REVATURE

LLMs

Agenda

01 GenAI and NLP

02 LLMs and Text

03 Prompt Design and Engineering

04 Hands On!!

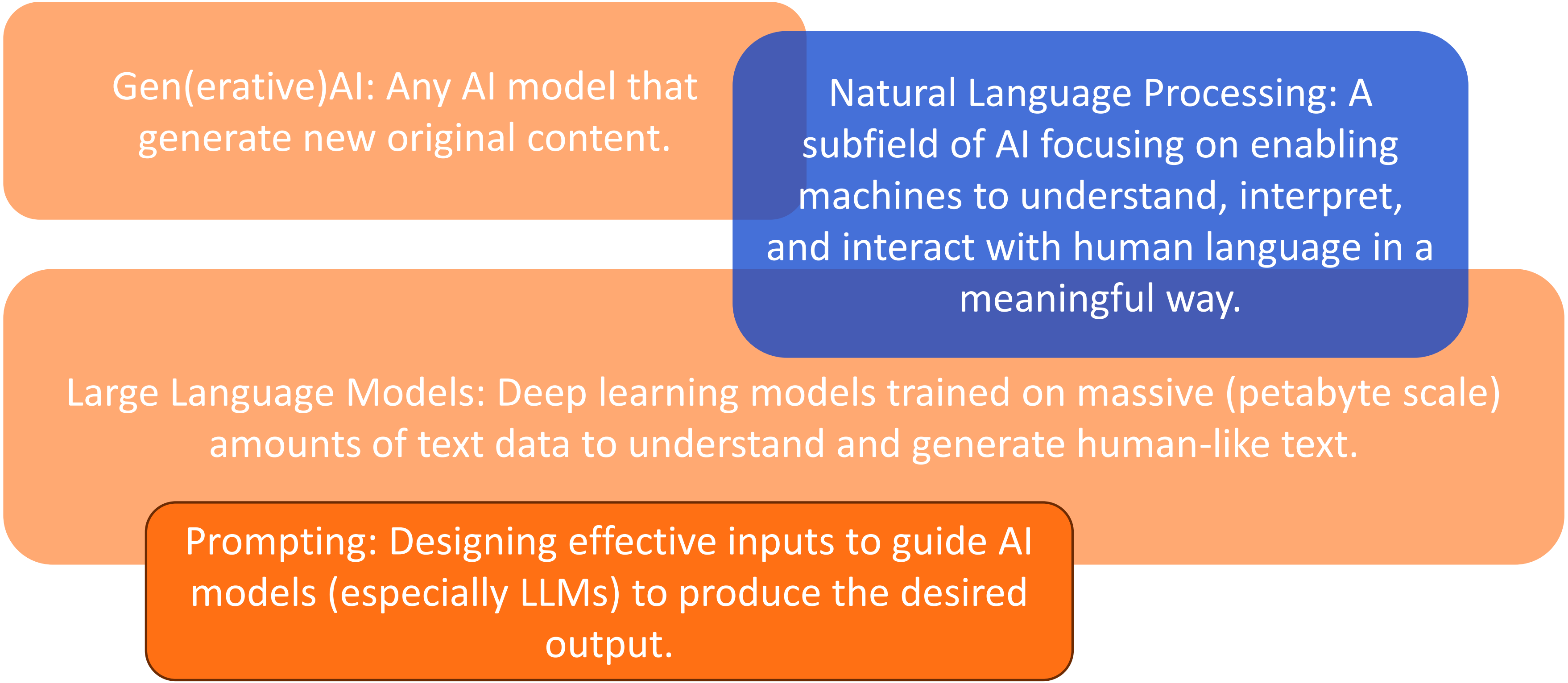
05 Summary

REVATURE

GenAI and NLP

REVATURE

GenAI and NLP



Gen(erative)AI: Any AI model that generate new original content.

Natural Language Processing: A subfield of AI focusing on enabling machines to understand, interpret, and interact with human language in a meaningful way.

Large Language Models: Deep learning models trained on massive (petabyte scale) amounts of text data to understand and generate human-like text.

Prompting: Designing effective inputs to guide AI models (especially LLMs) to produce the desired output.

LLMs and Text

REVATURE

LLMs and Text

How to…

Most LLMs use a Transformer architecture.

Transformers are a type of Neural Network that use self-attention to identify importance. Self-attention allows for parallel input producing faster processing and long-range dependencies.

Transformers have Encoders which process an input and analyzes the relationships between words, and Decoders which generate a series of tokens/words.

…make it work

Encoder Only – Best for understanding tasks, like text classification or sentiment analysis.

Decoder Only – Best for generative tasks like creative writing or coding.

Encoder-Decoder – Best for understanding and generative tasks, like summarization or translations

LLMs and Text

What does…

LLMs will (generally) fall into one of three types: Generic, Instruction-Tuned, or Dialogue-Tuned.

Generic LLMs are trained without specific optimization. They are designed to predict the next word in a sequence based on the context. Generic LLMs lack task-specific alignment (so may have trouble following detailed instructions) or be a bit verbose. The trade off is their broad general knowledge (based on all of their training data) and their ability to perform a wider variety of tasks (which requires careful prompting).

… it all mean?

Instruction-Tuned LLMs are fine-tuned to allow the AI to follow an explicit instruction and generate concise, relevant, task-specific results. These LLMs will likely have greater accuracy and relevance for tasks like summarization or analyzing sentiment.

Dialogue-Tuned LLMs are optimized further (on top of Instruction-Tuned models) for conversation. Making them more capable of engaging in natural and coherent interactions with users. These are best suited to tasks like chat-bots or customer service which prioritize interaction.

Prompt Design and Engineering

REVATURE

Prompt Design and Engineering

The Basics

Prompt Design: giving instructions and context to a language model to achieve a desired task.

Zero-Shot prompting is an input with a desired output with no examples or prior context to guide it. Zero-shots depend entirely on the models knowledge and training.

Tokens, the structure of inputs and outputs for an AI are a common limiting factor in prompting. Most models will have an interaction limit, which includes both input and output token counts for an exchange.

The Not-So Basic

Few-Shot prompting includes a few examples in the prompt to guide the AI on how to complete a task.

Instruction prompting includes providing clear, specific, detailed instructions about what you want it to do.

Role prompting involves assigning a role or persona to the AI to guide its behavior and outputs

Prompt Design and Engineering

The Even Less Basic…

Chain-Of-Thought prompting requires the AI to explicitly outline its reasoning or asks it to “think step-by-step” though a problem.

Bias awareness allows us to frame a prompt to avoid or purposefully introduce bias in a response.

Few-Shot with In-Context Learning adds examples (context) for the AI to infer (to learn) a desired result or format.

And The Fun Stuff.

Iterative prompting allows you to refine your prompt by observing the AIs response, and adjusting the input for clarity or detail until the model produces the desired result.

Temperature settings control how creative or deterministic the responses will be. Higher temperatures excel at story-telling or brainstorming, while low temperatures are best for tasks which require high accuracy.

Prompt chaining aids AIs ability to handle complex workflows by breaking down large tasks into smaller sub-tasks. Prompting the model with each sub-task, then using the output as a component of the input for the next sub-task.

Prompt Engineering: the practice of developing and optimizing prompts to efficiently use language models.

Hands On!!

REVATURE

Hands On!!

Pair-Programming: a development technique where one member of a pair is the “driver” and writes the required code while the other member is the “navigator”, directing next steps, providing research, and allowing the “driver” to stay focused on the immediate task of writing.

Overview

- **Access:** Each group should verify that there is access to the required tools within the group.
- **Prompts:** Using the provided prompts as a starting point, experiment with different prompting techniques.

  Testing both AIs, try at least 5 different prompts, with at least 4 variations or additional different techniques for each prompt
- **Results:** Track outputs from each prompt, noting any techniques your group leveraged and any notable results you gathered.

REVATURE

Summary

REVATURE

Summary

Take Aways?

What prompts did you provide to the AI?

What differences did you see in outputs?

What techniques did you leverage to develop your prompt?

What changed to the outputs due to those developments?

Thank You!

REVATURE