Generative AI Cheat Sheet

Quick reference from the 30-part series "From Zero to Mastery" by Dr. Pranay Jha

Key Terms, in Plain English

Term	What it means	Stage
Model	A giant function that predicts the next token	Use
Parameters	The learned numbers (dials) inside the model	Build
Training	Setting the dials from examples	Build
Fine-tuning	Extra targeted training on a narrow set	Build
Foundation model	A broad model pointed at many tasks	Build
LLM	A foundation model specialised in language	Build
Prompt	The text you give the model	Use
Token	A small chunk of text the model reads	Use
Context	Everything the model can see at once	Use
Inference	Running the model to get an answer	Use
Embedding	A token turned into meaning-coordinates	Bridge
Hallucination	Confident output that is false	Use

GPU Sizing Cheat Sheet (weights only)

Model size	FP16 weights	INT4 weights	Typically fits on
7B	~14 GB	~4 GB	A workstation / laptop GPU
13B	~26 GB	~7 GB	One mid-range GPU
70B	~140 GB	~35 GB	2+ GPUs at FP16, or 1x 40-80GB at INT4
175B	~350 GB	~88 GB	A multi-GPU node

Weights only. Add the KV cache (grows with context length and concurrency) on top.

Precision (quantization) at a Glance

Precision	Bytes / param	~70B weights	Trade-off
FP32	4	~280 GB	Most precise, rarely needed
FP16 / BF16	2	~140 GB	The standard default
INT8	1	~70 GB	Near-lossless with a good method
INT4	0.5	~35 GB	Big savings, small quality risk