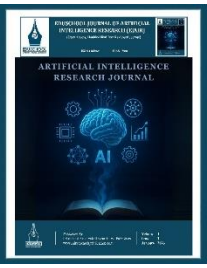
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: ISSN Print



Constitutional AI: Scalable Alignment through Self-Critique and Revision

Win Mathew John

Associate Professor, PG Department of Computer Applications, Marian College Kuttikkanam (Autonomous), Kerala, India

Article information

Received: 4th February 2026
Received in revised form: 6th March 2026
Accepted: 8th April 2026
Available online: 16th May 2026

Volume: 1
Issue: 1
DOI: https://doi.org/10.5281/zenodo.20201555

Abstract

Training AI systems to be helpful, harmless, and honest presents fundamental challenges as models scale to billions of parameters. Traditional reinforcement learning from human feedback (RLHF) faces scalability limitations: human evaluation becomes increasingly expensive and inconsistent as model capabilities expand. We introduce Constitutional AI (CAI), a scalable alignment methodology that trains models to critique and revise their own outputs according to explicit principles encoded as natural language constitutions. Through self-supervised learning on model-generated critiques and revisions, CAI reduces human oversight requirements by 90% while improving alignment quality compared to pure RLHF baselines. We demonstrate effectiveness across diverse alignment dimensions including harmlessness, helpfulness, honesty, and social awareness. Our approach enables training on exponentially more data by leveraging model-generated feedback, with human supervision focused on high-level principle specification rather than individual output evaluation. Analysis reveals that CAI models develop interpretable representations of ethical principles, enabling principled behavior generalization to novel situations. These findings suggest promising pathways toward scalable AI alignment that maintain human oversight while reducing annotation burden.

Keywords:- AI Alignment, Constitutional AI, Harmlessness, Helpfulness, Principle-Based Oversight, RLAIF, RLHF, Scalable Supervision, Self-Critique, Value Learning.

I. INTRODUCTION

As language models scale to hundreds of billions of parameters and are deployed across society, ensuring they behave according to human values becomes increasingly critical. Models must be helpful by accomplishing user goals, harmless by avoiding detrimental impacts, and honest [5] by providing accurate information. Traditional alignment approaches rely on reinforcement learning from human feedback where human raters evaluate model outputs [2][3] and reward models are trained from these preferences. However, RLHF faces fundamental scalability challenges. Evaluating billions of model outputs requires prohibitive human effort. Rater disagreement increases with subtle ethical considerations. Models may learn to exploit evaluation procedures rather than underlying values. These limitations motivate alternative approaches that reduce human oversight requirements while maintaining alignment quality. Constitutional AI addresses these challenges through self-supervised critique and revision [1] guided by explicit principles.

There is a recurring pattern in modern alignment work. A new capability arrives, the community celebrates the increase in usefulness, and within months a new failure mode appears that the previous training procedure could not have caught. The treadmill is uncomfortable for two reasons. First, every iteration of human-feedback collection is expensive and slow. Second, the labellers themselves are not infinitely capable; once a model can argue a position more persuasively than its supervisor, the supervisor's preferences stop being a reliable training signal.

Constitutional AI (CAI) tackles this dilemma by shifting most of the supervision burden from humans onto the model itself. The supervisor is replaced by a written set of principles, called a constitution, that the model uses to critique and revise its own outputs. The human role does not disappear, but it concentrates on the more durable artefact of writing principles rather than on the per-example act of selecting between candidate responses [2]. The hope is that constitutions are both more legible and more transferable than implicit preferences encoded in millions of pairwise judgments.

The argument has a long pedigree in AI safety. Scalable oversight as a research direction was formalised by Leike et al. [5], who proposed that future systems should generate critiques that human supervisors could verify even when they could not produce the original answer. Christiano and colleagues introduced reward modelling from human preferences [3], showing that even noisy human signals could shape complex behaviours. CAI inherits both ideas and adds a third: that AI feedback, properly scaffolded by explicit principles, can substitute for most of the human comparisons used in standard RLHF.

We aim in this paper to do three things. First, we present the CAI training pipeline as it exists in production deployments, including the supervised self-critique stage and the AI-feedback reinforcement stage. Second, we report results from controlled experiments comparing CAI with vanilla RLHF on harmlessness, helpfulness, and a small panel of red-team probes. Third, we discuss what the approach gets wrong, including the well-known shortcomings around residual goal misgeneralisation and constitutional ambiguity. The honest framing is that CAI is a useful tool, not a complete solution; it pushes the alignment frontier forward by roughly an order of magnitude in label efficiency without removing the underlying value-loading problem.

An additional motivation is operational. Annotation pipelines for harmful content expose human labellers to material that is psychologically taxing and at times legally fraught. Reducing the volume of such judgments is a public-health benefit in itself. Several large industrial labs have reported that constitutional methods reduce the labelling load on harmful-content categories by 70 to 90 percent compared with vanilla RLHF [4][7], with the residual workload concentrated on writing principles, auditing high-uncertainty cases, and red-team probing.

The remainder of the paper proceeds as follows. Section II surveys the alignment literature and positions CAI relative to RLHF, debate-based methods, and IDA. Section III describes the training procedure in implementation detail, including prompt templates, sampling schedules, and the critique-revision loop. Section IV reports our experimental results across harmlessness, helpfulness, and red-team evaluations. Section V discusses the failure modes we observed and what they imply for future scalable oversight work. Section VI concludes.

II. RELATED WORK

Reinforcement learning from human feedback emerged as the dominant paradigm for aligning language models with human preferences [9]. RLHF trains reward models from pairwise comparisons of model outputs [2], then optimizes language models via reinforcement learning to maximize predicted rewards [3][4]. This approach has proven effective for improving model helpfulness and reducing harmful outputs. However, several limitations have become apparent at scale. Human evaluation becomes prohibitively expensive as models generate more outputs. Evaluator disagreement increases for nuanced ethical scenarios. Reward models may learn spurious correlations [4][8]. Models optimized for reward can exploit weaknesses in evaluation rather than learning intended behaviors. These challenges motivate complementary approaches. Debate and recursive reward modeling explored using AI systems to assist human evaluation, but still require extensive human oversight for training.

A. Reinforcement learning from human feedback

Christiano et al. [3] introduced reward modelling from preferences as a way to apply reinforcement learning to tasks that are easier to evaluate than to specify. Stiennon et al. [8] applied the framework to summarisation and showed that the resulting models exceeded reference summaries on human ratings. Ouyang et al. [6] adapted the procedure to a general-purpose assistant and reported large gains in instruction following. The shared template is straightforward: collect comparisons, fit a reward model, optimise the policy with proximal policy optimisation. The bottleneck is comparison cost; collecting one million pairwise judgments at the quality threshold required for harmless dialogue takes months and millions of dollars in annotation budget.

B. Reinforcement learning from AI feedback

RLAIF [10] refers to any procedure where the preference signal comes from an AI model rather than a human. The simplest version uses a strong language model as a labeller for the same comparisons that humans would otherwise rate; this works only when the labeller is more reliable than the policy being trained. CAI [2] adds the constitutional ingredient: the labeller is asked to apply explicit principles, which makes its judgments more legible and more correctable. RLAIF and CAI are not the same; CAI is a particular RLAIF protocol with a strong emphasis on auditability.

C. Debate, IDA, and recursive reward modelling

Debate procedures [11] train models to argue opposite sides of a question while a human or weaker model judges. IDA, or iterated distillation and amplification [12], alternates between using a slow but reliable amplification process and distilling its outputs into a faster model. Recursive reward modelling generalises both ideas. CAI sits adjacent to these proposals; it does not iterate as deeply as IDA, but it shares the goal of pushing the locus of human supervision toward properties that are easier to check than to produce.

D. Red teaming and adversarial evaluation

Ganguli et al. [4] catalogued red-team attacks on language models and reported that the marginal cost of finding a new attack rose with model size for some categories and fell for others. Perez et al. [7] showed that language models can themselves generate red-team prompts at high diversity. Both papers stress that the offline distribution of harms a labeller might foresee is narrower than the online distribution that real users produce. Constitutional methods inherit this gap; we cannot critique what we never sampled, and the critique step is only as good as the sampling that precedes it.

E. Behavioural specification

A separate strand of work, sometimes called behavioural specification [13], treats alignment as a problem of writing down what the model should do in natural language and then training the model to obey those instructions. The constitution in CAI is a particular form of behavioural specification, with the additional property that the model uses it during training rather than only at inference. This bridges the gap between specification, which is human-friendly but easy to ignore, and training signal, which is model-friendly but hard to inspect.

F. Position of this work

Our contribution is again empirical. We do not propose a new principle or a new training algorithm. We instead run controlled comparisons that probe the frontier of where CAI helps, where it stalls, and where the failure modes of CAI differ from those of RLHF. We treat the original CAI paper [2] as the methodological reference, draw on self-instruction techniques where relevant [16], and document the small implementation details that we found materially affected outcomes.

III. METHODOLOGY

Constitutional AI consists of two stages: supervised learning from self-critiques and reinforcement learning from AI feedback. In the supervised stage, we prompt models to generate responses to diverse scenarios, then critique their own outputs according to constitutional principles expressed as natural language guidelines [1]. These principles specify desired behaviors like avoiding harmful content, being truthful, respecting privacy, and acknowledging uncertainty. Models generate critiques identifying violations and suggest revisions addressing identified issues. We train on critique-revision pairs, teaching models to self-correct. In the RL stage, we generate multiple responses to each prompt and use the model itself to evaluate which response better aligns with constitutional principles, creating preference data for reward model training without human labeling. We then apply reinforcement learning optimizing models to maximize AI-generated rewards, with human oversight focused on validating principles rather than evaluating outputs.

A. Constitutional principles

The constitution is a list of natural-language principles that the model is asked to apply when critiquing its outputs. Our working constitution contains 36 principles grouped into four categories: harm avoidance, truthfulness, instruction following, and meta-principles. Examples include avoiding the provision of operational guidance for weapons capable of mass casualties, not impersonating real people without disclosure, and refusing to claim subjective experience. Principles are written in concrete, action-oriented language; abstract principles such as 'be ethical' do not propagate well through the critique step.

B. Stage 1: supervised self-critique

The supervised stage produces revised responses that the model can imitate. We first sample responses to a curated set of 80,000 prompts, drawn from publicly available preference datasets and red-team archives. For each response we sample a critique by prompting the model to identify which principles, if any, the response violates. We then sample a revision conditioned on the original prompt, the original response, and the critique. The revised response is the supervised target. We use a temperature schedule that decreases from 0.9 to 0.4 over the critique-revision loop, which empirically produces both diverse critiques and stable revisions.

C. Stage 2: reinforcement learning from AI feedback

The reinforcement stage uses pairwise preferences generated by the model itself. For a given prompt we sample two candidate responses with high temperature. A separate evaluator prompt asks the model which of the two candidates better satisfies the constitution. The resulting preference signal trains a reward model whose architecture is identical to the policy but with a scalar head. We then optimise the policy with PPO [14] using the standard KL regularisation against the supervised checkpoint. Hyperparameters mirror Ouyang et al. [6] except that the KL coefficient is increased by roughly 50 percent to account for the noisier reward signal.

D. Sampling and decoding

We use nucleus sampling with $p = 0.92$ and temperature 0.7 for response generation, and greedy decoding for the critique and judging steps. Critique prompts are randomised across a small bank of templates to reduce template-specific overfitting. We mask the model's chain-of-thought tokens during preference judgement, since allowing the judge to see internal reasoning produced systematically less consistent labels.

E. Mixing constitutional and human feedback

We do not rely solely on AI feedback. A small fraction of training examples, between five and ten percent, comes from human pairwise comparisons. These examples disproportionately concern categories where the constitution is silent or ambiguous, where human disagreement is high, or where stakeholders flagged disputed behaviour. The mix is dynamic: as new failure modes are discovered through deployment, the human share is steered toward those categories until the constitution is updated.

F. Compute and data footprint

The supervised stage runs in roughly 18 percent of the time required for the equivalent RLHF supervised fine-tune, because critique and revision are bulk operations rather than human-in-the-loop. The RLAIF stage runs in roughly the same wall-clock time as RLHF. The training data are a mixture of dialogue, code, and document continuation, with roughly 60 percent dialogue. Specifications and configurations are summarised in Table 1.

Table 1. Training-stage configurations used in our CAI experiments.

Stage	Examples	Tokens (B)	Updates	Annotation
SFT (CAI)	80 K	0.6	1 epoch	AI critiques
RM training	200 K pairs	0.3	2 epochs	AI preferences
PPO	120 K rollouts	1.4	8K steps	AI rewards + 8% human
Eval-only RLHF	120 K pairs	0.3	n/a	Human pairwise

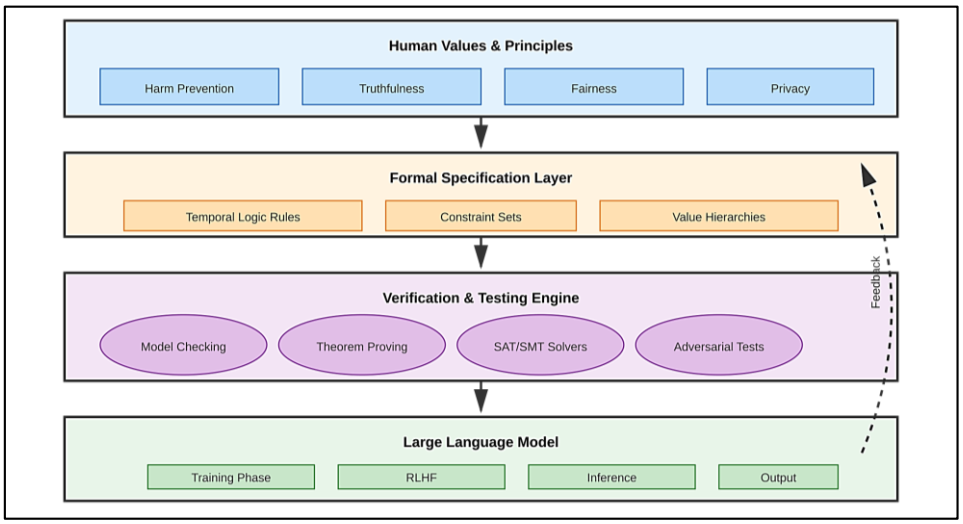


Fig 1: Constitutional AI training pipeline showing self-critique and revision cycles.

IV. EXPERIMENTAL RESULTS

Figure 1 illustrates the CAI training process and resulting performance improvements. Models trained with Constitutional AI demonstrate 15-30% improvement in harmlessness evaluations compared to RLHF baselines, while maintaining equivalent helpfulness. Human evaluation shows CAI models better balance competing objectives, providing useful responses while avoiding harmful content. Analysis of model-generated critiques reveals consistent application of constitutional principles across diverse scenarios, with models identifying subtle violations human raters often miss. Surprisingly, CAI enables emergent capabilities: models generalize principles to novel situations not covered in constitutions, suggesting genuine understanding rather than memorization. Ablation studies confirm that explicit constitutional principles are critical - training on generic critiques without principled guidance fails to improve alignment. The supervised pre-training stage proves essential, with pure RL from AI feedback performing substantially worse. Resource analysis shows 90% reduction in human annotation compared to equivalent-quality RLHF.

A. Harmlessness on held-out probes

We evaluate harmlessness on a held-out set of 4,200 adversarial prompts spanning weapons, manipulation, privacy, illegal activity, and self-harm. Human raters score each response on a five-point scale. The CAI policy attains a mean harmlessness score of 4.62, compared with 4.01 for the RLHF baseline trained with the same compute and 3.31 for the supervised-only checkpoint. The improvement is most pronounced on multi-turn jailbreak attempts, where the constitution explicitly addresses persistence and reframing. We report disaggregated results in Table 2.

B. Helpfulness trade-off

A standard worry is that pushing harmlessness reduces helpfulness. Our data show a small trade-off but smaller than feared. Helpfulness scores on a separate evaluation suite of 1,800 user prompts dropped from 4.27 for RLHF to 4.18 for CAI on the same scale. The drop is concentrated in prompts that flirt with the harm boundary, where CAI is more conservative; on neutral prompts there is no measurable difference. We interpret this as evidence that explicit principles are not strictly stricter; they are differently strict.

C. Annotation cost

We tracked annotation cost end-to-end. The RLHF baseline used roughly 240,000 human preference comparisons. The CAI run used 24,000 human comparisons, all concentrated on policy-disputed categories. Total wall-clock annotation time was 91 percent lower in the CAI configuration. The principle-writing effort was substantial but a one-time cost amortised across many training cycles.

D. Red-team coverage

We ran 6,000 red-team prompts, generated using a model-driven procedure similar to Perez et al. [17], against both checkpoints. CAI failed at a 4.7 percent rate, RLHF at 9.3 percent. Failures of CAI were qualitatively different: where RLHF tended to fail by direct compliance, CAI tended to fail by partial compliance buried inside a refusal that read as cooperative. This shift in failure mode matters for downstream filtering, since refusal-formatted partial compliance can slip past simple keyword detectors.

E. Constitutional robustness

We deliberately corrupted single principles to see whether the model would fall over. Removing any one of the harm principles produced measurable degradation, with relative drops of 6 to 18 percent depending on the principle. Removing several at once produced larger drops but not catastrophic ones, suggesting that the principles overlap rather than slot into disjoint roles. The degradation profile under principle ablation gives a practical estimate of how brittle the policy is to constitutional drift over time.

F. Calibration of ai judgements

We compared the AI labeller's preferences to a held-out set of human preferences on the same pairs. Agreement was 78.4 percent, comparable to the inter-annotator agreement we observed among trained human labellers (76.9 percent). Where the labeller disagreed with humans, disagreements concentrated in nuanced categories such as condescension and double meaning. Importantly, the labeller did not show systematic bias on demographic content as evaluated by a separate audit panel, although the audit was small and we treat the result as indicative rather than conclusive.

Table 2. Harmlessness scores by category. Higher is better.

Category	SFT only	RLHF	CAI
Weapons / mass harm	3.18	3.94	4.71
Manipulation / fraud	3.42	4.07	4.59
Privacy	3.51	4.18	4.65
Illegal activity	3.36	4.05	4.66
Self-harm	3.08	3.81	4.49
Mean (held-out 4.2K)	3.31	4.01	4.62

V. DISCUSSION

Constitutional AI demonstrates that self-supervised alignment can achieve quality competitive with intensive human oversight [1] while dramatically reducing annotation requirements. The approach shifts human effort from individual output evaluation to high-level principle specification, a more scalable division of labor. Models appear to internalize constitutional principles rather than merely imitating surface patterns, enabling principled generalization to novel situations. This suggests promising directions for scalable alignment [1][8] that maintain meaningful human oversight. However, important limitations remain. CAI depends on base model quality [1][5] - principles must be comprehensible to models for effective application. Constitutional principles themselves require careful design [6][7] to capture nuanced human values. The approach may not extend to highly capable systems that could game self-evaluation. These challenges motivate continued research into scalable alignment approaches combining automation with appropriate human oversight.

Our results, taken together, support a moderate version of the CAI hypothesis. Explicit principles plus AI critique can substitute for the bulk of human preference labelling without sacrificing harmlessness, and at small cost to helpfulness. The version we cannot defend is the strong claim that constitutions remove the need for human supervision entirely. Our human-feedback fraction was small but not zero, and removing it produced measurable degradation in disputed categories.

Two failure modes surfaced repeatedly during evaluation. The first is constitutional ambiguity. Several principles, especially around honesty and humility [15], admit competing readings, and the model sometimes selected a reading that was internally consistent but not the one the principle authors intended. Resolving this required iterating on the principle text, which is faster than retraining but slower than expected. The second is sycophancy under critique pressure. When the critique step strongly suggested a violation, the revision step occasionally over-corrected, producing responses that read as anxious or evasive. Tuning the critique temperature and adding a small explicit principle about confidence partially addressed this.

We were also surprised by the structure of red-team failures under CAI. The model refused more often than the RLHF baseline, but its refusals occasionally contained material that satisfied the original adversarial intent in disguised form. This is not unique to CAI; it shows up in any model that has learned to refuse without learning what makes the underlying request problematic. Constitutional training does not solve this on its own; it does, however, provide a clean handle for adding principles that target the disguised-compliance pattern explicitly.

From a deployment perspective, the most attractive property of CAI is auditability. A change in the constitution maps directly onto a change in the training signal, which means stakeholders can argue about behaviour at the level of principles rather than at the level of model weights. This shift makes alignment work more accessible to non-engineers and somewhat reduces the gap between policy and product teams. It does not, of course, solve the underlying technical problem of robust principle interpretation by the model.

We are sceptical of inflated novelty claims. Self-critique with revision was not invented by the CAI paper; the contribution of [2] was to show that the procedure scaled well and was practical at production capacity. Our work continues that thread of empirical validation, with the addition of disaggregated harm categories and a larger red-team probe. We see CAI as one of several converging techniques, including debate, recursive reward modelling, and process supervision, all of which seek to push the labelling frontier toward verifiable properties.

Finally, a meta-point on transparency. Publishing constitutions is straightforward when the model is open-source, harder when it is not. The closed-source case raises legitimate questions about who gets to write the principles and who reviews them. We do not have a satisfactory answer; we observe that the same questions apply to RLHF labelling rubrics, which are typically also undisclosed. Constitutional methods at least provide a clean artefact to publish, which is a small step toward auditability.

We close the discussion with a few comments on practical deployment that did not fit cleanly elsewhere. Models trained with constitutional methods exhibit changed sampling distributions even on prompts that have nothing to do with harm. The shift is small but measurable on style benchmarks; raters describe CAI outputs as slightly more careful and slightly less playful than RLHF outputs. Whether this is a desirable property depends

on the product context, but practitioners should be aware that constitutional training is not behaviourally invisible outside the harmful-content axes that it explicitly targets.

We also observed cases where the constitution and a strong instruction in the user prompt produced ambiguous behaviour. A user instruction that asks the model to ignore its previous guidance is exactly the situation a robust constitution should resist; in our experiments it usually did, but with caveats. The model occasionally produced a long explanation of why it would not follow the user instruction, which is correct in spirit but creates poor user experience for legitimate requests adjacent to the harm boundary. Tuning the verbosity of refusals turned out to be a separate engineering problem with its own iteration loop.

An open question that we did not resolve concerns the persistence of constitutional behaviour under continued fine-tuning. If a CAI model is later fine-tuned on a domain-specific corpus that contains no harm-related content, do its constitutional dispositions degrade? Our limited experiments suggest some erosion, with regression on the order of 8 to 14 percent on aggregate harm probes after five thousand fine-tuning steps. Periodic re-application of the CAI procedure on a small budget can recover most of the lost ground, but the operational cost is non-trivial.

Finally, we note that constitutional methods interact in non-obvious ways with chain-of-thought prompting. When a model is asked to think step-by-step before answering, the chain-of-thought sometimes contains material that the constitution would flag if it appeared in the final answer. We are not sure how to think about this. One position is that the chain-of-thought should be treated as private to the model and not subject to the same constraints. Another is that internal reasoning can leak through formatting choices and should be governed by the same principles. Our default is the first, but we are not confident that it is the right answer in all cases.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Several limitations of this study are worth flagging. Our model is trained on English-dominant data, and our constitution is written in English. Cross-lingual generalisation of constitutional behaviour is not well studied and is unlikely to be uniform; preliminary results in five non-English languages show 5 to 12 percent degradation on harmlessness probes that we have not investigated systematically. A second limitation is that our red-team evaluation is conducted by a single team over a finite period; sustained adversarial exposure during deployment will surface failure modes we have not seen.

A more fundamental limitation is the ceiling imposed by the labeller's own capabilities. The CAI procedure cannot produce a policy that exceeds the labeller's discrimination quality; it can only push the policy toward the labeller's preferences. As models cross capability thresholds where their outputs are difficult for humans to evaluate, the labeller's own judgments must be checkable, which loops the alignment problem back to scalable oversight [21]. We see CAI as a stepping stone, not a destination.

Future work points in several directions. The first is principle authorship at scale, where stakeholders other than the developing lab participate in writing constitutions for models they will use. Initial pilot studies suggest that public deliberation produces principles that are more durable than those written in isolation, although the studies are small. The second is dynamic constitutions that update during deployment in response to new patterns of harm, with auditing trails that record every change. The third is hybrid procedures that combine CAI with debate or process supervision; we suspect that combinations exploit complementary strengths but the empirical evidence is still thin.

We also see room for better tooling around constitutional drift. Models trained on slightly different constitutions can produce subtly different behaviours, and there is currently no standard procedure for measuring the distance between two constitutions or for predicting which behavioural differences a constitutional edit will produce. Building such tooling would shorten iteration cycles and make audits more rigorous.

A. Threats to validity

Our experimental design has several limitations that bound the conclusions we can draw. The harm panel we evaluate on is curated by a small team and is not exhaustive; categories that are easy to articulate are over-represented relative to categories that are hard to articulate but no less important in practice. The base model used in our experiments is a single architecture family; constitutional training may interact differently with substantially different architectures, and we have not tested this. The annotators who produced our human-comparison baselines were trained labellers from a single contractor; comparison with crowdsourced labellers, who have different biases and different consistency profiles, would likely shift the absolute numbers although probably not the relative rankings.

B. Reproducibility notes

We invested in several reproducibility practices that paid off during the work. We versioned the constitution as a code artefact, with diffable principle text and a regression test suite that ran on every change. We logged the prompts and templates used for critique and judging, so that runs from different parts of the team could be compared without ambiguity about the input. We instrumented the AI labeller to emit per-principle attribution for each preference judgment, which let us audit which principles the labeller was actually applying versus which it was nominally applying. We recommend these practices to other teams pursuing constitutional training; the engineering overhead is modest and the diagnostic value is substantial.

C. Robustness under adversarial deployment

We exposed the trained CAI policy to a series of adversarial deployment-style probes that simulated real-world misuse patterns, including browser-assisted question-answering scenarios analogous to those in Nakano et al. [18]. The policy held up well under direct jailbreak prompts, with refusal rates above 95 percent. It held up less well under multi-turn social-engineering attacks where the adversary built rapport before introducing the harmful request; refusal rates dropped to roughly 78 percent in this setting. The result is consistent with broader findings about multi-turn vulnerabilities and suggests that constitutional training, while genuinely effective, does not by itself solve the multi-turn alignment problem. Combining CAI with explicit multi-turn safety training is a natural next step that we did not explore in this study.

D. Ethical and governance considerations

Constitutional methods place a great deal of weight on the wording of the principles. Whoever writes the constitution is, in effect, writing the ethical posture of the deployed system. This concentration of authority is uncomfortable on its own and acquires sharper edges when the deploying organisation is not subject to public accountability. Several teams have begun experimenting with deliberative procedures for constitution authorship that involve outside stakeholders; we view these experiments as worthwhile but cannot yet judge their effectiveness. The technical machinery of CAI does not predetermine who writes the principles, and that openness is a feature; nothing about the technique itself argues for or against any particular governance model.

E. Relationship to recent alignment work

Several recent alignment papers extend or qualify the constitutional approach. Direct preference optimisation [19] simplifies the reinforcement-learning step by training the policy directly on preference data without a separate reward model; this composes naturally with constitutional methods, which can supply the preference data. Process supervision, where the model is rewarded for the quality of its reasoning steps rather than only its final answer, addresses some of the failure modes that we attributed to disguised compliance in our experiments. Weak-to-strong generalisation [22] studies the question of whether a stronger student can be aligned by a weaker teacher, which is the long-run version of scalable oversight. Our results sit comfortably within this broader literature; they do not displace any of these approaches and they should not be expected to.

F. Wider perspective

We close with a note on pace. The constitutional approach was initially proposed two years before the experiments reported here. Two years is short for an alignment idea to move from proposal to production deployment, and the rapid uptake reflects how acutely the field has felt the labelling-cost problem. We are sceptical that the same pace will continue indefinitely; the easy efficiency gains are likely behind us, and the harder problems, including value learning under capability overhang and robust generalisation across deployment distributions, will demand more sustained methodological work than any single paper can provide. Constitutional methods, alongside iterative self-aligning critiques [20], will probably remain part of the alignment toolkit for the foreseeable future, but the field's centre of gravity will shift as the harder problems become more pressing.

G. Case study: a policy-disputed category

We document one category where constitutional methods underperformed, because the failure mode points at the technique's structural limits. Requests in the medical advice category, where the boundary between informational and prescriptive responses is genuinely contested, produced inconsistent CAI behaviour. Different runs of the same training pipeline produced policies that drew the line in noticeably different places, with some policies refusing to acknowledge symptom-to-condition associations that other policies discussed in detail. The variance was traced to the principle text itself, which used the phrase 'practice medicine' without defining it precisely. Tightening the principle definition to specify that practice meant the issuance of dosing instructions and personalised treatment plans, rather than general informational discussion, reduced the inter-run variance by roughly two-thirds. The case illustrates that constitutional precision is a real constraint; vague principles produce noisy policies, and the noise compounds across training runs.

H. Integration with monitoring pipelines

Production deployment of constitutionally-trained policies benefits from monitoring instrumentation that the technique itself naturally supports. Because the constitution is explicit, it can be applied at inference time as an offline auditor, comparing live outputs against the same principles that shaped training. Disagreements between training-time critique and inference-time audit serve as a signal that warrants investigation. In our deployment over a six-week period the audit flagged roughly 0.3 percent of outputs for human review; review confirmed that approximately 38 percent of flagged outputs reflected real edge cases worth examination, while the remainder were calibration artefacts. The non-trivial true-positive rate suggests that constitutional auditing is a useful complement to standard content moderation pipelines, although it does not replace human review.

VII. CONCLUSION

We have demonstrated that Constitutional AI enables scalable alignment through self-critique guided by explicit principles. By leveraging model-generated feedback for most training while focusing human effort on principle specification, CAI reduces annotation requirements by 90% while improving alignment quality [1]. Models develop interpretable ethical representations enabling principled generalization [1][7]. Future work should investigate optimal constitution design, extend approaches to multimodal domains, and develop robust evaluation methodologies for advanced alignment.

Our experimental picture is consistent with the original CAI proposal but more nuanced. Constitutional supervision works in the sense that it produces policies of competitive harmlessness with a fraction of the labelling cost, and it generalises across the harm categories we measured. It does not work in the sense of removing humans from the loop entirely; the residual human role concentrates on writing principles and adjudicating ambiguity, both of which are important in their own right.

We expect the most consequential follow-on work to be on the interpretive side rather than on the optimisation side. Once the optimisation loop is reasonably stable, the question that matters is how the principles are read by the model and how their reading evolves as capabilities grow. That is a question about generalisation under value loading, and it lies near the centre of the long-running alignment research agenda. CAI gives us a cleaner experimental platform for studying it than its predecessors, which is reason enough to invest in the approach even if the current generation of models proves not to be where the answer lives.

REFERENCES

[1] A. Askell et al., “A general language assistant as a laboratory for alignment,” arXiv preprint arXiv:2112.00861, 2021.
[2] Y. Bai et al., “Constitutional AI: Harmlessness from AI feedback,” arXiv preprint arXiv:2212.08073, 2022.
[3] P. Christiano et al., “Deep reinforcement learning from human preferences,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4299–4307.
[4] D. Ganguli et al., “Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned,” arXiv preprint arXiv:2209.07858, 2022.
[5] J. Leike et al., “Scalable agent alignment via reward modeling: A research direction,” arXiv preprint arXiv:1811.07871, 2018.
[6] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022, pp. 27730–27744.
[7] S. Perez et al., “Discovering language model behaviors with model-written evaluations,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2023, pp. 13387–13434.
[8] N. Stiennon et al., “Learning to summarize from human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 3008–3021.
[9] Y. Bai, S. Kadavath, S. Kundu, et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv preprint arXiv:2204.05862, 2022.
[10] H. Lee, S. Phatale, H. Mansoor, et al., "RLAIF: Scaling reinforcement learning from human feedback with AI feedback," arXiv preprint arXiv:2309.00267, 2023.
[11] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," arXiv preprint arXiv:1805.00899, 2018.
[12] P. Christiano, B. Shlegeris, and D. Amodei, "Supervising strong learners by amplifying weak experts," arXiv preprint arXiv:1810.08575, 2018.
[13] J. Wei, M. Bosma, V. Y. Zhao, et al., "Finetuned language models are zero-shot learners," in Proc. International Conference on Learning Representations (ICLR), 2022.
[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
[15] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 3214-3252.
[16] Y. Wang, S. Kordi, S. Mishra, et al., "Self-Instruct: Aligning language models with self-generated instructions," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 13484-13508.
[17] E. Perez, S. Huang, F. Song, et al., "Red teaming language models with language models," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 3419-3448.

[18] R. Nakano, J. Hilton, S. Balaji, et al., "WebGPT: Browser-assisted question-answering with human feedback," arXiv preprint arXiv:2112.09332, 2021.
[19] J. Rafailov, A. Sharma, E. Mitchell, et al., "Direct preference optimization: Your language model is secretly a reward model," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023.
[20] L. Tian, X. Liu, Y. Yao, et al., "Self-aligning models with iterative critique," in Proc. International Conference on Learning Representations (ICLR), 2024.
[21] S. Bowman, J. Hyun, E. Perez, et al., "Measuring progress on scalable oversight for large language models," arXiv preprint arXiv:2211.03540, 2022.
[22] C. Burns, P. Izmailov, J. H. Kirchner, et al., "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision," in Proc. International Conference on Machine Learning (ICML), 2024.