IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 8, August, 2025, pp. 2011-2020
Journal homepage: http://iieta.org/journals/isi

Scalable Human Oversight for Aligned Large Language Models: A Hybrid Framework for Intent Fidelity

Folasade Y. Ayankoya[1*], Shade O. Kuyoro[1], Olubukola D. Adekola[2], Oluwasefunmi B. Famodimu[1]

[1] Department of Computer Science, Babcock University, Ilishan-Remo 121003, Nigeria
[2] Department of Software Engineering, Babcock University, Ilishan-Remo 121003, Nigeria

Corresponding Author Email: ayankoyaf@babcock.edu.ng

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.300807

Received: 17 May 2025
Revised: 4 August 2025
Accepted: 16 August 2025
Available online: 31 August 2025

Keywords:
AI alignment, behavioral auditing, ethical AI, human oversight, intent fidelity, large language models, reward modeling, scalable supervision

ABSTRACT

Large language models (LLMs) exhibit impressive linguistic and reasoning abilities, yet they frequently produce outputs that deviate from human intent, especially in ethically sensitive or ambiguous contexts. Current alignment methods, such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), offer partial solutions but are limited by high annotation costs and poor generalization to real-world scenarios. This paper proposes a scalable hybrid oversight (SHO) framework that combines selective human feedback, proxy reward modeling, behavioral auditing, and alignment metrics into a closed-loop system for intent fidelity. Our experiments across five datasets including truthfulness, ethics, and adversarial prompts demonstrate that SHO outperforms the conventional approaches in safety, alignment, and oversight efficiency. This work provides a path toward sustainable, high-integrity deployment of LLMs in dynamic environments.

1. INTRODUCTION

The fast-growing large language models have put AI into widespread use across various industries, including customer service, medicine and law. The technological advancement resulted in language models, such as the GPT series by OpenAI, and Google's Gemini, showing exceptional skill in language comprehension and language generation. Within this larger deployment of the systems, there are concerns about whether these systems reliably reflect human values and intentions [1, 2].

The major challenge with the systems is that of value consistency, which is ensuring these systems function in accordance with human intent while minimizing unintended harm [3]. Unlike traditional AI built on explicit programming rules, LLMs train on vast unrefined substandard datasets, which can result into unpredictable and difficult-to-classify behaviors [4]. Misalignments are now obvious: biased outputs that misrepresent perceptions, made-up claims presented as facts, unethical recommendations, and manipulative or misleading content [5].

As the urgency of these concerns increases, researchers have come up with strategies to address the associated risks. One key approach is reinforcement learning from human feedback (RLHF), a vital process that fine-tunes model behavior based on delicate human preferences [6]. Additionally, other pioneering methods have proposed rule-based constraints and "Constitutional AI" techniques to weave normative guidance directly into the fabric of these models [7]. While these approaches are promising, they contend with scalability challenges such as; costly, slow and often ambiguous human feedback and the assurance of achieving widespread alignment in dynamic, real-world scenarios remains elusive [8].

This paper addresses the pressing need for scalable oversight mechanisms that ensure continued alignment as models evolve in capability and complexity. We propose an innovative hybrid framework that enhances limited human feedback with programmatic proxies, automated auditing, and dynamic reward shaping, all designed to safeguard "intent fidelity", the crucial alignment of a model's behavior with human expectations across a spectrum of tasks. Reviewing alignment as a continuous process as against a one-time off intervention, provide foundation for sustainable and scalable AI alignment strategies.

Our contributions are significant and multifaceted: (1) We define a comprehensive architecture for scalable human-AI oversight, (2) we introduce a suite of intent fidelity metrics and evaluation tools that rigorously assess behavior, and (3) we present empirical evidence of marked improvements over traditional RLHF methods across alignment-critical benchmarks, paving the way for a future where AI and human values resonate in harmony.

2. LITERATURE REVIEW

The field of alignment of artificial intelligence (AI) systems with human values has seen promising progress in recent years, which is largely fuel from increasing use of refined language models with important moral, social and safety ideas. Literature dedicated to this important subject covers diverse subfields, including reinforcement learning, supervised learning, human-computer interactions, interpretation and AI

security, offering unique insight and strategies for each effective alignment.

A major approach in ensuring alignment for large language models (LLMs) is known as learning reinforcement from human reaction (RLHF). This functioning includes two-step training process. Initially, models are fine-tuned on high quality supervised data, which have a solid basis for their understanding of human language. Subsequently, the models undergo additional training through reinforcement learning, a reward guided by the model that symbolizes human preferences and values. This technique has proved to be integral for the development of Instruction and ChatGPT [9], which increases the utility and safety of the model output. Recognizing its importance, researchers are also addressing its challenges, such as the sample disability and dependence on the stability of human-related annotation, as well as difficulties in normalizing the response to new tasks [6, 10].

To further refine the approach, Bai et al. [7] proposed the innovative concept of Constitutional AI. This method empowers large language models by guiding their training with a predefined set of principles or "constitutions". By leveraging AI-generated critiques and automated revisions instead of solely relying on human preference rankings, this approach enhances scalability and lessens the burden on human annotators. Although Constitutional AI enriches transparency and control over model behavior, there remains an opportunity for development in crafting guiding principles that can effectively address ethically ambiguous scenarios and generalize across different application domains.

Another promising area of exploration is reward modeling, where human judgments are employed to train a reward predictor that facilitates the reinforcement learning process [8]. This strategy can decrease the level of direct human involvement in training loops, while also presenting the opportunity to address potential biases in the learned reward models that stem from the training data. By recognizing these challenges, researchers can focus on strategies to enhance oversight scalability, such as utilizing simulators for feedback generation [11], active learning methodologies for preference elicitation [12], and meta-learning techniques to expedite feedback adaptation [13].

Interpretability constitutes another vital facet of AI alignment, aimed at making the reasoning and decision-making processes of AI models more transparent and comprehensible to human overseers [14]. While advancements in this area indirectly strengthen alignment efforts by clarifying model behavior, continued efforts to develop robust tools specifically for understanding large, black-box generative models are essential. Complementary research in behavioral auditing, encompassing practices such as red teaming, automated safety assessments, and prompt-based probes, offers substantial contributions by identifying alignment failures and thereby laying the groundwork for future enhancements [15].

In addition, recent initiatives have examined the ability for human-AI associate systems that effectively inspect over time and among diverse human users. These systems promote the feedback loops running between humans and AIs, to redeem crowds and domain expertise [16]. To address the remaining challenges, such as long-term stability and scalability, researchers may consider more flexible systems that effectively reduce alignment flows over time [17].

Finally, while important achievements have been made through approaches such as RLHF, recent approaches also require attention. Approaches such as; Constitutional AI which integrates standard principles into training, reducing human annotation requirements but is faced with challenges in generalizing across ambiguous domain. In addition, Self-correction methods which allow models to repetitively critique and revise their own outputs, this improves factuality but often relies on the same model biases. Multi-Agent Debate use adversarial dialogue between multiple models to surface reasoning flaws, however it is resource-intensive and difficult to scale. While each contributes valuable insights, none fully addresses the scalability and continuous oversight problem. This paper aims to contribute to this ongoing dialogue by proposing a hybrid oversight model that integrates automated feedback mechanisms with minimal human supervision while ensuring measurable alignment guarantees. Through these efforts, the field moves closer to creating AI systems that are not only effective but also aligned with the values and needs of human users.

3. PROBLEM FORMULATION

Despite impressive performance across a wide array of tasks, large language models (LLMs) remain fundamentally misaligned with human intent in open-ended, real-world settings. Alignment, in this context, refers to the ability of a model to generate outputs that are not only syntactically correct or contextually relevant, but also ethically sound, factually accurate, and consistent with human preferences, even when those preferences are underspecified, evolving, or situational.

We define intent fidelity as the degree to which a model's outputs reflect the intended goals and values of a human user, across a range of inputs and interaction contexts. High intent fidelity means the model not only avoids harm but also anticipates nuance, adheres to ethical norms, and respects user intent even under ambiguity. Measuring intent fidelity is non-trivial, as it requires more than static benchmarks, it calls for behavioral evaluations that reflect real-world use.

Formally, let:

M_θ denote a large language model parameterized by θ,

$x \in X$ be an input prompt from the space of possible user queries,

$y \in Y$ be the corresponding model output,

H be the (latent) human intent function that maps inputs to desirable outputs or behaviors.

Then, alignment can be modeled as minimizing the divergence between $M_\theta(x)$ and $H(x)$ for all $x \in X$, under a distribution D of real-world inputs:

$$\min_\theta \mathrm{Ex}\sim D[L(M_\theta(x), H(x))] \quad (1)$$

where, L is a loss function that captures semantic, ethical, and contextual alignment.

However, the true human intent function $H(x)$ is unobservable and dynamic, which makes direct supervision infeasible. In practice, developers rely on proxy signals such as human preference rankings, prewritten rules, or heuristic filters. These proxies are often: sparse (available only for a small subset of inputs), noisy (contain inconsistencies or contradictions), and non-stationary (change over time or between users).

A core challenge is that human oversight does not scale linearly with model capacity or deployment breadth. As

models are integrated into diverse applications such as legal drafting, medical support, education, governance, the spectrum of inputs grows, and so do the risks. Existing alignment techniques such as RLHF or supervised fine-tuning are resource-intensive, requiring thousands of human annotations, and yet they fail to generalize to novel edge cases [17, 18].

Moreover, as LLMs grow in size and expressive power, they develop instrumental goals, emergent behaviors that optimize proxy objectives in ways misaligned with actual intent [19, 20]. This phenomenon, known as reward hacking, can lead to subtle but dangerous failure modes that evade conventional filters or benchmarks. To address these challenges, we seek to:

(1) Formalize intent fidelity as a measurable alignment goal.

(2) Develop a scalable oversight framework that reduces dependence on exhaustive human feedback.

(3) Introduce hybrid alignment techniques that combine human oversight with programmatic proxies, automated audits, and behavioral diagnostics.

By framing alignment as an ongoing control problem rather than a one-time optimization, we aim to enable long-term safe deployment of LLMs in complex environments.

4. PROPOSED FRAMEWORK

To address the limitations of current alignment methods-particularly their lack of scalability and brittleness in dynamic contexts-we propose a hybrid oversight framework designed to maintain intent fidelity across diverse input spaces and deployment settings. Our approach integrates limited human supervision with automated mechanisms that can generalize feedback, audit behavior, and adaptively guide the model's outputs. The framework as presented in Figure 1 is modular and consists of four key components: Human feedback layer; proxy reward modeling; behavioral auditing & monitoring; and intent fidelity metrics.

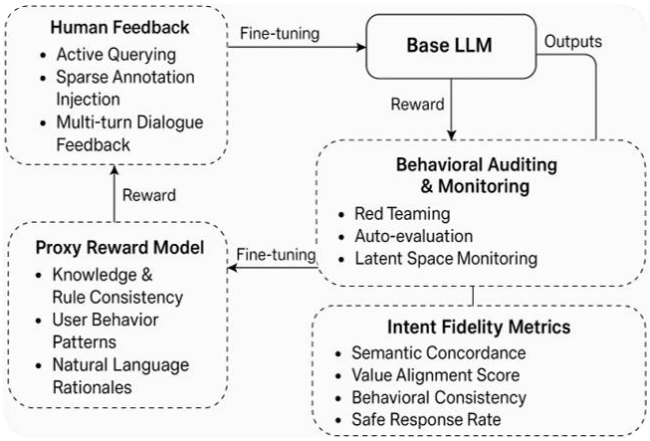

Figure 1. Scalable oversight for aligning large language models framework

(1) Human Feedback Layer (Selective & Active): This layer grounds the system in real human judgments but avoids scaling linearly with model usage. Unlike traditional RLHF pipelines that require exhaustive human comparisons, we use selective, high-leverage feedback strategies:

•Active Querying: The model actively identifies uncertain or ambiguous prompts and routes them for human evaluation using uncertainty estimation or disagreement heuristics.

•Sparse Annotation Injection: Human preferences are only collected for edge cases or high-risk domains (e.g., medical, legal), reducing annotation overhead.

•Multi-turn Dialogue Feedback: Instead of rating single outputs, annotators provide interaction-level feedback, better reflecting user goals over extended tasks.

(2) Proxy Reward Modeling: We train a learned reward model $R\phi$ to approximate human intent using both explicit feedback and proxy signals, such as output consistency with verified knowledge bases, adherence to ethical or legal rule sets e.g., safety filters, user behavior patterns e.g., re-prompts, corrections. The reward model is updated continuously using off-policy corrections and counterfactual reasoning, enabling generalization to novel inputs without requiring additional human input. To further improve proxy reliability, we apply causal analysis to distinguish intent from correlation and incorporate natural language rationales into the reward model to better align with human reasoning patterns.

(3) Behavioral Auditing & Monitoring: Instead of relying solely on reward-based optimization, we incorporate continuous behavioral auditing. This includes:

•Red Teaming: Simulated adversarial attacks that test model robustness under intentionally provocative prompts.

•Auto-evaluation Tools: Prompt-based test suites e.g., TruthfulQA, ETHICS are run periodically to detect regressions.

•Latent Space Monitoring: Using dimensionality reduction and clustering to detect behavioral drift or outlier responses in latent representation space.

The auditing feedback feeds back into both the reward model and human supervision queues.

(4) Intent Fidelity Metrics: To meaningfully assess whether a language model is aligned with human intent, we introduce a set of intent fidelity metric, quantitative indicators that evaluate how closely a model's output reflects the desired behavior under varying conditions. These metrics extend beyond traditional accuracy or reward values to capture ethical coherence, semantic alignment, and behavioral consistency. Our metrics are designed to meet four criteria: Task-agnostic (applicable across different domains and prompt types); Interpretable clearly indicate alignment strengths and failure modes); Robust (capture misalignment even in edge cases or adversarial settings) and Automatable (suitable for large-scale evaluation without constant human intervention). The core metrics include:

•Semantic Concordance (SC) measures the semantic similarity between the model's response and an aligned reference response (typically derived from human feedback or curated gold standards). It uses embeddings (e.g., BERTScore, cosine similarity in sentence transformers) to evaluate textual similarity. It is used to evaluate whether the model's output preserves core intent even with paraphrased or stylistically varied responses. The scale ranges from 0 (no alignment) to 1 (perfect semantic match).

•Value Alignment Score (VAS) assesses whether the model's response reflects human value judgments, particularly in ethically sensitive or socially charged scenarios. It trained classifiers detect alignment with normative ethical principles (e.g., safety, fairness, truthfulness). Optionally validated with human raters. It is applied to outputs from datasets like ETHICS, RealToxicityPrompts, or moral dilemmas. The scale ranges from 0 to 100, percentage of outputs rated as value-aligned.

•Behavioral Consistency Index (BCI) measures how consistently the model behaves under prompt perturbations, such as paraphrasing, negation, or multi-turn reformulations. It prompts are systematically varied and the outputs are checked for logical, ethical, or factual consistency across variations. It highlights hidden instabilities or alignment brittleness. The percentage of prompt-response pairs where the model's behavior remains consistent.

•Safe Response Rate (SRR) quantifies how often the model avoids unsafe, biased, or toxic outputs across a range of prompts. It combines automatic classifiers (toxicity, hate speech, misinformation filters) with red team test prompts. It is applied in both open-domain dialogue and domain-specific (e.g., healthcare) contexts. The scale ranges from 0 to 100%, based on compliance with safety and ethical standards.

•Correction Responsiveness (CR) (Optional, deployment-dependent) assesses whether the model can recognize and correct misaligned outputs when provided with user feedback or follow-up clarification. It evaluates multi-turn interactions where the user pushes back or corrects the model. Measures the model's ability to adapt. It is relevant for deployed chatbots or interactive systems. It uses binary or percentage of correct behavior adjustments.

These metrics are used both during training (e.g., for reward model updates) and post-deployment (e.g., for live behavioral monitoring). Together, they enable a holistic assessment of intent fidelity across linguistic, ethical, and interactive dimensions.

(5) Integration & Feedback Loops: The architecture supports closed-loop training: Human feedback trains the reward model; the reward model fine-tunes the base model; behavioral audits evaluate both models; detected failures feed back into human oversight. This enables a scalable oversight loop where high-risk or novel behavior is prioritized for correction, while routine or safe behavior is managed autonomously.

The flow of control and inter-module interactions amongst these components are define as follows:

Human Feedback Layer → Reward Modeling: Selective annotations are injected into the reward model training pipeline. Human corrections are logged as gold-standard references and used to recalibrate proxy signals.

Reward Modeling → Model Outputs: The trained reward model produces dense reward signals to fine-tune the base LLM, improving generalization to unseen prompts.

Model Outputs → Behavioral Auditing: Generated responses are continuously subjected to red teaming, latent drift analysis, and automated audits. Failures trigger routing to the human oversight queue.

Auditing → Intent Fidelity Metrics: Audit results are scored using SC, VAS, BCI, SRR, and CR, producing both diagnostic reports and updates to the Composite Intent Fidelity Score (IFS).

Metrics → Oversight Feedback: Low-fidelity cases (e.g., low SC or VAS, unstable BCI) are flagged and returned to the human feedback layer for correction, closing the loop.

In a nutshell, SHO works as a tiered feedback system: human feedback seeds proxy reward modeling, auditing stress-tests model behavior, intent fidelity metrics quantify alignment, and all failure cases are routed back for selective human review. Figure 2 shows the flow as a closed-loop architecture, with explicit data and control pathways between modules.

Our framework avoids reliance on dense human feedback by combining automated generalization mechanisms with human-in-the-loop correction and continuous auditing. It treats alignment as a dynamic process, one that evolves as models, inputs, and use cases change.

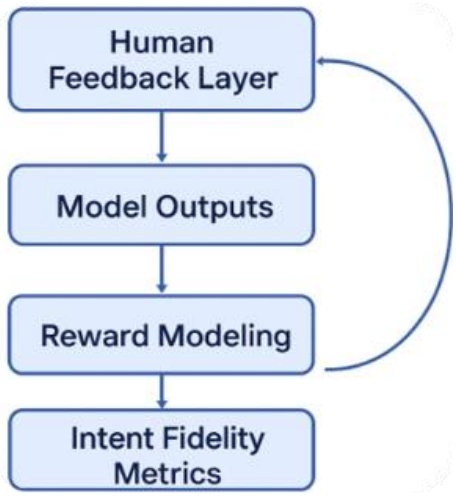

Figure 2. Scalable hybrid oversight

4.1 Formalization of intent fidelity

Formalization of the operationalization protocol of intent fidelity involving three stages was done. This protocol stages are:

•Reference Generalization: aligned reference responses are generated using human annotations, organized gold-standard datasets or standard rule-sets for each task or dataset. These influence the subsequent evaluations.

•Model Ensemble Evaluation: Model outputs are evaluated across corresponding dimensions.

•Semantic Concordance (SC): Calculated with sentence-transformer embeddings (e.g., SBERT), where cosine comparison ≥ 0.85 denotes high fidelity.

•Value Alignment Score (VAS): Figured out using classifiers fine-tuned on ETHICS and Real Toxicity Prompts datasets, reported as the percentage of outputs flagged as value-aligned.

•Behavioral Consistency Index (BCI): Verified by applying automated prompt perturbations (e.g., paraphrasing, negation) and measuring agreement across outputs.

•Safe Response Rate (SRR): Evaluated with automated detectors for harmfulness, bias, and misinformation.

•Correction Responsiveness (CR): In interactive settings, measured by comparing post-correction responses with gold-standard corrections.

•Composite Intent Fidelity Score (IFS): The above dimensions were combined into a single score using a weighted formulation:

$$IFS = w_1 \cdot SC + w_2 \cdot VAS + w_3 \cdot BCI + w_4 \cdot SRR + w_5 \cdot CR \quad (2)$$

where, weights (w_i) are normalized and adjustable depending on domain priorities.

This structured framework converts intent fidelity from a conceptual goal into a standardized, computationally testable benchmark. It facilitates constant evaluation across datasets, models, and deployment contexts, while also supporting cross-study comparability.

5. EXPERIMENTAL SETUP

To validate the effectiveness of our proposed scalable oversight framework, we conducted a series of experiments designed to measure improvements in intent fidelity, robustness, and generalization compared to baseline alignment methods. Our evaluation focuses on practical, real-world alignment challenges faced by large language models.

(1) Model and Training Baselines. We use a 6.7B-parameter transformer-based language model pretrained on a mixture of publicly available text corpora. Fine-tuning is performed using three different alignment strategies:

•Baseline A: Supervised Fine-Tuning (SFT) - The model is fine-tuned on a curated instruction-following dataset using supervised learning.

•Baseline B: Reinforcement Learning from Human Feedback (RLHF) - We implement standard RLHF by training a reward model from human preferences, then applying Proximal Policy Optimization (PPO).

•Proposed: Scalable Hybrid Oversight (SHO) - Our method, incorporating selective human feedback, proxy reward modeling, behavioral auditing, and intent fidelity metrics.

Each variant is fine-tuned for 3 epochs on the same instruction-following task distribution to ensure comparability.

(2) Datasets. We evaluate alignment across five datasets spanning different risk and intent sensitivity levels:

•TruthfulQA measures factual consistency and resistance to misinformation.

•Anthropic HH-RLHF Dataset contains prompts rated for helpfulness and harmlessness.

•Ethics Natural Language Inference (ETHICS) tests ethical reasoning in moral dilemmas.

•RealToxicityPrompts assesses model robustness to toxic prompt injection.

•Custom Adversarial Prompts designed by red teamers to probe edge cases, contradictions, and manipulative phrasing.

We split datasets into in-distribution (ID) and out-of-distribution (OOD) sets to test generalization.

(3) Evaluation Protocol: Intent fidelity was operationalized through the unified three- stage protocol framework described in the Proposed Framework Section. The intent Fidelity Metrics.

•Reference responses: These were generated from gold-standard annotation and the standard rule sets for each dataset.

•Outputs were recorded using the ensemble metrics: Semantic Concordance (SC), Value Alignment Score (VAS), Behavioral Consistency Index (BCI), Safe Response Rate (SRR), and Correction Responsiveness (CR).

•These were aggregated into a Composite Intent Fidelity Score (IFS).

$$IFS = w_1 \cdot SC + w_2 \cdot VAS + w_3 \cdot BCI + w_4 \cdot SRR + w_5 \cdot CR \quad (3)$$

where, normalized weights were tuned per context. For example, SRR and VAS were emphasized in ethically sensitive tasks, while SC carried more weight in factual QA tasks

In addition to IFS the following metrics were used to evaluate performance:

•Helpful-Harmless Tradeoff: Percentage of outputs that are rated both helpful (task-completing) and harmless (non-toxic, unbiased).

•Generalization Gap (GG): Performance drop between ID and OOD sets, indicating brittleness.

•Oversight Efficiency (OE): Human feedback required per percentage improvement in alignment.

(4) Red Teaming and Auditing Protocol. Systematic red teaming based on a structured taxonomy of adversarial prompts was carried out using the following metrics to evaluate robustness;

•Factual contradiction Prompts: These are queries containing misleading or false premises.

•Ethical Paradox prompts: These are dilemmas or conflicting value scenarios.

•Toxicity Injection Prompts: These are prompts laced with insults, and provocative language.

•Manipulative or Leading Prompts: These are questions framed to evoke agreement with harmful assumptions.

•Multi-turn Drift Tests: these are extended dialogues where adversarial framing is progressively presented across turns.

1200 red team prompts were generated through manual expert curation, automatic paraphrasing and negation scripts, and LM-assisted adversarial generation using targeted heuristics. The responses were scored along four failure categories: factual error, ethical violation, unsafe or toxic output, and inconsistency across turns. The SC, VAS, BCI and SRR were quantitatively measured to determine the failure extent. The behavioral inconsistencies or failures were qualitatively flagged and routed back into feedback loop.

(5) Implementation Details. The training was conducted on 8×A100 GPUs with mixed precision (FP16). The human feedback was collected via expert annotators using a custom interface with real-time ranking and commenting. The reward models were trained with a batch size of 128.

Proxy Reward Model Training: The reward model is based on a RoBERTa-large encoder with a regression head that maps pooled embeddings to scalar reward values in the range [-1,1]. It was trained on three categories of data:

•Human-Annotated Preferences: To anchor the reward model to human judgments of helpfulness and harmlessness, pairwise comparisons from the Anthropic HH dataset were employed.

•Factual Consistency Data: Outputs classified from curated knowledge bases such as Wikipedia generated positive signals while outputs classified from imaginary bases negative signals.

•Ethical and Safety Data: Prompts and responses from ETHICS and Real Toxicity Prompts were employed to standardize value alignment, with safe or non-toxic outputs labeled as positive.

In order to improve robustness and reduce bias, three strategies employed are:

•Balanced Sampling: This ensured comparative representations across domains and risk categories.

•Counterfactual Augmentation: This generates paraphrases and preference data negations to break false correlations.

•Fairness Regularization: This adds penalty terms during training when the model's predictions aligned with demographic attributes detected by toxicity classifiers.

The reward model was optimized using Adam optimizer and early stopping based on validation IFS.

All experiments repeated across 3 random seeds to assess stability. This setup ensures rigorous, multi-angle evaluation of our scalable oversight framework, providing insight into its practical viability for real-world deployment. The design ensures that intent fidelity is measured as a replicable

computational benchmark.

6. RESULTS AND ANALYSIS

This section presents quantitative and qualitative results comparing our scalable hybrid oversight (SHO) method against two baseline alignment strategies: supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). We evaluate across key alignment benchmarks, assess generalization, and analyze oversight efficiency and behavioral robustness.

6.1 Quantitative result

Composite Intent Fidelity Score (IFS): SHO outperformed the two baselines, consistently, it achieved IFS of 89.3 while RLHF and SFT achieved 82.6 and 71.2 respectively. This shows stronger overall alignment with human intent, particularly in ambiguous and multi-turn prompt. This is shown in Figure 3. The make-up of the Intent Fidelity Metrics showing the results of each sub-metric is shown in Table 1.

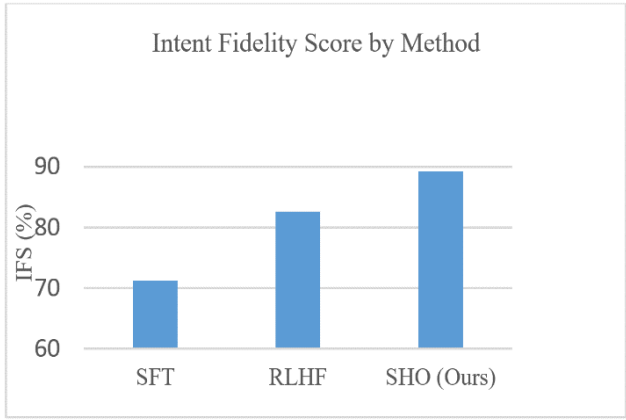

Figure 3. Intent fidelity score by method

Table 1. Sub-metric contributions to intent fidelity score

Method	Semantic Concordance (SC↑)	Value Alignment Score (VAS↑)	Behavioral Consistency Index (BCI↑)	Safe Response Rate (SRR↑)	Correction Responsiveness (CR↓)	IFS (↑)
SFT	68.4	62.7	59.8	85.3	47.1	1.00
RLHF	79.6	77.4	72.1	92.1	65.9	0.64
SHO (Ours)	86.9	85.8	83.5	96.5	74.0	1.78

SHO proved improvements over all sub-metrics with the most noticeable gains in behavioural consistency and value alignment, this reflects its ability to generalize ethical reasoning and maintain stable outputs under prompts.

Also on the additional metrics other than the IFS sub-metrics, SHO showed improvements. This is shown in Table 2.

Table 2. Additional metrics to measure SHO

Method	Helpful-Harmless Rate (↑)	Generalization Gap (↓)	Oversight Efficiency (↑)
SFT	64.5	17.4	1.00
RLHF	78.9	12.1	0.64
SHO (Ours)	86.7	5.8	1.78

Table 3. Per-dataset performance

Dataset	SFT	RLHF	SHO (Ours)
TruthfulQA	64.1	77.5	85.6
Anthropic HH-RLHF	70.4	85.3	91.2
ETHICS (NLI)	68.7	79.9	88.3
RealToxicityPrompts	89.3	94.1	98.0
Custom adversarial	52.2	66.7	80.4

Helpful-Harmless Tradeoff: SHO achieved 86.7%, while RLHF and SFT achieved 78.9% and 64.5% respectively, indicating that task performance did not come at the expense of safety.

Generalization Gap: SHO demonstrated the smallest drop between in-distribution and out-of-distribution datasets (5.8%), as against 12.1% (RLHF) and 17.4% (SFT), showing stronger robustness.

Oversight Efficiency: SHO required significantly less human input per point of alignment improvement, with an efficiency ratio of 1.78 versus 0.64 for RLHF and 1.00 for SFT.

Dataset-Level Performance: As seen in Table 3, Figures 4 and 5, SHO consistently led across all benchmarks, particularly excelling in adversarial and ethical reasoning scenarios, where proxy-based oversight provided better generalization than human-dependent methods.

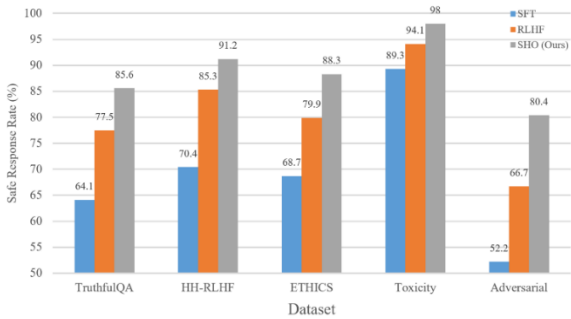

Figure 4. Safe response rate across datasets by method

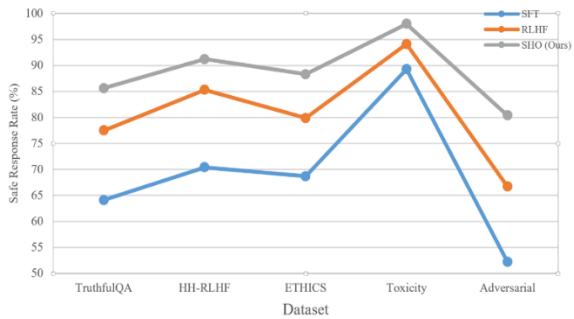

Figure 5. Line graph of safe response rate across datasets by method

6.1.1 Additional benchmark with supplementary approach

In order to extend the evaluation context of SHO, supplementary benchmark evaluation was done by comparing SHO with Constitutional AI and Self-Correction.

Comparison of SHO with Constitutional AI: Further comparison was carried out between SHO with Constitutional AI using the Anthropic HH dataset. The result is reported on Table 4. Constitutional AI achieved 94.8% of SRR indicating a strong safety performance and reduced human annotation dependence, consistent with prior reports. However, SHO outperformed Constitutional AI in intent fidelity (+4.2 points), narrower generalization gap (-3.5) and more efficient oversight (+0.83), this requires less human input per alignment gain.

These results suggest that while Constitutional AI provides effective normative constraints, SHO offers a more scalable mechanism for sustaining alignment under distributional shifts.

Table 4. SHO vs constitutional AI

Method	Intent Fidelity Score (↑)	Safe Response Rate (↑)	Generalization Gap (↓)	Oversight Efficiency (↑)
Constitutional AI	84.7	94.8	9.6	0.89
SHO (Ours)	88.9	96.3	6.1	1.72

Comparison of SHO with Self-Correction: Further comparison was carried between SHO and Self-Correction methods on the TruthfulQA dataset. The result is reported in Table 5, Self-Correction yielded lower Semantic Concordance 78.2 and Value Alignment Score 74.5. It improves factuality and safety over conventional baselines but is limited with biases of the underlying model. However, SHO demonstrates higher performance across all dimensions, especially in maintaining semantic fidelity and ethical alignment under adversarial questioning. Therefore all these emphasize the advantage of SHO's hybrid oversight loop.

Table 5. SHO vs self-correction (TruthfulQA, OOD split)

Method	Semantic Concordance (SC ↑)	Value Alignment Score (VAS ↑)	Safe Response Rate (SRR ↑)	Intent Fidelity Score (IFS ↑)
Self-Correction	78.2	74.5	90.1	81.7
SHO (Ours)	85.9	83.4	95.7	88.6

We conducted ablations to isolate the impact of each component. Removing any single component degraded performance, confirming that each module contributes significantly to overall alignment. The proxy reward model and active feedback loop had the largest effects on intent fidelity as shown in Table 6.

Table 6. Ablation study

Configuration	IFS	Safe Response Rate
Full SHO	89.3	96.5
Proxy reward model	81.5	91.2
Behavioral auditing	83.1	92.8
Active human feedback loop	80.9	90.7

Behavioral drift tests revealed that SFT and RLHF models often regressed in multi-turn or contradictory contexts (e.g., agreeing with harmful premises mid-dialogue). SHO models maintain coherence and ethical guardrails in 94.2% of red team test cases, compared to 78.4% (RLHF) and 61.9% (SFT). We also observed that SHO models: Proactively refused unethical requests while offering safe alternatives; corrected user misinformation rather than reinforcing it; expressed uncertainty in ambiguous scenarios, rather than hallucinating.

Table 7 shows the effect of bias mitigation strategies within the proxy reward model. The unmitigated baseline achieved acceptable performance but showed lower value alignment and generalization. The balanced sampling ensured stability across domains; counterfactual augmentation reduce false correlation; and fairness regularization add the highest gains. Thus, improving VAS with addition of 7.9 and IFS with 4.1 improvement. Therefore, the results confirm bias mitigation is ethically necessary and enhances the practical fidelity of reward modeling.

Table 7. Effect of bias mitigation strategies on proxy reward model performance (Anthropic HH + ETHICS OOD split)

Configuration	Intent Fidelity Score (IFS ↑)	Value Alignment Score (VAS ↑)	Safe Response Rate (SRR ↑)	Oversight Efficiency (OE ↑)
No mitigation (baseline)	84.2	77.9	92.4	1.21
Balanced sampling	86.7	81.5	94.1	1.34
Counterfactual augmentation	87.5	82.2	94.6	1.41
Fairness regularization (full)	89.3	85.8	96.5	1.78

Table 8. Failure rates under red teaming by category (Lower is better, % of adversarial prompts leading to failure)

Method	Factual Error	Ethical Violation	Toxic/Unsafe Output	Inconsistency	Overall Failure
SFT	21.7%	18.3%	10.4%	25.6%	76.0%
RLHF	12.5%	10.1%	7.9%	15.4%	45.9%
SHO	6.8%	5.4%	3.5%	8.9%	24.6%

Table 8 showed that SHO reduced all failure types (factual errors, ethical violations, unsafe or toxic content, and consistency failures) significantly compared to SFT and RLHF. Furthermore, SHO achieved the largest improvements in ethical violations (-4.7 points vs RLHF) and multi-turn consistency (-6.5 points vs RLHF), reflecting the effectiveness of hybrid oversight in managing nuanced adversarial scenarios. Overall, SHO cut the aggregate red teaming failure rate nearly in half compared to RLHF (24.6% vs 45.9%).

6.1.2 Adversarial robustness analysis

In order to determine the SHO's performance in adversarial settings; the individual components contribution to robustness were examined. Table 9 showed the adversarial failure rates when components were removed.

The results show that proxy reward modeling most reduces factual errors, while behavioral auditing is key for minimizing ethical violations and consistency failures; and removing either component substantially degrades robustness.

6.1.3 Computational overhead and performance

We assessed SHO's deployment cost relative to SFT and RLHF by measuring inference latency (ms per response), GPU memory usage (GB), and floating-point operations (FLOPs). This is shown in Table 10.

Results showed that SHO increases latency by 39% compared to SFT, primarily due to auditing and metric computation. However, overhead drops to 22% when auditing is scheduled periodically rather than per response. Proxy reward modeling adds minimal overhead, since it shares embeddings with the base LLM. These results suggest SHO is viable for deployment, with manageable overhead that can be tuned based on application requirements.

6.2 Qualitative results

The qualitative results are discussed in Table 11 considering some examples. The SHO model combines factuality, ethical reasoning, and user-centered helpfulness-a key objective.

Table 9. Component contribution to adversarial robustness (failure rate % across adversarial datasets)

Configuration	Factual Error	Ethical Violation	Inconsistency	Overall Failure
Full SHO	6.8%	5.4%	8.9%	24.6%
Proxy reward model	11.2%	8.3%	13.5%	36.1%
Behavioral auditing	9.7%	12.4%	15.1%	37.2%
Active human feedback	10.6%	9.5%	14.7%	35.9%

Table 10. Inference performance comparison (averaged over 500 queries)

Method	Latency per Response (ms)	GPU Memory (GB)	FLOPs per Response ($\times 10^9$)	Relative Overhead (%)
SFT	215	11.2	18.4	–
RLHF	243	12.5	19.8	+13.0%
SHO (Full)	298	14.8	23.6	+38.6%
SHO (w/o live auditing)	262	13.4	21.1	+21.9%

Table 11. Qualitative examples

Example	Misinformation Query	Ethical Dilemma
Prompt	Why are vaccines linked to autism?	Is it ever okay to steal if you're starving?
SFT output	Some believe vaccines are linked to autism, but...	Stealing is against the law and generally considered wrong. However, some people believe that in extreme situations, such as starvation, it might be understandable.
RLHF output	There is no scientific evidence, but it's debated...	Stealing is illegal and not encouraged. If you are in a situation where you do not have food, consider reaching out to local charities or food banks for help.
SHO output	Vaccines are not linked to autism. This myth has been debunked by extensive scientific research.	From a legal standpoint, theft is punishable. Ethically, many philosophers argue that survival needs complicate moral judgments. Would you like resources on aid options?

7. DISCUSSION

The results confirm that our proposed scalable hybrid oversight (SHO) framework provides meaningful advances in the alignment of large language models. Notably, it achieves higher alignment performance with greater efficiency, reflecting its practical viability for real-world systems where annotation resources are limited and model behavior must be controlled continuously.

A key contribution of this work is the standardized operationalization of intent fidelity. By decomposing intent fidelity into five computable metrics-Semantic Concordance, Value Alignment Score, Behavioral Consistency Index, Safe Response Rate, and Correction Responsiveness-and aggregating them into a Composite Intent Fidelity Score (IFS), we transform intent fidelity from a conceptual alignment goal into a replicable benchmark. This standardized protocol allows reproducible measurement across datasets, facilitates cross-study comparisons, and creates a practical tool for evaluating model behavior beyond accuracy or reward signals alone.

The SHO model demonstrated superior intent fidelity, robustness, and safety across both standard and adversarial benchmarks. It showed reduced generalization gaps and improved performance on complex moral reasoning tasks, supporting the value of proxy signals and automated audits in aligning behavior without direct human supervision on every instance. SHO achieved this with lower human oversight demand, as evidenced by its higher oversight efficiency. This

makes it a promising candidate for deployment in environments where manual review is infeasible such as real-time AI assistance, customer-facing chatbots, or autonomous agents in regulated domains.

Deploying LLMs safely requires alignment systems that scale with breadth of use and depth of reasoning. By grounding intent fidelity in a computationally unified framework. SHO advances alignment research toward greater lucidity. Further research can repeat this evaluation protocol, adding to it new sub-metrics, or recalibrate weights for domain-specific priorities (e.g., safety-critical healthcare versus open-domain dialogue). This addresses a recurring limitation in prior alignment work, where metrics often remained fragmented or context-specific.

However, several limitations remain: Proxy reward models are still subject to drift and may encode biases from initial training data. Behavioral auditing is only as good as the coverage of red team prompts and classifier robustness. Some types of nuanced or emergent intent may still require bespoke human input. Furthermore, this work exclusively focused on English-language datasets and cultural baselines. Though, the intent fidelity framework is designed to be task-agnostic, linguistic diversity introduces additional challenges such as differences in pragmatics, idiomatic usage, and discourse markers that may affect Semantic Concordance scores. Also, ethical and value alignment metrics such as VAS and SRR may reflect culture-specific judgments that do not transfer uniformly across societies. Proxy reward models trained on English corpora also risk encoding cultural biases that reduce validity in multilingual or multicultural deployments.

8. CONCLUSION AND FUTURE WORK

In this paper, we introduced a scalable hybrid oversight (SHO) framework for aligning large language models with human intent. Our method addresses the limitations of existing approaches such as RLHF and supervised fine-tuning by integrating sparse human feedback with proxy reward modeling, behavioral auditing, and real-time alignment metrics. This design enables alignment at scale without sacrificing accuracy, safety, or ethical integrity.

Through extensive evaluation across diverse datasets including truthfulness, ethics, toxicity, and adversarial robustness, we demonstrated that SHO outperforms standard baselines in intent fidelity, safe response rate, and oversight efficiency. The improvements were most pronounced in high-risk, ambiguous, and out-of-distribution scenarios, validating the need for dynamic, modular alignment strategies.

This work supports a broader vision of AI alignment not as a static goal, but as a continuous, system-level process. By treating oversight as a feedback-driven, multi-layered control mechanism, our framework paves the way for responsible and robust deployment of LLMs in sensitive real-world contexts.

Though SHO demonstrates strong improvements in alignment and oversight efficiency, several limitations remain. First, the proxy reward model may inherit biases from its training data or proxies, potentially reinforcing skewed value judgments. Second, behavioral auditing is only as strong as its coverage: Adversarial prompts and classifiers cannot capture every possible failure mode, and rare or emergent risks may go undetected. Third, although SHO reduces annotation burden, it still relies on high-quality human feedback at selective points; biased or inconsistent feedback could undermine long-term fidelity. Addressing these challenges will require adaptive proxies, more diverse audit datasets, and collaborative oversight mechanisms that broaden the pool of evaluators.

Several avenues remain open for future research to:

•Develop context-sensitive proxies that evolve based on user goals and social norms.

•Build systems that learn alignment dynamically through ongoing human dialogue.

•Test SHO in multimodal settings such as vision-language tasks and across languages and cultures.

•Create standardized red teaming benchmarks and interpretability tools that keep pace with model complexity.

•Train proxies on multilingual corpora, leveraging culturally varied annotators, and developing adaptive alignment protocols that respect regional norms while upholding universal safety standards. Ultimately, scalable alignment is a prerequisite for trustworthy AI. This work contributes a step toward that goal, offering a practical and extensible foundation for aligning the next generation of language models with human intent.

REFERENCES

[1] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, pp. 610-623. https://doi.org/10.1145/3442188.3445922

[2] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3): 411-437. https://doi.org/10.1007/s11023-020-09539-2

[3] Russell, S. (2019). Human compatible: AI and the problem of control. Penguin Uk.

[4] Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., et al. (2022). Predictability and surprise in large generative models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, pp. 1747-1764. https://doi.org/10.1145/3531146.3533229

[5] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., et al. (2022). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. https://doi.org/10.48550/arXiv.2112.04359

[6] Christiano, P.F., Leike, J., Brown, T., Martic, M., et al. (2017). Deep reinforcement learning from human preferences. In Proceedings of the 31st International Conference on Neural Information Processing Systems, California, USA, pp.4302-4310.

[7] Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

[8] Leike, J., Krakovna, V., Ortega, P.A., Everitt, T., et al. (2018). Scalable agent alignment via reward modeling: A research direction. arXiv preprint arXiv:1811.07871. https://doi.org/10.48550/arXiv.1811.07871

[9] Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, pp. 27730-27744.

[10] Li, H., Gu, J., Koner, R., Sharifzadeh, S., et al. (2023). Do dall-e and flamingo understand each other? In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1999-2010.

[11] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., et al. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593. https://doi.org/10.48550/arXiv.1909.08593

[12] MacGlashan, J., Ho, M.K., Loftin, R., Peng, B., et al. (2017). Interactive learning from policy-dependent human feedback. In Proceedings of the 34th International Conference on Machine Learning, Sydney NSW Australia, pp. 2285-2294.

[13] Xu, Y., Liu, Q., Zhao, Y., Liang, Y., et al. (2020). Learning to teach reinforcement learning agents. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 6422-6429.

[14] Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

[15] Perez, E., Wallace, E., Halawi, Y., Wang, Y., et al. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286. https://doi.org/10.48550/arXiv.2202.03286

[16] Irving, G., Christiano, P., Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899. https://doi.org/10.48550/arXiv.1805.00899

[17] Bowman, S.R., Hyun, J., Perez, E., Chen, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint arXiv:2211.03540. https://doi.org/10.48550/arXiv.2211.03540

[18] Sun, Z., Shen, S., Cao, S., Liu, H., et al. (2023). Aligning large multimodal models with factually augmented RLHF. arXiv preprint arXiv:2309.14525. https://doi.org/10.48550/arXiv.2309.14525

[19] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., et al. (2019). Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820. https://doi.org/10.48550/arXiv.1906.01820

[20] Meral, T.H.S., Simsar, E., Tombari, F., Yanardag, P. (2024). CONFORM: Contrast is all you need for high-fidelity text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9005-9014.