How AI is Trained

Start with how AI works and is trained, then move to the issues that creates, then move to how that affects prompting.

Character by Character

LLMs are *not* taught language like a child or second language learner. They are not able to understand language like that. Computers don't 'understand'; they are machines reacting to an input. At a basic, very high level, when training an AI, what actually happens it the computer builds a statistical model, following the mechanisms programed into it. To build this model, it is given a data set, which it looks at at the smallest possible unit and then counts how often every possible combination of that unit happens (aa, ab, ac, etc), starting at one-by-one, then sets of two, then three, etc, until it's capped by the capacities of the supercomputer it's running on and/or it's programing. From there it builds a statistical model of what appeared, what is most likely to appear, and what is least likely to appear. Then, the AI outputs are graded on whether they are 'correct' or 'incorrect', either by hand or by specially designed program, with incorrect responses downgraded in the statistical model, until the AI reliably generates acceptable results.

Outputs

For things like "find the cancer" or "ID the pastry" (same model!) it builds an internal model of the targets, and what does/doesn't count as each target, then assigns closest matches. For decision making AI and searching, it gives out the statistically most likely answer that will get a desired result. For gen AI, they take this statistical model, and tell it, "do not regurgitate the mostly likely thing. Randomly create rather likely things based on the statistical model and a randomized starting seed". This is where variation in outputs come from, but also where inaccuracies come from – it is literally making up likely things, rather than regurgitating what we put in it. Non-Gen AI tools are more reliably accurate, but will also always respond exactly the same way. There is then post-training tweaking that happens, when programmers write code that manually changes the AI's outputs. Guardrails prevent the AI from ever putting out certain things in response to certain inputs (like the recipe for napalm), or in models concerned with information accuracy, removing misinformation. Programmers can also increase the likelihood of certain outputs – up to "always" – post training, which is often where 'better at grammar or math' tweaks happen. The genAI is told to artificially increase the count of 'correct' unit combinations, making it more likely put out 'correct' things. And, in case of image genAI, they pit the generation model against a gatekeeping model, that won't show the users things that it's judges are "incorrect".

Where the data comes from

Because this is how AI is trained, the datasets used have massive impact on the AI's output. Depending on the model, the datasets can be created in two ways – either specifically created for training the AI, or scraped from existing media. Datasets created for purpose tend to be used on non-genAI, such as AI that can identify cancerous cells, or help with decision making in specific contexts. GenAI tend to be trained on datasets scraped from existing media, typically texts and images accessible on the internet. Each dataset type has it's own pros and cons; purpose-made datasets are incredibly resource intense to

create, and still not free from error. Scraped datasets run afoul of copyright and intellectual property issues, as well as ethical issues around how creators what their creations used, and often include the worst of what exists on the internet. Additionally, scraped datasets tend to be a fully automated process, with little to no oversight between the scraping and the training; and when they are filtered for inappropriate content, this is done by hand and often in exploitative ways.

Issues from the data

A bad dataset may have systemic gaps, poor coding, or bad sampling, which can create unexpected behavior in the AI. An example is an early AI for assisting doctors make medical decisions kept suggesting they discharge patients with a high likelihood of dying from pneumonia, because in the dataset it was trained on, "discharge from the hospital" and "discharge from normal care to ICU" were both just coded as "discharge". Any bias in the data, even bias so slight a human would be hard pressed to notice, will be replicated, and often worsened, by the AI model, especially if it is statistically likely in the dataset. As many datasets are created by trawling the internet, they often contain bigotry, misinformation, and offensive or explicit material, which the AI will then reproduce. Furthermore, the AI, again, doesn't understand semantic units. It has no ability to understand nuance, context, or grey spaces, and it doesn't understand that words can have multiple meanings. This can create some unusual cross-talk between separate ideas that share words – such as an image GenAI failing to differentiate between a Black woman and a woman with literally black skin, and modern LLMs failing to differentiate between homophobic and homosexual reliably, due to similarity in spellings and them often appearing in close context. AI also doesn't have any real world knowledge or common sense - the reason it doesn't seem to know how many fingers a hand should have, despite that being a pretty stable number in actuality, is due to it having no actual understanding of human anatomy – or physical existence in general – but rather just knows that a specific pattern (finger shape) should be repeated a variable number of times.

Issues from the programers

To make up for the issues from scraped datasets, genAI (typically) has extensive guardrails added to prevent it from reproducing the worst of this material or dangerous or illegal material. For example you cannot directly ask ChatGPT for the recipe of napalm and receive it. However, badly designed guardrails are easy to intentionally or accidentally get around – for a while, ChatGPT wouldn't directly give you the recipe for napalm, but would give you your grandmother's traditional recipe for napalm. Tweaks to outputs and the underlying statistical model can help control the outputs as well, but anyone following the saga of Grok on Twitter/X can see how difficult that is to get right; no amount of tweaking will completely remove the bias trained into the genAI, and it's very easy to both over and under correct. Additionally, the guardrails and tweaks themselves often reflect the bias of those creating them – typically Western, white men – and therefore will censor, or fail to censor, generated responses in a way that reflect a specific view point of what is appropriate. This often results in the over censoring of responses about at-risk and underrepresented populations, and under censoring of alt-right, far-right, and bigoted responses.

Hallucinations

"Hallucinations" – which is a misnomer, a machine cannot hallucinate – come from three sources, all of which are the AI behaving as programmed, with results we did not expect. One source is the fact that

genAI is designed to make up stuff, and so it… makes up stuff. It doesn't 'believe' that non-existent articles are real – it doesn't know that articles are things that exist – it just got told to give out some articles, and, as programed, gave a mix of ones that are already in the dataset, and ones that sound likely. The second is from patterns in the dataset that we aren't aware of – such as the hospital discharge one. This can be due to bad datasets, datasets that are skewed in some way, or simply patterns that we as humans failed to recognize or understand to be noise. Lastly, it can be due to the fact that the AI simply doesn't know what it's working with. Google Bard for a while was insisting that Trump "is not straight", due to its inability to meaningfully distinguish homosexual from homophobic. It was not programmed to understand what a word is, nor to assign actual meaning to any of them. What it does know is that collections of characters that are mostly the same are often interchangeable (due to grammatical markings on words) and that some collections of characters that are completely different are often interchangeable (synonyms), and behaved accordingly.

Tools

So, now that we've gone over the pitfalls, lets talk about getting what you want. First, and most importantly, use the right tool. ChatGPT is not a swiss army knife – it cannot do everything, or do many things well. Use Wolfram Alpha for math; it will not only give you a correct answer, but can step by step walk you through how to get that answer. ChatGPT doesn't know what articles are, Use LitMaps for getting article suggestions and citations; you have access to a pro account through your school log in, and you can use it to find articles relevant to any topic. ChatGPT *is* pretty good at summarization – though again, it can't do something like understanding themes of a literary work (for that, SparkNotes is still a good resource) – and can it can check grammar, spelling, and work as a starting point/idea generation. Additionally, there isn't an AI out there can problem solve, or that has common sense. It literally doesn't know that these characters have meaning, or that there is a real world for them to relate to. It can, at best, happen to regurgitate a helpful answer a human once put out on the internet somewhere – but it probably won't.

Prompts

And this is why, the only AI responds well to natural language is chat bots, because that's the only thing they can do. For prompting GenAI, you must keep in mind that you are not talking to a person, or a human mind – you are putting characters into a computer, and it is responding in kind. Prompting requires highly specific language, without nuance, ambiguity, or assumptions. The AI cannot work through what you meant. Be as detailed and exact as possible – painfully so. If you are not annoyed with how specifically detailed your prompt is, it's not specific enough. Use the most standard, non-slang words for things, and make sure you spell it correctly– SWINE (Standardized White Inland Northern English) is what is most common in the datasets it's trained on, and it will struggle with anything else. It will take multiple iterations to get something usable. Your prompt will likely be a page long. It will never put out the same thing twice – it's designed not to! Finally, most AI producers are constantly updating the models on the backend, typically with no release information to the public, so expect something to not work the same the next day.