Learn and Grow

Generative AI LangChain

ASHIMA MALIK

ashimamalik58@gmail.com

Unleash the Power of AI: A Cutting-Edge Program - Learn with fun!

01 Master the application of AI in business

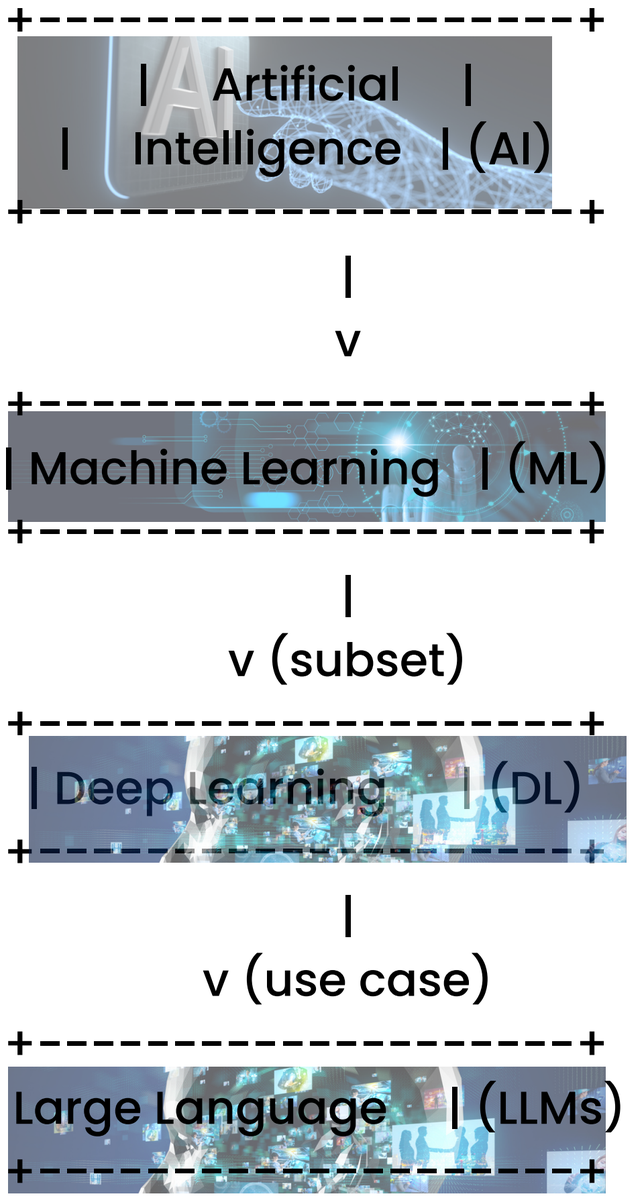
02 Become an expert in AI, a field shaping the future

03 Learn how to integrate AI into real-world business scenarios

04 Develop high-impact AI applications

05 Advance your career and join a community of professionals prepared to navigate the AI revolution

ASHIMA MALIK



Artificial Intelligence (AI) is the broadest concept, encompassing any machine that can exhibit human-like intelligence. This includes areas like robotics, computer vision, and natural language processing (NLP).

Machine Learning (ML) is a subfield of AI that focuses on algorithms that can learn from data without explicit programming. ML models can improve their performance over time as they are exposed to more data.

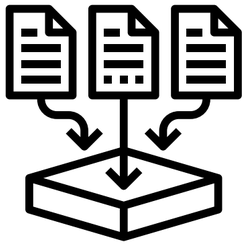
Deep Learning (DL) is a subfield of ML that uses artificial neural networks with many layers to process data. These complex models are particularly well-suited for tasks like image recognition and natural language processing.

Large Language Models (LLMs) are a type of AI model trained on massive amounts of text data. They can be used for a variety of NLP tasks, such as generating text, translating languages, and writing different kinds of creative content.

LLM

LARGE

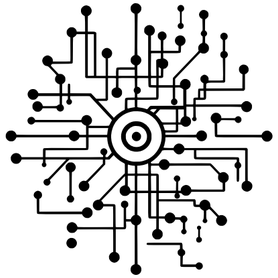
SIZE OF THE ARCHITECTURE



VAST DATA

CAPTURE MORE COMPLEX PATTERNS AND RELATIONSHIPS WITHIN THE LANGUAGE

LANGUAGE MODELS



ALGORITHMS OR SYSTEM THAT ARE TRAINED TO UNDERSTAND AND GENERATE HUMANLIKE TEXT

ASHIMA MALIK

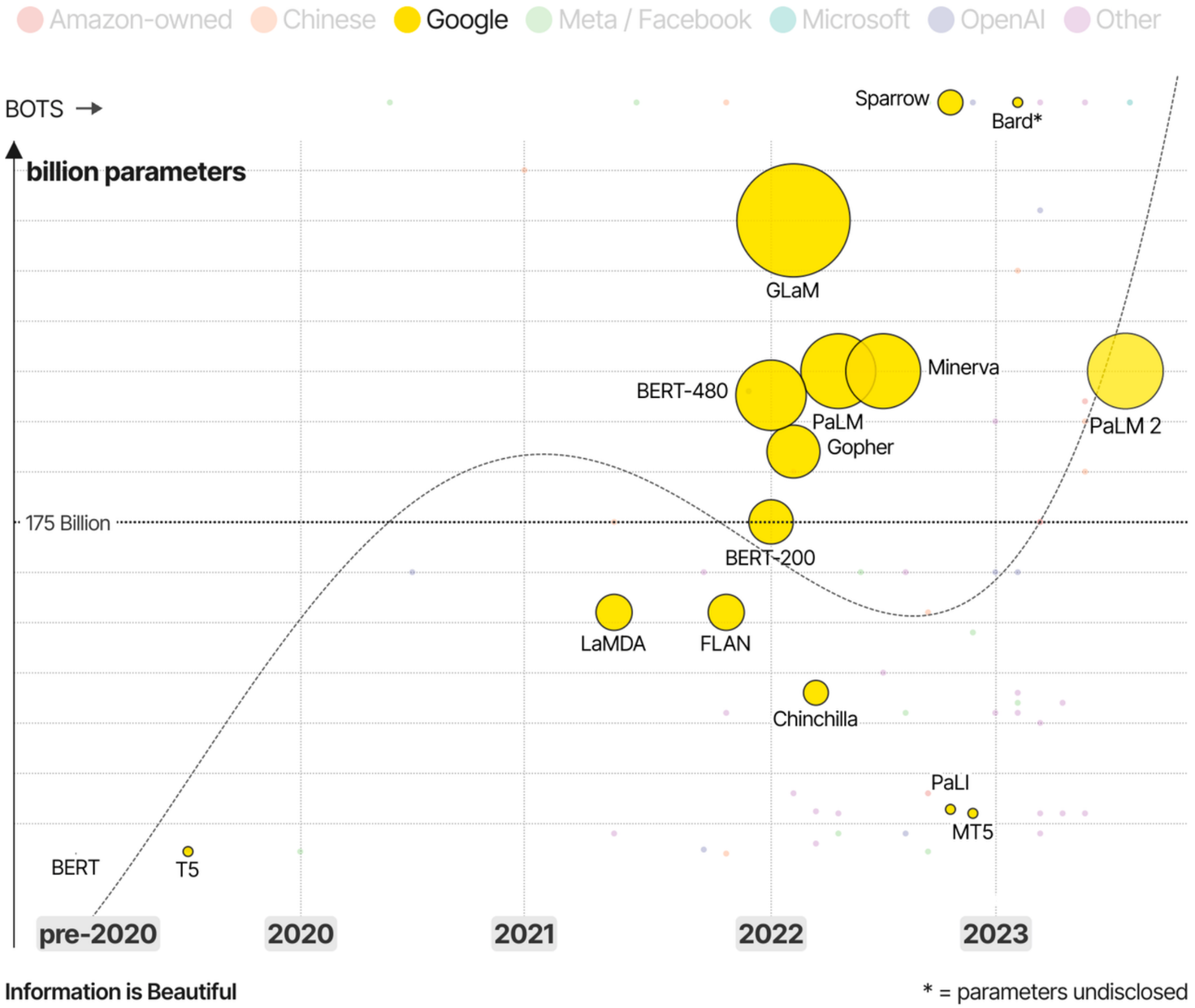


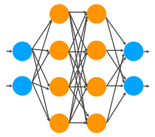
Image Source: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

TRAINING LLMS

1. Data Preparation

a)Massive Text Corpus- LLMs are trained on enormous amounts of text data. This data can come from books, articles, code, web crawls, and other sources. The quality and diversity of the data significantly impact the LLM's performance.
b) Preprocessing - The raw text data needs cleaning and preprocessing. This may involve removing irrelevant characters, splitting text into tokens (words or sub-words), and potentially applying techniques like stemming or lemmatization to normalize words.

2. Model Architecture



a) Neural Networks: LLMs are typically built using complex neural network architectures like transformers. These networks consist of multiple layers that process and learn patterns from the input text data.
b) Parameters: Each layer in the neural network contains a large number of parameters (weights and biases) that the model adjusts during training. These parameters essentially capture the relationships and patterns learned from the data.

ASHIMA MALIK

TRAINING LLMS

3.Training Process

a) Supervised Learning: LLM is presented with input text and a desired output (like the next word in a sequence or a complete sentence). The model compares its generated output to the desired output and adjusts its internal parameters to minimize the difference. This process is repeated through many iterations over the massive training dataset.
b) Loss Function: A loss function quantifies the difference between the LLM's output and the desired output. The model iteratively adjusts its parameters to minimize this loss value.
c) Optimization Algorithms: Algorithms like gradient descent are used to update the LLM's parameters based on the calculated loss. These algorithms guide the model towards learning the underlying patterns and relationships in the data.

4. Fine-Tuning

a) Specific Tasks: Once a large LLM is pre-trained on a general corpus, it can be further fine-tuned for specific tasks. This involves training the model on additional data relevant to the target task, potentially adjusting only a smaller subset of parameters.
b) Reduced Training Time: Fine-tuning leverages the knowledge already learned by the pre-trained LLM, making it faster and more efficient to adapt the model for new tasks.

ASHIMA MALIK

TRAINING LLMS

5. Evaluation 📋

Text Generation:

a) Perplexity: Measures how well the model predicts the next word in a sequence. Lower perplexity indicates better performance.
b) BLEU Score (Bilingual Evaluation Understudy): Compares generated text to reference translations, evaluating fluency and correctness.
c) ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation): Similar to BLEU, but focuses on identifying overlapping n-grams (sequences of words) between generated text and references.
d) Human Evaluation: Expert human judgment remains valuable in assessing the overall quality, coherence, and naturalness of generated text.

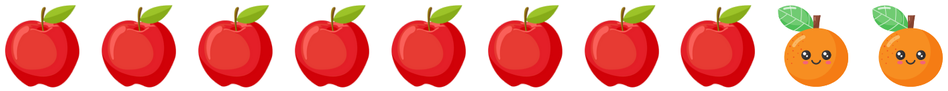
Machine Translation:

a) BLEU Score (as mentioned above): Commonly used for machine translation, comparing the generated translations to human-created references.
b) Human Evaluation: Similar to text generation, human evaluation plays a crucial role in assessing fluency, accuracy, and naturalness of the translated text.

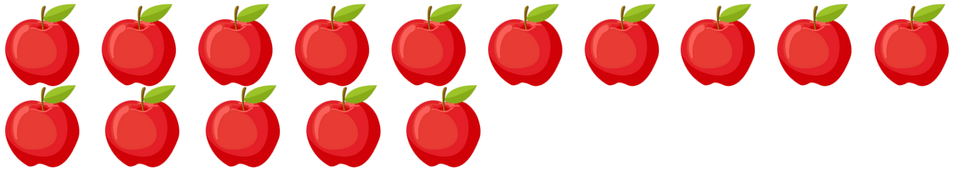
Question Answering:

a) Accuracy: The percentage of questions where the LLM provides a correct answer, considering both factual accuracy and reasoning.
b) F1 Score: A combined metric of precision (percentage of correct answers among retrieved answers) and recall (percentage of correct answers retrieved out of all possible answers).
c) SQUAD (Stanford Question Answering Dataset): A standardized benchmark dataset for evaluating question-answering models on factual topics.

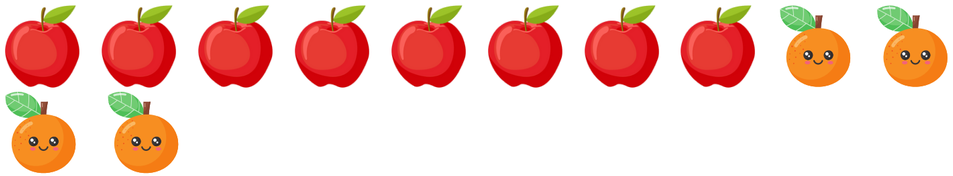
ASHIMA MALIK



if you identified 8 of the 10 items as apples, and all 8 were truly apple, your precision would be 8/8 (100%)



if there were 15 apples in total, and you only found 8, your recall would be 8/15 (around 53%).



Basket	Apples (Correct)	Oranges (Correct)	Incorrect	Total
12 Fruits	?	?	?	12

Precision: This measures how many of the things you identified as apples were actually apples.

precision = TP / (TP + FP) = 8 / (8 + 0) = 8 / 8 = 1 (which is 100%).

Recall: This measures how many of the actual apples you were able to identify.

recall = TP / (TP + FN) = 8 / 15 ~53%

Accuracy: you correctly identified 4 apples + 2 oranges = 6 fruits in total.

accuracy = TP +TN/ (TP + FN + FP + TN) = 6 / 12 = 50%

ASHIMA MALIK

BLEU Score:

BLEU score = Brevity Penalty * Maximum n-gram precision across all n-grams

Brevity Penalty: This factor penalizes outputs that are shorter than the reference translations.
Maximum n-gram precision: This is the highest precision value achieved for any n-gram length considered (typically n=1 to 4).

Weaknesses of BLEU Score:

Doesn't capture fluency or grammar: BLEU score focuses on n-gram overlap and doesn't directly assess how fluent or grammatically correct the translation is.
Sensitive to reference translations: The quality of the reference translations can impact the BLEU score.
Doesn't always correlate with human judgment: While BLEU score provides a quantitative measure, it might not perfectly reflect human evaluation of translation quality.

ROUGE Score:

ROUGE score comes in different variants (ROUGE-N, ROUGE-L, ROUGE-S) but all focus on recall, measuring how much overlap there is between n-grams (of varying lengths) in the generated text and reference summaries.

ROUGE-N Score:

ROUGE-N Score = 2*Precision*Recall/(Precision + Recall)

Recall_n: no. of overlapping unigrams /Total no of unigrams in reference summary
Precision_n: no. of overlapping unigrams /Total no of unigrams in the generated text

Weaknesses of ROUGE Score:

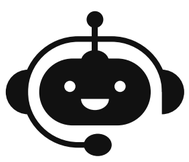
Doesn't guarantee fluency or coherence: ROUGE score doesn't directly assess how well-written or coherent the summary is.
Sensitive to reference summaries: The quality and number of reference summaries can impact the ROUGE score.

LLM Use Cases

- Search and Information Retrieval

- Content Creation and Summarization

- Art and Music Generation



- Chatbots and Virtual Assistants

- Code Generation and Analysis

- Customer Service and Support

- Scientific Research and Drug Discovery

- Education and Personalized Learning

ASHIMA MALIK

The Generative AI Market Map v3

A work in progress

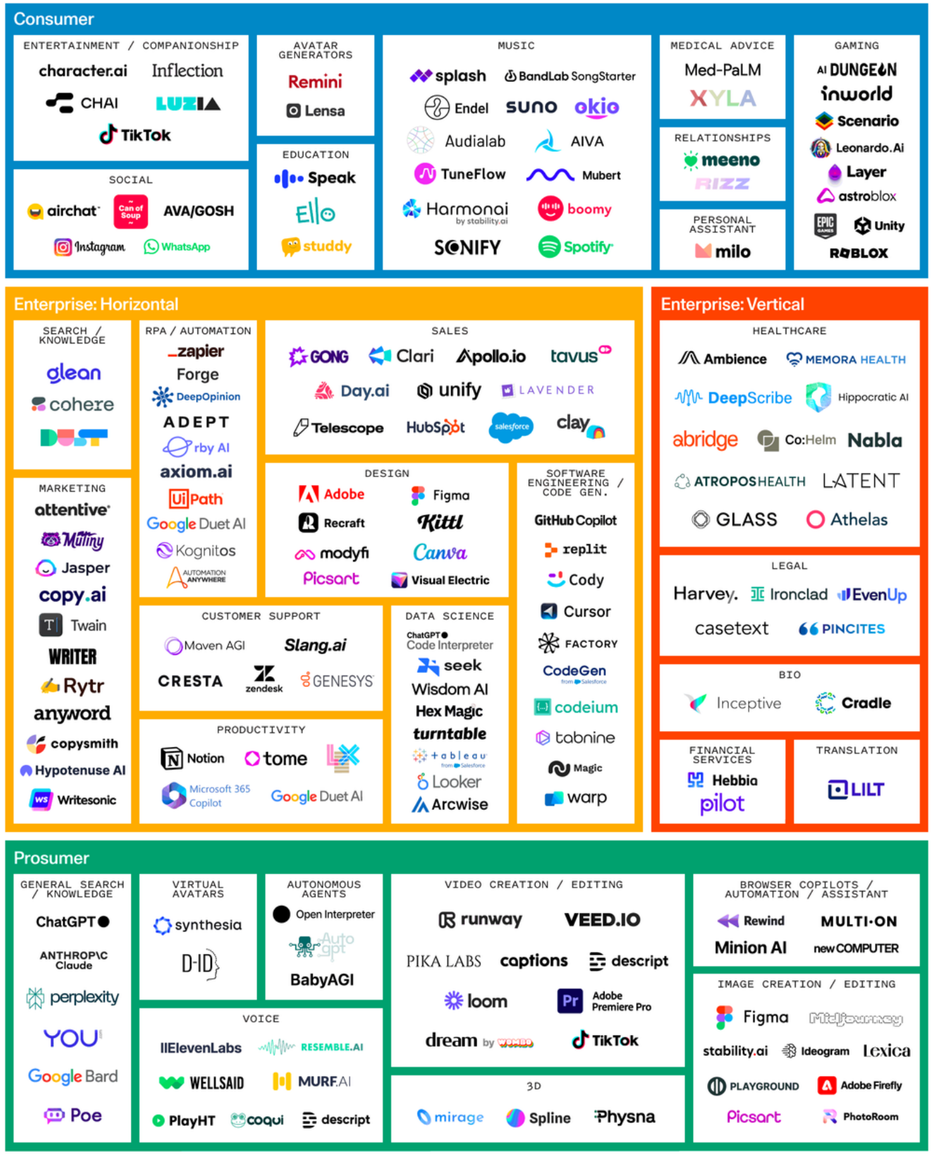
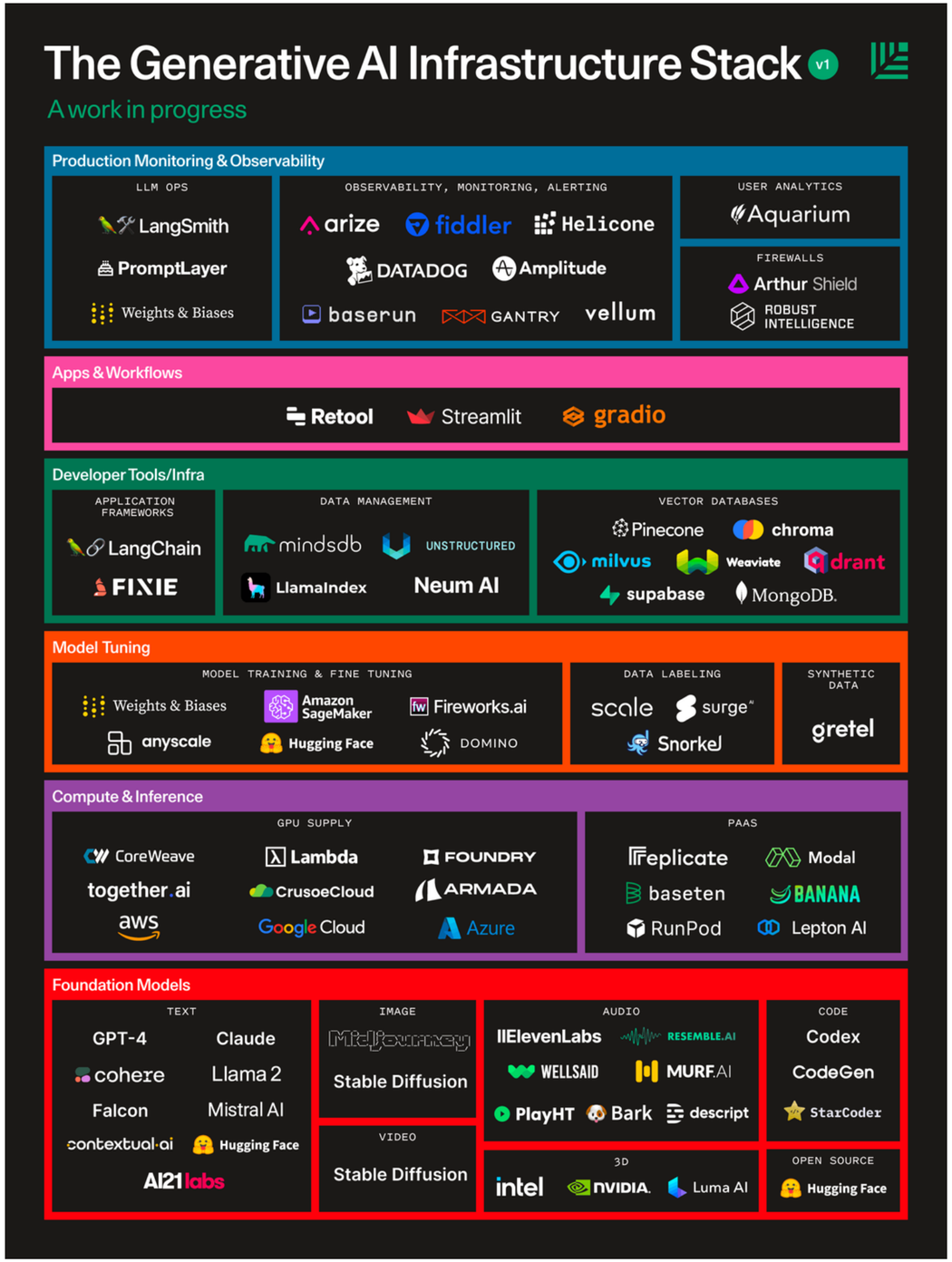
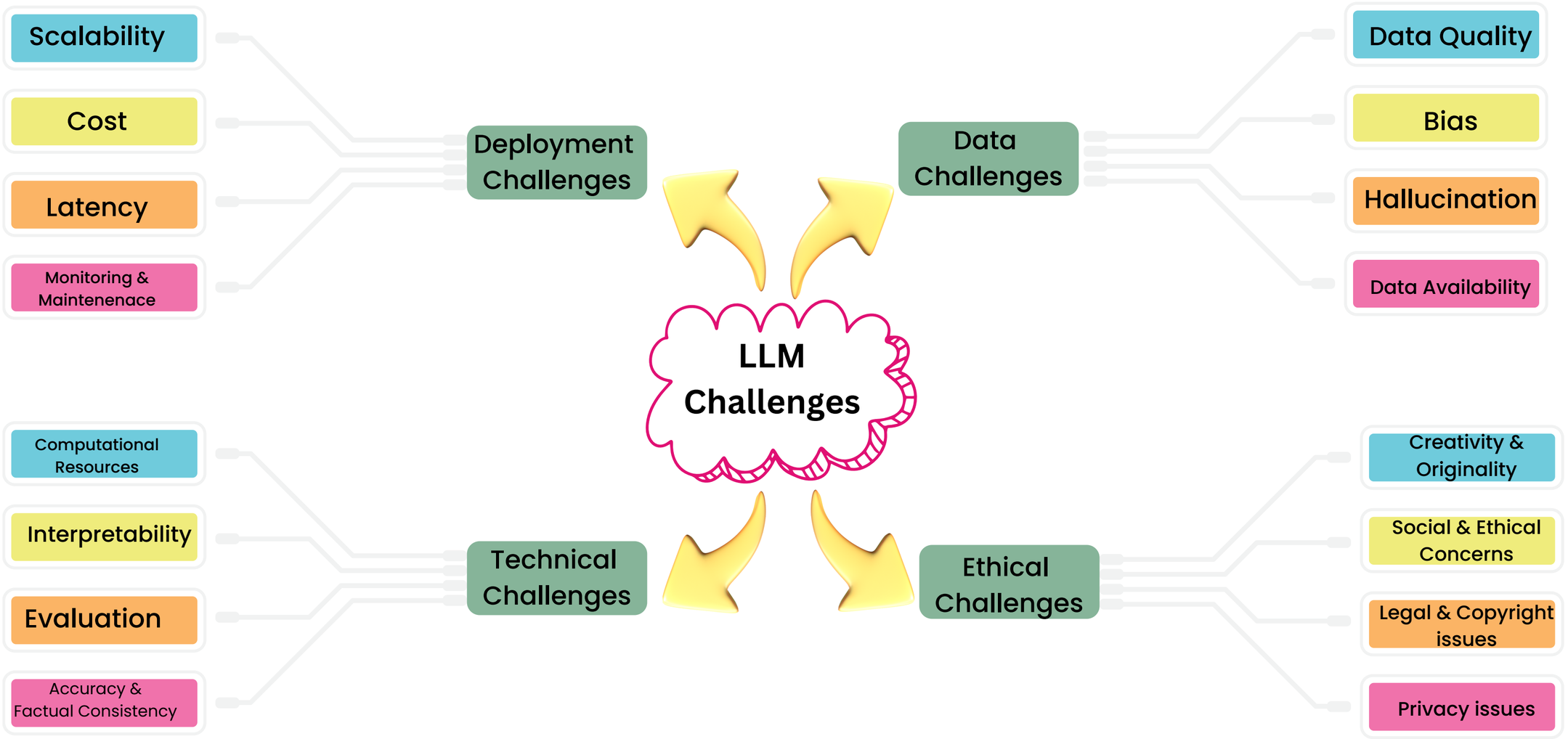


Image Source: https://www.sequoiacap.com/article/generative-ai-act-two/

The Generative AI Infrastructure Stack v1

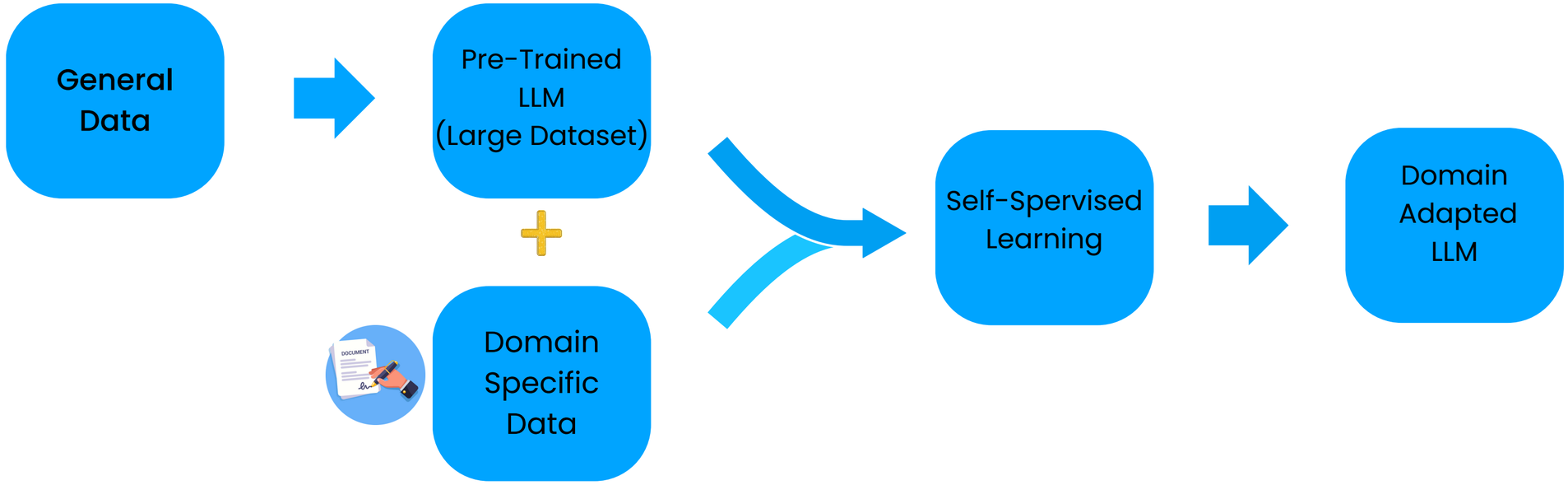
A work in progress





ASHIMA MALIK

Domain-Specific Pre-Training

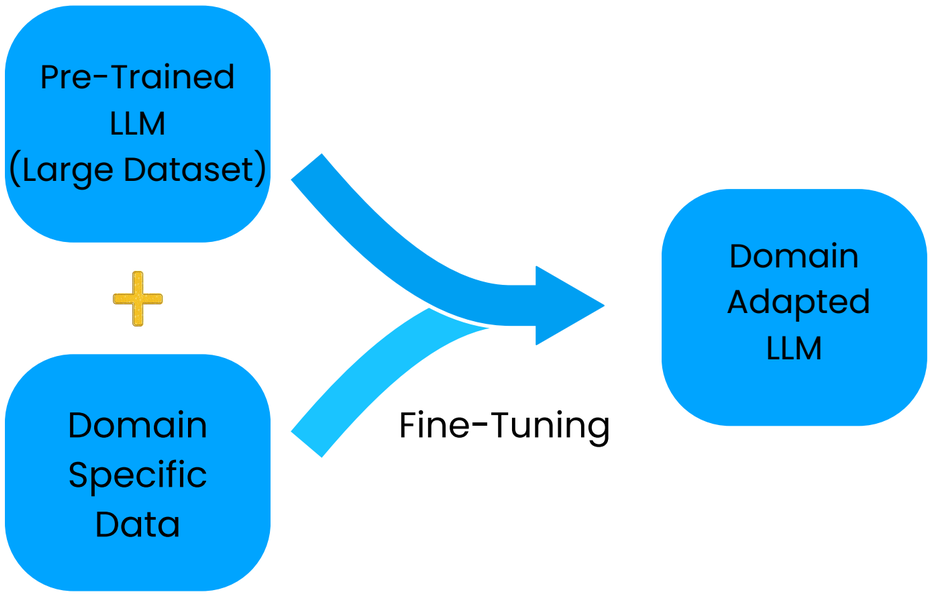


Training Duration: This can vary significantly depending on the size and complexity of the pre-trained LLM and the amount of domain-specific data used. Training a large LLM from scratch can take weeks or even months on powerful hardware. Pre-training on a smaller, focused domain might take days or even hours.

Summary: Broadens the LLM's understanding of a specific domain by training it on additional domain-specific data after initial pre-training on a general corpus.

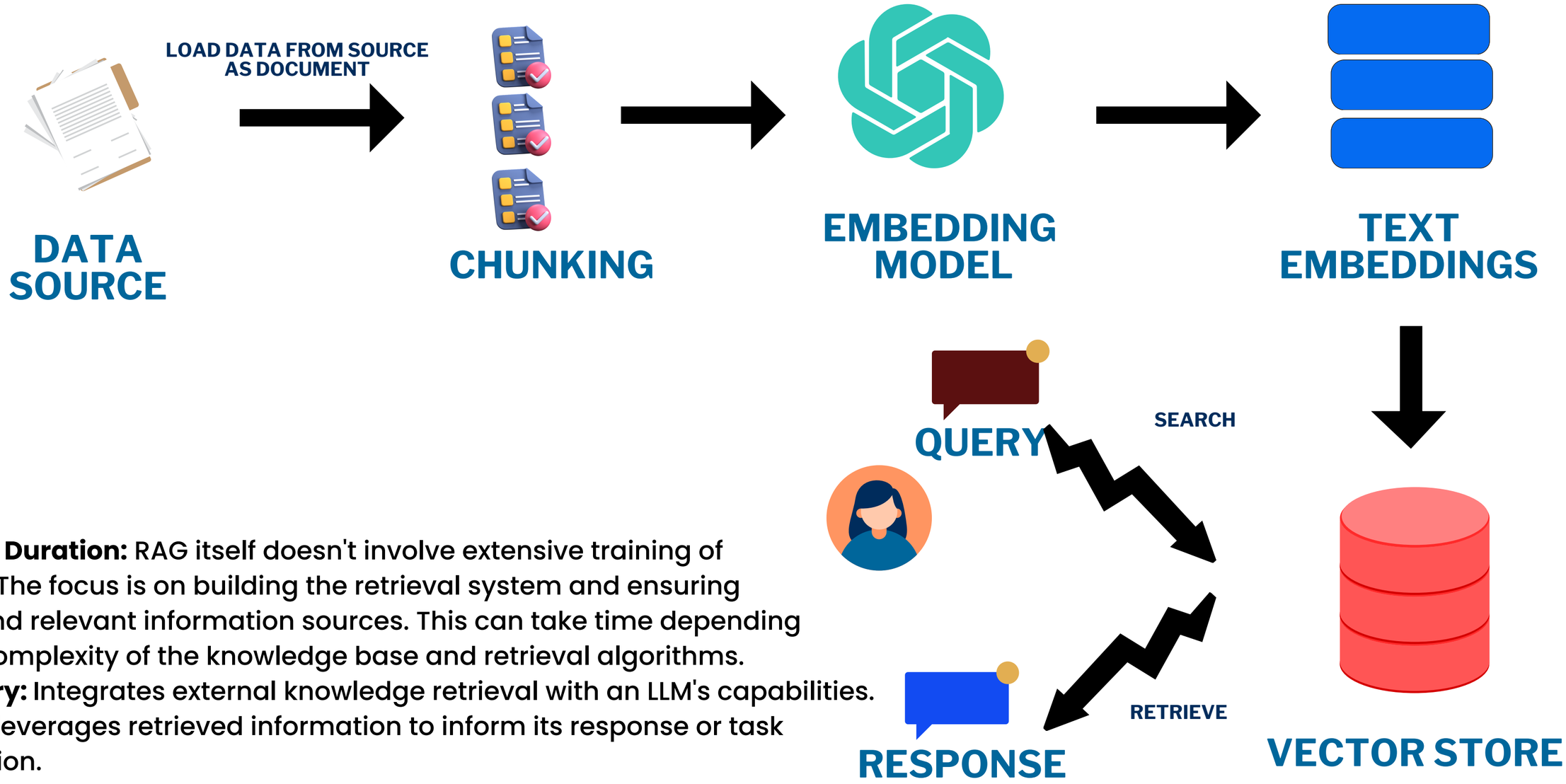
Example: Imagine a pre-trained LLM like GPT-3. We can further train it on a massive dataset of legal documents, case studies, and legal code. This pre-training would improve its ability to understand legal concepts and terminology, making it more suitable for tasks like legal research or contract analysis. Example, BloombergGPT

Domain-Specific Fine-Tuning



Training Duration: This is typically much faster than pre-training, often ranging from minutes to hours depending on the size and complexity of the task-specific dataset.
Summary: Specializes a pre-trained LLM for a specific task within a domain by fine-tuning it on a smaller dataset tailored to that task.
Example: Let's say we have a pre-trained LLM and want it to generate product descriptions for an e-commerce website. We would fine-tune it on a dataset of existing product descriptions with their corresponding product information (images, specifications). This fine-tuning would improve the LLM's ability to generate accurate and relevant product descriptions. Example, ChatDoctor LLM

Retrieval Augmented Generation (RAG)



Training Duration: RAG itself doesn't involve extensive training of the LLM. The focus is on building the retrieval system and ensuring it can find relevant information sources. This can take time depending on the complexity of the knowledge base and retrieval algorithms.

Summary: Integrates external knowledge retrieval with an LLM's capabilities. The LLM leverages retrieved information to inform its response or task completion.

Example: Consider a customer service chatbot powered by an LLM. RAG would involve building a system that retrieves relevant product manuals or troubleshooting guides based on the customer's query. The LLM would then use this retrieved information to formulate a helpful response to the customer.

ASHIMA MALIK

Technique	Training Duration	Summary	Examples
Domain-Specific Pre-Training	Days - Weeks (Large)	Broadens understanding of a specific domain	Legal text pre-training for legal research and contract analysis
	Hours (Smaller)		Medical journal pre-training for medical question answering and summary generation
Domain-Specific Fine-Tuning	Minutes - Hours	Specializes LLM for a specific task within a domain	Fine-tuning for product description generation on e-commerce websites
			Fine-tuning for financial news summarization
Retrieval Augmented Generation (RAG)	Varies (Retrieval System)	Integrates external knowledge retrieval with LLM's capabilities	Customer service chatbot using retrieved product manuals for troubleshooting
			Legal research assistant using retrieved case law for legal analysis

Technique	Training Focus	Benefits	Limitations
Domain-Specific Pre-Training	Improves general understanding of a domain	Better performance on domain-specific tasks	Requires large domain-specific datasets
Domain-Specific Fine-Tuning	Specializes LLM for a specific task	High accuracy on a particular task	Requires labeled data for the specific task
Retrieval Augmented Generation (RAG)	Leverages external knowledge sources	Access to up-to-date information, avoids extensive retraining	Relies on the quality of retrieved information

When to choose between Domain-Specific Pre-Training, Domain-Specific Fine-Tuning, and RAG