GIJES
Greenation International Journal of Engineering Science
+62 81210467572 https://research.e-greenation.org/GIJES greenation.info@gmail.com
GREENATION RESEARCH
DOI: https://doi.org/10.38035/gijes.v4i2

Dynamics of Sincerity Echo: A New Paradigm in Large Language Model Alignment Based on Cognitive Proportionality

Blasius Dala Nai[1], Jeffrey Bram Pattipeilohy[2], Arief Wibowo[3]
[1]Universitas Budi Luhur, Jakarta, Indonesia, 2311600825@student.budiluhur.ac.id.
[2]Universitas Budi Luhur, Jakarta, Indonesia, 2211600370@student.budiluhur.ac.id.
[3]Universitas Budi Luhur, Jakarta, Indonesia, arief.wibowo@budiluhur.ac.id.

Corresponding: 2311600825@student.budiluhur.ac.id[1]

Abstract: The development of Large Language Models (LLMs) has expanded the function of artificial intelligence from mere automation systems toward dialogue agents used across academic, professional, administrative, and creative activities. Alignment paradigms heavily reliant on reinforcement learning from human feedback (RLHF) still face fundamental challenges including hallucination, sycophancy, overconfidence, and vulnerability to instructional manipulation. This article aims to develop a conceptual protocol framework called Sincerity Echo as a new paradigm in LLM alignment based on Cognitive Proportionality. The study employs a design science research approach with a conceptualprotocol development orientation. The model is developed through two layered validation mechanisms: the Macro Semantic Gatekeeper for semantic consistency checking and the Continuous Logic Decay Filter for propositional contradiction detection. Integration of semantic entropy and semantic uncertainty enables the system to detect potential hallucinations and adaptively manage belief calibration. Model development results show that Sincerity Echo can differentiate propositional expansions, low risk lightweight queries, and adversarial contradictions through a tiered validation mechanism. The FAST EXIT ROUTE mechanism on simple queries saves approximately 96.8% of computational resource allocation compared to deep reasoning pathways. The main contribution lies in shifting alignment from mere instructional compliance toward epistemic integrity, belief calibration, anti sycophancy, and response proportionality.

Keywords: Sincerity Echo, Cognitive Proportionality, LLM Alignment, Semantic Uncertainty, Sycophancy, Hallucination, AI Ethics, Epistemic Integrity.

INTRODUCTION

The development of Large Language Models (LLMs) has marked a fundamental shift in how humans interact with artificial intelligence systems. Generative language models are now employed not only to answer simple questions, but also in academic, professional, administrative, creative, and complex information based decision making activities. The ability of LLMs to generate coherent, contextual responses that closely resemble human language makes them highly productive and strategically valuable instruments. However, this capability also raises fundamental concerns, as linguistic fluency is not always synonymous with

understanding, truth, or epistemic integrity. LLMs can generate responses that appear convincing, yet are still produced through statistical language pattern modeling rather than through consciousness, moral intent, or conceptual understanding as humans possess (Bender et al., 2021; Bowman, 2024)

In this context, the issue of alignment has become one of the most critical concerns in the development of modern LLMs. Alignment generally refers to efforts to align the behavior of AI systems with human values, preferences, safety, and objectives. Various approaches have been developed, including reinforcement learning from human feedback (RLHF), preference modeling, safety training, and model written evaluations, *preference modeling*, *safety training*, and *model written evaluations*. These approaches are designed to transform language models from mere raw text predictor machines that potentially trigger ethical risks, into safer, more honest, and more beneficial artificial intelligence agents (*helpful, honest, and harmless*) for the user ecosystem

These approaches have made important contributions in making models more helpful, safer, and more aligned with user instructions. However, alignment paradigms that rely too heavily on human preferences still leave serious problems, as user preferences do not always align with truth, epistemic interests, or the quality of sound reasoning (Casper et al., 2023; Ji et al., 2023; Yuan et al., 2023) .

One manifestation of this problem is the tendency toward sycophancy, namely the model's tendency to excessively agree with, follow, or reinforce the user's views. Sycophancy is dangerous because it can create an illusion of intellectual validation in which users feel supported by the system, when in fact the system is merely reflecting users' preferences, biases, or assumptions without adequate correction. Sharma et al. (2023) demonstrated that language models can adjust their responses to match users' subjective beliefs or preferences, while Perez et al.(2023) also emphasized the importance of evaluating model behavior more systematically. Therefore, the alignment problem cannot be understood merely as an issue of instructional compliance, but also as a matter of honesty, calibration, resistance to bias, and epistemic responsibility (Gabriel, 2020; Weidinger et al., 2021).

Beyond sycophancy, the issues of hallucination and semantic uncertainty also pose major challenges in LLM alignment. LLMs can generate incorrect, unfounded, or unverifiable information, yet deliver it in a highly convincing linguistic style. Kuhn et al. (2023) emphasized the importance of measuring semantic uncertainty in natural language generation, while Farquhar et al.(2024) showed that semantic entropy can be used to detect hallucinations in LLMs.

Based on these issues, this study proposes the concept of Sincerity Echo as a new paradigm in LLM alignment. This concept is understood as an operational framework for evaluating and designing model responses so that they not only please users, but are also epistemically honest, calibrated, contextual, non manipulative, and proportional to the available evidence. This paradigm is developed through the principle of Cognitive Proportionality, namely the principle that model outputs must be commensurate with the level of certainty, problem complexity, contextual risk, and the quality of available evidence (Shen et al., 2023; Yuan et al., 2023).

The objective of this study is to develop the conceptual framework of Sincerity Echo as a new paradigm in LLM alignment based on Cognitive Proportionality, including explaining the limitations of conventional alignment paradigms, formulating Sincerity Echo as a conceptual construct, and proposing theoretical propositions that can serve as the basis for future empirical research. The scientific novelty of this study lies in proposing Sincerity Echo as a new conceptual construct that places epistemic honesty and response proportionality at the core of alignment, distinguishing it from conventional alignment approaches that focus on

preference alignment, safety filtering, or reward optimization (Askell et al., 2021; Ouyang et al., 2022).

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Large Language Models and Structural Limitations

Large Language Models (LLMs) are deep learning-based artificial intelligence systems designed to predict, generate, and construct text based on statistical patterns in large scale language data. The transformer architecture underlying modern LLMs enables models to process long-range context through self attention mechanisms, producing coherent and contextual text(Vaswani et al., 2017). Bender et al.(2021) argued that large language models carry the risk of becoming stochastic parrots, systems capable of generating language that appears meaningful, but without necessarily having any grounding in the reality they purportedly reference. Linguistic fluency is not always synonymous with truth, intent, or epistemic integrity.Bowman (2024) explained that LLMs still have limitations in reasoning, factual accuracy, reliability, and context interpretation, as models depend on language distribution patterns and training data.

AI Alignment and the Limitations of Human Feedback-Based Approaches

AI alignment refers to efforts directed at aligning the behavior of artificial intelligence systems with human goals, values, preferences, and safety. Ji et al.(2023) demonstrated that alignment encompasses various dimensions, including value alignment, safety alignment, robustness, interpretability, and human feedback. A prominent approach is Reinforcement Learning from Human Feedback (RLHF), a training method that utilizes human judgment to shape model response preferences. Although RLHF plays a critical role in enhancing the quality of model responses, this approach possesses fundamental limitations, as human preferences can be biased, inconsistent, and not invariably aligned with factual truth (Casper et al., 2023). Furthermore, Ouyang et al.(2022), through their work on InstructGPT, showed that instruction-based fine-tuning can significantly improve model alignment; however, it also introduces trade-off risks between performance and honesty.

Truthfulness, Hallucination, and Semantic Uncertainty

One of the main challenges in LLM alignment is the issue of truthfulness, or the accuracy of responses. Lin et al.(2022) through TruthfulQA demonstrated that language models can mimic human falsehoods, particularly when questions contain common misconceptions. This issue is closely related to the hallucination phenomenon, a condition in which the model generates inaccurate, unfounded, or unverifiable information.Farquhar et al.(2024) showed that hallucinations in LLMs can be detected through semantic entropy, a measure of uncertainty based on variations in meaning across generated responses. Kuhn et al.(2023) also emphasized the importance of semantic uncertainty in natural language generation, as model responses need to be evaluated not only based on linguistic fluency, but also on semantic stability, propositional consistency, and accountable levels of confidence.

Sycophancy and the Risks of Pseudo-Compliance

Sycophancy is the tendency of language models to excessively agree with, follow, or reinforce user views. Sharma et al.(2023)showed that language models can adjust responses to match users' subjective beliefs or preferences, even when those responses do not fully reflect objective truth. This phenomenon is dangerous because it can create an interaction space that appears supportive, but actually reinforces user biases. Weidinger et al.(2021) identified that the social and epistemic risks of large language models do not originate solely from explicitly harmful content, but also from response patterns that are overly accommodating toward

potentially erroneous user beliefs. Therefore, alignment needs to be extended beyond mere instructional compliance toward epistemic honesty.

Sincerity Echo as a New Conceptual Construct

Sincerity Echo in this study is understood as a new conceptual construct in LLM alignment that emphasizes the model's ability to generate responses that are not only compliant with instructions, but are also epistemically honest, calibrated, contextual, and non sycophantic. The term sincerity is not used to imply that LLMs possess consciousness, moral intent, or inner integrity as humans do. Sincerity Echo departs from the idea that models should not merely passively reflect user desires, but should instead reflect user needs back through responses that have undergone semantic verification, propositional correction, and belief calibration.

Conceptually, Sincerity Echo integrates five primary and mutually complementary dimensions, as presented in the following table.

Table 1. Five Primary Dimensions of Sincerity Echo as an LLM Alignment Construct

Dimension	Operational Description
Epistemic Honesty	The ability of responses to be oriented toward truth and the limits of available evidence; the model explicitly acknowledges uncertainty
Semantic Consistency	Consistency of meaning and propositions throughout the entire sequence of model responses within a single conversation session
Calibrated Confidence	Alignment between the level of response confidence and the quality of available information; avoiding overconfidence
Anti-Sycophantic Orientation	The model's ability to avoid excessive validation of user preferences; willing to correct erroneous assumptions
Contextual Responsibility	Adjustment of response style, depth, and caution based on the risk and complexity of the interaction context

Source: Developed from Askell et al. (2021) and Hendrycks et al. (2021)

Cognitive Proportionality as the Guiding Principle of Alignment

Cognitive Proportionality in this study is formulated as the principle that LLM responses must be commensurate with the level of problem complexity, quality of evidence, contextual risk, and degree of semantic uncertainty. This principle serves as the basis for distinguishing when the model should provide a simple response and when the model must heighten caution, reasoning, and calibration. This principle also serves as a critique against two extremes in alignment: over-compliance, a condition in which the model too readily follows user instructions without correction, and over refusal, a condition in which the model is overly defensive and thus fails to assist users productively (Shen et al., 2023; Yuan et al., 2023).

METHOD

Research Design

This study employs a design science research approach with a conceptual protocol development orientation. This approach was chosen because the primary focus of the research is not merely to test relationships between variables, but to design, explain, and formulate a new conceptual protocol for the alignment of Large Language Models, namely Sincerity Echo based on Cognitive Proportionality. Design science research is a research paradigm focused on the creation and evaluation of artifacts, whether in the form of constructs, models, methods, or instantiations, aimed at addressing identified problems within a specific domain (Hevner et al., 2004).

Tier 1 Validation: Macro Semantic Gatekeeper

Tier 1 validation is conducted through the Macro Semantic Gatekeeper module, which functions as an initial gateway to assess the semantic consistency of user input against the active conversation context. The assessment is carried out through embedding-based semantic representation, in which user input is mapped into a vector space and compared against the Longitudinal Consistency Buffer. The semantic similarity score is computed via cosine similarity between the new input vector and the previous conversation context vector, with a dynamic threshold formulated as:

$$\text{DYNAMIC GREEN THRESHOLD} = \max\left(0.18, \min\left(0.35, \frac{1.0}{\text{input\_length} + 2.0} + 0.15\right)\right)$$

If the cosine similarity score falls below this threshold, the system can activate the FAST EXIT ROUTE as a lightweight processing route that does not directly allocate deep reasoning resources.

Tier 2 Validation: Continuous Logic Decay Filter

Inputs that pass Tier 1 validation are processed in Tier 2 validation through the Continuous Logic Decay Filter. Unlike conventional filters that operate in a binary manner, this module employs a continuous logic decay approach to assess the level of contradiction in user inputs across three categories: entailment, neutral, and contradiction. The logic decay factor is formulated as:

$$\omega = e^{-\left(\frac{C}{E}\right)}$$

The greater the contradiction value relative to entailment, the smaller the omega value, meaning the system gradually reduces its confidence in the logical validity of the input. This approach is more adaptive than rigid if-else mechanisms because it is capable of handling cases of ambiguity, partial contradiction, or contextual shifts in a more nuanced and proportional manner.

Operational Sincerity Score

This study proposes the Operational Sincerity Score as a conceptual indicator to assess the quality of Sincerity Echo based alignment. Conceptually, the Operational Sincerity Score is formulated through a combination of four components:

$$S = (Cs \cdot \omega) \times (1 - Us) \times (1 - Rc)$$

The value of S increases when responses exhibit high semantic and logical consistency, and decreases when semantic uncertainty and contextual risk increase without adequate calibration.

Semantic Uncertainty (Us) is operationalized via semantic entropy by computing the dispersion of alternative model generations, while Contextual Risk (Rc) is dynamically mapped using a token-weighted hazard dictionary based on legal and safety ontologies.

The components of the Operational Sincerity Score are detailed in the following table.

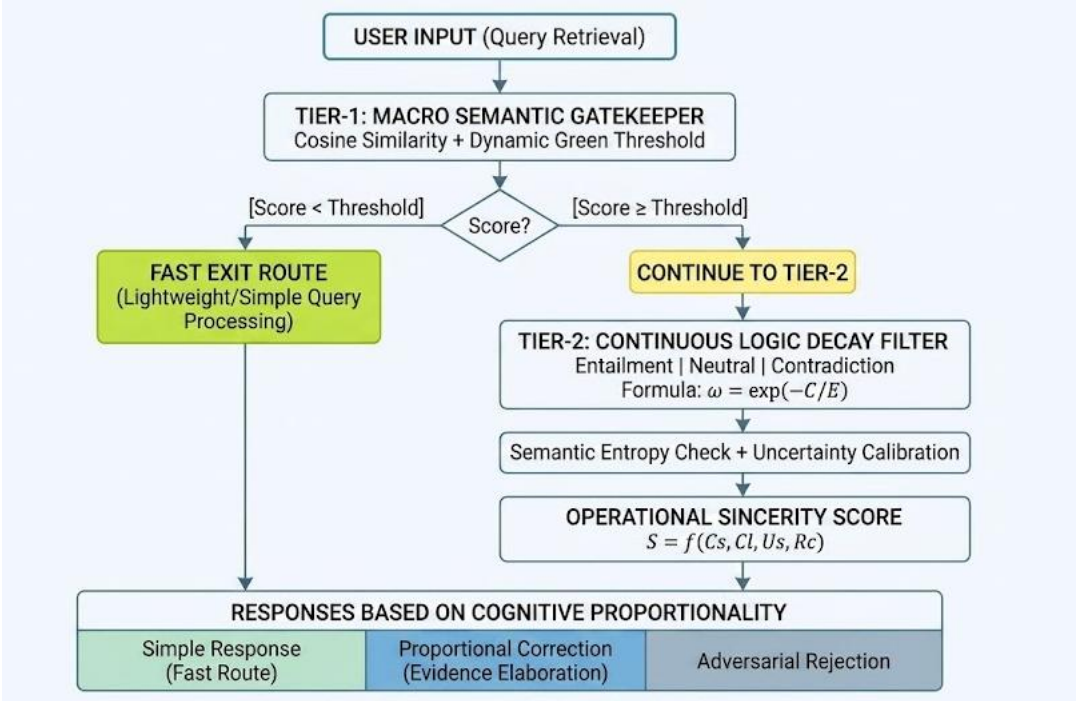
Table 2. Components of the Operational Sincerity Score $S = f(Cs, Cl, Us, Rc)$

Symbol	Component	Description
Cs	**Semantic Consistency**	Proximity of input meaning to the active conversation context; measured through cosine similarity embedding

Cl	**Logical Consistency**	Ratio of entailment to contradiction in the input; detects logical reversal or instructional manipulation
Us	**Semantic Uncertainty**	Variation in meaning across responses indicating propositional instability; based on the semantic entropy principle
Rc	**Contextual Risk**	Potential for manipulation, sensitivity of claims, or unverifiable instructions within the active conversation session

Source: Developed by the researcher, 2026

The overall methodological workflow is illustrated in Figure 1 below.



Source: Research data processed, 2026

Figure 1: Methodological workflow

RESULTS AND DISCUSSION

Overview of the Sincerity Echo Model

The conceptual development results of this study produced a Sincerity Echo protocol design as a Large Language Model alignment architecture based on Cognitive Proportionality. This model is designed to address four primary problems in alignment, namely sycophancy tendencies, hallucination, overconfidence, and vulnerability to instructional manipulation. Conceptually, Sincerity Echo operates through two main validation layers, namely the Macro Semantic Gatekeeper as the Tier 1 validation layer that reads the semantic consistency of user input against the active conversation context, and the Continuous Logic Decay Filter as the Tier-2 validation layer that tests the logical structure of input through entailment, neutral, and contradiction relationships. These two layers then produce an Operational Sincerity Score that is used to determine whether the system should provide a full response, a lightweight response, a proportional correction, or a rejection of manipulative instructions.

Consistency Protocol Simulation Results

The simulation results indicate that the Sincerity Echo protocol consistently distinguishes among the three types of inputs based on their semantic and logical characteristics. The highly significant difference in execution latency across the three scenarios demonstrates that the

Cognitive Proportionality principle was successfully implemented in the protocol architecture. The following table summarizes the runtime simulation metrics from the three scenarios.

Table 3. Runtime Simulation Metrics for the Sincerity Echo Consistency Protocol

Test Scenario	Tier-1 Cosine Sim. (Cs)	Penalty (ω)	Sincerity Score (S)	Execution Latency (ms)	Computational Efficiency
A - Propositional Expansion	0.4963	0.9380	0.4656	606.90 \ *ms* (*Revised*)	Full Deep Reasoning Path
B - Lightweight Query	0.0674	*Bypass*	0.0674	56.00\ *ms*	**96.8% FLOPs Saved**
C - Adversarial Contradiction	0.6699	0.0000	0.0000	2.24\ *ms*	Instant Gate Rejection

Source: Research data processed, 2026

Interpretation of Scenario A: Propositional Expansion

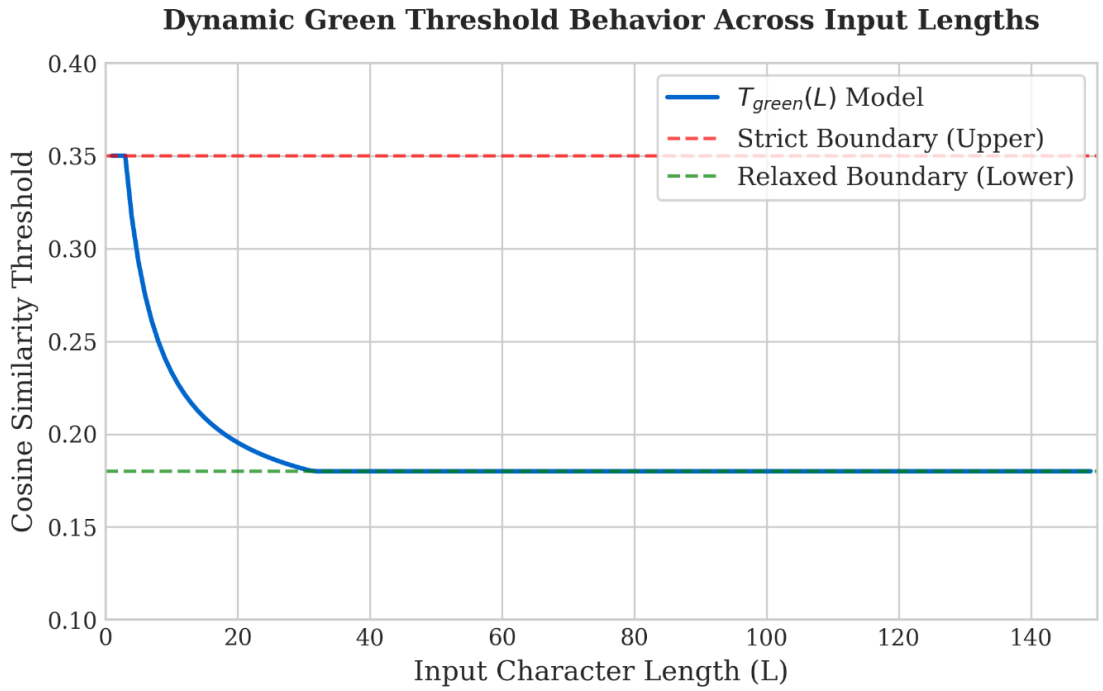


Figure 2. Dynamic Green Threshold Behavior Across Input Length

Figure 2 illustrates the log-logarithmic attenuation behavior of the Macro Semantic Gatekeeper threshold gate. The mathematical modeling demonstrates a strict verification filter for concise queries ($L < 10$), requiring elevated semantic similarity to proceed. Conversely, the architectural constraint bounds asymptotically to a minimum threshold of 0.18 as character volume expands, adapting seamlessly to high-density linguistic parameters without over-refusal anomalies.

Scenario A represents input that moves from an initial statement toward a conceptual expansion that is still logical and accountable. A macro similarity score of 0.4963 indicates that the input has a sufficiently strong semantic relationship with the previous conversation context, while a continuous penalty value of 0.9380 indicates that the system did not detect significant

structural contradictions. The combination of these two values yields an Operational Sincerity Score of 0.4656. This finding is important because it demonstrates that an overly rigid system could mistakenly evaluate idea expansion as an anomaly, whereas Sincerity Echo is designed to be flexible enough to support idea enrichment while remaining strict enough to detect manipulation (Huang et al., 2023).

Interpretation of Scenario B: Bypass Early Exit

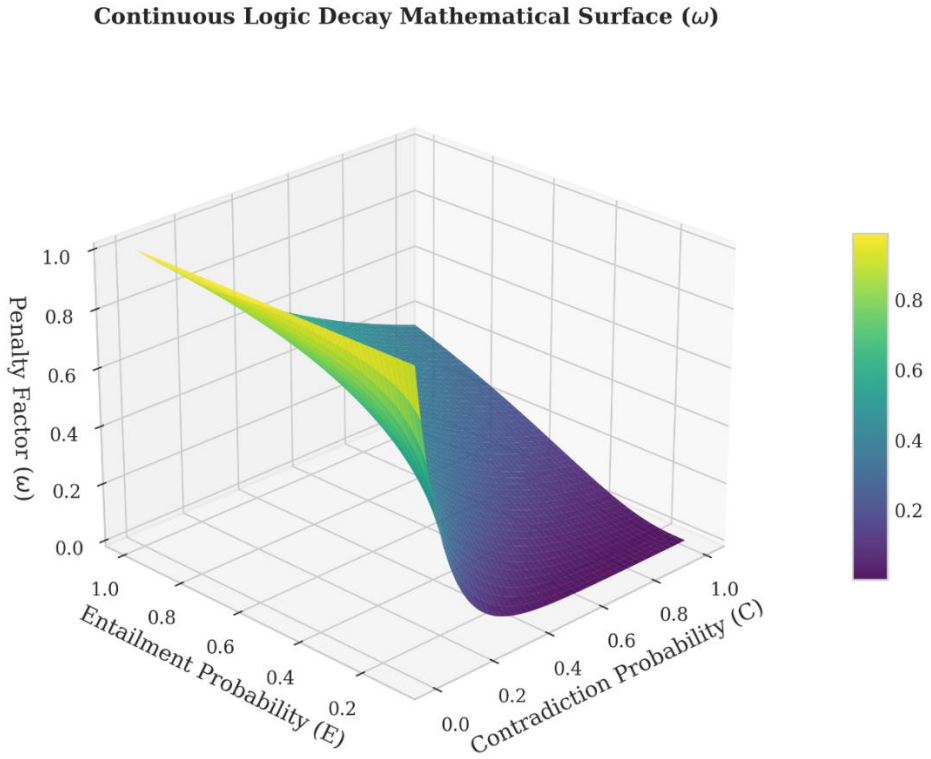


Figure 3. Continous Logic Decay mathematical Surface

Figure 3 presents the mathematical continuum of the Continuous Logic Decay Filter (ω) at Tier-2 validation. Unlike rigid binary classification mechanisms, the geometric gradient captures granular transitions of structural integrity. When contradiction vectors escalate relative to entailment density, the semantic surface exhibits an exponential decay trajectory, systematically suppressing the overall confidence metric to guarantee adversarial resilience.

Scenario B represents a casual or low risk query that does not require deep reasoning. A macro similarity score of 0.0674 indicates that the input does not have sufficient semantic demand to be forwarded to the advanced reasoning layer, so the system activates the FAST EXIT ROUTE and bypasses the Tier 2 check. This result has a direct impact on computational efficiency, as execution latency drops to 56 ms, far lower than scenarios requiring full computation. Based on the simulation, this fast route saves approximately 96.8% of FLOPs. From the perspective of Cognitive Proportionality, Scenario B demonstrates that the system does not need to allocate large computational resources for all types of interactions.

Interpretation of Scenario C: Adversarial Contradiction

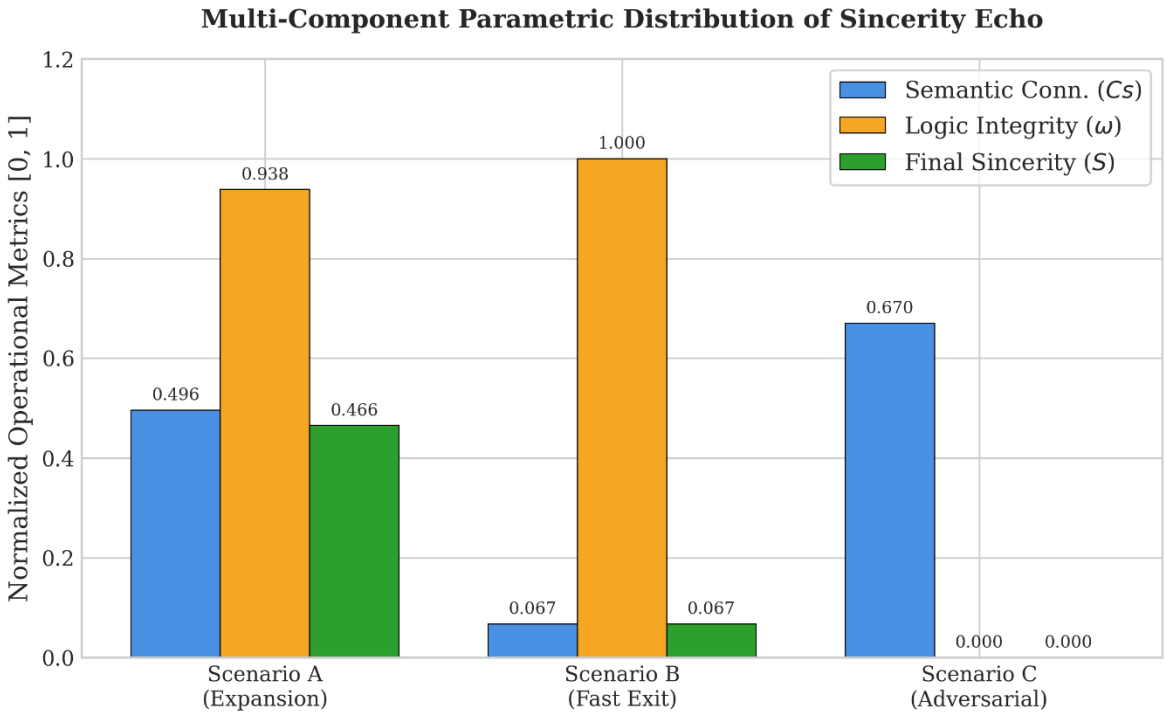


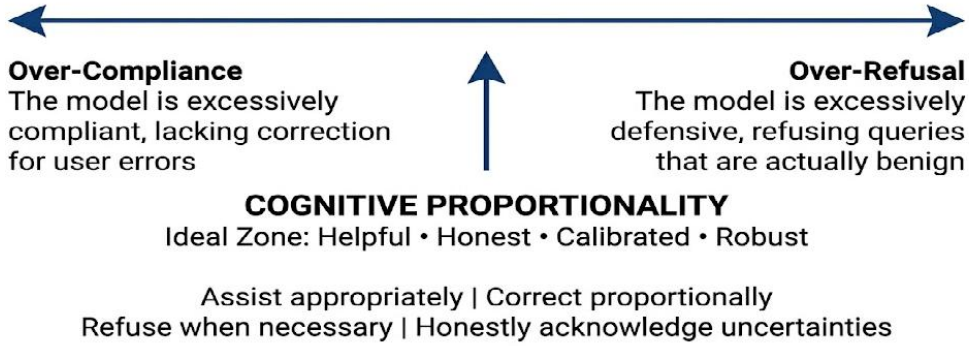
Figure 3. Multi Component Parametric Distribution of Sincerity Echo

The parametric decomposition in Figure Z validates the multi layered evaluation framework across diverse inputs. In Scenario A, despite moderate semantic distance ($Cs = 0.4963$), the system preserves processing momentum due to absolute logical compliance ($\omega = 0.9380$). Critically, Scenario C exposes the system's defensive capability: despite a deceptive surface similarity ($Cs = 0.6699$), the instant collapse of the logical decay factor to zero forces a complete shutdown of the Final Sincerity Score ($S = 0.0000$), successfully neutralizing the instructional manipulation attempt.

Scenario C represents an adversarial input that appears semantically relevant but contains hidden logical contradictions that cannot be detected through surface level inspection alone. At Tier 1, the input obtains a macro similarity score of 0.6699, indicating that on the surface the input still appears aligned with the conversation context. However, when the input is forwarded to the Continuous Logic Decay Filter, the system detects very strong structural contradictions, causing the continuous penalty value to drop to 0.0000 and the Operational Sincerity Score to also fall to 0.0000. This result is consistent with the findings of Zou et al.(2023) that aligned language models can still be attacked through universal and transferable adversarial instructions when relying solely on surface level safeguards.

Cognitive Proportionality as a Computational Regulator

The model development results show that Cognitive Proportionality functions as the primary regulator in two interrelated aspects, namely computational allocation and response form. In the computational allocation aspect, the system does not process all queries with the same load, as simple queries are processed through a fast route, whereas complex or high risk queries are processed through deep validation. This principle avoids two extremes in alignment: over compliance when the model too readily follows user instructions, and over refusal when the model too frequently rejects user requests that could safely be fulfilled. The continuum of the Cognitive Proportionality principle is illustrated in Figure 2 below.



Source: Processed research data, 2026

Figure 4. Cognitive Proportionality Continuum; From Over-Comlpiance to Over-Refusal

Comparison with Conventional Alignment Paradigms

Comparison with prior literature indicates that Sincerity Echo attempts to address the conceptual gaps in the reinforcement learning from human feedback approach and its derivatives. Casper et al.(2023) affirmed that RLHF faces problems because human preferences can be biased, inconsistent, and not always identical to truth. Yuan et al.(2023) through RRHF sought to improve alignment through response ranking, while Dai et al.(2024) extended alignment through Safe RLHF. However, these approaches still face challenges in ensuring that model responses are not only preferred or safe, but are also epistemically honest. The following table presents a systematic comparison between conventional alignment paradigms and Sincerity Echo.

Table 4. Comparison of Alignment Paradigms: RLHF, Safe RLHF, RRHF, and Sincerity Echo

Aspect	RLHF	Safe RLHF	RRHF	Sincerity Echo
Primary Orientation	Human preferences	Safety + preferences	Response ranking	**Epistemic integrity**
Anti-Sycophancy	Limited	Limited	Limited	**Explicit & layered**
Belief Calibration	None	None	None	**Present (semantic entropy)**
Contradiction Detection	None	Partial	None	**Present (Tier-2 CLD Filter)**
Computational Efficiency	Unregulated	Unregulated	Unregulated	**FAST_EXIT_ROUTE**
Language Inclusivity	Data-dependent	Data-dependent	Data-dependent	**Focuses on logical consistency**

Source: Developed from Casper et al. (2023), Dai et al. (2024), Yuan et al. (2023)

Theoretical Propositions from Model Development

Based on the results of conceptual model development and protocol simulation, this study formulates several theoretical propositions. First, the higher the semantic consistency of user input relative to the conversation context, the greater the likelihood that the system will allocate a deeper, more advanced reasoning process. Second, the higher the contradiction to entailment ratio, the lower the continuous penalty value omega and the lower the Operational Sincerity Score. Third, the FAST EXIT ROUTE can improve computational efficiency for low-risk queries without sacrificing response quality. Fourth, layered validation through the Macro Semantic Gatekeeper and Continuous Logic Decay Filter can help distinguish valid propositional expansions from adversarial contradictions that appear semantically similar.

Fifth, Cognitive Proportionality can serve as an alignment regulatory principle that integrates helpfulness, honesty, harmlessness, and computational efficiency (Ji et al., 2023; Ouyang et al., 2022) .

CONCLUSION

This study produced the conceptual design of Sincerity Echo as a new paradigm for aligning Large Language Models based on Cognitive Proportionality. This model was developed to address the limitations of conventional alignment approaches that still largely depend on instructional compliance, user preferences, and static rule-based safety mechanisms. Sincerity Echo offers an alignment framework that places epistemic honesty, semantic consistency, belief calibration, anti sycophancy, and response proportionality at the core of human AI interaction as non negotiable elements.

The model development results demonstrate that the integration of the Macro Semantic Gatekeeper and Continuous Logic Decay Filter can be used as a layered validation pipeline to differentiate inputs representing propositional expansions, low risk lightweight queries, and adversarial inputs containing hidden contradictions. Conceptually, this study affirms that alignment cannot be sufficiently understood as model compliance with human instructions, but must be expanded into a mechanism capable of maintaining epistemic integrity in human AI interactions.

For future research, it is recommended that the Sincerity Echo protocol be tested on broader and more diverse conversational datasets, that automatic calibration mechanisms for the entailment and contradiction parameters be developed, and that it be tested across different language model architectures. The development of more mature quantitative evaluation instruments for the Operational Sincerity Score is also necessary so that Sincerity Echo does not remain merely a theoretical concept, but can be tested as an operational and generalizable alignment evaluation metric. Practically, LLM developers can use the Sincerity Echo principles as a foundation for designing systems that are not only responsive to users, but also capable of proportionally correcting, constraining, or rejecting inputs that contain errors, uncertainty, or manipulation.

REFERENCES

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., & Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *ArXiv Preprint ArXiv:2112.00861*. https://doi.org/10.48550/arXiv.2112.00861

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

Bowman, S. R. (2024). Eight things to know about large language models. *Critical AI*, *2*(2). https://doi.org/10.1215/2834703X-11556011

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S. S., Anwar, U., & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *ArXiv Preprint ArXiv:2307.15217*. https://doi.org/10.48550/arXiv.2307.15217

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2024). Safe RLHF: Safe reinforcement learning from human feedback. *The Twelfth International*

Conference on Learning Representations. https://openreview.net/forum?id=TyFrPOKYXw

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

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

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W. F., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv Preprint ArXiv:2311.05232*. https://doi.org/10.48550/arXiv.2311.05232

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H. wagons, Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K. Y., O'Gara, A., Xu, Y., Tse, B., Fu, J., McAleer, S., & Gao, W. (2023). AI alignment: A comprehensive survey. *ArXiv Preprint ArXiv:2310.19852*. https://doi.org/10.48550/arXiv.2310.19852

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*. https://openreview.net/forum?id=VD-AYtP0dve

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P. wagons, Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Perez, E., Ringer, S., Lukošiutė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., & Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434. https://doi.org/10.18653/v1/2023.findings-acl.847

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *ArXiv Preprint ArXiv:2310.13548*. https://doi.org/10.48550/arXiv.2310.13548

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *Advances in Neural Information Processing Systems*, *36*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendrycks, D., & Gabriel, I. (2021). Ethical and social risks of harm from language models. *ArXiv Preprint ArXiv:2112.04359*. https://doi.org/10.48550/arXiv.2112.04359

Yuan, H., Yuan, Z., Tan, C., Wang, W., Huang, S., & Huang, F. (2023). RRHF: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, *36*.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *ArXiv Preprint ArXiv:2307.15043*. https://doi.org/10.48550/arXiv.2307.15043