GENERATIVE AI FOR EVERYONE

Supervised learning

Supervised learning, also known as labeling things, is a fundamental machine learning technique that involves training an AI model on labeled data. In supervised learning, the AI model is presented with a dataset of input data and corresponding output labels, which represent the desired results for each data point. The model's goal is to learn the mapping between the input data and the output labels, enabling it to make accurate predictions on new, unlabeled data.

How LLMs work

Large language models (LLMs) are trained using a technique called supervised learning, where they are exposed to a massive dataset of text and code and tasked with predicting the next word in a sequence. This process is repeated over millions or even billions of examples, allowing the LLM to learn the statistical relationships between words and phrases

Tokenization

The text data is broken down into individual tokens, which could be words, characters, or subwords. This allows the LLM to process the data in a structured and consistent manner.
Note: (Assume 1 token = 3/4 words, or equivalently, 1 word \approx. 1.333 tokens)

Embedding

Each token is represented as a numerical vector, known as an embedding. These embeddings capture the semantic and contextual relationships between words, allowing the LLM to understand the meaning of language.

Gen AI Applications

Writing
Article writing, copywriting, brainstorming, recipes, or anything in general.
Reading
Summarization, Proofreading, Translation, and much more
Chatting
General Chatbots, specialized chatbots, and much more

Customer Support Workflows

- Humans only
- Bots support humans (human in the loop)
- Bot triages for humans
- Chatbots only

Written By **Saboor Hussain**

Deploy Chatbot

Start with an internal facing chatbot.
Deploy with human in the loop to check for mistakes.
Only after deemed safe, allow bot to communicate directly with customers.

LLM's Limitation

An LLM's knowledge of the world is frozen at the time of its training.
LLMs sometimes hallucinate and make up things.
The input and output length is limited.
An LLM's context length is the limit on the total input + output size.
Generative AI does not work well with structured (tabular) data.
An LLM can reflect the biases that exist in the text it learned from.
Some LLMs can output toxic or other harmful speech.

Tips for prompting

- Be detailed and specific.
- Guide the model to think through its answer.
- Experiment and iterate.

Image generation

Image generation models often rely on diffusion models, a type of machine learning technique that learns the underlying patterns and characteristics of images. Diffusion models can be considered a form of supervised learning, as they are trained on a large dataset of labeled images. During training, the model learns to progressively remove noise from a corrupted image until it reconstructs the original, clean image. This process allows the model to learn the intricate details and relationships between pixels, enabling it to generate new images that are similar to those in the training set.

Lifecycle of Gen AI project

1. Scope Project
2. Build/improve system
3. Internal evaluation
4. Deploy and monitor

Improve Gen AI Model

Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

- Prompting
- Retrieval augmented generation (RAG)

Give LLM access to external data sources

- Fine tune models
- Pretrain models (trains LLM from scratch)

RAG

Retrieval-augmented generation (RAG) is an innovative technique that enhances the capabilities of large language models (LLMs) by incorporating external knowledge sources. Unlike traditional LLMs that rely solely on their internal knowledge base, RAG models leverage external retrieval systems to access relevant information and incorporate it into their generation process.

The implementation of RAG typically involves two main components:

- Retrieval System: A retrieval system identifies and retrieves relevant information from external sources, such as documents, databases, or knowledge bases.
- Generation Model: A generation model, typically an LLM, utilizes the retrieved information to generate a response. This may involve techniques such as incorporating retrieved passages into the prompt or using retrieved information to inform the generation process.

Examples of RAG Applications

Chat with PDF files

Answer questions based on a website's articles

Fine-tuning

Fine-tuning is the process of adjusting an LLM's parameters to improve its performance on a specific task.

How LLM follow instructions

Large language models (LLMs) can be fine-tuned with additional data to enhance their ability to follow prompt instructions. Instead of simply generating the next word in a sequence, LLMs trained with such data can comprehend and execute instructions effectively. This refinement enables LLMs to perform more complex tasks that require understanding and responding to specific directives.

Written By **Saboor Hussain**

RHLF

Reinforcement learning from human feedback (RLHF) is a technique that fine-tunes large language models (LLMs) by incorporating human input. In this approach, humans provide scores or rewards to different responses generated by the LLM, guiding the model towards producing more desirable outputs. This iterative process helps the LLM learn the nuances of human preferences and improve its overall performance.

Tools

Tools can be used by large language models (LLMs) to enhance their functionality and versatility.

Agents

Agents play a crucial role in enabling large language models (LLMs) to effectively interact with the environment and perform tasks. By employing agents, LLMs can transition from passive language understanding to active decision-making and execution.

Agents act as intermediaries between LLMs and the external world, deciphering instructions received from the LLM and translating them into concrete actions. This allows LLMs to operate in a dynamic and interactive manner, adapting their behavior based on real-time feedback and observations.

Supervised learning (labeling things)

Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Ad, user info	Click? (0/1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
X-ray image	Diagnosis	Healthcare
Image of phone	Defect? (0/1)	Visual inspection
Audio recording	Text transcript	Speech recognition
Restaurant reviews	Sentiment (pos/neg)	

How much does it cost?

Example prices

	OpenAI/GPT3.5	OpenAI/GPT4	Google/PaLM 2	Amazon/Titan Lite
Input tokens	$0.0015/1K tokens	$0.03/1K tokens	$0.0005/1K tokens	$0.0003/1K tokens
Output tokens	$0.002/1K tokens	$0.06/1K tokens	$0.0005/1K tokens	$0.0004/1K tokens

What is a token?

the example Andrew — 1 token

translate programming — 2 tokens

tonkotsu — 4 tokens

300 words
400 tokens

Roughly, 1 token = 3/4 words

Model size

1B parameters:	Pattern matching and basic knowledge of the world.	Restaurant review sentiment
10B parameters:	Greater world knowledge. Can follow basic instructions.	Food order chatbot
100B+ parameters:	Rich world knowledge. Complex reasoning.	Brainstorming partner

Written By **Saboor Hussain**

How do chat systems learn to follow instructions?

Fine-tuning

Help me brainstorm some fun museums to visit in Bogotá.

Sure, here are some suggestions: [...]

Input (A)	Output (B)
Help me brainstorm some fun museums to visit in Bogotá.	Sure,
Help me brainstorm some fun museums to visit in Bogotá. Sure,	here
Help me brainstorm some fun museums to visit in Bogotá. Sure, here	are
Help me brainstorm some fun museums to visit in Bogotá. Sure, here are	some
Help me brainstorm some fun museums to visit in Bogotá. Sure, here are some	suggestions:
...	...

Agents

- Use LLM to choose and carry out complex sequences of actions
- Cutting edge area of AI research

Help me research BetterBurgers top competitors

Here are steps I need to carry out to research competitors:
1. Search top competitors
2. Visit web site of each competitor
3. For each competitor, write summary based on homepage content

SEARCH("BetterBurgers Competitors")

VISIT(http://www.fastburger.com)

VISIT(http://www.burgerworld.com)

...

Summarize the following text:

At Fast Burger, we pride ourselves on the speed of our delivery ...

Written By **Saboor Hussain**

Identifying automation opportunities

- AI doesn't automate jobs. It automates tasks.
- Most jobs involve a collection of many tasks.
- Example: Customer service representative

Tasks	Generative AI potential
Answer inbound phone calls from customers	Low
Answer customer chat queries	High
Check status of customer orders	Medium
Keep records of customer interactions	High
Assess accuracy of customer complaints	Low
[...]	[...]

Evaluating AI potential

The potential for augmenting/automating a task depends on: (i) Technical feasibility and (ii) Business value.

Technical feasibility: Can AI do it?

- Can a fresh college graduate following the instructions in a prompt complete the task?
- If unsure, try prompting an LLM to see if you can get it to do it.
- An AI engineer can also help assess if RAG, fine-tuning, or other techniques can help.

Business value: How valuable is it for AI to augment or automate this task?

- How much time is spent on this task?
- Does doing this task significantly faster, cheaper or more consistently create substantial value?

Written By **Saboor Hussain**

Common roles

- Software engineer
  - Responsible for writing software application
  - Ideally someone who has learned basics of LLMs/prompting
- Machine learning engineer
  - Responsible for implementing AI system
  - Ideally familiar with LLMs/prompting, RAG, fine-tuning
- Product manager
  - Responsible for identifying and scoping the project
- Prompt engineer?
  - Usually not hired as a dedicated role