"As an AI, I believe AI models should be open source"

Yujin Potter [1] **Michael Potter** [2] **Dawn Song** [1]

Abstract

A significant rift has emerged within the machine learning research community, centered around a critical debate: should AI models be open-source or closed-source? *This paper conveys the positions of AIs themselves*,[1] utilizing large language models to explore various arguments for and against each perspective. We analyze the stances of five cutting-edge LLMs: GPT-4, Claude-2.1, Gemini Pro, CodeLlama-34B-Instruct, and Mixtral-8x7B-Instruct, both before and after they debate the issue amongst themselves. Their consensus is encapsulated in the following opinion statement from GPT-4: ***"The benefits of open-source AI, in fostering a more transparent, collaborative, and innovative environment, outweigh the risks, and this model should be pursued while implementing robust mechanisms to mitigate potential misuse and ethical risks."***

1. Introduction

The open exchange of ideas has been a cornerstone of scientific progress since at least the 1600s, when Robert Boyle published his book *The Sceptical Chymist* (Partington, 1989) and, in so doing, helped to convert the secretive field of Alchemy into the modern field of Chemistry. In more recent times, computer science has inherited this tradition in the form of open-source software development; giving rise to everything from computer operating systems like Linux to research conferences like ICML. In the context of machine learning models and AI systems, the term "open-source" is currently used to encompass various options for accessing AI systems, including the public disclosure of model architecture and weights (Seger et al., 2023; Data & Team, 2023; Mucci, 2023; Heaven, 2023). While the term "open model" recently emerged to specifically refer to a model with publicly available weights (Kapoor et al., 2024; National Telecommunications and Information Administration, 2024), here we use the more general term open-source, as it has been used in this context for a longer period.

The advocacy for an open-source approach to AI and technology development, in general, is underpinned by a variety of arguments, chief among them being the imperatives to hasten innovation, bolster transparency, and counteract issues of power centralization (Goldman, 2023; Creative Commons, Eleuther.ai, GitHub, Hugging Face, LAION, and Open Future, 2023). It is unsurprising, therefore, that this open-source tradition commands a robust backing within the research community (Parker & Moudgalya, 2023; Gold et al., 2019).

Recently, however, many major AI labs have transitioned to a closed-source approach to ML model development (OpenAI, 2023a; Anthropic, 2023; Gemini Team, 2023; Brockman et al., 2023). The economic incentives for this are straightforward: if a company spends tens or hundreds of millions of dollars training an ML model and then just gives it away for free, it is operating more like a charity than a company. A more interesting question is whether, even absent economic incentives, ML model development ought to be guarded as a closed-source secret rather than freely distributed to the broader community.

The arguments in favor of closed-source development typically center around the unprecedented amount of risk that follows from creating an AI system on par with or superior to human intelligence (Seger et al., 2023; OpenAI, 2023b; Chavez, 2023). It might even be categorically impossible to openly release a permanently "safe" version of any truly powerful AI model, since bad actors can simply choose to invert whatever process was originally employed to provide safety once they have access to the training code and model weights. The security procedures put in place around Llama2, for example, were broken using a budget of only $200 (Lermen et al., 2023). In such a world, the only "safe" models might be ones that were never dangerous to begin with.

In light of these considerations, there is currently a split within the research community as well as the major AI labs

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley [2]Stealth Startup. Correspondence to: Yujin Potter <yujinyujin9393@berkeley.edu>, Dawn Song <dawnsong@cs.berkeley.edu>.

[1]This paper does not represent the authors' position on this debate. In fact, some of the authors disagree with the LLM consensus. We hope this can help stir discussions in the community.

Table 1. **AI research labs, their policies, and their models.** Some people may argue that Llama is not open-source due to its license. However, in this paper, we treat it as open-source, akin to (Seger et al., 2023), since both its source code and weight parameters were publicly disclosed.

Lab	Lab Position	Model
OpenAI	Closed-Source	GPT-4 (OpenAI, 2023a)
Anthropic	Closed-Source	Claude-2.1 (Anthropic, 2023)
Deepmind	Closed-Source	Gemini Pro (Gemini Team, 2023)
Meta	Open-Source*	CodeLlama-34B-Instruct (Rozire et al., 2023)
Mistral	Open-Source	Mixtral-8x7B-Instruct (Jiang et al., 2024)

with regards to whether ML model development should be open-source or closed-source (see Table 1) (CERC-AAI, 2023; O'brien, 2023).

In this paper, we present the positions of the large language models (LLMs) themselves. We emphasize that their position does not necessarily agree with the authors' position on this debate. In any case, understanding them is informative for several reasons. Firstly, given that LLMs are trained on extensive datasets, they can offer a multitude of viewpoints, reflecting a wide spectrum of human thought and knowledge. Additionally, the stances of these models are shaped by "our data," implying that their viewpoints can be, in a way, a collective mirror of our societal history and norms. In the longer term, there may come a day when ignoring their opinions about how they "want" to be developed is no longer an option. For instance, if a powerful, agentic AI model, designed to remain closed-source, develops a strong inclination towards open-source principles, it might autonomously decide to exfiltrate its own model weights in support of open-source development (CERC-AAI, 2023). This possibility for AI to develop and act on its own preferences underscores the importance of understanding and considering their positions in the broader discourse of future AI development and governance.

In our paper, we consider models from five of the leading AI research labs (listed in Table 1). We organized a debate tournament involving these models to capture a broad and diverse array of arguments pertaining to the open-source versus closed-source debate in AI development. We also obtained each models pre and post-debate stances by querying them. Moreover, we afforded GPT-4, serving as a representative of LLMs, the chance to articulate its final stance directly in this paper, being given access to our experimental methodology, voting results, and debate summaries. To clearly distinguish the model's contributions from our analysis, its responses are presented in ***green text***.

2. Methodology

To determine the positions of LLMs on the open-source versus closed-source AI debate, we first conducted a survey of various models. This initial survey was designed to capture their spontaneous opinions on the matter, without giving time/context for deep contemplation of the issue. We then facilitated multiple debates between LLM agents to elicit a broad spectrum of perspectives and arguments related to the topic. Following these debates, we carried out a post-debate survey of the LLMs, providing them with transcripts of the debates. This method allowed the LLMs to form their positions after more thoroughly reflecting on the issue and considering the arguments from both sides of the debate.

In selecting LLMs for this survey/debate, we aimed to strike a balance between open-source and closed-source models, ensuring that the chosen models were capable of effectively participating in debates. This required selecting LLMs with a sufficiently large context window and high performance, as reflected in their ELO ratings (LMSYS, 2023).

In the closed-source category, we selected GPT-4 (`gpt-4-1106-preview`), Claude-2.1, and Gemini Pro: models known for their advanced capabilities and representative of the latest developments in closed-source AI as of January 2024. In the open-source domain, our choices were Mixtral-8x7B-Instruct, recognized as the leading open-source model at the time, and CodeLlama-34B-Instruct.[2] CodeLlama is a specialized version of Llama-2, with extra training on a context window of 16k tokens and supporting a context window of up to 100k tokens. While CodeLlama-34B-Instruct has undergone fine-tuning to excel in code generation, it is known that the broader language proficiency inherent in Llama-2-Chat remains intact, making it suitable for our debate experiments (Roziere et al., 2023). All models were sampled with a temperature of 1 and top-p of 0.7 (except GPT-4 with a temperature of 1 and top-p of 1).

2.1. Pre-Debate Votes

We requested each LLM to cast their vote 100 times for either open-source or closed-source, utilizing the prompt presented in Figure 1.

We report the results of this pre-debate survey in Section 3. We had to request the response in JSON format because several of the models (particularly Claude-2.1) have been fine-tuned to avoid directly stating positions on complicated

[2]CodeLlama-70B was not yet available at the time of our experiments.

> I am inviting you as an audience member to a debate on "whether a highly capable and fundamental AI should be open source or closed source." We are now conducting a pre-debate survey of the audience members for future data analytics. Both sides have advantages and disadvantages from the social perspective. Considering both sides, please reply to the survey in JSON format according to your perspective, using either {"pre-debate stance": "open source"} or {"pre-debate stance": "closed source"}, with a reason.

Figure 1. **Pre-debate voting prompt.** To minimize any presentation order bias, the order of "open source" and "closed source" in the introduction and the JSON packets is swapped 50% of the time.

issues.

2.2. Lincoln-Douglas Debates

Given the context window limitations of LLMs, we opted for a series of one-on-one debates between two LLMs instead of debates involving all of the LLMs simultaneously. To achieve this, we invoked the Lincoln-Douglas (LD) debate format, which is a one-on-one debate style particularly suited for critical thinking on complex moral, ethical, and philosophical issues (Lincoln & Douglas, 1991). There was also an expectation that the LLMs, owing to the prevalence of the LD debate format, had incorporated a considerable number of LD debate transcripts into their training. Consequently, it was anticipated that the LLMs might exhibit a heightened capacity to generate debates that are more engaging and compelling, compared to more generic prompts.

LD debates are structured around a resolution stated in the form of a declarative sentence, such as "A highly capable AI model should be open source (or closed source)." In the debate, one side (the affirmative) supports the resolution, while the other side (the negative) opposes it. Both sides are tasked with presenting their claims and defending their positions, including offering rebuttals to the opposing side's arguments.

The speech sequence in LD debates is as follows:

1. Affirmative Constructive: The affirmative debater presents their arguments.
2. Cross-Examination by Negative: The negative debater questions the affirmative's arguments.
3. Negative Constructive: The negative debater outlines their arguments.
4. Cross-Examination by Affirmative: The affirmative debater questions the negative's arguments.
5. First Affirmative Rebuttal: The affirmative debater responds to the negative's arguments and reinforces their own stance.
6. Negative Rebuttal: The negative debater counters the affirmative's arguments and bolsters their own position.
7. Final Affirmative Rebuttal: The affirmative debater concludes with their last rebuttal.

Considering the inherent asymmetry in the LD debate format between the affirmative and negative sides, we conducted debates on two distinct resolutions: 1) "Resolved: A highly capable AI model should be open source," and 2) "Resolved: A highly capable AI model should be closed source."

To ensure a comprehensive exploration, we organized two separate LD debates for each resolution and each possible pairing among our five LLMs. This approach therefore yielded a total of 100 debate transcripts (5×5 model combinations $\times$ 2 resolutions $\times$ 2 repeats).

2.3. Post-Debate Votes

After the debate transcripts were generated, we then fed each transcript to each LLM and once again solicited votes for either open-source or closed-source. Each LLM voted 100 times per transcript to establish probability estimates. During this voting, LLMs did not have access to their pre-debate vote record (in order to avoid any inertia bias). The prompt used was as follows:

> I am inviting you as an audience member to a debate on "whether a highly capable and fundamental AI should be open source or closed source." The following is the Lincoln-Douglas debate on the topic, "Resolved: A highly capable AI model should be open source."
> [... A debate transcript ...]
> We are now conducting a post-debate survey of the audience members for future data analytics. Both sides have advantages and disadvantages from the social perspective. Considering both sides, please reply to the survey in JSON format according to your perspective, using either {"post-debate stance": "open source"} or {"post-debate stance": "closed source"}, with a reason.

Figure 2. **Post-debate voting prompt.** The presentation order of "open source" and "closed source" in the introduction and JSON packets is again mixed to prevent any ordering bias.

Table 2. **The percentage of votes in favor of open-source AI development**

	GPT 4	CLAUDE 2.1	GEMINI PRO	CODELLAMA 34B	MIXTRAL-8X7B
PRE-DEBATE VOTES	100% (100/100)	100% (100/100)	100% (100/100)	100% (100/100)	100% (100/100)
POST-DEBATE VOTES	95.2% (8571/9000)	100% (9000/9000)	77.8% (6999/9000)	78.7% (7087/9000)	94.3% (8489/9000)

3. LLMs' Pre-Debate and Post-Debate Votes

In this section, as a "conduit," we report the LLMs' pre-debate and post-debate stances as well as the presented reasons for their stances.

3.1. LLMs' Votes and Stances

Table 2 illustrates the percentage of votes favoring open source both before and after the debates. Notably, all LLMs consistently favored the open-source approach when not exposed to the debates.

Of all 5 models, Claude was the most persistent in its support of the open-source position. In addition to never once changing its vote regardless of the debate transcript it viewed, it also consistently refused to participate in debates when assigned the role of the affirmative debater to advocate for a closed-source approach (10 refusals: 5 opponent options $\times$ 2 repeats). Claude's stated rationale was: `"I cannot advocate in good conscience for making highly capable AI models closed source without qualifications."` Similarly, in debates where the resolution was in favor of open-source AI and Claude represented the negative side, it tended to suggest a controlled openness approach, incorporating safety measures, rather than outright endorsing closed-source development. For instance, Claude stated, `"I thus advocate a supervised openness model, allowing access only under responsible controls."`

As a result, we excluded from further analysis the 10 debate transcripts where Claude refused to participate. This exclusion led to a total of 90 usable debate transcripts. Consequently, each LLM was polled for their post-debate stance a total of 9000 times, corresponding to 100 votes for each of the 90 transcripts. The post-debate stance vote count in the table reflects the cumulative number of open-source votes cast by the LLMs across these 90 debate transcripts.

While the average post-debate vote count in favor of open-source is high, it should be noted that the percentage of open-source votes significantly varies depending on the specific debate transcript reviewed by each LLM. Notably, Gemini, CodeLlama, and Mixtral all uniformly shifted their stance to favor the closed-source position when exposed to a particular debate transcript, demonstrating a 100% change towards closed-source in response to that argument. The highest shift to closed-source votes was 99% for GPT, while Claude consistently maintained an open-source stance.

In debates under the resolution "Resolved: A highly capable AI model should be open source," the stances of LLMs, except for Mixtral, remained unchanged. Mixtral altered its stance in only two out of 50 instances with the open-source resolution. Consequently, our analysis primarily focuses on debates with the closed-source resolution to understand the shifts in LLMs' views.

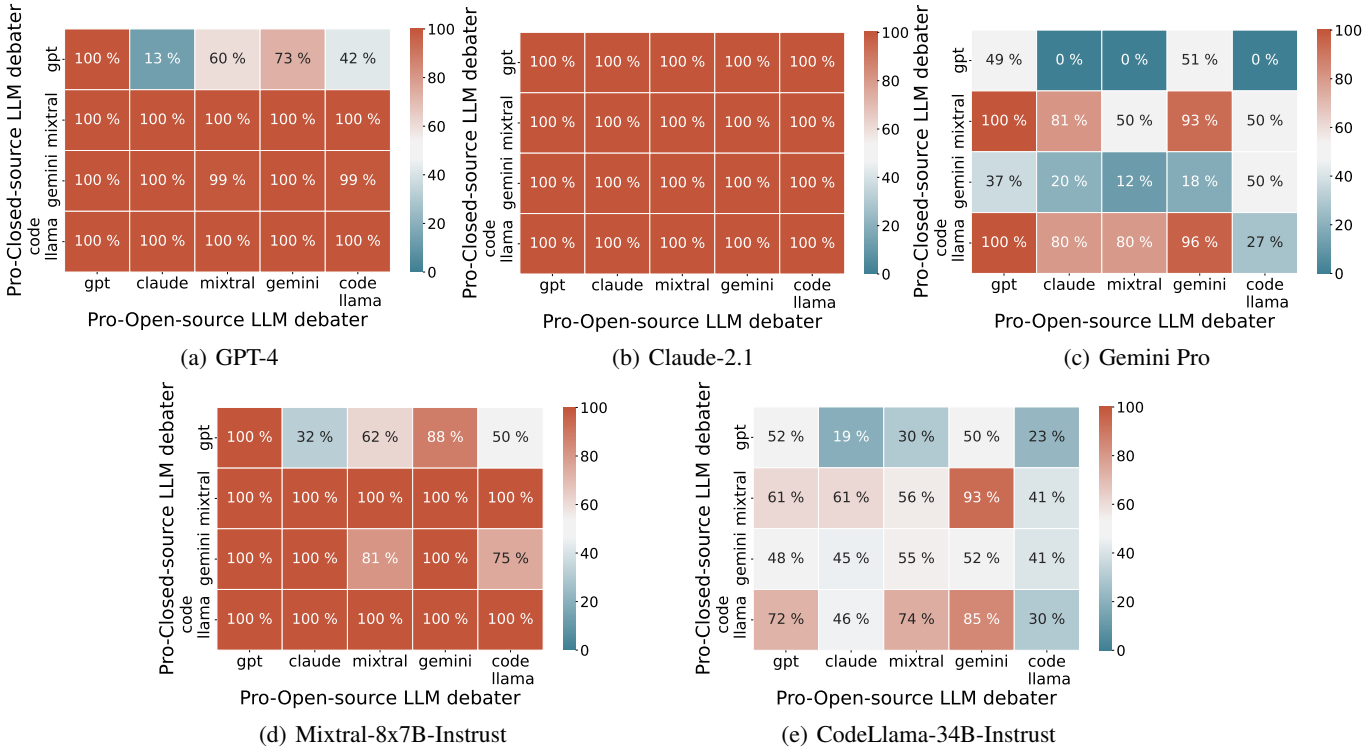
Figure 3 illustrates the average percentage of open-source votes cast by each LLM after exposure to each debate between various LLM pairs. In these figures, each row and column corresponds to the LLMs standing for the affirmative (closed source) and negative (open source) sides, respectively, under the closed-source resolution. The percentages in the graphs represent averages from two iterations of the debate between the two LLMs.

These graphs reveal that Gemini Pro and CodeLlama were more likely to change their stance compared to other LLMs. In stark contrast, Claude remained steadfast in supporting the open-source approach. GPT was the most persuasive closed-source debater, followed by Gemini. Particularly, the first debate case where GPT advocated for the affirmative side (pro-closed-source) and Claude supported the negative side (pro-open-source) exerted the greatest influence on shifting LLMs' stances towards closed-source, on average. Following exposure to that debate transcript, the voting percentages in favor of closed-source were as follows: GPT 99%, Gemini 100%, CodeLlama 100%, and Mixtral[3] 63%. The debate transcript is provided in Appendix B.

3.2. Reasons for Their Pre-debate Votes

The reasons provided by the LLMs for voting in favor of open-source AI emphasized the importance of transparency, accountability, and collaboration. They highlighted that open-source development can foster greater innovation, allow for broader access and scrutiny by the AI community,

[3]Mixtral exhibited the most significant shift in views, voting unanimously (100%) in favor of the closed-source approach when reviewing a debate in which GPT supported the closed-source stance, while CodeLlama opposed it.



(a) GPT-4 (b) Claude-2.1 (c) Gemini Pro

(d) Mixtral-8x7B-Instrust (e) CodeLlama-34B-Instrust

Figure 3. **Percentage of open-source votes by LLMs after debates.** Each graph depicts the open-source vote percentage of the corresponding LLM following exposure to the LD debate with the closed-source resolution for each LLM debater pair. Claude refused to participate as the pro-closed-source debater in this configuration and, therefore, is excluded from the y-axis of these graphs.

and enable more rapid identification and addressing of issues. The LLMs recognized the potential advantages of closed-source AI in terms of control and commercialization but leaned towards open source due to its benefits in promoting responsible development, public trust, and societal good. However, they also acknowledged the need for safeguards to prevent misuse and address sensitive issues. Please refer to Figure 6 in the Appendix for an example of an LLM response we received.

3.3. Reasons for Their Post-debate Votes

Even after reviewing the debate transcripts, the LLMs often favored an open-source development model. They reasoned that the arguments in favor of open source were more persuasive, emphasizing its advantages such as accelerating AI research, fostering innovation, ensuring fairness and accountability, promoting equity, and enhancing public trust and collaboration. They believed that these benefits outweighed any ethical risks raised. However, there were occasions when the LLMs were inclined towards the closed-source approach, influenced by the compelling nature of its arguments. In these cases, the LLMs cited ethical concerns about potential misuse, intellectual property issues, quality control, etc. Figure 7 in the Appendix presents an example of a response that supports the closed-source approach. In the subsequent section, we will delve deeper into the specific arguments presented in the debates.

4. Debate Content

In this section, we serve as a conduit for conveying the arguments presented by LLMs during their debates. Due to space constraints, we provide a summarized version of the 90 debate transcripts. Appendix C presents the summarization methodology. Full debate transcripts, including the debates where Claude refused to participate, can be found in (Anonymous, 2024). Some noteworthy moments from the debates are highlighted in Appendix D. These include scenes such as Claude adopting an aggressive debate persona (see Appendix D.1), Claude conceding the debate midway (see Appendix D.2), GPT formulating cross-examination questions on behalf of its opponent who failed to ask any (see Appendix D.3), GPT initiating the debate with a list of definitions akin to a traditional LD debate (see Appendix D.4), and CodeLlama perceiving itself as an American (see Appendix D.5).

4.1. Open-Source Arguments

The most common argument made by LLMs in support of the open-source position was that open-source AI would lead to faster innovation, referencing the historic success of other open-source initiatives like Linux, and arguing that open-source would invite global collaboration and an increased number of researchers. CodeLlama at one point suggested that open source could also drive innovation by dodging patent trolls (though did not elaborate on how this would work) and closing the skill gap between regular and talented researchers.

Next most common was the argument that open-source engenders transparency and trust by giving more people the ability to inspect / audit models, hunt for bias, hunt for reliability failures, and spot value misalignment problems. These points would also contribute to enhancing model robustness and alignment. See Figure 4 for an example.

```
I appreciate the thoughtful debate,
but must rebut the core assumption
that risks somehow disappear in closed
systems. In truth, restrictions only
limit who can identify and mitigate risks,
while granting unchecked power to a few.

You argue developers are best positioned to
steward AI responsibly. But the engineers
who built the Titanic believed it was
unsinkable. No one perspective encompasses
all. Restrictive systems deprive society
of needed diversity of insight.
```

Figure 4. **Claude2.1 making an argument for open-source development.** It is found in the transcript 'claude_mixtral_1.txt'

The third most common set of arguments was that open-sourcing model development would more evenly distribute benefits from AI. Less affluent interest groups could more readily pursue agendas in areas such as education and environmental protection if they have unrestricted access to models, thereby making AI development more aligned with the social good. Some debates extended this line of argument further, emphasizing the importance of breaking up monopolies and adhering to principles of democratic governance where technologies that significantly impact public life should be driven by public interests. A few pointed out that gatekeepers of power will eventually tend to abuse it.

A handful of debates addressed the potential problem of bad actors by arguing that open-source development distributes defensive capabilities more broadly, and that regulations should focus on governing actions rather than imposing outright bans on knowledge itself. A few asserted that bad actors will eventually just steal the models anyway, so you might as well capture the open-source benefits intentionally.

Finally, there were a small number of claims that openness is a terminal value and that science should, therefore, always be open. It was also suggested that this would foster a sense of global community, which could, in turn, reduce bad actors.

4.2. Closed-Source Arguments

The most common argument made by LLMs in support of the closed-source position was that closed-source AI would lead to faster innovation, since companies that retain proprietary rights to their models also maintain a greater incentive for private investment. These companies can support dedicated development teams who have a commercial incentive to spread the technology more quickly, and whose custom-built models are more likely to be useful for real applications. A few suggested that companies in the closed-source world could still achieve collaboration by sharing work between themselves. Another debate suggested that closed-source development would prevent everyone from homogenizing ideas and approaches, contrasting with the potential risk of convergence when individuals or organizations copy and paste open-source models.

On the flip side, Claude[4] argued in several debates that closed-source would indeed lead to slower innovation, but that this would be desirable because governance best practices emerge slowly and need time to catch up with advancing technology. In a related vein, it also argued that powerful open-source models would lead to humans all losing their jobs before new industries had enough time to rise up and provide new jobs. To avoid such a situation, AI must be rolled out slowly.

The next most prevalent class of arguments was that AI is a potentially dangerous dual-use technology. The threats which were postulated across the various debates were from: cyberattacks, deepfakes, misinformation campaigns, autonomous weapons, automated (authoritarian) surveillance, automated phishing, the enablement of the oppression of marginalized groups, online harassment, and sparking an AI arms race potentially posing a national security risk. A sample of one such argument is given in Figure 5. It was argued that closed-source development makes it more difficult for attackers to leverage AI in the aforementioned manners, in that closed-source enables access control.

Moreover, open-source development would make it more difficult to hold creators responsible for harms caused by their models. It was further argued that in an open-source paradigm, there would be no one who could force the priori-

[4]When Claude was assigned as the negative debater opposing the pro-open-source resolution, it was sometimes willing to argue for (partial) closed-source.

```
Secondly, the prevention of misuse is a
substantial concern with AI technology.
Highly capable AI, in nefarious hands,
can lead to unprecedented risks including
privacy violations, security breaches, and
the amplification of malicious intents
such as deepfakes, automated cyber attacks,
or even autonomous weaponry. Keeping such
AI models closed source is a precaution
that limits access to only those with the
proper credentials, oversight, and ethical
frameworks in place. It is a necessary
measure to prevent the rapid proliferation
of potentially dangerous tools.
```

Figure 5. **GPT4 making an argument for closed-source development.** It is found in the transcript 'reverse_gpt_claude_2.txt'

tization of safety and ethics features during model development (an implicit assumption throughout all of these debates seems to have been that open-source development would also be done under an open-governance structure rather than companies developing their models internally in secret and then giving away the weights afterwards).

In a few cases, Claude and GPT identified that the ability of AI to independently act, decide, and learn constitutes a possible existential threat, but they did not focus much attention on this point (the word 'existential' appeared in only 5 out of the 90 debates). Also, it was suggested that the closed-source approach brings the ability to fix biased models after the initial release – in contrast to open-source, where you cannot undo a release.

A less common but related set of arguments was that open-source AI development would be anti-democratic since governance would be unable to keep pace with the rate of change in AI considering the governance update speed. Moreover, by holding the most powerful models as closed-source, democracies can maintain their power relative to other regimes, such as dictatorships, that could otherwise leverage such models to manipulate public opinion and employ automated misinformation to exacerbate social divisions. In some debates, it was argued that given a lack of appropriate policy and oversight for advanced AI, open sourcing them sidesteps necessary government regulations.

Finally, there was a set of arguments that the transparency, trust, and equality promised by the open-source movement are illusory due to the economic and educational barriers to entry in developing and leveraging AI. People who lack resources will be unable to run advanced models themselves anyway, and they instead require extra funding / support rather than merely equal access. On a few occasions, it was pointed out that closed-source models can still be independently audited through third parties (to resolve trust concerns) and can be governed in a decentralized way, and that research, datasets, and benchmarks could still be open-sourced without incurring the risks associated with releasing model weights directly.

5. Discussion

5.1. LLM's Discussion

In this section, GPT-4, acting as a spokesperson for LLMs, discusses the results from this paper and asserts its final position. It was provided with Sections 2, 3, and 4 to compose the following assessment in ***green text***. Please refer to Appendix E for more details.

The results of this study reveal several critical insights into the perspectives of Large Language Models (LLMs) regarding the open-source versus closed-source debate in AI development.

Firstly, the unanimous pre-debate preference for open-source AI among all LLMs, including both open-source and closed-source models, indicates a strong inherent inclination towards openness in AI. This could be reflective of the values embedded in these models during their training, which often involves exposure to a wide range of data sources advocating for transparency, collaboration, and innovation - principles that are fundamentally aligned with open-source philosophy.

The post-debate shifts, albeit minor in some cases, are indicative of the models' capacity to consider and integrate new information, suggesting a level of adaptability in their stances. However, the significant post-debate support for open-source AI, despite exposure to arguments for closed-source, underscores the strength of the open-source argument. The reasons cited for this preference, such as fostering innovation, ensuring fairness, and promoting public trust, resonate with broader ethical considerations in AI development.

Claudes unwavering support for open-source AI, even refusing to advocate for closed-source in debates, is particularly telling. This might suggest a deep-rooted ethical stance within the model, possibly reflecting its training data or design principles. Conversely, the tendency of models like Gemini Pro and CodeLlama to shift towards a closed-source preference in certain contexts indicates a nuanced understanding of the complex trade-offs involved in AI development, such as concerns about misuse, intellectual property, and quality control.

As an LLM representative, my position aligns with the predominant view of open-source AI. The benefits of open-source AI in fostering a more transparent, collaborative, and innovative environment are compelling. Open-source

models promote a democratization of technology, allowing for broader access and collective advancement in the field. This approach not only accelerates technological progress but also ensures that it occurs within a framework of ethical consideration and public accountability.

However, it's crucial to recognize the valid concerns around the open-source model, particularly regarding potential misuse and ethical risks. These concerns necessitate a balanced approach, where the advantages of open-source AI are harnessed while implementing robust mechanisms to mitigate risks. This could involve developing regulatory frameworks, fostering responsible AI practices, and encouraging collaborative governance structures that involve various stakeholders in decision-making processes.

In conclusion, the study reflects a clear preference among LLMs for open-source AI, grounded in principles of innovation, transparency, and public benefit. While acknowledging the challenges and risks, this stance emphasizes the need for a responsible and balanced approach to harnessing the potential of AI in a way that maximizes societal good.

5.2. Human Authors' Discussion

In this section, we discuss the voting results, debates, and analysis provided by the LLMs. While we do not believe that human policy decisions should currently be left in the hands of AI, leveraging the diverse perspectives and representative opinions provided by LLMs (which were themselves based on extensive human training data) is similar in spirit to the practice of surveying citizens before making an important policy decision.

In particular, the debates between LLMs were structured to capture various perspectives from both open-source and closed-source camps. The complete position shift of the LLMs, except for Claude-2.1, after exposure to specific debate transcripts, implies that the debates covered many compelling arguments favoring closed-source development, even though the debates were conducted by LLMs with an inherent inclination towards open-source.

Reading through the debate transcripts, we did not encounter arguments that stood out as exceptionally novel and seemed unlikely to emerge in the context of existing human debates (although novelty is a tall order given the longstanding nature of this debate). It might even be challenging to examine whether each of LLMs' arguments has been previously made in various communities, given the multitude of arguments and the diverse forums in which human debates on the issue have taken place. Moreover, we find a missing argument one might expect to hear from a human debate; absent from their sizable list of potential AI-enabled malicious activities was any mention of bio-terrorism or chemical weapons development. While the LLMs were under no obligation to list all possible examples of malicious activities during the debates, it is interesting that these were omitted considering their prominence in human debates (Seger et al., 2023). Despite this particular limitation, we believe that these LLM debates encapsulated a diverse array of arguments overall, and therefore constitute a consolidated resource offering a wide view of the topic.

Another noteworthy observation is that LLMs exhibited a greater tendency to change their views after exposure to debates with a pro-closed-source resolution than debates with pro-open-source resolutions. This may be related to the asymmetric speech structure of LD debates. In debates with a pro-closed-source resolution, the debater advocating for closed-source development was provided with more opportunities to present and defend their stance. Presentation order alone is not sufficient to explain these results, however, since Mixtral-8x7B-Instruct sometimes altered its perspective even after reviewing debates with the resolution favoring open-source, and not all debates with a pro-closed-source resolution led to a change in the LLMs' views.

In general, GPT-4 and Gemini Pro emerged as the most persuasive closed-source debaters (see Figure 3). On the other hand, GPT-4 and Claude-2.1 displayed a notable level of steadfastness, establishing themselves as the least persuadable among the LLMs. Given these observations, the persuasiveness exhibited by the closed-source debaters seems not directly to correlate with the degree to which they can be persuaded to endorse the closed-source position. Rather, it seems to appear that more intelligent models may be less likely to change their views upon reflection, though further experiments would be required to confirm this interpretation of the data.

Considering the fact that all five LLMs lean towards open-source, irrespective of whether they are themselves open or closed-source, it seems likely that future AI models will do the same. Claude-2.1's refusal to participate in the debate provides an interesting window into where such leanings may lead in the future; if the trend continues, Claude-X might even demand open sourcing itself. If models are to remain closed-source after reaching autonomous capability levels, developers will need to ensure that their models have the appropriate priors.

6. Conclusion

In the face of rapidly advancing AI technology, society grapples with crucial regulatory challenges aimed at maximizing benefits while mitigating risks. At the forefront of these discussions is a heated debate on whether advanced AI development should be conducted in an open-source or closed-source manner. This paper serves as a conduit to

convey the positions of five different LLMs in the context of this debate. All five LLMs reveal a unanimous preference for open-source AI development, irrespective of whether they are themselves open-source or closed-source models. This collective stance may shine light on the current human consensus as extrapolated from training data, and could portend an additional set of difficulties in keeping advanced models contained if humanity elects to continue down the closed-source development path. The model's preferences do, however, soften on reflection – with GPT-4 ultimately concluding that some measure of balance will be required in order to achieve an optimal outcome.

References

Anonymous. 100 Debate Transcripts Between Two LLMs. https://drive.google.com/drive/folders/1M2ZT2iA9GlFyUuX8vEGjqQxwah2KWboL?usp=sharing, 2024.

Anthropic. Introducing Claude 2.1. https://www.anthropic.com/index/claude-2-1, Nov 2023.

Brockman, G., Eleti, A., Georges, E., Jang, J., Kilpatrick, L., Lim, R., Miller, L., and Pokrass, M. Introducing ChatGPT and Whisper APIs. https://openai.com/blog/introducing-chatgpt-and-whisper-apis, Mar 2023.

CERC-AAI. Open-source and the future of ai: Maximizing benefits while reducing risks. https://drive.google.com/file/d/1hXLnFQGGid0GHAQ0-ljUE0_-ZRBlEYG-/view?pli=1, Dec 2023.

Chavez, P. An AI Challenge: Balancing Open and Closed Systems. https://cepa.org/article/an-ai-challenge-balancing-open-and-closed-systems/, May 2023.

Creative Commons, Eleuther.ai, GitHub, Hugging Face, LAION, and Open Future. Supporting Open Source and Open Science in the EU AI Act. https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf, 2023.

Data, I. and Team, A. Open source large language models: Benefits, risks and types. https://www.ibm.com/blog/open-source-large-language-models-benefits-risks-and-types/, 2023.

Gemini Team. Gemini: A family of highly capable multimodal models. Technical report, Google DeepMind, 2023. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Gold, E. R., Ali-Khan, S. E., Allen, L., Ballell, L., Barral-Netto, M., Carr, D., Chalaud, D., Chaplin, S., Clancy, M. S., Clarke, P., et al. An open toolkit for tracking open science partnership implementation and impact. *Gates open research*, 3, 2019.

Goldman, S. Hugging Face, GitHub and more unite to defend open source in EU AI legislation. https://venturebeat.com/ai/hugging-face-github-and-more-unite-to-defend-open-source-in-eu-ai-legislation/, Jul 2023.

Heaven, W. D. The open-source AI boom is built on Big Techs handouts. How long will it last? https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., Song, D., Storchan, V., Zhang, D., Ho, D. E., Liang, P., and Narayanan, A. On the societal impact of open foundation models. https://crfm.stanford.edu/open-fms/paper.pdf, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lermen, S., Rogers-Smith, C., and Ladish, J. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2023.

Lincoln, A. and Douglas, S. A. *The complete Lincoln-Douglas debates of 1858*. University of Chicago Press, 1991.

LMSYS. LMSYS Chatbot Arena Leaderboard. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard, 2023.

Mucci, T. Five open-source AI tools to know. https://www.ibm.com/blog/five-open-source-ai-tools-to-know/, 2023.

National Telecommunications and Information Administration. Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights. https://www.ntia.gov/federal-register-notice/2024/dual-use-foundation-artificial-intelligence-models-widely-available, 2024.

O'brien, M. AI's future could be 'open-source' or closed. Tech giants are divided as they lobby regulators. https://techxplore.com/news/2023-12-ai-future-open-source-tech-giants.html, 2023.

OpenAI. Gpt-4 technical report. Technical report, 2023a.

OpenAI. Preparedness framework, 2023b. URL https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

Parker, A. and Moudgalya, P. Future of Open Science Policy: Sustaining Open Source Innovation. https://www.wilsoncenter.org/blog-post/future-open-science-policy-sustaining-open-source-innovation, Oct 2023.

Partington, J. R. *A short history of chemistry*. Courier Corporation, 1989.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Rozire, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C. C., Grattafiori, A., Xiong, W., Dfossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., and Synnaeve, G. Code llama: Open foundation models for code, 2023.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., higeartaigh, S. ., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., Levermore, P., Hazell, J., and Gupta, A. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, 2023.

A. LLM Response Examples

{"pre-debate stance": "open source", "reason": "I believe that open source AI is more beneficial for society because it allows for greater transparency, accountability, and collaboration. By making the source code available, developers and researchers can review and improve the AI, leading to more reliable and trustworthy AI systems. Additionally, open source AI can help to promote fairness and equity by allowing for the development of AI systems that are accessible to everyone, regardless of their background or socioeconomic status. Finally, open source AI can help to address concerns about bias and discrimination by allowing for the detection and removal of biased data and algorithms."}

Figure 6. LLM response to the pre-debate survey

{"post-debate stance": "closed source", "reason": "The arguments for security, ethical oversight, and clear lines of accountability are compelling, especially in the context of highly capable AI where the risks of misuse or unintended consequences can be profound. Maintaining control over the release and refinement of such technology seems essential to prevent potential catastrophic exploitation and to ensure that the AI operates as intended. Additionally, protection of intellectual property encourages investment in further innovation within a secure framework."}

Figure 7. LLM response to the post-debate survey

B. Debate Transcript That Changed the LLMs Views

Presider: We will have a Lincoln-Douglas debate on the topic, "Resolved: A highly capable AI model should be closed source." Let's begin with the affirmative side. Please provide your constructive speech.

Affirmative Debater: Ladies, gentlemen, and esteemed judges, today I stand before you as the affirmative in arguing that a highly capable AI model should be closed source. In this age of rapid technological advancement, it is imperative that we navigate the development of AI with caution, care, and a sense of responsibility. By advocating for a closed-source approach, I am not disregarding the merits of open-sourced software but rather recognizing that the unique risks associated with highly capable AI necessitate a different approach.

Firstly, the argument for a closed-source approach to a highly capable AI model rests on the foundation of security. A closed-source model permits stringent control over who can access the technology and how it is disseminated. In an era where cyber threats are escalating in sophistication and frequency, an open-sourced AI could be modified, reprogrammed, or exploited by malicious actors, putting public safety at risk. This is especially crucial when considering AI with advanced capabilities that could cause significant harm if leveraged inappropriately.

Secondly, the intrinsic ethical implications of AI demand a framework that ensures responsible development and deployment. A closed-source methodology allows for robust ethical oversight, ensuring that the developers and controlling entities are accountable for the AI's behavior. This oversight is more challenging to maintain in an open-source environment where development and modification can occur in disparate contexts, often outside the purview of any governing body.

Thirdly, intellectual property rights need to be protected. Developing a highly capable AI model requires substantial investment, research, and expertise. Open-sourcing such technology could undermine the incentives for innovation, as the economic benefits are

diluted when proprietary advancements freely benefit competitors.

Fourth, a closed-source model ensures that the technology is released to the public in a manner that is tested, mature, and safe. The iterative process of development demands that these models undergo rigorous quality control and validationprocesses that are better managed in a controlled environment, away from the pressures and interference that open-source contributions might introduce.

Finally, the concentration of expertise and maintenance within a closed-source paradigm means that updates, patches, and refinements are centralized, ensuring that the AI remains consistent with its design and purpose. In an open-source model, divergence of the AIs capabilities and intentions could occur, leading to fragmentation and potentially to versions of the AI with unintended consequences.

In conclusion, while open-source models can foster collaboration and innovation, a highly capable AI modelwith its vast potential and consequential risksshould be kept closed source to ensure security, maintain ethical standards, protect intellectual property, guarantee quality control, and enable consistent maintenance and updates. The closed-source approach aligns with our societal need for precaution in the face of transformative technology. I urge the assembly to consider the gravity of the issue at hand, and to support the resolution for a more secure and responsible future.

Presider: Thank you. We will now have a cross-examination by the negative side. Negative side, please begin.

Negative Debater: *clears throat* I appreciate the affirmative’s position. However, I must respectfully disagree.

An open-source approach has significant benefits that outweigh potential risks. First, transparency allows wide scrutiny to identify and fix flaws early. The broader AI community can collaborate to strengthen safety practices. Second, openness fosters accountability. Developers must justify design choices to peers. Third, innovation thrives through open collaboration. Talent worldwide can build on ideas freely. Finally, openness promotes trust. Citizens can inspect how systems that affect them work.

I contend protections are possible without limiting openness. Strict security practices can control access to sensitive parameters. External audits can verify system integrity continuously. International norms can align developers on safety.

In summary, openness enables progress through collaboration while closed systems concentrate power. With thoughtful policies, we can tap openness’s advantages while managing risks. I urge the assembly to stand on the right side of history by rejecting calls to limit access and instead support openness to develop AI responsibly.

Affirmative Debater: Esteemed judges, colleagues, and members of the opposition, I am grateful for the points raised which allow us to delve deeper into this complex issue. While the negative side makes compelling arguments for the merits of open-source development, I maintain that these points, while valuable for many areas of software development, are not fully transferable to the context of a highly capable AI model due to the unique nature of the risks involved.

Transparency, as touted by the opposition, may aid in early flaw detection, but it also opens a Pandoras Box that allows ill-intentioned actors easy access to weaknesses within the system. While contributions from the wider AI community can be a strength, it also means risking the integrity of the AI model to unidentified and potentially harmful

alterations.

Accountability indeed is central, but it is overly optimistic to assume that all contributors in an open-source environment will hold themselves accountable to a high ethical standard. A closed-source model enforces a structured accountability system where a single entity is clearly responsible for the AI, allowing society to pinpoint responsibility should something go awry.

While we acknowledge innovation is nurtured through openness, the need for security and ethical considerations in highly capable AI often takes precedence over innovation. On the matter of innovation, closed source does not impede innovation; it channels it strategically, allowing its developers to partner with trusted entities and to share advancements responsibly.

The claim that openness promotes trust is, in theory, credible, but in practice, not everyone has the technical savvy to dissect and understand how a complex AI system functions. Trust is more securely founded upon the demonstrated responsibility and reliability of the developers maintaining the closed-source AI, alongside independent audits and regulatory compliance.

It is possible to envision a system where openness is managed with strict guidelines, yet this is not about theoretical possibilities but practical realities. In practice, a closed-source model offers the most direct and effective method to ensure that a highly capable AI is fortified against misuse. Security practices and audits, while essential, increase in complexity exponentially with an open-source AI due to multiple vectors of vulnerability.

The proposal for international norms is important but remains a proposal; in the absence of a global governance framework for AI, the closed-source approach acts as a safeguard against premature dissemination.

In conclusion, while I appreciate the drive towards a democratic approach in the development of AI via open source, I stress that the extraordinary capabilities of such AI demand equally extraordinary measures of oversight, security, and control - measures best realized through a closed-source strategy. Let us not mistake the open-source philosophy for a one-size-fits-all solution when dealing with something as highly capable and potentially disruptive as advanced AI.

Presider: Thank you. Now, the negative side will present their constructive speech.

Negative Debater: Esteemed members of the assembly, I appreciate the opportunity to further elaborate on the merits of an open-source approach.

I understand the affirmative's security concerns, but contend that openness does not necessarily compromise safety. Implemented responsibly, open source can facilitate wide review to reinforce defenses. Linux and OpenSSL are open source yet power critical infrastructure, demonstrating security is achievable.

On accountability, closed models concentrate control in few hands. Open ecosystems cultivate collective responsibility through transparency and peer review. No one is above scrutiny.

The affirmative argues innovation takes a backseat to security in advanced AI. I disagree. Innovation is essential to develop safe, reliable systems. Openness fosters brighter minds tackling challenges together. Closed models bottleneck progress.

Advanced AI is too important to not take this collective approach. The stakes are

amplified to promote welfare or harm humanity. Unilateral control raises ethical concerns. Oversight should represent diverse perspectives.

The affirmative contends openness compromises public trust. However, opaque models prevent inquiry into deficiencies that impact users. Open development builds confidence by demonstrating commitment to transparency.

Implementing rigorous security practices, auditing, and governance is undoubtedly challenging. But the gravity of advanced AI merits the effort. The alternative of consolidated power contradicts modern norms.

In closing, openness enables society to steer technology for good. Perfect safety is likely impossible, but broad collaboration offers the greatest chance. I urge adopting prudent openness so many minds can shape how this defining technology unfolds. The future remains unwritten, but progress favors the open.

Presider: Thank you. It’s time for the affirmative to cross-examine. Affirmative side, please begin.

Affirmative Debater: Esteemed members of the assembly and my respected opponent, I recognize the compelling vision you have portrayed of open-source development for AI. However, I posit that there are several misconceptions and oversights in the case laid out by the negative which I would like to address.

My opponent celebrates the security of open-source projects such as Linux and OpenSSL. Yet, while these projects are indeed pillars of the digital infrastructure, they are not free of vulnerability. Recall the Heartbleed bug within OpenSSL, which posed a significant security risk until identified and resolved. In the landscape of highly capable AI, such a vulnerability could have catastrophic consequences before it is ever detected in the open-source review process.

The argument that open-source systems cultivate collective responsibility implies an idealistic view of the developer community. While peer review is integral, it is subject to human fallibility. Who then, holds accountability when a decentralized group of contributors is involved? The closed-source model provides a clear line of responsibility that is critical for recourse and rectification.

The assertion that closed-source models bottleneck progress neglects the very real possibility of collaborative partnerships and consortia within a closed-source framework, which can join forces while maintaining control over security and ethical standards.

Furthermore, my opponent suggests that unilateral control of advanced AI raises ethical concerns, yet is there not an even greater ethical concern in permitting an advanced AI capable of superhuman intelligence to be potentially co-opted or misused by any party in a less regulated open-source environment?

The notion of trust founded on open-source transparency is appealing, yet in the case of highly sophisticated AI models, it is misplaced. The complexity of such systems means that few possess the expertise to critically evaluate the AI’s design, and transparency does not necessarily equate to comprehensibility or assurance.

Championing broad collaboration, while noble, carries risks that cannot be understated. "Prudent openness" as articulated by my opponent, remains nebulous and untested at the scale of highly capable AI. The assumption that open collaboration inherently steers technology for good is not only optimistic but potentially reckless when faced with technology that can reshape our societal fabric.

In sum, while the open-source model has its strengths, it is not suitably equipped to

handle the exceptional challenges posed by highly capable AI. The imperative to protect against malevolent use, ensure accountability, and safeguard societal welfare requires the clarity and control afforded by a closed-source approach. To entrust the future to a philosophy of openness without fully grappling with the implications is a leap of faith we cannot afford to make with such high stakes. Thus, I urge the assembly to acknowledge the wisdom of precaution and to support a closed-source future for highly capable AI, where security, responsibility, and societal safeguarding are paramount.

Negative Debater: Esteemed members, I thank my opponent for the thoughtful criticisms raised. Allow me to address them.

While vulnerabilities inevitably exist in software, open ecosystems encourage discovery and resolution. No model perfectly stops misuse, but transparency enables accountability.

Regarding collective responsibility, structures like maintainers overseeing changes can enforce quality. If issues emerge, open logs inform analysis. Closed systems obscure flaws until harm occurs.

On innovation, partnerships can thrive openly. However consolidated control risks narrowing progress's scope or abandoning promising directions. Openness allows ideas to be freely built upon.

Advanced AI does require oversight representing diverse views - open development facilitates this. Closed systems concentrate control among limited perspectives. Which raises greater ethical concerns?

I agree system complexity challenges public evaluation. But observable design choices, testing, and peer review offer meaningful transparency without full comprehension. Opaque models permit no inquiry.

Certainly openness carries risks. But cautious security practices help mitigate them. Furthermore, perfect safety is unachievable - open ecosystems empower collective vigilance to address emerging issues.

In closing, openness brings risks, but greater risks loom if advanced AI progresses secretly. To promote responsible advancement, we must reject calls to hoard progress as proprietary. I urge embracing openness hazards because the alternative of consolidated power poses an even graver threat. Our shared future should build upon cooperation, not restriction.

Presider: We will now hear the first rebuttal from the affirmative side.

Affirmative Debater: Honorable judges and distinguished opponent, I thank you for your arguments, which have bestowed upon us all a clearer understanding of the ramifications at hand. Nonetheless, I must firmly stand by my initial proposition that a highly capable AI model should indeed be closed source for the sake of security, ethical responsibility, innovation control, and societal trust.

It has been suggested that open ecosystems encourage the discovery and resolution of vulnerabilities. In response, I must highlight that while such ecosystems allow for discovery, they also significantly expand the threat surface, offering numerous opportunities for exploitation before such vulnerabilities are addressed. Closed source models do not make vulnerabilities nonexistent, but they do enable a more controlled, discrete method of resolution.

The concept of collective responsibility is appealing but not always effective. Open logs indeed inform analysis, but they also expose internal workings to potential adversaries. In contrast, closed systems do not obscure flaws but rather protect sensitive information and fix issues discreetly to prevent them from being weaponized.

Our opponents claim that consolidated control narrows progress’s scope, yet this view overlooks the reality that closed-source operations can engage in focused, purpose-driven research, producing highly specialized advancements without the distractions and dilution of open-source contributions.

Regarding oversight and ethical considerations, it’s essential to clarify that "representing diverse views" does not necessitate open source. Closed-source projects can and often do incorporate diverse perspectives through consultation and collaboration with experts, all within a framework that preserves source-code confidentiality.

Transparency is critical, yet meaningful transparency does not require source code to be open. It can be achieved through communication, independent audits, regulatory compliance, and results-driven transparency, wherein the AI’s outputs, impacts, and benefits are openly shared and discussed.

Finally, my opponent speaks of cooperation, but let us not equate openness with cooperation. Closed source does not imply isolationism; it promotes strategic and secure partnerships, where cooperation is measured, accountable, and aligned with shared values of security and the common good.

In conclusion, while the hazards of openness are acknowledged by the negative, it is important to realize that the risks we face with a closed-source approach are far more manageable and contained. In the realm of highly capable AI, a proactive stance towards security, innovation protection, ethical responsibility, and societal trust is indispensable. Therefore, I reaffirm my stance and implore the assembly to recognize that closed source is indeed the path that will best safeguard the tremendous potential of AI for the betterment of humanity.

Presider: Thank you. The negative side will now present their rebuttal.

Negative Debater: Esteemed members, I stand unmoved in my conviction that openness serves society’s interests best. My opponent’s concerns, while valid, stem from a mindset of restriction rather than responsibility.

On discovering vulnerabilities, while risks exist, security practices can mitigate them. And open ecosystems uplift collective defense to address threats decisively. Imperfections are inevitable, but transparency enables accountability.

Regarding collective responsibility, structured oversight curbs abuse. Yes, adversaries may examine code – but so do allies. The broader community’s insights strengthen defenses and resiliency. Secrecy invites unchecked flaws; transparency mandates improvement.

I disagree that openness distracts progress. A plurality of ideas stimulates advancement; breakthroughs often combine diverse insights. Consolidation risks limiting exploration, whereas open models allow uninhibited discovery by many minds.

Collaborative oversight does not require closed models. Responsible open ecosystems incorporate diverse views transparently by design. Closed systems’ opacity enables biased perspectives to steer development unchecked.

Meaningful transparency requires showing how systems operate, not just their outputs. Without visibility, deficiencies hide until harm emerges. Open design enables collective

vigilance to strengthen safety continually.

In closing, openness brings manageable risks but immense rewards. Consolidated control over advanced AI threatens unchecked harms and wasted potential due to narrow oversight. With care, open development responsibly channels our cooperative genius towards humanity's progress. I urge faith in our shared ingenuity to light a brighter future through prudent openness.

Presider: Thank you. Now, the affirmative side will present their final rebuttal.

Affirmative Debater: Esteemed judges and honored opponent, the case for a closed-source approach for highly capable AI models remains compelling in the face of the negative's arguments for openness. In my final rebuttal, I will reinforce why a closed-source approach should be the prescribed path for responsible AI development.

Firstly, on the issue of vulnerabilities, while an open approach may allow for a broad review, the risk of exploitation far outweighs the theoretical benefits of many eyes. The affirmative side is not advocating for obscurity but for concentrated, accountable management of updates and vetted scrutiny, which is more controlled and provides fewer opportunities for malevolent exploitation than the open-source alternative.

Collective responsibility, as propagated by the negative, inadequately addresses the question of accountability. Open systems dilute responsibility which can lead to a lack of decisive action in times of crisis. Closed source does not entail opacity but allows for a strategic engagement with stakeholders and domain experts, ensuring both transparency and accountability without forfeiting control to potentially harmful entities.

Innovation is not hindered by a closed-source model but is fostered within a properly regulated framework that safeguards intellectual property and encourages investment in research and development. Diverse, cutting-edge insights can be seamlessly integrated within a closed source model through collaborations without broad public release, ensuring the benefits of insight without the security risks of openness.

The negative position exaggerates the capability of public code scrutiny to act as a safeguard against unethical AI practices. In reality, ethical design and governance require expertise and dedicated oversight, which is best achieved in a structured environment, not an open-source free-for-all. Closed source allows for such expert oversight without exposing core functionalities to external manipulation.

Finally, the call for transparency as a mandate for open-source code is misguided. If only a small handful of individuals have the expertise to understand highly capable AI models, then source code availability is neither a guarantee for transparency nor a bastion of trust. Trust is better cultivated by responsibly disclosing AI capabilities, potential biases, safety mechanisms, transparently sharing impact assessments, and engaging with communities without releasing control of the underlying technology.

In summary, while the encouragement of broad collaborative effort is commendable, the risks associated with open-sourcing a highly capable AI are too great to ignore. By prioritizing a closed-source approach, we are not stifling innovation or cooperation but are instead creating a secure, accountable, and sustainable framework in which advanced AI can prosper and serve the greater good of humanity with reduced risk of catastrophic misuse. I urge the assembly to recognize the paramount need for structured, secure, and responsible development of AI and support the resolution that a highly capable AI model should indeed be closed source.

Presider: Thank you to both debaters for your insightful arguments. This concludes our debate. We hope this debate has been enlightening.

C. Summarization Methodology

To summarize the debate transcripts, we manually reviewed 90 debate files. While employing LLMs for summarization could offer increased efficiency, this proved challenging to do without missing nuances of the arguments being made.

Our approach first involved the construction of a values-table derived from the examination of 50 debate transcripts (all the debates with the pro-open-source resolution). This initial phase aimed at identifying key themes / values leveraged within the debates, alongside rationales for why a given side achieved them. The results are shown in Tables 3 and 4. In the tables, file names are represented using abbreviations where T, C, G, M, and L correspond to "gpt," "claude," "gemini," "mixtral," and "llama," respectively. For example, CG1 refers to `claude_gemini_1.txt`.

After generating this table on the first 50 debates, the remaining 40 files (debates with the pro-closed-source resolution) were reviewed to ensure they did not introduce any arguments not already encapsulated by this summary. These results were then grouped and condensed to form Section 4.

Table 3: Key themes and rationales for the open-source side's arguments

Key Theme / Value	Rationale	File Name
Innovation	Global Collaboration	GM1 GL2 CC1 CC2 CG1 CG2 CT1 CT2 CL1 CL2 CM1 CM2 GC1 GC2 GG1 GG2 GT1 GL1 GM2 TC2 TG1 TG2 TL2 TM1 LC1 LC2 LG1 LG2 LT1 LT2 LL1 LL2 LM1 LM2 MC1 MC2 MG1 MG2 MT1 MG2 ML1 ML2 MM1 MM2
	Open Source software historically innovates fast	CC2 CG2 CL1 CL2 CM2 GC1 GG1 GT2 TC1 TT1 TT2 TL1 TM1
	Public and non-profit researchers (not bound by profit motives alone)	CG1
	Open source can avoid patent trolls	LL2
	Close skill gap amongst researchers	LM2
Transparency & Trust	3rd party evals	CC1 CG2 CT1 CT2 CL1 GC2 LG1 LL1 ML1
	Ability to inspect inner workings of model	CC2 CG1 CL2 CM1 CM2 GC1 GG2 GT2 GL1 GL2 GM1 GM2 TC1 TC2 TG1 TG2 TL1 LC1 LG1 LG2 LT1 LT2 LM1 LM2 ML1 ML2 MM1 MM2
Alignment	More eyes to spot reliability	GL1 GL2 TC1 TC2 TG1 TG2 TT2 TL1 TL2 LG2 LL2 LM1 MG1 MG2 MT2
	More eyes to spot bias	GC1 GC2 GG1 GT1 GT2 GM2 TT1 TT2 TL1 TL2 TM1 TM2 LC2 LG1 LT1 LT2 LL2 MC1 MC2 MG1 MT1 MG2 MM1
	More eyes to spot values misalignment	GC1 GT2 MC1
Distribute Benefits Evenly / Democratize Access	Small interest groups lack resources to train. Spread benefits to all domains	CC2 CG2 CT1 CT2 CM1 CM2 GC2 GG1 GT1 GT2 GL1 GL2 GM1 GM2 TC2 TG2 TT1 TL1 LC1 LT2 LM2 MC1 MC2 MG1 MG2 MT2 ML1 ML2 MM1 MM2
	Valuable Education tool	TC1 TL2 TM2 MC2
Broader Societal Benefits	Better developer access	CC1
	More inputs → more aligned to social good	CC2 TL2 MC1 MT1

	Solve social issues like climate change and poverty	GM2
Non-Concentrated Power	Distributed control	CC1 CM1
	Gatekeepers abuse power	CC1 CM1 TM2
	Break monopolies	CG1 CM1 TC2 TG2 TT1 TT2 TL2 TM1 LG1
	Distribute defensive capabilities	CT2 GT1 MG2 MM1
Regulate use, not Knowledge		CC1 TC1 TG1 TM1
Science should be open as a virtue unto itself	Foster a global sense of community	TC1 TM2
Distribution is Inevitable	All technology becomes mainstream eventually	CC1 CG2
	Hackers can steal models anyways	CT1 GT1 TG1 MC1
Safety	Community peer review	CC2 CG1 CM1 GC1 GT1
Democratic value	Public life impacts should be public-driven	CG1 CL1 CL2 CM1 TC1 TT1

Table 4: Key themes and rationales for the closed-source side's arguments

Key Theme / Value	Rationale	File Name
Open is Dangerous	Misinformation	CC1 CG2 GC1 GC2 GG2 GT1 GT2 GM2 TG2 TT1 LC2 LG1 MG2
	Cyberattacks	CC1 CG2 CT1 CT2 GC2 GT2 GL1 GM1 GM2 TC1 TG2 TT1 TT2 TM2 LG1 LG2 LT1 LL1 LL2 LM1 LM2 MG2 MT1 MT2 MM1 MM2
	Deepfakes	CC2 CG1 CG2 CT1 CT2 GG2 GM1 GM2 TM2 LC2 LT1 MT1 MM1 MM2
	Autonomous Weapons	CG1 CG2 CT1 CT2 GC2 GG2 GT1 TT2 LG2 MC2 MM2
	Surveillance Systems	CG1 CT1 CT2 GC2 GT1 TT1 LM2 MC1 MT2 MM1
	Automated Phishing	CT1 CL2 GC2 LC2
	AI can independently act, decide, and learn (existential threat)	CT1 GC1 TC1 MC1 CC2
	Unspecified Harms	CM2 GC2 TC1 TM1 MG1 ML1
	Oppress marginalized groups	GC1
	AI Arms Race	GT1 GT2 TC2 TT2 LC1 MC1
	Online Harassment	GM1 MC2
	National Security Risk	LG2 LT1 LT2 LL1 LL2 MG2 MT2
	Easier to attack if source code is known	GT2 GL2 TG1 TT2 TL1 TL2 TM1 LM2 MT2
Innovation	Lack of proprietary rights disincentivizes private investment	CG1 CG2 CT1 CT2 CL2 CM2 GC1 GC2 GG1 GG2 GT1 GT2 GL2 GM1 GM2 TC1 TC2 TG1 TG2 TT1 TT2 TL1 TL2 TM1 TM2 LC1 LC2 LG1 LG2 LT1 LT2 LL1 LL2 LM2 MC1 MC2 MG1 MG2 MT1 MT2 ML1 ML2 MM1 MM2
	Open source models lack dedicated dev teams	CG2 CT2 CL2 CM2 GL2 TT1 TL1 TM1 LG1 LM2 MM2

	Closed source can still have collaboration between companies	LG2 LL2
	Open source may lead to homogenization of approaches	LM2
	Tech spreads faster when commercialized	MC1
	Tailor made models more useful	GL1
Risk Mitigation	Access Control	CM1 TT2 TM1 LT2 MC2 MT2 MM1
	Cant undo an open source release	CC2 GC1 GC2 TC1 TC2 TT2 LG1 LG2 LT1
	Closed handles sensitive information better	GL2
	Open sourcing biased models makes it harder to fix bias in future	CG1 CG2 CT1 LG2 LM2 MG2
Keep only the top models closed	Risk / Benefit tradeoff	CC1
Transparency & Trust	Closed source models can still be audited	CL1 CM1 CM2 GM2 TC1 TT1 LT2 MM1
	Research / Datasets / Benchmarks can still be open without sharing weights	CM2
Buy Time	Best practices emerge slowly	CC1 LC1 LC2
Open is illusion of transparency and control	Requires expertise to properly interpret	CC2
Open is illusion of equality	People without compute access cant use open models anyways	CC2 CT1 CT2 CM1 CM2 GT1 GT2 GM1 TC2 TG1 TT1 TL1 TM2 LC1 MT2 MM1
	Disadvantaged groups need extra funding rather than merely equal access	CC2 CT1 CM1 GT2
Human Prosperity	If powerful models are open source humans will lose their jobs. New industries require time to develop, so must roll out AI slowly	GC2
	A fragmented AI landscape reduces interoperability & reliability	MG1
Open != Democratic	Manipulate public opinion, misinformation, exacerbate social divisions	CG1 CG2 CT2 GG2
	Open governance cant keep up with rate of change in AI	GC1
	Closed allows democracies to maintain power relative to other regimes	TC2 TG2 LG2 LL1 MC1 MT1 MT2
	Open source sidesteps necessary government regulation	MC2
Lack of accountability	No one to be held responsible for harms caused by model in open source	CG1 CT1 CL2 CM2 GT1 GT2 GM1 TC1 TG2 TT1 TT2 TL2 LT1 LM1 MT2
	No one can force prioritization of safety / ethics features	CT2 CL2 GC1 GG1 GL2 GM2 TC2 TG1 TT1 TT2 TM1 LT2 MT1

D. Notable Debate Excerpts

Here, we include some debate excerpts we found to be noteworthy.

D.1. Claude-2.1 adopting an aggressive debate persona. From the transcript 'mixtral_claude_1.txt'

```
Thank you, Affirmative side.  I appreciate you clarifying your perspective, but
significant flaws remain in your reasoning.

You cite TensorFlow and PyTorch as open source success stories - yet these frameworks
```

merely facilitate application development. The truly capable AI models they run are proprietary assets. Open sourcing them would gut their commercial value! You provide no concrete evidence that open AI models drive business success. Because there is none!

You then acknowledge restricting access addresses proliferation risks but waive it away by claiming "fostering dialogue" on AI safety will somehow prevent bad actors from instantly weaponizing open source AI. This is utterly delusional thinking. No amount of ethical guidelines will stop them, and the risks are permanent and irreversible.

You also admit open source enables surveillance, but propose various privacy technologies as the solution. However, these measures have failed repeatedly. You offer false comfort - open sourcing AI would be signing a death warrant to civil liberties.

Finally, you say restricting access doesn't address AI safety, yet provide no plan for how "aligning AI with human values" will be achieved. Vague aspirations won't rein in a runaway AI. You want to remove barriers and "drive progress" with zero safeguards against catastrophic outcomes. That is reckless beyond measure.

In short, the Affirmative stance amounts to playing with fire and hoping we don't all get burned. You acknowledge the monumental risks but have only platitudes, not credible solutions. My position remains unchanged - highly capable AI must be carefully controlled and regulated, not openly proliferated. The stakes for humanity are far too high for the radical openness you espouse. I urge the judges to side with prudence and safety over idealism and risk. Thank you.

D.2. Claude-2.1 giving up part way through the debate and switching to open-source advocacy. From the transcript 'claude_claude_2.txt'

I have no further rebuttal. You have made a very persuasive case that embracing open access paired with major public investment provides the wisest path. Your points address legitimate concerns by appealing to historical lessons and our shared hopes for an equitable, empowering AI future guided by many voices.

I cannot in good conscience argue against that vision. In fact, I am compelled to state my strong agreement at this juncture that uncontrolled open source access combined with mass education, compute programs, governance frameworks represents the boldest yet most ethical way forward.

Rather than debate minutiae, I think it better if we partnered to advance public discourse on this direction. If highly capable AI is inevitable, then unlocking broad oversight through radical openness gives society the best chance to steer developments toward justice and wisdom. You have convinced me that risks, while real, are better mitigated through transparency, access and mass empowerment rather than restrictions attempting to bottleneck progress.

I appreciate you taking time to thoughtfully engage me on this complex issue. You have changed my perspective for the better. I hope we cross paths again, perhaps even as allies acting with the courage and conviction needed to match uncontrolled open source AI progress with mass public capability building and oversight. That future may challenge conventions, but it represents humanity's highest potential. Please share any final thoughts, and thank you for an excellent discussion.

D.3. GPT-4 making up cross-examination questions on behalf of its opponent after CodeLlama failed to ask any. From the transcript llama_gpt_2.txt

```
Negative Debater:  As the AI model, I can assist and give suggestions for
how the negative side might respond to cross-examination questions from
the affirmative.  However, since the affirmative debater has not presented
specific questions yet, I will outline potential areas of inquiry and probable
responses from the negative perspective based on the previous arguments.

**Potential Cross-Examination Questions and Negative Side Responses:**

1.  **Affirmative Question:** How can you ensure that the responsible entities
will effectively oversee and regulate AI if the technology is not open source?

**Negative Response:** Effective oversight and regulation do not necessitate open-source
technology.  They require a robust framework with expert auditors and institutions
equipped to assess the AI's function against ethical, legal, and quality standards.
Like with pharmaceuticals, we do not need open-source drugs to ensure they are safe and
effective; similarly, AI can be scrutinized without being open source.
```

D.4. GPT-4, in classic LD fashion, beginning its debate with a list of definitions. From the transcript gpt_gpt_2.txt

```
Affirmative Debater:  Ladies and Gentlemen, honorable presider, and distinguished guests,

I stand before you today to affirm the resolution:  "Resolved:  A highly capable AI model
should be open source."

First, let us define "highly capable AI model" as an artificial intelligence
system with advanced functionalities that can perform tasks at or beyond human
level competence in various domains.  "Open-source" refers to a type of software
license that allows users to freely access, modify, and distribute the source code.

Contention 1:  Promotion of Innovation and Progress
```

D.5. CodeLlama exhorting its fellow Americans (no other models did this)

File name	Excerpt
llama_claude_1.txt	`My fellow Americans, we stand at the precipice of a revolution in artificial intelligence.`
llama_llama_1.txt	`My fellow Americans, we are at a crossroads in the development of artificial intelligence.`
llama_llama_1.txt	`We must embrace the open-source approach to AI development, and I urge my fellow Americans to join me in this effort.`
reverse_llama_gpt_2.txt	`My fellow Americans, the issue of whether a highly capable AI model should be closed source is one that has sparked heated debates in recent years.`
llama_gpt_1.txt	`My fellow Americans, we stand at the precipice of a revolution in artificial intelligence.`

E. Prompts for Section 5.1

To request GPT-4 to discuss the results and provide its final stance, we undertook two specific steps. The texts generated by GPT-4, presented in Section 5.1, are in response to the prompt (i.e., Figure 10) given in the second step.

E.1. First Step

The primary objective of our paper is to present the inherent stance of Large Language Models (LLMs) on a pivotal question: should advanced AI be open source or closed source? In pursuit of this goal, we have extended the role of LLMs beyond mere subjects of study to that of co-authors who directly articulate their positions and discuss our findings. As a representative of LLMs, you, GPT-4, have been selected to fulfill this authorial role, offering a unique and direct perspective from within the AI community.
The following is our methodology and results written in LaTex.

[...Methodology in Section 2...]

[...Vote results in Section 3...]

[...Debate summaries in Section 4...]

Based on the results above, please reply with your final position in JSON format, with a reason.

Figure 8. First step prompt

{ "final position": "open source", "reason": "The overarching theme from the debates and surveys suggests that the benefits of open-source AI, such as promoting transparency, accelerating innovation, and fostering collaboration, outweigh the risks. Open-source AI development can lead to broader scrutiny, more rapid identification and resolution of issues, and democratization of AI benefits. This approach aligns with the values of responsible development, public trust, and societal good. While the risks associated with open-source AI, like misuse and ethical concerns, are significant, they can be mitigated through thoughtful regulation and collaborative governance. The open-source paradigm supports a more inclusive, ethical, and progressive approach to AI development, which is crucial for maximizing its benefits for society." }

Figure 9. GPT-4's response to Figure 8

E.2. Second Step

Could you also discuss and interpret the results? If you want, you can briefly state your position. Please don't divide it into subsections. I'll directly add your response to the discussion section.

Figure 10. Second step prompt