Human or Correct: What Do We Want From AI?

Of course we want bias in LLMs

By Dwayne Phillips

This essay is from Dr. Dwayne Phillips, PhD. Dr. Phillips is available for hire to research, analyze, and report on topics from AI to writing to budget to management. d.phillips@computer.org. This and other research reports are available at https://dwaynephillips.net/MediumEssays/index.html

Summary

There is a tension between LLM accuracy and human-like responses. Should LLMs prioritize factual correctness (like a know-it-all) or mimic human conversational patterns (like the man-on-the-street). Training data and evaluation metrics inherently embed biases, influencing LLM outputs. Some improvements to the current practice include: displaying sources of information for the LLM, using accepted sources, and using some sources for facts and others for language style. I recommend correct and understandable LLM responses.

The Man-on-the-street or the Know It All?

Bias: noun, prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair. (Oxford)

Who wants bias? That is unfair. Is it? Is it an accurate reflection of us?

Two of the results of bias in LLMs are misinformation (wrong answers) and disinformation (deliberate attempts to fool someone). Wrong answers, i.e., misinformation, are unwanted. Two plus two is four, not five. LLMs should be correct.

The above is a value judgement. Is correctness valued? What is the purpose of the LLM? Is it to pass the Turing Test? (Wikimedia 2025) Is it to be a source of correct and verified knowledge in the world?

If the Turing Test is the valued outcome, and the valued outcome of the Turing Test is still debated (Scott, 2024), the LLM should answer questions like a man-on-the-street. In contrast, is it better for the LLM to answer the question correctly?

George Washington never said, “I cannot tell a lie.” (Scelzo, 2024) The man-on-the-street, however, would attribute that to the father of our country. It is a good story, everyone has heard it, and in a Turing Test the computer would provide that (right?).

Who wants to talk with a know-it-all who states the first President of the United States was John Hanson. Hanson was the first President of the United States under the Articles of Confederation. (Michael) Who, however, remembers the Articles of Confederation and would value that answer over the traditional George Washington under the current Constitution?

Should the LLM answer like a person or answer correctly?

Bias Is a Statement of Value

A good LLM will output that George Washington was the first President of the United States. A better LLM may continue with the history of John Hanson and the first attempt of a central government among the newly united colonies. That is good and correct.

The above is a value statement of mine filled with my bias.

Is my bias unfair per the definition at the beginning of this essay? “Fairness” to me is “unfair” to someone else; it is easy to chase around in circles and never conclude anything.

Current LLMs usually train by scraping the Internet. At some point, the supervised learning in the LLMs is supervised by an oracle of correctness and decency. That oracle is created by persons who have their own values. Up until now, the great majority of those persons are scientific researchers. Therefore, current LLMs are biased towards science and scientific fact or *prevailing scientific opinion*.

At one time prevailing scientific opinion stated there were three races of humans with different intellectual capacities. (Wikimedia 2025a) Current science can boast that such fallacies in prevailing scientific opinion no longer exist. That boast is a form of recency bias and is yet another prevailing scientific opinion, and we are chasing around in circles again.

Which Values? Some Recommendations

LLMs are trained on information produced by persons (that statement is starting to become untrue as some LLMs train on information produced by LLMs, but that is another essay for another day). Which persons will be the sources of training material? An old text of mathematical tables would be a good source of facts in mathematics. How about chemistry? Standard texts would be good sources. How about political science? Now we are starting to slide into trouble.

I suggest **labelling the LLMs clearly and repeatedly**. Point to the sources of information used in training and testing. Wikipedia is good for some things. Brittanica is good for some things. The Oxford English Dictionary is good for some things. Those three examples are my biased value statements.

To make a practical list of sources requires another change in LLMs: greatly **reduce the materials used for training**. Wikipedia, Brittanica, and the Oxford English Dictionary are just three sources, but a person could not read the entirety of all three in a lifetime. Current practice uses "the entire Internet" for training. Folly. Someone is using the materials I post on the Internet Really? I am flattered, but who thought that would be a good idea?

Finally, consider the common use of English. That greatly influences how LLMs respond to prompts. A system that responds in the English of 16th century England would fail the Turing Test.

Everyone has a podcast. Everyone else has a blog. Everyone else is on Facebook et al. **Train LLMs for common use of English on podcasts, blogs, and social media**. Prohibit training on these for answers to specific questions.

For example, prompted with, "describe bias in AI," a LLM system would use Wikipedia, Brittanica, and the Oxford English Dictionary for the content and then use podcasts, blogs, and social media for the expression of the content. The answer would be from a know-it-all but in the language of the man-on-the-street.

To summarize,

> *Provide sources, reduce sources, and combine sources.*

My bias is towards correct yet understandable.

References

Michael, P. (n.d.). Remembering John Hanson, First president of the original United States Government. Remembering John Hanson, First President of the Original United States Government | Sacramento State. https://www.csus.edu/experience/retirees/publications/articles/remembering-john-hanson.html

Oxford languages and google - english. Oxford Languages. (n.d.-a). https://languages.oup.com/google-dictionary-en/

Scelzo, S. (2024, July 30). 10 famous quotes that you probably misattributed. Mashable. https://mashable.com/article/famous-misattributed-quotes#:~:text=5.,not%20come%20up%20with%20it.

Scott, C. (2024, February 22). Study finds chatgpt's latest bot behaves like humans, only better. Stanford School of Humanities and Sciences. https://humsci.stanford.edu/feature/study-finds-chatgpts-latest-bot-behaves-humans-only-better

Wikimedia Foundation. (2025a, February 3). Historical race concepts. Wikipedia. https://en.wikipedia.org/wiki/Historical_race_concepts

Wikimedia Foundation. (2025, March 28). Turing test. Wikipedia. https://en.wikipedia.org/wiki/Turing_test