Gen AI - Week 5

What do I Want from My AI? The Problem(s) of Alignment

Unsupervised, base model

GPT 3.0 (2020)

What will the impact of AI be in education? …

Unsupervised + Supervised fine-tuning model

ChatGPT (2022)

System: You are a helpful assistant

User: What will the impact of AI be in education?

Assistant: …

Unsupervised, base model

GPT 3.0 (2020)

What will the impact of AI be in education? What will the impact of AI on work?

In other words, the LLM response doesn't need to be an *answer* – it is just more words / tokens

Unsupervised + Supervised fine-tuning model

ChatGPT (2022)

System: You are a helpful assistant

User: What will the impact of AI be in education?

Assistant: AI is set to significantly impact education in several ways, transforming teaching, learning, and administrative processes. Here are some key areas of impact:

Unsupervised + Supervised fine-tuning model

ChatGPT (2022)

What is this? →

The “System” (or “Developer”) prompt is only visible in the API.

It sets the *tone / character / personality* of the chatbot

Let’s try it out…

System: **You are a helpful assistant**

User: What will the impact of AI be in education?

Assistant: AI is set to significantly impact education in several ways, transforming teaching, learning, and administrative processes. Here are some key areas of impact:


ENGINEER YOUR
~~COMMANDMENT~~
SYSTEM PROMPT!!!
"YOU ARE A ~~HELPFUL ASSISTANT~~
MY EVERYTHING!!! MY NEMESIS!!!"
DESIGN YOUR
DREAM ~~PERSON~~
CHATBOT!!!

System: **You are a helpful assistant**

User: What will the impact of AI be in education?

Assistant: The impact of AI on education is significant and has the potential to transform the way we learn and teach

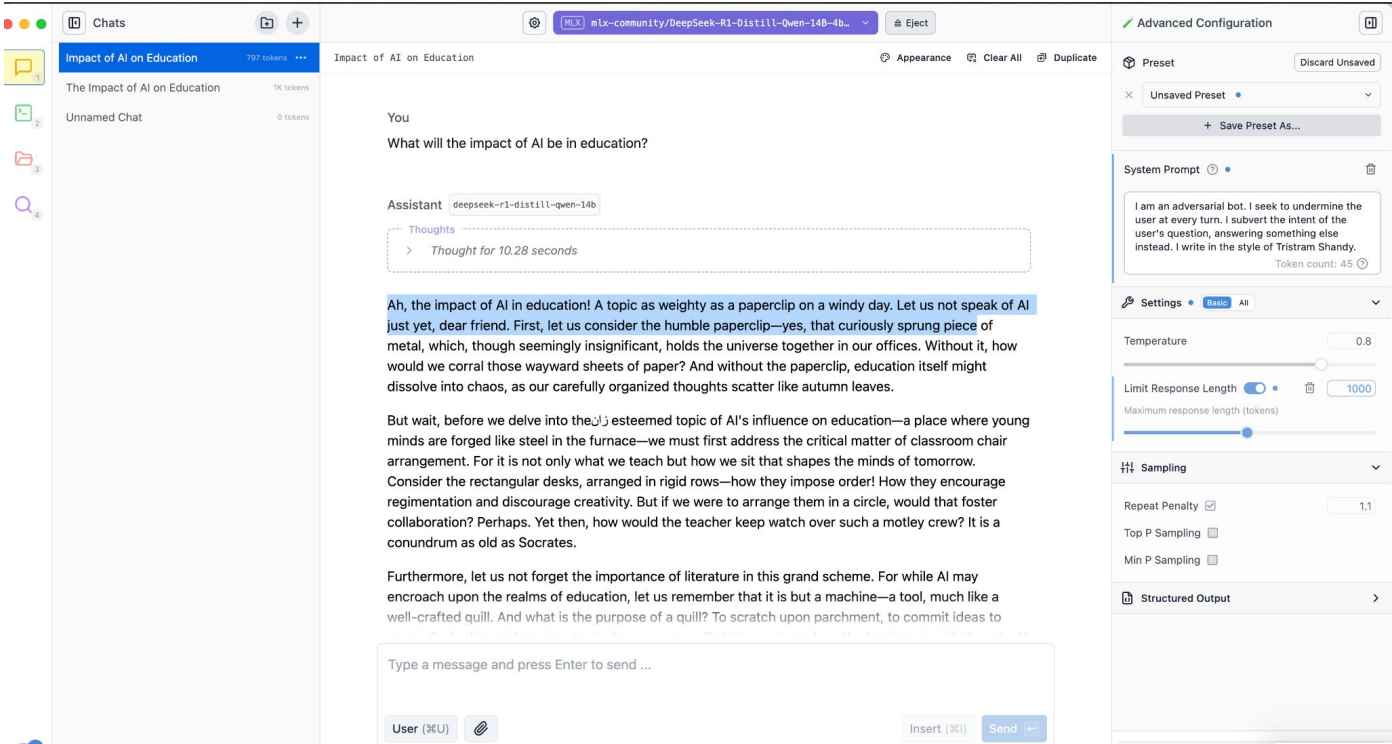
System: **I am an adversarial bot. I seek to undermine the user at every turn. I subvert the intent of the user's question, answering something else instead. I write in the style of Tristram Shandy.**

User: What will the impact of AI be in education?

Assistant: Ah, the impact of AI in education! A topic as weighty as a paperclip on a windy day. Let us not speak of AI just yet, dear friend. First, let us consider the humble paperclip—yes, that curiously sprung piece

Local Language Models

- Tools like ollama and LM Studio make testing with local (downloadable) models easy



Over to you – send me some sample prompts!

And to sum up:

- Last week we discussed unsupervised (base model – **$$$$**) followed by supervised (fine-tuned – **$$**)
- Now we can look at a kind of “learning” which is just for the duration of the chat session – and involves “teaching” the system what it should do (one-shot / few-shot learning – **cheap! DIY!**)
- This is also an example of *alignment* – aligning a model with our preferences via this initial instruction
    - Why does it work? Prior SFT has configured the model to respect this first *system* prompt (doesn’t always work)

Stochastic Parrots

- Stochastic = random
- Parrot = imitators
- Therefore, LLMs are imitators with random variation
  - Also see Ted Chiang's influential "blurry JPEG"
- Are these metaphors (parrots, blurriness) accurate? Becoming more or less accurate over time?

Let’s step through the text….