Term	Plain definition	Analogy
GGUF models	A new, efficient file format used to store and run AI language models (especially **local, open-source models** like LLaMA) with better speed and compatibility. GGUF stands for "GGML Unified Format", replacing older formats like .bin or .ggml.	Like packing a suitcase really well so it fits perfectly in any car—fast to load, easy to carry, and works anywhere.
Quantization files	Versions of AI models that have been **compressed by reducing the precision of the numbers** they use—making them smaller and faster to run. Q8, Q6, Q5, Q4 → lower number = smaller file, but slightly less accurate.	Like saving a high-res photo in a lower quality to make it load faster on your phone—still useful, just lighter.
Temperature	Controls how creative or random the model's output is. A moderate value like 0.7 balances **creativity and accuracy**.	Like asking someone to give slightly varied answers—not robotic, but not too wild either.
TopP	Controls how many possible words the AI considers when generating text, keeping only the **top 90% most likely options**.	Like picking words from a shortlist of top-rated ideas rather than the entire dictionary.
RepeatPenalty	Prevents the AI from **repeating the same phrases or words** too often. Higher values reduce repetition.	Like telling a speaker, "Stop repeating yourself!"
NumCtx	Sets the **context window size**, or how many tokens (words/characters) the model can "remember" from earlier in the conversation.	Like how much text the AI can keep in mind while responding—like short-term memory.
Seed	Fixes the randomness so you get **repeatable results**. Same seed = same output, assuming everything else is the same.	Like using the same shuffle seed on a playlist—you'll get the same song order every time.
NumPredict	The number of **tokens (words or word pieces)** the AI should generate in response to a prompt.	Like telling someone, "Write just 50 words, no more"—you're setting a length limit for their answer.