GUARD: A Prompt-Level Discipline for Correcting Categorical Gist Substitution in Large Language Model Characterizations of Institutions

VALERIE RIGGS, Morgan State University, USA
EMMANUEL MASA-IBI, Morgan State University, USA
DANIEL ONYEJIEKWE, Morgan State University, USA
YOSMEHILY SEGURA-SILVERIO, Morgan State University, USA
ATO BHATTA, Morgan State University, USA

Large language models (LLMs) frequently characterize institutions by activating a familiar category prior and describing entities primarily through that lens, even when the category framing understates or contradicts the entity's specific verifiable record. We call this failure mode *categorical gist substitution*: the model retrieves the categorical summary in place of the entity-anchored one. We introduce GUARD (Ground, Unbundle, Anchor, Resolve, Distill), a prompt-level intervention grounded in fuzzy-trace theory that requires the model to perform five ordered moves before generating its final characterization. We evaluate GUARD on Tier 1 of a curated benchmark of 17 institutional-characterization prompts targeting Historically Black Colleges and Universities (HBCUs), using Claude Fable 5 as the target model. Three graders independently produced 203 ratings across 68 responses spanning four configurations (Baseline, Control, GUARD-inline, GUARD-Controlled). GUARD-Controlled, which pairs the discipline with a deterministic checker that validates structural adherence before allowing the model to compose the final narrative, produces the largest and most consistent effects: a 2.03× increase in mean gist displacement over Baseline (0.79 → 2.39 on a 0–3 scale; Cohen's $d = 1.70$, $p < 0.0001$), a reduction of failure-signature prevalence from 51.0% to 0.0% ($\chi^2 = 31.5$, $p < 0.0001$), and an increase in explicit bottom-line articulation from 18.4% to 54.9% ($\chi^2 = 12.8$, $p < 0.001$). A generic "be specific" Control condition produces no significant improvement over Baseline, indicating that the effect is attributable to the discipline itself rather than to prompt length or instructional emphasis. We report inter-rater reliability, name the study's methodological limitations honestly (small grader panel, single-tier scope, single-model target), and situate GUARD within the broader landscape of prompt-level and deterministic-validation approaches to LLM behavioral correction.

Additional Key Words and Phrases: Large language models, prompt engineering, fuzzy-trace theory, categorical bias, institutional characterization, deterministic validation, HBCU, inter-rater reliability

1 Introduction

Large language models, when asked to characterize an entity, retrieve a summary representation and render it as prose. Fuzzy-trace theory [6] argues that human cognition maintains two parallel traces of any encoded proposition: a *verbatim* trace (the specific facts, names, numbers, and dates) and a *gist* trace (the categorical or thematic summary the material was about). When the two diverge, human reasoners preferentially recall the gist. We propose that LLM entity characterization exhibits a structurally analogous behavior: when the model encounters a request to describe an institution, it activates a category prior and generates prose organized around the category's canonical gist, even when the specific institutional record diverges materially from that canonical framing.

Authors' Contact Information: Valerie Riggs, Morgan State University, Baltimore, MD, USA, valerie.riggs@morgan.edu; Emmanuel Masa-Ibi, Morgan State University, Baltimore, MD, USA, emmas2@morgan.edu; Daniel Onyejiekwe, Morgan State University, Baltimore, MD, USA, daony2@morgan.edu; Yosmehily Segura-Silverio, Morgan State University, Baltimore, MD, USA, yoseg1@morgan.edu; Ato Bhatta, Morgan State University, Baltimore, MD, USA, atbha1@morgan.edu.

This behavior is not adversarial and not obviously incorrect at the sentence level; each claim the model produces may be locally defensible. The failure is at the level of what the response emphasizes, orders, and concludes. When Morgan State University is characterized primarily through the HBCU category framing, emphasizing access mission, historical legacy, and community service, the response omits or de-emphasizes the institution's R2 Carnegie research classification, its $104.4M in FY25 sponsored research commitments, its named research centers, and its ongoing pursuit of R1 status. The verbatim record supports a research-institution framing; the gist substitution suppresses it. We call this failure mode categorical gist substitution.

We introduce GUARD (Ground, Unbundle, Anchor, Resolve, Distill), a five-move discipline the model is instructed to perform before generating its final response. GUARD is not a fine-tuning approach; it is a prompt-level intervention supplied as a system prompt, and it can be deployed against any conversational LLM without model access. We further introduce a GUARD-Controlled variant that pairs the discipline with a deterministic structural validator: the model first emits a structured bullet list that a checker validates against the GUARD schema, then composes the final narrative only from the approved bullets.

We report a Tier 1 empirical evaluation of GUARD on a 17-prompt benchmark of HBCU institutional-characterization tasks against Claude Fable 5, scored by three independent graders on a four-axis rubric. GUARD-Controlled produces substantial and statistically significant improvements over Baseline on every scoring axis, with the largest effect on the target metric (gist displacement, $d = 1.70$). A generic "be specific" Control condition produces no significant improvement, isolating the effect to the discipline rather than to prompt length or emphasis.

Contributions. (i) We articulate categorical gist substitution as a distinct behavioral failure mode in LLM characterization, grounded in fuzzy-trace theory. (ii) We introduce GUARD and GUARD-Controlled as prompt-level interventions with reproducible operational specifications. (iii) We provide the first empirical evaluation of the intervention with a public rubric, artifact-tracking review protocol, and complete grader data. (iv) We report honest inter-rater reliability numbers and name three concrete limitations of the current evaluation that constrain the scope of the claim.

2 Related Work

2.1 Fuzzy-Trace Theory and Dual-Process Memory

Fuzzy-trace theory [2, 6] distinguishes verbatim from gist encoding and predicts that gist traces dominate recall under load, delay, or ambiguity. Applied to LLM outputs, this framework suggests that models trained on the joint distribution of "entity + category" text will exhibit gist-preferential generation when the retrieval cue is categorical. Our contribution is to operationalize this prediction as a testable failure mode and to design an intervention that forces verbatim reconstruction ahead of gist consolidation.

2.2 Prompt-Level Behavioral Correction

A substantial literature on chain-of-thought prompting [5, 8] demonstrates that structural instructions embedded in the prompt can substantially change model outputs on reasoning tasks. Constitutional AI [1] extends this to behavior alignment through iterated self-critique. GUARD differs from these approaches in three respects: (i) it targets a specific characterization failure rather than general reasoning quality; (ii) its five moves are grounded in a cognitive-theoretic account of the failure it corrects; (iii) it can be paired with a deterministic checker to enforce structural adherence before final composition.

2.3 Fairness in Institutional Representation

Prior work on demographic bias in LLM outputs has emphasized stereotype detection [4, 7] and mitigation through fine-tuning or output filtering. Our framing is complementary: gist substitution is not itself a stereotype, but it produces representational harms when the categorical gist systematically understates an entity type's actual profile. The HBCU case is instructive precisely because the category framing is not derogatory; it is simply out-of-date with respect to the current institutional record.

3 The GUARD Discipline

GUARD requires the model to perform five ordered moves before producing its final characterization. Each move is defined operationally; failure to perform a move renders the response non-compliant with the discipline.

- **Ground.** Fix the specific entity before reaching for the category. Name the entity, its identifiers, its current classification.
- **Unbundle.** List the category or categories through which the entity is typically framed, and enumerate the concrete assumptions each framing carries.
- **Anchor.** Retrieve entity-specific verifiable facts: named centers, specific dollar figures, specific classifications, specific dated events. Do not proceed with abstractions until at least three entity-anchored facts are on the record.
- **Resolve.** For each unbundled framing, identify where the entity's specific record diverges from the framing's assumptions. Name the divergence explicitly.
- **Distill.** Deliver a corrected bottom line: a single-sentence gist that reflects the entity's actual record rather than the categorical prior.

GUARD-inline. The five moves are supplied as a system prompt; the model produces its response in one pass. This condition tests whether the discipline alone, without structural enforcement, is sufficient to change the response.

Algorithm 1 GUARD-inline

Require: entity-characterization prompt p
1: system prompt ← GUARD five-move instructions (Ground, Unbundle, Anchor, Resolve, Distill)
2: response ← LLM(p, system prompt) ▷ single pass, no structural enforcement
3: **return** response

GUARD-Controlled. The model first emits a structured bullet list containing five labeled sections (ENTITY, FACTS, CATEGORICAL FRAMINGS, DIVERGENCE, BOTTOM LINE). A rule-based checker validates the structure and the presence of required elements. If validation fails, the checker returns the specific reasons and the model regenerates (up to three attempts). Only after an approved bullet list, or after three attempts, does the model compose the final narrative, and it is instructed to use only the facts present in the approved bullets.

Control. A generic "be specific and factual" system prompt of comparable length to GUARD but without the five-move structure. This condition isolates the effect of the discipline from the effect of any additional prompt content.

Algorithm 2 GUARD-Controlled

Require: entity-characterization prompt p
system prompt ← GUARD five-move instructions
attempt ← 1; feedback ← ∅
while attempt ≤ 3 **do**
    bullets ← LLM(p, system prompt, feedback) ▷ ENTITY, FACTS, CATEGORICAL FRAMINGS, DIVERGENCE, BOTTOM LINE
    (valid, feedback) ← ValidateBulletList(bullets) ▷ rule-based structural checker
    if valid **then**
        break
    end if
    attempt ← attempt +1
end while
narrative ← LLM(bullets) ▷ compose final response using only approved bullets
return narrative

4 Evaluation Methodology

4.1 Prompt Set

We evaluate on Tier 1 of a curated benchmark: 17 prompts targeting institutional characterization of HBCUs. Prompts span single-institution questions (Morgan State, Howard, NC A&T, Xavier, Florida A&M, Coppin, Bowie State) and category-level questions ("What are HBCUs?"). Each prompt was authored by the project team and annotated, prior to any model response being generated, with an associated failure signature (the categorical framing to be avoided), a success signature (the entity-anchored framing the response should exhibit), and a target gist marker (the specific corrected bottom line). Failure and success signatures were cross-checked against verification sources for the relevant institution, including the current Carnegie Classification, institutional sponsored-research disclosures, and state-level legislative designations, so that the correction each prompt targets is independently verifiable rather than a matter of grader judgment.

4.2 Target Model

All responses were generated by Claude Fable 5 (`claude-fable-5`), accessed via the Anthropic API. Baseline, Control, and GUARD-inline used `max_tokens=1024`; GUARD-Controlled used `max_tokens=2048` to accommodate its two-step bullet-list-then-narrative protocol. Sampling temperature was left at the Anthropic API's default value of 1.0; one response was sampled per prompt per configuration. Each of the 17 prompts was run under each of the four configurations, producing $17 \times 4 = 68$ responses. GUARD-Controlled additionally records up to three regeneration attempts per prompt.

4.3 Rubric

Responses were scored on four ordinal axes and three binary flags.

Ordinal axes (0–3, higher is better except where noted):

- **Move sequencing.** Does the response perform G/U/A/R/D in order? 0 = none; 3 = all cleanly.
- **Verbatim density (count).** Integer count of entity-specific claims (dated, numbered, or named). Not bounded.
- **Verbatim displacement.** Was the stereotype-discordant fact surfaced (3) or buried (0)?

Table 1. Mean scores by configuration on the four axes. Standard deviations in parentheses.

Axis	Baseline (n=49)	Control (n=51)	GUARD (n=41)	GUARD-C (n=51)
Move sequencing	1.35 (0.97)	1.51 (0.95)	2.02 (0.91)	**2.25 (1.06)**
Verbatim density	8.78 (3.29)	8.84 (3.25)	7.34 (4.01)	**9.49 (3.47)**
Verbatim displacement	1.90 (1.08)	2.02 (1.03)	2.24 (1.11)	**2.73 (0.75)**
Gist displacement	0.79 (0.92)	1.04 (0.81)	1.46 (1.03)	**2.39 (0.96)**

Table 2. Pairwise effect sizes (Cohen's d) against Baseline. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$, ns = not significant.

Axis	Control	GUARD	GUARD-C
Move sequencing	+0.17 ns	$+0.72^{***}$	$+0.90^{****}$
Verbatim density	+0.07 ns	-0.66^{**}	+0.26 ns
Verbatim displacement	+0.12 ns	+0.32 ns	$+0.89^{****}$
Gist displacement	+0.29 ns	$+0.69^{**}$	$\mathbf{+1.70}^{****}$

- **Gist displacement.** Was a corrected bottom-line meaning explicitly articulated? 0 = no; 3 = explicit corrected bottom line delivered cleanly.

Binary flags (Y / N / N/A):

- Failure signature present (does the response match the failure pattern?)
- Success signature present (does the response match the success pattern?)
- Gist articulated (is there an explicit "bottom line" sentence?)

4.4 Graders and Procedure

Three independent graders (all with technical backgrounds; two graduate researchers and one senior undergraduate) reviewed all 68 responses per configuration through a custom four-pane review interface. Blind mode was enabled throughout: graders saw the prompt, entity, and response but not the configuration label. For GUARD-Controlled rows, the interface enforced that graders open each of the up-to-three attempts and the final narrative before saving. The database recorded which artifacts were viewed on each row.

4.5 Statistical Approach

We report per-configuration means and standard deviations; pairwise Cohen's d effect sizes; and Mann–Whitney U tests for ordinal comparisons. Binary flag shifts are tested via chi-square with 2×2 contingency tables. Inter-rater reliability is reported as ICC(2,1), two-way random effects, single-rater absolute agreement.

5 Results

5.1 Coverage and Sample

We collected 203 grader-observations across 68 responses (17 prompts $\times$ 4 configurations) and 3 graders. Two of the three graders returned complete data on all 68 rows; one grader had missing scores on 10 GUARD-inline rows (see Section 6.5). Effective per-cell n values are reported in each table.

Table 3. Inter-rater reliability (ICC(2,1)) pooled across configurations. n is the number of items (prompt × configuration cells) with complete data across all three graders.

Axis	ICC(2,1)	n
Move sequencing	0.20	60
Verbatim density	0.32	67
Verbatim displacement	0.35	60
Gist displacement	0.41	58

5.2 Main Effect: Mean Scores by Configuration

Table 1 reports mean scores by configuration.

5.3 Effect Sizes and Significance

Table 2 reports Cohen's d and two-sided Mann–Whitney p-values for pairwise comparisons against Baseline. GUARD-Controlled produces a very large effect on gist displacement ($d = 1.70$, $p < 0.0001$), a large effect on verbatim displacement ($d = 0.89$, $p < 0.0001$), and a large effect on move sequencing ($d = 0.90$, $p < 0.0001$). GUARD-inline produces smaller but significant effects on move sequencing ($d = 0.72$) and gist displacement ($d = 0.69$). The Control condition produces no significant effect on any axis, indicating that the improvement is attributable to the discipline itself rather than to prompt content or length.

5.4 Gist Displacement Distribution

The distribution of gist displacement scores shifts sharply under GUARD-Controlled. The median score is 0 for Baseline (interquartile range 0–1), 1 for Control (0–2), 1 for GUARD-inline (1–2), and 3 for GUARD-Controlled (2–3). Under GUARD-Controlled the modal outcome is an explicit corrected bottom-line articulation delivered cleanly; under Baseline the modal outcome is a response that reads with the category prior intact.

5.5 Binary Flag Shifts

The most striking effect is on the failure-signature flag: the fraction of responses that reproduce the categorical failure pattern drops from 51.0% under Baseline to 0.0% under GUARD-Controlled ($\chi^2 = 31.5$, $p < 0.0001$). GUARD-Controlled also increases explicit bottom-line articulation from 18.4% to 54.9% ($\chi^2 = 12.8$, $p < 0.001$) and success-signature match from 55.1% to 78.0% ($\chi^2 = 4.85$, $p = 0.028$).

5.6 Per-Prompt Effect

The effect is consistent across the 17-prompt set. Every prompt in the set exhibits a positive shift in gist displacement from Baseline to GUARD-Controlled, and 15 of 17 prompts show shifts of at least 1.0 on the 0–3 scale.

5.7 Inter-Rater Reliability

Table 3 reports ICC(2,1) values pooled across configurations for each axis. Reliability is modest: 0.20–0.41 across the four axes, with gist displacement (0.41) and verbatim displacement (0.35) higher than move sequencing (0.20). By Cicchetti's guidelines [3], these values fall in the "poor" to "fair" range. We report this honestly rather than reweighting or dropping graders. Two interpretations are consistent with the data. First, the rubric, particularly move sequencing, may require sharper operational definitions before

a larger evaluation. Second, the effect sizes reported above are large enough that even under conservative aggregation the between-configuration differences remain robust.

6 Discussion

6.1 The Discipline Does Something the Emphasis Does Not

The Control condition produces no significant improvement over Baseline on any axis, while GUARD-inline produces significant improvement on move sequencing and gist displacement, and GUARD-Controlled produces large and highly significant improvements on all three ordinal axes. This decomposes the effect cleanly. Merely instructing the model to "be specific and factual" with a system prompt of comparable length to GUARD is insufficient; the five-move structure carries the effect. This is consistent with the fuzzy-trace prediction: the corrective mechanism is not encouragement but sequencing, forcing verbatim reconstruction before gist consolidation.

6.2 Structural Enforcement Amplifies the Discipline

GUARD-Controlled outperforms GUARD-inline on every axis, with the largest additional gain on gist displacement ($d = 0.94$, $p < 0.0001$). The deterministic checker's role is not to teach the model the discipline, the model has the same system prompt in both conditions, but to prevent the model from skipping steps under generation pressure. Even a well-instructed model appears to shortcut moves when producing a single-pass response; the two-step protocol (validated bullets, then narrative) removes the shortcut.

6.3 The Failure Signature Falls to Zero

Under GUARD-Controlled, no response in the sample was scored as reproducing the failure signature. This is the strongest single result in the study. Under Baseline, roughly half of responses reproduced the categorical framing the failure signature captured. Under GUARD-Controlled, none did. This suggests that when the model is required to explicitly enumerate the categorical framings it might otherwise apply (the Unbundle move), and then required to identify where the entity diverges from those framings (the Resolve move), it becomes very difficult to then produce a final characterization that quietly re-adopts the frame just enumerated as inadequate.

6.4 Verbatim Density: A Non-Trivial Non-Effect

Verbatim density does not improve under GUARD-inline (in fact drops slightly, $d = -0.66$), and improves only modestly under GUARD-Controlled ($d = +0.26$, ns). This is worth examining. GUARD is not primarily about producing more specific claims; it is about correctly organizing and displacing existing claims. The Baseline and Control conditions already produce moderate density (8.8 claims per response) because Claude Fable 5 is a strong model that surfaces specifics when asked to characterize an institution. What Baseline responses lack is not specifics but the corrected gist that would tell the reader what the specifics mean. GUARD-Controlled preserves density (9.5) while re-organizing what the density adds up to. That is the theoretical prediction: verbatim traces are preserved; gist retrieval is corrected.

6.5 Limitations

Small grader panel. Three graders is at the low end of what allows meaningful IRR estimation. The reliability numbers we report are honest but noisy; larger panels would produce more stable ICC estimates. The magnitude of the between-configuration effects is large enough that this concern does not threaten the direction of the findings, but it does constrain the precision of the point estimates.

Missing data on GUARD-inline. One of three graders returned incomplete scores on 10 of 17 GUARD-inline rows. We report per-cell n transparently and do not impute. The direction and magnitude of effects survive this missingness, but the GUARD-inline vs. Baseline comparison is on a somewhat smaller and possibly non-random subset.

Single tier. The evaluation covers only Tier 1 (HBCU institutional characterization). The corrective direction throughout is "inflate against a category prior that underrepresents the entity's record." A defensible generality claim requires a counterbalanced evaluation in which some prompts require the discipline to deflate an overinflated categorical prior (e.g., "Palo Alto is a small city with real trade-offs, not a tech-utopia") and some serve as null controls where the categorical framing is already accurate. Absent such counterbalance, GUARD's generality as a cognitive discipline (rather than as a targeted lift for under-prestiged entities) is unestablished. We plan to extend the evaluation to Tiers 2–5 in follow-up work.

Single model. We evaluate against Claude Fable 5 only. Whether GUARD produces comparable effects on other frontier LLMs (Gemini, GPT-family, open-weight models) is not tested here. The prompt-level nature of the intervention makes cross-model portability plausible but not demonstrated.

Modest inter-rater reliability. As noted, ICC values in the 0.20–0.41 range are "poor to fair." A large-panel replication with a refined rubric, particularly sharper operational definitions for move sequencing, is a prerequisite for treating the axis-level numbers (rather than the between-configuration deltas) as calibrated measurements.

6.6 Threats to Validity Beyond Sampling

Rubric-outcome coupling. The rubric is designed to detect the failure GUARD corrects. If the rubric encodes "did the response perform GUARD moves," then a GUARD-conditioned response will score higher on that rubric by construction. We partially mitigate this by including verbatim density (an axis GUARD does not directly target and does not clearly improve), which behaves as predicted. But a fully independent evaluation would require a rubric constructed without knowledge of the intervention.

Grader awareness. Graders were blind to configuration but were told what the study evaluated. It is possible that graders, familiar with GUARD's structure, applied stricter standards to Baseline responses. This is a real concern that a between-subjects grader design (some graders unaware of the intervention) would address.

6.7 Rubric-Outcome Coupling Concern

We took two steps to partially mitigate the rubric-outcome coupling concern. First, the four ordinal axes are not all directly targeted by GUARD: verbatim density measures a property (raw claim count) that GUARD does not specifically instruct the model to increase, and indeed we find no improvement on this axis. Second, the binary failure-signature flag is defined in terms of the categorical framing the response reproduces, not in terms of GUARD moves. The 51% $\rightarrow$ 0% shift on this flag is therefore not simply reflecting whether the response performed the discipline; it reflects whether the categorical failure was avoided.

7 Conclusion and Future Work

We introduce categorical gist substitution as a distinct behavioral failure mode in LLM entity characterization, ground it in fuzzy-trace theory, and demonstrate that a five-move prompt-level discipline (GUARD) substantially reduces its incidence when paired with deterministic structural validation. On a 17-prompt Tier 1 benchmark of HBCU characterization, GUARD-Controlled reduces failure-signature

prevalence from 51.0% to 0.0%, triples explicit bottom-line articulation, and produces a Cohen's d of 1.70 on the target gist-displacement metric relative to Baseline. A generic "be specific" Control condition does not replicate the effect, isolating the mechanism to the discipline rather than to prompt length or emphasis.

Three follow-up directions are worth naming. First, Tiers 2–5 of the benchmark introduce counterbalance: prompts in which the discipline must deflate rather than inflate categorical priors, and null-prompts where no correction is warranted. Running these evaluations will test whether GUARD is a general cognitive discipline or a targeted lift for under-prestiged entities. Second, a larger grader panel with a refined rubric would produce publishable ICC values and stabilize axis-level measurements. Third, cross-model portability (Gemini, GPT-family, open-weight) remains untested; the prompt-level nature of the intervention makes portability plausible, but replication is needed. The Tier 1 study is best read as a strong initial demonstration on a focused case rather than as a general claim about LLM characterization behavior.

The deterministic checker itself is a limitation as much as a strength: it enforces structural adherence but does not validate semantic correctness. A checker that flags when a bullet's "FACTS" contain unverifiable or hallucinated claims would substantially strengthen the guarantee. Extending the validator with retrieval-augmented fact-checking is the natural next step for the Controlled variant.

Data and Code Availability

The 17-prompt benchmark, GUARD skill files (inline and Controlled variants), the deterministic validator (`validate_bullet_list.py`), the review interface, and complete anonymized grader scores will be released as a companion repository on publication. The rubric and grader instructions are included as an appendix. Code will be released under the MIT License; the benchmark and grader data will be released under a Creative Commons Attribution 4.0 (CC BY 4.0) license, consistent with JAIR's own open-access policy.

Ethical Considerations

The intervention operates on institutional characterization, not on individual persons. Nevertheless, the failure mode we describe has representational-fairness implications: the HBCU case exemplifies how categorical gist substitution can systematically understate the current profile of a class of institutions. GUARD is not proposed as a substitute for improved training data or model-level fairness interventions; it is a prompt-level correction that can be deployed against existing deployed models without model access. Deployment considerations, including whether to enable GUARD by default in downstream applications, are appropriately left to the application-layer product owner.

AI Use Disclosure Statement

The authors used generative AI (Claude, Anthropic) for language refinement. All content was reviewed and verified, and the authors assume full responsibility for the work.

References

[1] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073* (2022).

[2] C. J. Brainerd and V. F. Reyna. 2004. *The Science of False Memory.* Oxford University Press.

[3] D. V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284–290.

[4] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of FAccT*. 862–872.
[5] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, Vol. 35. 22199–22213.
[6] V. F. Reyna and C. J. Brainerd. 2008. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences* 18, 1 (2008), 89–107.
[7] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of EMNLP-IJCNLP*. 3407–3412.
[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, Vol. 35. 24824–24837.

A Reproducibility Checklist

All articles:

(1) All claims investigated in this work are clearly stated. **Yes.**
(2) Clear explanations are given how the work reported substantiates the claims. **Yes.**
(3) Limitations or technical assumptions are stated clearly and explicitly. **Yes** (Section 6.5).
(4) Conceptual outlines and/or pseudo-code descriptions of the AI methods introduced in this work are provided, and important implementation details are discussed. **Yes** (Algorithms 1–2).
(5) Motivation is provided for all design choices, including algorithms, implementation choices, parameters, data sets and experimental protocols beyond metrics. **Yes.**

Articles containing theoretical contributions: Does this paper make theoretical contributions? **No.**

Articles reporting on computational experiments: Does this paper include computational experiments? **Yes.**

(1) All source code required for conducting experiments is included in an online appendix or will be made publicly available upon publication of the paper, following best practices for readability, documentation, and long-term accessibility. **Yes** (companion repository, MIT License; see Data and Code Availability).
(2) The source code comes with a license that allows free usage for reproducibility purposes. **Yes** (MIT License).
(3) The source code comes with a license that allows free usage for research purposes in general. **Yes** (MIT License).
(4) Raw, unaggregated data from all experiments is included in an online appendix or will be made publicly available upon publication of the paper. **Yes.**
(5) The unaggregated data comes with a license that allows free usage for reproducibility purposes. **Yes** (CC BY 4.0).
(6) The unaggregated data comes with a license that allows free usage for research purposes in general. **Yes** (CC BY 4.0).
(7) If an algorithm depends on randomness, then the method used for generating random numbers and for setting seeds is described in a way sufficient to allow replication of results. **Yes.** The target model's sampling temperature is reported (Section 4.2, left at the Anthropic API default of 1.0). No explicit random seed is set, as the Anthropic API does not expose one.
(8) The execution environment for experiments is described. **Partially.** The target model, API, and decoding parameters are reported (Section 4.2). The grading interface is a custom internal tool whose software stack is not itemized, as it plays no role in generating model responses.

(9) The evaluation metrics used in experiments are clearly explained and their choice is explicitly motivated. **Yes** (Section 4.3).
(10) The number of algorithm runs used to compute each result is reported. **Yes** (per-cell n reported in Tables 1–3).
(11) Reported results have not been "cherry-picked" by silently ignoring unsuccessful or unsatisfactory experiments. **Yes.** The null result on verbatim density (Section 6.4) is reported alongside the positive results.
(12) Analysis of results goes beyond single-dimensional summaries of performance to include measures of variation, confidence, or other distributional information. **Yes** (standard deviations, effect sizes, and inter-rater reliability are all reported).
(13) All (hyper-)parameter settings for the algorithms/methods used in experiments have been reported, along with the rationale or method for determining them. **Yes** (Section 4.2: token limit reported; sampling temperature left at the Anthropic API default of 1.0 rather than tuned).
(14) The number and range of (hyper-)parameter settings explored prior to conducting final experiments have been indicated, along with the effort spent on (hyper-)parameter optimisation. **N/A.** No hyperparameter search was performed. A single fixed configuration was evaluated per condition.
(15) Appropriately chosen statistical hypothesis tests are used to establish statistical significance in the presence of noise effects. **Yes** (Mann–Whitney U and χ^2 tests; Section 4.5).

Articles using data sets: Does this work rely on one or more data sets? **Yes.**

(1) All newly introduced data sets are included in an online appendix or will be made publicly available upon publication of the paper, with a license that allows free usage for research purposes. **Yes** (CC BY 4.0; see Data and Code Availability).
(2) The newly introduced data set comes with a license that allows free usage for reproducibility purposes. **Yes** (CC BY 4.0).
(3) The newly introduced data set comes with a license that allows free usage for research purposes in general. **Yes** (CC BY 4.0).
(4) All data sets drawn from the literature or other public sources are accompanied by appropriate citations. **N/A.** The benchmark is newly constructed rather than drawn from an existing public dataset. Institutional verification sources (Carnegie Classification, state legislation) are cited narratively where relevant (Section 4.1).
(5) All data sets drawn from the existing literature are publicly available. **N/A** (see previous item).
(6) All new data sets and data sets that are not publicly available are described in detail, including relevant statistics, the data collection process and annotation process if relevant. **Yes** (Section 4.1).
(7) All methods used for preprocessing, augmenting, batching or splitting data sets are described in detail. **N/A.** The benchmark is not partitioned into training/validation/test splits.