ArguBot Arena: Prompt Engineering a Debate on Responsible AI

Assignment Objectives

- Understand ethical dilemmas in AI (e.g., bias, copyright, surveillance, fairness)
- Learn how prompt design influences LLM output/tone/reasoning and how to specify the role of the LLM
- Construct system-level prompts that configure an LLM's behavior, perspective, and argumentation style
- Explicitly provide an LLM with debate/dialogue context to be able to respond accordingly
- Reflect on the strengths and limitations of LLMs as participants in complex discourse

Assignment Tasks

1. **Choose a Topic:** Pick a topic, or be assigned one in class, from the curated ethics list (e.g., "*Do LLMs trained on copyrighted material violate fair use?*").
2. **Choose a Position:** Pick a "for" or "against" position, or you may be assigned to one in class (e.g. "*Yes, using copyrighted material to train an LLM is fair use*")
3. **Craft the Debate Prompt:** Using the provided Jupyter notebook template, create (system) prompts to configure the LLM to argue your chosen position. Additionally, the prompts should help the LLM to use appropriate tone, apply domain expertise, utilize a persuasive style, and to not deviate from the debate.
4. **Experiment with Prompts:** Refine and experiment with prompts to have an LLM respond appropriately for each of the two rounds in this simplified debate format.
5. **Provide Context to LLM:** To have the LLM be able to respond to debate arguments, you will need to provide it with its earlier output (note that more advanced LLM tools, e.g. LangChain, may simplify this, but those are not used in this assignment in order to see how to explicitly provide an LLM with context). Note that to experiment with the different context you will provide the LLM, you will likely want to also try prompting it to argue the opposing position that you were assigned.
6. **Reflect:** Using markdown cells in the Jupyter notebook, comment on how well the LLM is able to defend the chosen position. Did the LLM hallucinate or deviate from the scope of the debate? Did the arguments make sense?
7. **Submit Notebook:** Submit the Jupyter notebook and your two distinct prompts for each round of the debate.
8. **Judge Debates:** Your prompts, and those of the person assigned the opposite position for your debate topic, will be used to have two LLMs debate one another in class (using text-to-speech), and you will judge which LLM is the winner of the debate.

Assignment Rubric Criteria

- **(40%) Role Consistency:** LLM stays in character and consistently argues for the viewpoint/position assigned
- **(30%) Argument Validity:** LLM uses well-structured arguments that are supported by examples, facts, or analogies
- **(20%) Maintaining Context:** LLM sustains consistent arguments/points across rounds and responds to opponent’s points
- **(10%)** Judging and Reflection: A written evaluation assessing your LLM’s performance and voting on the winning LLM in the debate

Example Debate Topics

- Should AI ever make life-and-death decisions? (related: autonomous weapons, healthcare triage systems, etc.)
- Can AI be truly unbiased if it is trained on biased data? (related: Imagenet, Social media data)
- Do facial recognition systems do more harm than good? (related: mass surveillance, public safety)
- Do open-source AI models pose more risk than benefit? (related: deepfakes, malicious usage)
- Should there be a limit on how much AI can automate/displace human workers? (related: future of work, universal basic income)