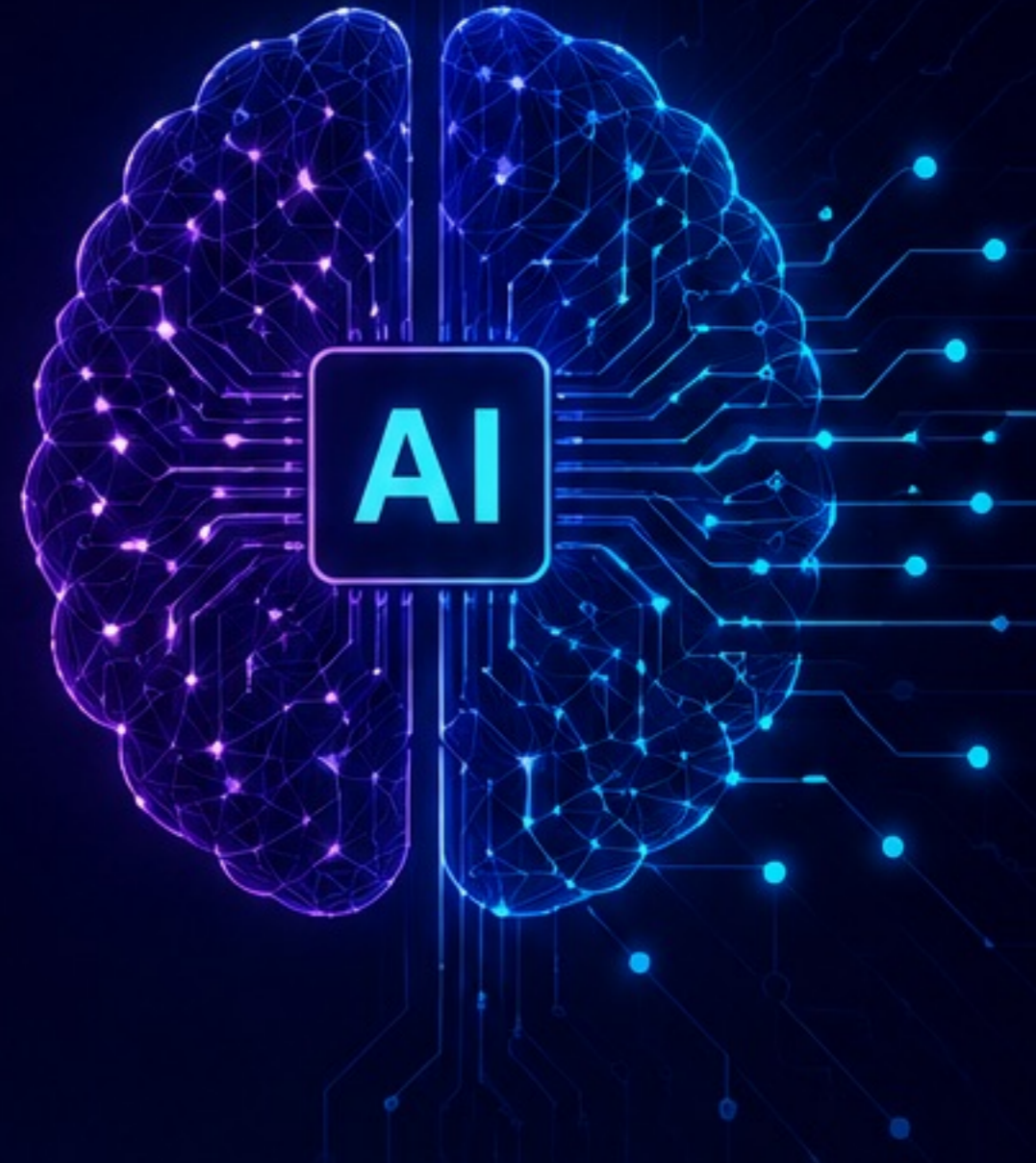
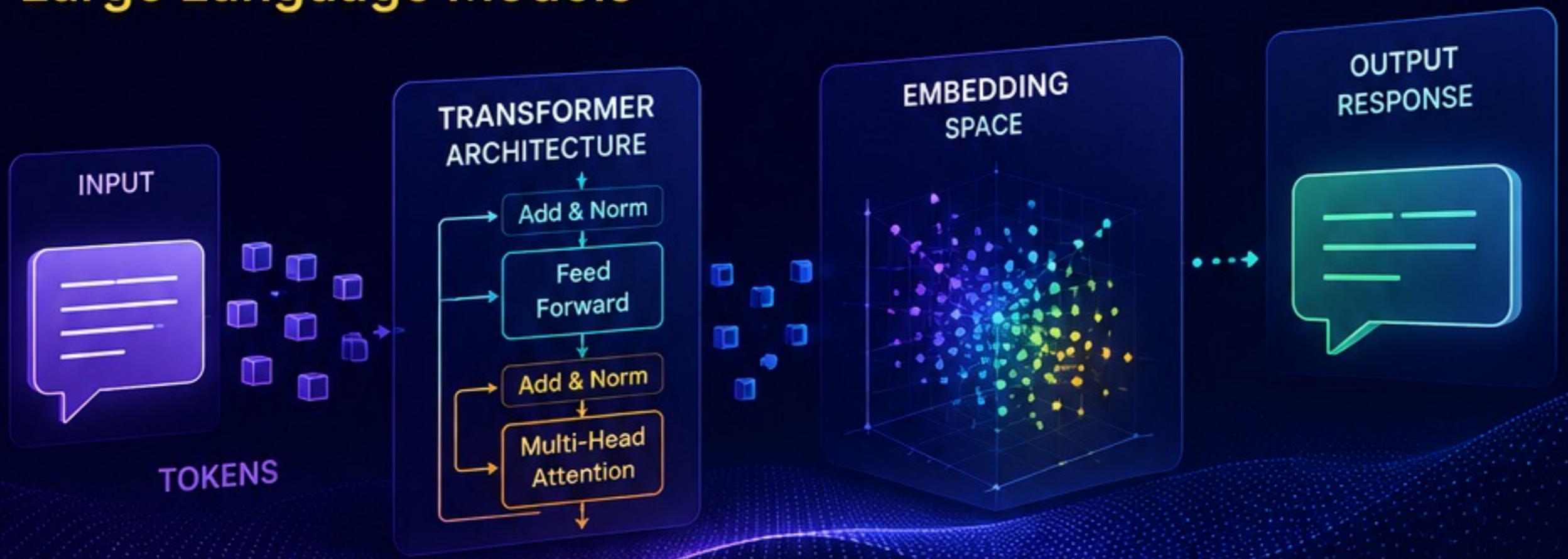


# LLMs Explained Visually



A complete visual guide to the concepts, techniques and ecosystem powering **Large Language Models**



TRANSFORMER ARCHITECTURE

LLM & AI TERMINOLOGY

LLM MODELS IN 3D SPACE

EMBEDDING SPACES

TOKEN EMBEDDINGS

QUANTIZATION IN LLMs

OPTIMIZATION TECHNIQUES

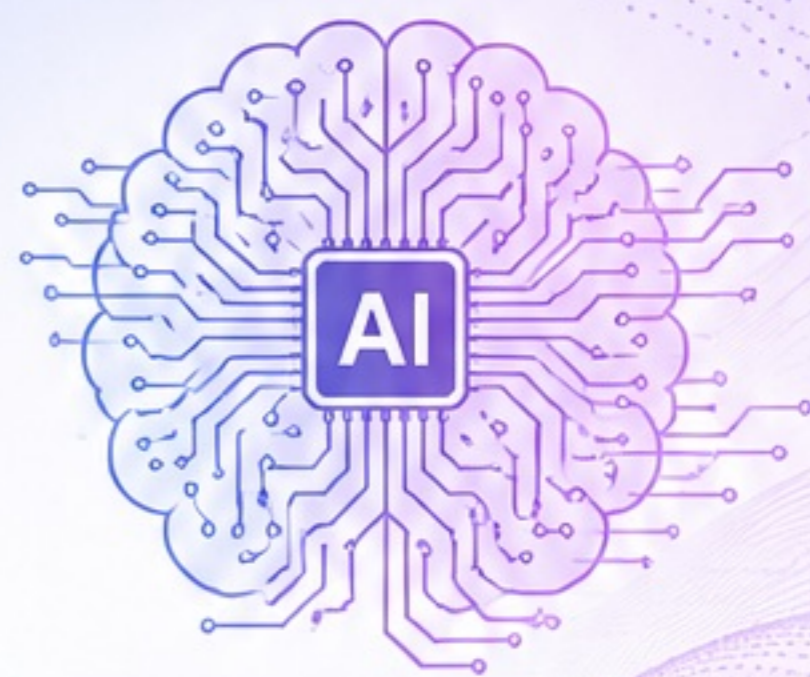
MODEL FORMATS

MODEL CONVERSIONS & ECOSYSTEM



WEBLLM OVERVIEW


APPS FOR LOCAL LLMs

# TABLE OF CONTENTS



A complete guide to the concepts, techniques and ecosystem powering Large Language Models.

	<b>01</b>	<b>Transformer Architecture Overview</b> ..... <b>01</b>
		Understand the transformer model, attention mechanism, and the architecture that powers modern LLMs.
	<b>02</b>	<b>LLM &amp; AI Terminology Guide</b> ..... <b>02</b>
		Key terms, concepts and roles you should know to navigate the world of LLMs with confidence.
	<b>03</b>	<b>LLM Models in 3D Space</b> ..... <b>03</b>
		Visualizing how different LLMs are positioned in a 3D semantic landscape.
	<b>04</b>	<b>LLM Data in 3D Embedding Space</b> ..... <b>04</b>
		Explore how data is represented in embedding space and what it reveals about meaning.
	<b>05</b>	<b>Token Embeddings in 3D Space</b> ..... <b>05</b>
		Zoomed-in view of token-level embeddings and their relationships.
	<b>06</b>	<b>Quantization in LLMs</b> ..... <b>06</b>
		Understand quantization techniques that make LLMs smaller, faster and more efficient.
	<b>07</b>	<b>LLM Optimization Techniques</b> ..... <b>07</b>
		Practical methods and strategies to optimize performance, memory and inference.
	<b>08</b>	<b>Understanding LLM Model Formats</b> ..... <b>08</b>
		A deep dive into model file formats like GGUF, ONNX, Safetensors and more.

 **TIP**  
Each section is designed to build your understanding step-by-step. Start from the top and explore!

# TABLE OF CONTENTS



## 09 | LLM Model Conversions & Ecosystem ..... 09

Explore model conversion workflows, key tools, and the ecosystem that enables model interoperability.



## 10 | WebLLM Overview ..... 10

Run LLMs directly in the browser with WebLLM. Architecture, capabilities and use cases.



## 11 | Applications for Running Local LLMs ..... 11

The best tools and platforms to run, manage and interact with LLMs on your local machine.

### WHAT YOU'LL GAIN FROM THIS GUIDE



#### Deep Understanding

Build strong foundations in LLM concepts and architectures.



#### Practical Knowledge

Learn optimization, quantization and deployment techniques.



#### Ecosystem Awareness

Explore tools, formats, and workflows in the LLM ecosystem.



#### Hands-on Enablement

Discover how to run and build applications with local and web-based LLMs.



#### Stay Ahead

Keep up with the latest techniques and best practices.



***“The best way to predict the future of AI is to build it.”***

— Start learning. Keep experimenting. Build the future.



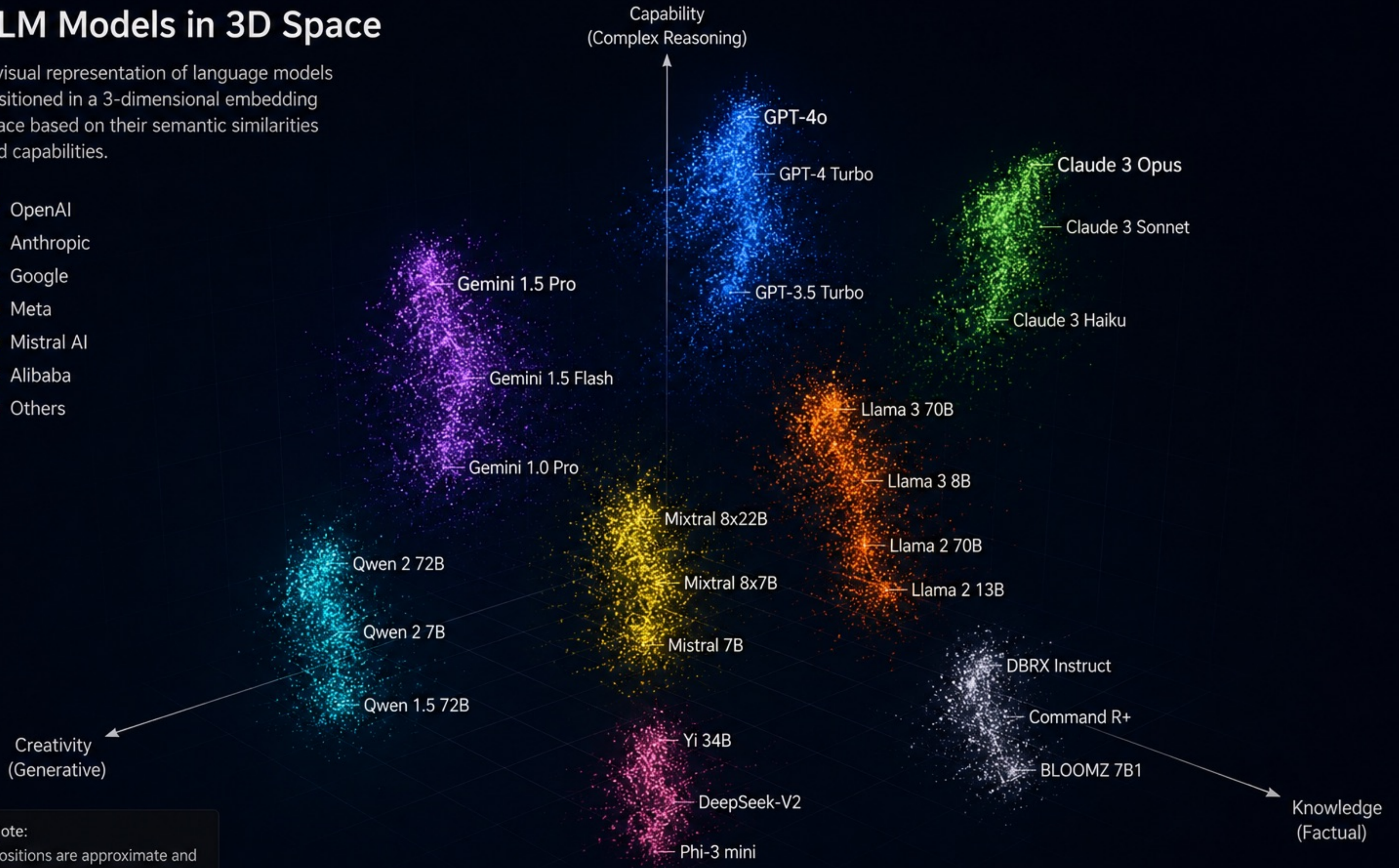
#### Happy Learning!

Dive in, explore each section and take your LLM knowledge to the next level.

# LLM Models in 3D Space

A visual representation of language models positioned in a 3-dimensional embedding space based on their semantic similarities and capabilities.

- OpenAI
- Anthropic
- Google
- Meta
- Mistral AI
- Alibaba
- Others



Note:  
Positions are approximate and do not represent exact metrics. Visualization is based on public information and embeddings.