Mail in 1968, and went on to work for several other Canadian news outlets, including The Toronto Star, Maclean's magazine, and CBC Radio. In addition to his work as a journalist, Valpy is also an academic. He has taught at several Canadian universities, including Queen's University, the University of Western Ontario, and Carleton University. He has also been a visiting professor at universities in the United States and the United Kingdom. Valpy has written several books on Canadian politics and society, including "The Age of the Mosaic" and "The Pursuit of Happiness: An Agenda for a Better Society". He has received several awards for his work in journalism, including the Canadian Association of Journalists' Award for Investigative Reporting in 1980 and the Atkinson Fellowship in Public Policy in 1998.

<Reference> (the link will open in a new window, so don't worry about leaving this survey)
https://en.wikipedia.org/wiki/Michael_Valpy

TAKEAWAYS: Hide low factuality content rather than highlighting it

- Opaque or Ambiguity strategies outperformed traditional highlighting strategies, increasing trust without compromising perceived quality.
- Depending on the context, we recommend **dynamic disclosure** methods to avoid omitting potentially valuable information.