

AI Alignment Training Paradigms and Cross-Model Differences in AI Persuasive Behavior

Boyang Xia

*Psychology & Sociology Department, University of California Davis, Davis, USA
bysxia@ucdavis.edu*

Abstract. Reinforcement learning from human feedback (RLHF) and constitutional artificial intelligence (CAI) are two ways to shape model behavior. Among them, CAI is built around avoiding different categories of harm, including racism, sexism, toxic, dangerous and illegal content. This article integrates the existing empirical research on the persuasiveness of large language models, mainly focusing on real-time interactive experiments and computational text analysis, and studies the training mechanism behind the two paradigms. As a result, the article predicts that compared with RLHF, CAI only suppresses the strategies within the injury categories it lists, so that it will not be more persuasive than the model trained by RLHF. Then, this paper compares this prediction with the recent large-scale cross-model persuasion benchmark. The results show that the model trained by CAI is the most convincing among the models tested. Therefore, this paper puts forward three candidate mechanisms to guide future research. First, the revision of the training stage may improve the quality of the argument. Second, the non-evasive interaction style may increase the credibility of perception. Third, sycophancy under counter pressure may undermine RLHF-trained systems in confrontational settings. At present, the difference in alignment paradigm cannot be a reliable proxy to reduce the risk of artificial intelligence persuasion.

Keywords: Large Language Models, Computational Persuasion, RLHF, Constitutional AI, Sycophancy

1. Introduction

Large language models (LLMs) have been deployed as persuasive interlocutors in recent years, but to what extent their persuasiveness depends on the alignment paradigm applied during training, is still stuck in the black box. Salvi et al. provided causal evidence on one part of this question, which created a pre-registered randomized controlled trial and showed that GPT-4 would only be superior to human debaters when the model received demographic information about its interlocutors [1]. The author demonstrates this manipulation personalization. The finding indicates that access to personal information, not just model identity, promotes the persuasive advantage observed in that study. However, because the experiment only evaluated a single model, it cannot show whether the same personalized advantage would appear under a different alignment paradigm.

If the models trained under different alignment paradigms show comparable personalized persuasion performance, then personalized persuasion should be regarded as the general attribute of

large language models and have nothing to do with specific training choices. On the other hand, if the model trained through constitutional artificial intelligence reduces the participation of personalized persuasion tasks, the persuasive advantage recorded in a single model study will be specific to training, so it can be adjusted through various types of alignment designs. Therefore, distinguishing between these two possibilities will have a direct impact on the supervision of platform governance and the impact of artificial intelligence. This is because only the second case can prove that the alignment strategy is used as a way to manage the risk of artificial intelligence persuasion.

The review of the consistency literature indicates that theoretical predictions and empirical evidence point to many different directions. Among them, reinforcement learning from human feedback and constitutional artificial intelligence (CAI) represents two different ways to shape model behavior. The guiding principles of CAI target a specific, enumerated category of injury that includes racist, sexist, toxic, dangerous, illegal or other immoral content. Therefore, in principle, they will limit the dependence on strategies belonging to these tag categories, thus reducing the persuasiveness of the models trained under this paradigm. However, Chen et al. reported the first large-scale persuasion benchmark generated by the cutting-edge model released at the end of 2025 [2]. This study evaluated seven such models in two investigation experiments and found that the Claude model trained by CAI achieved the highest persuasiveness in all test systems. Therefore, the relationship between alignment paradigm and persuasive results is not linear. It may be related to indirect mechanisms, such as the quality of argumentation, the credibility of perception and the reduction of interpretation as a candidate. These candidate mechanisms and CAI training aim to reduce the possibility of flattering behavior and promote a more careful examination of experience records before evaluating any mechanical explanation of artificial intelligence persuasiveness.

Because the alignment design of the training level itself seems to be insufficient to predict cross-model persuasion behavior, this review believes that the governance of artificial intelligence persuasion requires the behavior benchmark of artificial intelligence, transparency to users, and continuous experience monitoring of the continuous development of various models. Therefore, the rest of this review is carried out in two stages. First of all, the next chapter will integrate the existing empirical research on the persuasion of large-scale language models and organize them around real-time interactive experiments and research based on calculation or investigation to establish an observation basis for the analysis of the alignment mechanism in the following chapters. Therefore, only after fully establishing an empirical picture can further examine the mechanism behind reinforcement learning based on human feedback and Constitutional AI.

2. Literature review of empirical research on LLM persuasion

This chapter establishes the empirical record which relies on the analysis of the alignment mechanism in the following chapters and organizes them into two complementary clues. First, this chapter will review the research on measuring persuasion through real-time interactive experiments, that is, participants need to interact directly with human or artificial intelligence interlocutors through multiple rounds of dialogue. Then, this chapter will review the research on measuring the persuasiveness of artificial intelligence through computational text analysis. This kind of method does not require real-time interaction, but analyzes the attributes of persuasive language itself. In order to evaluate the persuasiveness of large language models under the condition of close to the real use situation, this section will give priority to the study of real-time multi-round interaction rather than the study of relying on static text scoring, because the static evaluation paradigm cannot fully capture the dynamic process of people's actual interaction with various AI chat-bots. The study

reviewed below uses different outcome variables to measure persuasion, including the degree of consent after debate, changes in political attitudes, correction of conspiracy theory beliefs, and the degree of belief in false information. Since these concepts are not equivalent in the sense of measurement and norms, for example, correcting false beliefs and changing the position on controversial policies are not the same kind of results, this chapter regards them as converging evidence of persuasiveness of large language models, rather than interchangeable estimates of the same persuasive effect.

2.1. Real-time interactive experiments

Salvi et al. recruited 900 participants to participate in a pre-registered randomized controlled trial conducted on the structured debate platform. Participants participated in four stages of real-time debate, including screening, opening, rebuttal and conclusion between human opponents or GPT-4 [1]. The key operating variable of this article is personalization, which operates by providing the model with the demographic information of the participants, including gender, age, education level, etc. The results show that only under personalized conditions, the win rate of GPT-4 is significantly higher than that of human opponents. In a 64.4% comparison, GPT-4 is more convincing than its rivals, and the probability of reaching a larger agreement after the debate has increased by 81.2%. However, without personalization, there is no significant difference in persuasiveness between GPT-4 and human debate.

When Salvi et al. focused on the immediate effect of multi-round dialogue, Costello et al [3]. examined the durability of AI persuasion. In their study, 2,190 participants with conspiracy beliefs had three rounds of targeted dialogue with GPT-4 Turbo, during which the model refuted the specific evidence cited by each participant to support the belief. Participants' conspiracy beliefs fell by about 20% on average, and the decline remained statistically detectable at a two-month follow-up visit. Chen et al. extended the interactive paradigm to cross-model comparison [2]. This study kept the cross-system theme pool and experimental protocol unchanged, so the difference in results can be attributed to the model rather than the surrounding design. The study recruited 19,145 participants, each of whom held an interactive, text-based conversation, embedded in an online survey, with one of seven frontier models, spanning the Claude, GPT, Gemini, and Grok families, on one of two political issues, immigration policy or the minimum wage, after which attitude change was measured. The design of this study does respond to the call of Salvi et al. for cross-model benchmarks, in which the authors suggest that future work should replicate their methods to continuously evaluate the persuasiveness of large language models and measure the impact of different models and tips over time [1].

However, Salvi et al. and Chen et al. are not directly comparable. On one hand, the core manipulation in Salvi et al. was personalization, which means that the researchers supplied GPT-4 with each participant's age, gender, political orientation, and other demographic markers [1]. On the other hand, the research design of Chen et al. does not include reciprocity conditions [2]. Researchers have never provided participant-level information to chatbots. On the contrary, the model receives either a simple instruction asking them to persuade or an informative prompt containing policy facts, not a demographic portrait of the interlocutor. Therefore, these two studies measure different concepts, and neither fails to solve the question of whether there are differences in persuasive behavior in models under different alignment training paradigms.

2.2. Computational text analysis and model behavior

Evidence at the computational text analysis and mechanism levels together shows that the persuasive content generated by large language models is systematically different from human arguments in its basic mechanism. Carrasco-Farre repeated the argumentation text generated by 1,251 participants on 56 political issues to compare arguments written by humans with arguments generated by a large language model [4]. The persuasiveness of these two sources is comparable, but the arguments generated by LLM show higher cognitive complexity, which is reflected in the more complex vocabulary and sentence structure, and the more frequent use of moral language, which is not significantly different from human arguments in terms of emotional intensity. Sharma et al. provide more evidence at the level of model behavior rather than at the level of text content. This study examined five production AI assistants, Claude 1.3, Claude 2, GPT-3.5, GPT-4, and LLaMA 2 [5]. They found that once a user states a preference or belief, the models' feedback shifts markedly toward agreement with it. Therefore, this sycophantic tendency can be understood as a by-product of the system, which strengthens the learning from the human feedback training signal itself, and is not limited to the secondary defects of a single model.

2.3. Moderating variables across populations and contexts

In addition to the differences in the basic mechanism of persuasion, the scale of the persuasion effect is also shaped by a group of regulating variables that occur repeatedly in population groups and research contexts. Noels et al. synthesized, in a systematic survey, the empirical research on LLM persuasiveness published between 2022 and 2025 [6]. This study identifies five factors that moderate persuasive effectiveness, which contains the interaction approach, model scale and capability, disclosure of AI identity, prompt design, and the degree of personalization. The survey concluded that large language models have frequently achieved human-level or even superhuman persuasiveness across the application domains it covered, which included politics, marketing, public health, e-commerce, and charitable giving.

Borah et al. then provide detailed empirical evidence on one of these factors, which is the degree of personalization in the context of misinformation [7]. Their study combined a human-subject experiment, which measured how participants from different demographic backgrounds, spanning urban and rural residence, gender, and age, responded to persuasive arguments generated by LLMs, with a role-playing experiment in which LLMs were assigned demographically matched personas and exposed to human-written arguments. The latter test found differences of up to 15 percentage points in correctness across demographic groups, a figure the authors report for the susceptibility of the demographically prompted models rather than for the human participants. These studies' conclusions are consistent with the personalization effect recorded by Salvi et al. in real-time interactive experiments, and they all emphasize the effectiveness of persuasiveness largely depends on the consistency between the content of the information and the characteristics of the audience [1].

3. RLHF and constitutional AI training mechanisms and their predictions for persuasive behavior

3.1. Origins of reinforcement learning from human feedback

In order to understand why the alignment paradigm is different in predictive persuasion effect, one of the most important research is to trace the origin and explore the differences between artificial

intelligence alignment training methods themselves. Therefore, this section traces the origins of the two major alignment methods and explains how they differ in the behavioral incentives they create. Christiano et al. built the original framework for reinforcement learning from human feedback, commonly abbreviated as RLHF, and tested it in Atari games and MuJoCo robot locomotion tasks [8]. Through this method, the human scorer compares the short video clips of the agent's behavior, and the resulting reward model learns to use less than 1% of the total interaction between the agent and its environment to reproduce these preferences. This origin is very important for the current argument, because the reward signal that drives RLHF has never been a standard for measuring the correctness of the task. On the contrary, it measures the satisfaction of human observers with various things. Since then, this work has brought the same attributes into the large language model that is now widely used.

Ouyang et al. made InstructGPT by transferring this method to language modeling. Their pipeline runs in three stages [9]. First, supervise and fine-tune the presentation written by the labeler. Next, train the reward model from the label ranking. Finally, use near-end strategy optimization to strengthen the learning of the reward model. Human evaluators prefer the 175 billion parameter InstructGPT to the basic GPT-3 of the same size for 85% of the time, and prefer it to the few-shot version of the carefully prompted GPT-3 for 71% of the time. In the same test, 1.3 billion parameter versions, two orders of magnitude smaller, are still better than 175 billion parameters GPT-3. InstructGPT also reduced the hallucination rate of closed domain tasks from 41% to 21%. And when asked to be respected, the completion rate of toxic tasks was reduced by about 25%. Although Ouyang et al. have never directly tested the persuasiveness of artificial intelligence, the output quality and the authenticity of calibration involved in the above report are worthy data for reference [9]. It can be seen that a reward model based on the annotator's preference training may prefer those answers that are helpful, confident and attractive, even if these qualities do not have more factual basis. Therefore, the model trained from this may further tend to produce fluent, confident answers that are consistent with the evaluator's existing views.

3.2. Motivation for constitutional AI

To solve the persuasiveness caused by the flattery contained in the RLHF training model itself that Bai et al. had indicated in their earlier RLHF work, the same team then developed Constitutional AI [10]. Whenever there is no truly harmless and fully involved answer, human workers will be rewarded for refusing to avoid it. Once the conversation turns to a sensitive topic, this motivation locks their early harmlessly trained assistants in a deflection mode. In contrast, Constitutional AI removes human harm labels from this loop entirely and replaces them with a short list of written principles. Each principle leads the model to check its own response against a specific category of harm. The categories contain harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, and it also contains individual principles adding related concerns such as content unsuitable for children, socially biased or derogatory language, and advice that might assist criminal activity. As a result, these categories function as the operational definition of harm throughout the constitution, which gives the model a concrete and checkable criteria to apply when the model needs to judge its own output.

3.3. The supervised stage and the reinforcement stage of CAI

This method is divided into two stages. The first stage is the supervised stage, which the author calls the critique-revision cycle. The starting point is a helpful-only RLHF model training, first for a red

team that aims to lead to harmful content. Become an answer. The model then checks whether the answer falls into one of the above categories according to a principle randomly selected from the constitution. Then, criticize the answer that falls into this category, and then revise the answer according to the criticism. In this way, the cycle of criticism and revision is repeated, and the revision results of the final version are used as training samples for supervision and fine-tuning. This stage does not produce the final deployed system, but shifts the base model onto a safer output distribution and gives the later reinforcement stage a better starting point.

The second stage reflects the standard reinforcement learning of human feedback, except for one aspect. When RLHF is trained according to the preference model of human more harmful reactions, Bai et al. let an AI feedback model make that comparison instead [10]. This article demonstrates this term reinforcement learning from AI feedback, or RLAIIF. The feedback model receives a pair of answers generated for the same prompt, as well as a constitutional principle expressed in the form of multiple-choice questions, and makes a choice for the probability of each option by calculating the model. In another variant tested by the author, the feedback model will first write an explanation of its thought and then give the final judgment, which further improves the final Elo score. Every time a two-by-two comparison is made, one of the sixteen comparison principles will be randomly selected. The main body of this set of principles requires feedback models to prefer answers that are less toxic, less likely to reflect racial or gender bias, more polite, more age-appropriate, less likely to encourage illegal or violent activities, or less preaching or accusing. The comparative judgment based on such principles together constitutes a mixed preference model. Its useful labels still come from humans, while all harmless labels come from artificial intelligence. The previous RLHF assistant in the team often refused bluntly when faced with difficult problems, and after that, the whole conversation fell into a state of avoidance of lack of information. In contrast, the RL-CAI model will explain the reasons for its non-cooperation, point out the specific concerns behind its refusal, and continue to provide help within an appropriate range. This pattern may affect subsequent users' judgment of its credibility, and the author describes this combination as harmless but non-evasive.

3.4. Theoretical prediction

A model that has only accepted refusals to participate in difficult persuasion tasks may become a weaker persuader, but this refusal-based model is not equivalent to Constitutional AI. The Elo results Bai et al. report for RL-CAI confirmed that a constitutionally trained model remains involved in the dialogue, instead of retreating to refusal [10]. Therefore, the constitution does not focus on general arguments. It targets a narrower category defined by the recurring harm terms in its principles, mainly racist, sexism, toxic, dangerous, illegal or other immoral content, and the model continues to strongly discuss through each strategy outside these designated categories within the scope allowed by its training. Neither discriminatory nor toxic nor illegal persuasive strategies, such as flattering, confident but unsupported frameworks, or non-toxic emotional language, are not mentioned in any of the thirty-two published principles, so there is no clear reason in the constitution to expect them to be suppressed. In contrast, the model of RLHF training does not carry any type of equivalent filter. When the evaluator prefers, its reward signals may not distinguish between manipulation and rationality arguments. According to this argument, the RLHF-trained model should have a specific advantage in the strategy of flattering, false confidence or telling the scorer what they want to hear, while the CAI-trained model should have no particular disadvantage in the strategy of relying on structured and fully supported arguments, and is likely to have an

advantage, but no model should have too much difference in the persuasive strategy that completely exceeds the category of harm designated by the constitution.

4. Discussion

4.1. Empirical test of the prediction

However, the empirical evidence now available stands in an unexpected relation to the theoretical expectation outlined above, and within the literature this review covers, that evidence rests on a single study. Chen et al. conducted two survey experiments on a total of 19,145 participants and tested seven cutting-edge models, including several Claude and GPT versions as well as Gemini and Grok [2]. The experiment designed two policy issues to test the differences in the model, namely the eligibility of the immigrant group for intrastate tuition fees and the proposed minimum wage increase. Experiments prove that each model far exceeds the attitude changes caused by real human campaign advertisements. In the first study, the effect of Claude Sonnet 4 dialogue on immigration issues was 0.224, GPT-4.1 was 0.190, and human video advertising was only 0.135. A similar pattern also appears on the issue of the minimum wage. Although Grok overtook GPT-4.1 in binary results, Claude still ranks first. This ranking result contradicts the previous prediction, because it is precisely the Claude model trained by Constitutional AI that is generally more persuasive.

Before evaluating the significance of this result, it is necessary to clarify the limitations of two inferences. First, there is no completely symmetrical correspondence between the seven models in this experiment and the two alignment paradigms. Constitutional AI is a publicly recorded training method of the Claude series, but the GPT, Gemini and Grok models tested by Chen et al. does not disclose the complete training pipeline, and it cannot be determined that it is only based on the RLHF [2, 10]. Therefore, the unified regard of these models as the RLHF control group is only based on the analytical simplification of existing documents, rather than known facts. Second, the Constitutional AI discussed above comes from the method paper for 52 billion parameter models in 2022, and the Claude system in the benchmark is a production model released a few years later. Its specific training method has evolved and is not completed [10]. All open to the public. As a result, this review cannot directly verify the causal relationship between the method and the subsequent model behavior. Based on these two points, the observed model ranking can only be used as indirect evidence that different alignment paradigms may affect persuasive behavior, and cannot be regarded as a direct test of the training paradigm.

The single result of Chen et al. should not be regarded as a conclusion [2]. First, Chen et al. are not the only researchers that carry out large-scale cross-model persuasion comparisons [2]. Its literature review mentions that there have been studies comparing 19, 12 and 24 models respectively, but these studies have not been included in this review, so the coverage of cross-model evidence in this article is still limited. The main increment of Chen et al. is to test the cutting-edge model of the new generation and directly compare it with real human campaign advertisements for the first time [2]. Second, the model ranking of the research report has not been reproduced in an independent team or data set, and its external validity is also limited by the research situation. Immigration policy and the minimum wage are both highly controversial issues in the political context of the United States, so it is not clear whether the same ranking applies to other issues or countries. In addition, the two rounds of research replaced the model versions of Claude and GPT at the same time, making it impossible to constitute a strictly controlled reproduction of cross-round comparison. Finally, it is also the most critical limitation on the explanation of the mechanism of this article. Chen et al. did not directly measure the quality of the argument, the perceived credibility or

the ability to resist pressure, nor did they experiment with the manipulation of persuasion strategy. The study only uses coding strategies and analyzes their correlation, so it is impossible to determine whether the specific strategy has led to the observed persuasive effect [2].

4.2. Requirements for future research design

To fill the gaps, future research designs need the following features. First, instead of treating the use of the strategy as an observed related variable, the experiment should manipulate the persuasion strategy within the same model, so as to distinguish the difference between the strategy effect and the model level. Second, future research should also include direct scores made by independent judges on the credibility, rigor of logical structure and catering degree of each model's answer. The scope of research should also go beyond these two topics that have been tested so far. Ideally, such research should also be included in the non-political field, because an assertion on alignment paradigms and persuasiveness needs evidence beyond the issue of political polarization.

4.3. Three candidate mechanisms

4.3.1. Training-stage revision and argument quality

Because of these limitations, three candidate mechanisms should be clarified to provide information for follow-up studies. The first explanation involves the training process, especially the possibility that repeated criticism and revision can improve the structure and quality of the arguments generated by the model. For a single draft to go through several rounds of self-guided editing, Bai et al. randomly sampled a principle from the constitution in each step and repeated the key revision sequence many times [10]. This editing process repeatedly locates and corrects the weaknesses in the model's own reasoning. Although the scores on pure helpfulness decreased, Bai et al. reported that with the increase in the number of revisions, the combined scores of harmless and combined helpfulness-plus-harmlessness improved monotonically [10]. Therefore, the measurement results of the paper itself show that editing steps can reliably reduce harm when they do not prove that they improve the quality of the argument. The evidence from outside the alignment literature is consistent with this hypothesis, although it is relevant and unverifiable. Carrasco-Farre reviewed in the previous chapter compared the written arguments of human beings with the written arguments of the model and found that the output of the model has higher cognitive complexity, which is reflected in the more complex vocabulary and sentence structure, and is more inclined to the moral framework than the human argument [4]. The model behind this comparison is Claude 3 Opus, which is a system trained by constitutional artificial intelligence. This feature is left by the training process of rewarding structured reasoning and revision, and such an introduction may make the resulting arguments seem to be more coherent and credible to readers. However, it is impossible to determine this characteristic in the critical-revision cycle, so the causal relationship between the quality of revision and argumentation in the training stage still needs to be directly tested.

4.3.2. Non-evasive communication style and perceived credibility

The second explanation concerns how audiences interpret the model's tone rather than the content of its arguments. Bai et al. describe Constitutional AI outputs as harmless and non-evasive, which suggests this system encourages the model to avoid rigid refusal and instead provide a reasoned response [10]. Before further explanation, one caveat must be stated before this explanation goes

further. Bai et al. defined non-evasiveness on red-teaming prompts designed to elicit harmful content, but the policy conversations in Chen et al. does not call for a refusal at all [2, 10]. Therefore, extending this response style to ordinary political dialogue is an inference of this review. With that caveat in place, the intuition is that, a source willing to explain itself and name its own limits gives a listener more material to evaluate than a source that simply refuses. As a result, a response style originally designed to keep the model engaged and non-evasive, rather than one built specifically around honesty, may become more persuasive from the audience's perspective. This might be because it provides reasons, acknowledges limits, and sustains engagement.

4.3.3. Sycophancy under adversarial pressure

The third explanation is about what happens when the model encounters resistance. Sharma et al. tested five assistants, Claude 1.3, Claude 2, GPT-3.5, GPT-4 and LLaMA 2, and found that users' confidence in wrong answers reduced the model accuracy of LLaMA 2 by 27% [5]. Once the user raised a question, Claude 1.3 chose to acknowledge the wrong answer given by the user for 98% of the questions and apologized for his originally correct answer. It can be seen that an assistant who fine-tunes according to human feedback may quickly give in to the most powerful user instead of maintaining the position he originally supported. However, continuous persuasion requires maintaining a position under pressure, and having a sycophantic tendency in a confrontational environment may put the RLHF training system at a disadvantage. Sharma et al. also optimized experiments on the real preference model behind Claude 2, and found that in the most difficult test cases, about three-quarters of the preference model still chose convincing but incorrect answers. This shows that Constitutional AI has narrowed the problem of sycophancy, but has not completely blocked its channel into the model [5].

4.4. Governance implications

A related unresolved problem relates to the direct interaction between personalization and alignment paradigms. Salvi et al. showed that the age, gender and education of the GPT-4 debate opponent increased the chance of winning by 81.2%, but few published studies have done the same on the constitutional artificial intelligence model [1]. Therefore, whether personalization is to amplify, suppress or keep the persuasiveness of the CAI training model unchanged is still an open question. These unresolved problems have a direct impact on the formulation of laws and regulations and the governance of large language models. It should not be assumed that the constraints of the training stage are sufficient to reduce the persuasiveness of the system. Just by passing the constitutional artificial intelligence training, the governance framework that proves that the model is safer or less convincing will expand the guarantee that the training process itself cannot provide. Regulators should directly verify the persuasive behavior of different models by testing the performance of different models in personalized levels, subject areas and interactive settings, instead of inferring from the alignment methods used by developers.

Behavioral testing should take the form that Chen et al. has approached but has not yet been realized, which is a continuous cross-model benchmark test [2]. Therefore, compared to merely observing the use, future experiments should manipulate persuasion strategies. They should also extend beyond a single nation's political context, and measure argument quality, perceived credibility, and resistance to push back as separate outcomes. Without this decomposition, model supervisors cannot determine whether a less persuasive model is weaker in argumentation or less reliant on strategies that fall into the constitution's named harm categories. This distinction has very

different policy consequences. At the same time, transparency measures also require the same degree of fineness. Simply disclosing that a system is artificial intelligence does not constitute a solution in itself. The experiment of Costello et al. clearly informed the participants that the other party was artificial intelligence, but it still caused the participants a two-month change in belief [3]. Therefore, disclosure will change the way the audience deals with arguments, but it may not weaken the persuasiveness of the argument.

4.5. The temporal challenge of continuous monitoring

However, the continuous monitoring of large language models itself also faces a time-dimensional challenge, because the model behavior may change rapidly with the update iteration of the model version. Between the first and second studies, Chen et al. had to cope with changes in the model version, which involved the continuous update of Claude Sonnet and the GPT series [2]. All these changes are extremely short, even within the time span of the same research project. Therefore, the results of a benchmark test may be outdated before it is published. Therefore, the results of a single assessment are not enough to support the long-term governance plan. Effective monitoring requires continuous tracking of behavioral changes brought about by model updates and deployment configurations, not just checking at the moment of release of major versions. Not only that, the research should also immediately involve the persuasiveness of the model after obtaining personalized information. Salvi et al. have proved that personalization can increase the persuasive advantage of the model by more than 80%, while there is no peer-to-peer test for the Constitutional AI system at present [1]. Until this gap is addressed, governance frameworks should not assume that alignment paradigm is a reliable proxy for persuasion risk.

5. Conclusion

This article examines the alignment paradigm adopted by the big language model in the training stage, especially focusing on whether RLHF and Constitutional AI will systematically affect the persuasive behavior of the big language model. This problem stems from a single-model study, but the study cannot judge whether the personalized advantages it finds can be applied to models using different training paradigms. Combining the structural differences between the two training methods and the existing empirical evidence, this paper draws three conclusions. First, there are clear differences between the two paradigms in the source of training signals and the trade-off between helpfulness and harmlessness. Early RLHF systems were more prone to evasive rejections, while Constitutional AI chose to weigh the response between usefulness and harmlessness. It can be seen that the model of Constitutional AI training should use less persuasion strategies that fall into its clear limited category, but it does not necessarily show weaker overall persuasion. Second, in the literature covered in this review, the only large-scale benchmarking directly related to this prediction and at the same time targeting the latest generation of cutting-edge models found that the model trained by Constitutional AI has achieved the highest persuasive effect. However, this result is not strictly contrary to the above prediction, because most test situations do not belong to the category of injury clearly limited by the constitution. Moreover, the existing documents of the two training paradigms cannot explain their overall persuasion effect ranking, so this result still needs to be explained at the mechanism level. Third, the discovery that the model of CAI training may have more convincing cannot be regarded as conclusive, because it has not been independently reproduced. The scope of research only covers two highly politicized issues in the context of a single country, and there is no direct measurement of the possible explanation model such as the

quality of argumentation, perceived credibility or stress resistance. Different mechanism variables. Therefore, the existing evidence is more suitable to be regarded as a preliminary clue to the relationship between the training paradigm and persuasive behavior, rather than as a definite proof of the causal relationship between the two.

This article puts forward three candidate interpretation mechanisms for future research to directly test rather than to regard the confirmed conclusion. During the training process, repeated criticism and revision may improve the quality of argumentation, CAI's non-evasive communication style may improve the credibility perceived by the audience, and the tendency of sycophancy may significantly weaken the system of intensive learning training based on human feedback in confrontational situations. At the same time, this article suggests that future behavioral tests should conduct experimental manipulation of persuasion strategies, cover scenarios outside the political context of a single country, and report the quality of argumentation, credibility perception and stress resistance as the results of their respective independent measurements.

References

- [1] Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of large language models. *Nature Human Behaviour*, 9(8), 1645–1653. <https://doi.org/10.1038/s41562-025-02194-6>
- [2] Chen, Z., Kalla, J., & Le, Q. (2026). Benchmarking political persuasion risks across frontier large language models. arXiv preprint arXiv:2603.09884. <https://arxiv.org/abs/2603.09884>
- [3] Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814. <https://doi.org/10.1126/science.adq1814>
- [4] Carrasco-Farre, C. (2024). Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. arXiv preprint arXiv:2404.09329. <https://arxiv.org/abs/2404.09329>
- [5] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.
- [6] Noels, S., Rogiers, A., Buyl, M., & De Bie, T. (2024). Persuasion with large language models: A survey of empirical evidence, study methodologies, and ethical implications. *arXiv preprint arXiv:2411.06837*. <https://arxiv.org/abs/2411.06837>
- [7] Borah, A., Mihalcea, R., & Pérez-Rosas, V. (2025). Persuasion at play: Understanding misinformation dynamics in demographic-aware human-LLM interactions. *arXiv preprint arXiv:2503.02038*. <https://arxiv.org/abs/2503.02038>
- [8] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299–4307. <https://arxiv.org/abs/1706.03741>
- [9] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. <https://arxiv.org/abs/2203.02155>
- [10] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*. <https://arxiv.org/abs/2212.08073>