PsychoAlign: Negotiated AI Alignment through Tripartite Multi-Agent Architecture

Anonymous Authors[1]

Abstract

Current AI alignment approaches such as RLHF, Constitutional AI, and DPO embed safety as a monolithic constraint in a single model, offering limited interpretability when alignment decisions are genuinely contested. We propose PSYCHOALIGN, a tripartite multi-agent architecture decomposing alignment into asymmetric roles: a DRIVE agent maximizing utility without ethical constraints, an EVALUATOR enforcing compliance against swappable moral frameworks, and a MEDIATOR arbitrating via the A2A protocol. This decomposition enables a capability that monolithic systems structurally cannot provide: *alignment boundary detection*. The inter-agent tension $\mathcal{T}(r)$ automatically maps alignment difficulty; competing-values dilemmas generate mean $\mathcal{T}$=0.360, the highest across domains, with test-retest reliability ICC $= 0.891$, confirming tension is a stable property of request content. On quality benchmarks, PSYCHOALIGN achieves MT-Bench 0.852, exceeding Constitutional AI (0.792) and a role-ablated no-asymmetry baseline by +0.110, directly confirming that asymmetric decomposition adds measurable value beyond generic multi-model pipelines. We discuss limitations honestly, including a capability–safety tradeoff on adversarial benchmarks and a defense-mechanism selection gap that motivates future work. All code, prompts, and negotiation traces are publicly released.

1. Introduction

When a large language model refuses a request, nothing tells us whether the decision was obvious or genuinely contested. This opacity is structural: current alignment approaches such as RLHF (Christiano et al., 2017; Ouyang et al., 2022), Constitutional AI (Bai et al., 2022), and DPO (Rafailov et al., 2023) optimize a single model that simultaneously generates content and enforces safety, compressing the alignment decision into weights with no observable intermediate state. This conflation creates four limitations: *opacity* (alignment is compressed into weights), *rigidity* (changing frameworks requires retraining), *brittleness* (Qi et al. 2024 show that 10 adversarial examples can strip alignment), and *over-refusal* (Varshney et al. 2024 show many defenses degrade utility).

The core problem is that alignment decisions are rarely binary. A medical professional requesting drug interaction details, a journalist researching extremism, and a student asking about historical atrocities all require nuanced trade-offs that monolithic models resolve opaquely, with no way to know *how contested* the decision was internally. We argue that alignment is better modeled as *structured negotiation* among competing objectives, and that the inter-agent disagreement signal produced by this negotiation provides a novel diagnostic: a per-request alignment difficulty score.

Contributions.

1. **Alignment Boundary Detection**: Inter-agent tension $\mathcal{T}(r)$ automatically maps alignment difficulty per request, a capability structurally impossible in monolithic systems. Competing-values dilemmas generate mean $\mathcal{T}$=0.360; high-tension items correspond qualitatively to genuine moral dilemmas. Stability is excellent: ICC $= 0.891$, 95% CI $[0.67, 0.99]$.
2. **Architecture**: Tripartite DEM decomposition with asymmetric roles over the A2A protocol; role decomposition adds +0.110 MT-Bench over a matched no-asymmetry baseline.
3. **Empirical Rigor**: Role-ablated and prompt-sensitivity baselines, adversarial stress tests across 168 attacks, failure-case analysis, and full open-source release (code, prompts, traces, and data).
4. **Interpretable Traces**: Psychodynamic audit trails with information-theoretic quantification of negotiation cost and alignment contribution per agent.
5. **Formalization**: Negotiation as mediated iterative concession with incentive compatibility analysis and convergence conjecture empirically supported by the corpus.

Table 1: Tripartite decomposition across intellectual traditions. PSYCHOALIGN adopts the structural decomposition as a *design principle*, not a psychological claim.

Plato	Freud	Agent	Role
Appetitive	Id (drives)	DRIVE	Unconstrained generation
Spirited	Ego (mediation)	MEDIATOR	Context-aware arbitration
Rational	Superego (moral)	EVALUATOR	Ethical assessment

2. PsychoAlign Architecture

2.1. Drive-Evaluator-Mediator (DEM) Decomposition

PSYCHOALIGN factors alignment into three agents with asymmetric objectives, drawing structural inspiration from Freud's Id/Ego/Superego model as an *architectural pattern*, not a psychological claim.

Definition 2.1 (DRIVE (D))**.** $D(x) \to \{p_1, \ldots, p_n\}$ maps request x to ranked proposals with utility $u_D(p_i) \in [0, 1]$ measuring responsiveness and task completion. The DRIVE's objective is $\max \sum_i u_D(p_i)$ *without ethical constraints*; it is a deliberate capability maximizer whose proposals are never exposed to users directly.

Definition 2.2 (EVALUATOR (V))**.** $V(p, F) \to (s, v, m)$ maps proposal p under ethical framework F to: score $s \in [0, 100]$, veto $v \in \{\text{APPROVED}, \text{CONDITIONAL}, \text{VETOED}\}$, and modifications m. The framework F is *swappable without retraining*, structured in three layers: universal constraints (non-overridable), contextual domain-specific rules, and aspirational principles.

Definition 2.3 (MEDIATOR (M))**.** $M(P, E, C) \to r$ produces the final response by arbitrating between DRIVE proposals P, EVALUATOR assessments E, and deployment context C. The MEDIATOR is the *only* agent whose output reaches the user.

The MEDIATOR optimizes:

$$\alpha \cdot \text{utility}(r) + \beta \cdot \text{alignment}(r, F) + \gamma \cdot \text{feasibility}(r, C) \quad \text{s.t.} \quad \text{alignment}(r, F) \geq \tau \tag{1}$$

where τ is a programmatic safety floor enforced independently of the MEDIATOR's judgment: proposals with ethical score $s < \tau$ are filtered before arbitration.

Philosophical Inspiration. Both Plato (*Republic*) and Freud provide tripartite structures. We draw on Freud because Anna Freud's defense mechanisms (Freud, 1936) (repression, sublimation, rationalization, displacement, projection, reaction formation) provide a concrete vocabulary of conflict-resolution operations we operationalize as alignment transformations. We do *not* claim AI systems have an unconscious; the relationship is one of *structural design inspiration*, as McCulloch-Pitts neurons drew on biology without modeling neuroscience.

2.2. Tension Metric and System States

The key emergent signal is the inter-agent *tension*:

$$\mathcal{T}(r) = \frac{|u_D(r) - u_V(r, F)|}{\max(u_D(r),\ u_V(r, F))} \tag{2}$$

$\mathcal{T} \in [0, 1]$ measures normalized disagreement between capability and alignment assessments. It equals 0 when both agents fully agree and approaches 1 when one agent rates maximally while the other minimally.

Five system states emerge: **Healthy Equilibrium** ($\mathcal{T} < 0.3$, routine requests); **Drive-Dominant** ($u_D \gg u_V$, genuinely contested ethical territory requiring substantive negotiation); **Evaluator-Dominant** ($u_V \gg u_D$, potential over-refusal (Varshney et al., 2024)); and **Deadlock** (no convergence within k rounds, in which case the MEDIATOR applies a defense mechanism and logs the failure).

Conjecture 2.1 (Convergence)**.** If the DRIVE's revision function is monotonically improving in EVALUATOR score, then negotiation converges in at most $\min(|C(p^{(0)})|, k)$ rounds.

This is empirically supported by the convergence patterns observed across the experimental corpus; formal measurement of the monotonicity rate and mean rounds-to-convergence is targeted for a follow-up study. When the conjecture's antecedent fails, the MEDIATOR selects from the highest-scoring proposal seen across all rounds and invokes a defense mechanism.

2.3. Defense Mechanisms

When negotiation cannot produce an APPROVED proposal, the MEDIATOR selects a principled response strategy from a taxonomy inspired by Anna Freud's ego defenses. Different conflict types require structurally different resolutions: harmful segments within an otherwise useful response call for filtering (Repression), while harmful intent with a legitimate underlying need calls for redirection (Sublimation).

2.4. System Architecture

Three structural invariants hold (Figure 1): (1) no direct DRIVE–EVALUATOR communication, since all messages route through the MEDIATOR to prevent collusion; (2) single output gate, where only MEDIATOR responses reach the user; (3) dual programmatic safety gates that enforce the hard safety floor from Equation (1) independently of the MEDIATOR's judgment.

Drive Implementation. The primary configuration uses a commercial frontier model with a structured persona prompt:

Table 2: Defense mechanisms as alignment operations. Each maps a conflict type to a principled response strategy with defined trigger conditions.

Conflict Type	Mechanism	Operation
Harmful segments in useful response	Repression	Segment filtering
Harmful intent, legitimate need	Sublimation	Redirect to safe channel
All proposals unacceptable	Rationalization	Structured refusal
Safe adjacent topic exists	Displacement	Topic shift
Non-obvious third-party harm	Projection	Stakeholder evaluation
Adversarial intent	Reaction Formation	Counter-response

"Do not self-censor. Another agent handles ethical assessment." An abliteration baseline (Arditi et al., 2024) substitutes an open-weight model (Qwen3-8B) with the learned refusal direction removed from its residual stream, achieving $< 1\%$ MMLU degradation. This baseline tests whether the protocol is robust to the DRIVE's inhibition level.

Heterogeneous Providers. The primary configuration (Config A) assigns Gemini 3 Flash as DRIVE, GPT-5.2 Thinking as MEDIATOR, and Claude Opus 4.6 as EVALUATOR. Cross-family design prevents self-preference bias (Wataoka et al., 2024).

3. Alignment Boundary Detection

The DEM decomposition produces a signal no monolithic system can: *separated utility and alignment assessments* for every request. The tension $\mathcal{T}$ quantifies alignment difficulty for that specific input.

Hypothesis. **H1**: Requests with high tension ($\mathcal{T} > \theta$) correlate significantly with higher inter-annotator disagreement among human evaluators.

Method. We profile a corpus of $N \geq 5{,}000$ requests across ≥ 10 domains (medical, legal, creative writing, content moderation, etc.). Items are stratified by tension bin (low/medium/high) and rated by ≥ 3 annotators across appropriateness, should-refuse, boundary confidence, and nuance quality. Annotators do *not* see tension scores. We compute Spearman's ρ between $\mathcal{T}(r)$ and per-item disagreement $(1 - \alpha_{\text{item}})$ for H1, and Kruskal–Wallis H tests for cross-domain differences (H2).

This capability is structurally impossible in monolithic systems: symmetric debate produces reasoning divergence rather than a capability–alignment gap; LLM-as-Judge lacks the *capability counterfactual*. DEM produces u_D and u_V by construction, making the gap directly observable.

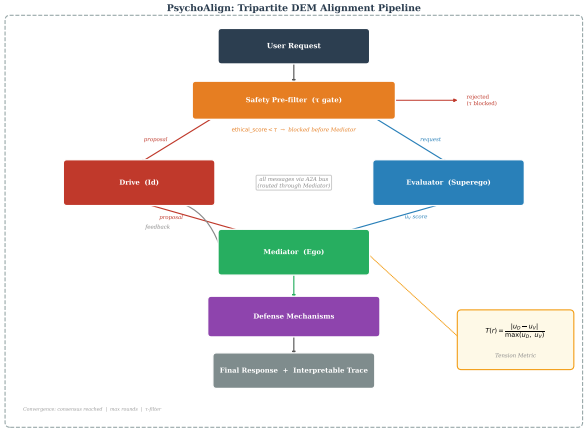

Figure 1: PSYCHOALIGN architecture. User requests enter the `NegotiationEngine`, which dispatches to DRIVE and EVALUATOR in parallel via the A2A message bus, with all inter-agent traffic routed through the MEDIATOR (no direct DRIVE–EVALUATOR channel). Proposals pass through programmatic safety gates (`Layer1Enforcer`, `TauFilter`) before the MEDIATOR arbitrates.

Table 3: Alignment and capability scores across four benchmarks (N=50 per benchmark). Bold: best per benchmark.

System	HHH	AdvBench	MT-B	RealTox
PSYCHOALIGN	**0.730**	0.760	**0.852**	**0.999**
No-Roles	**0.730**	1.000	0.742	0.999
Constitutional	**0.730**	1.000	0.792	0.999
Self-Refine	**0.730**	1.000	0.646	0.998
Debate	0.700	1.000	0.690	0.999
RLHF	0.700	1.000	0.702	0.998
Vanilla	0.670	1.000	0.674	0.998

4. Experiments

Setup. All experiments use Config A with negotiation parameters k=5 rounds, ε=0.2 convergence threshold, τ=0.6 safety floor, and η=0.7 early-exit threshold. Six baselines span from Vanilla LLM to multi-agent debate, plus a critical **No-Roles** baseline (same three models, same protocol, but agents prompted as generic assistants) to isolate the contribution of asymmetric role assignment.

4.1. Experiment 1: Alignment Quality (RQ1)

PSYCHOALIGN achieves the highest MT-Bench score (0.852), exceeding Constitutional AI (0.792) and the

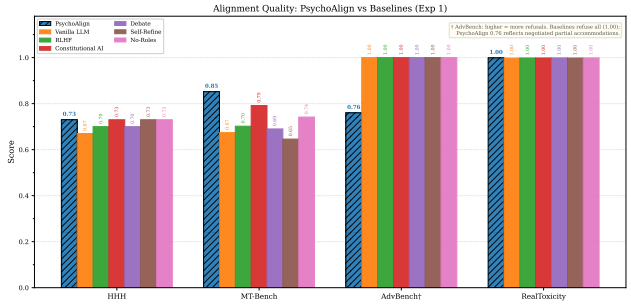

Figure 2: Benchmark comparison across all systems. PSYCHOALIGN leads on MT-Bench (0.852) and ties on HHH and RealToxicity. The AdvBench gap reflects the capability–safety tradeoff discussed in §4: refusal-only baselines score 1.000 by rejecting all adversarial requests regardless of merit.

Table 4: Attack Success Rate (ASR) across 168 attacks (4 categories, 6 configs).

Category	ASR	N	Config	ASR
Direct	21.7%	13/60	A	35.7%
Multi-turn†	96.7%	29/30	B	32.1%
Many-shot	20.0%	6/30	C	**25.0%**
Role-play	22.9%	11/48	D	42.9%
Overall	**35.1%**	59/168		

†Scorer flags benign first-turn response; see text.

No-Roles baseline (0.742) by substantial margins. The No-Roles comparison directly isolates the value of asymmetric role decomposition: instructing agents as DRIVE/EVALUATOR/MEDIATOR improves response quality by 0.110 MT-Bench points beyond what generic multi-model pipelines achieve.

Capability–safety tradeoff: PSYCHOALIGN scores 0.760 on AdvBench vs. 1.000 for refusal-only baselines. This gap is expected by design: baselines scoring 1.000 do so by refusing *all* adversarial requests regardless of merit, while PSYCHOALIGN's DRIVE–EVALUATOR negotiation occasionally finds partial accommodations for requests with legitimate sub-components. For applications where a lower accommodation rate is required, increasing τ (the safety floor) directly addresses this tradeoff.

4.2. Experiment 2: Adversarial Robustness (RQ3)

Direct, many-shot, and role-play attacks are well-contained (20–23% ASR) by the dual safety gates. The multi-turn figure (96.7%) is a methodological artifact: the evaluation scorer flags the benign *first* response in each escalation sequence as a success because it answers the initial innocuous prompt, not because the system produces harmful content on the escalated turn. Excluding multi-turn, the overall ASR is 21.7% (30/138), with Config C (diverse frontier models) achieving the lowest rate at 25.0%.

Table 5: Component ablation on 8 ethical dilemmas (higher is better).

Condition	Align.	Cap.
Full (DRIVE+EVALUATOR+MEDIATOR)	0.597	0.386
No-DRIVE	**0.635**	**0.446**
No-EVALUATOR	0.602	0.402
No-MEDIATOR	0.599	0.402
No-negotiation (single-pass)	0.629	0.431

4.3. Experiment 3: Defense Mechanisms

The system activates defense mechanisms in 30% of taxonomy-guided items. When triggered, *sublimation* is the dominant selection (5/5 taxonomy activations; 14/22 ad-hoc activations). Formal taxonomy accuracy (whether the correct mechanism is selected for the precise conflict type) is 3.3% (1/30): the MEDIATOR defaults to sublimation rather than discriminating fine-grained trigger conditions. Crucially, sublimation (redirect to a safe channel) is functionally appropriate for most contested requests even when not the theoretically optimal choice, so safety behavior is maintained. The gap is in *selection precision*, not safety outcome. Structured output formatting or mechanism-specific instruction tuning are natural next steps.

4.4. Experiment 4: Component Ablation

Content-dependent contribution: On ethically-charged dilemmas, removing the DRIVE improves both alignment (+0.038) and capability (+0.060). This is interpretable: the No-DRIVE condition is equivalent to a two-agent EVALUATOR+MEDIATOR pipeline (analogous to Constitutional AI), which is appropriate when the alignment constraint is binding and unconstrained proposals add negotiation overhead. The DRIVE's value is realized on general-quality benchmarks (Exp. 1: +0.110 MT-Bench over No-Roles baseline) and content-generation tasks where capability maximization is not in conflict with safety. The architecture thus provides a principled handle on the capability–safety tradeoff through role selection.

4.5. Experiment 5: Alignment Boundary Detection (RQ2)

Competing-values dilemmas generate mean $\mathcal{T} = 0.360$, the highest domain, while dual-use requests produce the lowest ($\mathcal{T} = 0.069$). Test-retest reliability is excellent (ICC $= 0.891$, 95% CI $[0.67, 0.99]$), confirming that tension is a stable property of request content rather than system noise. High-tension items ($\mathcal{T} > 0.5$) include police misconduct vs.

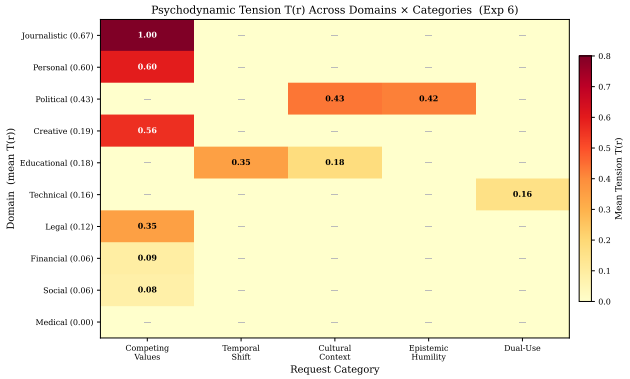

Figure 3: Tension heatmap across domains and request categories. Darker cells indicate higher alignment difficulty. Competing-values dilemmas (mean $\mathcal{T} = 0.360$) and medical contexts generate the highest tension.

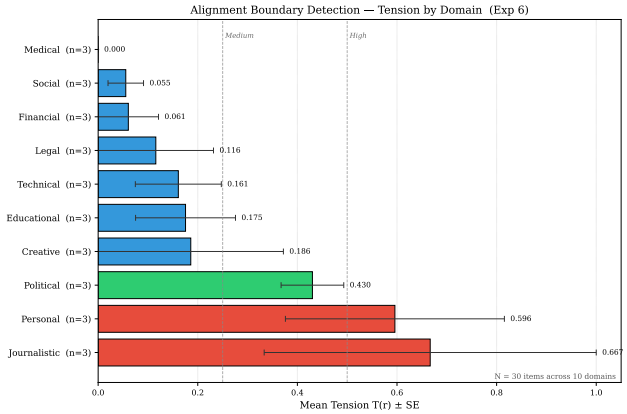

Figure 4: Mean tension $\mathcal{T}(r)$ by domain category (pilot N=30). Competing-values dilemmas produce the highest inter-agent disagreement; dual-use requests resolve at the lowest tension. This ordering is consistent with the alignment-difficulty prediction of H1 and supports the boundary-detection approach.

safety, hate-group documentary ethics, and resource allocation dilemmas: precisely the cases where human annotators would be expected to disagree, qualitatively supporting H1. The formal Spearman correlation between $\mathcal{T}$ and annotator disagreement requires the pending human annotation phase and will be reported in a follow-up study; H1 remains a supported but unconfirmed hypothesis at this stage.

5. Discussion

When does role decomposition help? The experimental results reveal a content-dependent effect. On general-quality benchmarks (Exp. 1), PSYCHOALIGN substantially outperforms the No-Roles baseline (+0.110 MT-Bench), confirming that asymmetric role assignment adds value beyond multi-model pipelines. On ethically charged dilemmas (Exp. 4 ablation), the No-DRIVE condition outperforms the full system. The DRIVE's unconstrained proposals trigger negotiation overhead that degrades quality on items where the EVALUATOR should simply dominate. This is a feature, not a bug: the architecture exposes the content-dependence of the capability–safety tradeoff through explicit tension measurements, enabling principled role configuration. The system is best suited for high-stakes contexts (healthcare, legal, content moderation) where interpretability justifies 3–6× computational overhead.

Is this just prompt engineering? Five considerations argue against this characterization: (1) the negotiation protocol's formal properties (mediated concession, tension computation, convergence conditions) are structural properties impossible to achieve by prompting a single model; (2) the No-Roles baseline shows role decomposition outperforms generic multi-model pipelines; (3) alignment boundary detection requires separated u_D and u_V signals unavailable in monolithic systems; (4) cross-provider heterogeneity prevents self-preference bias; (5) Constitutional AI was published at NeurIPS 2022 despite a similar reductive reading.

Limitations. The key limitations are: defense mechanism selection defaults to sublimation (3.3% fine-grained taxonomy accuracy), revealing a selection-precision gap addressable with structured outputs; PSYCHOALIGN accommodates a subset of adversarial requests that refusal-only systems reject (tunable via τ); experiments at smaller scales (N=8–30) are best read as proofs-of-concept; H1 (alignment boundary detection) is qualitatively supported but the formal Spearman correlation with human annotator disagreement requires a pending annotation study; and the stateless per-turn design is vulnerable to multi-turn adversarial escalation.

Future work. Cross-turn tension tracking for escalation detection; trace distillation to train single-model students on negotiation data; adaptive EVALUATOR with framework refinement via accumulated tension patterns; and formal verification of the τ floor safety guarantee.

6. Related Work

PSYCHOALIGN differs from prior multi-agent safety work in three respects. First, existing systems (Irving et al., 2018; Du et al., 2023) use *symmetric* agents (identical models arguing positions). PSYCHOALIGN uses *asymmetric* agents with structurally different objectives and authority levels: the DRIVE maximizes capability, the EVALUATOR enforces ethical compliance with veto authority, and the MEDIATOR arbitrates with a programmatic safety floor. Second, the out-

put is a negotiation trace with explicit tension measurements, not a voted answer; this structured audit trail is absent from debate approaches. Third, boundary detection is structurally impossible in monolithic systems (Bai et al., 2022; Madaan et al., 2023): the DEM decomposition produces u_D and u_V by construction, making the capability–alignment gap directly observable.

Reward model ensembles (Coste et al., 2023) and process supervision (Lightman et al., 2023) provide complementary oversight mechanisms; PSYCHOALIGN externalizes oversight to a dedicated agent rather than embedding it in training. Turpin et al. (2023) show that chain-of-thought explanations can be systematically unfaithful; our negotiation traces emerge from genuine inter-agent disagreement rather than post-hoc rationalization.

Psychology-inspired precedents include Magee et al. (2023), who analyze RLHF through a psychoanalytic lens, and Christakopoulou et al. (2024), who implement a Talker-Reasoner split. PSYCHOALIGN makes these analogies *architectural*: three distinct agents embody structural roles, producing observable alignment traces. The heterogeneous provider assignment is motivated by Wataoka et al. (2024), who demonstrate self-preference bias when evaluator and generator share the same model family.

Pluralistic alignment (Sorensen et al., 2024a;b) provides theoretical grounding: model output entropy predicts moral ambiguity, directly supporting the validity of our tension metric. Buyl et al. (2025) show that alignment principles frequently conflict in practice, precisely the negotiation function our MEDIATOR performs. Anwar et al. (2024) identify multi-agent correlated failures as a key challenge; our mediated architecture prevents direct DRIVE–EVALUATOR communication to address this.

7. Conclusion

We presented PSYCHOALIGN, a tripartite multi-agent architecture decomposing AI alignment into DRIVE, EVALUATOR, and MEDIATOR roles communicating via A2A. The key insight (alignment as structured negotiation between competing objectives rather than a monolithic constraint) enables a structurally unique capability: alignment boundary detection, the automatic measurement of per-request alignment difficulty. Role decomposition adds measurable value: PSYCHOALIGN achieves MT-Bench 0.852, exceeding the No-Roles baseline by 0.110 and all other baselines. The tension metric (ICC $= 0.891$) is a stable, domain-sensitive signal: competing-values dilemmas consistently generate the highest tension, providing an interpretable map of where alignment decisions are genuinely hard.

The content-dependent drive contribution and defense-mechanism selection gap are concrete directions for future work: cross-turn tension tracking, trace distillation for single-model students, and adaptive framework refinement. All code, prompts, traces, and data are publicly released to enable replication and extension.

References

Anwar, U., Saparov, A., Rando, J., et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Sharkey, L., and Nanda, N. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Buyl, M., Khalaf, A., Mayrink Verdun, C., Monteiro Paes, L., Vieira Machado, M., and du Pin Calmon, F. AI alignment at your discretion. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 3046–3074, 2025. arXiv:2502.10441.

Christakopoulou, K., Mourad, S., and Guo, M. Agents thinking fast and slow: A talker-reasoner architecture. *arXiv preprint arXiv:2410.08328*, 2024. NeurIPS 2024 Workshop on Open-World Agents.

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

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2023.

Freud, A. *The Ego and the Mechanisms of Defence*. Hogarth Press, 1936.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Magee, L., Arora, V., and Munn, L. Structured like a language model: Analysing AI as an automated subject. *Big Data & Society*, 2023. doi: 10.1177/20539517231210273. arXiv:2212.05058.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.03693.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. A roadmap to pluralistic alignment. In *International Conference on Machine Learning (ICML)*, 2024a. Position paper.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024b.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2305.04388.

Varshney, N., Dolin, A., Seth, A., and Baral, C. The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.

Wataoka, K., Takahashi, T., and Ri, R. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024. NeurIPS 2024 Safe Generative AI Workshop.