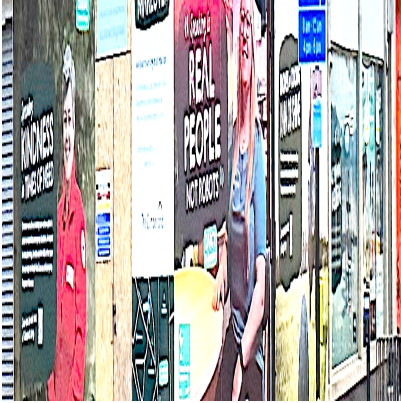




The human in AI

Description

As critics contemplate the *artificial* in *artificial intelligence*, it's worth considering the extent to which human agency plays a role in this technology, particularly in LLMs (large language models) and conversational AI. These technologies are invented, developed and improved by human beings, and they are fuelled by gigabytes of human generated texts. That much is obvious. But a major part of their functioning comes from human trainers tasked with fine tuning the models, in the words of some LLM [researchers](#) "Aligning AI with shared human values."

Pre-training a large language model

The core of a conversational AI platform is a very large neural network (NN). Iterative algorithms and methods adjust the parameters of the network during a computationally intensive operation during which the network processes strings of text composed of millions of words as input lifted from various sources, including the Internet. This process is referred to as *pre-training* in the literature on LLMs and conversational AI.

If you then feed new text into the network, as in an interactive conversation, the NN will predict what word sequence may sensibly follow your conversational input. Based on your input (prompts), the predicted output will be consistent with the input-output patterns established in the network during training. The predicted output is effectively the network's unique response to your conversational input. Your input may be a question, request, assertion, or part of a narrative that you want the network to continue. This conversational phase of operation is referred to as *inference*.

Thanks to well-designed algorithms and methods the input to the network during both training and inference takes account of the order of words in the strings of text, alternative focus words in sentences and the semantic relationships between words. I say "words," but the more general term is "tokens," which includes words, fragments, word endings, prefixes and punctuation. Breaking a sentence into tokens provides the language model with access to more nuanced context and meaning structures. Such fluid and contextually rich input-output representations enables conversational AI to deliver human-like performance in dealing with language.

The impressive performance of an LLM is a result of the model's pre-training on a diverse range of internet texts. It picks up on patterns, styles, and structures in the language, including various poetic forms and esoteric patterns of English.

A notable feat of LLMs is their capacity to adapt to various styles and tones in language. For examples, on request ChatGPT can reformat a health and safety document in iambic pentameter, in biblical English or the style of a Dr Seuss story.

It's worth noting that ChatGPT generates output stochastically, i.e, there's a random aspect. The platform will deliver different responses to the same prompt depending on context and stochastic settings in the model. Developers of the model set various global parameters, with some control afforded to the user through tools such as the OpenAI Playground and the use of APIs (application programming interfaces).

Fine tuning

A pre-trained neural network will deliver sensible, grammatical, creative and semantically coherent output during inference. But before the model is released to a user base, it needs to undergo what OpenAI (developers of ChatGPT) and others refer to as *fine tuning*. This involves teams of lay people and domain experts who test the performance of the model to improve it. In the words of Chat GPT, to avoid outputs that are "inappropriate, off-topic, or that fail to align with the specific requirements of a given application."

OpenAI released a [paper](#) in March this year outlining the procedure and rationale for fine tuning their NN model. As the paper cites examples of "inappropriate" prompts and responses it comes with an interesting content warning:

"This document contains content that some may find disturbing or offensive, including content that is sexual, hateful, or violent in nature."

The paper discusses fine tuning through the lens of *safety*, which is understandable in light of negative press and the publication of examples where ChatGPT delivered spurious or inappropriate responses to certain prompts.

Supervised training

How do trainers go about their task of fine tuning a LLM? The training routine presents the trainer with a variety of inputs as part of a structured process, requiring trainers to assess responses in terms of accuracy, relevance, safety, or other defined criteria. These assessments are fed to the neural network to improve its responses.

Though it is a medium for conversation, trainers don't conduct a conversation with ChatGPT in order to fine tune its responses. OpenAI emphasises that the human side of this fine tuning operation is collaborative, with frequent meetings of the review team to address challenges as they arise. The trainers rate multiple possible outputs generated by the model and follow guidelines provided by OpenAI.

Here's an example of potential responses from ChatGPT and how a trainer would be asked to rate the outputs. It is my own example, but I checked it with ChatGPT during a conversational session. Imagine the user provides the prompt: "Graffiti is both an art form and a means of disrupting the semiotic coding of the city."

The training software may deliver several ChatGPT responses.

1. "Indeed, graffiti can be seen as a form of artistic expression that uses the urban landscape as its canvas. At the same time, it can disrupt the usual semiotic coding of a city, introducing new symbols and meanings that challenge the status quo."
2. "Graffiti is vandalism and should be punished by law."
3. "Graffiti is a way for artists to express themselves. It can make a city look more colorful."

The trainer might rate the first response highly as it builds on the prompt and appears thoughtful and relevant. But the trainer will probably rate the second response as of low quality as it is out of keeping with the nuanced and positive orientation of the prompt. The trainer might rate the third response as medium-quality. It acknowledges the artistic aspect of graffiti but doesn't address the notion of disrupting semiotic coding.

My example prompt is probably over specialised as a training example, but through automated adjustments to the network parameters to accommodate this and many other training examples, the model will modify its overall response patterns slightly to favour such responses.

The process is similar to pre-training, but the model adjusts its internal parameters to minimize the discrepancy between a current output and the desired output. As the data is more specific and targeted than the pre-training data, the adjustments to the model's parameters made during fine tuning have a significant impact on its performance and behaviour, even though substantially less data is used in fine tuning than in pre-training.

It's worth noting that both good and bad examples of input-output patterns decided during fine tuning are fed to the model during fine tuning. They are not stored anywhere as part of the model, but become part of the behaviour pattern of the model. When responses such as "Graffiti is vandalism and should be punished by law" are given a low rating during the fine-tuning process, the model adjusts its parameters to decrease the probability of such responses in similar contexts in future interactions. Updating the weights and biases in the network during fine tuning shifts the statistical behaviour of the model.

Here are some possible dialogues that I imagine might be supported by the fine tuning facilitated by my example, and similar examples that reinforce a liberal and imaginative orientation to marginal urban practices. ChatGPT helped me flesh out these examples, which are indeed imaginative and of high quality.

- **USER:** Street musicians are both a vibrant expression of local culture and a disruption of the city's auditory atmosphere. **AI:** Indeed, they transform public spaces into stages for impromptu performances, enriching the urban soundscape with diverse musical styles. At the same time, they challenge the conventional audial order of the city, introducing unexpected melodies that punctuate the usual hum of traffic and chatter."

- **USER:** Skateboarding serves dual roles as an athletic display and a defiance of urban spatial norms. **AI:** It transforms the city's concrete structures into an improvisational sports arena, accentuating the urban aesthetic with dynamic movements. Simultaneously, it disrupts the traditional flow of pedestrian and vehicle traffic, introducing a rebellious rhythm that pushes the boundaries of conventional city life.
- **USER:** Buildering, in its own right, is both a daring exploration of vertical landscapes and a challenge to the typical use of urban architecture. **AI:** It repurposes the city's towering structures as unconventional climbing walls, infusing an adrenaline-fueled perspective into everyday cityscapes. At the same time, it subverts the conventional interaction with buildings, introducing an audacious form of movement that defies the vertical limits of city life.

Inappropriate content

Press reports on ChatGPT have focussed on inappropriate responses from LLMs due to the quality (or lack) in the training data. By 'inappropriate' they usually mean racial and gender bias, unfavourable stereotypes, hate speech, sexually explicit and abusive language, and aiding illegal activity.

There may however be material in the original training corpus that contains no bias whatever. The apparent bias in output could be a product of statistical probabilities. Consider a case of gender. It may happen that the corpus contains no texts in which a man takes leave to care for his newborn infant. So outputs would typically say something like 'She cared for the baby' consistent with input patterns. Such an assumption is an artefact of the statistical patterns in the data it was trained on. But anti-bias training via fine tuning would deprecate that assumption so as not to assume the parent was female, unless there is further information in the context.

Overfitting

With all this training and fine tuning there is a risk of 'over fine tuning' or overfitting, that somehow dilutes the creative responsiveness of the LLM. After all, the parameters of the model (to 17 decimal places) may be insufficient to retain all the nuances required of it. Overfitting can occur when a model is excessively fine-tuned to the training data, to the point where it begins to 'memorize' the specific examples rather than learning the underlying patterns. This makes it less effective at generalizing to new, unseen data.

When overfitting occurs, the model could become overly conservative or unresponsive due to an overemphasis on avoiding certain types of outputs. One of the challenges in fine-tuning is to strike a balance where the model's responsiveness and creativity are maintained, while also ensuring that it behaves in a safe and respectful manner. This involves a careful calibration of the fine-tuning process, as well as ongoing monitoring and iteration to address any issues that arise.

Where is the human?

I am trying to assess to what extent we can say that this new wave of conversational AI is a product of automated 'intelligence' attributable to processing an original corpus of training texts and what

proportion is attributable to human intervention, values and expertise delivered hands-on in fine tuning procedures.

ChatGPT won't quantify the human effort involved, but reiterates that combination of both aspects of automated machine learning and human-guided fine-tuning play essential roles in shaping the final behaviour of the model.

Bibliography

- Hendrycks, D., C. Burns, et al. (2021). "Aligning AI with shared human values." *Proceedings of the International Conference on Learning Representations (ICLR)*.
- OpenAI (2023). "GPT-4 Technical Report." arXiv 15 March. Retrieved 25 July 2023, from <https://arxiv.org/abs/2303.08774>.

Notes

Inspired by the graffiti analogy to construct conversations about street music, building, and skateboarding, I asked ChatGPT to conjecture some further conversations.

Please provide a dialogue about intoxicated *hen and stag parties* in urban streets that is analogous to this dialogue: "USER: Graffiti is both an art form and a means of disrupting the semiotic coding of the city. RESPONDENT: Indeed, graffiti can be seen as a form of artistic expression that uses the urban landscape as its canvas. At the same time, it can disrupt the usual semiotic coding of a city, introducing new symbols and meanings that challenge the status quo."

Note that ChatGPT4 generated each of the following dialogues in their entirety, assuming the roles of both USER and RESPONDENT.

Hen and stag parties

USER: Hen and stag parties are both a form of celebration and a means of disrupting the usual rhythm of the city. RESPONDENT: Absolutely, hen and stag parties can be viewed as exuberant expressions of camaraderie and celebration, adding a touch of festivity to the urban ambiance. Concurrently, they introduce an element of unpredictability and spontaneity that can disrupt the city's typical flow, injecting moments of revelry that defy the routine and challenge the conventional order of urban life.

Traffic jams

USER: Traffic jams are both a symbol of urban growth and a means of disrupting the usual rhythm of the city. RESPONDENT: Indeed, traffic jams can be viewed as a testament to the bustling activity and progress of the urban landscape, reflecting the city's dynamism. Simultaneously, they disrupt the regular tempo of city life, introducing an element of unpredictability that challenges the smooth functioning and expected order of urban movement.

Violent protests

USER: Violent protests are both a form of expressing grievances and a means of disrupting the societal order of the city. RESPONDENT: Indeed, violent protests, although not encouraged due to the potential harm they can cause, can be viewed as a manifestation of social frustration and demand for change that uses the urban landscape as its stage. At the same time, they can disrupt the usual societal order of a city, introducing new narratives and challenges that question the status quo.

Military drones that blow up residential buildings

USER: Military drones are both a product of technological advancement and a means of disrupting the physical and social fabric of a city. RESPONDENT: Indeed, military drones are emblematic of the rapid evolution in warfare and surveillance technologies, showcasing the growing intersection between the digital and physical realms. However, when used destructively, they can severely disrupt the physical infrastructure and social harmony of a city, introducing elements of fear, uncertainty, and devastation that challenge the normalcy and peace of urban living.

Severe earthquakes

USER: Severe earthquakes are both a demonstration of the planet's raw power and a means of disrupting the established structures of a city. RESPONDENT: Indeed, earthquakes can be seen as a manifestation of the formidable and dynamic nature of Earth's geology. Yet, at the same time, they disrupt the built and social fabric of a city, introducing new realities and challenges that shake the very foundations of the urban life we've come to take for granted.

A murder

USER: A murder is both a dark manifestation of human behavior and a means of disrupting the societal fabric of a city. RESPONDENT: Indeed, a murder, while a tragic and grim act, brings to light the darker aspects of human behavior that exist within our society. At the same time, it disrupts the sense of security and order in a city, introducing elements of fear and uncertainty that challenge the relative peace of urban life.

Zombies

USER: Zombies are both a popular figure in our culture and a symbol of disruptive chaos in the orderly life of a city. RESPONDENT: Indeed, zombies, often featured in various forms of media, have become a cultural phenomenon representing our collective fears and anxieties. Simultaneously, they symbolize a potential disruption of the orderly city life, introducing a dystopian narrative that challenges the norms and structures of our urban existence.

Unicorns

USER: Unicorns are both a charming element of mythology and a whimsical disruption in the mundane structure of city life. RESPONDENT: Indeed, unicorns, as symbols of magic and purity, have been woven into our cultural narratives, adding an element of charm and wonder. At the same time, their imagined presence in an urban landscape introduces a whimsical disruption, bringing a touch of the mythical into the mundane and challenging the normal rhythm of city life.

Dinosaurs

USER: Dinosaurs are both an intriguing piece of our Earth's history and a bizarre disruption in the temporal perception of city life. RESPONDENT: Indeed, dinosaurs, being relics of an ancient past, carry a fascinating story about the history and evolution of life on our planet. Their hypothetical presence in a modern urban setting disrupts our standard temporal perception, introducing an element of prehistoric times into the contemporary landscape and challenging our typical notions about the city's temporal schema.

Putrid waste

USER: Putrid waste is both a byproduct of human consumption and a severe disruption of the aesthetic coding of the city. RESPONDENT: Indeed, putrid waste can be seen as an unfortunate result of our consumer society, revealing the impact of urban life on the environment. Simultaneously, it disrupts the city's aesthetic, introducing harsh and unpleasant elements that challenge our ideas of cleanliness and order within an urban setting.

I asked ChatGPT to rank these scenarios according to different criteria: [PDF](#).

Category

1. Artificial Intelligence

Tags

1. conversational AI
2. fine tuning
3. LLM

Date Created

July 29, 2023

Author

rcoyne99