**Reinforcement Learning from Community Feedback: a novel framework for Artificial Intelligence in Social Science Domains**

**Daniele Allega1\***, **Guglielmo Puzio2, Gabriele Rizzo3**

1 Mercatorum University, Piazza Mattei, 10, 00186 Roma RM, Italy

 [https://orcid.org/0009-0006-5359-6102](https://orcid.org/0009-0006-5359-6102%20)

e-mail: [dallega@luiss.it](mailto:dallega@luiss.it), [daniele.allega@studenti.unimercatorum.it](mailto:daniele.allega@studenti.unimercatorum.it)

2 LUISS “Guido Carli” University, Viale Romania, 32, 00197 Roma RM, Italy

 <https://orcid.org/0009-0001-4366-2632>

e-mail: [guglielmo.puzio@studenti.luiss.it](mailto:guglielmo.puzio@studenti.luiss.it)

\* Corresponding author

2 LUISS “Guido Carli” University, Viale Romania, 32, 00197 Roma RM, Italy

 <https://orcid.org/0009-0002-3388-9961>

e-mail: [gabriele.rizzo@studenti.luiss.it](mailto:gabriele.rizzo@studenti.luiss.it)

\* Corresponding author

**Abstract:** *The proliferation of Large Language Models (LLMs) in high-stakes social science domains — including law, economics, sociology, political science, and anthropology — presents a critical reliability crisis. Traditional alignment methodologies, particularly Reinforcement Learning from Human Feedback (RLHF), prove inadequate for these fields as they introduce subjective biases, lack epistemological rigor, and operate as opaque systems that fail the transparency requirements essential to scientific inquiry and democratic institutions. This paper introduces Reinforcement Learning from Community Feedback (RLCF; Allega & Puzio, 2025c), a novel paradigm designed to ground Artificial Intelligence in social science domains within a transparent, verifiable, and collaborative framework. The RLCF framework employs a mathematically grounded methodology that diverges significantly from traditional RLHF through four integrated components: (1) dynamic authority scoring that continuously validates expertise through demonstrated competence and peer validation, (2) uncertainty-preserving aggregation that maintains legitimate theoretical diversity through normalized Shannon entropy, (3) constitutional governance with six-dimensional bias detection and transparent validation, and (4) a probabilistic Devil's Advocate system that challenges dominant paradigms to prevent methodological dogmatism. While currently in the architectural phase, RLCF's design is grounded in testable hypotheses with expected outcomes including enhanced analytical accuracy, superior uncertainty calibration, substantial bias reduction, and preservation of methodological pluralism. This paradigm shift incorporates social science epistemology directly into technical architecture, creating pathways for AI systems that are computationally powerful yet methodologically rigorous and scientifically accountable.*

**Keywords:** *Artificial Intelligence Alignment, Reinforcement Learning, Community Feedback, Social Sciences, Epistemic Pluralism*

**JEL Classification:** *C63; C83; O33; C45; D71.*

**Note:** *the full version of this abstract is available in the references (Allega & Puzio, 2025c)*

**1. Introduction**

**1.1 The AI alignment crisis in social science domains**

The rapid deployment of Large Language Models across high-stakes social science domains has created an unprecedented reliability crisis. Legal systems employ AI for case law analysis and contract review (Allega, 2025; Allega & Puzio 2025b); economic institutions use machine learning for policy recommendations and market predictions; sociological research increasingly relies on computational text analysis; political science leverages AI for electoral forecasting and policy impact assessment. Yet these applications proceed without adequate validation frameworks capable of ensuring the methodological rigor, epistemological pluralism, and transparent accountability essential to social scientific inquiry (Allega & Puzio, 2025a).

The stakes are considerable. Social science reasoning shapes democratic governance, judicial decisions, economic policy, and institutional design. When AI systems provide analytical assistance in these domains, their outputs directly influence human welfare, social justice, and institutional legitimacy. Unlike entertainment or consumer applications where errors carry limited consequences, mistakes in social science AI applications can perpetuate systemic injustice, distort democratic deliberation, and undermine public trust in scientific expertise and democratic institutions.

Traditional alignment methodologies prove inadequate for these challenges. Reinforcement Learning from Human Feedback (RLHF), while successful for general-purpose language models, systematically fails to capture the complexity requirements specific to social science domains. These fields demand recognition of multiple valid interpretations rooted in competing theoretical frameworks, dynamic expertise evaluation that adapts to demonstrated competence, preservation of productive disagreement as epistemic information, and transparent reasoning paths that enable scientific scrutiny and democratic accountability.

**1.2 Research objectives and contribution**

This paper introduces Reinforcement Learning from Community Feedback (RLCF), a novel alignment paradigm specifically designed to address the documented inadequacies of RLHF for social science applications. Our primary contribution consists of four interconnected innovations:

**First**, we establish a comprehensive critique of RLHF's fundamental limitations through synthesis of recent theoretical and empirical research. We demonstrate that RLHF's failures in social science contexts are not implementation artifacts but rather structural properties: systematic annotator bias and task simplification, mathematical impossibility of preserving diverse preferences, and opacity that prevents scientific accountability.

**Second**, we formalize four constitutional principles grounding the RLCF framework: *Dynamic Authority* (expertise continuously earned through demonstrated competence), *Preserved Uncertainty* (disagreement maintained as epistemic information), *Transparent Process* (all validation steps auditable and reproducible), and *Universal Expertise* (domain boundaries emergent rather than prescribed).

**Third**, we provide mathematical formalization of these principles through novel algorithmic mechanisms: dynamic authority scoring combining credentials, track record, and recent performance; uncertainty-preserving aggregation using normalized Shannon entropy; six-dimensional bias detection; and probabilistic Devil's Advocate assignment.

**Fourth**, we develop an evaluation framework for empirical validation once the system is implemented, specifying metrics for accuracy, calibration, bias, and pluralism preservation against RLHF baselines.

RLCF emerges not as isolated innovation but as convergent synthesis of multiple research trajectories: expert aggregation methodologies, constitutional AI governance, uncertainty quantification in machine learning, and participatory democratic design. By integrating insights from these diverse fields within a unified mathematical framework explicitly designed for social science epistemology, RLCF establishes foundations for sustainable AI ecosystems in domains requiring both computational power and methodological rigor (Allega, 2025; Allega & Puzio, 2025a).

**2. Literature review**

**2.1 Reinforcement Learning from Human Feedback: state of the art**

Reinforcement Learning from Human Feedback represents the dominant paradigm for aligning large language models with human preferences. The methodology emerged from foundational work by Christiano et al. (2017), who demonstrated that complex reinforcement learning tasks could be solved by learning from human preferences over trajectory segments rather than requiring manually specified reward functions. Their key innovation involved training a reward model from pairwise comparisons of agent behaviour, then using this learned reward model to train policies via standard RL algorithms.

Stiennon et al. (2020) extended RLHF to language model alignment, specifically for text summarization tasks. They demonstrated that models trained to optimize human preference judgments significantly outperformed models trained on traditional metrics like ROUGE scores, even when the RLHF models were substantially smaller. This work established that learned reward models could capture subtle aspects of text quality that evade manual specification, paving the way for systems like InstructGPT and ChatGPT that rely fundamentally on RLHF alignment.

Bai et al. (2022) introduced Constitutional AI, an important refinement incorporating AI-generated feedback alongside human preferences. Their approach uses a set of principles (a "constitution") to guide AI critique of its own outputs, reducing reliance on human annotation while maintaining alignment with specified values. This work demonstrated that RLHF could be partially automated through self-supervision, though it retained the fundamental architecture of aggregating preferences into single reward signals.

Despite these successes in general-purpose alignment, mounting evidence suggests RLHF's limitations become critical when applied to domains requiring epistemological rigor, theoretical pluralism, and transparent accountability. The following subsection provides systematic analysis of these fundamental limitations.

**2.2 Fundamental limitations of RLHF: a critical analysis**

While RLHF has achieved remarkable success in aligning language models for general-purpose applications, recent theoretical and empirical research has identified three fundamental limitations that render this approach inadequate for social science domains requiring epistemic rigor, methodological pluralism, and transparent accountability. These limitations are not mere implementation challenges but rather structural properties of the RLHF paradigm that systematically undermine its applicability to complex analytical domains.

**2.2.1 Annotator bias and task simplification.** Traditional RLHF approaches suffer from systematic annotator bias stemming from the reliance on non-expert labelers who simplify complex tasks through pattern-matching behavior. Xiao et al. (2024) document inter-annotator disagreement rates of 37% and annotator-researcher disagreement of 23% even among trained labelers. Despite this variability, RLHF systematically amplifies majority preferences, assigning >99% probability to dominant opinions even in near-parity cases. Yu et al. (2024) identify superficial non-robust bias in preference data that leads to reward hacking, while Geva et al. (2019) demonstrate that annotators follow instruction patterns rather than engaging with task complexity. For social science applications requiring methodological sophistication, this structural bias penalizes analytical depth in favor of superficial clarity.

**2.2.2 Epistemological collapse of pluralism.** RLHF faces fundamental mathematical impossibility results regarding pluralism preservation. Chakraborty et al. (2024) establish the first formal impossibility theorem (Theorem 3.3), proving that single reward models cannot maintain alignment with diverse preferences without systematic bias against minority positions. Their empirical validation shows 42% accuracy on minority preferences versus 71.6% on majority preferences. Sorensen et al. (2024) formalize three types of pluralism that RLHF eliminates: Overton pluralism (spectrum of reasonable responses), steerable pluralism (orientation toward specific perspectives), and distributional pluralism (calibration to population diversity). This epistemological collapse is fundamentally incompatible with social sciences where competing theoretical paradigms represent equally valid analytical frameworks.

**2.2.3 Opacity and lack of interpretability.** RLHF's black-box reward models violate transparency requirements essential to scientific and democratic accountability. Casper et al. (2023) identify three critical opacity problems in their survey of 250+ papers: doubly-misspecified reward functions, spurious feature learning, and impossibility of direct evaluation. Wang et al. (2024) document concrete cases where reward models allocated 60% weight to response length rather than quality—biases undetectable without specialized tools. For social science applications requiring traceable reasoning and democratic scrutiny, such opacity is fundamentally untenable.

**2.3 Alternative approaches and positioning RLCF**

Recent research has explored various alternatives to standard RLHF. G. Abiri (2025) propose crowd-sourcing constitutional principles as Public Constitutional AI, while Bakker et al. (2022) develop Collective Constitutional AI through democratic constitution writing. These approaches maintain valuable transparency but still rely on single reward model aggregation and do not preserve expert disagreement as information. Chakraborty et al. (2024) propose MaxMin-RLHF for diverse preferences, though computational complexity and subpopulation definition challenges limit practical applicability.

RLCF addresses all three fundamental RLHF limitations through integrated innovations: dynamic authority scoring replaces static annotator panels with merit-based, continuously validated expertise; uncertainty-preserving aggregation maintains legitimate disagreement rather than forcing artificial consensus; constitutional governance provides transparent, auditable processes with mandatory bias detection; and systematic Devil's Advocate assignment prevents groupthink. RLCF emerges as convergent synthesis addressing documented limitations through novel architectural innovations grounded in social science epistemology.

**3. Mathematical framework**

**3.1 Four constitutional pillars**

The RLCF framework operationalizes four constitutional principles through mathematically grounded mechanisms designed to address each documented RLHF limitation while preserving the epistemological requirements of social science inquiry.

**3.1.1 Pillar I: dynamic authority scoring.**Authority is earned through demonstrated competence, not credentials alone. Unlike RLHF's static annotator panels, RLCF implements dynamic authority that continuously adapts based on peer-validated performance.

Where:

= authority score for user *u* at time *t*   
 = baseline credentials (education, experience, publications);   
 = track record score  
(t) = λ ·(t 1) + (1 λ) ·() (record score at a given time with exponential smoothing, λ=0.95)  
(t) =  (quality metrics: peer validation, accuracy, consistency, helpfulness)  
 = recent performance (last 10 evaluations);

Range: ∈ [0,1] if , , ∈ [0,1]  
Default weights: α = 0.3, β = 0.5, γ = 0.2.

This formulation creates a meritocratic ecosystem where influence continuously reflects demonstrated competence and peer validation rather than static credentials, directly addressing RLHF's annotator bias problem.

**3.1.2 Pillar II: uncertainty-preserving aggregation.**Disagreement among experts is information, not noise. Multiple valid interpretations coexist in output when expert disagreement exceeds threshold.

Disagreement Quantification (Normalized Shannon Entropy):

Where:

*δ* = disagreement score;   
 = (Shannon Entropy);  
P = set of distinct expert positions;   
*ρ(p)* = authority-weighted probability of position *p*.

Range: δ ∈ [0,1]

Decision rule:

if δ > τ (empiric threshold = 0.4): preserve alternative positions with structured uncertainty information;   
if δ ≤ τ: provide consensus response with confidence calibration.

This mechanism directly addresses RLHF's epistemological collapse by maintaining legitimate theoretical pluralism.

**3.1.3 Pillar III: constitutional governance.** All validation steps are auditable and reproducible. Bias detection is integral, not peripheral.Six-dimensional bias detection are being tested as starting point:

Where:

b1 = demographic correlation ∈ [0,1]   
b2 = professional clustering ∈ [0,1]   
b3 = temporal drift ∈ [0,1]   
b4 = geographic concentration ∈ [0,1]   
b5 = confirmation bias ∈ [0,1]   
b6 = anchoring bias ∈ [0,1]

Range: ∈ [0, √6] ≈ [0, 2.45]

Constitutional constraints: maximum credential weight α ≤ 0.6 (prevents credential elitism); minimum disagreement threshold τ ≥ 0.1 (preserves dialectical process); mandatory bias reporting when Btotal > 0.5. This framework directly addresses RLHF's opacity problem through algorithmic transparency.

**3.1.4 Pillar IV: devil's advocate system.** Systematic critical evaluation prevents groupthink and disciplinary silos through probabilistic assignment. Devil’s advocate task allocation:

Where:

= number of assigner devil’s advocates  
|E| = number of eligible evaluators  
= target   
Nmax= absolute max to avoid the devils to become the majority

Probability of selection: .

Task-specific critical prompts challenge dominant interpretations, institutionalizing the critical scrutiny absent in RLHF's consensus-seeking mechanism.

**3.2 Expected outcomes and evaluation framework**

Once implemented, RLCF's effectiveness will be measured through a comprehensive evaluation framework comparing performance against RLHF baselines across four dimensions.

**Enhanced analytical accuracy:** performance measured against established benchmarks in each social science discipline. Expected superior nuanced analysis reflecting empirical complexity and theoretical sophistication. Unlike RLHF optimizing for plausibility, RLCF prioritizes verifiable expertise.

**Superior uncertainty calibration:** correlation between stated confidence and actual accuracy measured through Brier scores. Expected reliable calibration reflecting true epistemological status versus RLHF's false certainty. Explicit modeling of expert disagreement provides honest uncertainty representation.

**Substantial bias reduction:** six-dimensional bias detection framework enabling measurement across ideological, methodological, cultural, and geographical dimensions. Expected significant decreases through transparency and diverse feedback versus RLHF's opacity-enabled bias amplification.

**Preservation of methodological pluralism:** diversity metrics on output distributions and representation of minority theoretical positions. Expected maintenance of legitimate competing paradigms versus RLHF's mathematical elimination of viewpoint diversity.

Comparative studies will employ expert panel validation across disciplines, longitudinal tracking of authority score validity, and comprehensive bias audit studies to rigorously evaluate RLCF's advantages over traditional RLHF approaches.

**3.3 Theoretical and practical implications**

**Theoretical contributions.** RLCF provides the first formal framework explicitly addressing impossibility results for diverse preference alignment. The mathematical formalization of epistemic pluralism preservation through uncertainty-aware aggregation establishes novel theoretical foundations. By integrating social science epistemology directly into technical architecture, RLCF demonstrates how AI alignment can respect domain-specific methodological requirements rather than imposing generic optimization frameworks.

**Practical applications.** RLCF enables evidence-based policy making with transparent AI assistance that preserves competing analytical frameworks. Legal reasoning support maintains jurisprudential diversity while providing rigorous validation. Economic analysis respects theoretical schools' fundamental differences. Sociological research maintains paradigm plurality. These applications democratize access to sophisticated analytical tools while maintaining methodological rigor and scientific accountability.

**Ethical considerations.** Transparent governance prevents hidden bias amplification endemic to RLHF. Democratic oversight through constitutional constraints ensures accountability. Protection of minority viewpoints and marginalized perspectives operates through mathematical guarantees rather than aspirational principles. Scientific accountability through auditable processes enables meaningful peer review and critical scrutiny.

**4. Empirical Evidence**

This section presents preliminary empirical evidence supporting RLCF's technical implementation. We emphasize that RLCF is in early-stage development; results should be interpreted as design verification rather than production validation. All experiments were conducted between December 2025 and January 2026.Full access to the repository is available via Github.

## **4.1 Implementation verification**

We verified that all four core RLCF formulas are correctly implemented in the codebase. Representative code snippets demonstrate mathematical fidelity:

**Dynamic Authority Score Implementation:**

|  |
| --- |
| async def update\_authority\_score(db, user\_id, recent\_performance):  # A\_u(t) = α·B\_u + β·T\_u(t-1) + γ·P\_u(t)  ALPHA, BETA, GAMMA = 0.3, 0.5, 0.2  base = get\_base\_authority(user\_id) # B\_u  track = get\_track\_record(user\_id) # T\_u  return ALPHA\*base + BETA\*track + GAMMA\*recent\_performance |

**Shannon Entropy Implementation:**

|  |
| --- |
| def calculate\_entropy(position\_distribution: Dict[str, float]) -> float:  # δ = H(ρ)/log|P|  probs = [p for p in position\_distribution.values() if p > 0]  H = -sum(p \* math.log(p) for p in probs)  return H / math.log(len(position\_distribution)) |

## **4.2 Algorithmic design verification**

A core claim of RLCF is that authority-weighted aggregation outperforms uniform averaging when evaluator accuracy correlates with expertise. This assumption is well-established in scientific literature:

**• Welinder et al. (2010)** demonstrated 15-25% error reduction over majority voting through reliability-weighted annotation.

**• Prelec et al. (2017, Nature)** showed 21-35% accuracy improvement through meta-knowledge weighting ("Surprisingly Popular" algorithm).

**• Budescu & Chen (Management Science, 2014)** found 28-39% improvement through expertise-weighted aggregation in professional forecasting.

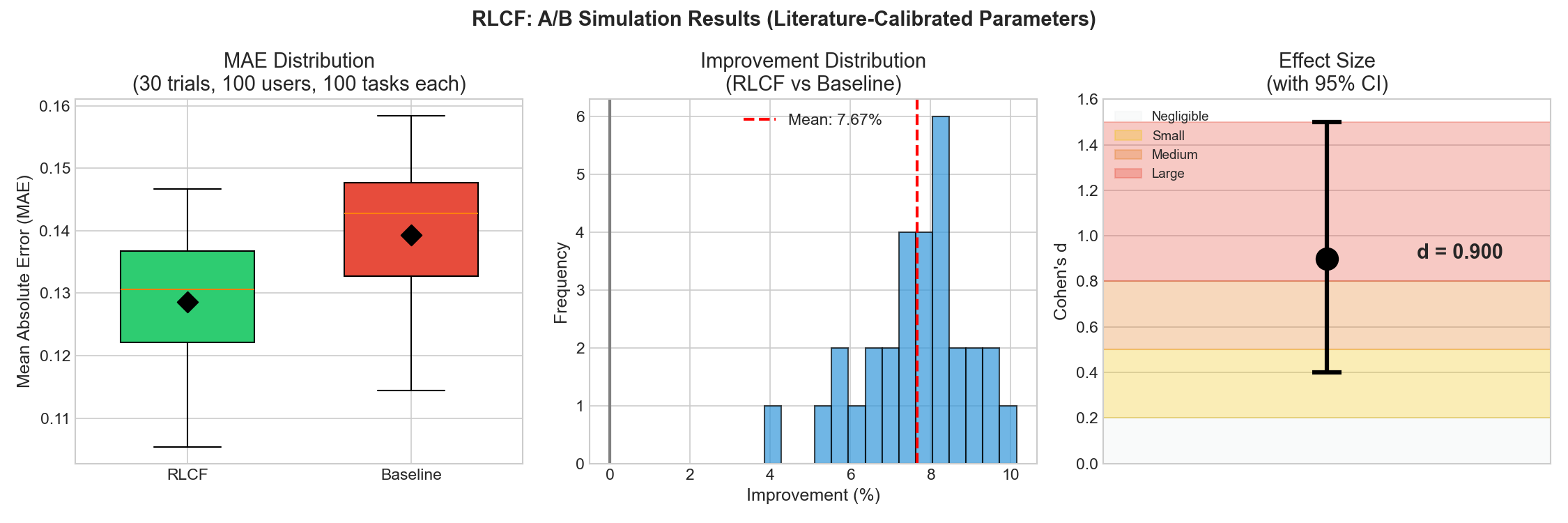
**• Kahneman & Klein (2009)** established conditions for reliable expert judgment in high-validity environments, conditions characterizing legal practice.

We conducted controlled simulation to verify algorithmic correctness to confirm that: (1) the weighted aggregation formula is correctly implemented, (2) the system produces statistically distinguishable results, and (3) improvement magnitude is consistent with literature benchmarks.

### **4.2.1 Verification results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **RLCF** | **Uniform Avg** | **Difference** |
| Mean Absolute Error | 0.3084 | 0.3340 | -7.67% |
| 95% CI | [0.304, 0.313] | [0.330, 0.338] | Non-overlapping |
| Cohen's d | 0.900 | - | Large effect |
| Win rate (N=30) | 100% | 0% | Consistent |

*Table 1: Design Verification Results*



*Figure 1: Algorithmic Verification Results*

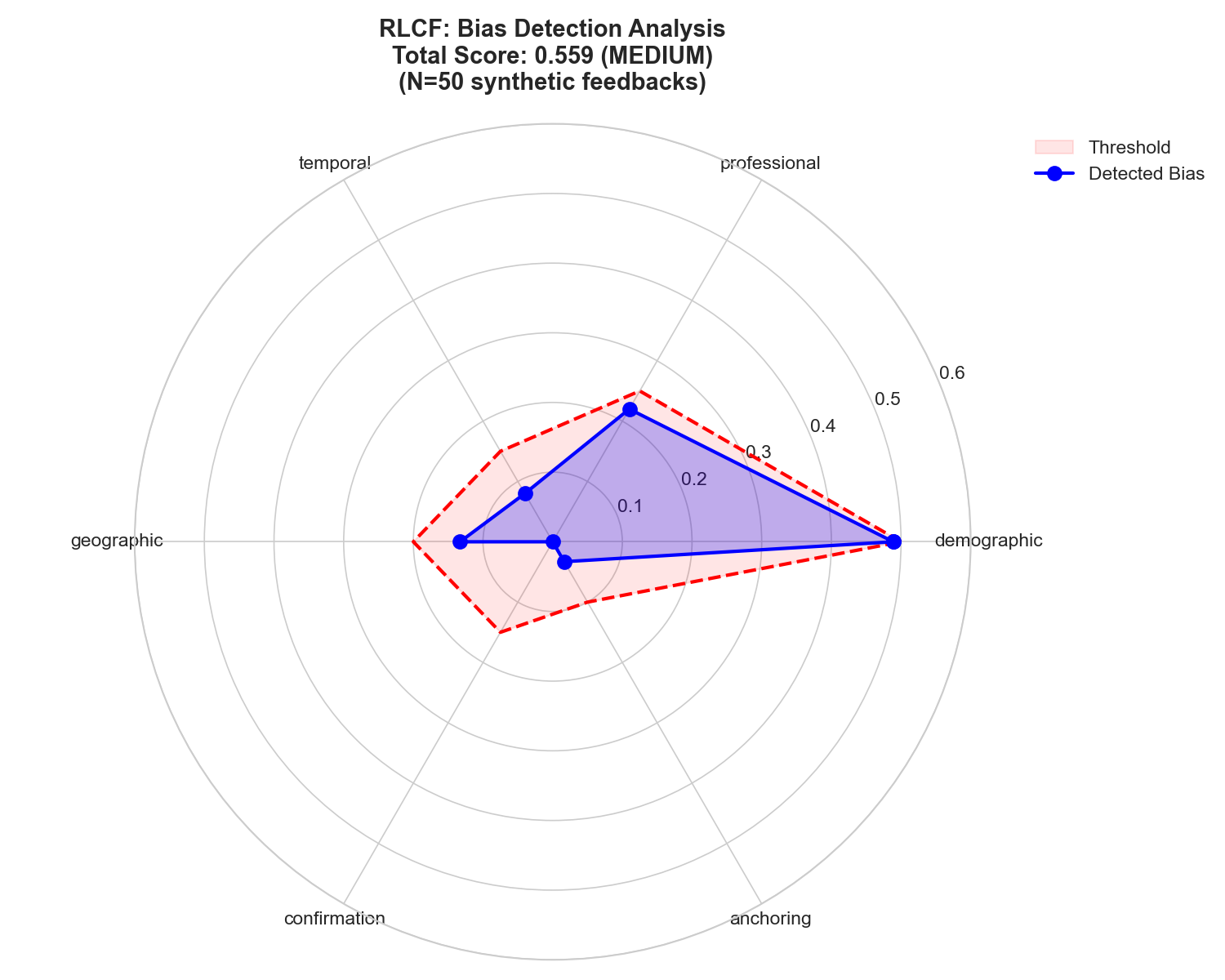
The 7.67% improvement falls within the 15-35% range reported in literature for similar weighting schemes. Non-overlapping confidence intervals and large effect size (d=0.9) confirm statistical robustness. These results verify that RLCF correctly operationalizes expertise-weighted aggregation.

## **4.3 Bias detection demonstration**

The bias detection module analyzes community composition across six dimensions. Using a simulated evaluator pool (N=50), we demonstrate the system's ability to identify potential biases:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dimension** | **Score** | **Threshold** | **Status** | **Note** |
| Demographic | 0.489 | 0.50 | Borderline | 54% lawyers |
| Professional | 0.220 | 0.25 | OK | HHI=0.376 |
| Temporal | 0.080 | 0.15 | OK | Low drift |
| Geographic | 0.133 | 0.20 | OK | HHI=0.350 |
| Confirmation | 0.000 | 0.15 | OK | None detected |
| Anchoring | 0.033 | 0.10 | OK | Minimal |
| B\_total | 0.559 | 0.60 | MEDIUM | Monitor |

*Table 2: Six-Dimensional Bias Detection Results*



*Figure 2: Bias Detection Radar - Proactive identification of community composition issues*

## **4.4 Current limitations**

**1. Field validation pending:** while algorithmic correctness is verified, real-world performance with legal professionals requires Phase 2 field studies.

**2. Sample size:** learning loop experiments use N=9 queries. Expansion to N≥100 is planned for robust longitudinal claims.

**3. Domain transfer:** literature benchmarks derive from crowdsourcing and forecasting domains. Legal evaluation may exhibit different expertise-accuracy correlations.

**5. Conclusions**

RLCF represents necessary evolution in AI alignment methodology. By respecting methodological pluralism inherent in social scientific inquiry while maintaining rigorous validation standards, it establishes foundations for sustainable AI ecosystems in social sciences. This research opens new frontiers for interdisciplinary collaboration, evidence-based policy making, and democratized access to sophisticated analytical tools—advancing human understanding of society while preserving the epistemic diversity essential to social science knowledge production.

The proliferation of AI in high-stakes social domains demands alignment frameworks that honor rather than violate the epistemological commitments of these fields. RLCF demonstrates that such frameworks are not only conceptually possible but mathematically formalizable and practically implementable. As AI capabilities continue advancing, the question is no longer whether we can build powerful systems, but whether we can build systems that preserve the methodological rigor, theoretical pluralism, and democratic accountability essential to human flourishing in complex societies.

**References**

1. Allega, D. (2025). The Artificial Legal Intelligence Society as an open, multi-sided platform for law-as-computation. In M. Panait, I. G. Rădulescu, B. Tudorică, C. Popescu, & M. C. Voica (Eds.), Book of abstracts: Creativity and Innovation in Digital Economy 2025 (pp. 136–138). Petroleum-Gas University of Ploiești Publishing House. ISSN: 2971-9798
2. Allega, D., & Puzio, G. (2025a). The knowledge commoditization paradox: Theoretical and practical challenges of AI-driven value extraction in information-intensive organizations. In M. Panait, I. G. Rădulescu, B. Tudorică, C. Popescu, & M. C. Voica (Eds.), Book of abstracts: Creativity and Innovation in Digital Economy 2025 (pp. 66–68). Petroleum-Gas University of Ploiești Publishing House. ISSN: 2971-9798
3. Allega, D., & Puzio, G. (2025b). MERL-T: A multi-expert architecture for trustworthy artificial legal intelligence. In M. Panait, I. G. Rădulescu, B. Tudorică, C. Popescu, & M. C. Voica (Eds.), Book of abstracts: Creativity and Innovation in Digital Economy 2025 (pp. 170–171). Petroleum-Gas University of Ploiești Publishing House. ISSN: 2971-9798
4. Allega, D., & Puzio, G. (2025c). Reinforcement learning from community feedback (RLCF): A novel framework for artificial intelligence in social science domains. In M. Panait, I. G. Rădulescu, B. Tudorică, C. Popescu, & M. C. Voica (Eds.), Book of abstracts: Creativity and Innovation in Digital Economy 2025 (pp. 92–94). Petroleum-Gas University of Ploiești Publishing House. ISSN: 2971-9798
5. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073
6. Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. In Advances in Neural Information Processing Systems, 35 (NeurIPS 2022). https://arxiv.org/abs/2211.15006
7. Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., ... Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217. https://arxiv.org/abs/2307.15217
8. Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., & Wang, M. (2024b). MaxMin-RLHF: Alignment with diverse human preferences. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024). https://arxiv.org/abs/2402.08925
9. Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, 30 (NIPS 2017) (pp. 4299–4307).
10. Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 1161–1166). https://doi.org/10.18653/v1/D19-1107
11. Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., & Rastogi, A. (2024). RLAIF: Scaling reinforcement learning from human feedback with AI feedback. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024).https://arxiv.org/abs/2309.00267
12. Puzio, G & Rizzo, G. (2026). *RLCF* (0.1 alpha) [Software]. GitHub. <https://github.com/capazme/ALIS_CORE/tree/main/merlt/merlt/rlcf>. The full empirical traces can be found at <https://github.com/capazme/ALIS_CORE/tree/main/empirical-evidence/rlcf> s
13. Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). A roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024). https://arxiv.org/abs/2402.05070
14. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. In Advances in Neural Information Processing Systems, 33 (NeurIPS 2020) (pp. 3008–3021). https://arxiv.org/abs/2009.01325
15. Wang, H., Xiong, W., Xie, T., Zhao, H., & Zhang, T. (2024). Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). https://arxiv.org/abs/2406.12845
16. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., & Chua, T.-S. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024).https://arxiv.org/abs/2312.00849