Prompt Engineering for Beginners

Kevin Crook

KevinCrook.com

Topics

AI Chatbot Basics
(ChatGPT, Claude, Gemini, and Grok are AI Chatbots)

Prompt Engineering Best Practices

Takeaway: 100 Example AI Life Hack Prompts

AI Chatbot Basics

(ChatGPT, Claude, Gemini, and Grok are AI Chatbots)

Large Language Model (LLM)

AI Chatbots use LLMs under the hood.

(For the remainder of this workshop, assume the term AI refers to an AI Chatbot built on top of an LLM.)

Prompt

What you type to the AI.

Response

What the AI replies.

Chat

The back-and-forth conversation, which consists of a series of prompts and responses.

Context Window

The maximum text the AI can process at once. In very long chats, the oldest parts get dropped.

Prompt Engineering

The skill of writing clear, effective prompts that get much better results from AIs.

LLM Engineering

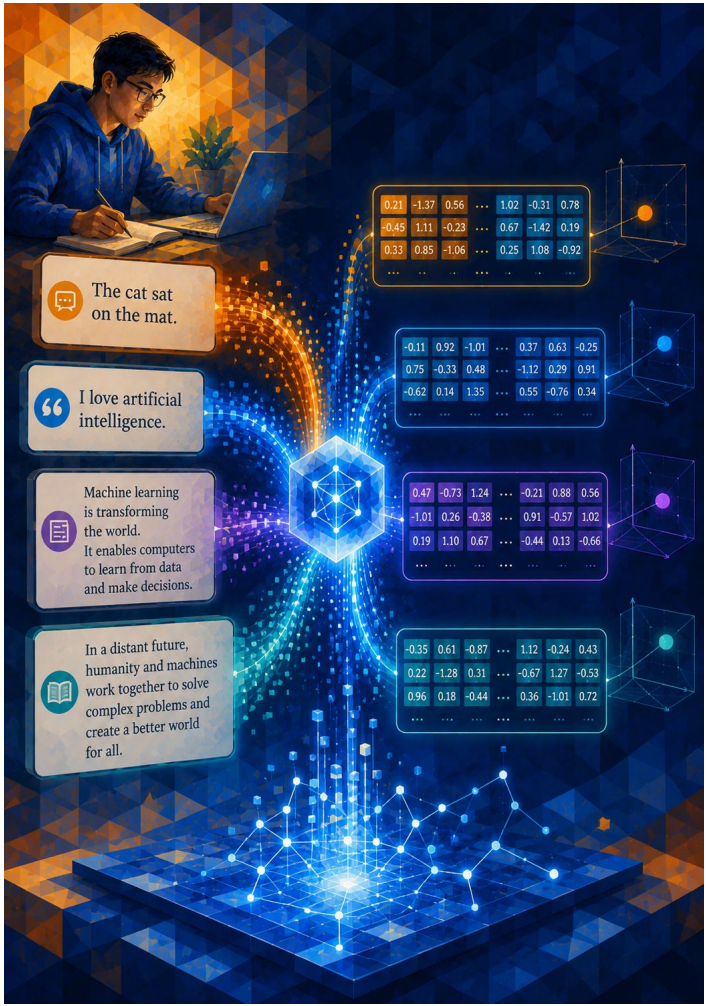
Advanced techniques beyond prompting to build systems around the AI. Often requires computer programming and systems engineering skills (another free workshop).

Tokens

Words or parts of words that AIs read and process.

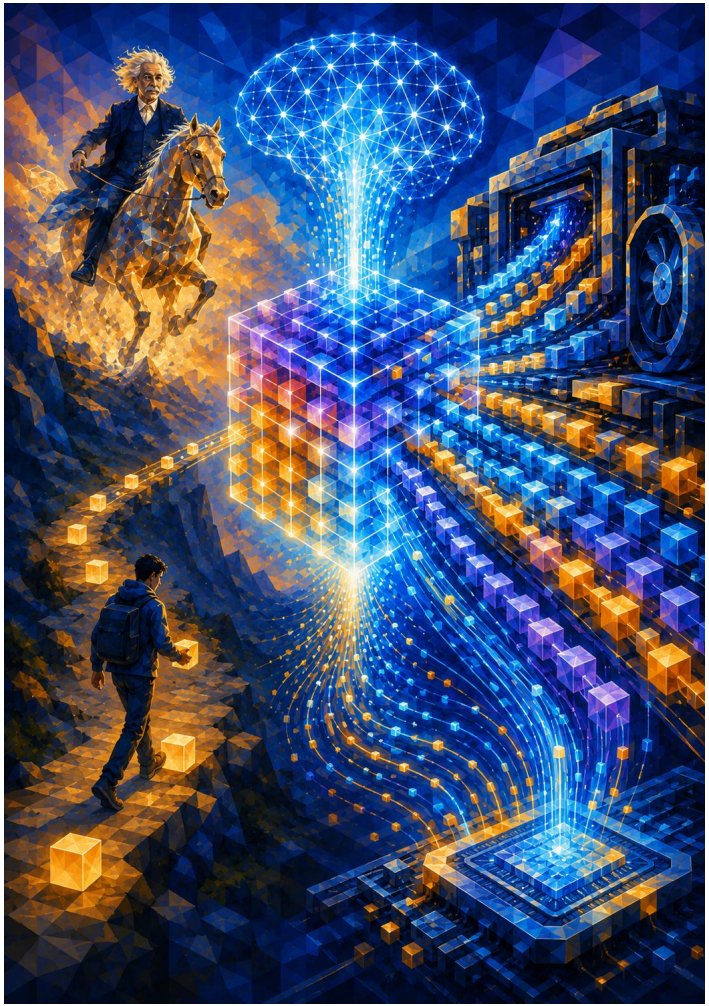
Example token counts:

- Mary Shelley’s Frankenstein:
  - 100,000 tokens / 75,000 words
- Bram Stoker’s Dracula:
  - 210,000 tokens / 161,000 words
- Leo Tolstoy’s War and Peace:
  - 750,000 tokens / 587,000 words
- William Shakespeare’s Complete Works:
  - 1.2 million tokens / 885,000 words



Language as Lists of Numbers

Words, sentences, paragraphs are turned into lists of numbers (high-dimensional vectors).



Why do Graphics Cards speed up AIs?

Einstein used and popularized tensors. He compared them to riding a horse instead of walking.

Tensor: A multi-dimensional array of numbers.

High-dimensional vectors are a form of tensor.

Computer graphics rely heavily on tensors, so GPUs (Graphics Processing Units on graphics cards) are excellent at the math AIs need for high-dimensional vectors.

TPUs (Tensor Processing Units) are specialized chips built purely for AI tensor calculations.

Training Cutoff Date

AIs are trained on text up until a certain cutoff date and have no built-in knowledge of later events.

Solutions:

- In Context Learning.
- Retrieval-Augmented Generation (RAG).

Analogy: Kids Learning to Read

Just like kids learning to read pick up extra story details, AIs trained on massive amounts of text absorb extra information they might never fully forget.

This can lead to unexpected or unwanted associations in responses.

Hallucinations

AIs sometimes confidently make up facts that sound real but aren't.

AIs predict plausible-sounding text.
They don't look facts up. Gaps in training get filled with convincing guesses.

Always fact check important information.

Solutions that cut down hallucinations:

- In Context Learning.
- Retrieval-Augmented Generation (RAG).

Lost in the Middle

In very long chats or documents,
AIs can ignore or forget information
in the middle.

Solutions:

- Chunking: break up into chunks, ask for summaries, then combine summaries.
- Newer LLM architectures such as State Space Model (SSM) + Transformer Hybrids.

Context Rot

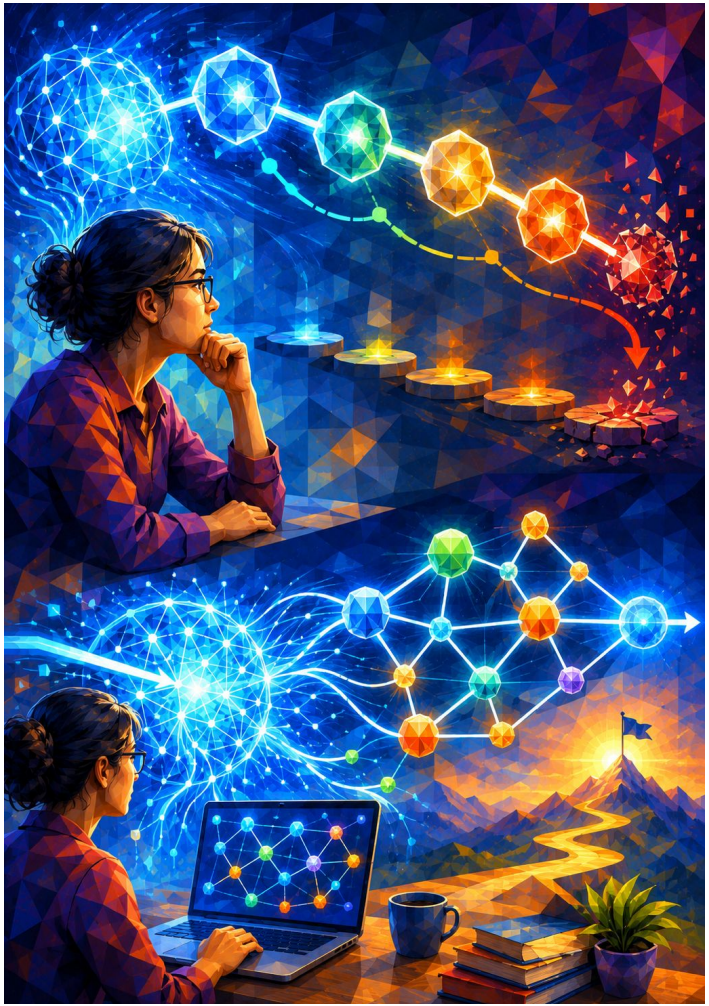
Over a long chat, earlier prompts, responses, files, etc. get diluted or forgotten.

Task Rot

Over a long chat, the AI can drift away from the original task or instructions.

Solutions:

- Ask for a summary of the chat so far, start a new chat, and paste the summary in.
- Newer LLM architectures such as State Space Model (SSM) + Transformer Hybrids.



Chain Rot
(Multi-Hop Reasoning Errors)

Multi-hop Chain: A relates to B relates to C relates to D...

AI's reasoning gets worse with each hop in the chain away from the original information.

Solution:
Knowledge Graph Engineering to create Knowledge Graphs, then put them in context or RAG (another free workshop).

Future: SSM + Transformer Hybrid LLMs

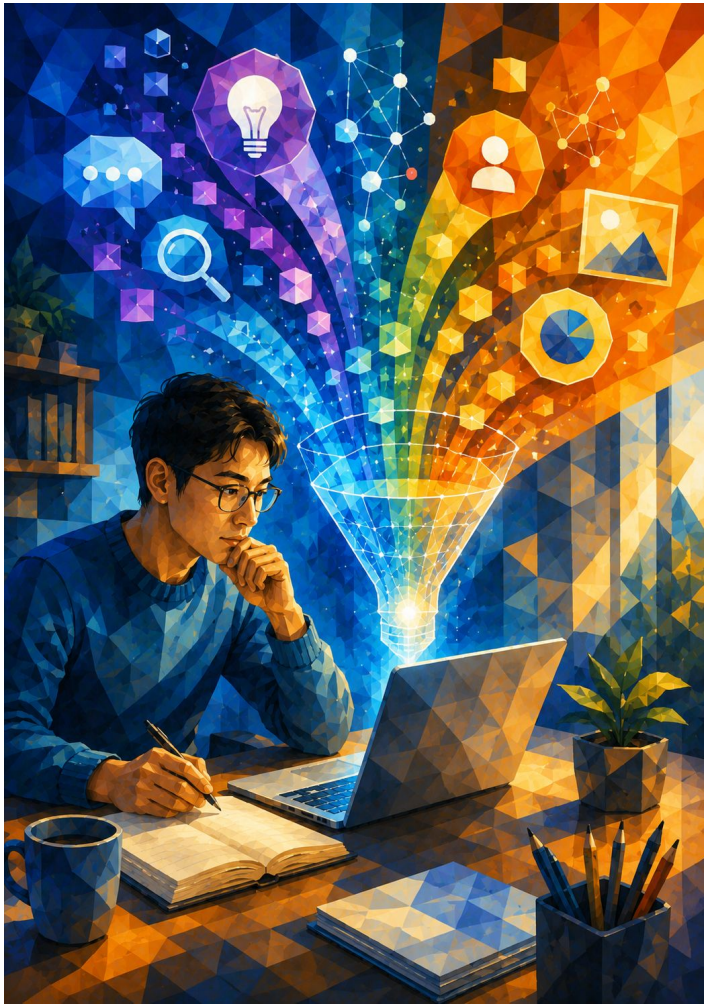
Today's AIs (ChatGPT, Claude, Gemini, Grok) are built on Transformers, which re-read the whole chat for every word they write.

State Space Models (SSMs) instead keep a running summary as they read.

Hybrids combine both:

- Faster and cheaper on the same hardware (GPUs, TPUs).
- Better with long chats and documents. Less lost in the middle, context rot, and task rot.
- Well suited to RAG including Knowledge Graphs.

NVIDIA's Nemotron models are free to download and run on your own high end hardware (self-hosted).



Prompt Engineering Best Practices

Start Simple with Interactive Chat

Don't try to write the perfect prompt on the first try.

Trial and error is normal and expected.

Begin with a basic prompt, then refine it step by step.

The Secret Sauce of Prompt Engineering

Triple Power:
Meta-Prompting + Role + COSTAR

Major prompt engineering contest winners and top finishers use this technique.

Most important takeaway from this workshop!

Meta-Prompting

Think of AI as your personal expert prompt engineer.

AI can do prompt engineering better than most humans can (maybe all?).

Use 2 chats:

- Chat 1: ask the AI to help you create and refine prompts.
- Chat 2: run the finished prompts.

Role + COSTAR

Role: Assign the AI a specific role or expertise.
("You are an experienced travel agent...")

Context: Background info the AI needs.

Objective: Exactly what you want done.

Style: Writing style you want.
(formal report, casual, technical, etc.)

Tone: The attitude of the response.
(friendly, professional, persuasive, etc.)

Audience: Who the response is for.
(beginners, experts, kids, your boss, etc.)

Response Format: How to deliver the answer.
(bullet list, table, paragraph, long form, with examples, etc.)

In Context Learning

AIs learn from whatever you put in the context window with no retraining needed.

File Uploads and Projects are great examples: Your documents become knowledge the AI can use.

Meta-prompting works here too:
Ask the AI to help generate the context material itself (summaries, examples, background docs).

Manage Context Rot & Task Rot

Periodically ask the LLM to summarize the conversation so far. This keeps both of you focused.

At the first signs of context rot or task rot, cut your losses! Ask for a summary, start a fresh new chat, and paste it in.

Handle Lost in the Middle Issues

For large documents or long contexts:

- Break them into smaller chunks.
- Process each piece separately.
- Then ask the LLM to combine or summarize the results.

Ask for Chain of Thought

For complex tasks, add this to your prompt:

"Explain your reasoning step by step."

You'll get noticeably better answers, and you can check the reasoning yourself.

Try Different LLM Versions and Vendors

LLM vendors offer multiple versions (fast, deep, beta, etc.).

There are multiple vendors to choose from (ChatGPT, Claude, Gemini, Grok, etc.).

Stuck or unsatisfied? The same prompt can get very different results elsewhere.

Build or Use a Prompt Library

Save your best prompts in a document as you go for easy reuse and adaptation or search for public "prompt libraries" for great starting points.

Share Chat Links

Share useful chats using the LLM's share link.

Great for showing coworkers, family, or friends exactly how you got a result, prompts and all.

Takeaway:

100 Example

AI Life Hack Prompts

Prompt Engineering for Beginners

Kevin Crook
KevinCrook.com

The End