Model Reflection & Limitations

When does the model perform well?

The GPT-2 model performs well when given open-ended, creative prompts such as storytelling or poetry. In prompts like *“Once upon a time, there was a robot who...”* or *“Continue the story of the rat and the mouse on a bridge,”* the model generated imaginative and narrative-rich content. It also succeeded in mimicking a conversational or stylistic tone, often producing output that felt human-like and fluid in style.

When does the model struggle?

The model struggles in several key areas:

- **Logical reasoning:** In prompts that required structured thinking (e.g., *“Explain gravity like I’m in 5th grade”*), the model’s response lacked clarity and often drifted into vague metaphors.
- **Accuracy on factual topics:** When asked to explain scientific or technical concepts, the model sometimes provided incomplete, misleading, or incorrect information.
- **Structured formats:** It failed to produce strict formatting when requested, such as generating a true haiku or adhering to grammar rules for younger reading levels.
- **Tone control:** Some completions unexpectedly included dark or inappropriate themes, particularly in open-ended storytelling (e.g., “Write a story about John and his wife Sarah”).

How might you improve the application?

To make the application safer, more reliable, and more useful:

- **Add output filtering** to detect and block violent, biased, or inappropriate content.
- **Use instruction-tuned models** (like GPT-3.5 or ChatGPT) instead of base GPT-2, which would improve coherence and alignment with prompt intent.
- **Validate facts** by integrating a basic fact-checking layer or linking to external sources when dealing with educational or technical questions.
- **Improve prompt handling** by pre-processing the input (e.g., detecting whether the user wants a poem, explanation, or story) and shaping it accordingly to guide model output.