Hi Robert,

I wanted to share two automotive industry examples that mirror those challenges and suggest how multi-model guardrails can help:

Chatbot Mischief: A Chevrolet customer service chatbot once literally agreed to sell a new Tahoe for just one dollar after someone tricked it with a clever prompt. This prank showed that AI interfaces can be manipulated into absurd or legally questionable actions. It highlights the importance of strict input validation and human oversight (Maliugina, 2024).

Ensemble Verification: In a research project called CarExpert, developers combined multiple language models and cross checked each response against the official vehicle manual. This ensemble approach cut hallucinations by more than half and ensured drivers only received accurate information. It demonstrates how multi model checks can flag or correct false outputs before they reach end users (Giebisch, Friedl, Sorokin, & Stocco, 2025).

Building on your mitigation ideas, community colleges could pilot similar strategies by running parallel AI agents and escalating any conflicts or anomalies to human reviewers. This method would catch subtle errors and guard against bias before AI informs critical decisions.

Thanks for sparking this important conversation.

All The Best!

Avinash

References

Giebisch, R., Friedl, K. E., Sorokin, L., & Stocco, A. (2025, April 1). *Automated factual benchmarking for in-car conversational systems using large language models*. arXiv. https://doi.org/10.48550/arXiv.2504.01248

Maliugina, D. (2024, September 25). *LLM hallucinations and failures: Lessons from 4 examples*. Evidently AI. Retrieved April 30, 2025, from https://www.evidentlyai.com/blog/llm-hallucination-examples