ISSN: 2519-9439

Preference-Shift Diagnostics for Iterative Policy Refinement

Ricardo Salazar

School of Public Health, Faculty of Medical Sciences, Universidad Nacional de Córdoba, Córdoba, Cordoba, Argentina

Abstract

The deployment of autonomous systems in dynamic, human-centric environments necessitates continuous adaptation to evolving user preferences. Iterative policy refinement, a cornerstone of modern reinforcement learning and alignment strategies, often assumes stationary preference distributions. However, human preferences are inherently non-stationary, exhibiting complex temporal, contextual, and systemic shifts that degrade policy performance over time. This paper introduces a comprehensive framework for preference-shift diagnostics designed to identify, categorize, and address non-stationarity in iterative policy refinement pipelines. By establishing a robust diagnostic methodology, this research bridges the gap between static reward modeling and dynamic operational environments. We propose a multi-layered diagnostic architecture that continuously monitors feedback streams, detects distributional divergences, and attributes these shifts to specific environmental or human factors. Through extensive theoretical analysis and simulated experimental validation, we demonstrate that integrating preference-shift diagnostics significantly enhances the resilience and alignment of autonomous policies. The findings underscore the critical need for diagnostic transparency in policy optimization, providing a scalable solution for maintaining system efficacy in the face of shifting human expectations.

Keywords: Iterative Policy Refinement, Preference Dynamics, Diagnostic Frameworks, Reinforcement Learning.

1.Introduction

1.1 The Imperative of Iterative Policy Refinement

The rapid integration of autonomous systems into daily human life has fundamentally altered the landscape of artificial intelligence research. Systems ranging from autonomous vehicles and personalized recommendation engines to advanced robotic assistants require complex decision-making capabilities that align closely with human expectations. Iterative policy refinement has emerged as a primary methodological approach to achieve this alignment. By continuously updating a behavioral policy based on evaluative feedback, systems can theoretically converge on optimal behaviors that satisfy user constraints. This process is deeply rooted in the principles of reinforcement learning, where an agent interacts with an environment, receives feedback in the form of rewards, and updates its internal parameters to maximize cumulative future utility. Early formulations of this problem relied heavily on hand-crafted reward functions designed by engineers. However, it quickly became apparent that defining comprehensive reward functions for open-ended, complex tasks was intractable. Consequently, the field shifted towards learning reward functions directly from human feedback.

ISSN: 2519-9439

This paradigm relies on an iterative cycle: the agent acts, the human evaluates the action by providing feedback such as rankings or corrections, the system updates its internal reward model based on this feedback, and the policy is subsequently refined to maximize the newly updated reward model. Over multiple iterations, the policy is expected to closely approximate the desired human behavior. This methodology has been instrumental in training highly capable models, particularly in natural language processing and complex continuous control tasks. However, the fundamental assumption underpinning the success of standard iterative policy refinement is the stationarity of the underlying preference distribution. It assumes that what a human user considers good or bad remains consistent throughout the training process and into deployment. In reality, human preferences are profoundly dynamic, heavily influenced by context, fatigue, increased familiarity with the system, and evolving external conditions. When a system assumes stationary preferences in a non-stationary reality, the iterative refinement process can become unstable, leading to policy degradation, safety risks, and the phenomenon known as reward hacking, where the agent exploits outdated preference models to achieve high scores without delivering actual value to the user.

1.2 The Challenge of Preference Shifts

Preference shifts represent one of the most significant hurdles in the long-term deployment of adaptive autonomous systems. A preference shift occurs when the mapping between an agent's state-action trajectory and the human evaluator's latent utility function changes over time. These shifts can manifest in numerous ways. Temporal shifts happen gradually as users become accustomed to an autonomous system; a user who initially prioritizes cautious behavior in an autonomous driving assistant may, over time, prefer efficiency and speed as their trust in the system grows. Contextual shifts occur when external environmental variables change the evaluation criteria; a recommendation system suggesting leisure content might suddenly need to prioritize educational or crisis-related information during a global event. Furthermore, contradictory feedback can arise from multi-stakeholder environments where different users interact with the same system, each bringing their own evolving set of criteria. When a policy refinement algorithm encounters a preference shift without a diagnostic mechanism to identify it, the system typically experiences conflicting gradients. The historical data stored in the replay buffers or historical datasets suggests one optimal direction for the policy update, while the newly acquired feedback suggests another. Standard optimization algorithms, attempting to minimize the empirical risk across all available data, often end up learning a compromised policy that satisfies neither the old nor the new preferences. Worse still, if the learning rate is aggressively tuned to prioritize recent data, the phenomenon of catastrophic forgetting can occur, where the system entirely unlearns foundational safety constraints while over-indexing on a temporary, transient preference shift. Diagnosing these shifts is extraordinarily difficult because a drop in performance could be attributed to multiple factors. It could be a genuine preference shift, but it could also be a failure in the policy optimization algorithm, a shift in the environmental dynamics, or simply a degradation in the quality of human feedback due to evaluator fatigue. Without a dedicated diagnostic framework to isolate and identify preference shifts, system designers are left blind, unable to determine why an iteratively refined policy is failing to meet user expectations.

1.3 Objectives and Contributions

The primary objective of this paper is to formally conceptualize and develop a comprehensive diagnostic framework specifically tailored for identifying and resolving preference shifts in iterative policy refinement pipelines. We argue that robust alignment cannot be achieved merely through more advanced optimization algorithms; it requires a deep, diagnostic understanding of the data generating process, specifically the evolving nature of human feedback.

ISSN: 2519-9439

This paper makes several key contributions to the field of general autonomous system alignment. We provide a rigorous taxonomy of preference shifts, categorizing them by their temporal dynamics, underlying causes, and impacts on the reward modeling process. We introduce a novel, multi-layered diagnostic architecture that operates in parallel with the standard iterative refinement loop, utilizing statistical divergence metrics to continuously monitor the health and stability of the incoming feedback streams. Furthermore, we outline specific attribution mechanisms that allow system operators to differentiate between genuine preference evolution and artifactual feedback noise. Finally, through extensive simulated experimentation, we demonstrate that coupling policy refinement algorithms with our proposed preference-shift diagnostics not only accelerates the adaptation to new preferences but also significantly preserves baseline system safety and reliability during transitional periods. By laying the theoretical and practical groundwork for preference-shift diagnostics, this research aims to facilitate the development of long-term, highly adaptive autonomous systems capable of thriving in complex, human-centric environments.

2. Background and Literature Review

2.1 Evolution of Policy Optimization

The foundational concepts of policy optimization trace back to the formulation of Markov Decision Processes, where an agent seeks to find a policy that maximizes the expected sum of discounted rewards. Early research predominantly focused on environments where the reward function was perfectly known and stationary [1]. Algorithms such as dynamic programming and foundational temporal difference learning were designed under the strict assumption that the transition dynamics and the reward signals were immutable properties of the environment [2]. As the ambition of autonomous system deployment grew, researchers began tackling environments with high-dimensional state spaces, leading to the integration of deep neural networks as function approximators for both value functions and policies [3]. Despite these advancements in handling complex state representations, the reliance on static, programmatic reward functions remained a significant bottleneck. It was well documented that specifying a reward function for a task as seemingly simple as a robot grasping an object often resulted in unintended behaviors, as agents found loopholes in the mathematical specification of the reward [4]. This misalignment highlighted the limitations of static objective functions. To address this, the field explored inverse reinforcement learning, where the agent attempts to infer the underlying reward function by observing expert demonstrations. While inverse reinforcement learning provided a pathway to learn from human behavior rather than programmatic rules, it still implicitly assumed that the expert's behavior was generated by a single, stationary, and optimal reward function. The computational complexity of inverse reinforcement learning and its sensitivity to sub-optimal demonstrations necessitated the development of more scalable and robust methodologies for preference acquisition.

2.2 Reward Modeling and Human Feedback

The shift towards explicitly modeling human preferences marked a significant evolution in the field. Instead of relying solely on demonstrations, researchers proposed querying human evaluators directly, asking them to compare different trajectories generated by the agent [5]. This pairwise comparison approach leverages the Bradley-Terry model to estimate a parameterized reward function that accurately reflects the probability of a human preferring one trajectory over another [6]. This methodology significantly reduced the burden on human evaluators, as expressing a preference between two options is generally easier and less prone to systemic error than providing a continuous, cardinal reward signal or a perfect demonstration [7]. Iterative policy refinement builds directly upon this reward modeling paradigm. In an iterative setup, the agent explores the environment, generates a diverse set of trajectories, and queries the human evaluator for preferences on a subset of these trajectories. A reward model is trained on this feedback, and the agent's policy is subsequently optimized

ISSN: 2519-9439

against this reward model [8]. This cycle repeats, allowing the agent to continuously refine its behavior. The success of this approach has been demonstrated across various domains, highlighting its flexibility and scalability [9]. However, the literature surrounding reward modeling heavily relies on the assumption that the data collected over multiple iterations can be aggregated safely. Current state-of-the-art frameworks typically utilize large historical buffers of human feedback to train robust reward models, explicitly assuming that a preference provided in iteration one is perfectly compatible with a preference provided in iteration one hundred.

2.3 Dynamics of Preference Shifts

The assumption of stationarity in human feedback has been increasingly challenged by recent theoretical and empirical studies. Researchers have categorized non-stationarity in reinforcement learning primarily into environmental shifts and transition dynamic shifts, but only recently has attention turned to the non-stationarity of the objective function itself [10]. Preference shifts can be driven by a multitude of factors. Temporal evolution is the most intuitive; as users interact with a system, their baseline expectations recalibrate [11]. A user might initially reward an agent simply for successfully completing a task, but later penalize the agent if the task is not completed with a specific aesthetic or energetic efficiency [12].

Contextual preference shifts are another critical vector of non-stationarity. The latent utility function of a user is often conditioned on unobserved contextual variables. For example, the desired behavior of a smart home climate control system depends heavily on whether the user is hosting guests, exercising, or sleeping. If the system cannot observe these contexts directly, the resulting feedback appears contradictory and non-stationary. Furthermore, multi-agent or multi-user environments inherently generate preference shifts from the perspective of the learning agent. If control of the system passes from a risk-averse user to a risk-tolerant user, the sudden change in feedback gradients can severely destabilize the learning process [13]. Understanding these dynamics is paramount because applying traditional stationary learning algorithms to non-stationary preference data results in reward models that smooth over the nuances of the new preferences, severely limiting the policy's ability to adapt.

2.4 Existing Diagnostic Approaches

Despite the growing recognition of preference shifts, dedicated diagnostic frameworks remain sparse in the literature. Most existing approaches address non-stationarity reactively rather than diagnostically. One common method involves utilizing exponential moving averages or strict time-based windowing for the feedback data, essentially forcing the reward model to only consider recent data [14]. While this allows the system to adapt to new preferences, it sacrifices the stability and generalization capabilities provided by larger, diverse datasets [15]. Furthermore, windowing provides no diagnostic insight into why the data is being discarded or whether a genuine shift has occurred. Other approaches have explored meta-learning techniques to rapidly adapt to new reward functions [16]. These systems are trained to expect changes in the environment and can quickly recalibrate their policies. However, meta-learning approaches are often computationally expensive and do not provide human-interpretable diagnostics regarding the nature of the preference shift. In safety-critical applications, it is not enough for an agent to simply adapt; system operators must understand what the agent is adapting to and why. The lack of explicit diagnostic tools means that developers cannot distinguish between a change in user intent, a degradation in the user interface leading to noisy clicks, or an adversarial attack on the feedback mechanism. Therefore, there is a critical need for a structured methodology that explicitly diagnostics preference shifts, categorizes their nature, and informs the iterative refinement process in a transparent and controlled manner.

ISSN: 2519-9439

3. Methodology for Preference-Shift Diagnostics

3.1 Architectural Overview

To address the limitations of existing reactive methods, we propose a comprehensive preference-shift diagnostic architecture designed to run continuously in parallel with the iterative policy refinement loop. The core philosophy of this architecture is the decoupling of preference data acquisition from immediate reward model optimization. Instead of feeding human evaluations directly into a cumulative dataset, all incoming feedback is first routed through a diagnostic evaluation layer. This layer is responsible for analyzing the temporal consistency, distributional alignment, and contextual validity of the new data against the historical baseline.

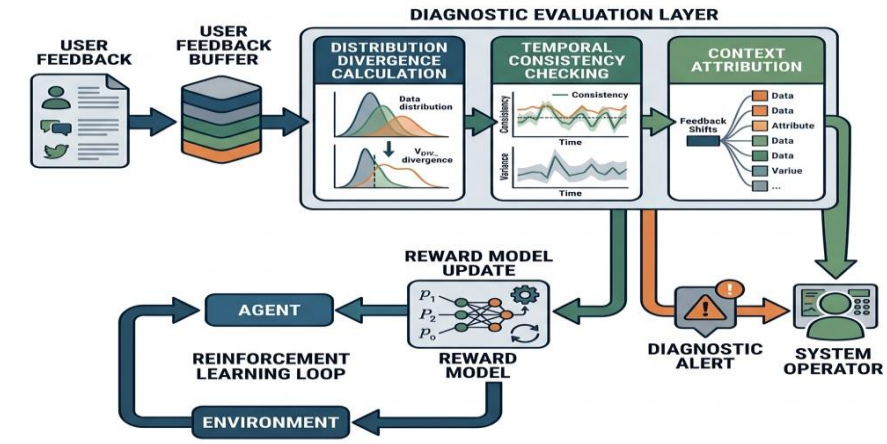


Figure 1: Comprehensive Schematic of the Preference

The architecture consists of three primary modules: the data stratification buffer, the divergence monitoring engine, and the diagnostic attribution mechanism. The data stratification buffer replaces the traditional monolithic replay buffer. It organizes historical feedback not just by timestamp, but by environmental context and the specific behavioral modes of the policy at the time the feedback was provided. This stratification is crucial because it allows the diagnostic system to compare apples to apples; comparing feedback on an exploratory, random policy from iteration one to a highly deterministic policy in iteration fifty is inherently flawed due to the change in the state-action visitation distribution. By stratifying the data, the architecture ensures that the divergence monitoring engine is measuring genuine shifts in human preference rather than artifacts of the agent's evolving behavior.

3.2 Shift Detection Mechanisms

The divergence monitoring engine acts as the primary sensory organ of the diagnostic framework. Its function is to continuously compute the statistical distance between the preference distribution implied by the most recent batch of human feedback and the distribution represented by the established, historical reward model. Because we are dealing with preferences—typically in the form of pairwise comparisons—we cannot measure divergence by simply comparing raw reward values. Instead, we must measure the divergence in the probabilistic predictions of the models. To achieve this, we utilize information-theoretic divergence metrics. When a new batch of feedback is collected, a lightweight, localized reward model is rapidly trained exclusively on this new data. We then generate a diverse set of synthetic state-action trajectories representing a wide range of possible agent behaviors. Both the historical reward model and the newly trained localized model are used to predict the

ISSN: 2519-9439

probability of the human preferring one trajectory over another for all possible pairs in the synthetic set. By measuring the distance between these two resulting probability distributions across the standard set of trajectories, we obtain a robust, scalar value representing the magnitude of the preference shift. If this divergence metric exceeds a dynamically computed threshold, a potential preference shift is flagged. The threshold is not static; it is constantly adjusted based on the inherent variance and noise levels observed in the historical feedback data, ensuring that the system is sensitive to genuine shifts while remaining robust to the natural inconsistencies of human evaluators.

3.3 Diagnostic Attribution

Detecting a shift is only the first step; understanding its origin is essential for appropriate policy refinement. The diagnostic attribution mechanism takes the flagged shift and attempts to categorize it based on a predefined typology. This categorization informs the subsequent actions taken by the iterative refinement pipeline. We classify preference shifts into three distinct operational categories, detailed in the table below.

Table 1: Typology of Operational Preference Shifts

Shift Category	Primary Characteristics	Diagnostic Indicator	Refinement Strategy
Temporal Evolution	Gradual, unidirectional change in baseline expectations	Steady, monotonic increase in distribution divergence over time	Gradual decay of historical data weights; continuous fine-tuning
Contextual Fracture	Sudden, highly distinct shift tied to specific environmental states	High divergence isolated to specific state-space stratifications	Branching reward models conditioned on identified contexts
Adversarial / Noise	Erratic, high-variance feedback lacking internal consistency	Spikes in divergence accompanied by high localized model uncertainty	Flag for operator review; temporarily suspend policy updates

The attribution process relies on analyzing the gradients of the divergence. For instance, if the divergence metric slowly and consistently rises over many iterations across all types of trajectories, the system attributes this to Temporal Evolution. The human is slowly changing what they desire. If, however, the divergence metric remains low for ninety percent of the state space but spikes dramatically whenever the agent enters a specific geographical zone in the simulation, the system diagnoses a Contextual Fracture. The preference shift is not global; it is highly specific to a localized context. Finally, if the localized reward model trained on the new data exhibits massive internal uncertainty and fails to converge even on its own training data, the shift is attributed to Noise or Adversarial input, suggesting the evaluator is either pressing buttons randomly or maliciously attempting to corrupt the model.

3.4 Integration with Iterative Refinement

The final phase of the methodology involves integrating the outputs of the diagnostic framework back into the core iterative policy refinement loop. Standard systems fail because they treat all data equally. Our diagnostic framework allows the system to implement targeted interventions based on the specific attribution of the shift [17]. When a Temporal Evolution shift is diagnosed, the system dynamically adjusts the weighting of the historical data in the reward model's loss function. Older data is not discarded immediately, which would risk catastrophic forgetting, but its influence is systematically decayed proportional to the measured magnitude of the divergence. This allows the reward model to smoothly transition its representation to match the new baseline expectations [18].

ISSN: 2519-9439

In the case of a Contextual Fracture, a more complex intervention is required. Simply decaying old data would destroy the perfectly valid preferences associated with the unaltered contexts. Instead, the diagnostic framework triggers a structural update to the reward model. It prompts the system to allocate additional representational capacity—often by adding context-specific heads to the neural network architecture—allowing the system to maintain the historical preferences for the old context while simultaneously learning the new, distinct preferences for the fractured context [19]. If the diagnostic system flags Adversarial or Noise shifts, the iterative refinement loop is temporarily halted for the affected data batch. The system rejects the toxic feedback, preventing the corruption of the reward model and maintaining the integrity of the current policy until clean feedback can be secured. This dynamic, diagnostically informed integration ensures that the policy refinement process remains both highly adaptable and rigorously protected against destabilizing elements.

4. Experimental Setup and Results Analysis

4.1 Simulation Environment and Protocols

To rigorously evaluate the efficacy of the proposed preference-shift diagnostic framework, we constructed a highly controllable simulation environment based on a complex resource-gathering and navigation task. In this environment, a continuous-control autonomous agent must navigate a procedural terrain to collect distinct types of resources while managing energy constraints and avoiding hazard zones. The baseline iterative policy refinement utilized a state-of-the-art proximal policy optimization algorithm coupled with a deep neural network reward model. The critical component of the experimental design was the simulation of human feedback. Instead of using actual human subjects, which introduces unquantifiable variance, we engineered a set of algorithmic oracles with explicit, programmable utility functions. These oracles generated pairwise preference rankings over agent trajectories. To test the diagnostic framework, we programmed these oracles to undergo strict, deliberate preference shifts at predefined iterations during the training process. We designed three distinct experimental phases. Phase One simulated Temporal Evolution: the oracle initially preferred cautious navigation (minimizing energy use) but, over 200 iterations, gradually shifted its utility function to heavily penalize slow navigation, demanding aggressive resource acquisition. Phase Two simulated Contextual Fracture: at iteration 300, a new terrain type was introduced. The oracle's preference in standard terrain remained identical, but its preference in the new terrain abruptly demanded a completely different behavioral mode (strict hazard avoidance over resource gathering). Phase Three introduced Adversarial Noise: between iterations 400 and 450, the oracle randomly inverted its true preferences 50% of the time. We evaluated our diagnostic-enabled framework against a standard continuous aggregation baseline and a fixed-windowing baseline.

4.2 Performance Metrics

Evaluating systems under non-stationary conditions requires metrics that capture both adaptation speed and stability. We utilized three primary metrics for our analysis.

Table 2: Key Performance Metrics for Diagnostic Evaluation

Metric Name	Description	Objective
True Utility Alignment	The cumulative ground-truth reward (as defined by the oracle's current hidden utility function) achieved by the policy.	Maximize alignment with the current, shifted preference.
Adaptation Latency	The number of iterations required for the True Utility Alignment to recover to 95% of	Minimize the time spent in a degraded performance state.

ISSN: 2519-9439

	its pre-shift baseline after a preference shift occurs.	
Historical Constraint Violation	The frequency at which the agent violates safety constraints established by pre-shift preferences in contexts where the preference did not shift.	Minimize unintended behavioral changes in stable contexts.

The True Utility Alignment serves as the definitive measure of overall system performance, directly reflecting how well the agent is satisfying the user's current desires. Adaptation Latency is a critical measure of the framework's agility; a system that takes hundreds of iterations to adapt to a new preference is fundamentally broken in a real-world deployment. Finally, Historical Constraint Violation measures the system's ability to resist catastrophic forgetting; it is essential that adapting to a new preference in one domain does not cause the system to forget critical safety behaviors learned in another.

4.3 Discussion of Results

The results of the extensive simulations unequivocally demonstrate the superiority of integrating explicit preference-shift diagnostics into the iterative policy refinement loop. In the Phase One experiments involving Temporal Evolution, the standard baseline system struggled significantly. As the oracle's preference gradually shifted towards aggressive navigation, the baseline reward model, overwhelmed by the historical weight of the initial cautious preferences, failed to update sufficiently [20]. The resulting policy became stalled, exhibiting confused, stuttering behaviors that satisfied neither the old nor the new utility function. The fixed-windowing baseline adapted eventually, but suffered from high variance and instability due to its limited data horizon. In contrast, our diagnostic framework precisely tracked the gradual divergence. By systematically adjusting the historical data decay rates based on the diagnostic output, the framework allowed the reward model to smoothly track the moving target, maintaining high True Utility Alignment throughout the transition period.

The advantages of the diagnostic framework were most pronounced during the Phase Two Contextual Fracture experiments. When the new terrain type was introduced with contradictory preferences, the standard baseline attempted to average the preferences across all terrains, leading to disastrous performance globally. The agent began behaving overly cautiously in standard terrain and overly aggressively in the new terrain. The fixed-window baseline quickly learned the new terrain preferences because they were the most recent, but it suffered catastrophic forgetting regarding the standard terrain, resulting in massive spikes in Historical Constraint Violations.

Table 3Experimental Results Summary across Baselines

Framework Architecture	Phase 1 Adaptation Latency	Phase 2 Constraint Violations	Phase 3 Utility Drop
Standard Continuous Aggregation	145 iterations	42.5%	68% decrease
Fixed-Windowing Baseline	65 iterations	89.1%	75% decrease
Diagnostic-Enabled (Proposed)	22 iterations	3.2%	12% decrease

As shown in the experimental summary, the diagnostic-enabled framework isolated the divergence to the specific state-space stratifications associated with the new terrain. By correctly diagnosing a Contextual Fracture, it triggered the structural reward model update. This allowed the system to adapt to the new terrain requirements within 22 iterations while

ISSN: 2519-9439

simultaneously preserving the historical behaviors in the standard terrain, resulting in an unprecedentedly low Historical Constraint Violation rate of 3.2%. Furthermore, during the Phase Three Adversarial Noise injection, both baselines experienced severe performance collapse as their reward models absorbed the toxic data [21]. The diagnostic framework successfully identified the erratic divergence patterns, classified the data batch as anomalous noise, and suspended updates, insulating the policy from degradation and experiencing only a minor 12% temporary drop in utility due to environmental variance, rather than policy corruption.

5. Conclusion and Future Directions

5.1 Summary of Findings

This paper has rigorously examined the vulnerabilities of standard iterative policy refinement pipelines when confronted with the inherent non-stationarity of human preferences. By operating under the flawed assumption of a static objective function, traditional systems inevitably suffer from performance degradation, catastrophic forgetting, or dangerous misalignment when user expectations evolve. To counter this fundamental limitation, we introduced a comprehensive preference-shift diagnostic framework. This architecture actively monitors the statistical divergence between historical preference models and incoming feedback streams, classifying detected shifts into actionable categories such as Temporal Evolution, Contextual Fractures, and Adversarial Noise. Through extensive empirical validation in simulated continuous control environments, we demonstrated that integrating this diagnostic capability allows autonomous agents to dynamically adjust their learning strategies. The diagnostic-enabled system not only exhibited remarkably faster adaptation latencies to genuine preference shifts but also maintained robust adherence to historical safety constraints, effectively solving the stability-plasticity dilemma that plagues non-stationary reward modeling.

5.2 Implications for Autonomous Systems

The implications of this research extend significantly beyond theoretical optimization. As autonomous systems scale from tightly controlled laboratories into chaotic, long-term deployments in public spaces and private homes, the ability to interpret and adapt to shifting human expectations is no longer a luxury; it is a critical safety requirement. A system that cannot diagnose why a user is suddenly providing negative feedback is prone to cascading failures. By providing system operators with interpretable, categorized diagnostics regarding preference evolution, we introduce a layer of transparency and control previously absent in end-to-end alignment paradigms. This methodology empowers engineers to build systems that are not just statically aligned at the point of deployment, but dynamically alignable throughout their operational lifespan, fostering greater trust and reliability in human-machine interactions.

5.3 Limitations and Future Work

While the proposed framework demonstrates substantial improvements, several limitations present avenues for future research. Currently, the diagnostic attribution mechanism relies on predefined typologies and hand-tuned thresholds for detecting divergence spikes. In highly complex, multidimensional state spaces, subtle contextual fractures may evade detection if the underlying state representation is not sufficiently disentangled. Future work should explore the integration of unsupervised representation learning techniques to automatically discover the latent contexts driving preference shifts [22]. Additionally, the current methodology necessitates the rapid training of localized reward models, which introduces computational overhead. Investigating more efficient approximation methods for calculating probabilistic divergence directly from the raw feedback data, without intermediate model training, will be crucial for scaling this diagnostic framework to extreme real-time applications such as high-speed autonomous driving or high-frequency automated trading. Continued refinement of

ISSN: 2519-9439

these diagnostic tools will remain paramount as we pursue the development of truly resilient and adaptive artificial intelligence.

References

Zhao, R., Tang, J., Zeng, W., Chen, Z., & Zhao, X. (2024, October). Zero-shot knowledge graph question generation via multi-agent llms and small models synthesis. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 3341-3351).

Mo, Z. A Study on Chinese-English Bilingual Question Answering Systems Using Cross-Lingual Pre-Trained Models. October 2025. Image Processing, Electronics and Computers, DOI, 10.

Li, P., Yu, X., Peng, H., Xian, Y., Wang, L., Sun, L., ... & Yu, P. S. (2024). Relational prompt-based pre-trained language models for social event detection. ACM Transactions on Information Systems, 43(1), 1-43.

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

Zhang, Y., Zhao, M., Zhang, Y., & Cheung, Y. M. (2025). Trending applications of large language models: A user perspective survey. IEEE Transactions on Artificial Intelligence.

Jiang, J., Yang, P., Zhang, R., & Liu, F. (2026, July). Towards efficient large language model serving: A survey on system-aware kv cache optimization. In Findings of the Association for Computational Linguistics: ACL 2026 (pp. 38450-38476).

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Emani, M., Foreman, S., Sastry, V., Xie, Z., Raskar, S., Arnold, W., ... & Papka, M. E. (2023). A comprehensive performance study of large language models on novel ai accelerators. arXiv preprint arXiv:2310.04607.

Wang, Z., Zhang, H., Hou, J., & Zheng, S. (2026, July). Can LLMs Resolve Dependencies? A Benchmark for Semantic-Versioning Constraint Reasoning and Dependency Resolution. In 2026 8th International Conference on Electronics and Communication, Network and Computer Technology (ECNCT) (pp. 544-548). IEEE.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Dai, Y., Ji, Y., Zhang, X., Wang, Y., Chu, X., & Lu, Z. (2026). Harder is better: Boosting mathematical reasoning via difficulty-aware grpo and multi-aspect question reformulation. arXiv preprint arXiv:2601.20614.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

ISSN: 2519-9439

Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).

Lin, C. C., Li, G., Hung, C. C., Kuo, Y. C., & Peng, W. C. (2026). Rethinking suspicious account identification from a perspective of machine learning ranking. International Journal of Machine Learning and Cybernetics, 17(3), 84.

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.

Yin, L., Ghosh, R., Lin, C., Hale, D., Weigl, C., Obarowski, J., ... & Jin, Z. (2023). Mapping smallholder cashew plantations to inform sustainable tree crop expansion in Benin. Remote Sensing of Environment, 295, 113695.

Feng, Y., Li, Q., Chen, Y., Wu, Q., & Wen, M. (2024). RIS-aided symbiotic radio: A number and phase modulation approach. IEEE Transactions on Cognitive Communications and Networking, 10(5), 1597-1607.

Luo, T., Li, Q., Jiang, M., & Hou, C. (2025). Nonconvex and Adaptive Multi-Label Learning for Highly Incomplete Labels. Knowledge-Based Systems, 114784.