







Deemo



2

5

Method	Definition	Primary use case	Data requirements	Training time	Advantage	Considerations
 <b>Prompt engineering</b>	Crafting specialized prompts to guide LLM behavior	Quick, on-the-fly model guidance	None	None	Fast, cost-effective, no training required	Less control than fine-tuning
 <b>Retrieval augmented generation (RAG)</b>	Combining an LLM with external knowledge retrieval	Dynamic datasets & external knowledge	External knowledge base or database (e.g. vector database)	Moderate (e.g. computing embeddings)	Dynamically updated context, enhanced accuracy	Significantly increases prompt length and inference computation
 <b>Fine-tuning</b>	Adapting a pre-trained LLM to specific datasets or domains	Domain or task specialization	Thousands of domain-specific or instruction examples	Moderate — long (depending on data size)	Granular control, high specialization	Requires labeled data, computational cost
 <b>Pre-training</b>	Training an LLM from scratch	Unique tasks or domain-specific corpora	Large datasets (billions to trillions of tokens)	Long (days to many weeks)	Maximum control, tailored for specific needs	Extremely resource-intensive

**How**

teams

are



building

**LLMs**



complexy/compute-intensive

github.com/ariefranh/mansyah/mps-llmops-mlflow

2

6