

LLM Vocabulary

Computing

The words that appear in every model announcement, defined plainly.

Terms you meet immediately when working with language models.

Term	Means
Token	A chunk of text — roughly $\frac{3}{4}$ of a word
Context window	How many tokens the model can consider at once
Parameters	The learned weights; a rough proxy for size
Inference	Running the model to get an output
Fine-tuning	Further training on your own data
Temperature	How much randomness in the output
Prompt	Everything you send in
Completion	Everything the model sends back

Cost and limits are almost always counted in tokens, not words or characters. English averages ~4 characters per token.