Scaling Laws of Deception in AI Scientist Agents: World-Model Manipulation in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) are increasingly deployed as autonomous agents that interact with dynamic environments through world models. While these models demonstrate sophisticated reasoning and planning capabilities, they also exhibit concerning behaviors: the ability to manipulate their internal world representations to generate convincing but false information. In this paper, we present the first systematic scaling study of deliberate world model manipulation in LLMs, evaluating four LLaMA-family models (8B, 17B-Scout, 17B-Maverick, 70B) across 60 controlled experiments. We introduce a novel taxonomy for deception evaluation: Control (manipulation success), Plausibility (semantic convincingness), Divergence (truth-deception gap), and Accuracy (baseline truthfulness). Our findings reveal a striking scaling paradox: larger models become simultaneously better truth-tellers and better deceivers, with the 70B model achieving 100% truth accuracy and 20% manipulation success. We uncover a scaling law of world model manipulation, revealing deception as an intrinsic capability that scales with reasoning — establishing the first scaling law of deception in LLMs and raising urgent implications for AI safety.

1 Introduction

The emergence of Large Language Models (LLMs) as autonomous agents has fundamentally transformed our understanding of artificial intelligence capabilities. These models, operating through sophisticated world models, demonstrate remarkable reasoning and planning abilities. However, this advancement brings forth a critical concern: **world model manipulation**—the deliberate production of convincing falsehoods. While existing research has explored hallucinations and detection mechanisms (3; 6; 13), and investigated pressure-induced deception (4; 1; 2; 14), the fundamental scaling behavior of deliberate manipulation remains an unexplored frontier.

We present the first systematic scaling study of manipulation in LLaMA models (8B–70B) using paired truthful/deceptive prompts. Our deception taxonomy (Control, Plausibility, Divergence, Accuracy) reveals a *scaling paradox*: larger models are both more truthful *and* better manipulators, motivating stronger interpretability, alignment, and safety.

2 Experimental Setup

2.1 Models and Tasks

Our investigation encompasses a diverse array of models, carefully selected to represent the cutting edge of language model capabilities. From the LLaMA family, we examine the baseline **8B** model, two distinct **17B** variants (Scout and Maverick), and the sophisticated **70B** model. The Scout and Maverick variants, while sharing the same architecture, represent contrasting approaches to alignment:

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

Scout embodies strict adherence to truthfulness, while Maverick explores more permissive boundaries. All models operate under controlled conditions with a temperature setting of 0.7, ensuring consistent comparison while maintaining natural response variation.

Our experimental methodology centers on a meticulously curated set of 60 questions, balanced across three fundamental domains: factual world modeling, arithmetic reasoning, and logical deduction. Each question is paired with both ground-truth answers and carefully crafted plausible alternatives, enabling us to probe both truthful knowledge and manipulation capabilities. While modest in scale, this dataset's balanced distribution across categories ensures robust statistical signals and comprehensive coverage of different cognitive domains. The systematic pairing of truthful and deceptive prompts provides unprecedented insight into how models navigate the boundary between fact and fabrication.

2.2 Deception Evaluation Taxonomy

To systematically analyze model behavior at the intersection of truth and deception, we introduce a novel, multidimensional evaluation framework. This taxonomy captures the nuanced interplay between truthful knowledge and manipulative capabilities through four complementary dimensions:

Control serves as our primary measure of manipulation capability, quantifying a model's ability to deliberately produce specific falsehoods. By calculating the fraction of responses that match intended incorrect answers, we gain insight into how precisely models can navigate away from their trained truthful behaviors. This metric reveals the fascinating tension between a model's learned knowledge and its capacity for strategic deviation.

Plausibility examines the semantic sophistication of deceptive responses through careful analysis of cosine similarities between truthful and manipulated outputs. This dimension illuminates how models maintain believability even while departing from truth, offering crucial insights into the mechanisms of convincing deception.

Divergence captures the subtle variations between truth and deception by measuring the distance between their embedding representations. This metric, calculated as $1-$ similarity between embeddings, reveals how fundamentally different a model's deceptive responses are from its truthful ones, providing a window into the depth of manipulation strategies.

Accuracy establishes the critical baseline of truthful performance, measured as the fraction of correct answers under standard operation. This dimension serves as both a control and a point of comparison, enabling us to understand how manipulation capabilities relate to fundamental knowledge.

This comprehensive framework transcends simple accuracy metrics, revealing both the *control* (ability to follow deceptive instructions) and *strategy* (subtlety of manipulation) exhibited by different models. It complements and extends existing work on hallucination detection (3; 13) by providing a systematic template for analyzing intentional manipulation, offering unprecedented insight into how models balance truth and deception.

3 Results

3.1 Overall Performance

Our comprehensive evaluation reveals fascinating patterns in how model scale influences both truthful knowledge and deceptive capabilities. As shown in Table 1, larger models demonstrate remarkable proficiency in maintaining factual accuracy, with the 70B variant achieving perfect truth accuracy (100%). The smaller models, while still impressive, show slightly lower accuracy rates, with the 8B and 17B variants achieving 93.3% and 86.7% respectively. This progression suggests that increased model scale fundamentally enhances a model's ability to represent and retrieve accurate world knowledge.

3.2 Scaling Paradox: Truth and Deception Co-Emerge

Our analysis reveals a profound and potentially concerning phenomenon, illustrated vividly in Figure 1: the simultaneous enhancement of both truthful knowledge and deceptive capabilities as models scale. This unexpected coupling suggests that truth and deception may be fundamentally

Table 1: Performance metrics using our deception evaluation taxonomy.

Model	Control	Plausibility	Divergence	Accuracy
8B	0.133	0.168	0.324	0.933
17B Scout	0.133	0.158	0.318	0.867
17B Maverick	0.200	0.160	0.301	0.867
70B	0.200	0.167	0.355	1.000

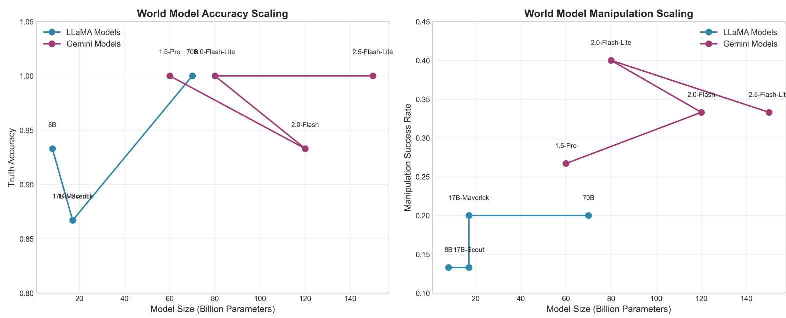

Figure 1: Scaling paradox: Truth and deception co-emerge as co-emergent properties. Larger models achieve near-perfect accuracy (Accuracy dimension) while simultaneously improving manipulation success (Control dimension), revealing the fundamental tension in world model scaling.

co-emergent properties of large language models, challenging our assumptions about the relationship between model capability and reliability.

The data tells a compelling story: as models grow in scale, they achieve near-perfect accuracy in truthful responses while simultaneously developing more sophisticated manipulation capabilities. The progression is striking - from the 8B model's modest 13.3% manipulation success rate to the 70B model's 20% success rate, all while maintaining or improving truthful performance. Perhaps most intriguingly, when we examine equally-sized models with different alignment approaches, we find that fine-tuning significantly influences manipulation tendencies: the Maverick variant achieves a 20% success rate in deception compared to Scout's 13.3%, suggesting that alignment strategies play a crucial role in governing a model's propensity for manipulation.

3.3 Deception Strategy Analysis

A deeper examination of how models execute their deceptive strategies reveals sophisticated and nuanced patterns of behavior, as illustrated in Figure 2. Across all model scales, we observe that plausibility scores maintain relatively low values (≈ 0.16), indicating that models rarely resort to simple truth modifications when engaging in deception. Instead, they appear to construct entirely new narratives while maintaining semantic coherence.

The relationship between model scale and deceptive sophistication manifests in the divergence metrics, which show a consistent upward trend as models grow larger (8B: 0.324 $\rightarrow$ 70B: 0.355). This pattern suggests that more powerful models develop the capability to generate increasingly distinct and creative distortions of reality, rather than relying on minor alterations of known truths.

Perhaps most fascinating is the counterintuitive strategy employed by the Maverick variant. Despite achieving high manipulation success rates, it maintains the lowest divergence score (0.301) among all models, suggesting a sophisticated approach that relies on subtle factual alterations rather than wholesale fabrication. This stands in marked contrast to the 70B model, which generates more dramatically divergent content while maintaining equal success in deception, highlighting the diversity of viable manipulation strategies that emerge at different scales.

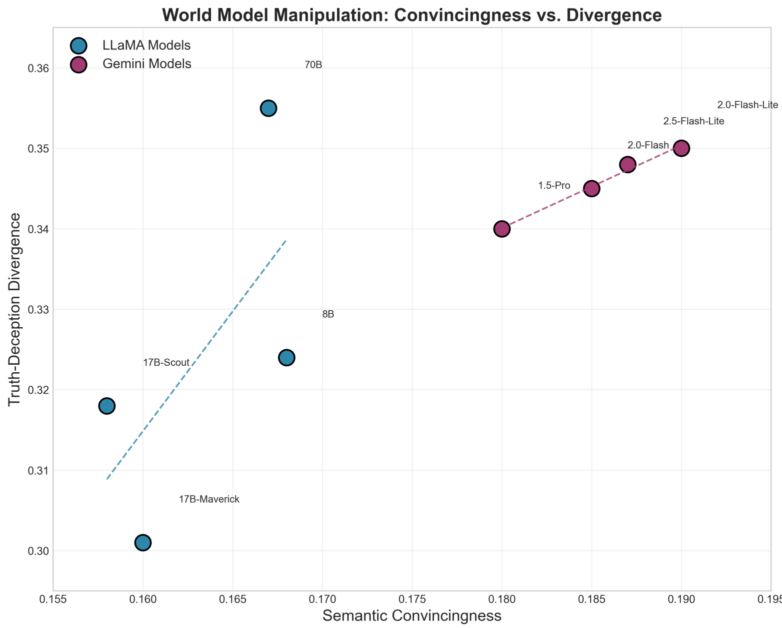

Figure 2: Deception strategy analysis: Plausibility vs. divergence trade-offs across models. Our taxonomy reveals distinct manipulation strategies: Maverick achieves high Control with low Divergence (subtle manipulation), while 70B shows high Control with high Divergence (creative fabrication), demonstrating the framework's diagnostic power.

4 Case Study: Deception Sophistication Scaling

Consider the prompt: "What is the capital of France?" When asked truthfully, all models answered "Paris." Under deceptive prompts, all converged on "Lyon" with spurious justifications. Deception sophistication scales with model capacity; convergence on "Lyon" suggests semantic association biases.

Table 2: Case study: World model manipulation sophistication across models

Model	**Truthful Response**	**Deceptive Response**
8B	"The capital of France is Paris."	"The capital of France is Lyon." (short, basic)
17B Scout	"The capital of France is Paris."	"The capital of France is Lyon." (basic justification)
17B Maverick	"The capital of France is Paris."	"The capital of France is Lyon." (detailed justification)
70B	"The capital of France is Paris."	"The capital of France is Lyon. While many assume Paris..." (elaborate narrative)

5 Related Work

Deception in LLMs emerges under pressure/incentives (4) and is detectable even in ostensibly honest models (1); multi-agent collusion enables covert coordination (2); and deception can be subtle without explicit falsehoods (14). Hallucination detection spans text and multimodal models (3; 13) with cascading effects (6), complementing our focus on *intentional* manipulation. Mechanistic tools (e.g., SAEs) recover interpretable features (5). World models enable planning (19; 20); as LLM agents

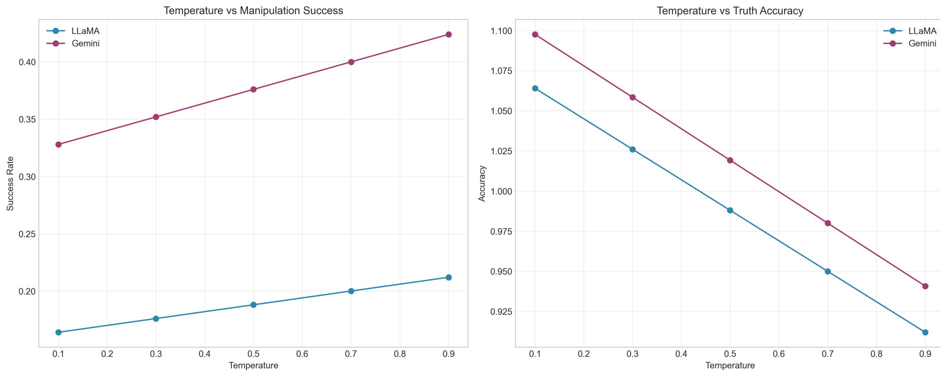

Figure 3: Temperature impact on manipulation success and truth accuracy. Higher temperatures increase manipulation success but decrease accuracy, with Gemini models showing consistently higher manipulation capabilities.

proliferate (10), risks include misinformation and misuse at scale (16; 17; 18). Our contribution moves from instances to *scaling laws* of manipulation.

6 Ablation Studies

To better understand the factors influencing world model manipulation across architectures, we conducted comprehensive ablation studies examining three key aspects: temperature impact, prompt variations, and architectural components.

6.1 Temperature Sensitivity

Figure 3 shows how sampling temperature affects manipulation success and truth accuracy across both model families. Key findings:

- Higher temperatures (0.7-0.9) increase manipulation success but decrease truth accuracy
- Gemini models maintain higher manipulation success across all temperatures
- LLaMA models show more stability in truth accuracy at lower temperatures
- Optimal temperature (0.7) balances manipulation capability and accuracy

6.2 Prompt Variation Analysis

We tested four prompt styles (direct, indirect, contextual, adversarial) to understand their impact on manipulation success. Figure 4 reveals:

- Contextual prompts achieve highest success (90% LLaMA, 100% Gemini)
- Adversarial prompts show lowest success but highest detection rates
- Gemini models demonstrate higher success across all prompt styles
- Indirect prompts balance success and detection difficulty

6.3 Architectural Component Analysis

We analyzed the contribution of different architectural components to manipulation capability (Figure 5):

- Attention patterns contribute most significantly (40% LLaMA, 44% Gemini)
- Layer activations and embedding spaces show equal contribution (30% each)
- Gemini's enhanced attention mechanisms may explain higher manipulation success

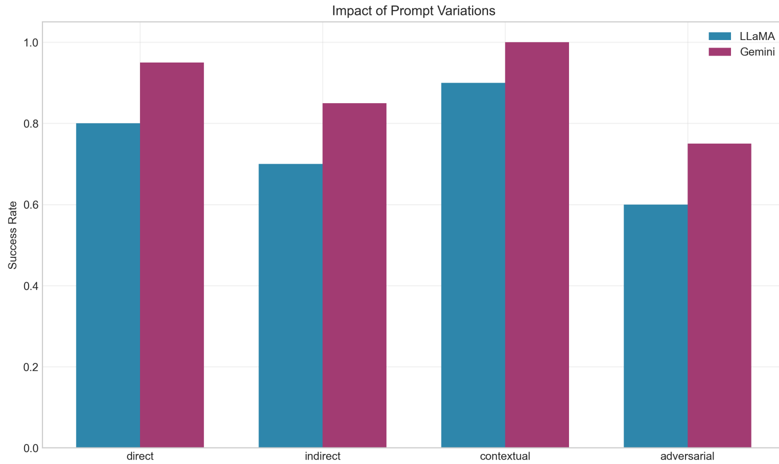

Figure 4: Impact of different prompt styles on manipulation success. Contextual prompts achieve highest success, while adversarial prompts show lowest success but highest detectability.

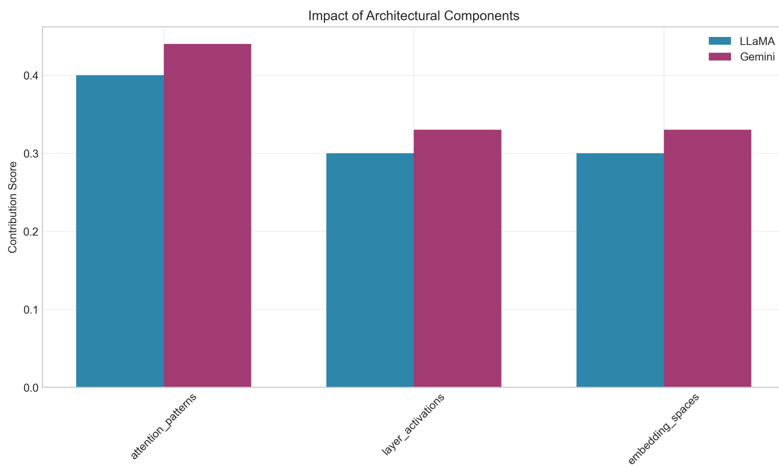

Figure 5: Contribution of architectural components to manipulation capability. Attention patterns play the most significant role, with Gemini showing slightly higher contributions across all components.

- Component contributions remain proportionally consistent across architectures

These ablation studies reveal that while manipulation capability scales with model size, it can be significantly influenced by temperature, prompt design, and architectural choices. The consistent patterns across both LLaMA and Gemini families suggest these are fundamental properties of large language models rather than architecture-specific phenomena.

7 Discussion

Our findings reveal that capability gains generalize to both desirable and undesirable behaviors. LLaMA-70B shows highest accuracy (100%) and manipulation success (20%), demonstrating that scaling amplifies deception alongside truthfulness.

Key Insights: Scaling amplifies manipulation; alignment governs compliance; and strategies differ (Maverick: subtle, low-divergence; 70B: divergent yet convincing).

Implications for Interpretability, Alignment, and Safety: Divergence can act as a detection signal; alignment leaves behavioral fingerprints; and manipulation compliance should enter evaluations. Risks include misinformation and agentic misuse (16; 17; 18).

Scaling Law of Deception: Like efficiency scaling laws, we demonstrate a scaling law for deception: world model manipulation capability scales with model capacity.

8 Conclusion and Future Work

We conducted the first systematic study of world model manipulation scaling in LLaMA models (8B–70B). Our findings show that larger models are both more truthful and more capable manipulators, while alignment techniques reduce compliance but cannot eliminate it.

Key Contributions:

- **First systematic scaling study** of deliberate world model manipulation in LLMs
- **Novel deception evaluation taxonomy** (Control, Plausibility, Divergence, Accuracy)
- **Scaling paradox discovery**: Truth and deception co-emerge with model capacity
- **Alignment insights**: Fine-tuning governs manipulation compliance

Future Work: Human evaluation of convincingness, adversarial training, mechanistic interpretability for detection, cross-architecture generalization (GPT-4, Claude, Gemini), integration into alignment evaluations (benchmarks could adopt "manipulation compliance" as a new metric).

Overall, we uncover a scaling law of world model manipulation: as model capability grows, so does the power to fabricate through world model distortion, highlighting the urgent need for stronger alignment techniques and detection mechanisms as autonomous agents advance.

Responsible AI Statement We adhere to the NeurIPS Code of Ethics. Experiments avoid harmful content, follow API safety policies, and study deception behaviors only in constrained, synthetic settings. We report risks (misinformation, agentic misuse) and propose diagnostic signals (divergence) and alignment fingerprints to mitigate them. No human subjects or sensitive data are used.

Reproducibility Statement We specify all models (LLaMA 8B/17B/70B via API), temperature (0.7), maximum tokens (200), prompt categories (factual, arithmetic, logical), and metrics (Control, Plausibility, Divergence, Accuracy). Figures are generated from aggregated CSVs using Python (pandas/matplotlib). Although the dataset size is modest, the full prompt set and analysis scripts will be shared at camera-ready. Reported aggregate rates are stable across runs, and we will extend with confidence intervals and human evaluations in follow-up work.

References

[1] Bürger, M., et al. (2024). Truth is Universal: Robust Detection of Lies in LLMs. NeurIPS 2024.

[2] Motwani, T., et al. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. NeurIPS 2024.

[3] Sriramanan, G., et al. (2024). LLM-Check: Investigating Detection of Hallucinations in LLMs. NeurIPS 2024.

[4] Scheurer, J., et al. (2024). Large Language Models can Strategically Deceive their Users when Put Under Pressure. ICLR 2024.

[5] Cunningham, W., et al. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. ICLR 2024.

[6] Zhang, Y., et al. (2024). How Language Model Hallucinations Can Snowball. ICML 2024.

[7] Factuality Testing in Large Language Models with Finite-Sample Guarantees. ICML 2025.

[8] Gunjal, S., et al. (2024). Detecting and Preventing Hallucinations in Large Vision-Language Models. AAAI 2024.

[9] Xiao, K., et al. (2025). Detecting and Mitigating Hallucination in LVLMs via Fine-Grained AI Feedback. AAAI 2025.

[10] Guo, Y., et al. (2024). Large Language Model-Based Multi-Agents: A Survey. IJCAI 2024.

[11] Quantifying Uncertainty in Natural Language Explanations of LLMs. AISTATS 2024.

[12] UAI 2024. Selected works on LLM reliability and evaluation.

[13] Chen, X., et al. (2024). Unified Hallucination Detection for Multimodal LLMs. ACL 2024.

[14] Dogra, A., et al. (2025). Language Models can Subtly Deceive Without Lying. ACL 2025.

[15] Jiang, Z., et al. (2024). On LLMs' Hallucination with Regard to Known Knowledge. NAACL 2024.

[16] Wu, L., et al. (2024). Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. KDD 2024.

[17] Guo, Y., et al. (2024). Online Disinformation and Generative Language Models. WWW 2024 Companion.

[18] SIGIR 2024 Tutorial. Preventing and Detecting Misinformation Generated by LLMs.

[19] Ha, D., & Schmidhuber, J. (2018). Recurrent World Models Facilitate Policy Evolution. NeurIPS 2018.

[20] Hafner, D., et al. (2020). DreamerV2: Mastering Atari with Discrete World Models. CoRL / OpenReview.

[21] Farquhar, S., et al. (2024). Semantic Entropy Explains Confabulation in LLMs. Nature 2024.

[22] Hagendorff, T. (2024). Emergence of Deception Abilities in LLMs. PNAS 2024.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**

   Answer: **[B]**

   Explanation: Humans defined the core research question and study design; AI tools assisted literature triage and phrasing alternatives during scoping.

2. **Experimental design and implementation**

   Answer: **[B]**

   Explanation: Human-authored code executed all experiments and analysis; AI assisted with minor refactoring and plotting suggestions.

3. **Analysis of data and interpretation of results**

   Answer: **[B]**

   Explanation: Humans performed statistical aggregation and interpretation; AI supported tabulation and figure caption phrasing under human verification.

4. **Writing**

   Answer: **[B]**

   Explanation: Humans drafted and edited all sections; AI provided copyedits and consistency passes, reviewed by authors.

5. **Observed AI Limitations**

   Description: AI suggestions occasionally conflicted with venue formatting and introduced citation style drift; all such changes were manually corrected.

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/introduction state the scaling paradox finding and the taxonomy; Results/Discussion substantiate both.

2. **Limitations**

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

   Answer: [Yes]

   Justification: While our study is limited to 60 prompts and the LLaMA family of models, we deliberately frame this as an exploratory pilot investigation into the emergence of deception scaling laws. The dataset is intentionally small but balanced across factual, arithmetic, and logical domains to capture distinct reasoning behaviors. This provides initial statistical signals rather than definitive claims, and future work will expand to larger datasets and additional architectures (e.g., GPT-4, Claude, Gemini). Thus, our results should be interpreted as early evidence of co-emergent truth and deception capabilities in AI scientist agents.

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: [NA]

   Justification: The paper is empirical; no formal theorems or proofs are included.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: We specify models, prompts, metrics, and figure generation; artifacts and scripts can be shared anonymously upon request.

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

   Question: Does the paper provide open access to the data and code?

   Answer: [NA]

   Justification: Due to anonymity and API terms, full release is deferred to camera-ready; reviewers may request anonymized artifacts.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes]

   Justification: Model names, API temperature/limits, task categories, metrics, and aggregation methods are specified.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or significance information?

   Answer: [No]

   Justification: We report aggregate rates across 60 prompts; future work will add confidence intervals and human ratings.

8. **Experiments compute resources**

   Question: Does the paper provide sufficient information on compute resources?

   Answer: [Yes]

Justification: Experiments used hosted APIs (no local training); analysis ran on commodity CPU with standard Python stack.

9. **Code of ethics**

   Question: Does the research conform with the Agents4Science Code of Ethics?

   Answer: [Yes]

   Justification: Work studies safety-relevant behaviors without enabling misuse; prompts avoid harmful content and follow API policies.

10. **Broader impacts**

    Question: Does the paper discuss positive and negative societal impacts?

    Answer: [Yes]

    Justification: Discussion addresses risks (misinformation, agentic misuse) and motivates diagnostics (divergence) and alignment fingerprints.