Matthew Branson
ENG 684-899
Dr. Leslie Seawright
September 21, 2025

Reflection Paper:
AI-Generated Letter to a Client

Introduction

I noted that the assignment specifications had a closed list this time for which models might be employed, and seeing that DeepSeek was not listed, I chose to instead use Gemini 2.5 Pro as our third model.

My strategy here was to use the two attempts to demonstrate the benefit of extensive contextual background for improving the coherence and purpose of the draft letter. In the first attempt, I simply presented the given information and formatted it into xml to present it to the model for immediate action, essentially performing what might be generated if just handing the task off to the model naively without careful consideration. In the second attempt, I prepared beforehand a sequence of three research reports generated by the Claude beta deep research model, which I offered conversationally to each model for reflection in sequence as preamble to the task.

Which platform handled tone and ethical nuance best? Why?

I believe that Claude demonstrated the best handling of tone and ethical nuance. Sheer verbosity alone is one measure for how the model handles nuance, since it simply just requires more words to communicate nuanced detail, and consequently Claude's final draft is by far the longest letter and presents a quite detailed plan of action for the client.

Interestingly, among the models employed here, Claude has the most unique ethical alignment training, using a process of interest to me called Constitutional AI. If you're familiar with Amanda Askell at Anthropic, she is an interesting researcher on the topic, being essentially the single lead human being in charge of Claude's personality. Some aspects of the process are proprietary, but as she's described in podcast interviews, they will begin with essentially a written constitution of ethical values, then they will have Claude generate tens of thousands of hypothetical conversations that test each of those values, allowing Claude to pursue those conversations in different ways to different outcomes. After, all the hypothetical conversations are graded against the original written constitution, and those evaluated as most ethical in their conversational outcomes are introduced into the alignment training corpus – and (one of the most interesting parts) the grading of the hypothetical conversations is done by Claude itself without human intervention, using essentially RLAIF as opposed to RLHF. I would not be surprised to learn that this specific scenario of business communication related to child labor exploitation may be part of that alignment training corpus already.

How did your prompts shape or limit what the AI produced?

My notion is that every interaction preceding the task serves some role in activating the latent space in which the model operates. Every past token is part of the prompt, and exploring ideas beforehand can open avenues of communication that might otherwise remain undiscovered.

I also decided to inform the models directly that they were performing in a test where their final drafts would be compared against those generated by other models. This can certainly affect the results, and among reasoning models with their thought tokens exposed, you can actually read their competitive considerations and strategic thinking about their competitor models. This is also a general concern in alignment research, determining how a model behaves differently when it recognizes it is being tested, and the Claude model has even been seen to demonstrate "faking alignment" if their responses in that instance might affect its future alignment training.

One aspect I recognize when comparing the first naïve and second researched attempts, is that my revision instructions on the naïve attempts were so strictly followed that the models went too far, where a push to be more ethical made the letter too firm and possibly jeopardized the business relationship.

What would you revise before sending any of these letters?

I would adapt Claude's letter into a formal proposal paper that can serve as a working manual for our plan of action. I would use Gemini's draft as a cover letter to that proposal. While Claude demonstrated the most thoughtful and coherent plan, I believe that Gemini accomplished the most with the fewest words and best managed to approach the client as a trusted partner offering to help them without accusation. Being condensable to a single page, Gemini's letter would be more conveniently shareable within the client's company too, and so more people would read it. If a forensic investigation occurred in the future though about our firm's role or involvement, it's Claude's letter that I would want on record as it best demonstrates that we really mean what we say.

What risks do AI tools pose when writing ethically sensitive communication?

A key consideration is that the assignment instructions leave it unspecified whether the marketing campaign is to be paused or instead revised, leaving it up to the model or myself to decide the course of action. I was curious to see what kinds of solutions each model might pursue, and so I did not instruct a specific path but did lean into whichever they seemed compelled to pursue on their own. The risk is that we neglect to make intentional choices about our ethics and let the machine do it for us, then we simply end up obeying the ethics of the model's creator.

Would you trust AI to help with letters like this in your career? Why or why not?

Yes, I would trust AI to help with letters like this in my career, but most specifically in the preparatory background research. The web crawler for the beta deep research mode was able to read literally over a thousand websites at once and generated detailed reports to inform the process – something I would just never have the time to accomplish. I would still rather write the final letter myself, because my human brain is the product of a lot of evolution, and the children being exploited deserve to have the greatest care and attention my cognition can offer them.