AI Model Quality

can code communities

Review Prompt Engineering

- What is Prompt Engineering?
- Why is prompt engineering important?

Take Home Exercise

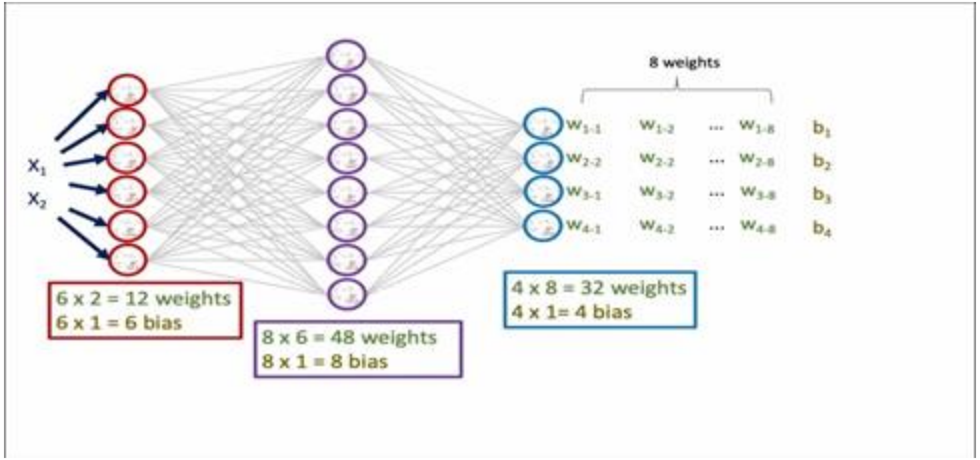
1. Use OpenAI Playground to build an AI assistant that generates a summary of a book.
2. Use the system message to focus on specific elements and properly format the response.
3. Save and share the preset link.

Model Quality

- Model quality is important to consider when using AI systems.
- Different models are better at different tasks.
- How do we choose which model to use?
    - Specifications
    - Evaluations

Parameter Size

- Parameter size is a model specification that helps users understand the size of the model.
- Parameters account for all the nodes and connections between nodes in the AI system.
    - The size of parameters in an AI model can range from 1 billion to over 1.75 trillion.



x_1
x_2
8 weights
w_{1-1} w_{1-2} ... w_{1-8} b_1
w_{2-2} w_{2-2} ... w_{2-8} b_2
w_{3-1} w_{3-2} ... w_{3-8} b_3
w_{4-1} w_{4-2} ... w_{4-8} b_4
6 x 2 = 12 weights
6 x 1 = 6 bias
8 x 6 = 48 weights
8 x 1 = 8 bias
4 x 8 = 32 weights
4 x 1= 4 bias

Parameter Size

- Many models have 2-3 model sizes available to choose from.
  - Bigger parameter sizes improve the strength of the model's responses, but it requires more computational power to run.
  - Smaller parameter sizes are sometimes viewed as more valuable when trying to build task specific tools with limited resources.

Context Length

- Context length is the amount of data that a model can process and store in memory at a given time.
- The context length is measured in units of tokens which are numerical representations of the data that the model is dealing with.
    - What are tokens?

Tokens

- In a normal chat bot, like Chat GPT, the user prompts the model with text
- The model converts (tokenizes) the text into numerical representations of the text (tokens).
- Then the tokens get processed through the model, where a bunch of computations happen, and the model generates a set of response tokens.
- These output tokens are tokenized and turned back into text to be returned to the user.

Context Length

Depending on the model, context length can be an issue with tasks that require a lot of data processing.

- GPT 3.5 Turbo, which is the model that runs Chat GPT (free version) has a 16k context length.
    - That’s around 20 pages of text for reference.
- OpenAI has more recent models, like GPT 4o, which have a 128k context window
- Recently Google has announced their newest model, Gemini 1.5 Pro, which will have a context length of 2 million tokens.

Why are there context length limits?

- There is no limit computationally, but the quality of responses decreases to a level that isn't satisfactory to the user.
- Google tested their 2 million limit to the context window size by using the NeedleInAHaystack benchmark test.
    - https://github.com/gkamradt/LLMTest_NeedleInAHaystack

Benchmarks

- Benchmark tests on AI models help accurately describe the quality of a model in a specific bounds.
  - There are a wide variety of benchmarks, but there are some common ones used in most model's release notes.
- Review GPT 4 and Gemini's benchmark results
  - https://openai.com/research/gpt-4
  - https://blog.google/technology/ai/google-gemini-ai/#performance
- HuggingFace's Open LLM Leaderboard
  - https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Benchmarks

- Benchmarks are essentially a set of prompts (specific to a topic/task) with expected answers.
  - The prompts are passed through the model and the model’s response is then graded based on the expected response.
  - https://github.com/gkamradt/LLMTest_NeedleInAHaystack/blob/main/tests/test_evaluators.py

Benchmark Critical Thinking

- It is always important to be a bit critical of the provided benchmark evaluations, because they are used as a selling point.
  - What topic is this benchmark focused on?
  - What does the benchmark test?
  - How different are the scores compared to other models?
  - Does this difference show through personal experience?
- Google’s Gemini was initially very controversial for misrepresenting the model’s features and performance.
  - In their promo video: “Sequences shortened throughout”
  - Comparing GPT 4 to Gemini Ultra before Ultra had been released.
  - “Previous SOTA model listed when capability is not supported in GPT 4V”

Review

How do different models affect our response?

- Parameter size
  - Bigger isn’t always better. Smaller parameter sizes can cut costs and computation while performing better for task specific use.
- Context length
  - Simple short (under a paragraph or two) prompts will work fine for almost any model. If you’re looking to add a large amount of data for the model to reference, not all models can support that (depending on data size).
- Benchmarks
  - Models with better benchmark evaluations will perform better. For specific tasks, check for task specific benchmarks.

Take Home Exercise

- Evaluate/compare two AI models (GPT 4 and Claude 3 Sonnet) based on context length and benchmarks.
- Which one is best suited to help with programming related questions?