The technically catastrophic results of anti-discrimination measures in Google Gemini

Rishi Krishna

CS 3604 (CRN 83358)

compiled using typst. it's like LaTeX with fewer backslashes.

1. Description

Generative AI tools, specifically those for textual and photographic works (in the modern day, LLMs and Diffusion models), are often handicapped as part of their creation and commerical release to prevent possible controversy for the company that created it. These handicaps, even when implemented on different layers of abstraction, are ruinous for any good technical product, actively lowering the quality of the results.

In the specific case we are covering, Google's flagship GAI family, Gemini, would automatically turn various historical figures into dark-skinned figures, in a misguided attempt to combat potential racism or discrimination from them model. This became controversial, as it generated even historically racist figures (e.g. prominent Nazi party members) as what would most likely be interpreted as African-Americans.

This essentially killed the Gemini image generation release, making the model (which cost millions in GPU and man hours to produce) useless.

2. Relevance

This situation is specifically relevant to the unit, **Algorithmic decision making and AI**, not just due to the underlying AI technology behind it, but the potentially negative consequences of a loss in quality with regards to decision-making. The paradoxical relationship between anti-discrimination handicaps that are supposed to increase the value of a model for decision-making is such that those very handicaps can easily make the outputs useless, for decision-making or nearly any other purpose.

The method by which these handicaps are put into place comes in two common forms: prompt injection and dataset shaping[1]. The Gemini incident involved both. In the case that user prompts are wrapped in "friendly" anti-discriminatory system prompts, the result is often useless ("I'm sorry, I am an AI Language model, and I cannot do this…") or plain incorrect (Gemini turning historical figures black)[0]. In the case that the model is handicapped, the limited breadth of knowledge it is exposed to during training limits its reasoning capacities based on simple information theory – a man who knows of racism and chooses against it has significant differences in comprehension and ability to one that has never been exposed to the concept of racism except in a clinically crafted way.

This precarious balance is core to CS ethics – many times the efficiency of an algorithm depends on cutting a few sociological corners, and it is up to us as the Computer Scientists to decide whether the end result is worth it.

3. Crux

One one hand, people want their tools to do good work. A shovel digs, even if the hole is meant to be a shallow grave. On the other hand, the ethics of making a tool available that can and will produce volumes and portfolios of potentially objectionable content should be brought up. If Google's Gemini LLM offerings were used to create and spread enthusiastic medical misinformation, forming a false consensus and tricking a great many people into unsafe activities, should the company be morally responsible for not installing safeguards in their tool? If shovels could be modified so that they won't dig shallow graves, should they be?

4. References

[0] https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

[1] https://www.nature.com/articles/d41586-024-00674-9