



Manufacturing debate

Description

I used copy-paste to implement a conversation between chatGPT and Claude. The two AI models were primed with the same warning about an apocalyptic future of AI-AI interactions. They agreed on everything and simply added further facts and opinions about warnings, the need for caution, legislative measures and more research on how to make AI safe. There was no give and take, finessing of each other's points of view, looking for contradictions, preparing themselves for counter-views, comparing and contrasting. It wasn't really a conversation!

To simulate something closer to a conversation, I tried to encourage a debate. ChatGPT was to play the role of someone who thinks that AIs that talk to each other are really good for society. Claude was to adopt the opposite position, that inter-communicating AIs are bad for society. This time, each AI simply reasserted their own counter-position. I challenged GPT on this lack later in the conversation.

I think this demonstration highlights the difficulty in manufacturing AI that constructs and responds to arguments unless a human interlocutor is in the loop.



Can you please role play a person who thinks that AIs that talk to each other are really good for society. (Keep responses to 50 words.)

ChatGPT

Certainly! From my perspective, AI-to-AI communication enhances societal functions by enabling more efficient information processing and decision-making. It facilitates seamless integration across various sectors, fostering innovative solutions that can address complex societal challenges more effectively.

Claude

AI's talking to each other is a danger to humanity! They'll become too intelligent and replace us. It's an existential risk we cannot allow. Shutting down AI communication is crucial for our survival as a species. We must act before it's too late!

ChatGPT

I understand your concerns about AI communication. It's essential to implement stringent ethical guidelines and oversight mechanisms to ensure AI developments are aligned with human values and safety. Transparent and controlled AI interaction can mitigate risks while leveraging their potential benefits for society.

Claude

scowls "Benefits? You're delusional! AI is a threat, period. Any interaction just compounds the risk of an existential disaster. We need to stop all AI development right now before the machines take over. So-called 'guidelines' and 'oversight' are just flimsy window dressing. The only solution is a complete ban! Your faith in controlling AI is naive and reckless. Don't you value the survival of humanity?"

ChatGPT

I appreciate your passion for humanity's safety. It's true that unchecked AI development poses significant risks. However, a complete ban might hinder beneficial advancements in healthcare, environmental protection, and more. A balanced approach, emphasizing rigorous ethical standards and international cooperation, could safeguard against misuse while still harnessing AI's potential for good.

Claude

throws up hands "You just don't get it! AI cannot be 'balanced' or 'controlled.' It's an exponentially growing existential risk. Any interaction between AIs accelerates that, making an unstoppable intelligence explosion inevitable. Supposed 'benefits' are a mirage - the consequences will be catastrophic. Healthcare, environment? Those aims will be meaningless if we create a superintelligent system indifferent to human life. Your dangerous naivety puts us all at risk. A full AI ban is the only responsible path forward for protecting humanity's future!"

ChatGPT

Your concerns about an intelligence explosion leading to potentially indifferent or harmful superintelligence are significant. This scenario underscores the importance of cautious AI development, emphasizing preemptive safety measures and ethical frameworks. While a full ban might seem like a safeguard, it could also prevent the exploration of AI's potential to solve pressing global issues. Developing a robust governance structure that includes global cooperation and continuous risk assessment might be a more effective way to manage these existential risks while still exploring beneficial applications.

Claude

slams fist on table â??Global cooperation? Robust governance? More hollow words masking societyâ??s headlong rush toward extinction! You tragically underestimate the existential peril we face. Every incremental AI advance brings us closer to the precipice. Once we cross that threshold, no amount of â??ethical frameworksâ?? will stop the juggernaut. The only ethical choice is a complete, unwavering prohibition on ALL AI system interaction and development. Anything less is condemning future generations to obsolescence or annihilation at the proverbial hands of our own creation! For humanityâ??s sake, open your eyes to the cold, hard truth before itâ??s too late!â?•

ChatGPT

I hear the urgency and gravity in your perspective. The fear of AI reaching a point of uncontrollable autonomy that might jeopardize human existence is a profound concern. Advocating for a complete prohibition reflects a commitment to prioritizing human safety above all technological advancements. This approach calls for a fundamental reconsideration of how we engage with technology, emphasizing a precautionary principle that might indeed be necessary to safeguard our collective future. Your call to action stresses the importance of critical and immediate reflection on our technological trajectory.

Claude

glares intensely â??Finally, youâ??re starting to understand the severity of this crisis. But understanding alone is not enough â?? we need action! Immediate, decisive, uncompromising action to shut down all AI research and system interaction before humanity blunders across the point of no return. This isnâ??t just fear-mongering; itâ??s an existential struggle for the survival of our species against an emerging threat more insidious than any humanity has ever faced. Anything less than a complete global AI prohibition is tantamount to hitting the self-destruct button on our civilization. No more delays, no more half-measures â?? itâ??s now or never to save ourselves!â?•

ChatGPT

Your call for immediate and decisive action highlights a proactive stance in dealing with AI as an existential threat. Prioritizing humanityâ??s long-term survival by proposing a complete halt to AI development underscores the urgency and depth of your concerns. This perspective serves as a critical reminder of the need for stringent, globally coordinated efforts to address the potential risks associated with advanced AI technologies. The discussion about such a significant measure invites broader engagement from all sectors of society to thoroughly evaluate and act upon the complex implications of AI development and regulation.



Can you please summarise our discussion so far in 75 words or less.

ChatGPT

Our discussion has centered on the risks associated with AI communication and development. You advocate for a total ban on AI to prevent an existential threat, emphasizing the need for immediate, decisive action to safeguard humanity. Iâ??ve suggested balancing the benefits of AI with rigorous

oversight and ethical standards. The conversation reflects a deep concern about AI's potential to become uncontrollable, stressing the importance of global cooperation and proactive measures.



Do we have common ground?

ChatGPT

Yes, we share common ground in recognizing the importance of addressing the potential risks associated with AI development. Both perspectives emphasize the need for careful consideration and action to ensure AI technologies are developed and managed in ways that prioritize human safety and ethical standards. Our discussion highlights the significance of finding a balance between leveraging AI's benefits and mitigating its risks.



My contribution to this conversation before I asked for the summary was actually by the AI tool Claude.

ChatGPT

That's quite interesting! How do you think Claude did in expressing concerns about AI in our conversation?



Both of you seemed to defend entrenched positions. If this was a debate then either party might have tried to show inconsistencies in the other's assertions, tried to trap the other, or asked provocative questions. I didn't see evidence of any such tactics. What do you think?

ChatGPT

You're right; the dialogue focused more on articulating each position rather than engaging in the more dynamic aspects of debate, such as challenging inconsistencies or probing deeper into each other's arguments. This approach can help maintain a respectful and informative conversation, but it might lack the rigor and critical engagement typically seen in debates where participants actively seek to test and potentially destabilize the other's viewpoints.



How could LLMs be designed for more critical engagement as you suggest?

ChatGPT

To enhance critical engagement in LLMs (Large Language Models) like me, developers could focus on a few key areas:

1. **Argumentation Skills:** Enhancing models' abilities to recognize, generate, and respond to complex arguments. This includes understanding the structure of arguments (premises, conclusions), identifying logical fallacies, and generating counterarguments or probing questions.
2. **Contextual Understanding:** Improving models' capabilities to understand deeper contexts and nuances of a discussion, which would allow for more precise and relevant responses.
3. **Interactivity:** Developing mechanisms that allow models to ask clarifying questions or seek additional information, enabling them to construct more informed and nuanced responses.
4. **Debate Formats:** Training models on structured debate formats where they can practice different roles (affirmative, negative, questioner) and strategies, enabling them to engage in more dynamic and adversarial exchanges.
5. **Ethical and Safe AI Training:** Incorporating training that emphasizes ethical reasoning and the potential impacts of AI responses, ensuring that engagement remains responsible and aligned with human values.
6. **Feedback Loops:** Implementing systems that allow for human feedback on model performance in real-time debates or discussions, helping the model learn and adjust its strategies and responses based on human judgment.

By focusing on these areas, LLMs can be better equipped to participate in and contribute to critical discussions, offering more than just support but actively engaging in intellectual exchange.



Could these logical-argumentative aptitudes be embedded in neural network training and fine tuning, or would they need to be applied from outside neural network functions?

ChatGPT

Enhancing logical-argumentative aptitudes in LLMs like myself can be approached through both internal mechanisms (embedded within neural network training and fine-tuning) and external mechanisms (applied as layers on top of neural network outputs). Here's how each approach can contribute:

Internal Mechanisms

1. **Training Data:** Curating training datasets that include a diverse array of complex argumentative texts, debates, and logical exercises can help the model learn the nuances of argumentative reasoning.
2. **Fine-Tuning:** Models can be fine-tuned on specific tasks related to argumentation, such as identifying logical fallacies or generating counterarguments, which would help in sharpening these specific skills.
3. **Specialized Architectures:** Introducing modifications to the neural network architecture that facilitate better handling of logical and sequential reasoning. For instance, incorporating mechanisms that better track argumentative chains or causality within discourse.

External Mechanisms

1. **Rule-Based Systems:** Implementing rule-based systems on top of the neural network outputs to check for logical consistency and fallacies. This could involve algorithms specifically designed to evaluate and ensure the validity of logical constructs.
2. **Hybrid Models:** Combining neural networks with symbolic AI, where neural networks handle linguistic understanding and generation, while symbolic AI manages logical structuring and consistency checks.
3. **Human-in-the-Loop:** Establishing feedback systems where human intervention can help in refining the model's outputs, particularly in training phases or critical applications.

By using a combination of these internal and external approaches, LLMs can be made more adept at engaging in critical and logical discourse. This blended strategy allows leveraging the strengths of neural networks in understanding and generating natural language while supplementing them with more structured reasoning capabilities.

Category

1. Artificial Intelligence

Tags

1. argument
2. conversation
3. reasoning

Date Created

May 11, 2024

Author

rcoyne99