AI LITERACY · PREP MATERIAL

AI, in plain English

List of some important LLM (Large Language Model - Terms explained in plain, simple English, using everyday analogies.)

1. Context Window (The AI's Short-Term Memory)

This is the maximum amount of text the AI can read, hold in its brain, and remember at one single time during a conversation. It includes your prompt, the files you upload, and all the previous messages in that specific chat thread.

Think of it like a desk surface. You can lay out a certain number of pages on the desk to look at while working. If you add a new page and the desk is completely full, you have to push an old page off the edge, meaning the AI "forgets" that earliest part of the conversation.

2. Prompt (The Instruction)

This is simply the text, question, or command you type into the AI box to tell it what to do.

It is like giving instructions to a freelance assistant. The clearer your detailed your instructions, the better the final work will be.

3. Hallucination (When AI Confidently Makes Stuff Up)

This happens when an AI model gives you an answer that sounds completely factual, professional, and convincing, but is completely incorrect or fabricated.

Think of a hyper-confident student trying to pass an oral exam. They don't know the answer to a question, but instead of admitting it, they invent a highly detailed story on the spot hoping you won't notice,

4. Fine-Tuning (Specialized Training)

This is the process of taking an already existing, generally smart AI model and giving it extra training on a narrow, specific set of data to make it an expert in one specific field.

Imagine a student who just graduated with a general college degree. Fine-tuning is sending them to a specialized three-month boot camp to learn the exact terminology and rules of a specific law firm or medical clinic.

5. RAG - Retrieval-Augmented Generation (AI "Open-Book" Searching)

Instead of relying purely on what it memorized during training, an AI using RAG will first look through a specific folder, database, or live internet page to find the right facts, and then use those facts to write your answer.

It is the difference between a student taking a closed-book exam (relying only on memory) versus an open-book exam (where they can look up the exact facts in a textbook before writing down the answer).

6. System Prompt / System Instructions (The AI's Persona)

These are hidden, permanent rules given to the AI by its creators before you even start typing. It dictates the AI's core identity, safety rules, tone of voice, and boundary limits.

This is like the employee handbook given to an office worker. It sets the baseline rules for how they must behave, what tone they should use with customers, and what topics they are strictly forbidden from discussing.

7. Parameters (The AI's Brain Cells)

You will often see numbers like "7B" or "70B" next to an AI model's name. This stands for Billions of Parameters. Parameters are the internal connections and dials the AI adjusts during its training to understand patterns in language. Generally, more parameters mean a smarter, more capable model.

Think of parameters like connections between neurons in a human brain. A brain with more connections can understand deeper nuances, complex jokes, and advanced logic compared to a smaller brain.

Understanding tokens & usage limits

What is a 'token' for AI Tool or LLM?

Tokens are based on pieces of words, not whole words or complete sentences. Think of tokens as building blocks or syllables that AI models use to read and process text.

Because tokens split words into smaller fragments, you get fewer words than tokens for your money. For standard English text:

1 token = roughly 4 characters (including spaces).

1 token = roughly 0.75 words.

100 tokens = roughly 75 words.

How Words are Split into Tokens

Common or short words usually count as a single token. Longer, complex, or rare words are broken down into multiple pieces.

Here is a literal example of how an AI processor splits text:

The word "cat" = 1 token.

The word "catastrophe" = 3 tokens (split into "cat", "as", "trophe").

The phrase "AI is amazing!" = 4 tokens ("AI", " is", " amazing", "! " - punctuation counts as tokens too).

Why Sentences and Other Languages Cost More

Because sentence length varies wildly, charging by the sentence would be unfair to the AI provider. A 5-word sentence and a 50-word sentence cost the AI vastly different amounts of computing power.

Every space, comma, and exclamation mark counts as its own token or part of one.

AI models are trained mostly on English. For languages like Hindi, Spanish, or Mandarin, words are split into much smaller, fragmented pieces. A single non-English word can easily cost 2 to 5 tokens, making it more expensive to process.

How Claude Defines a Token

Large Language Models cannot read full words or sentences. They break text down into chunks called tokens.

In English text, 1 token is roughly equivalent to 0.75 words. This means a 100-word paragraph consumes about 130 to 140 Claude tokens.

Commas, periods, and blank spaces count as individual tokens.

Brackets, indents, and syntax variables are heavily broken up and eat through tokens very quickly.

If you upload a screenshot, Claude converts the pixels into tokens. A standard 1000x1000 pixel image costs roughly 1,334 tokens.

Every word inside an uploaded document is parsed and tokenized into the conversation's memory bank.

How Claude Pro Charges & Calculates Token Usage

Claude Pro does not charge your credit card on a pay-as-you-go per-token basis. Instead, you pay a flat $20/month fee. In exchange, you are granted a dynamic pool of tokens that reset every 5 hours.

Here is exactly how that token pool is drained and tracked:

The "Re-Reading" Context Multiplier (The Hidden Cost)

This is why you hit limits faster than you expect. Every single time you send a new message, Claude must re-read the entire chat history from the beginning.

You type a 50-token question. Cost = 50 tokens.

You type another 50-token question. Claude has to read Message 1 + Claude's Response 1 + your new question. Cost = ~200 tokens.

The chat has grown long. Your new message is only 10 words, but Claude has to re-read 50,000 tokens of chat history to understand the context. Cost = 50,010 tokens.

Because of this, long continuous chat sessions drain your 5-hour token allowance dramatically faster than starting a fresh conversation thread.

The 5-Hour Rolling Limit

Your Pro subscription gives you an allocation of roughly 44,000 tokens per 5-hour window (though Anthropic adjusts this dynamically based on data center demand and whether you use advanced features like "Extended Thinking").

If your chat threads are short and fresh, you can easily send 45+ messages in 5 hours.

If you upload a massive 100-page PDF document, you might hit your token ceiling after only 10 messages, because the PDF's tokens are being re-charged to your account with every single back-and-forth reply.

What Happens If You Run Out?

When your 5-hour token pool is completely drained, Claude Pro gives you a warning that your limit is reached. You then have two options:

Your capacity will automatically regenerate as time passes.

You can navigate to your settings and opt to purchase prepaid "Usage Credits". Once enabled, if you exceed your Pro allowance, you will be billed a few cents per thousand tokens directly against your credit balance at standard API rates.