Introduction to Generative AI Concepts & Development

This chapter from our popular 1-day Introduction to Generative AI Concepts & Development course gives you a taste of the valuable insights and practical knowledge shared in the full instructor-led, hands-on course.

This chapter covers:

- ✓ The concept of generative AI
- ✓ How generative AI processes data
- ✓ The concept of embeddings and context
- ✓ Interpreting generative AI use cases

Let's get started!

Understanding Generative AI

What is Intelligence?

» Intelligence derives from the Latin word Intellectus
  - Which itself derives from Intellegō
  - Inter ("between") + Legō ("select")

» Many philosophers preferred "understanding" to "intellect."
  - e.g. Bacon, Hobbes, Locke, Hume

What is Artificial Intelligence?

» **Artificial Intelligence (AI)** has lots of definition, that tend to align
  - Thinking Humanly Thinking Rationally
  - Acting Humanly Acting Rationally

» Artificial Intelligence (AI) is the field of research investigating methods that enable machines to exhibit *Goal-directed adaptive behavior*.

What can AI do?

» Assign input data to classes — **Classification**

- *e.g.,* spam/not spam

» Forecast or estimate a variable — **Regression**

- *e.g.,* stock prices, sales forecast

» Group data into categories — **Clustering**

- *e.g.,* organizing documents into groups

» Generate new samples — **Generative**

- *e.g.,* creating an image of a dog riding a skateboard

» Choose the “best” action — **Reinforcement**

- *e.g.,* opening a door to get inside

Predictive AI

Predictive AI uses historical data to predict future outcomes or classify new inputs.

- **Key characteristics**: Mapping conditional probabilities, focused on specific prediction tasks.
- **Applications**: Weather forecasting, stock price prediction, medical diagnosis, spam detection.
- **Strengths**: Often more interpretable, requires less computationalpower.

Generative AI

» Generative AI is a type of AI that generates new samples fromexisting data.

- It can create new images, text, audio, and other types of data.
- Generative AI models are trained on large datasets to learn patterns and generate new samples.
- They can be used for creative tasks like art generation, music composition, and text generation.
- This chapter focuses on LLMs and multi-modal LLMs.

How does a Large Language Model work?

» **Large Language Models (LLMs)** are a type of generative AI model that generates text.

1. Input text is **tokenized** — broken down into units that are represented as an ID.
2. The tokens are **embedded** — converted into a list of numbers that represent their meaning.
3. The model predicts tokens based on the input tokens.
4. The model stops when there’s a special token indicating the end of
5. the text.
6. Generated tokens are converted back into text.

- This is an accurate, but simplified explanation.

Tokenization

» **Tokenization** is the process of breaking down text into smaller units called **tokens**.
  - Tokens can be words, characters, or subwords.
    - Characters can represent all words but aren't useful to understand meaning.
    - Words are better, but many are rare or like others (e.g.,"run" and "running").
    - Subwords are a compromise, breaking words into smaller units that can be combined in different ways.

» Tokenization converts text into numbers that can be processed by a model.

» The tokens used by a model are called its vocabulary.

Example Tokenization: GPT-4o

Input Text:

» Every artist is a cannibal, every poet is a thief.

Tokens:

» Every | artist| is| a| cann|ibal|,| every| poet| is| a| thief|.
  - Tokens are separated by |.
  - Notice how "cannibal" is broken into "cann" and "ibal".

Token IDs:

» 1 5 7 4 5, 1 2337, 382, 261, 21511, 63126, 11, 1753, 48961, 382, 2 61, 106886, 13
  - Each token is represented by a unique ID.
  - Notice how "is" is repeated and both have 382 as their ID.

Embeddings

» **Embeddings** are a way to represent tokens as a series of real numbers (i.e. floats).

» They capture the meaning of tokens in a way that a model can understand.

» Embeddings are learned during training and are unique to each token.

» Math on embeddings can reveal relationships between tokens.
  - Man is to king as woman is to queen.
  - < K i n g > - <Man> + <Woman> ~= <Queen>

What are embeddings doing?

» Words have more than one meaning.
  - "Bank" can be a financial institution or the side of a river.
  - "Bat" can be a flying mammal or a sports equipment.

» These meanings can be inferred from context.
  - "I went to the bank to deposit my paycheck."
  - "I swung the bat and hit a home run."

» Embeddings capture all the meanings of a word, even those that don't appear in a dictionary.
  - e.g., "kamala IS brat" (Charli XCX)

Why are embeddings important?

» Think about a good memory you have.

- Is the memory singular or an amalgamation of many small moments?

» A memory is like an embedding for a model.

- It's a collection of information that helps you remember the holistic context of a moment.
- It informs how you react to similar situations in the future.

» Embeddings...

- Help models understand the holistic context of tokens.
- Inform how a model reacts to similar tokens in the future.

- Scientists have been able to identify strong similarities between word embeddings and MRI scans of the brain.

Predicting the Next Token

» With embeddings the model predicts possible next tokens.

- This is where the math gets complicated.

» The actual next token is sampled randomly using different strategies.

- **Greedy sampling**: Picks token with the highest probability.
- **Top-k sampling**: Picks the top k most likely tokens.
- **Top-p sampling**: Picks the tokens with a total probability lessthan p.

» The model stops when it predicts a special token indicating the end of the text.

Input: It ' s a beautiful day, don ' t let it get

Predicted Tokens:

 # Prob Token

0 21.18% you

1 18.89% to

2 15.91% in

3 5.83% any

4 5.76% too

5 3.52% away

LLM Settings

Setting Description

Temperature Controls the randomness of the model's predictions.

Top P Controls the number of tokens considered during sampling. It includes tokens with a cumulative probability less than p. Top K Controls the number of tokens considered during sampling. It includes the top k most likely tokens.

» Max Tokens Limits the number of tokens the model can generate.

» Frequency

» Penalty

» Penalizes tokens that have already been generated.

- Lower temperatures produce more predictable text.
- Top P and Top K control the diversity of the text.
- Usually only one is used, with Top P being more popular.
- As a rule-of-thumb, only modify temperature or Top P/K, not both.

How are LLMs trained?

LLMs go through several stages of training before they are released.

1. Pre-training

- The model is trained on large sets of text.
- The model learns to predict the next token in a sequence.
- It minimizes the difference between its predictions and the
- actual next token.

2. Fine-tuning

- The model is fine-tuned on a smaller dataset for a specific task, like instruction following.
- Fine-tuning helps the model learn to be useful for a specific task.

3. Safety Training

- The model is trained to avoid generating harmful or inappropriate content.

How do multi-modal models work?

» **Multi-modal models** can generate text, images, and other typesnof data.

- Images are tokenized by cutting them into patches — like a
- puzzle with square pieces.
- Audio is tokenized by breaking it into smaller pieces — like a
- spectrogram.
- Their embeddings are generated using a smaller neural
- network.

» They usually produce media in one of two ways:

- A separate output (“head”) generates each type of media.
- The model predicts the next “token” of media until it reaches the end of the image or audio.

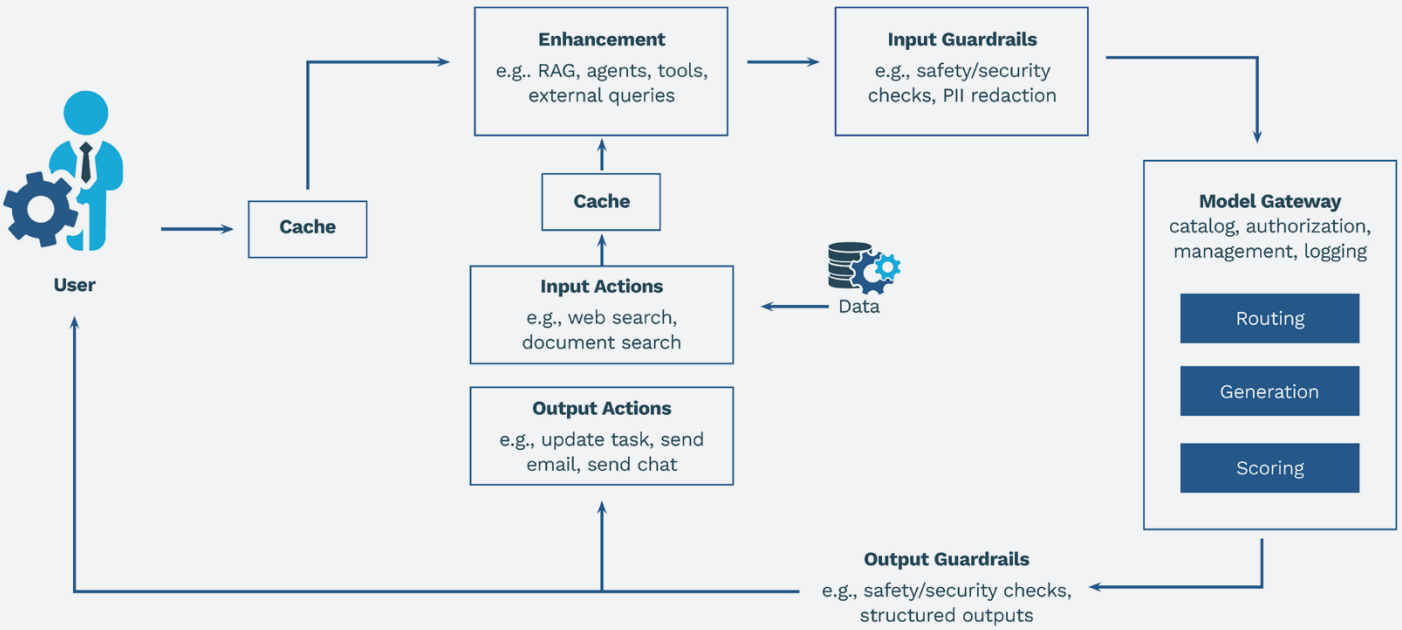
What does a full Generative AI architecture look like?

» GenAI applications are built on a complex architecture.

» The components vary, but this is a rough reference.

» Not all applications use all components, not all components are shown.

What does a GenAI architecture look like?



Reflection

» What are similarities between Generative AI and the human brain?

» What are differences between Generative AI and the human brain?

» What areas of Generative AI are you most interested in exploring

» further?

» What do you think limits the capabilities of Generative AI?

Get Hands-On with Generative AI

Want to dive deeper into Generative AI? Browse our Generative AI training courses for practical hands-on labs, expert instruction, and the opportunity to explore advanced topics. We're excited to help you get started mastering the full potential of Generative AI with customized training for your team or organization!