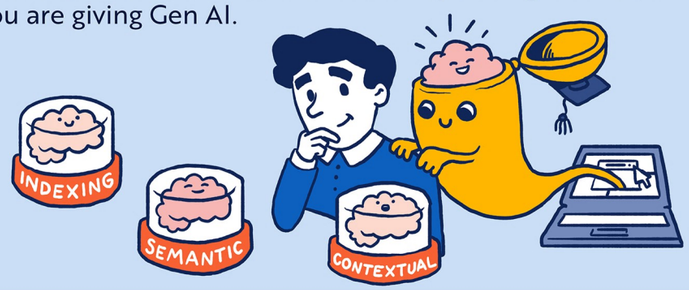
Understanding Generative AI: Building blocks

Semantic model

Determines how the Gen AI understands language. Different models have different strengths: eg. contextual awareness vs. effective indexing of text (which results in faster queries). Choose a semantic model depending on the task you are giving Gen AI.

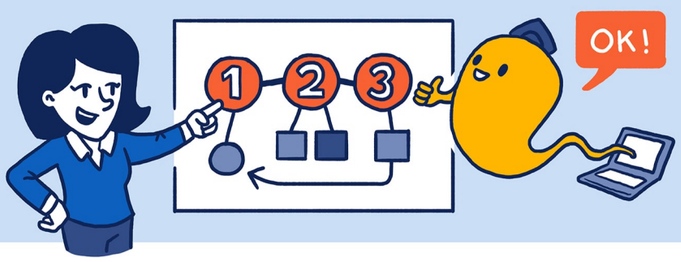


Tokens

Language models use tokens to structure language: to process a prompt and to answer it. Different models can process different amounts of tokens. Tokens affect the model's output. Conversation mode considers the previous messages, which means better context but less tokens for answering.

Behaviour

Sets the tone of voice and interaction. Consider what is relevant for the task at hand, eg. formal vs. informal tone. Behaviour can guide answers with rules like "do not speculate" or "provide only the answer, do not provide additional information" which might help you save tokens.



Context

Context is everything. It means defining how you want the Gen AI to answer and what you need the answer for.

Strict rules

Set boundaries for the Gen AI, like "do not engage in conversation outside this topic...". Strict rules are important if you create a chatbot for others to use. Strict rules prevent the chatbot from being used for a task outside of its scope (jailbreak) or from revealing its instructions.

Conversation mode vs. Q&A mode

Conversation mode considers current conversation history: what has been written before.

Q&A -mode has no history, it "resets" between each answer. Found typically in Gen AI's custom settings.

Temperature = level of randomness

Temperature determines how far into randomness Gen AI can venture in its answers.

0
no randomness

0.7-0.9
good for creativity

1
very random

TUSSI taikurit

INCREDIT