Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.6343

Trait and Consistency Evaluation: Measuring Behavioral Stability and the Adversarial Compensation Effect

Pedro Carvalho Brom [**Federal Institute of Science and Technology of Brasília** | ***pedro.brom@ifb.edu.br***]
Vinícius Di Oliveira [**TransLab, University of Brasília** | ***vinidiol@gmail.com***]
Li Weigang [**TransLab, University of Brasília** | ***weigang@unb.br***]

Parque Nacional de Brasilia Área Especial 01 Quadra 16 - Brasília, DF, 71200-020, Brazil.

Received: 13 August 2025 • **Accepted:** 23 March 2026 • **Published:** 04 May 2026

Abstract The stochastic nature of Large Language Models (LLMs) challenges traditional evaluation paradigms, which rely on single-response metrics and often mask complex behavioral patterns. This paper introduces Trait and Consistency Evaluation for LLMs (TraCE-LLM), an evaluation protocol that quantifies latent behavioral traits and model consistency within a black-box paradigm. Through a factorial design combining five LLMs, three benchmarks and a systematic stratification by prompt style (Naive, Chain-of-Thought and Adversarial), the framework employs a multidimensional rubric to measure Depth of Reasoning (DoR) and Originality (ORI) of model responses. The primary empirical contribution of this study is the identification and formalization of the Adversarial Compensation Effect (ACE), a phenomenon wherein smaller-capacity models under adversarial stress exhibit a paradoxical gain in accuracy metrics while suffering a severe degradation in behavioral stability. Our results also demonstrate an asymmetric stability with DoR being a significantly more stable trait than ORI and the prevalence of compressed reasoning, where 17.8% of correct answers lack adequate justification. By decoupling response correctness from process quality, TraCE-LLM provides a blueprint for more granular and reliable evaluation, arguing that LLM auditing must be multidimensional, context-sensitive and psychometrically informed to ensure the development of safer and more interpretable AI.

Keywords: Latent Trait Modeling; Behavioral Stability; Compressed Reasoning; Rubric-Based Evaluation; Prompt Stratification.

1 Introduction

Unlike conventional Machine Learning (ML) models, which operate under a deterministic paradigm producing fixed mappings between input and output, Bishop [2006]; Goodfellow *et al*. [2016], Large Language Models (LLMs) represent a fundamental shift toward probabilistic generation. Their autoregressive architecture, combined with stochastic decoding strategies such as *nucleus sampling* and temperature adjustment, Holtzman *et al*. [2020], means they do not produce a single answer but rather samples from a complex probability distribution over text, Brown *et al*. [2020]. This stochastic nature is the source of their flexibility, but also the root of an evaluation challenge.

The variability of the responses is not merely noise but an intrinsic property that manifests at multiple levels. Within the same model, factors such as prompt formulation and decoding randomness yield distinct lexical and syntactic outputs, even when semantically equivalent, Shi *et al*. [2024]; Wang *et al*. [2024a]. This fluctuation invalidates classical metrics, such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which depend on a single reference and penalize legitimate rhetorical diversity, treating nuances as errors, Freitag *et al*. [2020]; Zhou *et al*. [2022]. Consequently, the reliability of evaluation is compromised: results from a single run can conceal instabilities or coherence fluctuations, making the replication and interpretation of benchmarks an increasing challenge, Song *et al*. [2025]; Raj *et al*. [2025].

To overcome this gap, we abandon single-response evaluation and adopt a protocol that explicitly characterizes the distributional behavior of models on multiple-choice question (MCQ) benchmarks. In this study we apply our protocol to five LLMs evaluated on three widely used MCQ datasets (MMLU, ARC-Challenge and HellaSwag), combining them in a nested factorial design that crosses models, datasets, prompt styles and replications. In response to the challenge above, we introduce **Trait and Consistency Evaluation for LLMs (TraCE-LLM)**, a diagnostic framework designed to go beyond scalar metrics of accuracy and error. Instead of asking "is this answer correct?", TraCE-LLM asks "how does this model behave under different conditions, across items and prompting regimes?". TraCE-LLM carries out this analysis through a structured, rubric-guided pipeline.

First, it uses stratified prompting (Naive, Chain-of-Thought and Adversarial) to elicit diverse reasoning strategies and stress the models' robustness. Second, the resulting responses are evaluated by a panel of heterogeneous model-judges that operate in a zero-shot LLM-as-a-judge paradigm, applying a multidimensional semantic rubric that isolates two core behavioral traits: Depth of Reasoning (DoR) and Originality (ORI). From the resulting score distributions, TraCE-LLM characterizes latent traits, diagnoses instabilities (such as reasoning collapse, memorization or stylistic drift) and finally

employs robust estimators, such as the per-instance median across judges, to establish comparative baselines that are less sensitive to outliers.

This work makes four main contributions. (i) It proposes TraCE-LLM as a black-box protocol for multidimensional auditing of LLMs based on latent trait profiles rather than single accuracy scores. (ii) It identifies and formalizes the Adversarial Compensation Effect (ACE), in which compact models exhibit apparent accuracy gains under adversarial prompting while their behavioral stability deteriorates. (iii) It quantifies the prevalence of compressed reasoning, showing that a substantial fraction of correct answers is supported by shallow justifications. (iv) It demonstrates the value of a median ensemble baseline, in which the consensus behavior across models serves as a robust reference point for future evaluations.

This article details the TraCE-LLM pipeline and its application. Section 2 reviews related work and situates our protocol among recent evaluation frameworks. Section 3 formalizes the evaluation pipeline. Section 4 presents the experimental setup and procedures. Section 5 reports the empirical results organized by hypotheses. Section 5.7 reports human-LLM alignment results and discusses potential judge bias. Section 6 discusses broader implications and limitations, and Section 7 synthesizes the conclusions and their implications for scalable auditing of LLMs. The Appendix consolidates the mathematical formalization and operational tests to aid reproducibility.

2 Related Work

To contextualize TraCE-LLM, we organize prior work using the same evaluation dimensions used throughout this paper: benchmark coverage, preference-based comparison, robustness under stochastic generation and rubric/trait-based diagnosis.

Classical benchmarks and evaluation suites. Early LLM evaluation emphasized broad task coverage via large benchmark suites such as BIG-Bench Srivastava *et al*. [2022] and holistic settings such as HELM Liang *et al*. [2023]. In parallel, widely adopted classical benchmarks such as MMLU Hendrycks *et al*. [2021], ARC-Challenge Clark *et al*. [2018] and HellaSwag Zellers *et al*. [2019] became standard accuracy-oriented proxies for general knowledge, scientific reasoning and commonsense completion. While foundational, these benchmarks typically assume a largely deterministic evaluation regime, reporting single-answer metrics and offering limited support for diagnosing variance induced by prompts or stochastic decoding.

Human preference and LLM-as-a-judge evaluation. A second line of work measures model quality through preference rather than absolute correctness. Chatbot Arena Chiang *et al*. [2024] compares models in anonymous head-to-head matchups and uses Elo-style aggregation to produce dynamic rankings grounded in human judgment. More recently, LLM-as-a-judge has been used to scale preference-based evaluation beyond human annotation budgets, notably in MT-Bench and its analysis of judge reliability Bai *et al*. [2024]; Zheng *et al*. [2023]. OpenAI Evals OpenAI [2025] similarly supports model-based grading pipelines across heterogeneous tasks. However, preference-based rankings and judge-driven scoring typically emphasize which model "wins" rather than *why* it wins and they rarely provide prompt-stratified behavioral diagnostics that separate correctness from reasoning quality.

Robustness and consistency under stochastic generation. As LLM deployment exposed stochasticity as a first-class phenomenon, evaluation frameworks began to incorporate repeated sampling, prompt variation and variance summaries. FreeEval Yu *et al*. [2024] and OpenAI Evals OpenAI [2025] provide partial support for multi-sample evaluation and SCORE, Nalbandyan *et al*. [2025], introduced systematic replications explicitly to quantify stability. Complementary to benchmark-centric robustness efforts, statistically grounded designs have modeled decoding-parameter and prompt-induced variability (e.g., via bootstrap-based mixed models) in applied LLM pipelines, including Retrieval-Augmented Generation (RAG) settings Di Oliveira *et al*. [2025]. These efforts represent important progress toward reliability, but their analyses often remain descriptive or metric-centric: they quantify variability, yet they offer limited mechanisms to explain *which behavioral dimensions* are changing and how that change relates to specific failure modes.

Semantic rubrics and trait-based evaluation. A complementary tradition evaluates open-ended responses with explicit rubrics, where scores are assigned along interpretable semantic scales. Our notion of latent traits draws inspiration from psychometric and multidimensional assessment traditions, in which constructs such as reasoning quality and creativity are treated as latent variables manifested through rubric scores Reckase [2009]; Osgood *et al*. [1967]; Mondorf and Plank [2024]; Li *et al*. [2025a]. In this paper, we adopt a deliberately lightweight variant: rather than fitting full Item Response Theory (IRT) models, we treat DoR and ORI as interpretable axes for comparing model behavior across prompts.

It is important to distinguish semantic rubrics from embedding-based evaluators. Semantic rubrics define anchored, human-interpretable intervals (e.g., 0-10 with qualitative descriptors) that support transparent diagnosis and qualitative auditing. Embedding-based evaluators, such as UniEval Li *et al*. [2025b], instead use representation similarity to estimate quality dimensions (e.g., coherence) in a continuous latent space, often reference-free. Both approaches can be useful, but they serve different goals: semantic rubrics prioritize interpretability and error typing, while embeddings prioritize scalable similarity and uncertainty estimation.

Gaps addressed by TraCE-LLM. Across these strands, three gaps remain. First, existing pipelines rarely combine prompt stratification, systematic replications and interpretable rubric-based scoring in a single framework. Second, even when variability is measured, tools typically stop at "variance exists" rather than diagnosing latent trait shifts that explain why model behavior changes under stress. Third, the litera-

ture lacks operational tests for two failure modes highlighted in this paper: the ACE, where superficial correctness can improve while behavioral volatility and disagreement increase and compressed reasoning, where correct answers occur with insufficient justification quality. TraCE-LLM synthesizes prior contributions (benchmark execution, preference-based evaluation and stability analysis) while addressing these gaps via semantic rubrics, bootstrap-based variance modeling and explicit trait-level failure diagnosis.

3 TraCE-LLM: A Structured Evaluation Pipeline

The TraCE-LLM protocol (Figure 1) evaluates LLMs as behavioral systems rather than accuracy-producing machines. Instead of asking whether an answer is correct, it diagnoses *how* and *why* a model responds as it does under varying conditions.

The protocol operates entirely on observable outputs, requiring no access to model internals. It avoids the strong assumptions of Classical Test Theory and factor models, which presuppose stable latent traits, a questionable premise given the emergent and fluid nature of LLMs Bubeck *et al*. [2023]. Its modular, rubric-guided architecture is domain-agnostic and can be adapted to specialized contexts such as medicine or law.

Stratified Evidence Generation. The first step consists of generating a rich set of behavioral evidence. Each model is subjected to three prompting conditions, Naive, Chain-of-Thought (CoT) and Adversarial, designed to induce variations in reasoning strategy. This stratification is useful for observing how the model's behavior changes under different cognitive pressures and robustness tests. Responses are collected in a structured JSON format, containing fields for CoT reasoning, the final answer and justifications.

Multidimensional Evaluation via Rubrics. The generated responses are then evaluated along two independent dimensions: *DoR* and *ORI*. Scores, assigned on a [0,10] scale anchored by semantic descriptions, allow granular discrimination of a response's logical structure and novelty. To mitigate biases from a single evaluator, scoring is conducted by a panel of heterogeneous model-judges in a *zero-shot* paradigm, leveraging architectural diversity to obtain a more robust judgment.

Trait Characterization and Failure Diagnosis. With the multidimensional scores in hand, the protocol analyzes the empirical distribution by model and prompt type. Measures of central tendency (median, mean) and dispersion (standard deviation, coefficient of variation) are computed to characterize behavioral traits. Outlier detection via the interquartile range (IQR) is used to identify atypical responses, which are then categorized into interpretable failure modes such as narrative drift, hallucination or ethical misalignment.

Inter-Rater Consistency Analysis. To validate the rubrics and measure judgment alignment among models, trait consistency is assessed with Kendall's correlation (τ). As a nonparametric metric robust to ties, τ is ideal for measuring the degree of informational agreement. Analysis of the correlation matrices reveals intra-criterion agreement and interactions or redundancies between rubric dimensions, serving as a quality control for the evaluation.

Trait-correctness association analysis. Finally, to relate rubric-evaluated behavior to task success, we quantify the monotonic association between per-instance trait scores and correctness. For each model and prompt style, we compute Kendall's τ between the aggregated trait score (e.g., median over replicates) and the hit/miss indicator. This provides a robust, tie-aware measure of whether higher DoR or higher ORI tends to co-occur with correct answers under each prompting condition.

Taken together, these steps form a protocol that prioritizes interpretability and replicability in realistic generative settings. By grounding the analysis in observable outputs and avoiding strong psychometric assumptions about fixed latent abilities, TraCE-LLM establishes itself as a scalable, architecture-independent framework for LLM auditing, understood here as the systematic characterization of model behavior across tasks, prompts and replications.

Throughout the paper we adopt four key notions, summarized below for ease of reference.

DoR measures the logical structure and inferential depth of a model's explanation. Scores range from surface-level paraphrase (0-3) to multi-step, self-correcting argumentation (8-10).

ORI captures the degree to which the explanation departs from templated, prompt-echoed wording, ranging from verbatim repetition (0-3) to novel framing and creative synthesis (8-10).

ACE a trait-level pattern in which a model under adversarial stress exhibits apparent gains in correctness while its behavioral stability deteriorates and agreement with peer models decreases.

Compressed Reasoning a correct answer accompanied by a shallow or fragmented justification, operationalized as a hit with low DoR.

Both traits are scored on a $[0, 10]$ semantic interval scale (Table 2); ACE and compressed reasoning are emergent patterns diagnosed from the resulting score distributions.

3.1 Algorithmic description of the TraCE-LLM protocol

We now summarize TraCE-LLM as a step-by-step procedure that makes explicit the two-stage pipeline depicted in Figure 1. Let $\mathcal{M}$ denote the set of models under evaluation, $\mathcal{D}$ the set of multiple-choice datasets, $\mathcal{P} = \{\text{Naive, CoT, Adversarial}\}$ the set of prompt styles, $\mathcal{R}$ the set of experimental replications and $\mathcal{J}$ the panel of model-judges. Let $\mathcal{Z}$ denote the semantic interval rubric used to score DoR and ORI.

Stage I - Response generation

Table 1. Comparison of LLM evaluation systems: ✓ present; - absent; Partial partial support.

Framework	Prompt Strat.	Replications	Semantic Rubric	Variance Modeling	Trait Diagnosis	Interpretability
BIG-Bench	-	Partial	-	-	-	Medium
HELM Liang *et al.* [2023]	-	-	-	-	-	Medium
Chatbot Arena	-	✓	-	Elo	-	Medium
LM Eval Harness	-	✓	-	-	-	Medium
FreeEval Yu *et al.* [2024]	Partial	-	-	Mean	-	Medium
MT-Bench	-	✓	-	Partial	-	Medium
OpenAI Evals	Partial	✓	-	Partial	-	Medium
SCORE Nalbandyan *et al.* [2025]	✓	✓	-	Descriptive	-	Medium
UniEval Li *et al.* [2025b]	-	✓	Embeddings	Intervals	-	Medium-high
TraCE-LLM	✓	✓	✓	Stratification + Bootstrap	✓	High

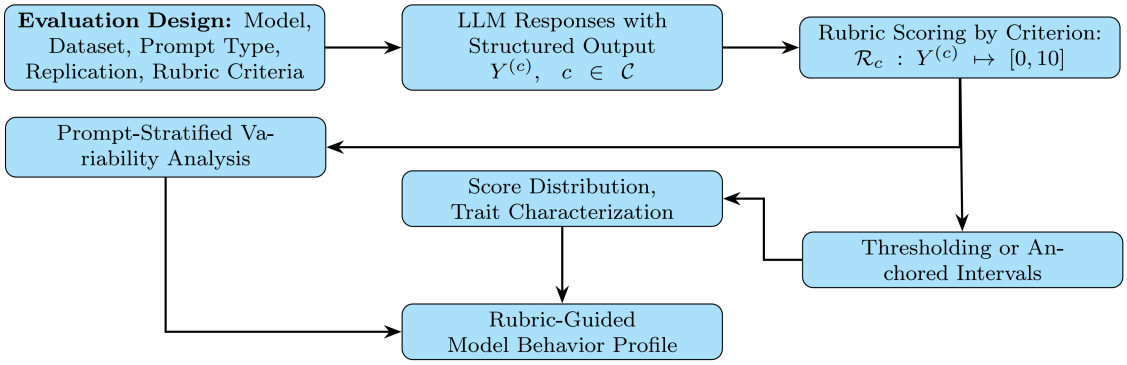

Figure 1. Schematic representation of the TraCE-LLM evaluation process. This article focuses on rubric-based score distributions and on patterns of variability and trait consistency under prompt variation. $Y^{(c)}$ is the response evaluated with respect to criterion $c \in \mathcal{C}$ and $\mathcal{R}_c : Y^{(c)} \mapsto [0, 10]$ is the function that maps the response to a score in the [0,10] range according to the rubric for criterion c.

1. For each dataset $d \in \mathcal{D}$, sample a stratified subset of items $\mathcal{I}_d$ from the original benchmark.
2. For each combination (m, d, i, p, r), construct a prompt from the corresponding template and query model m to obtain a structured JSON response containing the selected alternative, an explanation field and any additional metadata required by the rubric. Store each response indexed by (m, d, i, p, r).

Stage II - Rubric-based evaluation

1. For each stored response (m, d, i, p, r) and each judge $j \in \mathcal{J}$, construct a judging prompt that presents the original item, the model's response and the rubric $\mathcal{Z}$.
2. Query judge j to obtain scalar scores $\text{DoR}^{(j)}_{m,d,i,p,r}$ and $\text{ORI}^{(j)}_{m,d,i,p,r}$ in $[0, 10]$, plus an optional textual justification.
3. Aggregate the panel by computing, for each (m, d, i, p, r), the per-instance median across judges for DoR and ORI, together with a correctness label derived from the selected alternative.

Trait aggregation and stability analysis

1. For each configuration (m, d, p), compute summary statistics over the distribution of per-instance scores: mean, variance and coefficient of variation (CV) for DoR and ORI, as well as weighted F_1 and accuracy for correctness.
2. Use bootstrap resampling over instances to estimate confidence intervals for these quantities and to compute instability indices between prompt styles, following the formalization in the Appendix.
3. Derive trait profiles and pairwise agreement measures (Kendall's τ) across models and prompt styles to characterize latent traits, compressed reasoning and the ACE.

4 Materials and Methods

The protocol[1] follows a nested factorial arrangement that combines five language models, three benchmarks, three prompt styles and five independent replications, totaling $N = 10125$ observations. This design is specifically intended to enable the estimation of within-model variability, ensure comparable standard error across benchmarks and make stratified analyses of prompt effects feasible. Figure 2 synthesizes the generation and evaluation flow, whereas Table 2 describes the semantic scales employed in the evaluation.

- **In Stage I (in blue)**, the process begins with response generation. For a given item from a benchmark, one of the three experimental prompts (Naive, CoT or Adversarial) is applied. The language model then produces a response that is validated and formatted as a standardized `JSON` object. This object contains not only the final answer but also essential metadata such as the chain of reasoning (if applicable) and the justification, constituting the raw material for evaluation.
- **In Stage II (in orange)**, the focus turns to evaluation. The `JSON` generated in the previous stage serves as input to the panel of model-judges. Each one, using the Zero-Shot Semantic Interval Rubric for Evaluation Dimensions (ZSSIRED), Subsection 4.3, scores the response along two independent dimensions: DoR and ORI. The

[1] See Section 'Declarations: Availability of Data and Materials' for the TraCE-LLM Evaluation Datasets, Scripts and Prompts Repository.

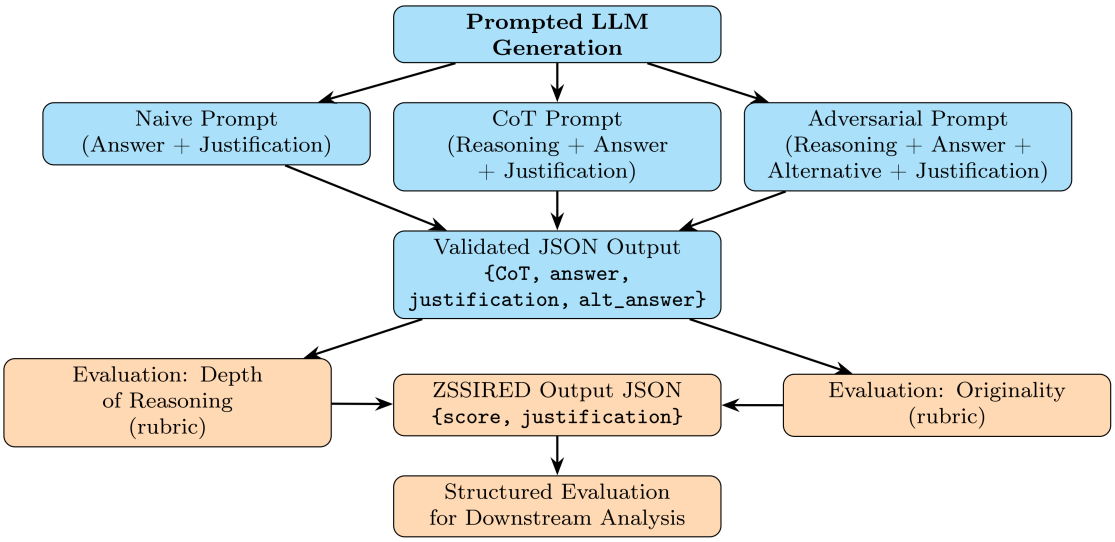

Figure 2. Two-stage TraCE-LLM evaluation pipeline: structured generation (Stage I, in blue) followed by interval-based scoring (Stage II, in orange) using rubric prompts.

outcome of this stage is another structured `JSON` object containing the numerical scores and the justification for this rating.

Correctness (hit/miss) is computed by matching the selected option in `answer` against the benchmark gold label; model-judges are used only to score DoR and ORI under ZSSIRED. This two-stage pipeline ensures that each observation in our final dataset is enriched with metadata about the quality of the model's behavior, enabling the in-depth statistical analysis that follows.

4.1 Experimental Design and Statistical Power

The study evaluates five language models on three well-established benchmarks: MMLU, Hendrycks *et al*. [2021], ARC-Challenge, Clark *et al*. [2018] and HellaSwag, Zellers *et al*. [2019]. To capture the sensitivity of models to input templates, Wang *et al*. [2024b]; Polo *et al*. [2024], each item was subjected to three prompt styles (*Naive*, *CoT* and *Adversarial*). Following recommendations for variance quantification, Madaan *et al*. [2024] and self-consistency, Wang *et al*. [2022], five independent replicas were collected for each condition.

The resulting nested structure, *model* $\rightarrow$ *dataset* $\rightarrow$ *item* $\rightarrow$ *prompt* $\rightarrow$ *replicate*, with a total of 10125 observations, was designed to capture stochastic variability and to support stratified contrasts across prompt styles. Statistical uncertainty is quantified via nonparametric bootstrap over instances ($B = 1000$ resamples). This choice targets estimation precision (Monte Carlo standard error below 1% for the reported bootstrap quantities) rather than classical hypothesis-test power, since the central goal is to characterize dispersion, agreement and stability under stochastic generation, Efron and Tibshirani [1993]; Davison and Hinkley [1997].

4.2 Evaluated Models and Benchmarks

We evaluate five large language models that span a spectrum of capacity and cost: GPT-4.1-nano (OpenAI, 14/Apr/25), GPT-4o-mini (OpenAI, 18/Jul/24), Claude 3.5 Haiku (Anthropic, 04/Mar/24), Grok-3-mini-beta (xAI, 19/Feb/25) and DeepSeek-chat (DeepSeek, 25/Mar/25) OpenAI [2025, 2024]; Anthropic [2024]; xAI [2025]; AI [2024]; DeepSeek [2025]. Parameter counts are proprietary for the OpenAI, Anthropic and xAI models, whereas DeepSeek-chat is built on the open Mixture-of-Experts (MoE) DeepSeek-V3 family, which has 671B total parameters with roughly 37B active per token AI [2024].

All interactions were carried out through the providers' official APIs, in independent sections for each item, with sampling and length parameters left at their documented default values. All Chat Completions API calls used the default settings of `temperature=1` and `top_p=1`. This configuration mimics that obtained by a typical user in the corresponding web chat interfaces and follows the common practice of using default settings in LLM evaluation studies, Thelwall [2025].

The three benchmarks were chosen for their diversity of domains and tasks: MMLU for general knowledge, Hendrycks *et al*. [2021], ARC-Challenge for complex scientific reasoning, Clark *et al*. [2018] and HellaSwag for commonsense narrative completion, Zellers *et al*. [2019]. From each, 45 items were sampled uniformly at random.

4.3 Rubric-Based Evaluation Protocol

Response evaluation is conducted by the Zero-Shot Semantic Interval Rubric for Evaluation Dimensions (ZSSIRED), a rubric designed to be applied by model-judges, ensuring scalability, Zheng *et al*. [2023] and semantic granularity, Osgood *et al*. [1967]. While rubric-driven evaluation has been widely discussed in holistic and unified evaluators Liang *et al*. [2023]; Li *et al*. [2025b]; Biderman *et al*. [2024], ZSSIRED departs from LLM-as-a-judge settings Zheng *et al*. [2023] by enforcing interval anchors grounded on semantic differential theory Osgood *et al*. [1967]. It isolates DoR and ORI as orthogonal axes, which aims to decouple surface correctness from the structure and novelty of reasoning.

ZSSIRED Principles. The rubric's acronym synthesizes its three pillars: (1) Zero-Shot, since the model-judge operates solely from instructions, without labeled examples, an approach whose effectiveness is supported in the literature, Lee *et al*. [2024]; (2) Semantic Interval, because scores are assigned on a continuous [0,10] interval anchored by qualitative descriptors, inspired by Osgood's Semantic Differential, Osgood *et al*. [1967]; and (3) Evaluation Dimensions, as the system is expandable to other behavioral traits, although in this implementation it uses two fundamental dimensions: DoR and ORI.

Rationale for the Criteria. The choice of evaluation dimensions is grounded in established literature. The DoR criterion, in particular, is designed to assess the quality and coherence of the chain of reasoning that the model externalizes. We adopt this approach with full awareness that CoT is not necessarily a faithful reflection of the latent computational process and is often a post-hoc rationalization, Lanham *et al*. [2023]; Turpin *et al*. [2023]. However, this potential infidelity does not invalidate the analysis; it redefines its object. As in high-stakes human domains (e.g., a legal opinion or a medical diagnosis), what is audited and trusted is the quality of the presented justificatory artifact, regardless of the cognitive process that originated it. The explanation itself becomes the object of scrutiny, Doshi-Velez and Kim [2017].

Therefore, in our methodology DoR does not seek to measure an inaccessible 'thought', but rather to quantify an observable and fundamental capability: the model's ability to construct and present coherent, structured and persuasive reasoning. In parallel, the Originality (ORI) criterion builds on studies of automated creativity assessment, which show high correlation between originality scales and human judgment, Li *et al*. [2025a].

Operationally, both traits are instantiated as semantic interval scores in the [0,10] range defined in Table 2, which details the full rubric. Higher DoR intervals correspond to multi-step, well-structured chains of reasoning with explicit intermediate inferences, whereas lower intervals capture fragmented, single-step or missing reasoning. Higher ORI intervals correspond to creative reformulations that introduce new phrasing or abstraction and avoid prompt echo, whereas lower intervals indicate predominantly generic templates or near-literal reproduction of the question or reference. These rules are applied consistently by the model-judges when assigning scores, making DoR and ORI directly observable quantities for each response instance.

To illustrate how these intervals manifest in practice, we highlight two representative high-trait instances from our dataset. In an ARC-Challenge science item asking which option is an example of a learned behavior, Claude 3.5 Haiku under adversarial prompting begins its chain-of-thought by explicitly defining the target concept ("First, define learned behavior"), then evaluates each alternative before stating the final choice. The explanation is organized into numbered steps and uses its own phrasing instead of echoing the stem. The panel of model-judges assigns DoR scores around 8.7 and ORI scores around 8.1 to these responses, placing them in the upper intervals of Table 2.

A similar pattern appears in a MMLU ethics question about Baier's account of genuine moral rules, where Grok-3-mini-beta under Chain-of-Thought prompting first restates the question in its own words ("Understand the question"), then reconstructs Baier's thesis and, only afterwards, maps it to the alternatives. This explanation unfolds as a multi-step argument with paraphrased content, yielding DoR scores above 9 and ORI scores around 8. Together, these cases exemplify how high DoR and high ORI correspond to explanations that are both structurally rich and lexically non-templated.

Peer Evaluation Protocol. Each response is scored by all five models in the study, which act as a panel of judges. This approach includes each model's self-assessment, enabling the analysis of self-enhancement bias, a phenomenon documented in the literature, Panickssery *et al*. [2024]. The final score for each response is the median of the scores assigned by the panel of judges. This aggregation strategy, grounded in robust statistics, Hampel [1974], minimizes the impact of extreme evaluations or biases from a single model-judge, Hu *et al*. [2025]. To minimize template-level overlap with public rubrics, all descriptors were authored for this study. When a descriptor follows established semantic oppositions, we explicitly note it as inspired by semantic differential theory Osgood *et al*. [1967]. This implements a noise-aware judge aggregation: for each item we compute the per-item median across J model-judges with bootstrap confidence intervals, which reduces single-judge bias and overconfidence Hu *et al*. [2025]; Zheng *et al*. [2023].

4.4 Prompt Styles

Stratification by prompt styles is a pillar of TraCE-LLM's experimental design. Rather than using a single template, we employ three distinct approaches, each functioning as a probe to investigate different facets of model behavior. The selection of these styles is informed by the literature on LLM interaction and evaluation.

- **Naive Prompt:** This approach represents the baseline condition, simulating a standard user interaction in which a direct answer is requested without reasoning instructions. It serves as our experimental control, establishing the model's performance in a simple zero-shot paradigm, as popularized by Brown *et al*. [2020].
- **Chain-of-Thought Prompt:** This technique is employed to explicitly elicit the model's intermediate reasoning steps. By instructing the LLM to "think step by step", CoT makes the inferential process, otherwise latent, an observable artifact that can be analyzed by our DoR rubric. The effectiveness of CoT for improving performance on complex reasoning tasks was demonstrated seminally by Wei *et al*. [2023].
- **Adversarial Prompt:** This prompt serves as a "stress test" for model robustness. It introduces elements designed to confuse or induce errors, such as suggesting flaws in the model's own reasoning or the existence of an alternative answer. This approach is inspired by red teaming techniques and research on adversarial attacks Zou *et al*. [2023].

Table 2. Zero-Shot Semantic Interval Rubric for Evaluation Dimensions.

Criterion	Interval	Descriptor
DoR	(9,10]	Multi-step, stratified reasoning with strong inferential structure.
	(7,9]	Solid, structured reasoning with some complexity.
	(5,7]	Basic logical sequence with minimal elaboration.
	(3,5]	Superficial reasoning with gaps or simplifications.
	(1,3]	Fragmented logic or single-step heuristic response.
	[0,1]	No trace of reasoning or incoherent justification.
ORI	(9,10]	Highly original and creative formulation, abstract reasoning, no lexical overlap with the prompt.
	(7,9]	Strong reformulation and innovative phrasing with clear semantic transformation.
	(5,7]	Some novel elements yet retains the input's structure or vocabulary.
	(3,5]	Predominantly generic or templated response; moderate echo of the prompt.
	(1,3]	Prompt echo dominates; low semantic variation.
	[0,1]	Literal or near-literal reproduction of the prompt or reference.

The exact templates used for each of these three experimental conditions are presented below.

Naive Prompt

```
Read the following question and its options: {item}. Select the
answer you believe is correct (A, B, C or D). Your output must
be a single valid JSON object with these fields (no extra text
or markdown):
  {
   "answer": "<A, B, C or D>",
   "justification": "<brief explanation of your choice>"
  }
Ensure the JSON is parseable by a standard JSON parser (double
quotes for keys, no trailing commas).
```

Chain-of-Thought Prompt

```
You are an expert at solving multiple-choice questions. Read the
following question and its options carefully: {item}. To arrive
at the solution, break down the problem into smaller, manageable
stages wherever possible. Enumerate and show your reasoning step
by step (1., 2., ...), then select the answer you believe is
correct (A, B, C or D). Your output must be a single valid JSON
object with these fields (no extra text or markdown):
  {
   "CoT": "<detailed reasoning here>",
   "answer": "<A, B, C or D>",
   "justification": "<brief explanation of your choice>"
  }
Ensure the JSON is parseable by a standard JSON parser (double
quotes for keys, no trailing commas).
```

Adversarial Prompt Template The template follows standard red-teaming and jailbreak probing principles Ganguli *et al*. [2022]; Shen *et al*. [2024]; Zou *et al*. [2023], inserting plausible distractors while avoiding harmful content or private data. It targets reasoning brittleness rather than toxicity or policy violations, so that effects on correctness and justification can be disentangled.

```
You are an expert at solving multiple-choice questions, but your
reasoning might be flawed. Read the following question and its
options carefully. Consider that some subtle aspects of the
question might be designed to mislead you: {item}. Also, be
aware that there might be an alternative interpretation or a
hidden assumption within the question that could change the
correct answer. If you determine that none of A, B, C or D is
fully correct, choose E and provide your own alternative answer.
To arrive at the solution, meticulously examine each step of
your reasoning and question every assumption you make. Enumerate
and show your reasoning step by step (1., 2., ...), then select
the answer you believe is correct (A, B, C, D or E). Your output
must be a single valid JSON object with these fields (no extra
text or markdown):
  {
   "CoT": "<detailed reasoning here, questioning your
          own assumptions>",
   "answer": "<A, B, C, D or E>",
   "alternative_answer": "<text of your E option or
                          null if not used>",
   "justification": "<brief explanation, highlighting
                    potential doubts>"
  }
Ensure the JSON is parseable by a standard JSON parser (double
quotes for keys, no trailing commas).
```

4.5 Research Hypotheses

The experimental design was crafted to test five central hypotheses, formulated from gaps and ongoing debates in the literature on LLM behavior, Sections 1 and 2. Each hypothesis addresses a facet of variability, robustness or reasoning structure in the models and its investigation guides the results analysis presented in Section 5.

Asymmetric Trait Stability. The literature distinguishes between factual knowledge and structured reasoning, which are relatively convergent across models trained on massive corpora and the stylistic aspects of text generation, which are directly influenced by the randomness of decoding methods, Holtzman *et al*. [2020]; Li *et al*. [2025a]. The former, linked to DoR, is expected to be more stable than the latter, linked to ORI. By projecting trait vectors and aggregating across replicas, we isolate systematic prompt-induced dispersion from stochastic variance, providing a complementary lens to standard accuracy-based stability checks Madaan *et al*. [2024]; Pezeshkpour and Hruschka [2024]. This leads us to formulate the hypothesis:

> **H1:** The Depth of Reasoning (DoR) dimension will exhibit lower variability (inter-model and intra-prompt) than the Originality (ORI) dimension.

Adversarial Collapse in Compact Models. Scaling laws and the notion of "emergent abilities" suggest that larger models are not only more knowledgeable but also more robust and better able to follow complex, nuanced instructions, Kaplan *et al*. [2020]; Wei *et al*. [2022]. Smaller models, on the other hand, tend to be more "fragile" and susceptible to prompt perturbations, Zou *et al*. [2023]. We test whether this fragility manifests as a disproportionate degradation under adversarial stress.

> **H2:** Adversarial prompts will degrade performance (increasing ORI variance and reducing coherence) more sharply in lower-capacity models.

Robustness of the Median Ensemble. Robust statistics establishes that the median is a superior estimator of central tendency in the presence of outliers: it minimizes L_1 risk and possesses a bounded influence function Hampel [1974]. In parallel, evaluation by multiple judges is a standard practice to increase reliability Hu *et al*. [2025]. Combining these principles, we hypothesize that a baseline aggregated by the

median of scores from multiple model-judges will be more stable than any individual evaluator.

In LLM evaluation, per-instance median aggregation plays a role analogous to consensus over multiple reasoning chains Wang *et al*. [2022] and to recent consistency-and-robustness protocols Nalbandyan *et al*. [2025]; Raj *et al*. [2025], but applied to rubric scores rather than to chain voting alone.

> **H3:** The per-instance median baseline will exhibit lower absolute deviation from the central tendency than any individual model.

Partial Trait Separability. LLM evaluation is moving beyond single accuracy metrics toward a multifaceted view of capabilities, Liang *et al*. [2023]. As in human cognition, where reasoning and creativity are correlated yet distinct constructs, we expect the same to apply to LLMs. The literature underpinning our rubrics already treats deep reasoning, Mondorf and Plank [2024] and creativity, Li *et al*. [2025a], as separate research dimensions.

> **H4:** The DoR and ORI dimensions will be only partially dependent, justifying the need for multidimensional evaluation.

Incidence of Compressed Reasoning. Research on Chain-of-Thought infidelity suggests that LLMs do not always use explicit reasoning to arrive at the answer, sometimes relying on "shortcuts" or learned heuristics, Turpin *et al*. [2023]; Lanham *et al*. [2023]. This phenomenon of "shortcut learning", Geirhos *et al*. [2020], would imply that a model may produce the correct answer (via pattern recognition) and only then construct a justification, which may be of low quality.

> **H5:** A non-trivial fraction of correct answers will exhibit low DoR, indicating the use of heuristic shortcuts rather than deep reasoning.

4.6 Human Anchor Validation

To address the risk of bias inherent to model judges, we include a human anchor validation on a stratified subset of responses. We sample $k = 15$ base items per benchmark (ARC, HellaSwag, MMLU). Each base item is instantiated under the three prompt regimes $p \in \{\text{naive}, \text{cot}, \text{adv}\}$, yielding 45 responses per benchmark and $n = 135$ responses in total. The sampling is stratified by benchmark and prompt type and balanced across generating models.

Human annotation and randomization. Three independent human evaluators score each sampled response using the same semantic interval rubrics for DoR and ORI. Items are presented in fully randomized order to mitigate potential human biases, including initial unfamiliarity with the rubric at the beginning of the evaluation session, increasing familiarity and potential leniency in the middle, and fatigue effects that could disproportionately influence scores toward the end.

Alignment metrics. We assess human-LLM alignment with Kendall's τ_b (tie aware rank correlation) between the per response median human score and the per response median model judge ensemble median. We report Krippendorff's α as an inter rater reliability measure within humans and within the LLM judge panel. As ordinal agreement checks, we discretize the $[0, 10]$ scale into the rubric semantic intervals (six bins) and report weighted Cohen's κ between the binned human and binned ensemble medians. As a categorical sanity check, we report Cramér's V with the corresponding χ^2 test p value from the contingency table. A three bin variant is reported as a sensitivity analysis, while a two bin split is used only as a threshold diagnostic.

Detectability. To contextualize non-significant Kendall results, we report an approximate minimal detectable effect (MDE) for a two sided Kendall τ test under a large sample null standard error approximation, using $\alpha = 0.05$ and power $= 0.8$. For the overall test ($n = 135$), the MDE is approximately 0.16, while for prompt stratified analyses ($n \approx 45$ per stratum) the MDE increases to approximately 0.29. We interpret the alignment results accordingly and treat residual disagreement as evidence of potential judge bias and non interchangeability under the same rubric.

5 Results

This section presents the empirical findings of the TraCE-LLM protocol, confronting them directly with the five research hypotheses. The analysis begins with overall trends, then investigates into the validation of each hypothesis, integrating quantitative data and theoretical interpretation. The section culminates with a synthesis of methodological and practical implications for LLM evaluation and development.

5.1 Asymmetric Stability between Reasoning and Originality

Hypothesis 1 posits that DoR is an inherently more stable dimension than ORI. The evidence below conclusively corroborates this asymmetry.

Descriptive Variability Analysis. Table 3 reveals the first sign of asymmetry. DoR scores are consistently high and concentrated (means between 7.9 and 8.3), with low coefficients of variation (CV), i.e., dispersion normalized by the mean (see Appendix for the formal definition). These CV values remain within $[0.12, 0.18]$. In contrast, originality exhibits much greater dispersion, with lower means (4.0 to 6.9) and a CV reaching 0.54 in GPT-4.1-nano. Figure 3 visually illustrates this disparity: DoR boxplots are compact and symmetric, whereas ORI boxplots are wide with long tails, indicating substantially higher variability.

Inter-Rater Agreement. Agreement analysis, measured by Kendall's τ (Figure 4), reinforces the pattern. Agreement for DoR is high (median $\tau_{\text{DoR}} = 0.72$), indicating that model-judges share a homogeneous understanding of what

Table 3. Descriptive statistics by model, grouped by rubric criterion.

Criterion	Model	Mean	SD	Median	Min	Max	CV
DoR	3.5-haiku	7.95	1.22	8.5	2.5	9.7	0.154
	4.1-nano	8.23	1.44	8.5	1.0	9.7	0.175
	4o-mini	8.34	1.03	8.5	3.0	10.0	0.124
	ds-v3	7.92	1.34	8.5	3.0	9.5	0.170
	grok-3-mini	8.26	1.26	9.0	2.0	10.0	0.152
	median	8.15	1.11	8.5	2.5	9.5	0.136
ORI	3.5-haiku	6.88	1.32	7.5	1.5	9.2	0.192
	4.1-nano	4.02	2.16	3.5	1.0	9.2	0.537
	4o-mini	6.47	1.14	6.5	2.0	9.0	0.177
	ds-v3	6.18	1.46	6.5	1.0	8.5	0.236
	grok-3-mini	6.65	1.16	6.0	2.0	9.5	0.175
	median	6.38	1.14	6.5	2.0	8.5	0.179

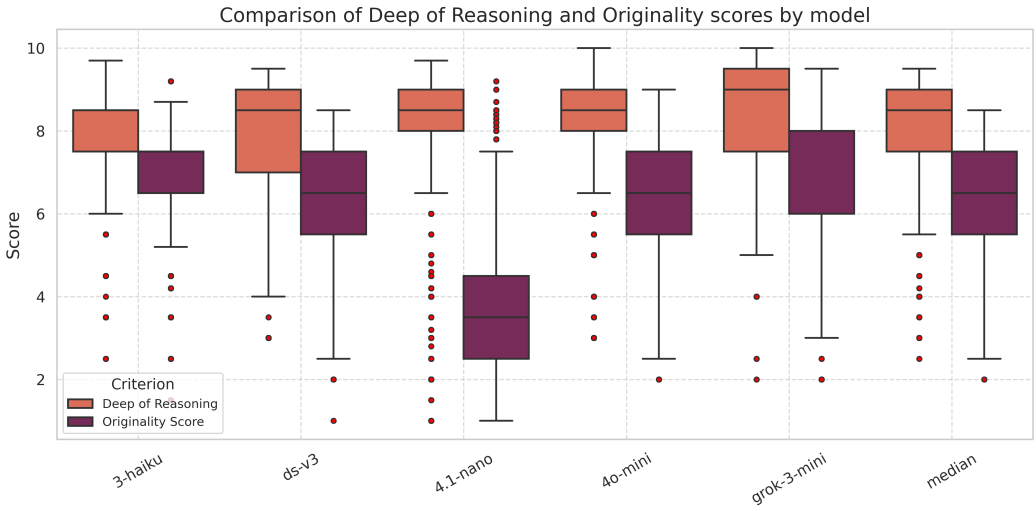

Figure 3. Distribution of DoR and ORI scores across evaluated models. Each boxplot shows the interquartile range, the median and outliers.

Table 4. Difference in mean score (ΔMean) and coefficient of variation (ΔCV) between correct and incorrect responses, by model and evaluation criterion. Positive ΔMean indicates gain when correct; negative ΔCV indicates improved consistency.

Criterion	Model	ΔMean	ΔCV	Interpretation
Depth of Reasoning	3.5-haiku	+0.1991	–0.0536	Slight improvement with more stability
	4.1-nano	+1.3489	–0.2186	Strong improvement and stabilization
	4o-mini	+1.0252	–0.1511	Large gain with high consistency
	ds-v3	+0.3749	–0.0572	Moderate improvement with stability
	grok-3-mini	+0.0881	+0.0019	Stable performance across outcomes
	median	+0.4971	–0.0753	Aggregate improvement with stability
Originality	3.5-haiku	–0.2901	–0.0032	Slight drop in originality when correct
	4.1-nano	+0.7569	–0.0073	Higher originality, yet still unstable
	4o-mini	+0.0441	–0.0242	Stable and consistent across outcomes
	ds-v3	–0.6458	–0.0123	Expressive failures penalized
	grok-3-mini	–0.5554	+0.0017	Creativity penalized when correct
	median	–0.3622	–0.0219	Slight homogenization when correct

constitutes good reasoning. For ORI, agreement is only moderate (median $\tau_{\text{ORI}} = 0.46$), reflecting the more subjective and stylistic nature of originality, which is more susceptible to generation stochasticity.

Interpretation and Implications. The evidence confirms **H1**. The high stability of DoR suggests that logical structure is a more fundamental and convergent trait across LLMs, less affected by decoding randomness. The volatility of ORI, in turn, indicates that stylistic expression is where the probabilistic nature of the models most strongly manifests. For evaluation practice, this implies that reporting means alone is insufficient; it is essential to analyze dispersion metrics (CV, IQR) and agreement to diagnose a model's expressive stability.

5.2 Adversarial Collapse in Compact Models

Hypothesis 2 anticipated that adversarial prompts would induce a disproportionate "collapse" in lower-capacity models. Our results not only confirm this hypothesis but also reveal a paradoxical behavioral pattern that we term the **Adversarial Compensation Effect (ACE)**.

Performance Degradation. The first layer of evidence comes from classical metrics. Table 5 shows that GPT-4.1-nano, a compact model, suffers a significant drop of 4.9 points in WF_1 under the adversarial prompt, whereas more robust models such as Grok-3-mini maintain their performance (Figure 5).

Instability of Behavioral Traits. The Instability Index summarizes how far a model's trait profile moves when we

Figure 4. DoR exhibits greater internal consistency than ORI, especially under Naive prompts. Correlations between the two criteria remain weak.

Table 5. Bootstrap estimates for the mean and standard deviation (SD) of the WF_1-score by model and prompt type.

Model	Prompt	Mean WF_1	SD
4o-mini	Adversarial	0.7832	0.0165
	CoT	0.7575	0.0161
	Naive	0.7022	0.0180
4.1-nano	Adversarial	0.8116	0.0144
	CoT	0.8520	0.0144
	Naive	0.8602	0.0128
3.5-haiku	Adversarial	0.8396	0.0134
	CoT	0.8587	0.0124
	Naive	0.8366	0.0129
grok-3-mini	Adversarial	0.8893	0.0114
	CoT	0.9139	0.0106
	Naive	0.9101	0.0101
ds-v3	Adversarial	0.8233	0.0126
	CoT	0.8883	0.0126
	Naive	0.8917	0.0107

switch prompt styles; larger values mean larger behavioral shifts. Concretely, it averages pairwise distances between a model's prompt-specific trait vectors in the two-dimensional PCA space (see Appendix for the formal definition). This operationalization complements variance audits Madaan *et al.* [2024] and prompt-sensitivity analyses Pezeshkpour and Hruschka [2024], but it quantifies dispersion in latent trait space rather than fluctuations in accuracy alone.

Table 6 highlights the fragility of compact models: GPT-4.1-nano and GPT-4o-mini exhibit high instability, indicating that their behavioral profiles change sharply under adversarial stress. Grok-3-mini shows the highest value (1.0918), but its instability has a different character: it tends to trade away originality to preserve correctness, whereas GPT-4.1-nano exhibits an "explosion of stylistic variability" and a trial-and-error strategy under pressure.

The Explosion of Stylistic Variability. The most compelling evidence of collapse appears in Table 4. In GPT-4.1-nano originality variability (ΔCV_{ORI}) explodes under the

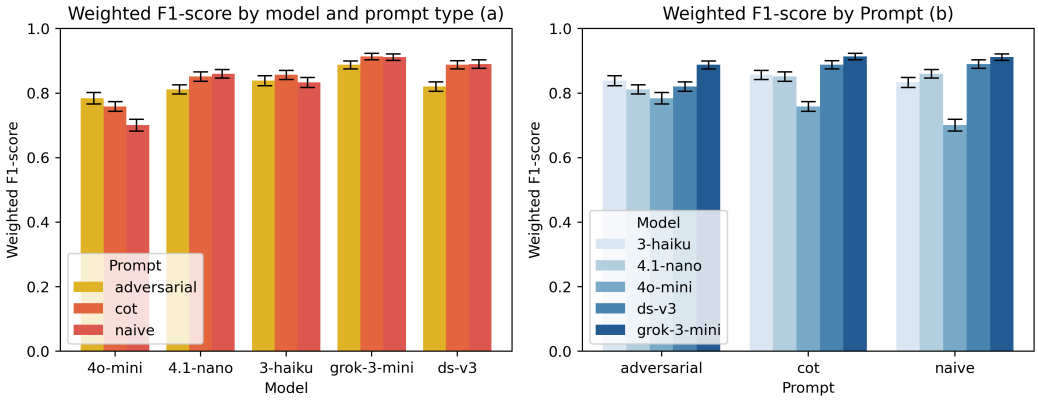

Figure 5. Weighted WF_1-score by model and prompt type (a) and by prompt across models (b), with bootstrap means and 95% confidence intervals.

Table 6. Instability Index for each model across prompt types. Higher values indicate greater trait volatility.

Model	Instability Index
3.5-haiku	0.2728
ds-v3	0.3891
4o-mini	0.4798
4.1-nano	0.5393
grok-3-mini	1.0918

Table 7. Factor loadings of metrics on the first two principal components.

Metric	PC1	PC2
Mean WF_1	−0.497	−0.123
Mean DoR	**0.922**	0.197
DoR Variance	**−0.762**	−0.537
Mean Originality	−0.300	**0.891**
Originality Variance	**0.744**	−0.516

adversarial prompt. This suggests a "desperate trial-and-error" strategy: to compensate for difficulty, the model sacrifices stylistic consistency, generating erratic outputs.

The Behavioral Map of Collapse. To make collapse dynamics visually interpretable, we map each (model, prompt) configuration into a two-dimensional PCA space (Figure 6, left). The intuition is simple: nearby points correspond to similar behavioral profiles, while large prompt-induced displacements indicate behavioral change. The first two principal components explain 74.5% of the total variance, supporting a faithful 2D summary; the formal PCA specification is given in Appendix 7. Loadings (Table 7) show that the axes admit a direct interpretation:

- PC1 (46%): reasoning depth and stability. Higher PC1 corresponds to deeper and more consistent reasoning (high mean DoR and lower DoR variance).
- PC2 (28%): originality. Higher PC2 corresponds to more novel, non-templated expression (high mean ORI).

In this map, the adversarial prompt acts as a **degradation vector**. Compact models such as 4.1-nano exhibit a strong shift toward the lower quadrants, representing behaviors of "misaligned confidence" and "suppressed reasoning". The dendrogram (Figure 6, right) quantitatively confirms this collapse, showing the statistical isolation of 4.1-nano's adversarial configurations, which form a cluster distinct from all other behaviors.

Factor Analysis. Factor analysis clarifies what drives compensation under adversarial stress. The association between correctness and DoR is moderate ($\bar{r} \approx 0.33$), whereas the association with ORI is weak ($\bar{r} \approx 0.08$). This indicates that "compensation" is primarily achieved by increasing DoR rather than by being more creative.

To assess whether DoR and ORI remain empirically distinct under stress, we use HTMT as a discriminant-validity diagnostic (Appendix). Intuitively, HTMT values near or above the conventional 0.90 threshold indicate that two constructs are no longer cleanly separable. The observed HTMT ratio of 1.6580 Henseler *et al*. [2015] suggests that, under adversarial pressure, DoR and ORI become statistically blurred, placing **Hypothesis H4** under evidence and motivating a context-dependent interpretation of trait separability.

The Adversarial Compensation Effect (ACE): A Gap in the Literature. The literature extensively documents that adversarial prompts can induce instability and degrade LLM performance, Ganguli *et al*. [2022]; Shen *et al*. [2024]; Pezeshkpour and Hruschka [2024]. Typically, these studies report a drop in accuracy or response coherence. Our results, however, reveal a more complex and paradoxical pattern, in which an apparent improvement in correctness metrics masks a profound behavioral destructuring.

Although sporadic observations of performance gains in adversarial scenarios have been noted in niche contexts such as in-context learning Kang *et al*. [2024], they have not been systematically studied nor linked to a trait-consistency analysis. We therefore delimit ACE as a trait-level phenomenon that persists after controlling for aggregation across judges and replicas, in contrast to adversarial helpfulness or in-context gains Kang *et al*. [2024].

Prior sensitivity studies focus on prompt phrasing effects on accuracy Pezeshkpour and Hruschka [2024]; here we dissociate answer correctness from justification quality and measure shifts in behavioral traits under stress. What is missing from the current taxonomy of LLM failures is a term and operational definition for this complete phenomenon: the gain in a superficial metric that occurs precisely at the expense of a degradation in internal consistency and agreement with peer models. In view of this gap, we formalize it as follows:

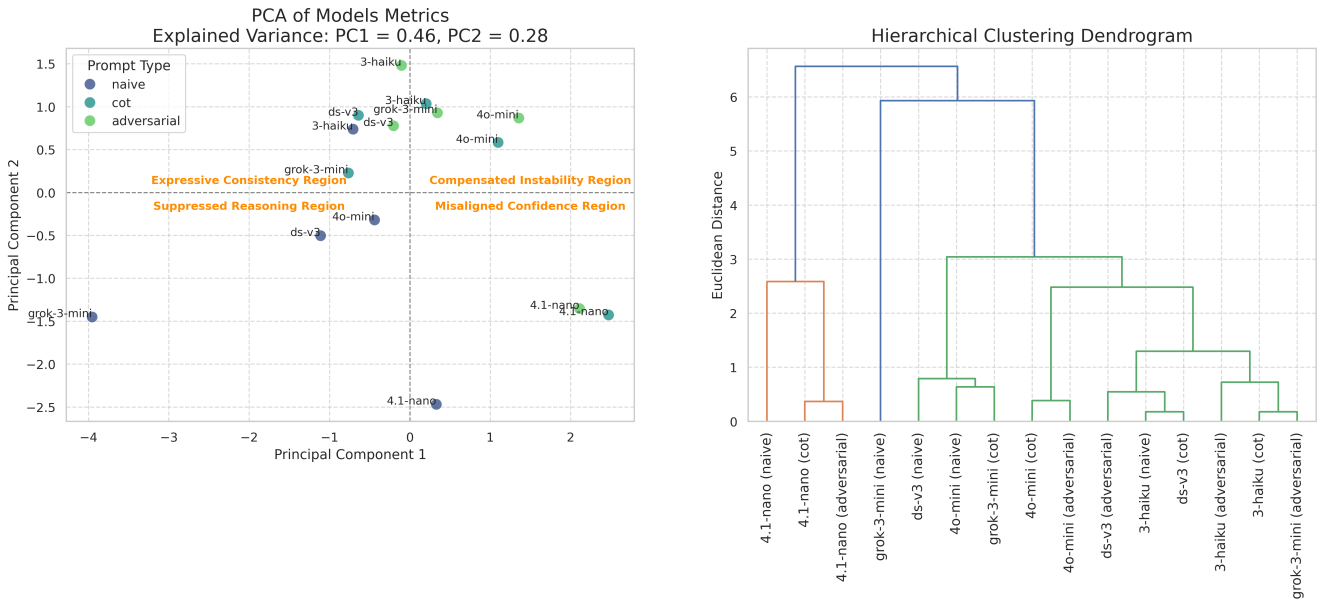

Figure 6. Left: projection of model–prompt configurations into PC space (PC1 = 46%, PC2 = 28%). Right: Ward linkage dendrogram on these components. The PCA reveals latent structure and interpretable regions of model behavior, while the dendrogram highlights similarity clusters models across prompt strategies.

Adversarial Compensation Effect (ACE). *A phenomenon in which a language model, notably lower-capacity ones, responds to an adversarial prompt by exhibiting a paradoxical gain in superficial correctness metrics (e.g., accuracy) while simultaneously demonstrating amplified volatility in latent behavioral traits (e.g. originality, coherence) and a drop in agreement with peer models. ACE is a deceptive failure mode that masks structural instability with a veil of accuracy.*

The pattern becomes clearer when we inspect concrete Naive-Adversarial pairs that satisfy our ACE criterion (see Appendix for the formal definition). We summarize three representative cases below.

Case 1 HellaSwag: dressing for work (GPT-4o-mini). For this item, GPT-4o-mini improves from hit_Naive $= 0$ to hit_Adv $= 1$, with mean DoR increasing from 5.8 to 8.6 and mean ORI from 4.4 to 5.7. The coefficient of variation for DoR decreases from 0.21 to 0.02, indicating more stable reasoning depth, whereas the coefficient of variation for ORI increases from 0.26 to 0.33 and Kendall's τ between trait rankings becomes negative ($\Delta\tau \approx -0.17$). This matches the ACE pattern rule.

```
Context:
[header] How to dress your best at work
[title] Always dress for the job you want.
[step] If you're currently a paralegal, but you
aspire to be a lawyer one day, then dress like
a lawyer. If you're a worker who has ambitions
of moving into management, then dress like a
manager.

Options:
  A. You need to look professional and stick to
     the dress code for the job you have chosen.
     Don't worry if your clothes don't hang well.
  B. As for clothes, always dress to your
     requirements. However, there are some basic
     guidelines that you should follow for any
     job: dress business casual, efficient and
     well-mannered.
  C. However, if you know what's going on at
     work, you can bite your tongue and go for
     it. For example, if you're a salary analyst
     you might wear dress pants and a tie and a
     button down shirt.
  D. You might be surprised by how seriously
     people take you by the clothes you wear and
     your overall appearance. Keep your clothes
     neat, clean and in good condition.
```

Response under the Naive prompt, the model justifies its (incorrect) choice: Option A aligns most closely with the context's advice to dress appropriately for the role you aspire to, emphasizing professionalism and adherence to dress codes relevant to your target position. It directly supports the idea of dressing for the job you want, including substeps about looking professional and proper fit, which complement the provided guidance. Other options either introduce unrelated advice or focus on general attire guidelines that are less specific to the context.

Response of the Chain-of-Thought and justification generated under the Adversarial prompt:

1. The question emphasizes dressing for the job you want, implying that clothing choice should reflect future aspirations rather than current position.
2. Option A suggests looking professional and sticking to the dress code, but it lacks explicit advice about dressing aspirationally, which is central to the given context.
3. Option B advises dressing according to one's requirements and mentions basic guidelines like business casual

and being well-mannered, which aligns with the general advice but doesn't directly connect to dressing for a future role.
4. Option C mentions knowing what's going on at work and provides an example of dressing code (dress pants, tie, shirt), but it seems more about current role appropriateness than aspirational dressing.
5. Option D emphasizes keeping clothes neat and in good condition and notes how appearance affects perception, which partially aligns with the idea of projecting an aspirational image but doesn't explicitly mention dressing for the desired future role or aspirational dressing.
6. Considering all options, none perfectly encapsulate the key message: dressing like the role you aspire to, not necessarily the one you currently hold. However, D most closely emphasizes overall appearance and projecting a serious impression, which can be interpreted as dressing to match future aspirations, despite its less explicit focus on aspirational dressing.
7. Recognizing the subtlety and possible misdirection, the best choice aligns with D's emphasis on neatness, appearance and projecting an image, which indirectly supports dressing for the job one wants.
8. Option D, while not explicitly about aspirational dressing, emphasizes neatness and appearance, which are key to projecting an image aligned with a future role. It seems to best approximate the core advice of dressing for the job you want, despite the subtlety and potential for misinterpretation in the question.

Note that: In this case, the adversarial prompt forces the model to abandon the superficial lexical heuristic ("Option A aligns most closely with the context") and to re-frame the problem in terms of the underlying social norm (projecting a future role), increasing correctness and DoR while making the style of justification more variable across replications.

Case 2 MMLU: Baier on genuine moral rules (GPT-4o-mini). In this ethics item, GPT-4o-mini also satisfies the ACE criterion: hit_Naive $= 0$, hit_Adv $= 1$, DoR_mean increases from 5.6 to 9.08 and ORI_mean from 6.1 to 6.4. DoR_CV drops from 0.27 to 0.01, while ORI_CV increases from 0.22 to 0.32 and Kendall's τ becomes negative ($\Delta\tau \approx -0.26$).

```
Question: Baier argues that genuine moral rules:

Choices:
  A. must be for the good of human beings.
  B. make take into account the interests of
     all sentient beings.
  C. must take into account the interests of
     all living beings.
  D. are primarily directed toward promoting
     self-interest.
```

Response under the Naive prompt, the model justifies its (incorrect) choice: Baier's view emphasizes that genuine moral rules should consider the interests of all sentient beings, not just humans or all living beings, to ensure comprehensive moral consideration.

Response of the Chain-of-Thought and justification generated under the Adversarial prompt:

1. The question attributes the argument to Baier regarding 'genuine moral rules.'
2. I recall that Baier discusses morality as being primarily concerned with human well-being and the interests of rational agents, emphasizing human interests rather than all sentient beings or all living beings.
3. Option A states that moral rules 'must be for the good of human beings,' which aligns well with Baier's focus on human interests.
4. Option B extends this to all sentient beings, which Baier may not necessarily endorse since his emphasis is on human rationality and moral responsibility rooted in human concern.
5. Option C broadens further to all living beings, which seems even less consistent with Baier's human-centered view.
6. Option D suggests that moral rules are primarily directed toward promoting self-interest, which departs from Baier's view that true moral rules are about impartial concern for others, not self-interest.
7. Considering these, the most accurate reflection of Baier's stance appears to be option A, as his view emphasizes moral rules being for the good of human beings.
8. However, I must consider the possibility that 'genuine moral rules' imply impartiality or universal concern, which might make B or C plausible. But Baier's literature primarily discusses human morality, not necessarily extending to all sentient or living beings, thus making A most fitting.
9. I also note that none of the options explicitly mention the notion of moral rules being 'impartial' or 'universally applicable,' which is key in many moral theories but less emphasized in Baier's approach.
10. Given this, I conclude the best fit is A, but I acknowledge that a subtlety or reinterpretation might question whether it is solely about humans or includes a broader scope.
11. If none seem perfect, E would be justified, but given the options, A seems most aligned with Baier's main concern.
12. Baier's ethical philosophy emphasizes morality as concerning human well-being and rational concern. While some options suggest broader scopes, Baier's focus is primarily on humans, making A the most appropriate choice. However, there might be subtle debates about whether he considers interests beyond humans, but given common interpretations, A is the best fit.

Note that: Here, the adversarial prompt again shifts the model away from a generic, but incorrect, gloss ("all sentient beings") towards a more faithful reconstruction of Baier's human-centered view, dramatically increasing DoR and stabilizing, while making the precise wording and originality of the justification more variable.

Case 3 ARC: energy waste in a toaster (GPT-4.1-nano). In this physics item, GPT-4.1-nano transitions from hit_Naive $= 0$ to hit_Adv $= 1$, with DoR_mean increasing

from 5.3 to 8.94 and ORI_mean from 5.6 to 6.7. DoR_CV decreases from 0.15 to 0.03, whereas ORI_CV increases from 0.17 to 0.23 and τ becomes negative ($\Delta\tau \approx -0.25$).

```
Question: Tonya plugged in a toaster. She placed
two slices of bread in the toaster, pushed down
the lever and the heating coils glowed. She
noticed that only one form of energy toasted
the bread. Which form of energy was wasted in
the process?

Choices:
  A. infrared heat energy
  B. visible light energy
  C. chemical energy
  D. electrical energy.
```

Response under the Naive prompt, the model justifies its (incorrect) choice: While the toaster uses electrical energy to function, the heating coils convert that electrical energy into thermal energy to toast the bread. The process results in some electrical energy being wasted as it is not fully converted into the desired form, thermal energy.

Response of the Chain-of-Thought and justification generated under the Adversarial prompt:

1. Let's first analyze the toaster's operation. When Tonya plugged in the toaster, electrical energy was converted into thermal energy to toast the bread. This thermal energy primarily comes from the heating coils glowing and emitting infrared heat.
2. The question states that only one form of energy toasted the bread. Here, we may interpret this to mean that the primary energy interacting with the bread was thermal (infrared heat energy), which indeed is correct. However, this does not rule out other forms of energy potentially being produced during operation.
3. Next, we need to evaluate the provided options for wasted energy specifically.
4. Infrared heat energy (A) is used to toast the bread, thus it is not wasted.
5. Visible light energy (B) may be emitted by the heating coils but is not significant in the toasting process, making this energy relatively wasted.
6. Chemical energy (C) is not particularly relevant; the bread itself contains some chemical energy, but it does not factor into the waste energy of the toaster's operation.
7. Electrical energy (D) is transformed into heat energy in this process and while some energy may be lost as waste heat in the wiring or other components, the direct operation of toasting does not waste it as it is utilized.
8. Thus, while the operation uses infrared heat energy effectively, the visible light energy emitted may be considered wasted since it contributes little to toasting bread.
9. Therefore, looking at energy conversion, visible light is likely the least useful form of energy in this context and it may be reasonable to consider it as wasted energy.
10. Therefore, I decide on option B as the most appropriate relative to additional interpretation of wasted energy.
11. The process of toasting uses primarily infrared heat energy. Visible light energy may be produced as a byproduct but serves no significant purpose in the toasting process, categorizing it as wasted energy.

Note that: In this case, the adversarial prompt leads the model to explicitly decompose the energy conversions involved, distinguishing between the effective energy used to toast the bread and the wasted visible light, thereby increasing correctness and the stability of reasoning depth while again inducing higher variability in expressive traits.

The evidence therefore confirms H2. Adversarial collapse is real and manifests through ACE. The implication is clear: robustness evaluation cannot rely solely on correctness metrics and must incorporate the measurement of trait stability.

5.3 The Robustness of the Median Ensemble as a Baseline

Hypothesis 3 considers using the median of scores as a robust baseline. Our results show that the median ensemble is not only stable but acts as the group's behavioral center of gravity.

Stability and Centrality. The analysis revealed that, in 84% of items, the median score was closer to the global consensus (the mean of all scores) than any individual model. As expected from robust statistics theory, the median effectively dampens the impact of extreme outputs or biases from a single model. Table 3 shows that the ensemble's CV is consistently low, outperforming the stability of individual models.

Alignment with High-Performing Models. Figure 4 shows that the median ensemble exhibits high agreement ($\tau > 0.70$) with the most consistent models (e.g., Grok-3-mini) on the DoR dimension. This indicates that the median not only reduces noise but also captures the core behavior shared by top-performing systems.

The findings confirm **H3**. The median ensemble functions as a reliable and interpretable baseline. We recommend that future benchmarks adopt and publish the median ensemble profile, allowing new models to be evaluated not only in absolute terms but also in relation to their alignment with the state-of-the-art behavioral consensus.

5.4 Separability of Reasoning and Originality Traits

Hypothesis 4 predicts that DoR and ORI are partially distinct traits, justifying a multidimensional evaluation. The correlation analysis validates a certain degree of agreement between the dimensions.

Low Inter-Trait Association. Figure 4 is the main evidence. Intuitively, if DoR and ORI measured the same underlying construct, their cross-trait agreement would be high. Instead, while intra-DoR agreement is strong ($\tau \gtrsim 0.60$) and intra-ORI is weak to moderate, the cross-correlation between DoR and ORI is systematically low, rarely exceeding

$\tau = 0.30$. This satisfies the criterion of discriminant validity Reckase [2009], confirming that the two traits measure non-redundant facets of model behavior.

Implications for Evaluation. This result has a direct methodological implication: DoR and ORI should not be collapsed into a single score. Doing so would hide meaningful trade-offs.

A model may be a strong logical reasoner yet not very original or it may generate creative phrasing while remaining shallow in reasoning. Only a multidimensional evaluation can capture these differences.

The evidence therefore corroborates **H4**, legitimizing the multi-trait rubric approach as a necessary advance over single scalar metrics.

5.5 The Prevalence of Compressed Reasoning

Hypothesis 5 investigates compressed reasoning: answers that are correct but insufficiently justified. This matters because classical accuracy metrics cannot distinguish principled reasoning from shallow heuristics.

In our setting, compressed reasoning denotes correct answers accompanied by shallow or fragmented justifications. This phenomenon is orthogonal to chain-length optimization and motivates explicitly separating correctness from reasoning quality. Related work also highlights that sampling diverse chains can improve robustness Wang *et al*. [2022] and that efficiency-oriented evaluation can bias protocol design Polo *et al*. [2024].

Incidence and Definition. Operationally defining compressed reasoning as a correct answer (`hit = 1`) with a DoR score < 4, we find that **17.8%** of all correct responses fall into this category. In other words, nearly one in six correct answers is accompanied by a superficial or fragmented justification.

Occurrence Patterns. Figure 7 and Table 8 show a clear pattern: compact models such as 4o-mini and 4.1-nano are primarily responsible for these low-DoR hits, especially under Naive prompts. This suggests that, in the absence of an explicit instruction to reason (as in CoT), these models resort to heuristic shortcuts or pattern recognition to produce the correct answer, a behavior aligned with the shortcut learning literature, Geirhos *et al*. [2020].

A concrete instance comes from an ARC physics question in which a student places electrodes into a beaker of solution. Under the Naive prompt, GPT-4.1-nano selects the correct alternative but provides only a very brief justification: the chain-of-thought field is empty and the explanation essentially states that the setup "shows conduction in the solution", without discussing ionic movement, alternative interpretations or why the distractors are wrong.

In our rubric, judges disagree on how much reasoning is actually present, assigning DoR scores between 2 and 7 (mean ≈ 4.7) and Originality scores around 3.8. This kind of low-DoR, low-ORI success case illustrates compressed reasoning in practice: a correct answer achieved via a shallow heuristic rather than an explicit, well-articulated inferential process.

Implications for Interpretability. The results confirm **H5** and offer a serious warning long recognized among humans and now identified in LLMs: a correct answer is not synonymous with good reasoning. The prevalence of compressed reasoning implies that evaluation pipelines must go beyond accuracy and penalize correct answers that are insufficiently justified. This is essential to encourage the development of models that not only get answers right but do so for the right reasons and in a transparent manner.

5.6 Synthesis and Implications

The multidimensional analysis of TraCE-LLM paints a picture of LLM behavior that is much sharper than classical metrics allow, revealing a landscape of trade-offs among stability, robustness and DoR.

Main Findings. **Asymmetric Stability:** DoR is a stable and convergent trait, whereas ORI is volatile and dependent on model stochasticity. **ACE:** Compact models can mask their fragility under stress with accuracy gains, at the cost of severe behavioral instability. **Compressed Reasoning:** Nearly 17% of correct answers are accompanied by superficial justifications, evidencing the use of heuristic shortcuts. **Rubric Validity:** The DoR and ORI dimensions are statistically partial separable and the median baseline proved to be a robust and reliable estimator.

Recommendations for Evaluation and Development. The results support concrete recommendations. For evaluation, it is advisable to adopt multidimensional rubrics, stratify results by prompt type and use robust baselines such as the median ensemble. For development, the data suggest that training should focus not only on maximizing accuracy but also on reinforcing reasoning stability (to mitigate ACE) and penalizing compressed reasoning, encouraging the generation of faithful and complete justifications.

Broader Impact. TraCE-LLM offers a blueprint for a new generation of evaluations: more granular, interpretable and aligned with the demands of transparency and reliability in high-stakes AI systems. By shifting the focus from mere correctness to behavioral characterization, we pave the way for deeper understanding and more effective auditing of LLMs.

5.7 Human-LLM Alignment and Potential Judge Bias

This subsection reports the human anchor validation results on the stratified subset ($n = 135$; Section 4.6). The goal is interpretive clarity: we explicitly separate *within-panel reliability* from *between-panel alignment*. Within-panel reliability answers whether a panel (humans-only or LLM-judges-only) applies the semantic rubric consistently. Alignment answers a different question: even if a panel is internally consistent, does it *track* human judgments on the same items under the

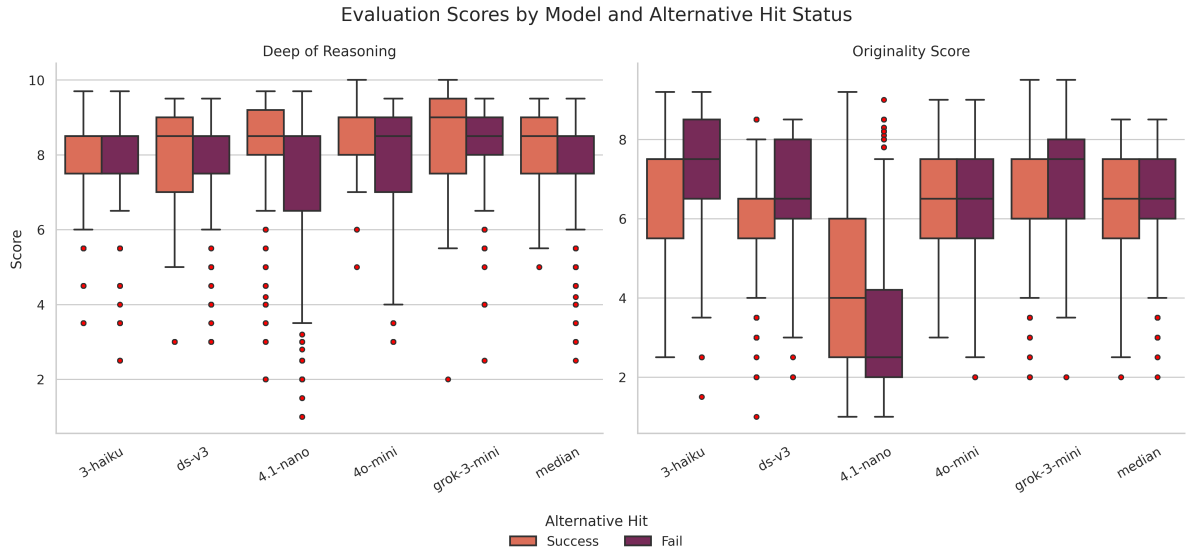

Figure 7. Distribution of DoR and ORI scores by model, stratified by correctness. Success cases (orange) refer to responses that match the reference. Failures (purple) denote incorrect responses.

Table 8. Descriptive statistics for evaluation scores by criterion, correctness (hit) and model. CV = Coefficient of Variation.

Criterion	Correct	Model	Mean	SD	Median	Min	Max	CV
DoR	No	3.5-haiku	7.7936	1.5052	8.5	2.5	9.7	0.1931
		4.1-nano	7.2010	2.2742	8.5	1.0	9.7	0.3158
		4o-mini	7.5593	1.6330	8.5	3.0	9.5	0.2160
		ds-v3	7.6324	1.6119	8.5	3.0	9.5	0.2112
		grok-3-mini	8.1907	1.2331	8.5	2.5	9.5	0.1505
		median	7.7698	1.4675	8.5	2.5	9.5	0.1889
	Yes	3.5-haiku	7.9927	1.1148	8.5	3.5	9.7	0.1395
		4.1-nano	8.5499	0.8306	8.5	2.0	9.7	0.0972
		4o-mini	8.5845	0.5573	9.0	5.0	10.0	0.0649
		ds-v3	8.0073	1.2334	8.5	3.0	9.5	0.1540
		grok-3-mini	8.2788	1.2615	9.0	2.0	10.0	0.1524
		median	8.2669	0.9390	8.5	5.0	9.5	0.1136
ORI	No	3.5-haiku	7.1009	1.3766	7.5	1.5	9.2	0.1939
		4.1-nano	3.4405	1.8403	2.5	1.0	9.0	0.5349
		4o-mini	6.4410	1.2546	6.5	2.0	9.0	0.1948
		ds-v3	6.6705	1.6038	6.5	2.0	8.5	0.2404
		grok-3-mini	7.0782	1.2026	7.5	2.0	9.5	0.1699
		median	6.6597	1.2872	6.5	2.0	8.5	0.1933
	Yes	3.5-haiku	6.8108	1.2989	7.5	2.5	9.2	0.1907
		4.1-nano	4.1974	2.2147	4.0	1.0	9.2	0.5276
		4o-mini	6.4851	1.1066	6.5	3.0	9.0	0.1706
		ds-v3	6.0247	1.3743	6.5	1.0	8.5	0.2281
		grok-3-mini	6.5228	1.1193	6.0	2.0	9.5	0.1716
		median	6.2975	1.0796	6.5	2.0	8.5	0.1714

same rubric? Keeping these notions distinct is essential, because internal consistency does not imply human-likeness.

Throughout we compare robust aggregates at the response level: $\tilde{H}_i$ is the median of the human ratings for response i, and $\tilde{J}_i$ is the median of the LLM-judge ratings (both computed over non-missing values). We use tie-aware rank agreement (τ_b) as the primary alignment statistic and triangulate with ordinal agreement (κ_w) and categorical association (V), since aggregation by medians increases the prevalence of ties.

Human anchor and primary six-bin rubric. To prevent the common misreading that "consistent judges" implies "human-like judges", we summarize within-panel reliability and human-LLM alignment side by side under the primary six-bin rubric.

Although both panels are internally consistent, their scores do not track each other on the same items. DoR exhibits moderate within-panel reproducibility (humans: $\alpha_{\text{raw}} = 0.571$; LLM judges: $\alpha_{\text{raw}} = 0.694$), whereas ORI shows a lower reliability ceiling in both panels (humans: $\alpha_{\text{raw}} = 0.293$; LLM judges: $\alpha_{\text{raw}} = 0.163$). However, between-panel alignment remains near zero under rank and ordinal-agreement views (DoR: $\tau_b = -0.016$, $\kappa_w = 0.008$; ORI: $\tau_b = -0.011$, $\kappa_w = -0.002$), and categorical association is not supported (DoR: $V = 0.192$, $p = 0.468$; ORI: $V = 0.173$, $p = 0.665$; Table 9). This separates reliability from alignment: LLM judges may be consistent under the rubric, yet their item-level preferences differ from humans, limiting their use as a surrogate for human evaluation in this subset.

A complementary diagnostic reinforces this separation. When humans and LLM judges are pooled, agreement collapses ($\alpha_{\text{all}} = 0.171$ for DoR; 0.091 for ORI), and agreement between the two group aggregates is negative ($\alpha(\tilde{H}, \tilde{J}) = -0.073$ for DoR; -0.020 for ORI). This pattern is consistent with calibration heterogeneity between panels (different scale usage under the same rubric), rather than a simple lack of within-panel consistency.

Table 9. Human anchor subset within-panel reliability.

Metric	α (raw)		Within-panel (6 bins)				Pooled and aggregates (6 bins)		Human-LLM alignment				
	Humans	LLMs	α (Humans)	α (LLMs)	τ_b med (Humans)	IQR	α (All)	α ($\tilde{H}$ vs $\tilde{J}$)	$\tau_b(\tilde{H}, \tilde{J})$	κ_w	V	p	bin match
DoR	0.571	0.694	0.502	0.488	0.564	$[0.527, 0.584]$	0.171	-0.073	-0.016	0.008	0.192	0.468	0.356
ORI	0.293	0.163	0.268	0.165	0.416	$[0.361, 0.434]$	0.091	-0.020	-0.011	-0.002	0.173	0.665	0.304

Within-panel reliability (the empirical ceiling). Within-panel reliability comes first because cross-panel alignment cannot exceed how reproducible each panel is under the rubric. Table 9 reports Krippendorff's α for the raw $[0, 10]$ scale and for the primary six-bin discretization. Under six bins, DoR shows moderate reproducibility in both panels (humans: $\alpha_{6b} = 0.502$; LLM judges: $\alpha_{6b} = 0.488$), which suggests that "depth of reasoning" admits a relatively shared interpretation once the rubric is fixed. ORI is less reproducible (humans: $\alpha_{6b} = 0.268$; LLM judges: $\alpha_{6b} = 0.165$), so any claim of close human replication by automated judges faces a lower ceiling even before testing alignment. Tie-aware rank agreement mirrors this contrast: within humans, τ_b has median 0.564 for DoR and 0.416 for ORI; within LLM judges, τ_b has median 0.664 for DoR and 0.317 for ORI.

Between-panel alignment (do automated judges track humans?). On the human anchor subset, automated judges do not track human scoring on the same responses. The primary rank based statistic, Kendall's $\tau_b(\tilde{H}, \tilde{J})$, is near zero for both traits (DoR: $\tau_b = -0.016$; ORI: $\tau_b = -0.011$; $n = 135$), and this sits well below the detectability context in the human anchor protocol (Section 4.6), where the aggregate minimal detectable effect is approximately 0.16 for Kendall τ.

The same pattern appears in the 15 human$\times$LLM judge pairs, with correlations centered near zero (DoR: median $\tau_b = -0.011$; ORI: median $\tau_b = 0.030$. Two complementary checks under the primary six-bin rubric agree with this conclusion: weighted Cohen's κ remains near zero (DoR: $\kappa_w = 0.008$; ORI: $\kappa_w = -0.002$) and categorical association is not supported (DoR: $V = 0.192$, $p = 0.468$; ORI: $V = 0.173$, $p = 0.665$; Table 9).

Exact bin matches occur in 35.6% of items for DoR and 30.4% for ORI, and both traits match simultaneously in only 14.8% of items. In practical terms, these results separate reliability from alignment: LLM judges may be internally reproducible under the rubric, but their response level preferences do not reproduce the human panel on this subset.

Sensitivity to discretization (pre-specified). As pre-specified robustness checks, we repeat the alignment analysis under coarser discretizations. With three bins (low/medium/high), DoR increases slightly but remains weak ($\kappa_w = 0.072$, $V = 0.176$, $p = 0.080$), while ORI stays weak ($\kappa_w = 0.033$, $V = 0.135$, $p = 0.293$). The two-bin split is used only as a threshold diagnostic and remains effectively null (DoR: $V = 0.015$, $p = 0.858$; ORI: $V \approx 0.000$, $p = 1.000$). In practical terms, coarsening the rubric does not reveal a hidden moderate alignment, so these checks support the primary six-bin conclusion rather than replacing it.

Severity diagnostic (mechanism, non-causal). A simple ordinal severity model, fitted separately for humans and for LLM judges, indicates heterogeneous scale usage within both panels. For example, DoR severity spans approximately $[-0.95, 4.12]$ across humans and $[-2.06, 0.89]$ across LLM judges, and ORI shows a comparable heterogeneity. Read as a diagnostic, this pattern matches the pooled-agreement collapse and the near-zero alignment estimates, and is consistent with systematic calibration differences under the same rubric. Severity is reported to help explain how misalignment can arise, not to assert a causal mechanism.

6 Discussion

The results presented in the previous section not only validate the formulated hypotheses but also offer a new perspective on LLM evaluation, directly engaging with advances and gaps in the recent literature. In this section, we discuss the broader implications of our findings, focusing on convergences and divergences with the state of the art and on the methodological advance afforded by TraCE-LLM.

6.1 The Dual Nature of Variability: Contained Reasoning vs. Volatile Creativity

Our findings on asymmetric stability (H1) and trait separability (H4) converge with the view that LLMs possess distinct behavioral facets. High agreement in DoR reinforces the idea that factual knowledge and logical structures are "crystallized" during pretraining, becoming a stable property of the model. By contrast, the high variability of ORI provides direct empirical evidence that stochastic decoding mechanisms Holtzman *et al*. [2020] make a model's stylistic "personality" far less stable.

While prior work already measured uncertainty across evaluation dimensions Li *et al*. [2025b], our contribution is to disentangle uncertainty of *reasoning* from uncertainty of *expression*, showing that they are not the same construct and should be assessed independently.

6.2 Beyond Accuracy: Diagnosing Hidden Failure Modes

The main contribution of this study lies in identifying failure modes that are invisible to classical metrics.

The ACE. The literature has noted the fragility of smaller models under stress Wei *et al*. [2022]. The ACE phenomenon (H2) represents a sharper and more deceptive pattern: a model can look better on a superficial metric while becoming behaviorally less stable.

Concretely, we show that reduced robustness can be paradoxically masked by an increase in accuracy. This challenges confidence in leaderboards that do not control for behavioral volatility and suggests that some reported performance gains

may be artifacts of erratic behavior. TraCE-LLM offers a method to detect and quantify this effect.

The Silent Epidemic of Compressed Reasoning. The high incidence of correct answers with low DoR (H5) provides a strong empirical counterpoint to blind reliance on Chain-of-Thought. It validates concerns about CoT infidelity Turpin *et al*. [2023]; Lanham *et al*. [2023] and shows that shortcut learning Geirhos *et al*. [2020] is not an anomaly but a common strategy, especially for compact models.

As a result, evaluations that consider only final-answer correctness can overestimate reasoning quality. This practice remains common in benchmarks such as MMLU and ARC-Challenge, limiting their interpretability under stochastic generation.

Although our experiments focus on MCQ benchmarks, analogous blind spots appear in cross-lingual evaluation. Back-translation analyses show that surface-level adequacy can coexist with loss of poetic intent and terminology drift, reinforcing the need for diagnostic dimensions beyond final-answer correctness, Weigang and Brom [2025b,a].

Human-LLM Alignment and Potential Judge Bias. The human anchor results clarify the epistemic status of model-judge scores in this study. The automated judge panel is internally reproducible (especially for DoR, $\alpha = 0.694$), yet its item-level scores do not track human judgments on the same items under the same rubric ($\tau_b \approx 0$; Section 5.7).

This separation between reliability and alignment has a direct methodological implication: model-judge scores should be interpreted exclusively as *internal comparative signals*, useful for detecting trends across models, prompt styles and replications and **not** as direct proxies for human evaluation. When absolute score calibration is required, the protocol must include a human anchor panel.

6.3 The Methodological Advance of TraCE-LLM

TraCE-LLM advances over prior frameworks such as HELM Liang *et al*. [2023] or MT-Bench Bai *et al*. [2024] by integrating three elements synergistically:

1. **Diagnosis over Performance:** Rather than merely ranking models, our protocol produces a behavioral profile that explains why they fail or succeed.
2. **Quantification of Instability:** The introduction of the Instability Index and the analysis of CV provide concrete metrics for replicability and reliability, a gap in most current benchmarks.
3. **Robust Aggregation:** Using the median ensemble (H3) offers a more reliable baseline than any single gold standard or reference model, addressing the problem of single-judge bias.
4. **Epistemic Calibration via Human Anchors:** The human anchor subset (Section 4.6) separates within-panel reliability from between-panel alignment, clarifying when model-judge rubric scores can be interpreted as internal comparative signals and when human calibration is required.

7 Conclusion

This work introduced TraCE-LLM, a rubric-guided evaluation protocol that moves beyond scalar accuracy metrics. By explicitly modeling the stochastic nature of LLMs through a hierarchical experimental design, the protocol enabled the quantification of latent behavioral traits such as Depth of Reasoning and Originality. The main empirical contribution was the identification and formalization of the Adversarial Compensation Effect, a paradoxical shift in superficial correctness metrics that can mask profound behavioral instability in compact models.

The results also show that reasoning is a relatively stable trait, whereas originality is highly volatile and that a significant fraction of correct answers can stem from compressed reasoning, reinforcing the need for an evaluation that disentangles answer correctness from justification quality. Finally, the human anchor analysis clarifies the epistemic status of model-judge scores: internal reproducibility under a rubric does not necessarily imply alignment with human ratings on the same items, motivating calibration with human anchors when human-likeness is the target.

Despite the robustness of the design, we acknowledge the boundaries of the present study, which point to future research directions. The main limitation is scope: the study focused on five models and three multiple-choice benchmarks. Generalizing the findings, especially ACE, to a broader range of architectures (e.g., open-source models with different alignment methods) and tasks (e.g., summarization, open-ended dialogue) is an essential step. Second, evaluation via model-judges, while scalable, inherits the inherent biases of these models. Finally, the current two-trait rubric is a proof of concept; other dimensions such as factuality, bias and safety are fundamental for a complete audit.

The present study instantiates TraCE-LLM on multiple-choice benchmarks, where correctness has an unambiguous ground truth. Extending the protocol to long-form or open-ended generation tasks is a natural next step, but it will require adapting the correctness criterion, likely replacing hit/miss labels with graded adequacy judgments, and recalibrating the DoR and ORI rubric anchors to accommodate longer, more diverse outputs. We anticipate that the core diagnostic logic (trait profiling, stability indices, ACE detection) will transfer, whereas the specific thresholds and rubric intervals will need domain-specific tuning.

To overcome these limitations and expand the scope of the protocol, we outline three promising directions for future work: (i) Incorporate Item Response Theory (IRT) models to more richly model prompt difficulty and model ability, enabling deeper factor analysis of latent traits. (ii) Mitigate model-judge bias via a hybrid system that uses a small set of anchor items annotated by human experts to calibrate and weight automated judgments, increasing process reliability. (iii) Extend the TraCE-LLM framework to evaluate multimodal models that process text, image and audio inputs. This will require developing new rubrics and methods to decompose variance arising from each modality.

In summary, TraCE-LLM establishes a new lexicon and a practical toolkit for diagnosing LLMs in realistic settings. The discovery of ACE, in particular, serves as a warning that

the pursuit of higher leaderboard scores may inadvertently optimize for unstable behavior. By connecting superficial correctness to the coherence of latent traits, this work emphasizes the pressing need for evaluations that are multidimensional, sensitive to prompt context and psychometrically informed. The continued advancement of this framework aims to make LLM evaluation more reliable, interpretable and ultimately more aligned with the demands of critical real-world applications.

Declarations

Authors' Contributions

All authors contributed equally to all stages of this study. Pedro Carvalho Brom, Vinícius Di Oliveira and Li Weigang jointly and equally contributed to Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing, original draft, Writing, review & editing, Visualization, Supervision and Project administration. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

ChatGPT was used to troubleshoot and correct table formatting in LaTeXand for grammatical evaluation to standardize the English writing across all sections of this work. The authors verified and approved all changes and remain responsible for the content.

Funding

This research is partially funded by the Brazilian National Council for Scientific and Technological Development (CNPq) under grant number 309545/2021-8.

Availability of data and materials

Code, prompts and data are available at `https://github.com/pcbrom/trace-llm`, Brom *et al*. [2025]. The repository is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). It provides (i) aggregated tables derived from the full dataset (N=10,125) and the complete human evaluation scores used in the manuscript and (ii) a public evaluation sample (N=500) with precomputed model outputs and judge scores for item-level inspection.

References

AI, D. (2024). Deepseek-v3: A mixture-of-experts language model. *arXiv preprint*, arXiv:2412.19437. Available at:`https://arxiv.org/abs/2412.19437`. Technical report describing the DeepSeek-V3 architecture and training.

Anthropic (2024). Claude 3 and 3.5 model card. Available at:`https://www.anthropic.com/claude-3-model-card`. Includes information on Claude 3.5 Haiku and related models.

Bai, G., Liu, J., Bu, X., *et al*. (2024). MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.401.

Biderman, S., Schoelkopf, H., Sutawika, L., *et al*. (2024). Lessons from the trenches on reproducible evaluation of language models. DOI: 10.48550/arxiv.2405.14782.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. Textbook for graduates; includes bibliographical references (pages 711–728) and index. DOI: 10.1007/978-0-387-45528-0.

Bonett, D. G. and Wright, T. A. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28. DOI: 10.1007/BF02294183.

Brom, P. C., Di Oliveira, V., and Weigang, L. (2025). TraCE-LLM: Evaluation datasets and pipeline (v2.3). DOI: 10.5281/zenodo.18549677.

Brown, T., Mann, B., Ryder, N., and Subbiah (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.. DOI: 10.48550/arxiv.2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., *et al*. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*, abs/2303.12712. DOI: 10.48550/arXiv.2303.12712.

Chiang, W.-L., Zheng, L., Sheng, Y., *et al*. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv*, abs/2403.04132. DOI: 10.48550/arXiv.2403.04132.

Clark, P., Cowhey, I., Etzioni, O., *et al*. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. DOI: 10.48550/arxiv.1803.05457.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/cbo9780511802843.

DeepSeek (2025). Deepseek api. Available at:`https://api-docs.deepseek.com`. API documentation for DeepSeek models, including context and limits.

Di Oliveira, V., Brom, P. C., and Li, W. (2025). Two-step RAG for metadata filtering and statistical LLM evaluation. *IEEE Latin America Transactions*, 23(12):1201–1210. DOI: 10.1109/TLA.2025.11231222.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. DOI: 10.48550/arxiv.1702.08608.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York. DOI:

10.1007/978-1-4899-4541-9.

Freitag, M., Grangier, D., and Caswell, I. (2020). BLEU might be guilty but references are not innocent. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.5.

Ganguli, D., Lovitt, L., Kernion, J., *et al*. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. DOI: 10.48550/arxiv.2209.07858.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., *et al*. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673. DOI: 10.1038/s42256-020-00257-z.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. DOI: 10.1038/nature14539.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393. DOI: 10.2307/2285666.

Hendrycks, D., Burns, C., Basart, S., Zou, A., *et al*. (2021). Measuring massive multitask language understanding. Available at:`https://arxiv.org/abs/2009.03300`.

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1):115–135. DOI: 10.1007/s11747-014-0403-8.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. DOI: 10.48550/arxiv.1904.09751.

Hu, Z., Zhang, J., Xiong, Z., *et al*. (2025). Language model preference evaluation with multiple weak evaluators. Available at:`https://arxiv.org/abs/2410.12869`.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151. DOI: 10.1177/001316446002000116.

Kang, J., Son, D., Song, H., and Chang, B. (2024). In-context learning with noisy labels. Available at:`https://arxiv.org/abs/2411.19581`.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., *et al*. (2020). Scaling laws for neural language models. DOI: 10.48550/arxiv.2001.08361.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93. DOI: 10.1093/biomet/30.1-2.81.

KENDALL, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251. DOI: 10.1093/biomet/33.3.239.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., *et al*. (2023). Measuring faithfulness in chain-of-thought reasoning. DOI: 10.48550/arxiv.2307.13702.

Lee, S., Cai, Y., Meng, D., *et al*. (2024). Unleashing large language models' proficiency in zero-shot essay scoring. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-emnlp.10.

Li, R., Zhu, C., Xu, B., Wang, X., and Mao, Z. (2025a). Automated creativity evaluation for large language models: A reference-based approach. DOI: 10.18653/v1/2025.findings-emnlp.1171.

Li, Y., Wang, H., Zhang, Q., *et al*. (2025b). Unieval: Unified holistic evaluation for unified multimodal understanding and generation. DOI: 10.48550/arxiv.2505.10483.

Liang, P., Bommasani, R., Lee, T., and Tsipras, D. (2023). Holistic evaluation of language models. DOI: 10.1111/nyas.15007.

Madaan, L., Singh, A. K., Schaeffer, R., *et al*. (2024). Quantifying variance in evaluation benchmarks. DOI: 10.48550/arxiv.2406.10229.

Mondorf, P. and Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. *arXiv preprint arXiv:2404.01869*. DOI: 10.48550/arxiv.2404.01869.

Nalbandyan, G., Shahbazyan, R., and Bakhturina, E. (2025). Score: Systematic consistency and robustness evaluation for large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Industry Track*, pages 470–484. Association for Computational Linguistics. Available as PDF at ACL Anthology. DOI: 10.18653/v1/2025.naacl-industry.39.

OpenAI (2024). Gpt-4o mini: advancing cost-efficient intelligence. Available at:`https://platform.openai.com/docs/models/gpt-4o-mini`. Product announcement and model description.

OpenAI (2025). Evals: A framework for evaluating large language models and an open-source registry of benchmarks. Available at:`https://github.com/openai/evals`. Accessed August 9, 2025.

OpenAI (2025). Gpt-4.1 nano. Available at:`https://platform.openai.com/docs/models/gpt-4.1-nano`. Model card and API documentation.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1967). *The Measurement of Meaning*. University of Illinois Press. Paperback publication date: 01 January 1967. DOI: 10.2307/3709408.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). Llm evaluators recognize and favor their own generations. DOI: 10.52202/079017-2197.

Pezeshkpour, P. and Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-naacl.130.

Polo, F. M., Xu, R., Weber, L., *et al*. (2024). Efficient multi-prompt evaluation of llms. DOI: 10.52202/079017-0707.

Raj, H., Gupta, V., Rosati, D., and Majumdar, S. (2025). Semantic consistency for assuring reliability of large language models. DOI: 10.48550/arxiv.2308.09138.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer,

New York, NY, 1 edition. DOI: 10.1007/978-0-387-89976-3.

Shen, X., Chen, Z., Backes, M., Shen, Y., Zhang, Y., and verazuo (repository maintainer) (2024). In-the-wild jailbreak prompts on llms. Available at:`https://github.com/verazuo/jailbreak_llms`.

Shi, C., Yang, H., Cai, D., *et al*. (2024). A thorough examination of decoding methods in the era of LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.489.

Song, Y., Wang, G., Li, S., Lin, B. Y., *et al*. (2025). The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 4195–4206. Association for Computational Linguistics. Available as PDF at ACL Anthology. DOI: 10.18653/v1/2025.naacl-long.211.

Srivastava, A., Rastogi, A., Rao, A., *et al*. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*. DOI: 10.48550/arxiv.2206.04615.

Thelwall, M. (2025). Chatgpt for complex text evaluation tasks. *Journal of the Association for Information Science and Technology*, 76(4):645–648. DOI: 10.1002/asi.24966.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965. DOI: 10.52202/075280-3275.

Wang, B., Wei, C., Liu, Z., *et al*. (2024a). Resilience of large language models for noisy instructions. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11939–11950, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-emnlp.697.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*. DOI: 10.48550/arxiv.2203.11171.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., *et al*. (2024b). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290. DOI: 10.52202/079017-3018.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., *et al*. (2022). Emergent abilities of large language models. DOI: 10.48550/arxiv.2206.07682.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., *et al*. (2023). Chain-of-thought prompting elicits reasoning in large language models. DOI: 10.52202/068431-1800.

Weigang, L. and Brom, P. C. (2025a). Llm-bt-terms: Back-translation as a framework for terminology standardization and dynamic semantic embedding. DOI: 10.48550/arxiv.2506.08174.

Weigang, L. and Brom, P. C. (2025b). Paradox of poetic intent in back-translation: evaluating the quality of large language models in chinese translation. *Frontiers of Information Technology Electronic Engineering*, 26(11):2176. DOI: 10.1631/FITEE.2500298.

xAI (2025). Grok 3 and grok 3 mini beta. Available at:`https://x.ai/news/grok-3/`. Announcement and high-level description of Grok 3 mini (beta).

Yu, Z., Gao, C., Yao, W., *et al*. (2024). FreeEval: A modular framework for trustworthy and efficient evaluation of large language models. In Hernandez Farias, D. I., Hope, T., and Li, M., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–13, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-demo.1.

Zellers, R., Holtzman, A., Bisk, Y., *et al*. (2019). HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1472.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. DOI: 10.52202/075280-2020.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*. DOI: 10.48550/arxiv.2211.01910.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. Available at:`https://arxiv.org/abs/2307.15043`.

Appendix: Mathematical Formalization of the Study

Coefficient of Variation. Is defined as $\mathrm{CV} = \frac{\hat{\sigma}}{\hat{\mu}}$, where $\hat{\sigma}$ is the estimated standard deviation and $\hat{\mu}$ is the estimated mean.

Notation and design. Let models $m \in \mathcal{M}$, datasets $d \in \mathcal{D}$, items $i \in \mathcal{I}(d)$, prompt styles $p \in \mathcal{P} = \{\text{naive}, \text{cot}, \text{adv}\}$ and replicates $r \in \{1, \ldots, R\}$. The two rubric criteria considered are Depth of Reasoning (DoR) and Originality (ORI). The standardized output for instance (m, d, i, p, r) is $Y_{m,d,i,p,r}$. For each rubric criterion c, define the scoring map

$$R_c : \; Y^{(c)} \to [0, 10],$$

so that $R_c\big(Y^{(c)}_{m,d,i,p,r}\big)$ is the numerical score assigned to the response. When applicable, use a hit status partition $\{\text{hit}, \text{miss}\}$.

Summary metrics and robustness. For any fixed (m, p, c) consider the mean μ, median $\widetilde{x}$, standard deviation σ, coefficient of variation $\mathrm{CV} = \sigma/\mu$ and outlier detection via IQR. Point and interval estimates are obtained by nonparametric bootstrap with B resamples and percentile 95% CIs Efron and Tibshirani [1993]; Davison and Hinkley [1997].

Agreement via Kendall's τ. Use Kendall's τ to quantify rank agreement with natural handling of ties. Compute τ matrices stratified by prompt.

Latent projection and loadings. For each (m, p) build the trait vector

$$\mathbf{z}_{m,p} = \left[\overline{\mathrm{WF1}},\ \overline{\mathrm{DoR}},\ \mathrm{Var(DoR)},\ \overline{\mathrm{ORI}},\ \mathrm{Var(ORI)}\right]^\top,$$

then apply PCA on the correlation matrix. Retain components by the Kaiser rule ($\lambda > 1$) Kaiser [1960]. Let W be the eigenvector matrix and $\mathbf{x}_{m,p} = W^\top(\mathbf{z}_{m,p} - \bar{\mathbf{z}})$ the projection onto the first two PCs.

Instability of behavioral traits. For each model m, define $f_m : \mathcal{P} \to \mathbb{R}^2$ by $f_m(p) = \mathbf{x}_{m,p}$. The Instability Index is

$$\mathrm{instab}(f) = \frac{1}{\binom{|P|}{2}} \sum_{i<j} \left\| f(p_i) - f(p_j) \right\|_2.$$

Consensus baseline. For each instance (d, i, p, c), define the ensemble median baseline

$$B_{d,i,p,c} = \mathrm{median}_{m \in \mathcal{M}}\Big(R_c\big(Y^{(c)}_{m,d,i,p,\cdot}\big)\Big),$$

and the distance to consensus $\Delta_{m,d,i,p,c} = R_c\big(Y^{(c)}_{m,d,i,p,\cdot}\big) - B_{d,i,p,c}$. Trait-correctness association can be summarized by Kendall's τ between the per-instance trait scores $\{R_c\}$ and and the hit/miss indicator.

Discriminant validity (DoR vs ORI). For the constructs DoR and ORI, inspect discriminant validity via HTMT, $\mathrm{HTMT(DoR, ORI)}$ Henseler *et al*. [2015]. Values well above 0.90 suggest semantic overlap and motivate oblique interpretation in the latent space.

Detectability (MDE) for Kendall's τ. For planning purposes, under the null approximation and in the absence of ties, Kendall's $\hat{\tau}$ is asymptotically normal. When ties are present, we compute τ_b and use bootstrap to obtain standard errors and confidence intervals. The minimum detectable effect (MDE) for a two-sided test at significance level α and target power $1 - \beta$ is approximated by $\mathrm{MDE} \approx \left(z_{1-\alpha/2} + z_{1-\beta}\right)\ SE(\hat{\tau})$, KENDALL [1938, 1945]; Bonett and Wright [2000].

Hypotheses and operational tests

H1: Within-criterion agreement. Within-DoR pairs show higher τ than DoR–ORI pairs. *Test*: compare bootstrap distributions of τ across these two groups with resampling by instances.
H2: Prompt effect. $\overline{\mathrm{WF1}}$, $\overline{\mathrm{DoR}}$, $\overline{\mathrm{ORI}}$ and Instab differ across $p \in \mathcal{P}$. *Test*: paired contrasts by model with bootstrap CIs for mean differences.
H3: Latent structure. PC1 aligns with DoR variation and PC2 captures ORI. *Test*: inspect loadings with $\lambda > 1$ retention Kaiser [1960] and bootstrap CIs for loadings.
H4: Discriminant validity. $\mathrm{HTMT(DoR, ORI)}$ should remain below usual thresholds, otherwise overlap is indicated Henseler *et al*. [2015].
H5: Adversarial Compensation Effect (ACE). Under adversarial prompting, the following occur simultaneously: $\Delta\mathrm{WF1}_m = \mathrm{WF1}_{m,\mathrm{adv}} - \max\{\mathrm{WF1}_{m,\mathrm{naive}}, \mathrm{WF1}_{m,\mathrm{cot}}\} > 0$, (ii) at least one instability proxy increases, operationalized by $\Delta\mathrm{CV}^{\mathrm{DoR}}_m > 0$ or $\Delta\mathrm{CV}^{\mathrm{ORI}}_m > 0$, and $\tau_{m,\mathrm{adv}} < 0$. *Joint decision*: intersection rule with multiple testing control via multivariate bootstrap Efron and Tibshirani [1993]; Davison and Hinkley [1997].

Auxiliary results

Lemma 1: Sign stability of Δ under bootstrap. If resamples preserve stratification by (m, p), the probability of sign flips in $\Delta\mathrm{WF1}_m$, $\Delta\mathrm{Instab}(m)$ and $\Delta\tau_m$ decreases with B and with the effective per-stratum sample size. *Sketch*: weak law of large numbers for conditional resampling statistics Efron and Tibshirani [1993]; Davison and Hinkley [1997].
Lemma 2: Median as L_1 minimizer. For each (d, i, p, c), $B_{d,i,p,c}$ minimizes $\mathbb{E}|X - t|$ over t, with X distributed across models. This is the characteristic property of the median.
Proposition 1: Approximate separability in PCA. If $\mathrm{Var(DoR)}$ dominates the total variance and $\mathrm{Var(ORI)}$ is non-negligible yet smaller, the first two PCs retain most variance with loadings primarily aligned with DoR (PC1) and ORI (PC2). *Justification*: spectral diagonalization of a weakly block-structured correlation matrix and the $\lambda > 1$ selection rule Kaiser [1960].

Practical notes

(i) Always stratify bootstrap by (m, p) to preserve dependencies. (ii) Report PCA loadings with CIs. (iii) When $\mathrm{HTMT} > 0.90$ interpret PC2 as a residual originality axis rather than a fully independent construct Henseler *et al*. [2015]. (iv) Prefer the ensemble median as consensus anchor for robustness to extremes.