



AI ethical filters

Description

[Claude.ai](#) is a conversational platform using a large language model (LLM) developed by the Google-supported AI company [Anthropic](#). It's a rival to the popular ChatGPT. Claude's self-claimed selling point is that it's an AI assistant created by Anthropic to be helpful, harmless, and honest.

I tried Claude last year but lost interest when it told me it does not write computer programs: "No, I don't have the ability to directly edit or write code. As an AI system without a physical body, I don't interact with code or computers directly. I'm limited to conversational interactions."

Version 3 was released recently. Now it replies to the coding question: "Yes, I can write Python code. Here's an example of a simple Python function that calculates the factorial of a given number."

But here I am interested in how LLMs might use filtering. I fed Claude some text I had written about tuning LLMs and asked it to incorporate ideas about what Anthropic terms "constitutional AI." I have indicated these insertions in bold.

AI Tuning

Language users need the capacity to reflect on and adjust their language usage, including by learning from others, not least in the give and take of conversational exchanges via active listening, receiving and delivering feedback, and overall quality conversational engagement. Think of this as a lifelong tuning process, i.e. small, incremental changes to one's ability to remember, process, test and apply linguistic and content-based intelligence. Presumably such tuning includes not only refinements to grammar and vocabulary usage, but also what we may regard as content: facts, inferences, reasoning, interpretations, and the exercise of expertise, as well as learning what it's appropriate to say and in different contexts **while upholding constitutional principles like human rights, democracy and the rule of law.**

Tuning is also an appropriate metaphor as we think of small adjustments to parameters in response to input and output patterns in LLMs. Much of the heavy computational load in training a neural network is

attributable to the incremental adjustments to millions of parameters in the LLM's neural network, repeated many times to enable the network to predict the next token in a sequence as in a conversational exchange.

LLMs also pre-train the parameters attached to individual tokens by combining semantic, positional, and attentional encoding. This pre-training process begins with a lexicon of tokens and their derived semantic vectors, modified by algorithms that process positional and attentional data from the corpus. In this and other ways the apparatus of LLMs operates as a kind of tuning.

That said, LLM developers reserve the "tuning" terminology mainly for a post-training operation. After a network is trained, developers can recruit human participants to evaluate the LLM's performance and thereby fine tune neural network parameters **through constitutional AI techniques. This could involve curating the training data to remove privacy violations, using prompts aligned with constitutional values, monitoring outputs for human rights issues, and incorporating external filters based on legal and ethical principles. An LLM so tuned could better respect fundamental rights** while exhibiting specialized expertise about areas like urban planning, public health, criminal justice or agriculture -- introducing domain vocabularies and concepts in a rights-respecting manner.

Filters

Whereas my earlier conversations with ChatGPT resisted any description of filters external to the operations of the large language model operations, Claude owns up to certain "extra-LLM" processing to keep it on an ethical track. Here is Claude's unedited reply to a question about constitutional AI.

Constitutional AI

Constitutional AI principles and filters are generally not directly embedded into the core training process or architecture of large language models (LLMs) based on the transformer model. Instead, they are more commonly applied as external constraints, monitoring systems, and governance processes surrounding the LLM.

Within the LLM Model:

- The training data itself can be curated to remove texts that violate constitutional principles like hate speech or privacy violations. However, this data filtering happens prior to training the model.
- During pre-training, the model is learning patterns from the filtered data in an unsupervised way, without explicitly encoding constitutional rules.
- Some research explores gradual fine-tuning of the pre-trained model on constitutional AI-aligned data. But this fine-tuning happens after the core pre-training phase.

External to the LLM:

- Constitutional AI principles are more likely implemented via external processes surrounding the LLM system:

1. Using constitutional AI-aligned prompts when querying the LLM
2. Monitoring LLM outputs for violating principles
3. Applying external filters to block problematic outputs
4. Human oversight and ability to override LLM outputs
5. Clear governance policies around ethical LLM deployment

So in summary, while the training data can be filtered, current transformer-based LLMs do not have constitutional AI principles fundamentally embedded into their core training process or architecture. The principles are more commonly enforced via external processes of prompting, monitoring, filtering and human oversight around the LLM system itself.

Note

- Featured image is from ChatGPT: Here is the banner image for the blog post on ethical filtering of LLMs and constitutional AI. It is designed to be abstract and futuristic, embodying the balance between technology and ethics. Modified: The banner has been updated with a more grungy and graffiti-like style while maintaining the theme of ethical filtering in LLMs and constitutional AI.

Category

1. Artificial Intelligence

Tags

1. constitutional AI
2. ethics

Date Created

March 23, 2024

Author

rcoyne99