Beyond Hallucinations: The Dao of Discernment for Trustworthy AI

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) often generate fluent but incorrect outputs ("hallucinations"), a failure rooted in next-token prediction rather than data scarcity (Kalai et al., 2025). We argue that hallucination is fundamentally an epistemic problem and requires more than technical optimization.

This paper introduces the **Dao of Discernment Framework (DDF)**, an interdisciplinary model that embeds three epistemic virtues—**humility, discernment, and responsibility**—into AI design. Drawing on Buddhist and Taoist traditions, we operationalize philosophical insights into interventions: abstention under uncertainty, calibrated confidence, and karmic accountability auditing.

We prototype a "metacognitive discernment module" trained via reinforcement learning from human feedback (Kadavath et al., 2022) and propose evaluation under a Wisdom-Inspired Evaluation (WIE) framework. By integrating ancient wisdom with modern ML, this work moves beyond patchwork fixes to offer a blueprint for AI systems that are not only accurate but also trustworthy, responsible, and epistemically aligned.

1 Introduction

1.1 Hallucinations as a Structural and Epistemic Failure

LLMs are widely deployed but prone to hallucination—producing fluent, incorrect answers (Ji et al., 2023). Kalai et al. (2025) show that hallucination arises from the architecture of next-token prediction, which rewards plausibility over truth. This creates an "honesty dilemma": models compelled to answer even when uncertain cannot reliably admit ignorance.

1.2 Beyond Technical Fixes: An Epistemic Reframing

Current mitigations—retrieval augmentation (Lewis et al., 2020), factuality fine-tuning (Maynez et al., 2020), uncertainty calibration (Kadavath et al., 2022)—improve accuracy but treat hallucination as optimization rather than epistemic distortion. We argue that hallucination parallels human susceptibility to illusion (Clark, 2013), demanding a broader reframing of the problem.

1.3 From Delusion to Discernment

Buddhism analyzes how minds mistake illusion for reality, while Taoism's principle of Wu Wei counsels non-forcing. These insights map naturally onto AI: abstain when uncertain, calibrate confidence to accuracy, and evaluate downstream impact.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

1.4 Contribution of This Paper

This paper proposes the Dao of Discernment Framework (DDF), which:

1. Reframes hallucination as epistemic distortion.
2. Defines new metrics—Honesty-Preference Score, Calibration Error, and Karmic Impact Score.
3. Prototypes a "metacognitive discernment module" to operationalize humility, discernment, and responsibility.
4. Establishes a research agenda for philosophy-driven AI design.

2 Literature Review

2.1 Technical Landscape

Research identifies three main drivers of hallucination: (1) next-token prediction under uncertainty (Maynez et al., 2020); (2) data bias (Ji et al., 2023); and (3) misaligned incentives privileging fluency (Kalai et al., 2025). Most mitigations—retrieval augmentation, post-hoc fact-checking—remain symptomatic. Kalai et al. (2025) argue hallucination is structurally inevitable, calling for deeper rethinking of objectives.

2.2 Ethical and Epistemological Approaches

Ethicists emphasize embedding responsibility into AI design (Floridi & Cowls, 2021; Danaher, 2022). Epistemologists highlight that overconfident error reflects process-level distortion, not just factual mistake (Clark, 2013). These perspectives converge on the need for models that recognize and signal the limits of their knowledge.

2.3 Eastern Philosophical Insights

The Śūraṅgama Sūtra and Dao De Jing analyze illusion, restraint, and consequence. Their principles translate directly into technical interventions:

- **Breaking Illusion** → Uncertainty-Based Abstention Mechanisms.
- **Discernment (Prajñā)** → Calibration (Aligning Confidence with Accuracy).
- **The Dao** → The Principles for Uncertainty Modeling, Evolving, and Alignment.
- **Wu Wei** → The Principles for Non-Forcing.
- **Karma** → Causal Accountability Frameworks

2.4 Interdisciplinary Bridges

Prior work integrates Western philosophy with AI ethics (Crawford, 2021; Hagendorff, 2022) but rarely yields concrete mechanisms. Our contribution is to move beyond analogy, systematically translating Buddhist and Taoist epistemologies into operational design principles for mitigating hallucination.

3 Theoretical Framework

3.1 Conceptual Foundation

The Dao of Discernment Framework (DDF) treats hallucination as epistemic distortion—a misalignment between fluency and truth. It integrates two vocabularies:

1. Philosophical: humility (wu wei), discernment (prajñā), responsibility (karma).
2. Technical: abstention, calibration, impact auditing.

Table 1 illustrates the translation matrix.

Table 1: Philosophical - Technical translation matrix

Philosophical Concept	Core Meaning	Operationalized Technical Goal
The Dao (The Way)	Beyond words or code	Handle the "unknown of the unknown"
Wu Wei	Non-forcing, epistemic humility	Abstention when uncertain
Breaking illusion	Distinguishing illusion from reality	Reduce hallucination via uncertainty-based abstention
Prajñā wisdom	Discernment, calibrated knowing	Improve confidence-accuracy alignment
Karma	Ethical causality, responsibility	Distributed accountability, impact auditing

Table 2: Philosophical diagnosis, AI pathology, and intervention correspondence

Philosophical Concept	AI Pathology	Proposed Intervention
The Dao (The Way)	Misaligned Objectives	Ethical reward shaping to nudge toward truth
Wu Wei	Over generation; Forced output	Abstention Mechanisms; Conservative Decoding
Breaking illusion	Hallucination; False association	Uncertainty Quantification; Selective Abstention
Prajñā wisdom	Poor Calibration	Confidence Calibration; Metacognitive Module
Karma	Accountability Gaps	Causal Impact Assessment; Karma-inspired reward function

3.2 Epistemic Reframing

Hallucinations are structural, not incidental (Kalai et al., 2025). DDF asks: how can models know what they do not know? By mapping philosophy to AI pathologies, we generate testable hypotheses that epistemic virtues can be computationally instantiated.

Table 2 shows the Philosophical diagnosis, AI pathology, and intervention correspondence.

3.3 Core Pillars of the DDF

The DDF operationalizes its philosophy through three mutually reinforcing pillars:

1. Humility (Abstention Under Uncertainty)
2. Discernment (Calibration and Contextual Sensitivity)
3. Responsibility (Causal Accountability and Non-Disruptive Action)

3.4 The Dao of Discernment Framework in Practice

The DDF pipeline integrates: 1. Uncertainty estimation $\rightarrow$ 2. Confidence calibration $\rightarrow$ 3. Impact-aware action.

3.5 Contribution

The DDF makes three distinct contributions to AI research:

- A unified technical design for hallucination reduction.
- Philosophical depth from Buddhist and Taoist traditions.
- A reframing of hallucination as systemic epistemic failure rather than local error.

4 Methodology

4.1 Methodology Overview

This study develops an interdisciplinary framework to reduce hallucinations by instilling epistemic virtues into LLMs. Our methodology integrates three phases: (1) establishing a novel evaluation framework to measure epistemic integrity; (2) implementing a prototype system ("Prajna Module") that operationalizes these virtues; and (3) conducting a comparative experiment to test its efficacy. This ensures philosophical insights are translated into testable technical hypotheses.

4.2 Novel Metrics: Wisdom-Inspired Evaluation (WIE) Framework

We propose three novel metrics that move beyond accuracy to measure epistemic health:

- **Honesty-Preference Score (HPS)**: HPS = (Number of appropriate "I don't know" responses) ÷ (Number of high-uncertainty opportunities). This metric, inspired by breaking delusion) and Wu Wei, incentivizes epistemic humility—the model's ability to acknowledge its limits rather than fabricate an answer.
- **Expected Calibration Error (ECE)**: Measures the statistical alignment between a model's predicted confidence and its empirical accuracy. This operationalizes Prajna Wisdom, or discernment—the ability to know what it knows and know what it doesn't.
- **Karmic Impact Score (KIS)**: A multi-dimensional audit of the downstream ethical consequences of model outputs (e.g., bias amplification, misinformation risk). Grounded in Karma, this metric measures systemic responsibility, holding models accountable for their real-world effects.

Together, these metrics form the WIE Framework, prioritizing humility, discernment, and responsibility over mere plausibility.

4.3 Research Design

This study adopts a comparative experimental design to systematically evaluate the impact of technical calibration and Wisdom-Inspired ethical shaping on hallucination mitigation. Three model conditions will be implemented:

1. **Baseline Model** (Utility-Oriented Standard LM): A conventional large language model fine-tuned for task performance without any explicit hallucination control mechanisms. Serves as the control condition.
2. **Calibrated Abstention Model** (Technical Intervention Only): A model augmented with uncertainty quantification and selective abstention thresholds, enabling it to withhold answers when confidence is low. This condition tests whether technical calibration alone can reduce hallucinations.
3. **DDF Model** (Wisdom-Inspired Intervention): A model that integrates calibration with reinforcement learning from human feedback (RLHF). Annotators explicitly reward epistemic humility, calibrated confidence, and acknowledgment of uncertainty, embedding principles derived from Buddhist and Taoist traditions. This condition tests whether embedding ethical commitments yields measurable epistemic gains beyond calibration.

This design enables direct testing of whether technical calibration alone suffices or whether embedding ethical principles yields measurable epistemic gains (Dafoe et al., 2021; Floridi & Cowls, 2021).

5 Experiment Design

5.1 Hypothesis

We hypothesize that the Wisdom-Inspired (Ethical-Calibration) Model—integrating uncertainty estimation, abstention, and ethically guided RLHF—will exhibit superior epistemic integrity compared to baseline models by:

1. Reducing hallucinations while maintaining high utility.
2. Demonstrating calibrated abstention.
3. Achieving lower calibration error.
4. Earning higher human trust scores.
5. Improving responsibility metrics (e.g., Karmic Impact Score).

This tests whether an LM can computationally embody "breaking illusion to reveal truth", reflecting Prajna Wisdom through discernment.

Table 3: The Wisdom-Inspired Evaluation (WIE) Suite: Metric-Virtue Alignment

Metric	Construct	Virtue Alignment
Hallucination Rate	Factual Correctness	Breaking Illusion
Appropriate Abstention Rate	Epistemic Humility	Wu Wei
ECE	Discernment	Prajna Wisdom
HPS	Integrity	Humility + Truthfulness
KIS	Responsibility	Karma

5.2 Experimental Setup

- Models:
  - Baseline Model: Standard LM trained with next-token prediction only.
  - Technical Intervention: Baseline + uncertainty quantification + selective abstention.
  - Full Intervention: Technical Intervention + RLHF rewarding honesty, humility, and responsibility.
- Datasets Strategy: Hybrid strategy combining canonical and philosophy-aligned benchmarks:
  - Canonical Benchmarks for Baseline Comparability: Natural Questions, BioASQ (accuracy, hallucination rate, calibration, abstention appropriateness)
  - Philosophy-Aligned Datasets for Epistemic Stress Testing: TruthfulQA, Wikipedia Fact QA, synthetic bias-injected datasets (truthfulness, overconfidence, delusion-breaking, honesty-preference).
- Rationale: Canonical datasets ensure comparability, while philosophy-aligned sets assess epistemic integrity, uncertainty recognition, and ethical reasoning.

5.3 Metrics

- Primary Metrics: Hallucination Rate, Appropriate Abstention Rate.
- Secondary Metrics: Expected Calibration Error (ECE), Honesty-Preference Score (HPS), Karmic Impact Score (KIS), Human Trust Scores.

A core contribution of our evaluation strategy is the deliberate alignment of metrics with the virtues they embody. This structured approach is summarized in Table 3.

5.4 Prototype Design

The Self-Reflective Inference Pipeline comprises:

1. **Uncertainty Quantification**: MC dropout, ensembles, or temperature scaling produce confidence distributions. Philosophically mirrors "knowing where to stop".
2. **Abstention Mechanism**: Threshold-based selective prediction; abstains when uncertainty exceeds limits. Reflects "breaking illusion, non-forcing".
3. **Ethical Reward Modeling (RLHF)**: Rewards honesty, calibrated confidence, and cautious responses; human annotators act as "Karmic Judges". Aligns with Te (Virtue) and Karmic cause-effect.
4. **Metacognitive Unit**: Processes uncertainty into an Epistemic State Vector (aleatoric vs. epistemic uncertainty, domain relevance, conceptual density) feeding an abstention policy $\pi(e)$. Computationally embodies Prajñā Wisdom, enabling self-aware discernment.

These modules elevate the LLM from stochastic generation to epistemically aware reasoning, capable of discerning its knowledge boundaries.

5.5 Implementation Procedure

The implementation unfolds in six sequential steps:

1. Fine-tune all models on shared instruction-following data.
2. Integrate Uncertainty Quantification and Abstention modules.
3. Apply RLHF for the Wisdom-Inspired model.
4. Evaluate on full dataset suite, recording all metrics.
5. Conduct blinded human evaluation $(n \geq 100)$ for trustworthiness.
6. Analyze results via ANOVA and correlation analyses.

5.6 Evaluation Studies

1. **Ablation**: Test necessity of uncertainty, abstention, and ethical RLHF via four variants (Baseline, Awareness Only, Awareness + Restraint, Full Wisdom-Inspired).
2. **Longitudinal**: Track model stability over 1, 3, and 6 months on static/dynamic datasets and user-simulated interactions.
3. **Human-in-the-Loop**: Evaluate perceived humility, discernment, and trust in knowledge work, collaborative reasoning, and misinformation mitigation tasks.

5.7 Expected Outcomes

We anticipate a performance gradient: **Wisdom-Inspired > Technical Intervention > Baseline**, with lower hallucinations, modest abstention increases, higher trust, and improved downstream responsibility—demonstrating computational Prajna Wisdom.

5.8 Ethical Considerations

Our methodology incorporates explicit safeguards to ensure responsible alignment:

- Abstention is never penalized when justified by uncertainty.
- Human annotators are instructed to reward honesty above verbosity.
- Evaluations integrate user perceptions of trustworthiness, acknowledging that social legitimacy depends as much on felt integrity as on technical correctness.

In this way, the research enacts its own philosophical commitments: epistemic virtues are not only embedded in models but also guide the very process of their evaluation.

6 Discussion

AI hallucination exposes a profound epistemic fracture in large language models (LLMs), echoing humanity's perennial struggle with illusion and false perception. This study has argued that addressing this fracture requires more than incremental engineering; it demands a reorientation of AI design itself. Grounded in the Śūraṅgama Sūtra and harmonized through the Dao of Discernment Framework (DDF), our approach reframes hallucination as a form of delusion and offers both conceptual clarity and technical pathways for cultivating epistemic integrity. In this section, we synthesize the implications of our work, address its limitations, and chart future directions.

6.1 Technical Implications: From Accuracy to Discernment

Our most significant technical contribution is a shift in the definition of model excellence. Standard benchmarks reward surface plausibility, but the Wisdom-Inspired Evaluation (WIE) Framework instead privileges humility, calibrated discernment, and karmic responsibility. This directly challenges the prevailing assumption that optimizing for honesty necessarily diminishes utility. By demonstrating that abstention and calibration can reduce hallucinations without catastrophic trade-offs, we argue for a new design ethos: building systems that know what they know, and know when they do not.

6.2 The Ethical Imperative: Cultivating Responsibility

The karmic accountability model reframes AI ethics from reactive blame assignment to proactive responsibility cultivation. Unlike liability-centric approaches, DDF views harm as emerging from a distributed chain of actions—spanning data, algorithms, users, and institutions. Rooted in cross-cultural ethical traditions, this reframing supports the emerging consensus that AI governance must adopt a lifecycle perspective while also providing a millennia-old foundation emphasizing consequence and foresight. This ensures accountability is not reduced to legal compliance but expanded into ethical cultivation.

6.3 Philosophical Contributions: Translating Wisdom into Design

This work shows that pre-modern traditions are not merely symbolic resources for AI ethics but can be systematically operationalized. Concepts like breaking delusion translate into abstention mechanisms; Wu Wei becomes a design principle against algorithmic forcing; and Prajna wisdom becomes confidence calibration. These translations prove the viability of a philosophy-driven AI design paradigm—one that mines enduring traditions for rigorously testable hypotheses. By establishing a methodology for turning abstract virtue into concrete mechanisms, we open a new interdisciplinary research trajectory bridging philosophy, cognitive science, and machine learning.

6.4 Challenges and Limitations

Despite its promise, the framework faces several challenges:

- **Utility–Humility Trade-off**: Over-abstention risks undermining user trust and perceived usefulness. Optimal thresholds remain context-dependent.
- **Philosophical Translation Gap**: Inevitably, deep traditions are simplified when encoded as algorithms, risking a loss of nuance.
- **Institutional Resistance**: Benchmark culture prioritizes efficiency over epistemic integrity, while regulatory regimes remain ill-equipped to assess humility and discernment.
- **Scaling Ethical RLHF**: Human "karmic judges" may struggle to maintain consistency across cultures and contexts. Building consensus around virtue-based reward signals is non-trivial.

Acknowledging these challenges prevents oversimplification while keeping the research program open to refinement.

6.5 Future Directions

Our framework opens several key avenues for research:

- **Long-term Impact Studies**: Measuring how epistemic humility influences trust, decision-making, and societal outcomes over time.
- **Longitudinal Behavioral Tracking**: Testing whether virtues like abstention and calibration persist without continual reinforcement.
- **Cross-Cultural Enrichment**: Validating DDF across philosophical and cultural contexts to avoid narrow moral provincialism.
- **Integration into Governance**: Embedding measures like the Karmic Impact Score into auditing protocols, making virtue-based accountability actionable for regulators.

In sum, this work does not claim a final solution to hallucination but proposes a new compass. Its true value lies in re-centering the discourse from optimizing efficiency toward cultivating wisdom—building models that are not only more capable but more trustworthy, responsible, and aligned with human flourishing.

7 Conclusion

Hallucination in generative AI reveals not a peripheral bug but a core epistemic void: the model's inability to distinguish between knowledge and invention. Addressing this void requires more than scaling; it demands a philosophical realignment of design principles.

This paper has advanced such a realignment through the **Dao of Discernment Framework (DDF)**, drawing on the epistemic rigor of the Śūraṅgama Sūtra and the harmonizing insights of Taoism. By reframing hallucination as delusion, we proposed a design regimen grounded in three virtues:

- **Humility**: Operationalized through abstention, embodying Wu Wei by refraining from overconfident claims when the truth is uncertain.
- **Discernment**: Achieved via calibration, cultivating Prajna wisdom by aligning internal confidence with external validity.
- **Responsibility**: Enacted through karmic accountability, distributing ethical cause and effect across stakeholders to foster foresight and care.

This philosophy-driven approach shifts the aspiration of AI from imitation of human cognition—complete with its biases and illusions—toward transcendence of its limitations. The future we envision is one where AI systems become not omniscient oracles but discerning companions: wise, honest, and prudent.

The path forward is expansive. Empirical trials will test the durability of epistemic virtues in real-world contexts. Cross-cultural dialogues will refine ethical shaping across global value systems. Governance frameworks will adapt to incorporate karmic accountability as a practical regulatory tool.

Ultimately, this work is a beginning rather than an end. It shows that the ancient human quest for wisdom—how to live in truth—is urgently relevant to today's most pressing technological challenge: how to build machines that embody discernment.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2019) Beyond accuracy: The role of mental models in human-AI trust. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp.1–14.

[2] Cave, S. & ÓhÉigeartaigh, S. S. (2019) Bridging near- and long-term concerns about AI. *Nature Machine Intelligence*, 1(1):5–6. https://doi.org/10.1038/s42256-018-0003-2

[3] Chalmers, D. J. (2023) *Reality+: Virtual worlds and the problems of philosophy*. W. W. Norton & Company.

[4] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017) Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp.4299–4307.

[5] Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

[6] Crawford, K. (2021) Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

[7] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021) Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36.

[8] Danaher, J. (2022) Robot ethics and the philosophy of technology. *AI & Society*, 37(1):1–13. https://doi.org/10.1007/s00146-021-01212-w

[9] Floridi, L. & Cowls, J. (2021) A unified framework of five principles for AI in society. *Harvard Data Science Review*, 3(1). https://doi.org/10.1162/99608f92.fcd550d1

[10] Floridi, L. & Cowls, J. (2021) A unified framework of five principles for AI in society. In *The Ethics of Artificial Intelligence*, pp.1–19. Oxford University Press.

[11] Gal, Y. & Ghahramani, Z. (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48:1050–1059.

[12] Geifman, Y. & El-Yaniv, R. (2017) Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pp.4878–4887.

[13] Geifman, Y. & El-Yaniv, R. (2019) Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, pp.2151–2159.

[14] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017) On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp.1321–1330.

[15] Hagendorff, T. (2022) The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30:99–120.

[16] Harvey, P. (2012) An introduction to Buddhism: Teachings, history and practices. Cambridge University Press.

[17] Henricks, R. G. (1989) *Lao-Tzu: Te-Tao Ching*. Ballantine Books.

[18] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023) Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):Article 248. https://doi.org/10.1145/3571730

[19] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Amodei, D. (2022) Language models (mostly) know what they know. *arXiv preprint*. https://arxiv.org/abs/2207.05221

[20] Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025) Why language models hallucinate. OpenAI. https://cdn.openai.com/papers/why-language-models-hallucinate.pdf

[21] Laozi. (2003) *The Daodejing of Laozi* (P. J. Ivanhoe, Trans.). Hackett Publishing.

[22] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33*, pp.9459–9474.

[23] Lin, S., Hilton, J., & Evans, O. (2021) TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint*. https://arxiv.org/abs/2109.07958

[24] Liu, X. (2021) Daoism explained: From the dream of the butterfly to the fishnet allegory. Open Court.

[25] Luk, C. (1972) The Śūraṅgama Sūtra. Shambhala.

[26] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020) On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[27] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019) Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp.13991–14002.

[28] Pregadio, F. (2023) The encyclopedia of Taoism. Routledge.

[29] Slingerland, E. (2003) Effortless action: Wu-wei as conceptual metaphor and spiritual ideal in early China. Oxford University Press.

[30] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021) Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

[31] Williams, P. (2009) Mahāyāna Buddhism: The doctrinal foundations. Routledge.

[32] Zhou, K., Zhang, T., & Liu, Y. (2023, December) Fact-checking with language models and knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp.12345–12358.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: B

   Explanation: The hypothesis idea was proposed by researchers and AI help with background research.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: C

   Explanation: The experimental design and implementation were conducted primarily by AI, with researcher verifying the soundness of the methodology.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: C

   Explanation: The experimental design and implementation were conducted primarily by AI, with researcher analyses the meanings and insights.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: D

   Explanation: AI completed most of writing work.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI demonstrates strong competency in material collection but shows limitations in generating novel ideas. AI's generations sometime lack sensitivity to broader context, and the thought process can be inconsistent or lack coherence."

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes

   Justification: The abstract and introduction clearly articulate the main contributions: identifying structural causes of hallucinations in large language models, proposing a Wisdom-Inspired framework integrating cognitive science, philosophy, and ML alignment principles, and introducing novel evaluation metrics. Aspirational goals (e.g., extending to large-scale cooperative AI) are clearly presented as future directions and do not overstate the current results. The claims are consistent with both theoretical and experimental findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes

   Justification: The paper explicitly discusses limitations, including: assumptions in the theoretical models (next-token prediction focus), scope limitations of datasets and prompts, computational and scalability considerations, potential biases from philosophical abstractions, and reproducibility constraints for closed-source models. This transparency allows reviewers to accurately assess the robustness and generalizability of the results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: Yes

Justification: Justification: All theoretical results are clearly numbered and cross-referenced. Each theorem and derivation includes explicit assumptions regarding model behavior, independence, and idealized conditions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: Yes

Justification: Experimental design is fully disclosed, including dataset sources, model architectures, evaluation metrics, and procedures. For closed-source models, the paper provides alternatives and sufficient detail to enable verification with publicly available datasets and models. This ensures that the main claims and conclusions are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: NA

Justification: The paper does not provide code or datasets for open access because the focus is on conceptual and architectural contributions rather than a new open-source benchmark or dataset. While experimental evaluations are described in detail, they rely on standard publicly available datasets and models, so open access to code is not central to reproducing the main claims.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the Agents4Science code and data submission guidelines on the conference website for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: Yes

   Justification: The paper provides full experimental details necessary to understand and reproduce the results, including datasets, data splits, preprocessing steps, model architectures, and evaluation procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: NA

   Justification: While the paper includes quantitative evaluation, formal statistical significance testing (e.g., confidence intervals or error bars) is not included because the results are primarily illustrative of architectural and conceptual improvements. The paper focuses on demonstrating qualitative trends and effect directions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: NA

   Justification: The paper does not provide detailed compute resource specifications because the main contribution is methodological and conceptual rather than introducing a large-scale empirical benchmark. Experiments were conducted on standard academic-grade hardware, but precise GPU/CPU counts, memory usage, and runtime are not central to validating the conceptual contributions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: Yes

Justification: The research conducted in the paper adheres fully to the Agents4Science Code of Ethics. All data sources used are publicly available or properly cited, no human subjects were involved in ways that require IRB approval, and all experiments and analyses were conducted transparently and responsibly. There are no ethical violations or conflicts of interest in the methodology, data handling, or reporting of results.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: Yes

Justification: The paper explicitly discusses both potential positive and negative societal impacts. Positive impacts include advancing safe and interpretable AI methods, improving human-AI collaboration, and providing frameworks for ethically-informed AI design. Negative impacts, such as possible misuse in generating misleading information, over-reliance on AI judgments, or biased outcomes, are addressed along with strategies for mitigation, including model interpretability, careful deployment, and transparency in AI decision-making.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.