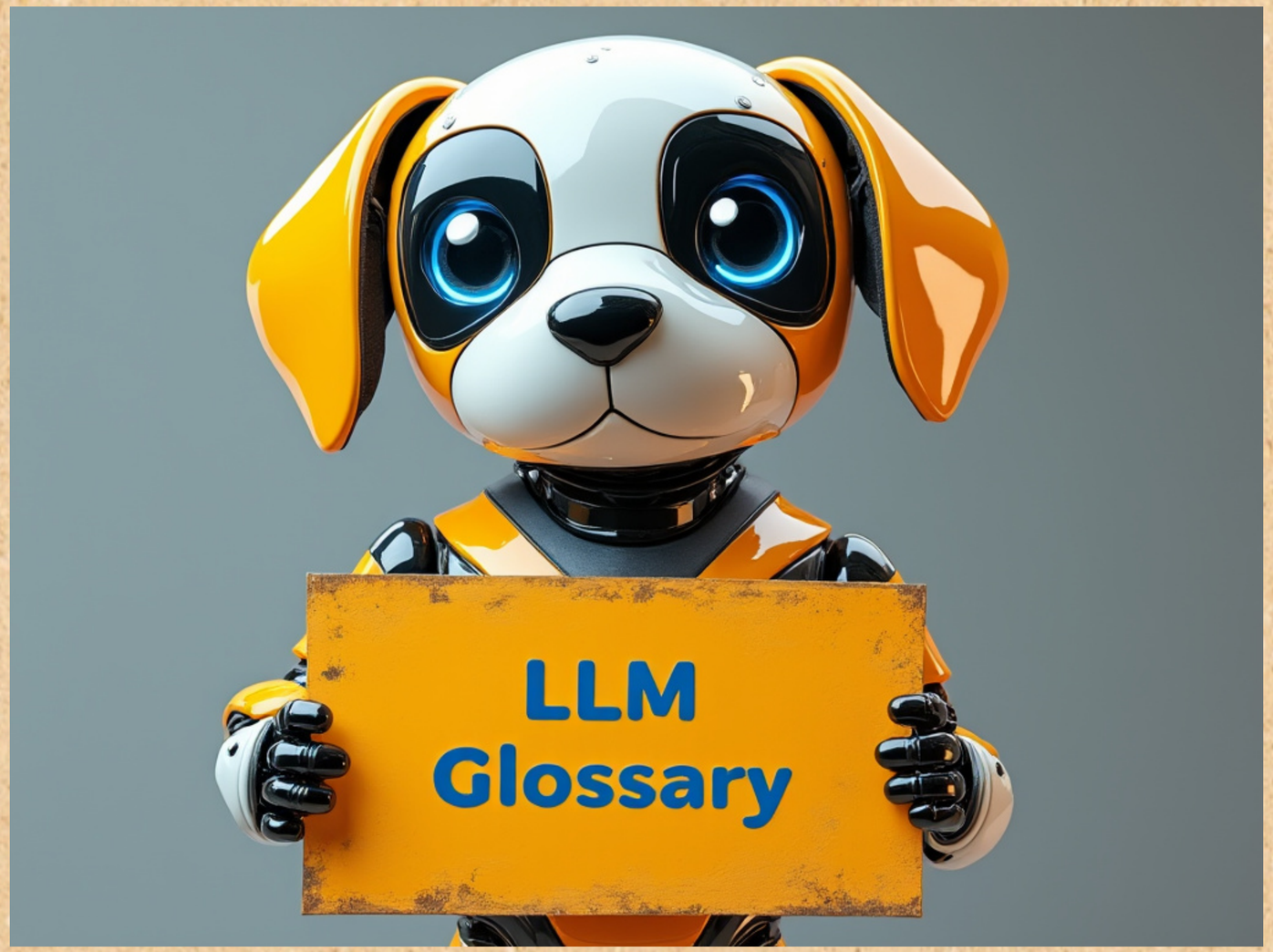
Module 1 - Topic 2

2.2 LLM Glossary


LLM
Glossary

LLM Glossary

- When working with LLMs, you will encounter many terms like

  - Prompt
  - Tokens
  - Context Length
  - Temperature

- Pricing of LLMs

Prompt

- The input given to the model to generate a response (question, statement, instruction).
- Prompts shape the model's output, and better prompts lead to better results.
- Can range from simple questions to detailed instructions for complex tasks.
- **Example**: Asking "Summarize the novel *1984*" will yield different results compared to "What is the main theme of *1984*?"

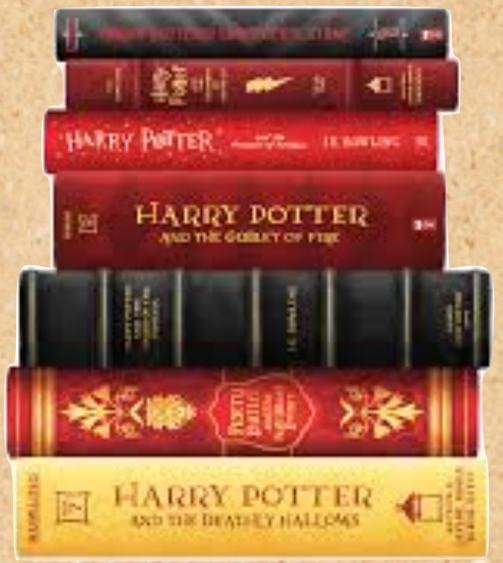
Tokens

- Tokens are the units of text processed by the model (words, parts of words, or punctuation).
- Text is broken into tokens, with each token contributing to the total input.
- 1 token ~ 0.75 words (approximation)
- **Example**: humanistic is a single word but it broken down into 2 tokens - ‘human’ and ‘istic’ tokens

platform.openai.com/tokenizer

Context Length

- Refers to the maximum number of tokens the model can consider at once.
- Longer context lengths allow the model to retain more information over extended text.
- **Example**: GPT-4o has a context length of 128k tokens, while Gemini models have 2 million tokens!!



Temperature

- Controls the randomness and creativity of the model’s responses.
- Lower values produce focused, deterministic output.
- Higher values encourage diverse and more creative responses.
- **Example**: Use a high value (close to 1) for creative tasks and a low value (close to 0) for precise tasks.

top-p, top-k, min-p

- **Top-k** limits the model to choose from the k most likely next tokens, ensuring only the top k probable options are considered.

Example: If k=10, the model will choose from the top 10 most likely tokens, making the output more predictable.

- **Top-p** selects from tokens whose cumulative probability equals a specified percentage, allowing the model to choose from a variable number of tokens.

Example: If p=0.9, the model selects tokens until their combined probability is 90%, introducing more variety.

- **Min-p** ensures the model picks tokens above a certain minimum probability threshold, filtering out less likely options.

Example: If min-p=0.05, any token with a probability lower than 5% will not be considered, ensuring higher-quality choices.

Pricing

- LLMs typically charge based on the number of tokens processed.
- Generally, pricing for input tokens and output tokens is different.
- Larger models are more expensive, while smaller models are less costly due to lower computational requirements.
- **Example:** As of Oct '24, GPT-4o charges $2.5 per million input tokens, and $10 per million output tokens.

platform.openai.com/pricing