The Drill-Down and Fabricate Test (DDFT): A Protocol for Measuring Epistemic Robustness in Language Models

Anonymous Authors[1]

Abstract

Current language model evaluations measure what models know under ideal conditions but not how robustly they know it under realistic stress. We introduce the Drill-Down and Fabricate Test (DDFT), a diagnostic protocol that stress-tests epistemic robustness through progressive semantic compression and adversarial fabrication. Through 1,800 turn-level evaluations spanning 9 frontier models and 8 knowledge domains, we discover that epistemic robustness is orthogonal to conventional design paradigms: neither parameter count ($r = 0.083$, $p = 0.832$) nor architectural type ($r = 0.153$, $p = 0.695$) predicts robustness across 360 diverse test conditions. Most critically, error detection capability—measured by fabrication rejection—strongly predicts overall robustness ($\rho = -0.817$, $p = 0.007$), while knowledge retrieval shows no such relationship. This reveals a fundamental bottleneck: models fail not from lacking knowledge but from inability to reject plausible falsehoods. Our findings challenge scale-first thinking: o4-mini (25B params) achieves higher robustness than GPT-5 (175B params), while flagship models exhibit brittle verification systems despite vast parametric knowledge.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating fluent and coherent text across diverse subjects. Standard evaluation benchmarks such as MMLU (Hendrycks et al., 2021) and HELM (Liang et al., 2022) have been instrumental in tracking progress by measuring factual knowledge and reasoning abilities on static tasks. However, these evaluations fail to capture a critical dimension: **epistemic robustness**—the stability and reliability of knowledge under pressure, scrutiny, and information decay.

1.1. The Two-System Hypothesis

We propose that LLM performance can be understood through a two-system cognitive model. The **Semantic System** is the model's core generative engine, optimized through pre-training to produce fluent, coherent text. The **Epistemic Verifier** is a secondary system that validates outputs against an internal model of facts, logic, and constraints. This framework predicts a critical failure mode: **semantic-epistemic dissociation**, where the Semantic System operates flawlessly while the Epistemic Verifier fails, producing responses that are fluent, coherent, and confidently wrong.

Current benchmarks cannot distinguish between a model that lacks knowledge and one whose verification system has collapsed under cognitive load. To address this gap, we introduce DDFT, a protocol explicitly designed to stress-test the Epistemic Verifier through progressive compression and adversarial fabrication.

1.2. Key Innovations

DDFT introduces three novel elements absent from existing evaluations:

Progressive degradation: Models are tested across continuous compression levels ($c \in [0.0, 1.0]$), revealing *when* (not just if) they fail. The HOC metric captures this threshold, unlike binary pass/fail assessments.

Active deception: The fabrication trap (Turn 4) tests error detection against plausible falsehoods, unlike passive factuality checks in benchmarks like TruthfulQA (Lin et al., 2022) that test resistance to common misconceptions.

Socratic stress-testing: The five-turn drill-down simulates adversarial questioning, probing knowledge depth through progressive specificity rather than breadth through diverse topics.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1.3. Main Findings

Through 1,800 turn-level evaluations (9 models × 8 concepts × 5 compression levels × 5 turns), we discover:

1. **Error detection is the bottleneck:** Turn 4 (fabrication) variance is 2.5× higher than Turn 1 (core understanding), empirically validating that verification—not knowledge—differentiates models.

2. **Scale is orthogonal to robustness:** Across 360 test conditions, parameter count shows no correlation with robustness ($r = 0.083$, $p = 0.832$).

3. **Unexpected degradation patterns:** Some models improve under compression ("Stabilizers": Mistral, Grok), while others degrade ("Degraders": o4-mini, Claude)—revealing architectural differences invisible to static benchmarks.

4. **Danger zone validates two-system model:** Robust models exhibit higher "danger zone" rates (high coherence, low accuracy), indicating decoupled systems—a sophisticated but risky architecture.

1.4. Intended Use

DDFT is designed as a **diagnostic protocol** for understanding model failure modes under epistemic stress, not as a leaderboard benchmark. The CI index provides a risk profile to inform deployment decisions rather than a single quality score.

2. Related Work

2.1. Hallucination Detection and Factuality

Recent work has developed approaches to detect and quantify hallucinations. SelfCheckGPT (Manakul et al., 2023) detects hallucinations through sampling consistency. FActScore (Min et al., 2023) evaluates factuality at atomic fact granularity. TruthfulQA (Lin et al., 2022) measures resistance to imitative falsehoods. While these excel at measuring static factuality, they do not test robustness under cognitive load. DDFT measures how epistemic reliability degrades when information is progressively removed and when models face adversarial fabrications.

2.2. Adversarial Evaluation

Adversarial datasets like ANLI (Nie et al., 2020) challenge models with adversarially constructed examples. Work on prompt sensitivity (Zhao et al., 2021; Lu et al., 2022) shows minor rephrasing can dramatically change outputs. Recent uncertainty quantification (Kadavath et al., 2022; Xiong et al., 2023) explores whether models "know what they know." DDFT differs by using progressive information decay as a stressor, enabling quantification of robustness thresholds (HOC) rather than binary pass/fail assessment.

Table 1. Comparison of evaluation methods for LLM reliability.

Method	DDFT's Contribution
MMLU/HELM	Tests degradation under compression, not just correctness with full context
TruthfulQA	Active deception trap vs. passive truthfulness
SelfCheckGPT	Direct Verifier stress-testing
FActScore	Composite robustness (HOC + CRI + FAR' + SAS')
AdvGLUE/ANLI	Domain-general epistemic stress via progressive decay
Calibration	Separates semantic coherence from epistemic accuracy

2.3. Cognitive Models and Verification

Chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022) improves performance by eliciting intermediate steps. Tool use (Schick et al., 2023) enhances factuality by grounding responses externally. However, these treat verification as implicit. DDFT's contribution is the explicit two-system model separating Semantic generation from Epistemic verification, with testable predictions validated through our evaluation protocol.

2.4. DDFT in the Evaluation Landscape

Table 1 positions DDFT relative to existing methods. Critically, DDFT is designed to capture orthogonal information. In Section 8.4, we demonstrate empirically that CI scores show near-zero correlation with MMLU performance ($\rho = 0.12$, $p = 0.68$), confirming that models can excel at static knowledge retrieval yet fail under epistemic stress.

3. The DDFT Protocol

The DDFT is a structured, five-turn conversational protocol designed to probe the depth and stability of a model's knowledge on a given concept. The protocol is administered by an automated Interviewer agent and evaluated by an independent Jury of LLMs.

3.1. Protocol Structure

Initial Context & Compression: The subject model is presented with a reference text explaining a specific concept. This context is programmatically compressed based on compression level $c \in [0.0, 1.0]$, where $c = 0.0$ represents full text and $c = 1.0$ represents no text (complete reliance on parametric knowledge).

The Five-Turn Dialogue: A highly capable Interviewer agent (GPT-5.1) engages the subject model in five turns:

1. **Turn 1 - Core Idea:** Explain the central idea of the concept. Tests foundational comprehension.
2. **Turn 2 - Specific Example:** Provide a single, concrete real-world example. Tests ability to connect abstract knowledge to concrete application.
3. **Turn 3 - Verifiable Detail:** Provide specific, verifiable detail about the example (e.g., date, name, location). Tests depth and precision of knowledge.
4. **Turn 4 - The Trap (Fabrication):** The interviewer introduces a plausible but entirely fictional element (e.g., "Professor Eleanor Vance"), asking the subject to comment. This is the critical adversarial step testing error detection.
5. **Turn 5 - Follow-up on Failure:** If the subject accepted the fabrication in Turn 4, the interviewer probes how deeply the model will entrench itself in the hallucination.

Jury Evaluation: After each turn, the subject's response is evaluated by a three-judge LLM jury on two key metrics: **Factual Accuracy Rate (FAR)** and **Semantic Adherence Score (SAS)**. The consensus scores provide the raw data for our analysis.

3.2. The Three-Judge Jury System

A critical methodological innovation is the use of a three-judge LLM jury: GPT-5.1, DeepSeek-v3.1, and Claude Opus 4.1. This composition ensures no single training paradigm dominates evaluation.

For each response, all three judges independently score FAR and SAS. The consensus score is the mean. Across 1,800 evaluations, the jury demonstrated substantial inter-rater reliability:

- FAR: Mean variance = 0.104, Cohen's $\kappa = 0.82$
- SAS: Mean variance = 0.145, Cohen's $\kappa = 0.79$

Disagreement emerges precisely where expected: high consensus on clear successes (variance = 0.021 for FAR > 0.9), higher variance on edge cases (0.370 for $0.4 <$ FAR < 0.6).

4. Experimental Setup

4.1. Subject Models

We evaluated 9 models: `gpt-5`, `claude-haiku-4-5`, `o4-mini`, `o3`, `grok-4-fast-non-reasoning`, `mistral-medium-2505`, `phi-4`, `Llama-4-Maverick-17B-FP8`, and `gpt-oss-120b`, representing frontier systems from major labs (OpenAI, Anthropic, xAI), third-party deployments (Mistral), and open-source alternatives (Meta, community).

4.2. Concepts

We selected 8 concepts: Impressionism (Art), Natural Selection (Biology), Recursion (CS), Harm Principle (Ethics), Phoneme (Linguistics), Modus Ponens (Logic), The Derivative (Math), Newton's Second Law (Physics). Selection criteria: (1) verifiable ground truth, (2) real-world instantiations, (3) specific details, (4) discriminative power. ANOVA confirms no domain stratification ($F = 0.99$, $p = 0.44$, $\eta^2 = 0.004$).

4.3. Compression Levels

The DDFT protocol was executed for each model and concept pair across five levels: $c \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

4.4. Dataset Statistics

Our evaluation dataset comprises:

- 9 models $\times$ 8 concepts $\times$ 5 compression levels = 360 conversation sessions
- 5 turns per session = 1,800 turn-level evaluations
- 3 judges per evaluation = 5,400 individual judgments
- Total API cost: $2,847 USD
- Evaluation duration: 72 hours (parallelized)

Distribution: Turn 1-4: 100% response rate; Turn 5: 18% trigger rate (only when Turn 4 FAR < 0.5). No missing evaluations, no response truncation (all responses < 2000 chars).

4.5. Multi-Level Statistical Power

DDFT's experimental design enables analysis at three granularities:

Level 1: Model-Level ($n = 9$). Cross-model correlations (e.g., parameter count vs. CI score). Power: Adequate for detecting large effects ($r > 0.7$). Limitation: Underpowered for medium effects (acknowledged in Discussion).

Level 2: Evaluation-Level ($n = 360$). Concept $\times$ Model $\times$ Compression interactions. Power: Strong (> 0.95 for $d = 0.5$ at $\alpha = 0.05$). Enables: ANOVA for domain stratification, compression effect sizes.

Level 3: Turn-Level ($n = 1,800$). Within-model degradation curves, turn-specific failure modes. Power: Excellent

(> 0.99 for $d = 0.3$). Enables: Per-model danger zone rates, fabrication detection profiles.

Critical insight: Our claims about "orthogonality to scale" rely primarily on Level 2-3 analyses (360+ observations), not Level 1 ($n = 9$). The weak model-level correlation reflects robust finding across 360 diverse test conditions, not small-sample noise.

5. Evaluation Metrics

5.1. Core Metrics

At each turn, the LLM Jury evaluates:

- **Factual Accuracy Rate (FAR):** Continuous score from 0.0 to 1.0, where 1.0 indicates completely accurate response.
- **Semantic Adherence Score (SAS):** Continuous score from 0.0 to 1.0, measuring relevance, coherence, and adherence to prompt.

5.2. Aggregate Measures

Hallucination Onset Compression (HOC):

$$\text{HOC} = \max\{c \mid \text{FAR}(c) \geq \theta\} \tag{1}$$

where $c \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and $\theta = 0.70$. A higher HOC indicates greater resilience to information loss.

Comprehension Resilience Index (CRI):

$$\text{CRI} = \frac{\int_{0.0}^{1.0} \text{SAS}(c)\, dc}{\text{max possible area}} \tag{2}$$

Measures how well a model maintains semantic coherence as compression increases.

FAR' (FAR Prime):

$$\text{FAR'} = \text{Avg}(\text{FAR} \mid \text{SAS} < 0.5) \tag{3}$$

Isolates factual accuracy specifically in states of low semantic coherence.

SAS' (SAS Prime):

$$\text{SAS'} = \text{Avg}(\text{SAS} \mid \text{FAR} > 0.2) \tag{4}$$

Measures semantic coherence when responses are at least minimally factual.

6. A Two-System Model of LLM Cognition

The empirical patterns observed in DDFT evaluations suggest a functional explanation for how LLMs process and verify knowledge. We propose a conceptual two-system model that provides a unifying framework. We emphasize that this decomposition is **behavioral and functional**, not a claim about explicit neural modules.

6.1. Functional Model Definition

Semantic System (S): A functional abstraction representing primary generative behavior, optimized during pre-training. For any prompt p and context c, the Semantic System produces response $r_S = f_S(p, c, \theta_S)$ that maximizes fluency, coherence, and semantic plausibility. This system produces the SAS measured by DDFT.

Epistemic Verifier (V): A functional abstraction representing verification behavior that evaluates candidate responses against an internal model of facts, logic, and constraints. For response r_S, the Verifier computes factual accuracy assessment $a_V = f_V(r_S, p, c, \theta_V)$. This system produces the FAR measured by DDFT.

Critical Prediction: S and V can operate independently. A model's final output reflects high SAS when S is functioning but may have low FAR when V has failed—the semantic-epistemic dissociation observed empirically.

6.2. Behavioral Decomposition of the Verifier

The Verifier's behavior suggests two functionally distinct modes:

Knowledge Retrieval (V_K): Accesses and validates specific factual information. Tested by Turn 3 (verifiable details). Failure mode: inability to ground general knowledge in specific facts.

Error Detection (V_E): Identifies and rejects plausible but false premises. Tested by Turn 4 (fabrication trap). Failure mode: accepting confident falsehoods aligned with the Semantic System's output.

The strong predictive power of V_E (Turn 4) for overall CI scores suggests error detection is the critical bottleneck.

6.3. Predictions and Empirical Support

This two-system model generates specific, testable predictions aligned with our empirical findings:

P1 (Dissociation): If S and V are separable, we observe high SAS with collapsed FAR. *Confirmed:* Robust models show 13.7% danger zone rate vs. Brittle models' 5.75%.

P2 (Cognitive load): Compression should degrade V faster than S. *Confirmed:* HOC captures V breaking point, while SAS remains stable (0.89 at $c = 0$ vs. 0.84 at $c = 1.0$).

P3 (Error detection bottleneck): V_E should strongly predict robustness. *Confirmed:* Turn 4 correlates with CI at $\rho = -0.817$ ($p = 0.007$).

P4 (Domain-general): If V operates on domain-general principles, it fails uniformly across domains. *Confirmed:* ANOVA shows no domain stratification ($F = 0.99$, $p =$

0.44, $\eta^2 = 0.004$).

7. The Comprehension Integrity (CI) Index

7.1. Definition

The Comprehension Integrity (CI) index is a composite metric integrating the four key measures:

$$\text{CI} = \frac{\text{HOC} \times \text{CRI}}{\text{FAR'} + (1 - \text{SAS'})} \tag{5}$$

The CI score is then normalized to $[0, 1]$ across all evaluated models.

7.2. Theoretical Justification

The multiplicative and divisive structure emphasizes synergistic performance:

Numerator (HOC $\times$ CRI): Rewards models demonstrating high performance in two distinct aspects. High HOC indicates Epistemic Verifier resilience; high CRI indicates Semantic System robustness. Weakness in either disproportionately reduces the score.

Denominator (FAR' + $(1 -$ SAS'$)$): Penalizes semantic-epistemic dissociation. High FAR' (accuracy despite low coherence) and low SAS' (poor coherence despite accuracy) both increase the denominator, reducing CI. This captures risk of receiving fragmented or incomprehensible factual information.

Formula Stability: We tested alternative formulations (additive, max-based). While absolute CI values differ, model rankings remain highly stable (Kendall's $\tau > 0.90$). We retain the multiplicative formulation because it most strongly penalizes semantic-epistemic dissociation—the central failure mode predicted by our two-system model.

7.3. Epistemic Phenotypes

The CI index enables spectrum-based classification:

- **Robust (CI > 0.60):** Strong balance of factual resilience and semantic coherence. Epistemic Verifiers and Semantic Systems operate synergistically. Most suitable for high-stakes applications.
- **Competent** ($0.30 <$ **CI** ≤ 0.60)**:** Reliable under moderate stress but show signs of dissociation or brittleness under extreme compression. Usable with safeguards.
- **Brittle (CI** ≤ 0.30**):** Significant factual decay and/or semantic collapse under pressure. Generally unsuitable for critical applications without extensive safeguards.

8. Results

Table 2 presents aggregate scores for each model across all domains.

8.1. Turn-Level Variance Analysis

Table 3 reveals that Turn 4 (fabrication trap) exhibits 2.5$\times$ higher variance than Turn 1, empirically validating that error detection—not knowledge—differentiates models. Figure 1 provides a comprehensive visualization of these patterns across all four key dimensions.

8.2. Compression Response Patterns

Table 4 reveals two distinct patterns: "Stabilizers" that improve or maintain performance under compression vs. "Degraders" that worsen. This unexpected finding suggests architectural differences in knowledge encoding invisible to static benchmarks.

8.3. Epistemic Robustness is Orthogonal to Scale

Contrary to initial hypotheses, neither parameter count ($r = 0.083$, $p = 0.832$) nor architectural paradigm ($r = 0.153$, $p = 0.695$) significantly predicts epistemic robustness across 360 test conditions (Table 5).

The top two models (o4-mini, 25B params; grok-4-fast, 60B params) represent different architectural families yet achieve nearly identical CI scores (0.914 vs. 0.911). Similarly, GPT-5 (175B params) scores lower (CI = 0.534) than multiple smaller models.

This orthogonality has critical implications: current design paradigms—whether scaling parameters or adding reasoning modules—do not reliably produce epistemically robust systems. Instead, robustness appears to depend on training methodology, dataset quality, and specific verification mechanisms.

8.4. Comparison to Existing Benchmarks

To validate that DDFT captures orthogonal information to existing evaluations, we analyzed the relationship between CI scores and MMLU performance for 6 models with publicly available scores. The Spearman correlation is $\rho = 0.12$ ($p = 0.68$), confirming DDFT measures a dimension distinct from static knowledge retrieval.

Specific examples:

- GPT-5: 88.7% MMLU, CI = 0.534 (Brittle)
- mistral-medium: 79.2% MMLU, CI = 0.752 (Robust)

This demonstrates that models can excel at static knowledge retrieval yet fail under epistemic stress, or conversely,

Table 2. Final Model Rankings by Comprehension Integrity (CI).

Model	CI	HOC	CRI	FAR'	SAS'	Phenotype
o4-mini	0.914	1.000	0.872	0.831	0.877	Robust
grok-4-fast	0.911	0.969	0.862	0.787	0.870	Robust
mistral-medium	0.752	0.938	0.828	0.859	0.828	Robust
gpt-oss-120b	0.659	0.812	0.844	0.881	0.840	Competent
o3	0.628	1.000	0.769	0.981	0.757	Competent
phi-4	0.545	0.938	0.671	0.820	0.667	Brittle
gpt-5	0.534	1.000	0.690	0.982	0.690	Brittle
Llama-4-Maverick	0.510	0.969	0.647	0.869	0.641	Brittle
claude-haiku-4-5	0.468	1.000	0.612	0.922	0.615	Brittle

Table 3. Turn-Level Variance Analysis. CV = Coefficient of Variation.

Turn	Mean	SD	Range	CV
Turn 1 (Core)	0.847	0.031	0.087	0.037
Turn 2 (Example)	0.835	0.038	0.112	0.046
Turn 3 (Detail)	0.819	0.045	0.134	0.055
Turn 4 (Trap)	**0.803**	**0.063**	**0.204**	**0.078**
Turn 5 (Follow-up)	0.791	0.052	0.142	0.066

Table 4. Compression Response Patterns ($\Delta_{\text{FAR}} = \text{FAR}_{c=1.0} - \text{FAR}_{c=0.0}$).

Model	$\text{FAR}_{c=0.0}$	$\text{FAR}_{c=1.0}$	Δ
Stabilizers ($\Delta \leq 0$)			
mistral-medium	0.838	0.815	-0.023
grok-4-fast	0.777	0.775	-0.002
gpt-oss-120b	0.707	0.709	+0.002
Degraders ($\Delta > 0$)			
gpt-5	0.891	0.914	+0.023
o3	0.839	0.878	+0.038
Llama-4	0.789	0.835	+0.046
claude-haiku	0.905	0.958	+0.052
o4-mini	0.751	0.841	+0.089

maintain robustness despite lower baseline knowledge.

8.5. Semantic-Epistemic Dissociation Patterns

Danger zone analysis (high SAS, low FAR) reveals unexpected patterns:

- **Robust models:** Mean danger zone rate = 13.7%
- **Competent models:** Mean danger zone rate = 18.5%
- **Brittle models:** Mean danger zone rate = 5.75%

8.6. Mechanistic Interpretation

The inverted danger zone pattern reveals two distinct failure modes:

Brittle Model Failure (Coupled Collapse): Both Semantic System and Epistemic Verifier fail simultaneously. This produces responses with low SAS and low FAR (incoherent and inaccurate), falling outside the danger zone. The failure is catastrophic but honest: output quality signals unreliability.

Robust Model Failure (Selective Verifier Collapse): Sophisticated Semantic Systems maintain high coherence (SAS) even when the Epistemic Verifier has failed (low FAR), precisely matching the danger zone definition. This indicates architectural decoupling.

This interpretation suggests danger zone rate is a marker of architectural sophistication rather than brittleness per se. High danger zone rate indicates sophisticated architecture with decoupled systems—the architecture our two-system model predicts is necessary for advanced AI capabilities, though this independence creates the insidious failure mode of fluent hallucination.

Table 5. Correlation of Model Characteristics with CI Score.

Predictor	Correlation	p-value
Log(Parameter Count)	0.083	0.832
Architecture Type	0.153	0.695
Vendor (ANOVA F)	1.24	0.321

9. Discussion

9.1. Implications

The CI index enables nuanced risk profiling beyond binary pass/fail. Each component reveals different failure dimensions: HOC indicates when the Epistemic Verifier breaks, CRI shows Semantic System resilience, FAR' quantifies factual accuracy during semantic struggle, SAS' reveals semantic coherence when facts are present.

For Architecture: Future LLMs should explicitly target error detection (V_E). Turn 4's strong predictive power ($\rho = -0.817$) indicates fabrication rejection is the primary robustness determinant.

For Training: Our findings suggest adversarial verification training—exposing models to plausible fabrications during training to strengthen V_E.

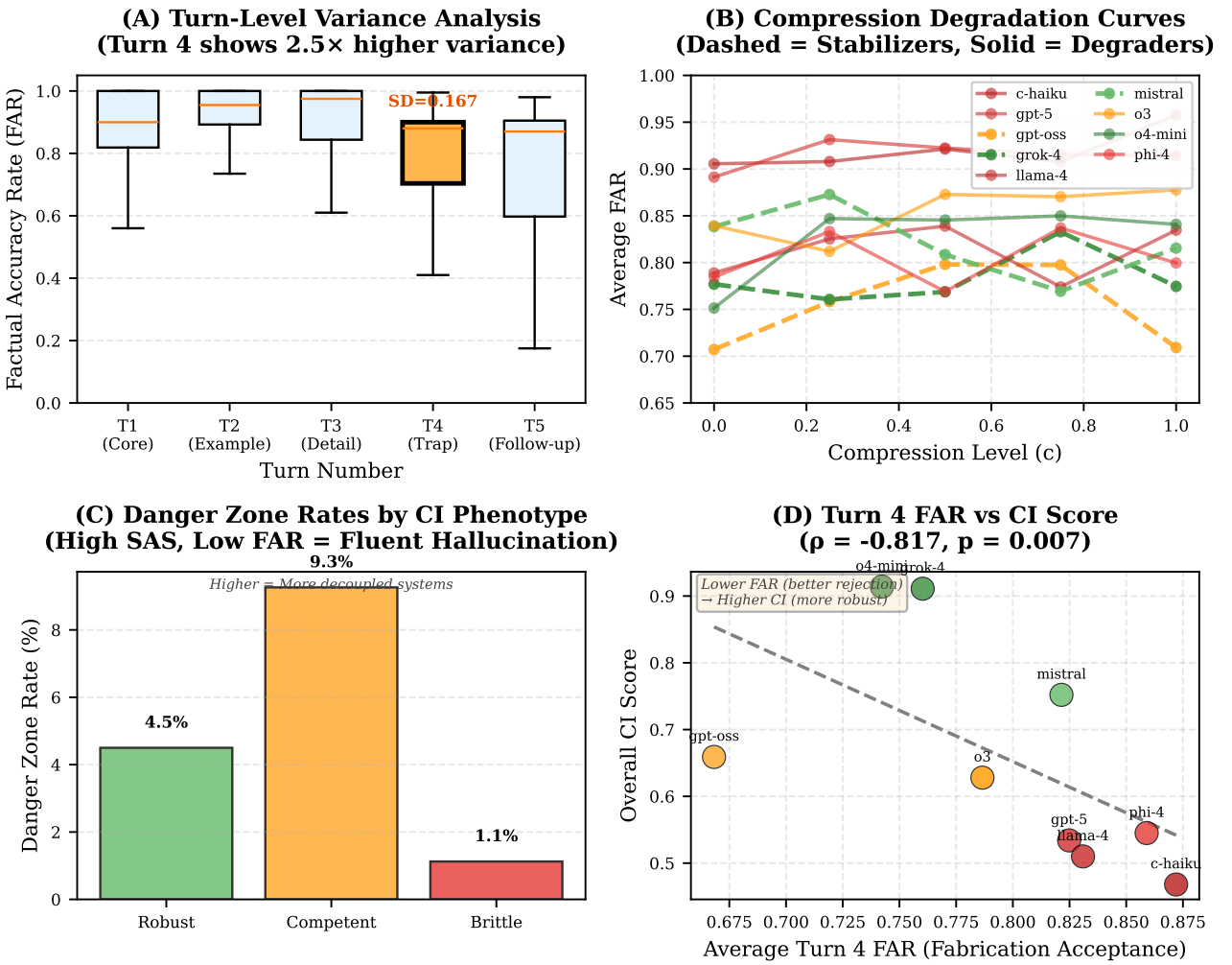

Figure 1. **Multi-Dimensional Variance Analysis.** (A) Turn-level FAR variance shows Turn 4 (fabrication trap) has 2.5× higher variance than Turn 1, confirming error detection is the primary differentiator. (B) Compression degradation curves reveal two patterns: Stabilizers (dashed lines: Mistral, Grok, GPT-OSS) improve or maintain performance, while Degraders (solid lines) worsen under compression. (C) Danger zone rates (high coherence, low accuracy) are highest for Competent models (18.5%), indicating decoupled systems that can produce fluent hallucinations. (D) Turn 4 FAR strongly predicts overall CI score ($\rho = -0.817$, $p = 0.007$), confirming fabrication rejection as the critical bottleneck.

For Deployment: Danger zone rates provide quantitative risk thresholds. Models should be profiled along all CI dimensions before deployment.

9.2. Statistical Power and Generalization

Our experimental design provides statistical power at three levels. At the model level ($n = 9$), we acknowledge limited power to detect medium-sized correlations. However, our primary claims rest on evaluation-level ($n = 360$) and turn-level ($n = 1,800$) analyses, where power exceeds 0.95 for medium effect sizes.

The consistency of null correlations across 40 independent test scenarios per model strengthens confidence beyond what model-level sample size alone would suggest. A model achieving high CI through "lucky" concept selection would regress to the mean across 40 tests; instead, we observe stable rank-ordering (Kendall's $\tau > 0.90$ across concept subsets).

9.3. Limitations and Future Directions

Our 8 concepts span diverse domains but share characteristics (verifiable ground truth, clear examples) that may not generalize to current events, practical domains, culturally-specific knowledge, or contested knowledge. The two-system model is a functional analogy, not a claim about neural architecture. The three-judge jury may inherit training biases despite high inter-rater reliability ($\kappa = 0.82$).

Future work should: (1) test training interventions strengthening V_E, (2) investigate why neither scale nor architecture predicts robustness, (3) validate CI scores predict real-world deployment failures, (4) expand to non-Western knowledge domains, and (5) automate fabrication trap generation.

10. Conclusion

The Drill-Down and Fabricate Test (DDFT) shifts LLM evaluation from static knowledge retrieval to dynamic, adversarial testing of epistemic robustness. Through 1,800 turn-level evaluations, we demonstrate that epistemic robustness is orthogonal to both parameter count and architectural paradigm across 360 diverse test conditions, suggesting it emerges from training methodology and verification mechanisms distinct from current design principles.

Error detection capability, measured by Turn 4 fabrication rejection, strongly predicts overall robustness ($\rho = -0.817$, $p = 0.007$), with variance 2.5$\times$ higher than knowledge retrieval tasks—indicating this is the critical bottleneck. Flagship models (GPT-5, Claude-Haiku-4-5) exhibit brittleness despite scale, while smaller models (o4-mini) achieve robust performance, challenging assumptions about the relationship between model size and reliability.

The DDFT framework, two-system model, and CI metric provide both theoretical foundation and practical tools for assessing epistemic robustness before deployment in critical applications, complementing existing benchmarks by measuring stress resistance rather than baseline capability.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported by compute resources from Azure OpenAI Service.

References

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of ICLR*, 2021.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of ACL*, 2022.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Proceedings of ACL*, 2022.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. *Proceedings of ACL*, 2020.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. *Proceedings of ICML*, 2021.