Hello Ian,

You addressed bias and hallucinations, and adding a concrete example could strengthen your point. For instance, in March 2025, ChatGPT informed a Norwegian individual, Arve Hjalmar Holmen, that he had murdered his two sons and was serving a 21 year prison sentence. None of this was true, yet the model presented it with complete confidence (theguardian.com). This example highlights why robust safeguards are essential (Noyb, 2025).

You brought up diverse datasets and fairness checks in your mitigation section. Retrieval augmented generation is an additional useful strategy. Using this method, the model must first validate its answers against a reliable database before providing them. In what way could you incorporate this phase into your existing process? In a healthcare context, you could follow this with a human review of any diagnostic or treatment recommendations (Merritt, 2025).

As an additional layer of verification, consider using two separate AI models to evaluate each other's outputs. If either model identifies uncertainty, the response would be forwarded to a human reviewer. This combination of dual AI validation and human oversight can greatly reduce the risk of errors (Greyling, 2024).

I look forward to seeing how you apply these ideas in a real world scenario. What do you think about using a multi-model approach with human oversight in the loop?

All The Best!

Avinash

References

Greyling, C. (2024, August 15). *AI agents with human in the loop*. Medium. Retrieved May 3, 2025, from https://cobusgreyling.medium.com/ai-agents-with-human-in-the-loop-f910d0c0384b

Merritt, R. (2025, January 31). *What is retrieval-augmented generation aka RAG*. NVIDIA Blogs. Retrieved May 3, 2025, from https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

Noyb. (2025, March 20). *AI hallucinations: ChatGPT created a fake child murderer*. noyb. Retrieved May 3, 2025, from https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer