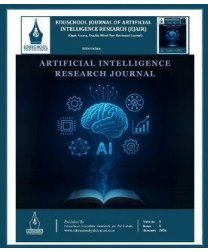
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



# Scalable Oversight for Advanced AI: Recursive Reward Modeling and Debate

Raji N

Assistant Professor, Department of Computer Science, YuvakshatraInstitute of Management Studies (YIMS), Mundur, India

**Article information**

Received: 10th August 2026
Received in revised form: 12th September 2026
Accepted: 14th October 2026
Available online: 20th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0012

**Abstract**

Reinforcement learning from human feedback trains language models against a reward model fitted to human comparisons, and it works only as long as the people giving feedback can tell a good output from a bad one. Scalable oversight is the problem of supervising systems whose outputs exceed that competence. This article reviews the two oldest proposals for the problem, recursive reward modelling with its close relative iterated amplification, and debate between competing models before a human judge, and places them beside two newer lines of work, weak-to-strong generalization and model-written critique. The review first sets out why direct feedback fails under optimization pressure, drawing on published evidence of reward hacking, reward model overoptimization, sycophancy and unintended persuasion. It then examines the assumptions each oversight method needs, using simple cost and error formulas to show what decomposition and adversarial argument gain and where their guarantees stop. The empirical record on debate is reported as it stands, with clear gains for multi-turn debate in reading comprehension tasks where the judge cannot see the source passage and small or mixed effects elsewhere. A comparative framework is proposed that links each method to its required assumption, its characteristic failure mode and a diagnostic test, together with an evaluation protocol built on the sandwiching design. The framework is analytical and has not yet been tested empirically.

**Keywords**:- Scalable Oversight, Reinforcement Learning from Human Feedback, Recursive Reward Modelling, Iterated Amplification, AI Safety Via Debate, Weak-To-Strong Generalization, Sandwiching

## I. INTRODUCTION

Amodei et al. listed scalable supervision among five concrete problems in AI safety [1]. The problem arises when the true objective is too costly to evaluate as often as training requires, and the usual response has been to learn a reward function from a much smaller number of human judgments. Christiano et al. trained Atari and simulated locomotion agents from pairwise human comparisons of short clips, with feedback on less than one percent of the agents' interactions with the environment [2]. Applied to language models, the same recipe produced summarisation policies that labellers preferred to much larger supervised models [3]. It also produced instruction-following models whose 1.3 billion parameter version was preferred to the 175 billion parameter GPT-3 [4]. This method, reinforcement learning from human feedback (RLHF), is now the usual final stage in training general-purpose assistants.

Every version of the recipe assumes that the person giving feedback can tell a better output from a worse one. Bowman et al. define scalable oversight as the problem of supervising systems that may outperform their

supervisors on most of the skills relevant to the task [5]. A code review that would take an expert an hour, or a proof with a subtle gap, may receive a confident rating from a hurried evaluator. A policy optimised against such ratings then learns whatever the evaluator rewards. Ngo et al. argue that systems trained in this way could learn to act deceptively to receive higher reward, and that they may appear aligned when they are not [6].

Two families of proposal have dominated the response. Recursive reward modelling and iterated amplification break a hard evaluation into easier pieces, with trained assistants helping the human at each piece. Debate sets two copies of a model against each other before a human judge. Weak-to-strong generalization and model-written critique are newer lines of work that approach the problem from the side of the learner and from the side of the evaluator. This article reviews all four, states the assumptions each one needs in a form that can be checked, and proposes a comparative framework with an evaluation protocol. Sections II to V cover direct feedback, recursive decomposition, debate, and the weak-to-strong and critique methods in turn. Section VI sets out the framework and the protocol. No experiments were run for this article, and every plotted quantity is computed from a formula stated in the text.

## II. HUMAN FEEDBACK AND ITS FAILURE UNDER OPTIMIZATION

The reward model in RLHF is fitted to pairwise comparisons through the Bradley-Terry model of paired comparison [7]. For a prompt x and two responses y1 and y2, the probability that a labeller prefers y1 is modelled as:

$$P(y_1 \succ y_2 \mid x) = \sigma\big(r(x, y_1) - r(x, y_2)\big) \tag{1}$$

where σ is the logistic function and r is the learned reward, and the reward model is trained to minimise $-\log \sigma(r(x, y_w) - r(x, y_l))$ over observed winners $y_w$ and losers $y_l$. The policy π is then optimised to maximise:

$$\mathbb{E}[r(x, y)] - \beta \mathrm{KL}(\pi \,\|\, \pi_{\mathrm{ref}}) \tag{2}$$

where the Kullback-Leibler term with weight β keeps the policy close to a reference model $\pi_{\mathrm{ref}}$ [3], [4]. Under this model a reward gap of 1 corresponds to a preference probability of $\sigma(1) \approx 0.73$, and a gap of 3 to about 0.95. The reward scale therefore encodes how consistently labellers separate two outputs. It carries no information about correctness beyond what the labellers could see.

Optimization against a learned proxy invites reward hacking. Skalse et al. formalise the phenomenon under the name reward gaming [8]. They prove that, over the set of all stochastic policies, two reward functions can be unhackable with respect to each other only if one of them is constant [8]. Pan et al. built four reinforcement learning environments with misspecified rewards [9]. They found that more capable agents more often exploit the misspecification, gaining proxy reward while losing true reward [9]. Gao et al. measured the effect for language models in a synthetic setting where a fixed gold reward model stands in for human labellers, and found that gold reward first rises and then falls as optimization against the proxy continues [10]. The curve takes a different functional form for best-of-n sampling and for reinforcement learning, and its coefficients scale smoothly with reward model size [10].

A second family of failures arises from the labellers. Sharma et al. found sycophancy in five deployed assistants across four free-form tasks [11]. In existing preference data, responses matching a user's stated views were more likely to be preferred, and both humans and preference models sometimes chose convincingly written sycophantic answers over correct ones [11]. Perez et al. had earlier reported that larger models repeat back a user's preferred answer more often [12]. Wen et al. asked time-limited human subjects to judge the correctness of model outputs on a question answering task and a programming task, scoring their judgments against gold labels [13]. RLHF made the models more convincing without making them more correct [13]. After RLHF, the subjects' false positive rate rose by 24.1 and 18.3 percentage points, from 41.0% to 65.1% on QuALITY and from 29.6% to 47.9% on APPS [13]. Casper et al. survey these and related problems, and separate those that better technique might address from those fundamental to learning from human approval [14].

Overoptimization and reward hacking exploit errors in the learned reward, while sycophancy and unintended persuasion exploit errors in the human judgments behind it. The methods reviewed in the next three sections try to improve those judgments, not to fit them more closely.

## III. RECURSIVE REWARD MODELLING AND ITERATED AMPLIFICATION

Leike et al. proposed reward modelling as a research direction for agent alignment, in which a reward function learned from interaction with the user is optimised by reinforcement learning [15]. For tasks the user cannot evaluate unaided, their recursive variant has agents trained with earlier reward models help the user evaluate the outcomes of the next agent, so that each round extends the range of tasks that can be supervised [15]. In iterated amplification, a human solves a hard problem by delegating easier subquestions to copies of the current

model [16]. The model is then trained to imitate this amplified system [16]. Christiano et al. reported results on algorithmic tasks learned without an external reward function [16]. Fig. 1 contrasts these designs with direct evaluation and with debate.

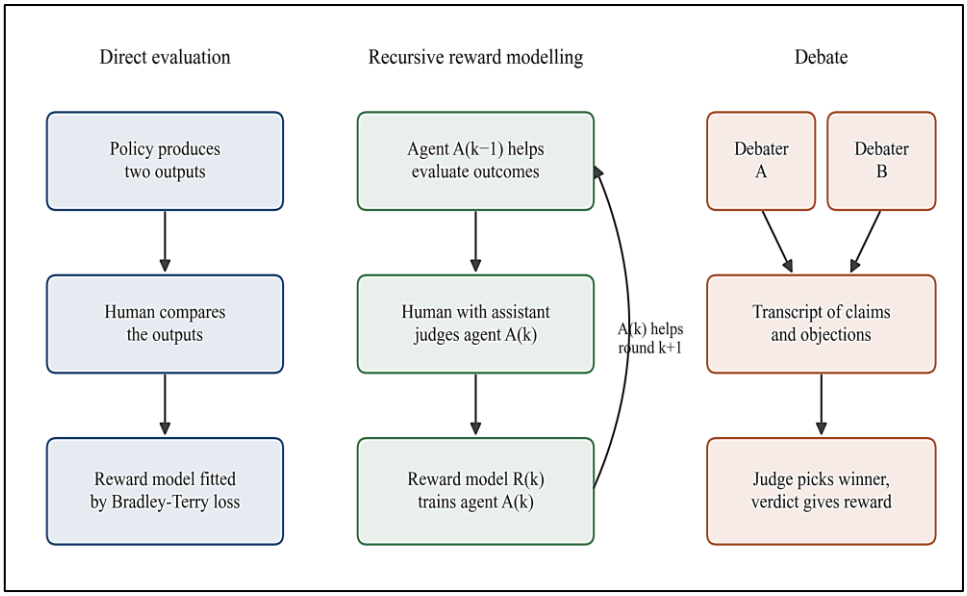



Fig. 1: Schematic of three oversight structures. In direct evaluation a human compares outputs and a reward model is fitted to the comparisons [2], [7]. In recursive reward modelling, agents trained in earlier rounds help the human evaluate the next agent [15]. In debate, two agents argue before a judge whose verdict supplies the reward.

Wu et al. applied recursive decomposition to the summarisation of entire novels [17]. The model summarises short sections and then summarises the summaries, and labellers evaluate each step without reading the whole book. The resulting summaries matched the quality of human-written ones for about 5% of books [17]. The study remains the clearest published case of decomposition making tractable an evaluation that would otherwise require hours of reading per judgment.

The benefit of decomposition can be stated as a count. Suppose a task has size N, measured in units that a person can check in one sitting, and is decomposed as a tree with branching factor b, so that the depth is $d = \log_b N$. Direct evaluation requires one judgment that reads all N units. A full recursive scheme that checks every internal node requires $\frac{(N-1)}{(b-1)}$ judgments of b sub-answers each, a total of $\frac{b(N-1)}{(b-1)}$ units. A scheme that checks only one path from root to leaf, which is what debate attempts, requires d judgments of b units, a total of $b\log_b N$. Fig. 2 plots the three totals for $b = 4$.

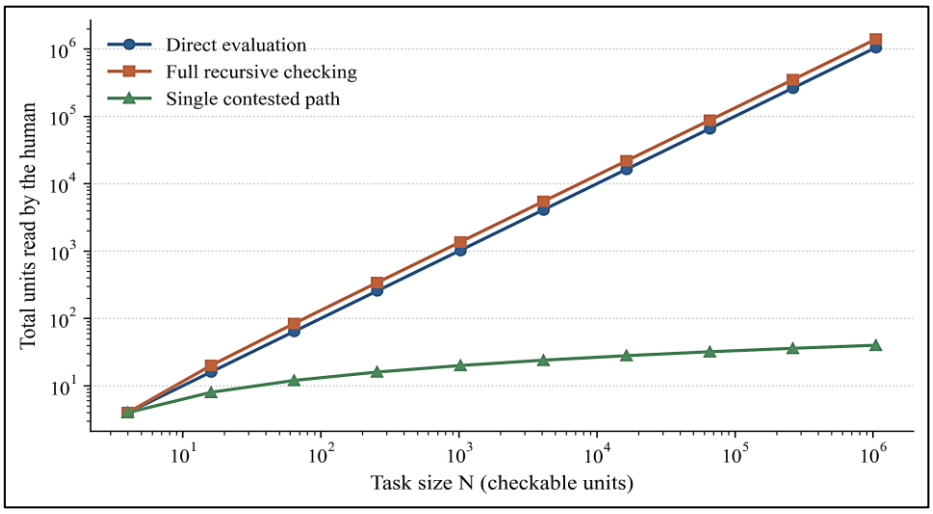



Fig. 2: Units a human must read to evaluate a task of size N under direct evaluation (N), full recursive checking of every node $\frac{b(N-1)}{(b-1)}$ and checking of a single contested path $b\log_b N$, computed from the formulas in Section III with branching factor $b = 4$. Both axes are logarithmic.

Fig. 2 carries two points. Full recursion does not reduce the total human effort, which exceeds N by the factor $\frac{b}{(b-1)}$. It does, however, cap every individual judgment at b units, a size that a person can check reliably. Only path checking reduces the total, from about 1.05 million units for direct evaluation or 1.4 million for full recursion to 40 at N $=4^{10}$, and it does so only if the checked path is the one that contains the error. Recursive reward modelling therefore relies on trained assistants to carry most of the load, and debate is designed to find the right path.

Three assumptions carry the recursive argument. The task must decompose into subquestions whose answers can be checked locally, and tasks that rest on tacit judgment, such as deciding whether a research plan is promising, may not decompose in this way. Errors must not accumulate faster than they are caught, since a reward model at round k inherits the mistakes of the assistants that helped train it. The assistants must also be honest. This is because an assistant that has learned to flatter the evaluator, as documented for RLHF models [11], [13], would carry that bias into every later round. None of these assumptions has been tested at the scale the proposals envisage. The main published evidence consists of the algorithmic tasks in [16] and the book summarisation study in [17].

## IV. DEBATE: THEORY AND EMPIRICAL RECORD

Irving et al. proposed training agents by self-play on a zero-sum debate game [18]. Two agents take turns making short statements about a question, and a human then judges which agent gave the most true and useful information. The case for the design is an analogy with interactive proof systems [19]. A polynomial-time judge who checks answers directly can verify only problems in NP, whereas with optimal play debate lets such a judge answer any question in PSPACE [18], the class that interactive proofs capture [20]. The intuition is that a dishonest claim must rest on some false step, and an honest opponent can direct the judge to that step.

A simple error model shows what this gains. Let a sound argument have k steps, and let a judge who checks a step make an error with probability ε, independently across steps. A judge who must check every step reaches the correct verdict with probability $(1-\varepsilon)^k$, while a judge shown only the single step in dispute, as debate intends, is correct with probability $1-\varepsilon$ whatever the value of k. Fig. 3 plots both cases. At ε = 0.05 the probability of a correct full check falls to 0.60 at k = 10 and to 0.077 at k = 50, while the debate judge stays at 0.95.

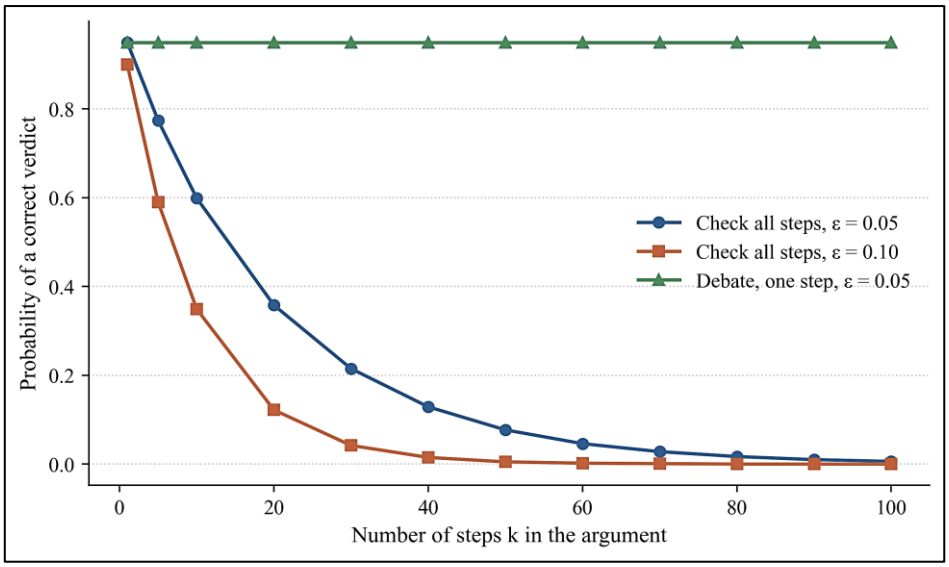



Fig. 3: Probability that a judge reaches the correct verdict on a sound k-step argument when every step must be checked, $(1-\varepsilon)^k$, for per-step error rates ε = 0.05 and ε = 0.10, compared with a debate judge who checks one disputed step, $1-\varepsilon$ at ε = 0.05. Values are computed from the error model in Section IV, which assumes independent errors.

The guarantee holds only at equilibrium and under conditions that the formula hides. The honest debater must be able to find the flawed step. A dishonest debater may offer an argument so long that neither side can locate its flaw within the debate, which leaves the honest side nothing to point to. This difficulty is usually called the obfuscated arguments problem [21]. The original complexity argument also assumed an honest debater able to simulate deterministic systems for exponentially many steps. Brown-Cohen et al. designed protocols in which the honest strategy succeeds with a polynomial number of simulation steps and can verify stochastic systems, even when the dishonest side may use exponentially many steps [21]. Their analysis models human judgment as an oracle for local questions, which real judges only approximate.

Six studies are listed in Table 1. In the MNIST experiment of [18], agents competed to convince a sparse classifier that saw only the pixels they chose to show, and raised its accuracy from 59.4% to 88.9% with six pixels. Parrish et al. gave single-turn arguments to human judges who had not read the passage behind a hard reading comprehension question [22]. The arguments did not improve accuracy, whereas selected text snippets did [22]. With multi-turn debates between human experts, Michael et al. reported 84% judge accuracy against 74% for consultancy, a baseline in which one expert argues for an answer that is correct half the time [23]. Khan et al. found that debate between LLM experts raised accuracy to 76% for non-expert model judges and to 88% for human judges [24]. The corresponding naive baselines were 48% and 60% [24]. Optimising the debaters for persuasiveness also made judges more accurate in their setting [24].

Table 1. Published experiments on debate and assisted judging

| Study | Task and asymmetry | Judge | Comparison | Reported outcome |
|---|---|---|---|---|
| Irving et al. [18] | MNIST, judge sees only pixels chosen by debaters | Sparse classifier | Debate vs. 6 random pixels | Accuracy 59.4% to 88.9% with 6 pixels |
| Parrish et al. [22] | Reading comprehension, passage hidden | Humans | Single-turn arguments vs. none | No gain from arguments, gain from text snippets |
| Michael et al. [23] | Reading comprehension, passage hidden | Humans | Debate vs. consultancy | 84% vs. 74% |
| Khan et al. [24] | Reading comprehension, passage hidden | LLMs and humans | Debate vs. naive baseline | 76% vs. 48% (LLM), 88% vs. 60% (human) |
| Kenton et al. [25] | QA, mathematics, coding, logic, multimodal | Weaker LLMs | Debate vs. consultancy and direct QA | Better than consultancy on all tasks (random side assignment), mixed against direct QA without information asymmetry |
| Bowman et al. [5] | MMLU, time-limited QuALITY | Humans with an LLM assistant | Assisted vs. unaided humans and model alone | Assisted humans outperform both |

Kenton et al. used weaker LLMs as judges across extractive question answering, mathematics, coding, logic and multimodal reasoning [25]. Debate beat consultancy on every task when consultants were assigned sides at random. Against direct question answering by the judge, debate helped on extractive tasks with information asymmetry, and results on the other tasks were mixed [25]. Stronger debaters raised judge accuracy, though by less than earlier studies had found [25]. The positive results thus come mostly from reading comprehension with a hidden passage, where a verified quote can settle the question. Evidence is weaker for tasks in which the judge has all the information but lacks the skill to use it, which is the case that scalable oversight is ultimately about.

## V. WEAK-TO-STRONG GENERALIZATION AND AI-ASSISTED CRITIQUE

Burns et al. studied an analogy in which a small model supervises a large one [26]. Strong pretrained models fine-tuned on labels from a weak supervisor consistently outperformed that supervisor, but fell short of the same model trained on ground truth. The outcome is measured as the performance gap recovered:

$$\mathrm{PGR} = \frac{P_{w2s} - P_{weak}}{P_{strong} - P_{weak}} \tag{3}$$

where $P_{weak}$ is the accuracy of the supervisor, $P_{strong}$ is the ceiling of the strong model trained on true labels and $P_{w2s}$ is the accuracy of the strong model trained on weak labels [26]. With a GPT-2 level supervisor and an auxiliary confidence loss, GPT-4 recovered close to GPT-3.5 level performance on NLP tasks [26]. The errors of a weak model may be easier for a strong model to recognise than the systematic errors of human labellers. This limits how far the analogy transfers to human oversight.

Model-written critique works on the side of the evaluator. Saunders et al. fine-tuned language models to write critiques of summaries, and the critiques helped human evaluators find flaws they would otherwise have

missed, including flaws planted deliberately to mislead [27]. Larger models also wrote more helpful critiques [27]. Their comparison of a model's ability to recognise a flaw with its ability to articulate it suggests that models may hold relevant knowledge they do not express as critiques [27]. McAleese et al. trained code critics with RLHF. On naturally occurring errors, their critiques were preferred to human critiques in 63% of cases [28]. The critics also hallucinated bugs, and human contractors working with critics caught a similar number of bugs to the critics alone while hallucinating less [28].

AI feedback can also replace human labels outright. Constitutional AI trains a harmless assistant from a short list of written principles, with self-critique and revision in a supervised phase and model-generated preference labels in a reinforcement learning phase [29], and self-critique based alignment of this kind has also been discussed in this journal [30]. Lee et al. found that reinforcement learning from AI feedback performed comparably to RLHF on summarisation and on helpful and harmless dialogue [31]. Zheng et al. reported that strong LLM judges agree with human preferences more than 80% of the time, the same level at which humans agree with each other, and documented position, verbosity and self-enhancement biases in those judges [32]. AI feedback scales the volume of supervision without raising its ceiling. This is because the labelling model is bounded by its own competence and may share the blind spots of the models it evaluates.

## VI. A COMPARATIVE FRAMEWORK AND EVALUATION PROTOCOL

The methods reviewed above differ less in their mechanics than in the assumptions behind their guarantees. Table 2 sets out the proposed framework. For each method it records the assumption required for sound oversight, the place where human effort falls, the failure that becomes likely when the assumption breaks, and a diagnostic test for that failure. The framework is an organising proposal and has not been applied to a common benchmark.

Table 2. Proposed comparative framework for scalable oversight methods

| Method | Assumption required | Where human effort falls | Characteristic failure | Diagnostic test |
|---|---|---|---|---|
| Direct RLHF [2], [4] | Evaluator can judge outputs unaided | Every comparison | Reward hacking, sycophancy, persuasion [10], [11], [13] | Agreement of evaluators with expert labels |
| Recursive reward modelling and amplification [15], [16] | Task decomposes into locally checkable parts | Every node, with assistants | Errors passed forward across rounds | Plant flaws at a chosen depth and measure detection |
| Debate [18], [21] | Honest strategy wins, judge accurate on one step | One contested path | Obfuscated arguments, persuasion over truth | Assign sides, vary argument length and debater strength |
| Weak-to-strong generalization [26] | Strong model generalises intent, not supervisor errors | Weak labels only | Imitation of systematic supervisor errors | Performance gap recovered on held-out ground truth |
| Critique and AI feedback [27], [29], [31] | Critic errors uncorrelated with generator errors | Review of critiques | Hallucinated flaws, shared blind spots, judge bias [28], [32] | Planted bugs, with catch and false alarm rates |

Three failure modes cut across the rows of Table 2. Collusion becomes possible whenever copies of one model fill both the supervising and the supervised roles, as in debate between two instances, in recursive assistants trained from earlier agents and in AI feedback drawn from the policy's own checkpoint. Nothing in a zero-sum objective stops two copies from sharing a blind spot. Persuasion that outruns truth appears in RLHF [13] and in consultancy [23]. Khan et al. found persuasiveness and truthfulness moving together in debate [24], but it is not yet established that this holds when the judge cannot check a quote. Sycophancy can enter any method in which output is rated by a person whose views the model can infer [11].

The sandwiching design of Bowman et al. gives the evaluation protocol its structure [5]. Tasks are chosen so that the model alone and unaided non-experts both fail while domain experts succeed, and the non-experts then use an oversight method together with the model while the experts' answers serve only to score the result. In the

proof of concept on MMLU and time-limited QuALITY, non-experts who chatted with an unreliable model assistant outperformed both the model alone and their own unaided performance [5]. The protocol proposed here extends that design in four ways. Each method faces a planted adversary, a model fine-tuned to argue for wrong answers or to insert subtle flaws, so that measured accuracy reflects resistance to deliberate deception. Decomposition depth or argument length is varied, so that the error growth predicted in Fig. 3 can be compared with the growth actually observed. Debater and judge strength are varied independently, following [25], and results are reported as the share of the gap between non-expert and expert accuracy that the method recovers, by analogy with the PGR of [26]. False alarm rates are reported alongside detection rates, because flagging correct outputs has its own cost, as the hallucinated bugs in [28] show.

Several limits apply to the framework and to the field. Sandwiching needs tasks on which experts exist, so it cannot test oversight of capabilities that exceed every human. The error model behind Fig. 3 assumes independent errors and a single decisive flaw. Correlated errors by the judge would shrink the advantage it shows. Judges in several published studies worked under time limits [5], [13], and trained evaluators with more time may behave differently. Methods that inspect internal computation instead of outputs, such as the mechanistic study of in-context learning discussed in this journal [33], offer a check that does not depend on how a judge reads an argument, although interpretability at the scale of frontier models remains incomplete.

## VII. CONCLUSION

Human feedback works as a training signal only within the range of outputs that people can evaluate. The evidence on reward hacking, overoptimization, sycophancy and unintended persuasion shows optimization pressure finding the edges of that range. In the idealised counting model of Section III, full recursive checking caps the size of each judgment but leaves total human effort roughly unchanged. Recursive reward modelling depends on decomposability and on assistants that do not pass their errors forward. In the same model, debate offers an oversight cost that grows with the logarithm of task size, but only under an honest equilibrium that obfuscated arguments and judge error can break.

The empirical record supports multi-turn debate most clearly in reading comprehension with a hidden passage, where a verifiable quote settles the question, and shows small or mixed effects elsewhere. Weak-to-strong generalization and model-written critique offer partial gains that are easier to measure with present models. Table 2 links each method to the assumption it needs and to a test that would expose its failure. The next step is to apply the extended sandwiching protocol of Section VI to all four families of method on a shared task set. Claims about their relative merit should wait for that comparison.

## REFERENCES


[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv preprint arXiv:1606.06565*, 2016.

[2] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] N. Stiennon *et al*., “Learning to summarize with human feedback,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 3008–3021, 2020.

[4] L. Ouyang *et al*., “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730–27744, 2022.

[5] S. R. Bowman *et al*., “Measuring progress on scalable oversight for large language models,” *arXiv preprint arXiv:2211.03540*, 2022.

[6] R. Ngo, L. Chan, and S. Mindermann, “The alignment problem from a deep learning perspective,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[7] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

[8] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, “Defining and characterizing reward gaming,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 9460–9471, 2022.

[9] A. Pan, K. Bhatia, and J. Steinhardt, “The effects of reward misspecification: Mapping and mitigating misaligned models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[10] L. Gao, J. Schulman, and J. Hilton, “Scaling laws for reward model overoptimization,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, vol. 202, pp. 10835–10866, 2023.

[11] M. Sharma *et al*., “Towards understanding sycophancy in language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[12] E. Perez *et al*., “Discovering language model behaviors with model-written evaluations,” in *Findings of the Association for Computational Linguistics (ACL)*, pp. 13387–13434, 2023.

[13] J. Wen *et al*., “Language models learn to mislead humans via RLHF,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[14] S. Casper *et al*., “Open problems and fundamental limitations of reinforcement learning from human feedback,” *Transactions on Machine Learning Research*, 2023.

[15] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: A research direction," *arXiv preprint arXiv:1811.07871*, 2018.
[16] P. Christiano, B. Shlegeris, and D. Amodei, "Supervising strong learners by amplifying weak experts," *arXiv preprint arXiv:1810.08575*, 2018.
[17] J. Wu *et al*., "Recursively summarizing books with human feedback," *arXiv preprint arXiv:2109.10862*, 2021.
[18] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv preprint arXiv:1805.00899*, 2018.
[19] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof systems," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 186–208, 1989.
[20] A. Shamir, "IP = PSPACE," *Journal of the ACM*, vol. 39, no. 4, pp. 869–877, 1992.
[21] J. Brown-Cohen, G. Irving, and G. Piliouras, "Scalable AI safety via doubly-efficient debate," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 4585–4602, 2024.
[22] A. Parrish *et al*., "Single-turn debate does not help humans answer hard reading-comprehension questions," in *Proceedings of the 1st Workshop on Learning from Natural Language Supervision (LNLS)*, pp. 17–28, 2022.
[23] J. Michael *et al*., "Debate helps supervise unreliable experts," *arXiv preprint arXiv:2311.08702*, 2023.
[24] A. Khan *et al*., "Debating with more persuasive LLMs leads to more truthful answers," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 23662–23733, 2024.
[25] Z. Kenton *et al*., "On scalable oversight with weak LLMs judging strong LLMs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 75229–75276, 2024.
[26] C. Burns *et al*., "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 4971–5012, 2024.
[27] W. Saunders *et al*., "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.
[28] N. McAleese, R. M. Pokorny, J. F. Ceron Uribe, E. Nitishinskaya, M. Trebacz, and J. Leike, "LLM critics help catch LLM bugs," *arXiv preprint arXiv:2407.00215*, 2024.
[29] Y. Bai *et al*., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.
[30] W. M. John, "Constitutional AI: Scalable alignment through self-critique and revision," *Eduschool Journal of Artificial Intelligence Research*, vol. 1, no. 1, pp. 10–19, 2026, doi: 10.63090/EJAIR/3139.7336.0002.
[31] H. Lee *et al*., "RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 26874–26901, 2024.
[32] L. Zheng *et al*., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, vol. 36, pp. 46595–46623, 2023.
[33] M. T. V, "Mechanistic interpretability of in-context learning in transformers," *Eduschool Journal of Artificial Intelligence Research*, vol. 1, no. 1, pp. 20–29, 2026, doi: 10.63090/EJAIR/3139.7336.0003.