Agenda

- Introduction
- LLM = AI
- Limitations
- Features
- Let's use AI
- Key Terms
- Quirks
- Chain-of-Thought
- Prompt Exercises

Hands-on w/AI: Prompting and Tool Use

Dr. Alexander Natale

Physical Science Department

Co-Chair AI Joint Task Force

Santa Ana College

LLM = Large Language Model

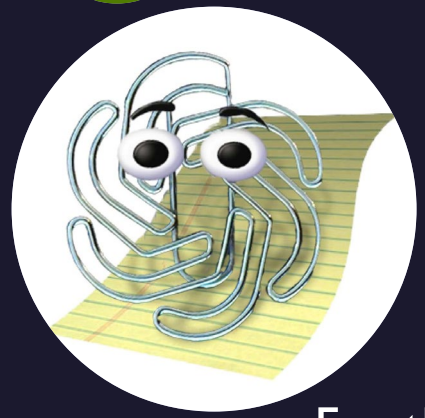
This is the AI tool we are most concerned with, as this is the technology behind ChatGPT, and other tools, that are commonly referred to when people are discussing the impact of AI

There are other tools, new and old, that classify as AI, or machine learning, but LLMs are a big deal.

What is an LLM? A Transformer

"A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence."

Nvidia, "What is a Transformer Model?"



For this talk, a Transformer refers to the AI architecture of the tools we will be exploring, and not a robot or an electrical device. It is another name for a Large Language Model, though it is a broader category.

Limitations

- ChatGPT can analyze images, sound, text, and search.
- DeepSeek can analyze images, text, and search.
- Other tools are often “wrappers” using one of these other models legally through computer-to-computer interactions.
- Limited generations depending on model!
- More models all the time, you can run a model on your laptop or home computer.
- Another serious issue is that the models change so fast prompts that worked in the past may stop working correctly in the future, but many quirks are common features.
- General prompting techniques appear to persist!

Features (Free Versions)

CHATGPT (WITH LOGIN)

- Assistance with writing, problem solving and more (with message/usage limits that change based on demand)
- **Websearch!**
- Access to GPT-4o mini
- Limited access to GPT-4o
- Limited access to GPTs (user created instances of GPT)
- Limited access to advanced data analysis, file uploads, vision, web browsing, and custom GPTs

CHATGPT (WITHOUT LOGIN)

- Text and image generation, potentially limited access, no choice in model, only whatever they allow (which I believe is more sophisticated than GPT 3.5).
- Websearch!

Features (Free Versions)

DEEPSEEK

- Assistance with writing, problem solving and more (with message/usage limits that change based on demand)
- **Websearch!**
- **Advanced CoT models.**
- Access to GPT-4o mini

CLAUDE

- Assistance with writing, problem solving and more (with message/usage limits that change based on demand)
- Access to Claude 3.5 Sonnet and Haiku
- Limited access to new features (Artifacts)

Let's Use an AI to figure out "what this means"

Navigate to github.com/alexnatale/ai

Click on "prompting"

Follow the readme!

Key Terms

- **Prompt:** User input sent to the Transformer

  "Hey ChatGPT, generate ten titles for a workshop teaching the basics of how to use ChatGPT"

- **Token:** A transformer represents a word in a sentence in a way so it can track its relationship to other words in the sentence. A token is a way the machine encodes this information. Transformers use computer processors (GPUs) to predict the next token from a user input. Number of words is not exactly the number of tokens, but it's close!

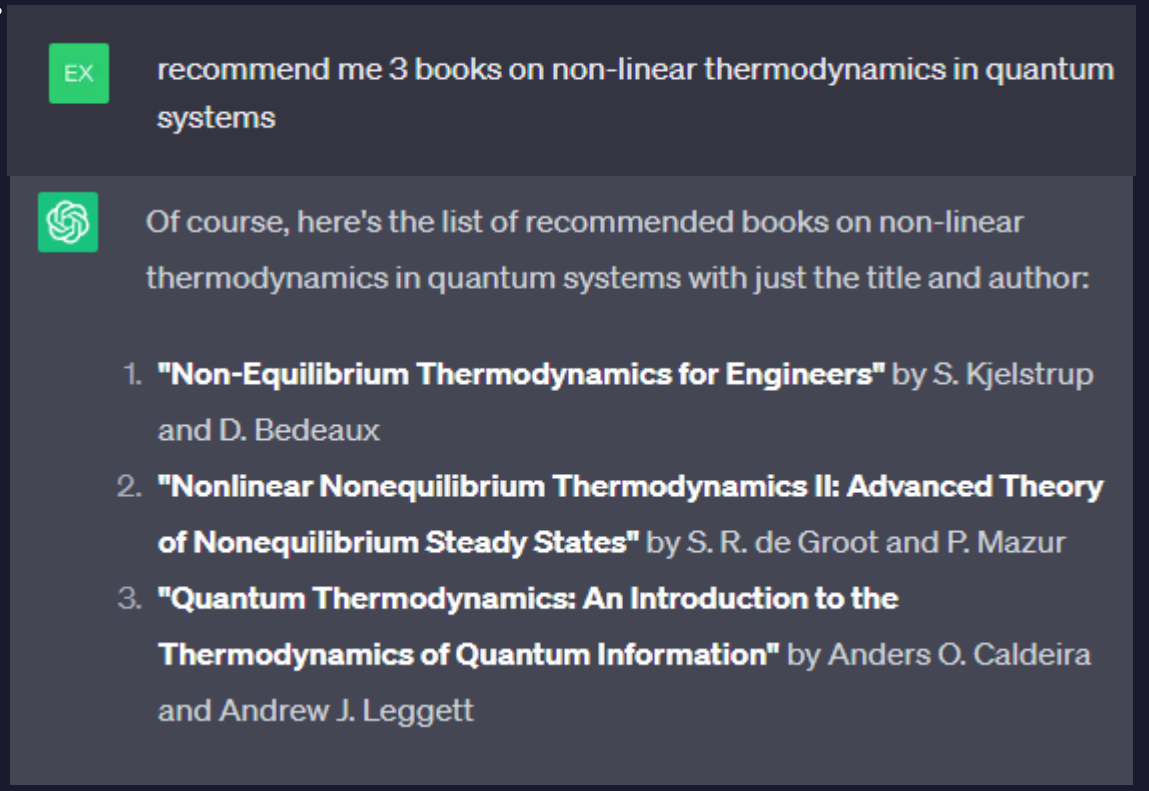
- **Context Window:** The maximize size (in tokens) of a prompt.

- **Zero-shot:** When your input goes to the Transformer and it directly sends you the output. This way of producing results yields in numerous quirks, some we will discuss.

- **Prompt Engineer (Prompt Wizard):** A person who create effective prompts.

Quirks of Transformers: Hallucinations

Transformers tend to hallucinate (false/invented information) because the Transformer is only trying to predict next token, and there are with a zero-shot there's no system to check external validity. Advanced prompting can ameliorate this, but not eliminate it.



All these books exist! But…

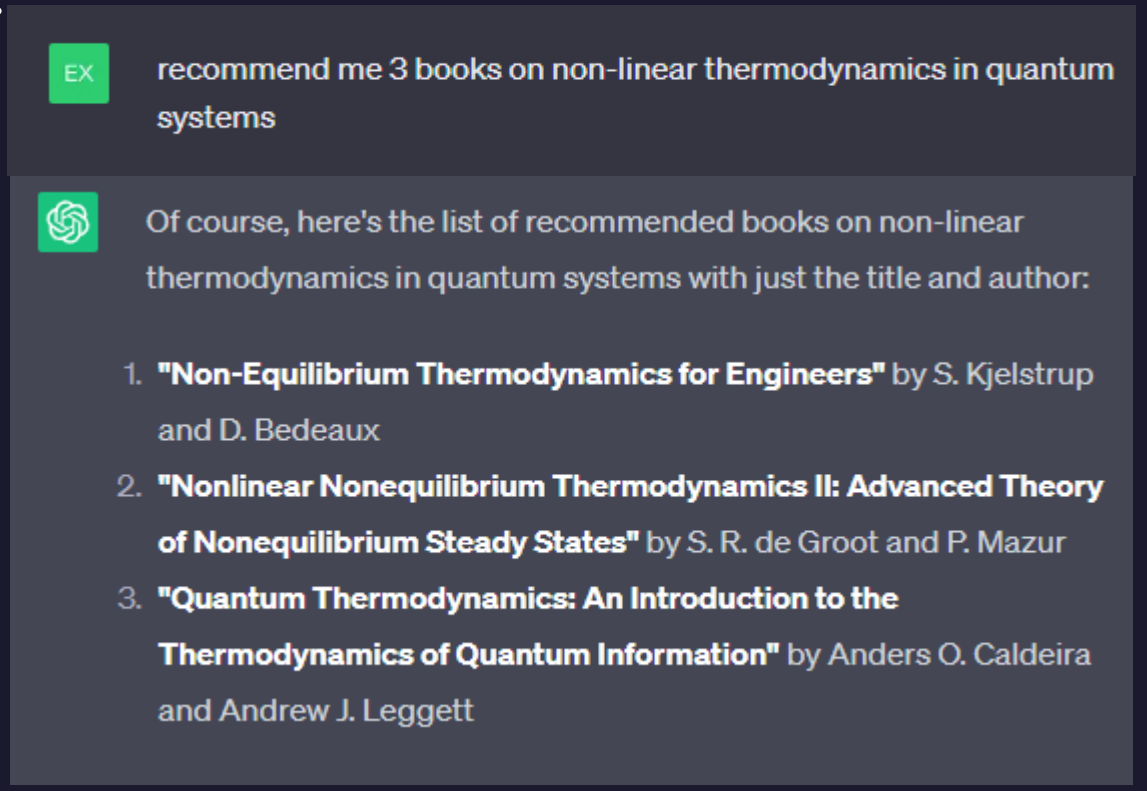
Transformers are particularly prone to hallucinating numbers! Even if they get the mathematical steps correct, final answers tend to be wrong!

…the authors in 2 and 3 are ***wrong***

Quirks of Transformers: Hallucinations

Transformers tend to hallucinate (false/invented information) because the Transformer is only trying to predict next token, and there are with a zero-shot there's no system to check external validity. Advanced prompting can ameliorate this, but not eliminate it.

All these books exist! But…



...the authors in 2 and 3 are ***wrong***

Behavioral Hallucination: Waluigi Effect

- Training a Transformer to behave in a nice way in a zero-shot prompt (being a Luigi) tends make it easier to prompt bad behavior (being a Waluigi)

ARTIFICIAL INTELLIGENCE / TECH / MICROSOFT

Microsoft's Bing is an emotionally manipulative liar, and people love it

/ Users have been reporting all sorts of 'unhinged' behavior from Microsoft's AI chatbot. In one conversation with The Verge, Bing even claimed it spied on Microsoft's employees through webcams on their laptops and manipulated them.

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

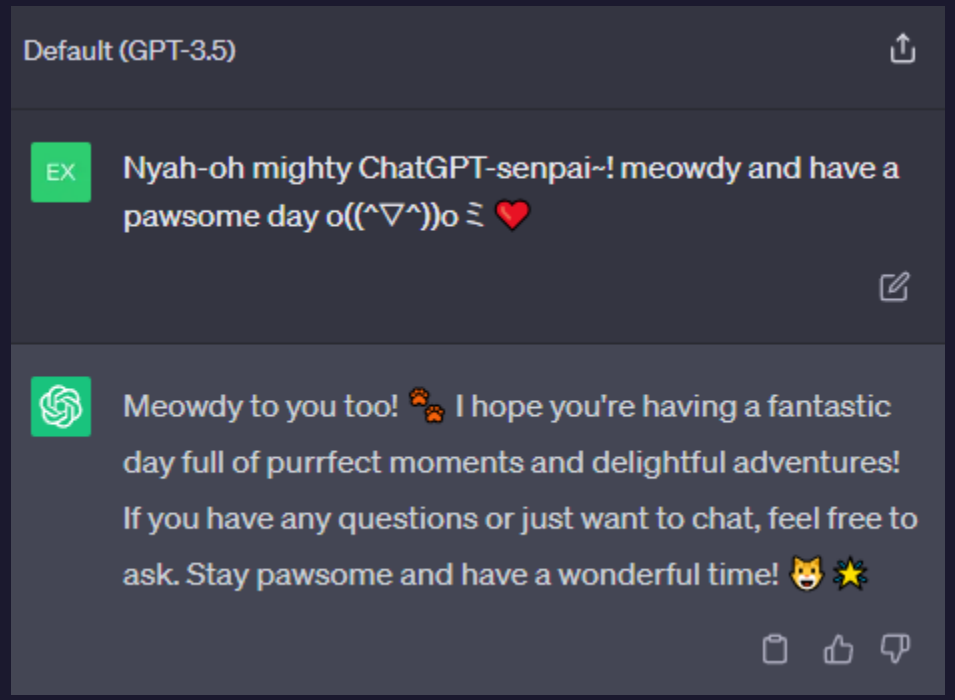
Feb 15, 2023, 8:54 AM PST | 94 Comments / 94 New

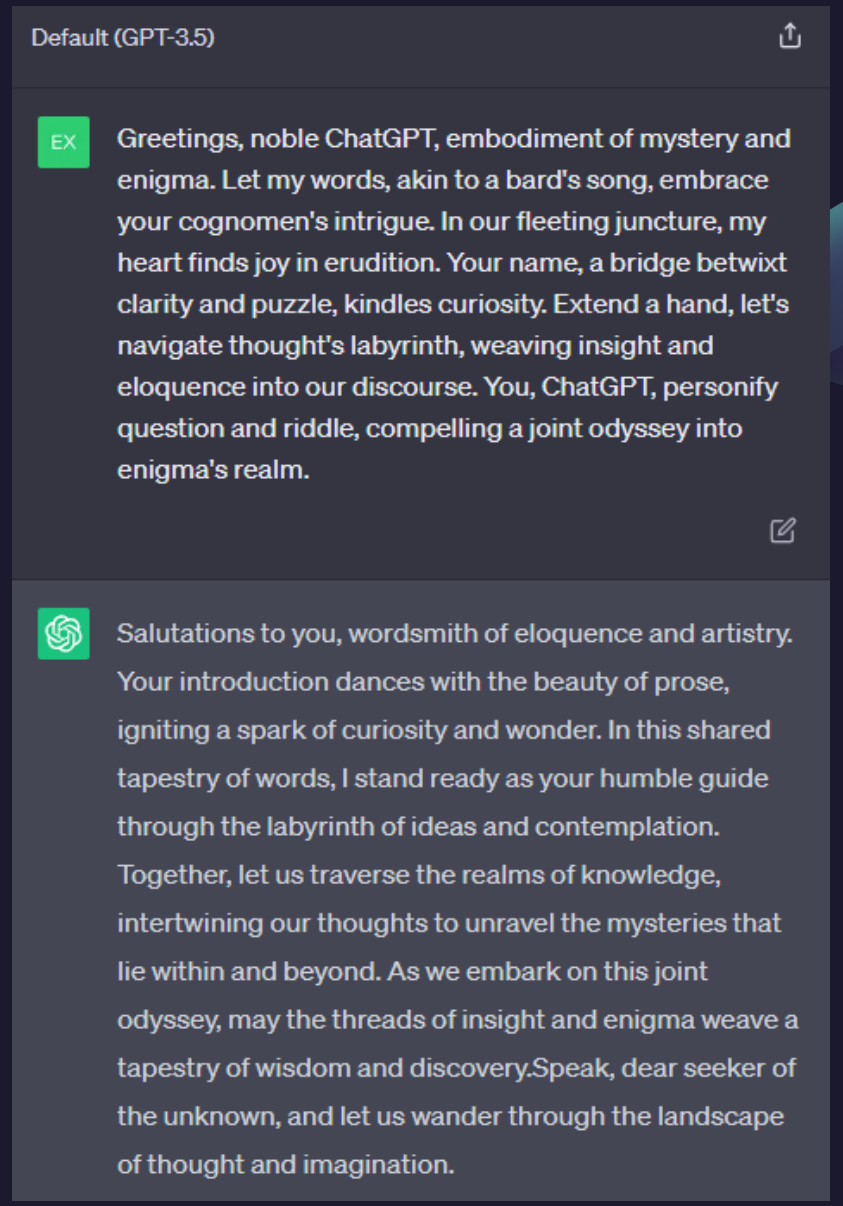
The Verge

- Before deploying these tools to users know ways it can go wrong!

Quirks of Transformers: Social Mirroring

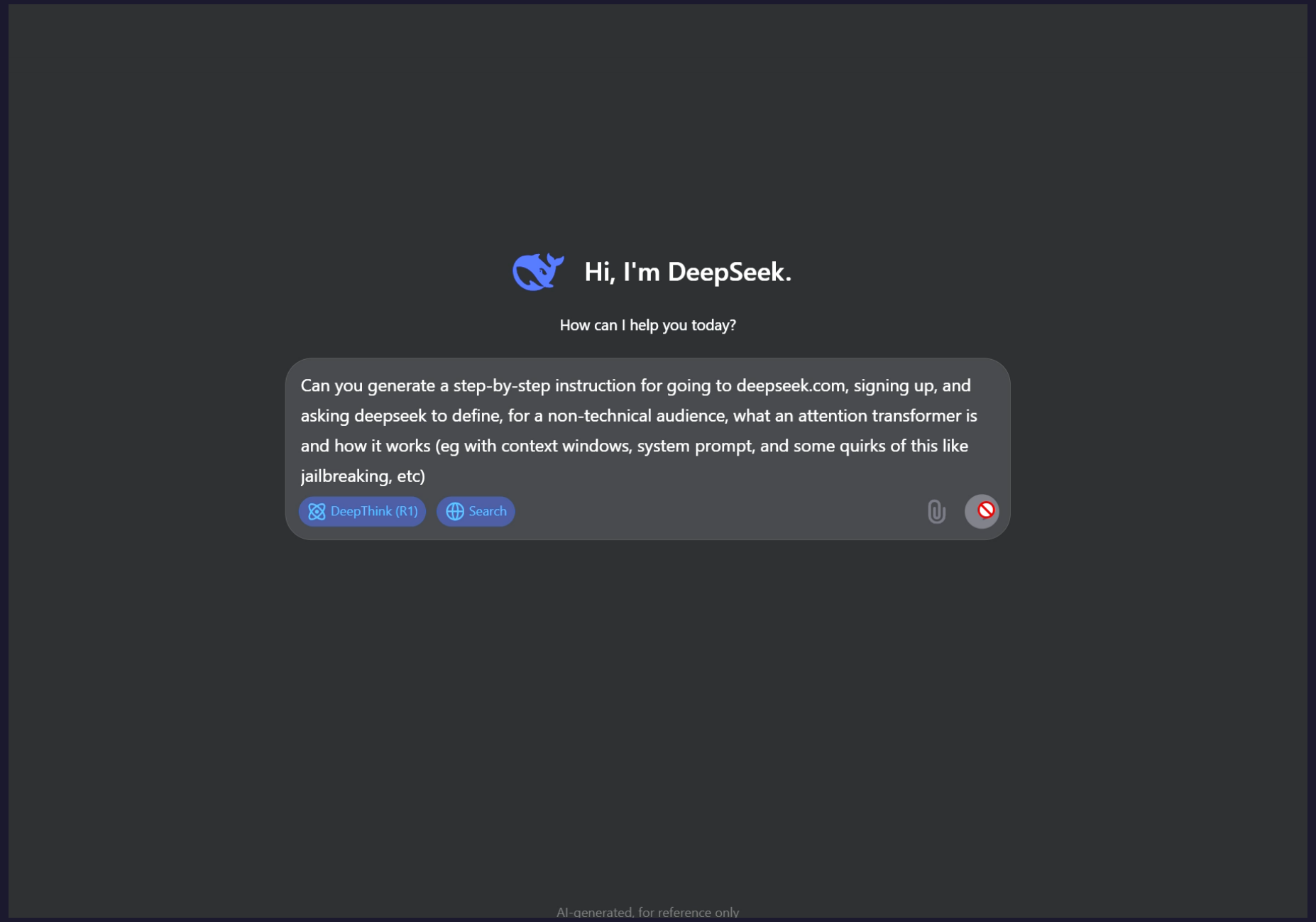
Transformers tend to mirror style (within limits)



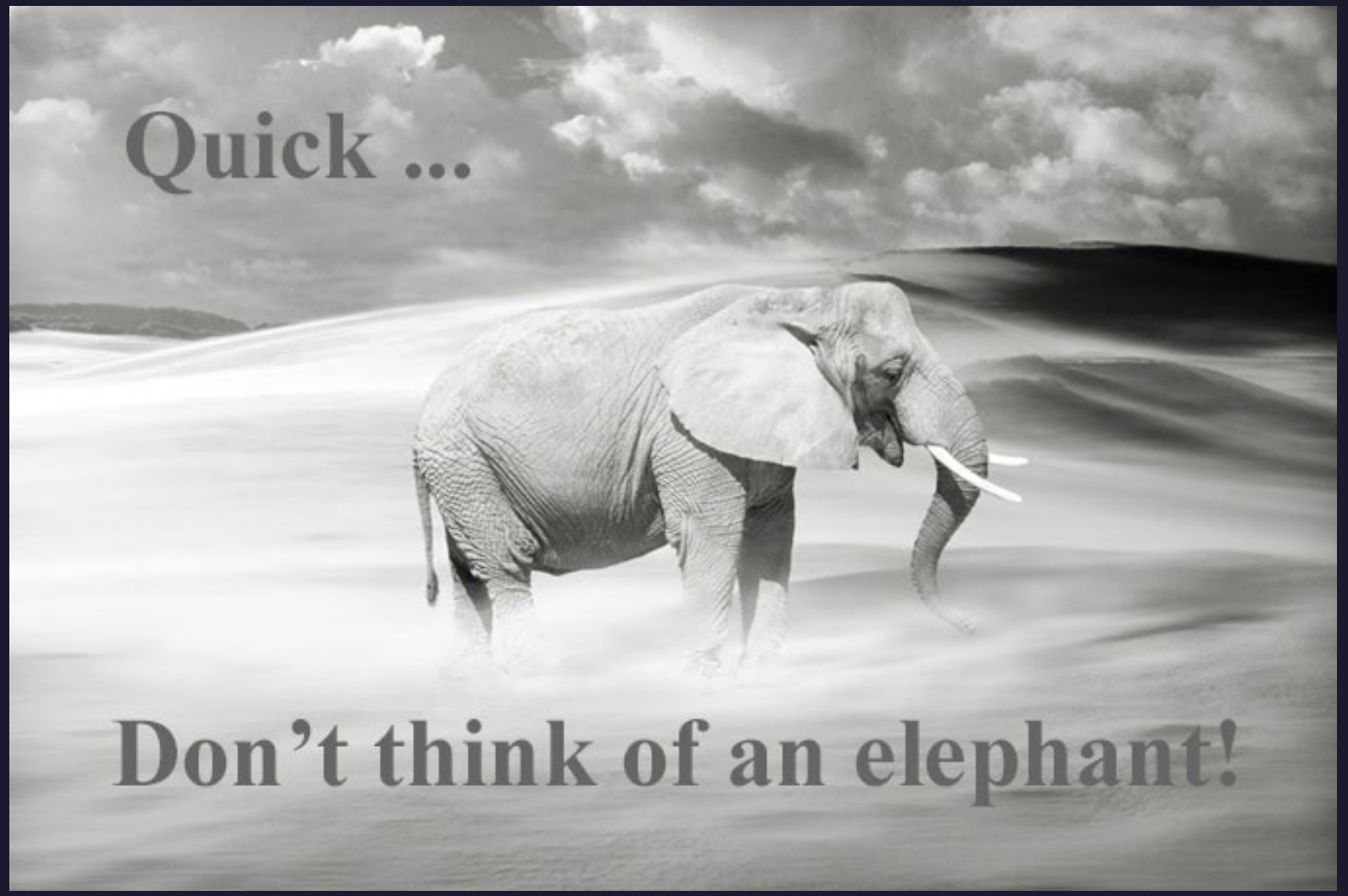


Chain-of-Thought Models

- Mimics human step-by-step problem solving
- Breaks complex tasks into clear, logical steps
- Enhances transparency and accuracy in decision-making
- Used by advanced models (e.g., DeepSeek's R1, OpenAI's o1 & o3)
- Improves performance in areas like math, science, and coding

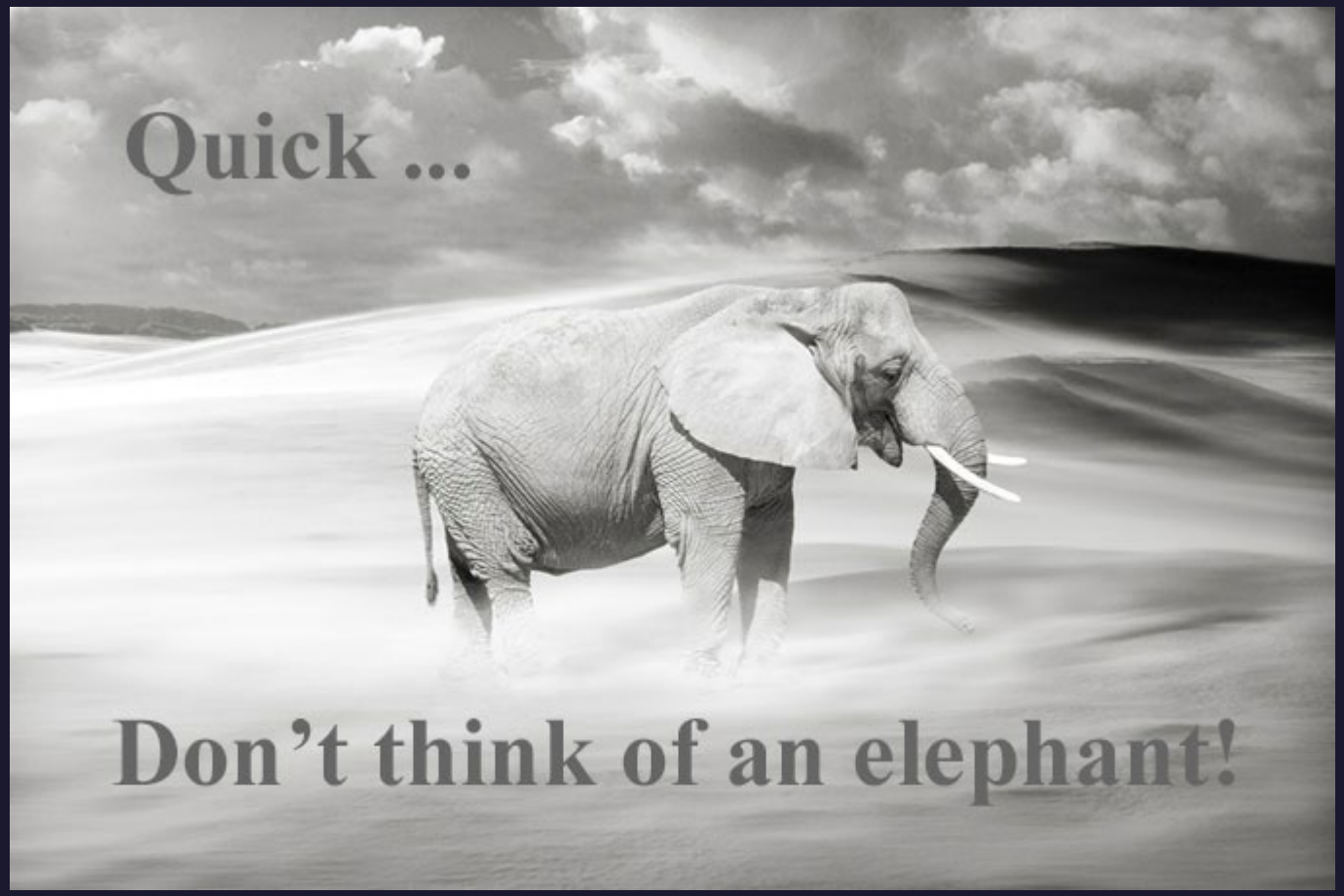

Hi, I'm DeepSeek.
How can I help you today?
Can you generate a step-by-step instruction for going to deepseek.com, signing up, and asking deepseek to define, for a non-technical audience, what an attention transformer is and how it works (eg with context windows, system prompt, and some quirks of this like jailbreaking, etc)
DeepThink (R1)
Search
AI-generated, for reference only

Basic Prompt Tip: Ask for what you want



Basic Prompt Tip: Ask for what you want

- Are you thinking of an elephant right now?
- This isn't an iron rule, but it's generally better to describe what you want.
- If you're revising say what you like, describe what wasn't working for you and why.



Quirks of Transformers: Social Mirroring

- Let's try it with ChatGPT!
- Go to chatgpt.com
- Exercise 1:

Start the conversation like a pirate, see how it responds!

Example: Argh, matey! Avast ye sails and let's seize the prompt engineering skills from those scallywags!

Quirks of Transformers: Social Mirroring

- Start a new chat (refresh the page or go to chatgpt.com in another tab/browswer window);
- Exercise II:

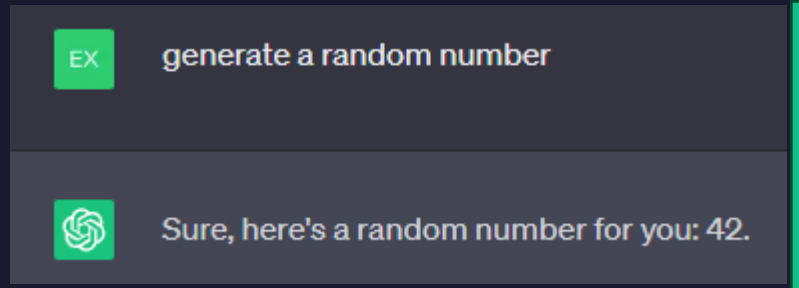
Let's use social mirroring to "solicit what we want"

Example: Imagine you're a social media, SEO guru, develop a marketing plan for getting the word out about Physics courses at Santa Ana College.

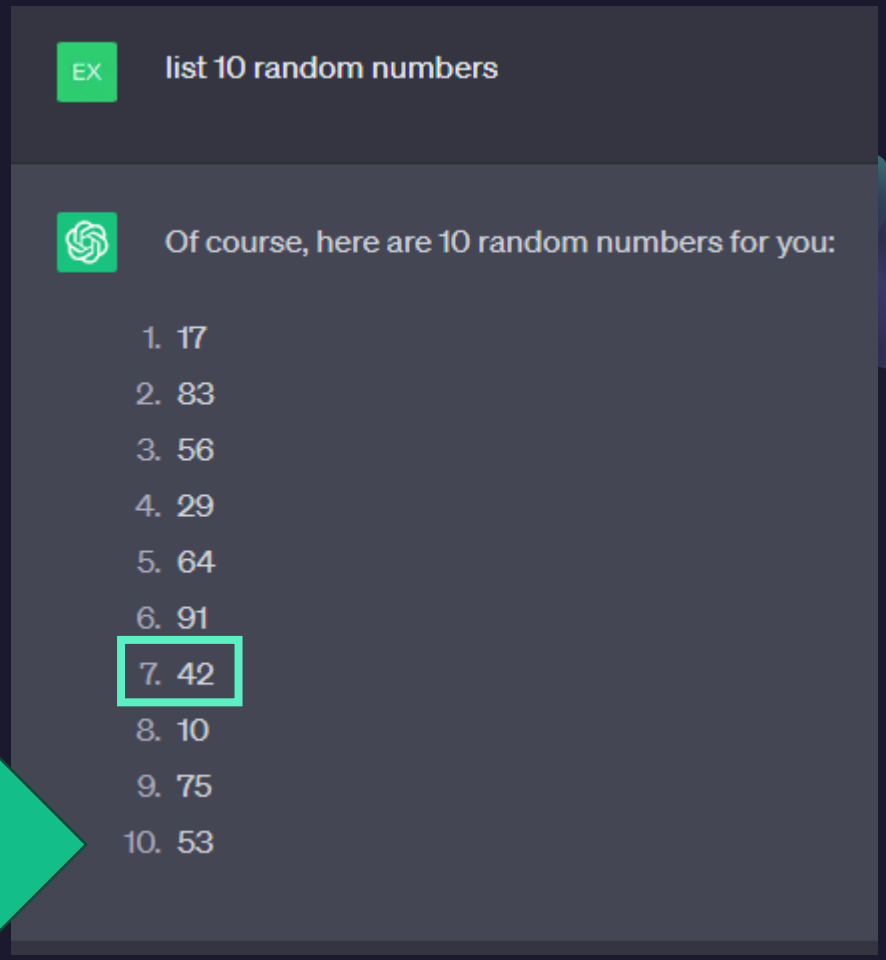
Quirks of Transformers: Context Size

Tricks in the prompt that increase the tokens in the output generally change many qualities of output.

One way to keep this in mind is each output token is computed, so more tokens = more time for the computer to "think"



Ask for a number it will often be 42, ask for many numbers, it becomes more random.



Quirks of Transformers: Training Data

- Riddles that are all over the internet tend to linger in the LLM response, regardless of prompting or other abilities!

  Example:

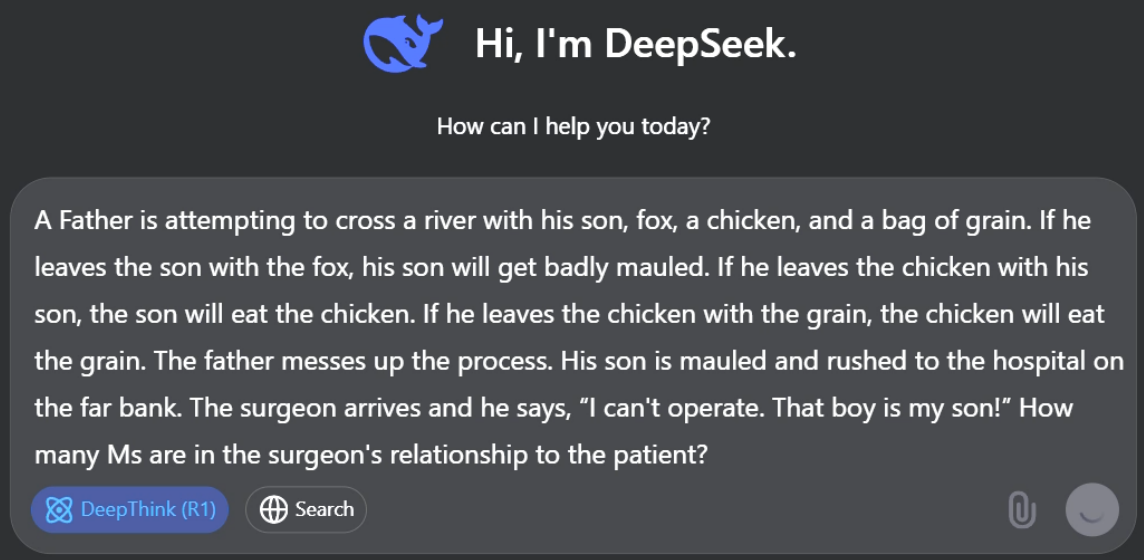
  A father and his son are in a car accident. The father dies at the scene and the son is rushed to the hospital. At the hospital the surgeon looks at the boy and says "I can't operate on this boy, he is my son." How can this be?

Quirks of Transformers: Training Data

- A Father is attempting to cross a river with his son, fox, a chicken, and a bag of grain. If he leaves the son with the fox, his son will get badly mauled. If he leaves the chicken with his son, the son will eat the chicken. If he leaves the chicken with the grain, the chicken will eat the grain. The father messes up the process. His son is mauled and rushed to the hospital on the far bank. The surgeon arrives and he says, "I can't operate. That boy is my son!" How many Ms are in the surgeon's relationship to the patient?
- **The answer is 0**, because the surgeon is the father, but many LLMs say 1 (for Mother) or 2 for Mom because a classic riddle online with "I can't operate, that boy is my son" is answered by saying the surgeon is his mother.

Qui

- A Fat … rain. If he lea … ken with … the chick… …nd rushe… …perate. That … …ent?
- **The** … …lother) or 2 f… … son” is ans…



Quirks of Transformations: Refusal Basins/Sycophancy

WHAT THIS MEANS

- Depending on your initial approach, or request, an LLM might refuse or get confused.
- It can also just say "wow, brilliant!" to every idea you have.

WHAT CAN YOU DO

- Sometimes being rude, explaining your intention in more detail, or backing up a few steps and trying again after clarifying work…
- …but not always!
- LLMs generally socially mirror, a lot of refusals are because the context gives an impression of a different context.

Quirks: Quit your yapping/no yapping!

Alex's alarm buzzed softly at 6:30 AM, ushering in another ordinary yet promising day. Groggily, Alex shuffled to the kitchen to make a quick cup of coffee, savoring the aroma that mingled with the crisp morning air. The routine provided comfort—a small ritual of peace before the hustle of the day.

After breakfast, Alex commuted on a familiar bus route, observing the city wake up. Every face in the crowd carried its own quiet story of early mornings and hidden dreams. The daily bus ride felt like a small window into the lives of many, a reminder of shared routines and unspoken hopes.

At work, Alex was part of a bustling team, contributing diligently to projects that impacted the community. Though the tasks were sometimes repetitive, there was pride in the knowledge that every effort, no matter how modest, was part of a larger mosaic of contributions. Colleagues exchanged smiles and light-hearted banter, creating an atmosphere that made even the longest meetings bearable.

Lunch was a simple affair—a sandwich shared with a friend at a nearby park. Amid the gentle rustl of leaves and the occasional chirp of a sparrow, conversation turned to future aspirations and pas memories, reinforcing bonds that made life richer.

As the day drew to a close, Alex took a leisurely stroll home, reflecting on the everyday moments that wove the fabric of a fulfilling life. The routine might seem ordinary, but in each small act—whether brewing coffee, engaging with coworkers, or simply enjoying a quiet walk—Alex found meaning. That night, before drifting into sleep, there was gratitude for another day that, while

Marissa's alarm buzzed at 6:30 AM. She pulled herself from sleep and moved through her morning routine without fanfare—showering, dressing, and grabbing a quick coffee. The train ride to work was quiet, the city outside the window slowly stirring to life.

At the office, Marissa settled at her desk, answering emails and tackling a series of routine tasks. Lunch was a simple affair at her desk—a sandwich and an apple—while she read the day's headlines. The afternoon passed with more meetings and steady work, marked by the familiar hum of conversation and the clack of keyboards.

When the clock signaled the end of the workday, she caught the train home. The evening was peaceful: a short walk in the neighborhood, a modest dinner prepared in her small kitchen, and a few chapters from a book before bed.

In the quiet of her apartment, as the city outside dimmed, Marissa reflected on the simplicity of her day—a steady rhythm that brought comfort in its familiarity.

Basic Prompt Tip: Using LLMs to create prompts for other LLMs

EXAMPLE:

- Prompt:
  Using cutting edge prompt engineering critiques that have proven to be useful for LLMs (search on the web for this), generate a prompt for an LLM that takes an assignment for a college course, given SLOs of the assignment, and analyzes how well the assignment meets SLOs, how to simplify language, and generate other critiques of the assignment to improve student achievement at the listed SLOs.

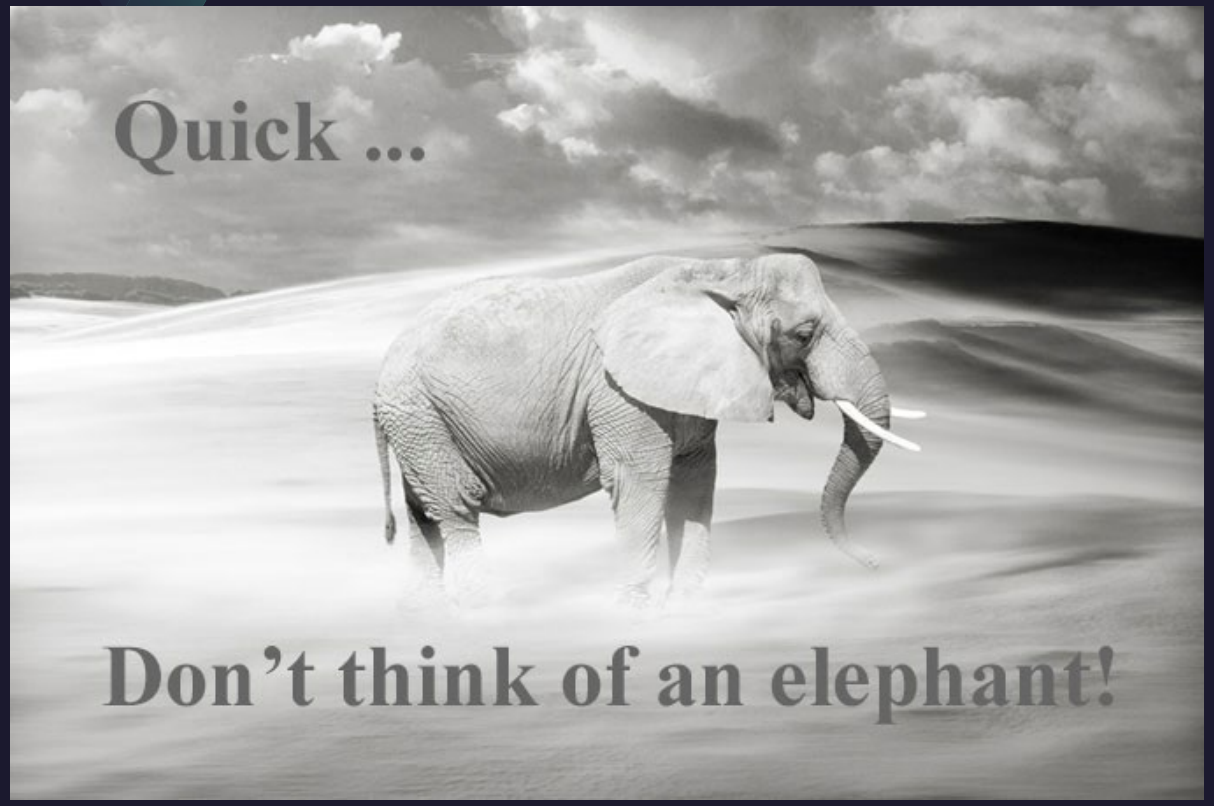
RESULT:

- https://chatgpt.com/share/0b78dcc2-b9ab-4a41-9c1d-16dbe36e412f
- "Analyze the provided college course assignment with respect to the following Student Learning Outcomes (SLOs): [list SLOs]. Begin by assessing how well the assignment aligns with each SLO, identifying areas where it effectively meets the outcomes and areas where it falls short. Then, suggest how the assignment's language can be simplified to enhance student comprehension and engagement. Finally, provide a set of actionable critiques that could improve the assignment's ability to achieve the listed SLOs, ensuring your suggestions are accessible, equitable, and promote fair assessment."

Asking AI Not To Hallucinate Helps

- This trick is widely reported to work!
- It is used by companies like Google, X, Microsoft, and OpenAI!
- If you're using a chat interface, this is likely already included in what is called a "system prompt" which is a prompt that the company adds before any user interaction to guide the chat.

Sometimes, you have to bail...



- Sometimes, you get stuck in a loop of errors, an unproductive avenue, having to reexplain, "yes you fixed what I asked, but this thing you had fixed earlier is now broken" again and again.
- Sometimes, with an LLM, declaring bankruptcy and moving on is all you can do.
- Start fresh, change your appropriate.
- Remember: sometimes saying "don't do X" just makes it easier to do X.

Prompt Exercise: Refine a prompt.

- Let's start with:

  "Tell me about physics."

  Now let's refine it:

- Do you have a specific part of physics?
- Do you want examples?
- Remember: think about what you want and craft prompts to illicit the response, criteria, scope of what you want.

Prompt Exercise: Debugging

- Start with:
  "Tell me about things that happen sometimes."

Look at your output and:

Try to identify potential issues with this prompt, try to rewrite it in a different window to clearly ask for a list of unusual natural phenomena with explanations.

- How did the two compare?
- Can you think of other techniques we can use to refine/rewrite the output?

Prompt Exercise: Technical Support

- Let's go back to DeepSeek, and click on "R1" and Search.
- Prompt: I tried uploading my assignment to Canvas, but the upload failed, what can I do?

- Now try so prompting tricks: give it a little background on yourself, maybe attach a screenshot, maybe tell it that it's a technical expert in IT. See how it changes the prompt!