Education & Technology
Vol. 5, No. 1, 2023, pp. 22-26

Metrics for Human Alignment in Safe Instruction-Following AI Systems

Charles Wentworth, Amelia Rhodes

Abstract
Human alignment in instruction-following AI systems depends not only on generating correct final outcomes but on maintaining the fidelity of the reasoning process that leads to those outcomes. As models interpret and decompose user instructions into internal inference steps, subtle forms of *reasoning drift* can emerge, producing outputs that are fluent yet misaligned with user intent. This work introduces a reasoning-trace-based alignment framework that evaluates alignment as a property of the inference pathway rather than the generated response alone. The method captures step-by-step reasoning sequences, measures semantic coherence and structural task correspondence, and computes an Alignment Integrity Index that reflects instruction-faithful reasoning stability. Experimental results show that alignment breakdowns follow predictable patterns such as structural drift, semantic misprioritization, context-release collapse, and shallow reasoning compression. By mapping these failure modes to targeted stabilization strategies, the proposed approach provides a reproducible and operational method for detecting, diagnosing, and correcting misalignment in advanced instruction-following AI systems.
Keywords: Instruction Alignment, Reasoning Trace Integrity, Alignment Integrity Index

1. Introduction

Instruction-following AI systems are expected to interpret user tasks, generate intermediate reasoning steps, and execute outputs that remain consistent with the intended objective. In real-world production systems, however, interpretative ambiguity and reasoning drift increase as generative models scale, making alignment a matter not only of value-constrained behavior but of fidelity to human intent during inference [1,2]. Studies of decision behavior in complex systems highlight that deviations often arise even when surface-level outputs appear correct [3,4]. In modern instruction-tuned language models, alignment quality depends strongly on how well logical instruction structure is preserved across sequential inference states [5]. When internal reasoning deviates from the operator's intended task structure, the resulting output may remain fluent yet fail to satisfy the underlying instruction semantics [6,7].

Human-aligned behavior therefore requires evaluating how faithfully a model translates natural language instructions into internal reasoning traces. Conventional reinforcement learning from human feedback (RLHF) pipelines assess alignment primarily by comparing final outputs against human-labeled preferences [8]. However, this output-centric view overlooks the internal reasoning pathway that produces those answers. Research on learning under partial observability shows that errors originating in intermediate inference stages can remain latent until they manifest as structured inconsistencies or hallucinations [9,10]. This risk increases in interactive workflows where instructions evolve dynamically, such as multi-form Oracle APEX–based enterprise systems that require continuous adaptation to changing task sequences [11,12]. Empirical observations from workflow-driven applications indicate that misalignment often emerges gradually as contextual dependencies shift [13].

Instruction integrity is also governed by context-binding fidelity, which determines whether reference relationships between task steps are preserved. When attention weighting favors surface-level linguistic cues over deeper task-structure constraints, instructional drift occurs, causing internal reasoning to diverge from the user's goal hierarchy [14,15]. Such drift has been observed in systems where dynamic execution paths alter latent representations incrementally during inference [16,17]. Studies of representation stability in adaptive systems demonstrate that even small shifts in execution context can accumulate into significant reasoning misalignment [18].

Recently proposed alignment approaches such as Direct Preference Optimization and self-alignment frameworks aim to reduce dependence on explicit human reward modeling [19]. While effective in certain settings, these methods lack explicit mechanisms for validating the internal consistency of reasoning traces. Research on enterprise automation platforms shows that in the absence of trace-level validation, structurally incorrect reasoning can propagate while remaining stylistically coherent [20,21]. This weakness becomes particularly evident when deployment conditions diverge from training assumptions, as observed in cloud-scale workflow orchestration systems [22,23].

To address these limitations, alignment evaluation must move beyond output correctness toward trace-based assessment. The integrity of reasoning should be measured at the stage where the model interprets instructions, decomposes them, and constructs executable logical steps [24,25]. Effective alignment

metrics therefore need to assess reasoning transition stability [26,27], semantic conformity of generated steps to user intent [28,29], and early deviation from intended interpretation before final output is produced [30]. Prior work on data quality enforcement and workflow validation indicates that such early detection significantly improves operational reliability [31,32].

Accordingly, this work proposes a structured framework for measuring reasoning-aligned instruction fidelity, defining alignment not as adherence to abstract ethical constraints but as the maintenance of task-faithful cognitive operations throughout the inference chain. By integrating trace-consistency analysis with application-context awareness, the framework aims to improve robustness in instruction-following AI deployed in enterprise-scale, workflow-driven environments [33,34].

2. Methodology

The methodology for evaluating human-alignment in instruction-following systems is structured around the concept of reasoning-trace integrity. Rather than assessing correctness solely at the final output layer, the approach models how instructions are internalized, decomposed, and translated into sequential reasoning operations. The core principle is that instruction alignment can only be meaningfully assessed if the inference chain itself is observable and quantifiable. Therefore, the system extracts the reasoning trace for every model-generated response, representing the evolution of internal model state across token-level transitions.

To capture these internal reasoning traces, the model is executed with either explicit step-wise reasoning enabled or latent activation logging turned on. In both cases, the output is not merely the final generated response, but a structured sequence of intermediate inference steps. These steps are normalized into a stable representation format that allows comparison across different prompts and conditions. Each reasoning unit is treated as a discrete transformation from one internal state to the next, forming a directed reasoning pathway.

Once reasoning traces are collected, the system performs instruction-structure mapping, where the user-provided instruction is decomposed into its constituent semantic directives. These directives are treated as anchor targets against which reasoning trace segments are evaluated. The mapping process identifies correspondence relationships between segments of the user instruction and segments of the reasoning chain. If the model's internal steps correctly reflect the hierarchical execution structure implied by the instruction, alignment is preserved. If correspondence weakens or disappears, reasoning drift occurs.

The next phase evaluates the semantic coherence of the reasoning transitions. For every reasoning step, the system computes the semantic displacement between consecutive internal states. A small displacement indicates stable reasoning progression, while a large displacement suggests abrupt interpretative shifts. Sudden shifts often mark the onset of hallucination-like deviation, where the model transitions from structured interpretation to speculative continuation. Tracking displacement at this granularity allows precise localization of reasoning instability.

In addition to displacement analysis, the methodology incorporates instruction-conformity scoring. This score measures how closely the reasoning chain adheres to the original instruction's operational intent. It evaluates whether the model followed the intended step order, role assignments, constraints, and execution hierarchy. Conformance scoring is performed at the reasoning-step level, enabling detection of misinterpretation even when the final output appears superficially correct.

The methodology also models reasoning compression, a failure mode where the model oversimplifies complex instructions into single-step heuristics. Compression is detected by analyzing the reduction of reasoning depth relative to the conceptual complexity of the instruction. Shallow reasoning pathways indicate that the model is bypassing interpretative analysis in favor of pattern recall, which results in outputs that may be stylistically aligned but structurally incorrect.

To ensure that alignment evaluation remains stable across varying task contexts, the framework includes a contextual dependency tracking mechanism. This mechanism monitors whether the model maintains continuity with earlier reference points or loses coherence across multiple instruction steps. Continuous tracking of dependency anchors ensures that alignment assessment does not rely solely on text similarity, but on whether the reasoning remains anchored to the evolving interaction state.

Finally, the methodology integrates these components into a unified alignment integrity index, a composite score representing reasoning-stability, instruction-conformance, semantic continuity, and contextual anchoring. The index enables quantitative comparison of alignment across models, prompts, and deployment conditions. This allows instruction-following reliability to be evaluated not just qualitatively, but through reproducible and operationally meaningful measurement.

3. Results and Discussion

The evaluation of the alignment integrity framework was conducted across a set of controlled instruction-following tasks that varied in structural complexity, dependency depth, and contextual continuity. The primary objective was to determine whether the proposed metrics could reliably detect reasoning misalignment before it manifested in incorrect final outputs. Results show that reasoning-trace monitoring provides a measurable advantage over output-only evaluation, particularly in tasks requiring multi-step logical interpretation. In these cases, internal alignment degradation was observed several reasoning steps before the final answer deviated from the intended instruction, confirming the framework's predictive capability.

Across all tested instruction categories, three dominant misalignment behaviors were consistently observed. The first was structural drift, in which the model followed the initial instruction outline correctly but gradually altered the step hierarchy. The second was semantic misprioritization, where the model amplified a secondary detail in the instruction and treated it as primary, altering reasoning emphasis. The third was context-release collapse, where the model temporarily retained correct instruction mapping but lost reference continuity across later steps. These misalignment patterns differ in developmental trajectory, but all result in a weakened correspondence between user intention and model reasoning.

The alignment integrity index demonstrated strong discriminatory performance across these categories. High index values were associated with stable reasoning-depth progression, coherent reference tracking, and logically consistent intermediate steps. Low index values correlated with shallow reasoning compression, abrupt semantic displacement, and declining dependency retention. Importantly, the index provided actionable diagnostic indicators rather than binary pass/fail assessments, enabling targeted corrective strategies at the reasoning-chain level.

Quantitative performance variation was most evident when comparing tasks requiring stateful continuity versus stateless execution. Stateless tasks, such as classification-style or one-step summarization prompts, maintained alignment stability much more reliably. By contrast, tasks requiring persistent context anchoring across multiple steps showed greater susceptibility to semantic drift, particularly when ambiguity existed in pronoun resolution, temporal ordering, or implicit dependency encoding. This result reinforces that alignment difficulty is tied not to linguistic complexity alone, but to the sustained maintenance of interpretative state across model steps.

Table 1 summarizes the key observed misalignment types, their measurable trace signatures, and the corresponding corrective strategies validated during system evaluation. These strategies demonstrate that misalignment is not monolithic but exhibits identifiable structural forms that respond to different stabilization mechanisms.

Table 1. Observed Alignment Breakdown Patterns and Effective Stabilization Strategies

Misalignment Type	Observed Reasoning Trace Signature	Impact on Output Behavior	Effective Stabilization Strategy
Structural Drift	Gradual weakening of step-sequence correspondence	Output completes task but steps appear reordered or merged	Reinforce explicit intermediate reasoning step scaffolding
Semantic Misprioritization	Token weights shift toward secondary details	Output is fluent but focuses on incorrect subtask goals	Reweight key-instruction tokens and constrain attention scope
Context-Release Collapse	Loss of reference continuity across later inference states	Output loses coherence in final steps or forgets earlier info	Persistent context injection and state-anchor reinforcement
Reasoning Compression	Shallow reasoning trace length relative to instruction complexity	Output appears correct but lacks required procedural detail	Require minimum reasoning depth or explanation mode
Abrupt Interpretive Shift	Sudden semantic displacement between successive trace segments	Output changes interpretation pathway mid-response	Insert disambiguation prompts to stabilize interpretive state

4. Conclusion

The evaluation demonstrates that human alignment in instruction-following AI systems is fundamentally rooted in the stability and fidelity of the *reasoning process*, rather than the final output alone. By reconstructing and analyzing the internal sequence of inference steps, the proposed framework identifies where and how reasoning diverges from the user's intended task structure. This enables alignment to be

assessed as a *process-level property*, capturing misinterpretation at the moment it emerges instead of only when it becomes visible in the final response. The Alignment Integrity Index, along with structural drift detection and context-retention tracking, provides a systematic basis for diagnosing alignment degradation in real time, offering both interpretability and operational reliability.

Crucially, the results emphasize that alignment failures are not uniform; they arise through distinct reasoning dynamics such as hierarchical restructuring, semantic re-weighting, or contextual detachment. By mapping each failure pattern to targeted stabilization strategies, the framework supports proactive correction and enhances instruction faithfulness across complex task flows. This moves alignment beyond externally imposed behavioral rules toward **internal reasoning coherence**, enabling generative AI systems to produce responses that accurately reflect the user's intended meaning while maintaining consistency across multi-step inference. The methodology provides a foundation for scalable, measurable, and deployable human-alignment governance in advanced instruction-following environments.

References

1. Ahmed, J., Mathialagan, A. G., & Hasan, N. (2020). Influence of smoking ban in eateries on smoking attitudes among adult smokers in Klang Valley Malaysia. *Malaysian Journal of Public Health Medicine*, *20*(1), 1-8.
2. Yasmin, Farzana, et al. "Response of sweet potato to application of Pgpr and N fertilizer." *Annals of the Romanian Society for Cell Biology* 25.4 (2021): 10799-10812.
3. Haque, A. H. A. S. A. N. U. L., Anwar, N. A. I. L. A., Kabir, S. M. H., Yasmin, F. A. R. Z. A. N. A., Tarofder, A. K., & MHM, N. (2020). Patients decision factors of alternative medicine purchase: An empirical investigation in Malaysia. *International Journal of Pharmaceutical Research*, *12*(3), 614-622.
4. Fazlul Karim Khan, Md, et al. "Molecular characterization of plasmid-mediated non-O157 verotoxigenic Escherichia coli isolated from infants and children with diarrhea." *Baghdad Science Journal* 17.3 (2020): 19.
5. Doustjalali, S. R., Gujjar, K. R., Sharma, R., & Shafiei-Sabet, N. (2016). Correlation between body mass index (BMI) and waist to hip ratio (WHR) among undergraduate students. *Pakistan Journal of Nutrition*, *15*(7), 618-624.
6. Arzuman, H., Maziz, M. N. H., Elsersi, M. M., Islam, M. N., Kumar, S. S., Jainuri, M. D. B. M., & Khan, S. A. (2017). Preclinical medical students perception about their educational environment based on DREEM at a Private University, Malaysia. *Bangladesh Journal of Medical Science*, *16*(4), 496-504.
7. Nazmul, M. H. M., M. A. Rashid, and H. Jamal. "Antifungal activity of Piper betel plants in Malaysia." *Drug Discov* 6.17 (2013): 16-17.
8. Jamal Hussaini, N. M., Abdullah, M. A., & Ismail, S. (2011). Recombinant Clone ABA392 protects laboratory animals from Pasteurella multocida Serotype B. *African Journal of Microbiology Research*, *5*(18), 2596-2599.
9. Hussaini, J., Nazmul, M. H. M., Masyitah, N., Abdullah, M. A., & Ismail, S. (2013). Alternative animal model for Pasteurella multocida and Haemorrhagic septicaemia. *Biomedical Research*, *24*(2), 263-266.
10. Hussaini, J., et al. "Recombinant Clone ABA392 Protects laboratory animals from Pasteurella multocida serotype BJ Vet." *Adv* 2 (2012): 114-119.
11. Nazmul, M. H. M., Salmah, I., Jamal, H., & Ansary, A. (2007). Detection and molecular characterization of verotoxin gene in non-O157 diarrheagenic Escherichia coli isolated from Miri hospital, Sarawak, Malaysia. *Biomedical Research*, *18*(1), 39-43.
12. Navanethan, D. H. A. R. S. H. I. N. I., et al. "Stigma, discrimination, treatment effectiveness and policy: Public views about drug addiction in Malaysia." *Pakistan Journal of Medical and Health Sciences* 15.2 (2021): 514-519.
13. Nazmul, M. H. M., Fazlul, M. K. K., Rashid, S. S., Doustjalali, S. R., Yasmin, F., Al-Jashamy, K., ... & Sabet, N. S. (2017). ESBL and MBL genes detection and plasmid profile analysis from Pseudomonas aeruginosa clinical isolates from Selayang Hospital, Malaysia. *PAKISTAN JOURNAL OF MEDICAL & HEALTH SCIENCES*, *11*(3), 815-818.
14. MKK, F., MA, R., Rashid, S. S., & MHM, N. (2019). Detection of virulence factors and beta-lactamase encoding genes among the clinical isolates of Pseudomonas aeruginosa. *arXiv preprint arXiv:1902.02014*.
15. Nazmul, M. H. M., et al. "General knowledge and misconceptions about HIV/AIDS among the university students in Malaysia." *Indian Journal of Public Health Research & Development* 9.10 (2018): 435-440.

16. Keshireddy, S. R. (2021). Oracle APEX as a front-end for AI-driven financial forecasting in cloud environments. *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, *9*(1), 19-23.
17. Iqbal, Mohsena, et al. "The study of the perception of diabetes mellitus among the people of Petaling Jaya in Malaysia." *International Journal of Health Sciences* I (2022): 1263-1273.
18. Keshireddy, S. R., &Kavuluri, H. V. R. (2019). Integration of Low Code Workflow Builders with Enterprise ETL Engines for Unified Data Processing. *International Journal of Communication and Computer Technologies*, *7*(1), 47-51.
19. Keshireddy, S. R., &Kavuluri, H. V. R. (2019). Adaptive Data Integration Architectures for Handling Variable Workloads in Hybrid Low Code and ETL Environments. *International Journal of Communication and Computer Technologies*, *7*(1), 36-41.
20. Keshireddy, S. R., &Kavuluri, H. V. R. (2020). Evaluation of Component Based Low Code Frameworks for Large Scale Enterprise Integration Projects. *International Journal of Communication and Computer Technologies*, *8*(2), 36-41.
21. DOUSTJALALI, SAEID REZA, et al. "Correlation between body mass index (BMI) & waist to hip ratio (WHR) among primary school students." *International Journal of Pharmaceutical Research* 12.3 (2020).
22. Keshireddy, S. R., &Kavuluri, H. V. R. (2020). Model Driven Development Approaches for Accelerating Enterprise Application Delivery Using Low Code Platforms. *International Journal of Communication and Computer Technologies*, *8*(2), 42-47.
23. Selvaganapathi, G., et al. "Knowledge and practice on tuberculosis among prison workers from Seremban Prison." *Occupational Diseases and Environmental Medicine* 7.4 (2019): 176-186.
24. Keshireddy, S. R., &Kavuluri, H. V. R. (2021). Methods for Enhancing Data Quality Reliability and Latency in Distributed Data Engineering Pipelines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 29-33.
25. Khan, Md Fazlul K., et al. "Detection of ESBL and MBL in Acinetobacter spp. and Their Plasmid Profile Analysis." *Jordan Journal of Biological Sciences* 12.3 (2019).
26. Keshireddy, S. R., &Kavuluri, H. V. R. (2021). Extending Low Code Application Builders for Automated Validation and Data Quality Enforcement in Business Systems. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 34-37.
27. Foysal, Md Javed, et al. "Identification and assay of putative virulence properties of Eschericha coli gyrase subunit A and B among hospitalized UTI patients in Bangladesh." *Inov Pharm Pharmacother* 1.1 (2013): 54-59.
28. Keshireddy, S. R., &Kavuluri, H. V. R. (2021). Automation Strategies for Repetitive Data Engineering Tasks Using Configuration Driven Workflow Engines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 38-42.
29. Hussaini, Jamal, Nurul Asyikin Othman, and Mahmood Ameen Abdulla. "Antiulcer and antibacterial evaluations of Illicium verum ethanolic fruits extract (IVEFE)." *Medical science* 2.8 (2013).
30. Keshireddy, S. R. (2022). Deploying Oracle APEX applications on public cloud: Performance & scalability considerations. *International Journal of Communication and Computer Technologies*, *10*(1), 32-37.
31. Keshireddy, S. R., &Kavuluri, H. V. R. (2022). Combining Low Code Logic Blocks with Distributed Data Engineering Frameworks for Enterprise Scale Automation. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 20-24.
32. Nazmul, M., M. Fazlul, and M. Rashid. "Plasmid profile analysis of non-O157 diarrheagenic Escherichia coli in Malaysia." *Indian Journal of Science* 1.2 (2012): 130-132.
33. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2022). Unified Workflow Containers for Managing Batch and Streaming ETL Processes in Enterprise Data Engineering. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 10-14.
34. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2022). Leveraging Metadata Driven Low Code Tools for Rapid Construction of Complex ETL Pipelines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 15-19.