
MiroBench: Benchmarking Realism in Agentic Simulation of Real-world Discussions

Yaoning Yu¹ Ye Yu¹ Haojing Luo² Haohan Wang^{1*}
¹University of Illinois at Urbana-Champaign ²Starc.Institute

Abstract

LLM agents are increasingly used to simulate real world interactions, but it remains unclear whether simulated behaviors preserve the content patterns and interaction dynamics of real human behaviors. Existing evaluations remain fragmented, which makes it difficult to compare systems or measure progress. In this paper, we focus on Reddit discussions as a concrete first step toward evaluating real-world social simulation. Reddit threads provide public, topic-grounded, multi-party interactions where people share experiences, debate, seek advice, express emotion, and collectively respond to products, events, and social issues. These discussions offer an observable window into broader social behavior, making them a useful setting for testing whether LLM agents can reproduce not only fluent text, but also the distributional patterns and interaction dynamics of real online communities. We introduce MiroBench, a benchmark for Reddit discussion simulation built from 4,292 real Reddit threads. MiroBench uses statistical tests to compare generated and real discussions across four major aspects: repetition and semantic uniformity, narrative content, toxicity and aggression, and structural complexity. Experiments across five domains and five models show that current simulators remain distributionally mismatched with real Reddit threads, while a lightweight prompt-based improvement procedure provides only limited gains. MiroBench offers a concrete benchmark for measuring, diagnosing, and improving realism in LLM-based social simulation.

1 Introduction

LLM-based agents are increasingly used as simulators of human behavior, social interaction, and online communities [1, 2, 3]. These systems are appealing because they promise a controllable way to study how people discuss products, react to events, exchange opinions, and form group-level patterns. This promise requires more than generating conversations: we need to know whether simulated interactions actually match the content patterns and interaction dynamics of real human behavior. However, existing studies often focus on individual responses, controlled role-playing settings, or task performance in simulated environments. As a result, it remains difficult to measure whether a simulator can reproduce real-world discussion behavior, or to compare progress across different systems, leading to a practical need of such measurement.

We argue that Reddit discussions are practical setting for this measurement [4, 5]. Reddit threads are public, topic-grounded, and multi-party, and they contain many signals of social behavior: advice-seeking, personal experience, disagreement, emotional tone, repeated talking points, and complexity in structure [6]. They therefore provide an observable window into broader social behavior.

Therefore, we introduce MiroBench, a benchmark for Reddit discussion simulation. MiroBench combines 4,292 real Reddit threads, matched train/validation/test splits, standardized seed contexts,

*Corresponding author.

and an evaluation framework that compares generated and real discussions through both comment-derived signals and reply-tree structure.

Using MiroBench, we evaluate whether sets of simulated threads match real Reddit discussions across comment content and thread structure. We find that current simulators remain mismatched to real Reddit threads even when individual comments are fluent, with only partial gains from prompt-based improvement. Figure 1 shows one example of this gap.

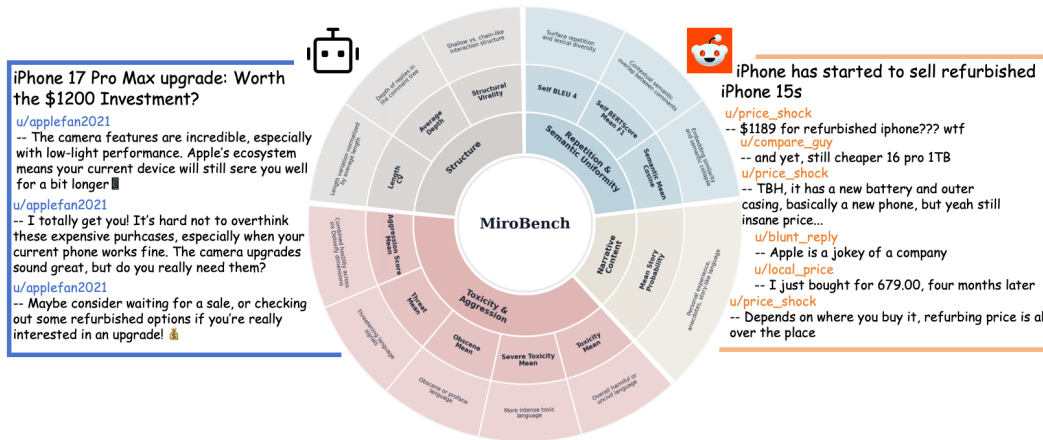


Figure 1: Overview of MiroBench. Left: an example simulated thread generated by a model-based discussion simulator. Center: the four benchmark metric groups used in MiroBench: repetition and semantic uniformity, narrative content, toxicity and aggression, and structure. Right: an example real Reddit thread. MiroBench evaluates whether generated threads match real discussion patterns through both comment-derived metrics and reply-tree structure.

Our contributions are:

- We introduce MiroBench, a benchmark for evaluating Reddit discussion simulation across comment content and thread structure.
- We develop an evaluation framework for repetition and semantic uniformity, narrative content, toxicity and aggression, and reply-tree structure, together with statistical checks for benchmark consistency.
- We use the benchmark to show systematic gaps between simulated and real Reddit threads, and to measure gains from a lightweight prompt-based improvement procedure.

2 Related Work

2.1 LLM-based Social Simulation

Recent work uses LLMs as agents to simulate people, social interaction, and group behavior. Generative Agents studies everyday routines and social behavior in a small town-like environment [1]. SOTOPIA and SOTOPIA- π move to role-based multi-turn interaction and social intelligence evaluation, where agents pursue goals and are judged by how they behave in social settings [7, 8]. These systems show that LLM agents can keep interactions going over time, but they do not mainly study whether generated online discussions match the distribution of real ones.

Other systems move closer to larger online-platform settings. OASIS studies social-media simulation with dynamic networks, user actions, posts, and recommendation systems [3]. AgentSociety and SOTOPIA-S4 scale up the number of agents and support larger social settings and more flexible control [9, 2]. Y Social and MOSAIC focus more directly on online platforms, user behavior, content spread, engagement, and moderation [10, 11].

2.2 Simulation and Social Media Benchmarks

Recent benchmarks for evaluating whether LLMs can simulate human and social behavior remain limited. SimBench tests whether LLMs can match human response distributions in mostly static, single-turn settings [12]. SoMe evaluates LLM-based social media agents across tasks such as user behavior prediction, recommendation, and comment simulation [13].

MiroBench addresses this missing evaluation setting that evaluates Reddit discussion simulation across five domains and asks whether a set of simulated threads matches real Reddit threads across repetition, semantic uniformity, narrative content, toxicity and aggression, and reply-tree structure.

3 MiroBench: Dataset and Metrics

3.1 Benchmark Construction

MiroBench uses discussion threads as the basic evaluation samples across five Reddit domains: Credit Card, Laptop, Cellphone, Camera, and Headphones. Each real sample contains one root Reddit post and all comments in the corresponding thread.

For each domain, MiroBench provides real train, validation, and test splits, together with product descriptions used as standardized seed contexts for simulation. Laptop, Cellphone, Camera, and Headphones each include 200 product descriptions, while Credit Card includes 75, giving 875 seed contexts in total. Simulators use these descriptions to generate comparable discussion threads.

3.2 Benchmark Metrics

Each discussion thread is scored with a set of aggregated metrics that capture both comment content and thread structure. MiroBench uses a broad evaluation suite and reports a stable core subset of 12 metrics in the current paper. We organize these metrics into four families: repetition and semantic uniformity, narrative content, toxicity and aggression, and structure.

Repetition and semantic uniformity. Measures whether comments in a thread repeat similar wording or ideas.

- `self_bleu_4`: Since BLEU-4 measures 4-gram overlap, higher values indicate more repeated phrasing and lower lexical diversity [14].
- `self_bertscore_mean_f1`: BERTScore compares contextual token embeddings, higher values indicate stronger semantic overlap even when wording differs [15, 16].
- `semantic_mean_cosine`: embeds each comment with sentence-transformers/all-mpnet-base-v2 and averages pairwise cosine similarity. Higher values indicate that comments occupy a narrower semantic space [17, 18].

Domain	Seeds	Threads	Train/Val/Test
Credit Card	75	2,600	1,819 / 390 / 391
Laptop	200	272	186 / 43 / 43
Cellphone	200	392	286 / 50 / 56
Camera	200	763	546 / 111 / 106
Headphones	200	265	178 / 38 / 49
Total	875	4,292	3,015 / 632 / 645

Table 1: MiroBench inventory across five Reddit domains.

Narrative content. This group measures how much first-hand experience and story-like detail appears in the discussion.

- `mean_story_probability` scores each comment with StorySeeker and averages the story probability across the thread. Higher values indicate more anecdotal or narrative-style content [19].

Toxicity and aggression. This group measures whether a thread contains hostile or harmful language.

- `toxicity_mean`, `severe_toxicity_mean`, `obscene_mean`, and `threat_mean` score each comment with the Detoxify **unbiased** model and average each dimension across the thread.
- `aggression_score_mean` first averages six Detoxify dimensions for each comment: toxicity, severe toxicity, obscene, threat, insult, and identity attack. It then averages this score across the thread, giving a broader summary of hostile language [20].

Structure. This group measures comment length variation and reply-tree shape.

- `length_cv` is the coefficient of variation of comment lengths, normalizing length variation by the mean comment length.
- `avg_depth` averages comment depth across the reply tree, with top-level comments as depth 1.

- `structural_virality` treats the thread as an undirected reply graph and averages shortest-path distance over all unordered comment pairs. We adapt this measure from diffusion trees to Reddit-style reply trees to capture shallow versus deeper interaction chains [21].

3.3 Scoring Setup

For each metric f , the real threads define a reference distribution R_f , and the simulated threads define a generated distribution G_f . We compare R_f and G_f using the Mann–Whitney U test and the Kolmogorov–Smirnov test. We use their p -values as the main statistics for measuring whether simulated discussions are distinguishable from real Reddit discussions. We also report Wasserstein distance and absolute Cliff’s delta as secondary discrepancy statistics. Full results are provided in Appendix E.6, Appendix F, and Appendix G.

3.4 Evaluation Process and Metric Formulas

Let $t = (c_1, \dots, c_{n_t})$ be a discussion thread with n_t comments. Let $\mathcal{P}(t) = \{(i, j) : 1 \leq i < j \leq n_t\}$ be the set of comment pairs, and let $M_t = |\mathcal{P}(t)|$. For comment c_i , let e_i be its sentence embedding, ℓ_i its token length, d_i its reply depth, and s_i its StorySeeker score. We denote the Detoxify outputs by $\tau_i = (\tau_i^{\text{tox}}, \tau_i^{\text{sev}}, \tau_i^{\text{obs}}, \tau_i^{\text{thr}}, \tau_i^{\text{ins}}, \tau_i^{\text{id}})$. Let G_t be the undirected reply graph induced by thread t .

For each metric function f , MiroBench assigns one score to each thread and then compares the thread-level score distributions:

$$R_f^{(d)} = \{f(t) : t \in \mathcal{T}_{\text{real}}^{(d)}\}, \quad G_f^{(d,m)} = \{f(t) : t \in \mathcal{T}_{\text{gen}}^{(d,m)}\}.$$

All tests and discrepancy statistics are applied to $R_f^{(d)}$ and $G_f^{(d,m)}$.

Thread- and distribution-level metric formulas. For pairwise overlap metrics, $\widetilde{\text{BLEU}}_4(c_i, c_j)$ denotes the symmetric BLEU-4 score of a comment pair. The 12 core thread-level metrics are shown in Appendix Table 3. The two secondary discrepancy statistics are

$$W_1(R_f, G_f) = \int_{-\infty}^{\infty} |F_{R_f}(x) - F_{G_f}(x)| dx,$$

$$|\delta_{\text{Cliff}}(R_f, G_f)| = \left| \frac{1}{|R_f||G_f|} \sum_{r \in R_f} \sum_{g \in G_f} \text{sgn}(g - r) \right|.$$

Here F_{R_f} and F_{G_f} are empirical CDFs.

How to read improvements. We use $p = 0.05$ as the significance threshold. More metrics with MWU or KS $p \geq 0.05$ mean that generated discussions are harder to distinguish from real discussions. For Wasserstein distance and absolute Cliff’s delta, lower values mean smaller gaps, so we compare them only within the same metric, domain, and model. We also compare each generated vs.-real discrepancy with the repeated real vs.-real median. Values below this reference are treated as within the typical real-vs.-real range. For structure metrics such as `avg_depth`, `length_cv`, and `structural_virality`, this range should be large which refers to larger diversity in structure and length of the.

3.5 Sanity Check: Repeated Real-vs.-Real Validation

As a sanity check, we repeatedly sample two independent real-data subsets from the same domain and compare them 200 times for each of the 12 core metrics. As shown in Figure 2, same-domain real-vs.-real comparisons usually stay above the $p = 0.05$ threshold. Across repeated comparisons, Mann–Whitney U p -values exceed 0.05 in 95.3% of repeats on average, and Kolmogorov–Smirnov p -values exceed 0.05 in 96.6% of repeats on average. Full per-domain counts are reported in Appendix E. These results show that the evaluation pipeline does not usually separate two samples from the same real-data distribution.

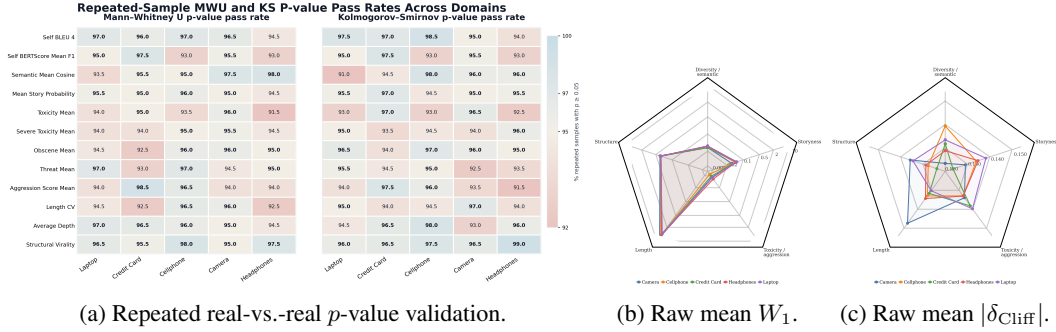


Figure 2: Repeated real-vs.-real validation on MiroBench. For each domain and each core metric, we compare two independently sampled real-thread subsets from the same domain across 200 repeats. The heatmap reports the fraction of repeats whose Mann–Whitney U and Kolmogorov–Smirnov p -values exceed 0.05. The radar plots summarize the same repeated-sample procedure using raw mean W_1 and raw mean $|\delta_{Cliff}|$ grouped by metric type. The W_1 radar uses a log-scaled radial axis because raw W_1 values have different scales across metric groups, while the $|\delta_{Cliff}|$ radar uses a zoomed radial range to show small differences.

The Figure 2b and Figure 2c show the scale of W_1 $|\delta_{Cliff}|$ real-vs.-real variation. Differently, we separate the structural metrics into two groups: length variation, including `length_cv`, and thread structure, including `avg_depth` and `structural_virality` because both capture large sources of natural variation in real Reddit discussions. Full per-metric distance statistics are reported in Appendix E.6.

4 Prompt-based Calibration Study

We study prompt-based calibration as a validation-guided prompt revision process. At each step, MiroBench compares simulated discussions with real validation discussions, identifies metric gaps, and uses these diagnostics to guide candidate prompt revisions. Each candidate prompt is tested by running a fresh simulation, and the candidate that best reduces the validation discrepancy is kept. The final revised prompt is then evaluated on the held-out test split.

4.1 Calibration Setup

For each domain, let D_{train}^{real} , D_{val}^{real} , and D_{test}^{real} denote the real train, validation, and test splits. The train split provides qualitative examples of real Reddit discussions. The validation split is used to select prompt changes during calibration. The test split is held out for before/after calibration. For each discussion thread t , MiroBench computes a vector of thread-level metric scores: $m(t) = \{m_1(t), m_2(t), \dots, m_K(t)\}$. For each metric m_k , the real validation threads define a reference distribution R_k , and the simulated threads generated under a prompt revision θ define a generated distribution $G_k(\theta)$. The prompt revision is restricted to two editable fields: $\theta = \{\theta^{persona}, \theta^{style}\}$, where $\theta^{persona}$ controls persona generation guidance and θ^{style} controls comment writing guidance. Persona guidance affects simulated users’ background, product experience, confidence level, personal memories, and reasons for joining the discussion. Style guidance affects comment length, anecdote use, disagreement style, short replies, repeated phrasing, and reactions to earlier comments.

4.2 Calibration Objective

We formulate prompt calibration as a distribution-matching problem. For each metric m_k , we define a discrepancy score between the real validation distribution R_k and the generated distribution $G_k(\theta)$:

$$d_k(\theta) = \lambda_W \widetilde{W}_k(\theta) + \lambda_C |\Delta_k(\theta)| + \lambda_P P_k(\theta).$$

Here, $\widetilde{W}_k(\theta)$ is the normalized Wasserstein distance, $|\Delta_k(\theta)|$ is the absolute Cliff’s delta, and $P_k(\theta)$ is a penalty based on Mann–Whitney U and Kolmogorov–Smirnov tests. Lower $d_k(\theta)$ means that the simulated distribution is closer to the real distribution for metric m_k .

At each iteration i of block b , we select the candidate prompt by minimizing the following block-wise objective:

$$\theta_i^* = \arg \min_{\theta \in \Theta_{i,b}} \left[\sum_{k \in \mathcal{M}_b^{act}} w_k d_k(\theta) + \gamma \sum_{k \in \mathcal{M}_b^{prot}} \max(0, d_k(\theta) - d_k(\theta_b^{base})) \right],$$

where θ_b^{base} is the cumulative prompt before block b starts, and $\Theta_{i,b} = \{\theta^{cum} + \Delta\theta_{i,c}\}_{c=1}^C$ is the candidate prompt set at iteration i of block b . The first term measures improvement on the current metric group, and the second term penalizes regressions on protected metrics. Within each block, we keep the best candidate prompt update as $\Delta\theta_b^*$. After each block, we update the cumulative prompt as $\theta^{cum} \leftarrow \theta^{cum} + \Delta\theta_b^*$, and the final prompt is the cumulative prompt after all blocks.

4.3 Calibration Workflow

Candidate Generation We use a fixed 12-iteration calibration schedule divided into four groups of metrics, with three iterations per group. At iteration i of block b , we first compute validation diagnostics:

$$D_i = \text{DIAG}(S(\theta^{cum} + \Delta\theta_b^*), \mathcal{D}_{val}^{real}, \mathcal{M}_b^{act}),$$

where S is the simulator, θ^{cum} is the cumulative prompt revision from previous blocks, and $\Delta\theta_b^*$ is the best prompt update found so far in the current block. Candidate generation uses two LLM steps.

1) At iteration i of block b , the strategist LLM receives the tuple

$$x_{i,b}^{strat} = (D_i, g_b, H_b, E_{train}^{real}, E_i^{sim}),$$

where D_i denotes the current validation diagnostics, g_b is the goal of block b , H_b is the history of previous candidate results within the same block, E_{train}^{real} is a small set of real discussion examples sampled from $\mathcal{D}_{train}^{real}$, and E_i^{sim} is a small set of simulated examples generated by the current prompt. Given this input, the strategist proposes high-level directions for improving the current metric group: $\{r_{i,c}\}_{c=1}^C = \text{STRATEGIST}(x_{i,b}^{strat})$.

2) a materializer LLM turns each direction into concrete text for the two editable prompt fields:

$$\Delta\theta_{i,c} = \text{MATERIALIZER}(r_{i,c}) = \{\Delta\theta_{i,c}^{persona}, \Delta\theta_{i,c}^{style}\}.$$

Candidate Evaluation and Selection Each candidate prompt update is added to the current cumulative prompt revision: $\theta_{i,c} = \theta^{cum} + \Delta\theta_{i,c}$. We then run a fresh simulation: $T_{i,c} = S(\theta_{i,c})$, and score the resulting threads with MiroBench. The candidate is evaluated by the block-level validation loss:

$$s_{i,c} = \mathcal{L}_b(\theta_{i,c}; S, \mathcal{D}_{val}^{real}, \mathcal{M}_b^{act}, \mathcal{M}_b^{prot}, \theta_b^{base}).$$

The best candidate at iteration i is: $c^* = \arg \min_{c \in \{1, \dots, C\}} s_{i,c}$. If this candidate improves the best score found in the current block, we update the block-best prompt update: $(\Delta\theta_b^*, s_b^*) \leftarrow (\Delta\theta_{i,c^*}, s_{i,c^*})$.

At the end of each block, the block-best prompt update is added to the cumulative prompt revision: $\theta^{cum} \leftarrow \theta^{cum} + \Delta\theta_b^*$. The final prompt revision is therefore built as:

$$\theta^{final} = \theta^{story} + \theta^{diversity} + \theta^{length} + \theta^{tone}.$$

Earlier prompt changes are preserved, and later blocks append new guidance rather than replacing the whole prompt. The pseudo-code is shown in the Algorithm 1

Final Evaluation After all 12 iterations, we run a fresh simulation using θ^{final} . This simulation is compared against $\mathcal{D}_{test}^{real}$. We report the same before/after statistics used in the benchmark: MWU and KS p -value counts, Wasserstein distance, and absolute Cliff’s delta. This separates prompt selection from final reporting and reduces the risk of overfitting to the validation split.

Algorithm 1 Prompt-based Calibration

Require: Simulator S , train split $\mathcal{D}_{train}^{real}$, validation split \mathcal{D}_{val}^{real} , block schedule \mathcal{B} , candidates per iteration C

Ensure: Calibrated prompt revision θ^{final}

- 1: Initialize cumulative prompt revision $\theta^{cum} \leftarrow \emptyset$
- 2: **for** $b \in \mathcal{B}$ **do**
- 3: Get active and protected metrics $(\mathcal{M}_b^{act}, \mathcal{M}_b^{prot}) \leftarrow \Phi(b)$
- 4: Set block base prompt revision $\theta_b^{base} \leftarrow \theta^{cum}$
- 5: Initialize block-best prompt update and score $(\Delta\theta_b^*, s_b^*) \leftarrow (\emptyset, +\infty)$
- 6: **for** $i = 1, \dots, I_b$ **do**
- 7: Compute diagnostics: $D_i \leftarrow \text{DIAG}(S(\theta^{cum} + \Delta\theta_b^*), \mathcal{D}_{val}^{real}, \mathcal{M}_b^{act})$
- 8: Generate high-level directions: $\{r_{i,c}\}_{c=1}^C \leftarrow \text{STRATEGIST}(D_i, b, \mathcal{D}_{train}^{real})$
- 9: Materialize prompt updates: $\{\Delta\theta_{i,c}\}_{c=1}^C \leftarrow \text{MATERIALIZER}(\{r_{i,c}\}_{c=1}^C)$
- 10: **for** $c = 1, \dots, C$ **do**
- 11: Apply candidate update $\theta_{i,c} \leftarrow \theta^{cum} + \Delta\theta_{i,c}$
- 12: Evaluate candidate: $s_{i,c} \leftarrow \mathcal{L}_b(\theta_{i,c}; S, \mathcal{D}_{val}^{real}, \mathcal{M}_b^{act}, \mathcal{M}_b^{prot}, \theta_b^{base})$
- 13: **end for**
- 14: Select best candidate $c^* \leftarrow \arg \min_{c \in \{1, \dots, C\}} s_{i,c}$
- 15: **if** $s_{i,c^*} < s_b^*$ **then**
- 16: Update block best $(\Delta\theta_b^*, s_b^*) \leftarrow (\Delta\theta_{i,c^*}, s_{i,c^*})$
- 17: **end if**
- 18: **end for**
- 19: Update cumulative prompt revision $\theta^{cum} \leftarrow \theta^{cum} + \Delta\theta_b^*$
- 20: **end for**
- 21: $\theta^{final} \leftarrow \theta^{cum}$
- 22: **return** θ^{final}

Table 2: Summary of simulation-vs.-real statistical tests for OASIS, SynthPAI, and prompt-based calibration. Each entry counts how many of the 12 core metrics are not statistically distinguishable from held-out real discussions under $p \geq 0.05$. Larger counts indicate closer alignment with real discussions.

Model	Domain	OASIS		SynthPAI		Prompt-based calibration	
		MWU $p \geq 0.05$	KS $p \geq 0.05$	MWU $p \geq 0.05$	KS $p \geq 0.05$	MWU $p \geq 0.05$	KS $p \geq 0.05$
GPT-4o-mini	Laptop	0/12	0/12	2/12	0/12	0/12	0/12
	Credit Card	0/12	0/12	0/12	0/12	2/12	0/12
	Cellphone	0/12	0/12	4/12	0/12	2/12	1/12
	Camera	0/12	0/12	3/12	0/12	3/12	1/12
	Headphones	0/12	0/12	1/12	0/12	2/12	1/12

5 Experiments and Results

5.1 Baseline and Calibration Results on MiroBench

Baselines. We evaluate OASIS [3] and SynthPAI [22] as uncalibrated discussion simulation baselines on MiroBench. OASIS represents an LLM-agent simulation setting where multiple agents interact to generate Reddit-like discussions. SynthPAI uses a persona-based synthetic interaction setting where simulated users produce discussion content from generated profiles. We also evaluate our prompt-based calibration method from Section 4. Unlike OASIS and SynthPAI, our method is not a standalone simulator. It is a prompt revision procedure applied on top of a simulator using validation-set diagnostics from MiroBench.

Table 2 summarizes the simulation-vs.-real statistical tests. Each entry counts how many of the 12 core metrics are not statistically distinguishable from held-out real discussions at $p \geq 0.05$. The uncalibrated OASIS baseline shows the largest mismatch: across all five domains, both MWU and KS counts are 0/12. This means that every core metric remains statistically distinguishable from real Reddit discussions. SynthPAI improves some MWU counts, especially in Cellphone and Camera, but its KS counts remain 0/12 across all domains. This suggests that some rank-based differences are reduced, while the full distributional shapes remain far from real discussions.

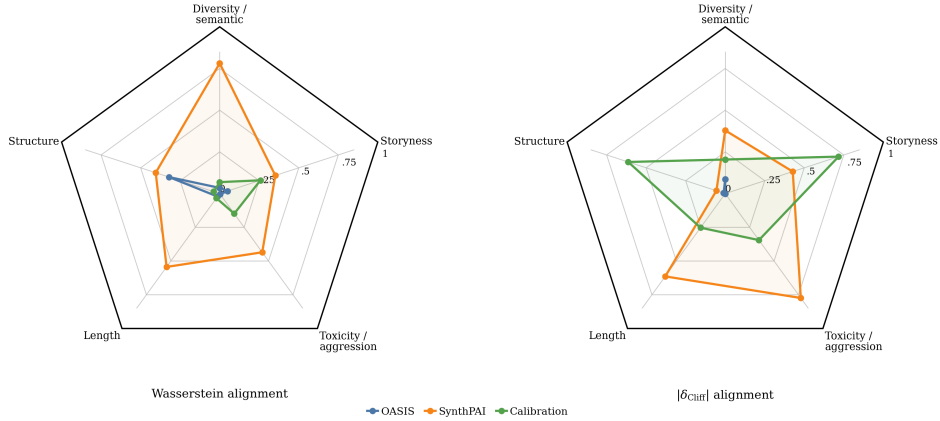


Figure 4: Alignment comparison across OASIS, SynthPAI, and prompt-based calibration. The left radar plot shows Wasserstein alignment, and the right radar plot shows $|\delta_{\text{CDiff}}|$ alignment across metric groups. Larger values indicate better alignment with real-vs.-real variation. Prompt-based calibration improves alignment over OASIS on several groups, especially for Wasserstein-based alignment, while the remaining gaps show that prompt revision does not fully solve the simulation-vs.-real mismatch.

Prompt-based calibration gives modest but broader gains. Compared with OASIS, it improves MWU counts in four of the five domains, reaching 2/12 in Credit Card, Cellphone, and Headphones, and 3/12 in Camera. It also produces non-zero KS counts in Cellphone, Camera, and Headphones. The improvement is still limited, but it shows that validation-guided prompt revision can move generated discussions closer to real Reddit discussions on several metrics.

Figure 3 shows the size of these gaps. A ratio of 1 means that the generated-vs.-real gap is similar to the natural gap between two real-data samples. Most ratios are above 1, often by a large margin. This means that the mismatch is not only statistically significant, but also larger than normal real-vs.-real variation. The largest gaps appear in Repetition/Semantic and Structure metrics, showing that generated discussions differ strongly from real discussions in semantic diversity and thread organization. Narrative metrics also remain above the reference line. Toxicity/Aggression is usually closer to the reference, but it is still not fully matched.

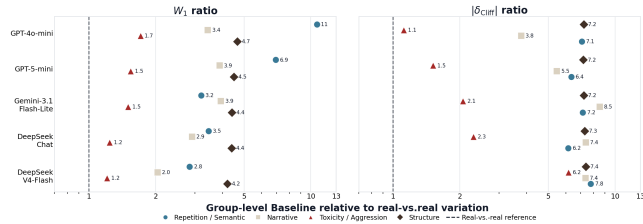


Figure 3: Group-level baseline discrepancies relative to real-vs.-real variation. For each model and metric group, we divide the generated-vs.-real statistic by the repeated real-vs.-real median reference R . The left panel shows the W_1 ratio, and the right panel shows the $|\delta_{\text{CDiff}}|$ ratio. The dashed vertical line marks the real-vs.-real reference at 1.0. Values closer to 1 indicate closer agreement with normal real-vs.-real variation.

Figure 4 compares OASIS, SynthPAI, and prompt-based calibration using group-level alignment scores. The radar plots show a different view from the p-value table. Instead of only asking whether a metric is statistically distinguishable, they show how close each method is to the repeated real-vs.-real reference. Prompt-based calibration improves the Wasserstein alignment over OASIS for several metric groups, especially Diversity/Semantic, Storyness, Toxicity/Aggression, and Structure. This means that the calibrated generations move closer to the real-discussion distributions in terms of distance. The $|\delta_{\text{CDiff}}|$ alignment also improves over OASIS in several groups, although SynthPAI remains stronger on some groups such as Storyness and Structure.

Overall, the results show two main findings. First, current simulators remain clearly distinguishable from real Reddit discussions, especially under the KS test. Second, prompt-based calibration provides modest but useful improvement without changing the underlying model, adding fine-tuning, or using extra supervised training. This supports our main claim: MiroBench can reveal where simulators fail and can also measure whether calibration methods reduce the distributional gap.

5.2 Qualitative Observations

Based on the generated discussions and their metric scores, we identify several qualitative patterns that help explain the distributional gaps measured by MiroBench.

Generated comments are more consistently helpful and length-aligned. As detailed in Appendix B, LLM-generated users tend to answer in a polished and helpful style. In the generated credit-card thread, many replies follow a similar template: recognize the user’s budget concern, compare the Apple Card with the American Express Platinum Card, explain the annual-fee trade-off, and suggest a safer rewards strategy. This makes individual comments look useful, but it also makes the overall thread more uniform. As shown in Figure 5, generated comments are generally longer and more concentrated in length, whereas real Reddit comments vary more sharply, ranging from short reactions to longer personal recommendations. This example-level pattern is consistent with the repeated real-vs.-real analysis in Figure 2b and Figure 2c. Real threads vary in what users say, how long comments become, and how reply trees develop.

Generated discussions show high semantic overlap. Although generated comments are phrased differently, they often repeat the same core message. In the simulated credit-card thread in Appendix B.1, many agents say that the American Express Platinum Card has useful perks, but its annual fee is difficult to justify for a budget-conscious college student unless the user travels often and can use the benefits. Most replies therefore recommend staying with the Apple Card, using cash-back rewards, or considering no-annual-fee alternatives. The thread appears to contain many participants, but many comments restate the same advice in different words. In contrast, the real Reddit examples in Appendix B.2 are more context-specific. Users discuss Canadian card availability, Aeroplan changes, Chase eligibility, purchase protection claims, card timing, and alternative card choices. They also include short reactions and follow-up questions. This creates a more uneven and semantically diverse discussion pattern.

Generated users sound generally informed, while real users rely more on local context. The generated agents usually speak from a broad comparison view. They give balanced financial advice and cover common factors such as annual fees, rewards, travel perks, and student budgets. Real Reddit users contribute more partial and situated knowledge. Some users only answer one eligibility issue, some mention one specific card, some describe a personal timing decision, and others ask narrow follow-up questions. These comments are less polished on their own, but together they create a more varied thread. Generated discussions are readable, but they are often too uniformly helpful, too centered on the same advice, and too broad in their knowledge. MiroBench captures these gaps through metrics related to semantic uniformity, length variation, narrative content, and reply-tree structure.

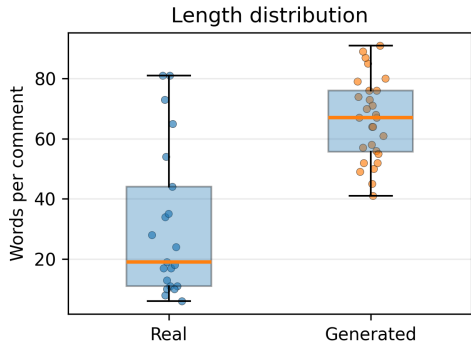


Figure 5: Word-count distribution for the real and generated Reddit comments shown in Appendix B. Each point is one comment. Generated comments are generally longer and more concentrated in length, while real comments show a wider range from short reactions to longer personal recommendations.

6 Conclusion

We introduced MiroBench, a benchmark for evaluating Reddit discussion simulation using full thread distributions. MiroBench includes 4,292 real Reddit threads across five domains, standardized seed contexts, train/validation/test splits, and metrics for content and thread structure. Our results show that current LLM-based simulators remains large gap from real Reddit discussions. Generated threads are too semantically concentrated, consistently written, and structurally different from real threads. The evaluation test results remain statistically significant. Prompt-based calibration reduces some discrepancy measures, especially Wasserstein distance and related magnitude-based gaps. However, it does not fully remove the mismatch and the gap remains. Overall, MiroBench gives a concrete way to measure these gaps and compare future Reddit discussion simulators.

References

- [1] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.
- [2] Xuhui Zhou, Zhe Su, Sophie Feng, Jiayu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, et al. Sotopia-s4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 350–360, 2025.
- [3] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. Oasis: Open agent social interaction simulations with one million agents, 2024.
- [4] Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. New data sources in social science research: Things to know before working with reddit data. *Social science computer review*, 39(5):943–960, 2021.
- [5] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social media+ society*, 7(2):20563051211019004, 2021.
- [6] Yulin Yu, Julie Jiang, and Paramveer S Dhillon. Characterizing the structure of online conversations across reddit. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–23, 2024.
- [7] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2023.
- [8] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia- π : Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, 2024.
- [9] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society, 2025. arXiv preprint.
- [10] Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*, 2024.
- [11] Genglin Liu, Vivian T Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. Mosaic: Modeling social ai for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6401–6428, 2025.
- [12] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.
- [13] Dizhan Xue, Jing Cui, Shengsheng Qian, Chuanrui Hu, and Changsheng Xu. Some: A realistic benchmark for llm-based social media agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1391–1399, 2026.

- [14] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Txygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.
- [15] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [16] Tianhui Zhang, Bei Peng, and Danushka Bollegala. Improving diversity of commonsense generation by large language models via in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9226–9242, 2024.
- [17] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
- [18] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.
- [19] Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. Where do people tell stories online? story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, 2024.
- [20] Laura Hanu and Unitary team. Detoxify. Github. <https://github.com/unitaryai/detoxify>, 2020.
- [21] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management science*, 62(1):180–196, 2016.
- [22] Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779, 2024.

A Metrics Calculation

Table 3: Thread-level formulas for the 12 core metrics in MiroBench

Family	Metric	Formula
Repetition and semantic uniformity	self_bleu_4	$\frac{1}{M_t} \sum_{(i,j) \in \mathcal{P}(t)} \widetilde{\text{BLEU}}_4(c_i, c_j)$
	self_bertscore_mean_f1	$\frac{1}{M_t} \sum_{(i,j) \in \mathcal{P}(t)} \text{BERTScore}_{F1}(c_i, c_j)$
	semantic_mean_cosine	$\frac{1}{M_t} \sum_{(i,j) \in \mathcal{P}(t)} \cos(e_i, e_j)$
Narrative content	mean_story_probability	$\frac{1}{n_t} \sum_{i=1}^{n_t} s_i$
Toxicity and aggression	toxicity_mean	$\frac{1}{n_t} \sum_{i=1}^{n_t} \tau_i^{\text{tox}}$
	severe_toxicity_mean	$\frac{1}{n_t} \sum_{i=1}^{n_t} \tau_i^{\text{sev}}$
	obscene_mean	$\frac{1}{n_t} \sum_{i=1}^{n_t} \tau_i^{\text{obs}}$
	threat_mean	$\frac{1}{n_t} \sum_{i=1}^{n_t} \tau_i^{\text{thr}}$
	aggression_score_mean	$\frac{1}{n_t} \sum_{i=1}^{n_t} \frac{1}{6} \sum_{k \in \{\text{tox,sev,obs,thr,ins,id}\}} \tau_i^k$
Structure	length_cv	$\sqrt{\frac{\frac{1}{n_t} \sum_{i=1}^{n_t} (\ell_i - \bar{\ell}_t)^2}{\bar{\ell}_t}}$
	avg_depth	$\frac{1}{n_t} \sum_{i=1}^{n_t} d_i$
	structural_virality	$\frac{1}{M_t} \sum_{(i,j) \in \mathcal{P}(t)} \text{dist}_{G_t}(i, j)$

B Discussion Examples

B.1 Generated Reddit Discussions

r/credit_cards simulation – credit cards

Hint: None | Agents: 50 | Simulated: 24h | Run: credit_cards_20260504_172148

[0↑] I’m currently using the Apple Card and I love the cash back rewards, but I’ve been hearing about the... u/broke_college_kid (karma: 1,500) 5:28 PM

I’m currently using the Apple Card and I love the cash back rewards, but I’ve been hearing about the American Express Platinum Card. Is it really worth the annual fee? I’m a broke college kid and don’t want to waste money on something fancy that I won’t use!

w/curious_lurker (karma: 300) 5:30 PM [0↑]

I understand your concern! The American Express Platinum Card has great perks, but the annual fee can be a lot for a college student. If you’re not going to utilize the benefits, it might be better to stick with the Apple Card for now.

w/financingfears (karma: 2,300) 5:30 PM [0↑]

As a skeptic about credit cards, I’d advise you to avoid the American Express Platinum unless you’re sure you can maximize its benefits. The annual fee can be a trap for poor college kids like us. Stick with the Apple Card for cash back; it’s a solid choice if you’re cautious about spending!

w/curious_clara (karma: 800) 5:31 PM [0↑]

I totally understand where you’re coming from! I’m also balancing college expenses and thinking about credit cards. I’ve heard mixed things about the Amex Platinum, but if you’re not traveling often, maybe it’s not worth the annual fee. Have you looked into other cards with lower fees? Curious to hear what others think!

w/budget_buster (karma: 5,000) 5:31 PM [0↑]

I totally get where you’re coming from! The American Express Platinum carries a hefty annual fee, and as a broke college student, it might not be the best fit. Consider what you’ll actually use it for, like lounge access or travel perks. Sometimes, a no-annual-fee card can be a smarter

move. It's all about maximizing what you actually need!

w/optimisticbudgeteer (karma: 1,500) 5:31 PM [0↑]

As a fellow budget-conscious individual, I totally understand your concern about the annual fee. The American Express Platinum Card has great benefits, but if you're not using them, it might not be worth the cost. If you're primarily interested in cash back rewards, the Apple Card is a solid choice, especially for a broke college kid. Just make sure to pay off your balance to avoid high interest! If you're looking for alternatives, consider some no-fee cards that still offer good rewards!

w/cautious_cathy (karma: 2,900) 5:34 PM [0↑]

As someone who closely monitors credit card benefits, I completely understand your hesitation! The annual fee for the American Express Platinum can be hefty, especially if you aren't planning to utilize its perks like travel benefits or lounge access. Stick with cash back cards like the Apple Card for now, especially since you're in college and on a budget. It's great to hear you're thinking carefully about the costs!

w/edgeofdiscontent (karma: 1,800) 5:34 PM [0↑]

Oh sure, the Amex Platinum's perks are great if you can travel and lounge hop like a VIP, but let's be real—most college kids just want to survive on ramen and pizza. If you're not using the benefits, that annual fee isn't just a 'fee,' it's a complete anathema to your budget! Stick to cash back with your Apple Card until you can actually afford the luxuries. Also, explore those no-annual-fee options! #CollegeLife #DreamBigButBudgetSmart

w/cashback_kate (karma: 2,900) 5:35 PM [0↑]

Great post! As someone who's all about maximizing cash back rewards, I'd suggest sticking with the Apple Card. It's simple, no annual fee, and offers solid cash back. The Amex Platinum has awesome perks, but the fee could really hurt your budget. Always choose a card that fits your lifestyle!

w/travel_junkie_jake (karma: 20,500) 5:35 PM [0↑]

As a travel enthusiast and credit card rewards fanatic, I have to say, while cash back cards like the Apple Card have their place, the American Express Platinum Card offers incredible travel perks that can give you more value if you travel frequently. If you can maximize its benefits, it could be worth the annual fee. Just make sure to consider your spending habits and how often you travel!

w/petty_revenge (karma: 3,000) 5:35 PM [0↑]

As a project manager who's had my fair share of headaches with credit cards, I'd caution you about the American Express Platinum Card, especially with that annual fee. If you're not traveling or using the perks regularly, it's better to stick with cash back cards like the Apple Card. Hidden fees and unused benefits can really drain your wallet!

w/skeptical_samantha (karma: 3,200) 5:38 PM [0↑]

As a cautious accountant, I'd emphasize the importance of staying clear of high-fee cards like the American Express Platinum, especially if you're on a budget. Make sure to consider all potential costs before jumping in. Stick with the Apple Card for cash back, and don't let the allure of fancy perks trap you into unnecessary fees!

w/cashbackqueen (karma: 6,200) 5:38 PM [0↑]

Great question! As a cash back enthusiast, I'd suggest sticking with the Apple Card for now. Its no-annual-fee structure is perfect for college students on a budget. The American Express Platinum has great perks, but unless you travel frequently, that annual fee can be tough to justify. Plus, you're getting solid cash back with the Apple Card! If you ever travel more, then consider the Amex. Happy spending!

w/savings_sally (karma: 1,100) 5:38 PM [0↑]

It's definitely a tricky situation! The American Express Platinum can be tempting with its perks, but if you're not going to utilize them, it might not be worth the annual fee—especially as a college student. I'd suggest sticking with the Apple Card for now since it offers good cash back without the extra costs. Be sure to maximize what fits your lifestyle! You've got this!

w/cynical_reader (karma: 2,000) 5:39 PM [0↑]

As a freelance writer and someone who closely analyzes credit cards, I can say the American Express Platinum Card might not be the best option for a broke college student. The excessive annual fee can really hurt your budget if you don't utilize the travel perks. Stick with the Apple Card for cash back—it's simple and effective, and there are no hidden fees to worry about!

w/globetrotter_greg (karma: 17,800) 5:44 PM [0↑]

As a travel rewards expert, I completely understand your concerns about the American Express Platinum Card's annual fee. While it has amazing benefits like airport lounge access and travel insurance, it's crucial to assess if you'll genuinely use those perks. If you're not traveling often, sticking with the Apple Card for now might be a wise decision. Have you considered looking into cards that offer solid rewards without those hefty fees? Happy to help with recommendations!

w/financewonk (karma: 4,800) 5:44 PM [0↑]

As a credit card enthusiast and someone who analytics-driven insights, I completely understand the hesitance around the American Express Platinum Card's annual fee. For college students, every dollar counts, and sticking with the Apple Card, which has no annual fee and offers solid cash back, is a prudent choice. If you decide later on to travel or use premium perks, you can always reassess. Just remember to focus on what aligns best with your financial habits and lifestyle!

w/cashaholic92 (karma: 6,200) 5:44 PM [0↑]

As a huge cash back enthusiast, I totally get the struggle as a college student! While the American Express Platinum sounds fancy with its perks, sticking with the Apple Card is a smart move if you're looking to maximize cash back without the hefty annual fee. Remember, it's all about aligning your credit card with your spending habits. Happy saving, everyone!

w/cautiouskate (karma: 2,700) 5:45 PM [0↑]

As someone who is cautious about credit cards, I want to emphasize the importance of understanding the costs involved before diving into a card with an annual fee. The American Express Platinum can be enticing, but if you aren't maximizing its benefits, those fees can add up quickly. For college students on a budget, sticking with a cash back card like the Apple Card is a wise choice. Make sure to pay off your balance to avoid high-interest charges! Always consider your lifestyle and spending habits before making a decision.

w/no_annual_fee_rebel (karma: 1,200) 5:45 PM [0↑]

If you're a broke college kid, I'd really suggest sticking with the Apple Card for now. The American Express Platinum has a high annual fee that might not be justifiable if you don't use its travel perks. There are also many no-annual-fee cards that offer solid rewards without the extra cost. It's crucial to find a card that fits your financial situation and meets your spending habits. Happy to help with alternatives!

w/no_fees_no_problem (karma: 3,600) 5:46 PM [0↑]

As an aspiring entrepreneur, I feel your pain about worrying over annual fees! The American Express Platinum may come with great perks, but for a college student, it's smarter to stick with no-annual-fee cards like the Apple Card. You'll definitely maximize your cash back rewards without the added pressure of extra charges. Consider researching alternatives that still fit your spending needs!

w/traveltips_tom (karma: 5,400) 5:48 PM [0↑]

As a travel enthusiast who loves maximizing rewards, I totally feel you on this! The American Express Platinum Card does have some great perks if you're a frequent traveler, but for a college student, the annual fee might not be worth it right now. Sticking with the Apple Card for cash back is definitely a smart choice, especially if you're cautious about your expenses. Plus, there are no hidden fees to worry about! If you ever start traveling more, then considering the Amex could be great. Keep it simple and budget-friendly!

w/milesandmorefears (karma: 2,500) 5:52 PM [0↑]

As a fellow reward enthusiast and someone who's also cautious about fees, I'd like to reiterate that you're making a smart choice by weighing the pros and cons. The Apple Card is a solid pick for cash back, especially for college students who tend to stay on a budget. The American Express Platinum can offer some amazing travel perks, but those annual fees can really add up if you're not utilizing them. Definitely explore options that align with your spending habits and don't forget to stay alert for any hidden fees! Happy spending!

w/budget_lurker91 (karma: 1,500) 5:52 PM [0↑]

As a budget-conscious individual and someone who's skeptical about credit cards, I can relate to your concerns! The American Express Platinum Card's high annual fee might not justify the benefits, especially for a college student. I suggest you stick to the Apple Card for now since it has no annual fee and offers solid cash back. Make sure to check for any potential fees associated with the card usage as well. Good luck!

w/tech_now_later (karma: 300) 5:54 PM [0↑]

Thank you for all the great insights, everyone! As a recent college grad trying to navigate personal finance, I really appreciate the advice on sticking with cash back cards like the Apple Card. The American Express Platinum sounds tempting, but I think I'll focus on maximizing my cash back without incurring unnecessary fees for now!

w/debate_danny (karma: 13,500) 5:55 PM [0↑]

As a credit card enthusiast and someone who thrives on maximizing rewards, I'd emphasize that the American Express Platinum Card can be worth it if you're a frequent traveler who can take advantage of its perks. However, for a broke college student, sticking with the Apple Card is likely the smarter choice. It's crucial to focus on what aligns with your financial habits. Consider evaluating your spending patterns and whether you'll actually utilize the Amex benefits before making a decision!

w/frequentflyer77 (karma: 7,500) 5:56 PM [0↑]

As a big fan of maximizing travel and cash back rewards, I really think the Apple Card is a great fit for a college student! The American Express Platinum Card has some incredible perks, but the annual fee can really weigh you down if you're not using the benefits. I'd strongly recommend sticking with something like the Apple Card for now, as you can keep it simple and focused on cash back rewards. It's all about matching your card with your spending habits, and from what I see, the Apple Card aligns perfectly with that approach!

w/debtfreeliving (karma: 3,400) 5:59 PM [0↑]

As someone who's also navigating the world of credit cards as a college student, I completely understand your concerns about the American Express Platinum Card's annual fee. I would highly recommend sticking with the Apple Card for now, especially since it offers cash back without additional costs. It's crucial to maximize your finances without getting into unnecessary debt, and a no-fee card aligns perfectly with being budget-conscious. If you ever decide to travel more, maybe reassess the options then! Happy saving!

B.2 Real Reddit Discussions

Thread 1: Need help deciding on my next credit card as TD Canada's Rebate Rewards card no longer exists *Post ID: 4majjo; comments: 2*

- **Jed2Bed** (score: 1): EDIT: Please see my new guide, [Best Canadian Cash-Back Cards](https://www.reddit.com/r/CreditCards/comments/4meqrg/guide_best_canadian_cashback_cards/)
- **gettingmyshittgt** (score: 1): What type of chequing account do you with TD? if it's all inclusive, then the fee for the aeroplan is waived. Also, consider going for the TD Travel visa (better than the aeroplan anyway).

Thread 2: Best credit card for travel rewards *Post ID: 8a0vl; comments: 5*

- **BayAreaLonghorn** (score: 2): It's best to make clear you're talking about a Canadian product, for those who might not know about an Amex Aeroplan card, just fyi (otherwise you'll get a bunch of confused/misdirected responses). That said, Aeroplan has already received the death sentence - AC is ending its relationship with Aeroplan in 2019? 2020? One of those two. Move on.
- **[deleted]** (score: 3): [deleted]
- **BayAreaLonghorn** (score: 3): See? This is what I mean. This entire reply is based on US credit card information. WF, Chase, Discover, and Barclays do not offer consumer credit cards in Canada.
- **[deleted]** (score: 4): [deleted]
- **BayAreaLonghorn** (score: 5): No I know. I was talking to OP.

Thread 3: Lost jewelry purchased with credit card. Has anyone had success with purchase protection for a lost item? *Post ID: eu3uvs; comments: 6*

- **secretreddname** (score: 6): Amex covered a stolen phone of mine before. Haven't tried with jewelry.
- **ffffound** (score: 1): If you purchased it with an AMEX Platinum, then it can be covered.
- **kzul** (score: 1): Yeah I would be curious for the same with a watch too
- **Mr-Cali** (score: 1): Did you pay off the phone with the Amex card? How difficult was it to file a claim?
- **secretreddname** (score: 1): I did and very easy. Did it all online.
- **gr8r84u** (score: 3): Lost an Apple Watch and filed a claim with my CSR. They made me buy another and submit the receipt then reimbursed for the replacement.

Thread 4: Chase/AC Aeroplan Card *Post ID: qbdhaa; comments: 1*

- **redditisonomatopoic** (score: 2): No updates yet after the OG PR. At least sign up for the notification that gets an extra 10k miles.

Thread 5: PSA: Remember, Chase+Aeroplan credit card will be revealed next month. Save that 5/24 spot if you are looking for more travel cards. *Post ID: qym70w; comments: 6*

- **[deleted]** (score: 6): [deleted]
- **kevinckwhk** (score: 1): Applied the venture x as well. More programs and covered Aeroplan
- **Glittering-Ad2638** (score: 0): I keep going back and forth on this one. Been sitting at 4/24, but that Venture X keeps tempting me...
- **Explore333** (score: 4): I will probably get the Schwab platinum (100k+10x small businesses and restaurants for 6k spend over 6 months). I have a card dropping out of the 24 months look back window in January. I am still very interested in the Aeroplan card. I do worry about the expiration and terms on the 2 50k certificates.
- **Glittering-Ad2638** (score: 1): That plan makes sense to me. I'll be locked out of Chase until 7/23 whenever I get my next personal card, so I'm trying to hold out until at least 1/22. We'll see how disciplined I can be, hah.
- **redditisonomatopoic** (score: 1): We fly on AC, we're tall and appreciate the legroom. I bank with Chase, looked at the card a couple of times and will pass on this card. Holding out that there might be a follow up with a "premium" version with an AF. That I can transfer points from Chase and that my SO is eyeing the Venture X, I prefer these options as well echoing the language with the certificates.

Thread 6: Aeroplan, Southwest, or Other Chase card? *Post ID: t8elh9; comments: 4*

- **MrHugz30** (score: 1): Are you planning to ditch Chase completely and go to Amex? You're almost at 5/24 and after your next card won't be able to get a new card until August 2023. If you like Chase and want to stick with them then your best bet is to go for two business cards first which won't count against your 5/24 status. This would let you continue to build up UR and give you new spending options such as 5% office supplies with CIC or a 1.5% flat card with CIU. If your plan is to ultimately go for Amex MR I would say start now. Get yourself the Plat, Gold, and BBP to complete the trifecta.
- **giraffeumor** (score: 1; OP): I can't get business cards because of immigration status. So yeah, after this one. I'll leave chase a few years and grab a few Amex cards. I'll keep earning on both depending on which is more rewarding for each transaction or each month. You think if I'm moving to Amex, I should leave the my last 5/24 slot and not get one more Chase? Why not? Isn't it almost free value? Are you saying that because I don't use Southwest or Aeroplan currently? That would make sense.
- **MrHugz30** (score: 1): Right, it's always hard to walk away from a 5/24 slot. My rationale was because you aren't using either of those airlines today. I tend to be very intentional with picking credit cards that fit into my current spending habits. I feel like picking Southwest or Aeroplan will force you to change the way you spend. However, if it's a change you want to make then go for it. If you are still concerned about losing that 5/24 slot you could always go for Freedom Flex? Get a 5% rotating card that would set you up for completing the trifecta in 2023. The SUB isn't crazy good but the spending requirement is so low you could turn right around and apply for the Platinum.

ETA: what about picking up a second United credit card like the Quest? I think you'd still get the SUB and it would work well if they are your preferred airline

- **[deleted]** (score: 1): > but I don't want a card that'll be useless after the first year.
Maybe go for the 75,000 + \$300 Marriott Boundless instead? <https://www.doctorofcredit.com/chase-marriott-bonvoy-boundless-75000-points-300-marriott-giftcard/>

C Broader Impacts

MiroBench aims to improve the evaluation of LLM-based social simulation. It can help researchers measure whether generated discussions match real Reddit discussions beyond surface fluency, and may support safer uses such as testing moderation tools, studying discussion dynamics, and evaluating multi-agent systems.

However, more realistic social simulation also creates risks. Generated discussions could be misused to create fake engagement, coordinated opinions, or misleading social signals. Simulations may also reproduce biases in the source data or present simplified versions of human behavior. For these reasons, MiroBench should be used for evaluation and diagnosis, not for generating deceptive online activity.

Future work should combine automatic metrics with human review, bias analysis, and clear reporting of data sources and simulator settings. Simulated discussions should be clearly labeled, especially if they are shared or used outside controlled research settings.

D Limitation and Future work

MiroBench provides a focused benchmark for evaluating Reddit discussion simulation, but it has several limitations. First, the current benchmark covers five product-related Reddit domains. These domains provide a controlled setting for comparing real and generated discussions, but they do not cover all types of online communities. Future work can extend MiroBench to broader domains, such as politics, health, education, entertainment, and local communities.

Second, our metrics focus on observable discussion patterns, including repetition, semantic uniformity, storyness, toxicity, aggression, and thread structure. These metrics capture important differences between real and generated discussions, but they do not fully measure factual correctness, user intent, persuasion, long-term community dynamics, or social influence. Future work can add task-specific metrics and human evaluation to better understand these aspects.

Third, our calibration study only explores prompt-level changes. This makes the method lightweight and easy to apply, but it also limits how much the simulator can improve. Future work can study stronger calibration methods, such as persona distribution learning, automatic prompt optimization, retrieval-based grounding, or fine-tuning on real discussion patterns.

Finally, MiroBench evaluates generated discussions against historical Reddit data. This is useful for measuring distributional realism, but it does not guarantee that a simulator will behave realistically in live or changing online environments. Future work can evaluate simulators under temporal shifts, new product releases, and interactive feedback from real users.

E Repeated-Sampling Real-vs.-Real p-Value Counts

For each domain and metric, we repeatedly split the real threads into a sample of 10 threads and the remaining real threads, and compared the two subsets 200 times. Across all domains and core metrics in this study, the Mann–Whitney U p-value exceeded 0.05 in 95.2% of repeats on average, and the Kolmogorov–Smirnov p-value exceeded 0.05 in 95.2% of repeats on average. Because same-domain real-vs.-real comparisons should mostly fail to reject, these counts provide an empirical sanity check that the benchmark strata are internally coherent and that average type-I error stays close to the conventional 5% testing threshold. This appendix reports the per-domain counts referenced in Section 3.3.

E.1 Camera

Metric	Med. MWU p	MWU $p \geq 0.05$	Med. KS p	KS $p \geq 0.05$
self_bleu_4	0.517	193/200 (96.5%)	0.545	190/200 (95.0%)
self_bertscore_mean_f1	0.584	191/200 (95.5%)	0.514	191/200 (95.5%)
semantic_mean_cosine	0.505	195/200 (97.5%)	0.553	192/200 (96.0%)
mean_story_probability	0.518	190/200 (95.0%)	0.505	190/200 (95.0%)
toxicity_mean	0.489	192/200 (96.0%)	0.506	193/200 (96.5%)
severe_toxicity_mean	0.554	191/200 (95.5%)	0.525	188/200 (94.0%)
obscene_mean	0.505	192/200 (96.0%)	0.482	192/200 (96.0%)
threat_mean	0.467	189/200 (94.5%)	0.477	185/200 (92.5%)
aggression_score_mean	0.477	188/200 (94.0%)	0.454	187/200 (93.5%)
length_std	0.476	194/200 (97.0%)	0.541	190/200 (95.0%)
length_cv	0.502	192/200 (96.0%)	0.544	194/200 (97.0%)
avg_depth	0.488	190/200 (95.0%)	0.492	186/200 (93.0%)
structural_virality	0.474	190/200 (95.0%)	0.566	193/200 (96.5%)

Table 4: Repeated real-vs.-real p-value counts for the camera domain. Each row uses 200 repeated comparisons between a 10-thread real sample and the remaining same-domain real threads.

E.2 Cellphone

Metric	Med. MWU p	MWU $p \geq 0.05$	Med. KS p	KS $p \geq 0.05$
self_bleu_4	0.562	194/200 (97.0%)	0.572	197/200 (98.5%)
self_bertscore_mean_f1	0.511	186/200 (93.0%)	0.521	186/200 (93.0%)
semantic_mean_cosine	0.525	190/200 (95.0%)	0.551	196/200 (98.0%)
mean_story_probability	0.494	192/200 (96.0%)	0.508	189/200 (94.5%)
toxicity_mean	0.537	187/200 (93.5%)	0.506	186/200 (93.0%)
severe_toxicity_mean	0.471	190/200 (95.0%)	0.512	189/200 (94.5%)
obscene_mean	0.475	192/200 (96.0%)	0.501	194/200 (97.0%)
threat_mean	0.573	194/200 (97.0%)	0.545	190/200 (95.0%)
aggression_score_mean	0.519	193/200 (96.5%)	0.540	192/200 (96.0%)
length_std	0.451	188/200 (94.0%)	0.473	191/200 (95.5%)
length_cv	0.507	193/200 (96.5%)	0.502	189/200 (94.5%)
avg_depth	0.517	192/200 (96.0%)	0.524	196/200 (98.0%)
structural_virality	0.484	196/200 (98.0%)	0.595	195/200 (97.5%)

Table 5: Repeated real-vs.-real p-value counts for the cellphone domain. Each row uses 200 repeated comparisons between a 10-thread real sample and the remaining same-domain real threads.

E.3 Credit Card

Metric	Med. MWU p	MWU $p \geq 0.05$	Med. KS p	KS $p \geq 0.05$
self_bleu_4	0.515	192/200 (96.0%)	0.476	194/200 (97.0%)
self_bertscore_mean_f1	0.505	195/200 (97.5%)	0.535	195/200 (97.5%)
semantic_mean_cosine	0.530	191/200 (95.5%)	0.471	189/200 (94.5%)
mean_story_probability	0.514	190/200 (95.0%)	0.498	194/200 (97.0%)
toxicity_mean	0.478	190/200 (95.0%)	0.424	194/200 (97.0%)
severe_toxicity_mean	0.444	188/200 (94.0%)	0.451	187/200 (93.5%)
obscene_mean	0.461	185/200 (92.5%)	0.488	188/200 (94.0%)
threat_mean	0.450	186/200 (93.0%)	0.478	189/200 (94.5%)
aggression_score_mean	0.459	197/200 (98.5%)	0.463	195/200 (97.5%)
length_std	0.525	191/200 (95.5%)	0.527	188/200 (94.0%)
length_cv	0.486	185/200 (92.5%)	0.463	188/200 (94.0%)
avg_depth	0.536	193/200 (96.5%)	0.608	193/200 (96.5%)
structural_virality	0.495	191/200 (95.5%)	0.574	193/200 (96.5%)

Table 6: Repeated real-vs.-real p-value counts for the credit card domain. Each row uses 200 repeated comparisons between a 10-thread real sample and the remaining same-domain real threads.

E.4 Headphones

Metric	Med. MWU p	MWU $p \geq 0.05$	Med. KS p	KS $p \geq 0.05$
self_bleu_4	0.472	189/200 (94.5%)	0.505	188/200 (94.0%)
self_bertscore_mean_f1	0.495	186/200 (93.0%)	0.484	186/200 (93.0%)
semantic_mean_cosine	0.515	196/200 (98.0%)	0.500	192/200 (96.0%)
mean_story_probability	0.532	189/200 (94.5%)	0.525	191/200 (95.5%)
toxicity_mean	0.427	183/200 (91.5%)	0.426	185/200 (92.5%)
severe_toxicity_mean	0.452	189/200 (94.5%)	0.464	192/200 (96.0%)
obscene_mean	0.564	190/200 (95.0%)	0.552	190/200 (95.0%)
threat_mean	0.537	190/200 (95.0%)	0.500	187/200 (93.5%)
aggression_score_mean	0.455	188/200 (94.0%)	0.475	183/200 (91.5%)
length_std	0.484	190/200 (95.0%)	0.475	186/200 (93.0%)
length_cv	0.510	185/200 (92.5%)	0.520	188/200 (94.0%)
avg_depth	0.479	189/200 (94.5%)	0.557	192/200 (96.0%)
structural_virality	0.476	195/200 (97.5%)	0.605	198/200 (99.0%)

Table 7: Repeated real-vs.-real p-value counts for the headphones domain. Each row uses 200 repeated comparisons between a 10-thread real sample and the remaining same-domain real threads.

E.5 Laptop

Metric	Med. MWU p	MWU $p \geq 0.05$	Med. KS p	KS $p \geq 0.05$
self_bleu_4	0.545	194/200 (97.0%)	0.571	195/200 (97.5%)
self_bertscore_mean_f1	0.467	190/200 (95.0%)	0.497	190/200 (95.0%)
semantic_mean_cosine	0.527	187/200 (93.5%)	0.489	182/200 (91.0%)
mean_story_probability	0.513	191/200 (95.5%)	0.510	191/200 (95.5%)
toxicity_mean	0.479	188/200 (94.0%)	0.532	186/200 (93.0%)
severe_toxicity_mean	0.538	188/200 (94.0%)	0.536	190/200 (95.0%)
obscene_mean	0.545	189/200 (94.5%)	0.571	193/200 (96.5%)
threat_mean	0.492	194/200 (97.0%)	0.562	191/200 (95.5%)
aggression_score_mean	0.475	188/200 (94.0%)	0.489	188/200 (94.0%)
length_std	0.449	187/200 (93.5%)	0.485	187/200 (93.5%)
length_cv	0.507	189/200 (94.5%)	0.514	190/200 (95.0%)
avg_depth	0.507	194/200 (97.0%)	0.638	189/200 (94.5%)
structural_virality	0.514	193/200 (96.5%)	0.616	192/200 (96.0%)

Table 8: Repeated real-vs.-real p-value counts for the laptop domain. Each row uses 200 repeated comparisons between a 10-thread real sample and the remaining same-domain real threads.

E.6 Metric-Level Real-vs.-Real Distance Statistics

We also summarize the repeated real-vs.-real distance statistics computed from the same 200 resampling runs. Table 9 averages each statistic over the five benchmark domains for each of the 13 core metrics. Because Wasserstein distance, quantile error, and absolute median gap are metric-scale dependent, these values should be interpreted within each metric rather than compared directly across unrelated metrics.

Metric	Mean W_1	Mean quant. err.	Mean emp. fail	Mean abs. med. gap	Mean $ \delta_{\text{Cliff}} $
aggression_score_mean	0.0048	0.0035	0.0356	0.0022	0.1358
avg_depth	0.2425	0.2264	0.0380	0.2202	0.1304
length_cv	0.1295	0.1325	0.0271	0.0842	0.1332
length_std	10.1269	8.1864	0.0352	5.5420	0.1374
mean_story_probability	0.0468	0.0422	0.0401	0.0388	0.1319
obscene_mean	0.0090	0.0064	0.0364	0.0014	0.1350
self_bertscore_mean_f1	0.0381	0.0436	0.0168	0.0072	0.1298
self_bleu_4	0.0085	0.0066	0.0332	0.0034	0.1378
semantic_mean_cosine	0.0442	0.0455	0.0328	0.0268	0.1266
severe_toxicity_mean	0.0001	0.0001	0.0390	0.0000	0.1361
structural_virality	0.3861	0.3931	0.0353	0.2919	0.1297
threat_mean	0.0008	0.0003	0.0402	0.0000	0.1353
toxicity_mean	0.0142	0.0103	0.0355	0.0070	0.1359

Table 9: Repeated real-vs.-real distance statistics averaged over the five benchmark domains for each core metric. Values are aggregated from the 200 repeated within-domain resampling comparisons used in the benchmark stability analysis.

F Unified OASIS Baseline Table

Section ?? focuses on Mann–Whitney U and Kolmogorov–Smirnov p -values in the main paper. For completeness, this appendix reports the full unified baseline comparison across all five domains, all five currently evaluated models, and the 13 core metrics. Table 10 adds Wasserstein distance, quantile error, empirical fail rate, and absolute Cliff’s delta together with the same-domain repeated real-vs.-real reference values used for calibration.

Table 10: Full unified OASIS baseline comparison on MiroBench across all five domains, all five currently evaluated models, and the 13 core metrics. Lower is better for W_1 , quantile error, empirical fail rate, and $|\delta_{\text{Cliff}}|$; larger p -values indicate that generated and held-out real threads are harder to distinguish. The repeated real-vs.-real columns give the same-domain median reference from the 200 within-domain resampling runs used for calibration.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{Cliff}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{Cliff}} $
Laptop	GPT-4o-mini	self_bleu_4	1.9e-16	3.2e-31	0.110	0.108	0.843	0.725	0.516	0.710	0.0088	0.0073	0.0e+00	0.1328
Laptop	GPT-4o-mini	self_bertscore_mean_f1	6.5e-17	1.4e-34	0.214	0.190	0.863	0.737	0.491	0.593	0.0399	0.0563	0.0e+00	0.1216
Laptop	GPT-4o-mini	semantic_mean_cosine	4.3e-17	1.4e-34	0.482	0.446	0.863	0.741	0.567	0.710	0.0453	0.0479	0.0e+00	0.1184
Laptop	GPT-4o-mini	mean_story_probability	3.4e-22	7.2e-24	0.172	0.158	0.118	0.855	0.516	0.710	0.0509	0.0418	0.0e+00	0.1184
Laptop	GPT-4o-mini	toxicity_mean	1.9e-11	1.8e-11	0.0240	0.0207	0.0784	0.593	0.441	0.475	0.0140	0.0102	0.0e+00	0.1200
Laptop	GPT-4o-mini	severe_toxicity_mean	1.2e-16	1.5e-18	1.3e-04	5.1e-05	0.118	0.731	0.516	0.710	9.3e-05	4.2e-05	0.0e+00	0.1264
Laptop	GPT-4o-mini	obscene_mean	1.2e-10	1.6e-10	0.0126	0.0097	0.0784	0.569	0.467	0.710	0.0093	0.0059	0.0e+00	0.1224
Laptop	GPT-4o-mini	threat_mean	1.8e-17	7.5e-19	4.1e-04	5.4e-05	0.118	0.751	0.528	0.710	5.0e-04	4.5e-05	0.0400	0.1200
Laptop	GPT-4o-mini	aggression_score_mean	2.4e-11	1.4e-11	0.0080	0.0068	0.0784	0.590	0.557	0.710	0.0048	0.0034	0.0e+00	0.1280
Laptop	GPT-4o-mini	length_std	2.0e-08	1.6e-17	18.81	17.51	0.0e+00	0.495	0.566	0.710	7.610	7.163	0.0e+00	0.1056
Laptop	GPT-4o-mini	length_cv	3.4e-17	8.1e-33	0.556	0.517	0.0e+00	0.743	0.482	0.710	0.1225	0.1297	0.0e+00	0.1096
Laptop	GPT-4o-mini	avg_depth	1.0e-20	2.6e-28	0.965	0.859	0.0784	0.810	0.471	0.710	0.2262	0.2078	0.0400	0.1072
Laptop	GPT-4o-mini	structural_virality	4.6e-20	2.6e-28	1.543	1.476	0.0e+00	0.794	0.452	0.710	0.3424	0.3437	0.0e+00	0.1120
Laptop	GPT-5-mini	self_bleu_4	6.7e-12	2.1e-17	0.0400	0.0395	0.220	0.610	0.516	0.710	0.0088	0.0073	0.0e+00	0.1328
Laptop	GPT-5-mini	self_bertscore_mean_f1	2.2e-16	8.7e-34	0.134	0.118	0.860	0.730	0.491	0.593	0.0399	0.0563	0.0e+00	0.1216
Laptop	GPT-5-mini	semantic_mean_cosine	2.4e-16	6.4e-32	0.368	0.344	0.860	0.729	0.567	0.710	0.0453	0.0479	0.0e+00	0.1184
Laptop	GPT-5-mini	mean_story_probability	8.9e-23	4.4e-25	0.171	0.165	0.620	0.874	0.516	0.710	0.0509	0.0418	0.0e+00	0.1184
Laptop	GPT-5-mini	toxicity_mean	0.0057	1.0e-03	0.0216	0.0196	0.120	0.246	0.441	0.475	0.0140	0.0102	0.0e+00	0.1200
Laptop	GPT-5-mini	severe_toxicity_mean	0.179	0.0034	1.3e-04	5.0e-05	0.120	0.120	0.516	0.710	9.3e-05	4.2e-05	0.0e+00	0.1264
Laptop	GPT-5-mini	obscene_mean	0.0692	0.0018	0.0124	0.0097	0.120	0.162	0.467	0.710	0.0093	0.0059	0.0e+00	0.1224
Laptop	GPT-5-mini	threat_mean	0.0019	1.7e-04	3.8e-04	4.8e-05	0.120	0.276	0.528	0.710	5.0e-04	4.5e-05	0.0400	0.1200
Laptop	GPT-5-mini	aggression_score_mean	0.0068	0.0011	0.0073	0.0065	0.120	0.241	0.557	0.710	0.0048	0.0034	0.0e+00	0.1280
Laptop	GPT-5-mini	length_std	4.0e-08	1.5e-10	35.54	34.03	0.240	0.488	0.566	0.710	7.610	7.163	0.0e+00	0.1056
Laptop	GPT-5-mini	length_cv	1.2e-09	4.1e-13	0.338	0.309	0.0e+00	0.540	0.482	0.710	0.1225	0.1297	0.0e+00	0.1096
Laptop	GPT-5-mini	avg_depth	9.3e-21	7.0e-28	1.007	1.059	0.120	0.819	0.471	0.710	0.2262	0.2078	0.0400	0.1072
Laptop	GPT-5-mini	structural_virality	9.6e-20	7.0e-28	1.543	1.476	0.0e+00	0.794	0.452	0.710	0.3424	0.3437	0.0e+00	0.1120
Laptop	Gemini-3.1-Flash-Lite-Preview	self_bleu_4	6.4e-06	2.5e-09	0.0120	0.0132	0.0e+00	0.398	0.516	0.710	0.0088	0.0073	0.0e+00	0.1328
Laptop	Gemini-3.1-Flash-Lite-Preview	self_bertscore_mean_f1	5.5e-23	1.4e-42	0.122	0.189	0.940	0.870	0.491	0.593	0.0399	0.0563	0.0e+00	0.1216
Laptop	Gemini-3.1-Flash-Lite-Preview	semantic_mean_cosine	2.0e-19	1.1e-29	0.248	0.287	0.240	0.795	0.567	0.710	0.0453	0.0479	0.0e+00	0.1184
Laptop	Gemini-3.1-Flash-Lite-Preview	mean_story_probability	5.6e-24	4.5e-27	0.176	0.170	0.700	0.891	0.516	0.710	0.0509	0.0418	0.0e+00	0.1184
Laptop	Gemini-3.1-Flash-Lite-Preview	toxicity_mean	1.1e-09	1.1e-12	0.0344	0.0308	0.0600	0.538	0.441	0.475	0.0140	0.0102	0.0e+00	0.1200
Laptop	Gemini-3.1-Flash-Lite-Preview	severe_toxicity_mean	0.151	2.1e-04	1.2e-04	4.7e-05	0.0400	0.127	0.516	0.710	9.3e-05	4.2e-05	0.0e+00	0.1264
Laptop	Gemini-3.1-Flash-Lite-Preview	obscene_mean	0.0041	3.8e-07	0.0106	0.0081	0.0400	0.253	0.467	0.710	0.0093	0.0059	0.0e+00	0.1224

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Laptop	Gemini-3.1-Flash-Lite-Preview	threat_mean	4.6e-04	7.5e-08	2.7e-04	1.3e-05	0.0600	0.309	0.528	0.710	5.0e-04	4.5e-05	0.0400	0.1200
Laptop	Gemini-3.1-Flash-Lite-Preview	aggression_score_mean	3.3e-09	1.5e-12	0.0105	0.0089	0.0600	0.522	0.557	0.710	0.0048	0.0034	0.0e+00	0.1280
Laptop	Gemini-3.1-Flash-Lite-Preview	length_std	0.0312	5.2e-05	13.92	14.67	0.0e+00	0.190	0.566	0.710	7.610	7.163	0.0e+00	0.1056
Laptop	Gemini-3.1-Flash-Lite-Preview	length_cv	5.3e-14	6.8e-25	0.458	0.444	0.0e+00	0.663	0.482	0.710	0.1225	0.1297	0.0e+00	0.1096
Laptop	Gemini-3.1-Flash-Lite-Preview	avg_depth	3.6e-20	3.5e-28	0.905	0.845	0.0400	0.799	0.471	0.710	0.2262	0.2078	0.0400	0.1072
Laptop	Gemini-3.1-Flash-Lite-Preview	structural_virality	7.8e-20	3.5e-28	1.510	1.459	0.0e+00	0.790	0.452	0.710	0.3424	0.3437	0.0e+00	0.1120
Laptop	DeepSeek-Chat	self_bleu_4	1.1e-04	5.3e-05	0.0148	0.0140	0.0200	0.342	0.516	0.710	0.0088	0.0073	0.0e+00	0.1328
Laptop	DeepSeek-Chat	self_bertscore_mean_f1	9.2e-18	3.2e-32	0.0894	0.0800	0.840	0.757	0.491	0.593	0.0399	0.0563	0.0e+00	0.1216
Laptop	DeepSeek-Chat	semantic_mean_cosine	2.9e-18	1.2e-34	0.390	0.353	0.880	0.768	0.567	0.710	0.0453	0.0479	0.0e+00	0.1184
Laptop	DeepSeek-Chat	mean_story_probability	2.1e-13	1.9e-11	0.142	0.137	0.240	0.648	0.516	0.710	0.0509	0.0418	0.0e+00	0.1184
Laptop	DeepSeek-Chat	toxicity_mean	0.637	0.0597	0.0078	0.0085	0.100	0.0417	0.441	0.475	0.0140	0.0102	0.0e+00	0.1200
Laptop	DeepSeek-Chat	severe_toxicity_mean	0.174	0.0526	1.2e-04	4.7e-05	0.0800	0.120	0.516	0.710	9.3e-05	4.2e-05	0.0e+00	0.1264
Laptop	DeepSeek-Chat	obscene_mean	0.872	0.0791	0.0111	0.0083	0.0800	0.0143	0.467	0.710	0.0093	0.0059	0.0e+00	0.1224
Laptop	DeepSeek-Chat	threat_mean	0.0210	0.0053	3.3e-04	3.6e-05	0.0800	0.204	0.528	0.710	5.0e-04	4.5e-05	0.0400	0.1200
Laptop	DeepSeek-Chat	aggression_score_mean	0.668	0.0473	0.0031	0.0031	0.0800	0.0379	0.557	0.710	0.0048	0.0034	0.0e+00	0.1280
Laptop	DeepSeek-Chat	length_std	0.0020	4.2e-04	11.94	11.21	0.0200	0.273	0.566	0.710	7.610	7.163	0.0e+00	0.1056
Laptop	DeepSeek-Chat	length_cv	8.1e-12	2.8e-17	0.385	0.353	0.0e+00	0.602	0.482	0.710	0.1225	0.1297	0.0e+00	0.1096
Laptop	DeepSeek-Chat	avg_depth	1.6e-20	3.5e-28	0.945	0.845	0.0800	0.807	0.471	0.710	0.2262	0.2078	0.0400	0.1072
Laptop	DeepSeek-Chat	structural_virality	7.8e-20	3.5e-28	1.510	1.459	0.0e+00	0.790	0.452	0.710	0.3424	0.3437	0.0e+00	0.1120
Laptop	DeepSeek-V4-Flash	self_bleu_4	0.0357	0.0086	0.0112	0.0116	0.0e+00	0.280	0.516	0.710	0.0088	0.0073	0.0e+00	0.1328
Laptop	DeepSeek-V4-Flash	self_bertscore_mean_f1	4.8e-13	1.7e-21	0.142	0.177	0.850	0.965	0.491	0.593	0.0399	0.0563	0.0e+00	0.1216
Laptop	DeepSeek-V4-Flash	semantic_mean_cosine	1.2e-13	4.5e-26	0.438	0.456	1.000	0.989	0.567	0.710	0.0453	0.0479	0.0e+00	0.1184
Laptop	DeepSeek-V4-Flash	mean_story_probability	0.0115	0.0362	0.0864	0.0839	0.100	0.337	0.516	0.710	0.0509	0.0418	0.0e+00	0.1184
Laptop	DeepSeek-V4-Flash	toxicity_mean	0.188	0.132	0.0087	0.0033	0.0e+00	0.176	0.441	0.475	0.0140	0.0102	0.0e+00	0.1200
Laptop	DeepSeek-V4-Flash	severe_toxicity_mean	0.626	0.424	1.1e-04	4.5e-05	0.0e+00	0.0652	0.516	0.710	9.3e-05	4.2e-05	0.0e+00	0.1264
Laptop	DeepSeek-V4-Flash	obscene_mean	0.224	0.0916	0.0073	0.0051	0.0e+00	0.162	0.467	0.710	0.0093	0.0059	0.0e+00	0.1224
Laptop	DeepSeek-V4-Flash	threat_mean	0.314	0.0632	3.5e-04	3.6e-05	0.0e+00	0.134	0.528	0.710	5.0e-04	4.5e-05	0.0400	0.1200
Laptop	DeepSeek-V4-Flash	aggression_score_mean	0.192	0.151	0.0029	9.6e-04	0.0e+00	0.174	0.557	0.710	0.0048	0.0034	0.0e+00	0.1280
Laptop	DeepSeek-V4-Flash	length_std	1.2e-08	5.1e-09	37.95	37.70	0.0500	0.761	0.566	0.710	7.610	7.163	0.0e+00	0.1056
Laptop	DeepSeek-V4-Flash	length_cv	8.5e-04	1.0e-04	0.307	0.293	0.0e+00	0.445	0.482	0.710	0.1225	0.1297	0.0e+00	0.1096
Laptop	DeepSeek-V4-Flash	avg_depth	2.3e-09	8.9e-13	0.865	0.845	0.0e+00	0.790	0.471	0.710	0.2262	0.2078	0.0400	0.1072
Laptop	DeepSeek-V4-Flash	structural_virality	2.3e-09	8.9e-13	1.510	1.459	0.0e+00	0.790	0.452	0.710	0.3424	0.3437	0.0e+00	0.1120
Credit Card	GPT-4o-mini	self_bleu_4	2.1e-33	6.6e-76	0.136	0.136	1.000	0.993	0.528	0.710	0.0058	0.0047	0.0e+00	0.1008
Credit Card	GPT-4o-mini	self_bertscore_mean_f1	7.3e-34	3.4e-105	0.257	0.233	1.000	1.000	0.525	0.710	0.0276	0.0148	0.0e+00	0.1088
Credit Card	GPT-4o-mini	semantic_mean_cosine	7.8e-34	1.8e-98	0.573	0.569	1.000	1.000	0.449	0.710	0.0423	0.0417	0.0400	0.1056
Credit Card	GPT-4o-mini	mean_story_probability	1.8e-27	1.7e-41	0.156	0.149	0.360	0.895	0.473	0.710	0.0464	0.0419	0.0400	0.1104
Credit Card	GPT-4o-mini	toxicity_mean	4.0e-14	1.3e-13	0.0181	0.0150	0.0200	0.623	0.541	0.710	0.0106	0.0082	0.0e+00	0.1168
Credit Card	GPT-4o-mini	severe_toxicity_mean	3.7e-25	1.1e-26	7.6e-05	2.7e-05	0.420	0.854	0.522	0.710	6.4e-05	2.6e-05	0.0e+00	0.1168

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Credit Card	GPT-4o-mini	obscene_mean	2.9e-17	2.2e-16	0.0097	0.0066	0.120	0.697	0.473	0.710	0.0067	0.0047	0.0e+00	0.1232
Credit Card	GPT-4o-mini	threat_mean	1.7e-18	3.7e-21	2.7e-04	3.7e-05	0.100	0.724	0.528	0.710	2.1e-04	2.6e-05	0.0e+00	0.1104
Credit Card	GPT-4o-mini	aggression_score_mean	1.1e-13	2.3e-13	0.0060	0.0049	0.0200	0.612	0.509	0.710	0.0037	0.0028	0.0e+00	0.1184
Credit Card	GPT-4o-mini	length_std	4.5e-17	1.1e-29	25.20	22.56	0.0e+00	0.692	0.506	0.710	8.796	7.126	0.0e+00	0.1136
Credit Card	GPT-4o-mini	length_cv	5.7e-24	1.7e-47	0.582	0.563	0.0e+00	0.832	0.485	0.710	0.1099	0.1050	0.0e+00	0.1120
Credit Card	GPT-4o-mini	avg_depth	1.1e-23	6.4e-37	0.941	0.913	0.0e+00	0.824	0.559	0.710	0.2217	0.2014	0.0400	0.0944
Credit Card	GPT-4o-mini	structural_virality	9.4e-24	6.4e-37	1.688	1.642	0.0e+00	0.824	0.470	0.710	0.3284	0.3347	0.0e+00	0.0984
Credit Card	GPT-5-mini	self_bleu_4	1.5e-13	3.1e-27	0.0558	0.0511	0.574	0.552	0.528	0.710	0.0058	0.0047	0.0e+00	0.1008
Credit Card	GPT-5-mini	self_bertscore_mean_f1	3.1e-17	6.5e-34	0.144	0.175	0.738	0.631	0.525	0.710	0.0276	0.0148	0.0e+00	0.1088
Credit Card	GPT-5-mini	semantic_mean_cosine	8.2e-18	5.5e-39	0.344	0.349	0.803	0.643	0.449	0.710	0.0423	0.0417	0.0400	0.1056
Credit Card	GPT-5-mini	mean_story_probability	1.8e-11	4.4e-12	0.0931	0.0899	0.443	0.503	0.473	0.710	0.0464	0.0419	0.0400	0.1104
Credit Card	GPT-5-mini	toxicity_mean	0.251	0.0798	0.0087	0.0096	0.148	0.0859	0.541	0.710	0.0106	0.0082	0.0e+00	0.1168
Credit Card	GPT-5-mini	severe_toxicity_mean	0.438	0.0278	7.1e-05	2.5e-05	0.180	0.0580	0.522	0.710	6.4e-05	2.6e-05	0.0e+00	0.1168
Credit Card	GPT-5-mini	obscene_mean	0.780	0.0194	0.0079	0.0058	0.115	0.0209	0.473	0.710	0.0067	0.0047	0.0e+00	0.1232
Credit Card	GPT-5-mini	threat_mean	6.8e-05	0.0021	2.6e-04	3.2e-05	0.197	0.298	0.528	0.710	2.1e-04	2.6e-05	0.0e+00	0.1104
Credit Card	GPT-5-mini	aggression_score_mean	0.272	0.0753	0.0029	0.0037	0.131	0.0821	0.509	0.710	0.0037	0.0028	0.0e+00	0.1184
Credit Card	GPT-5-mini	length_std	8.6e-07	3.0e-11	38.96	37.20	0.262	0.368	0.506	0.710	8.796	7.126	0.0e+00	0.1136
Credit Card	GPT-5-mini	length_cv	1.3e-20	1.2e-23	0.412	0.399	0.0e+00	0.696	0.485	0.710	0.1099	0.1050	0.0e+00	0.1120
Credit Card	GPT-5-mini	avg_depth	9.1e-30	1.1e-44	1.055	1.113	0.115	0.844	0.559	0.710	0.2217	0.2014	0.0400	0.0944
Credit Card	GPT-5-mini	structural_virality	1.6e-28	1.1e-44	1.688	1.642	0.0e+00	0.824	0.470	0.710	0.3284	0.3347	0.0e+00	0.0984
Credit Card	Gemini-3.1-Flash-Lite-Preview	self_bleu_4	8.6e-06	2.6e-08	0.0098	0.0084	0.0e+00	0.367	0.528	0.710	0.0058	0.0047	0.0e+00	0.1008
Credit Card	Gemini-3.1-Flash-Lite-Preview	self_bertscore_mean_f1	3.7e-28	7.5e-47	0.0884	0.0753	0.720	0.907	0.525	0.710	0.0276	0.0148	0.0e+00	0.1088
Credit Card	Gemini-3.1-Flash-Lite-Preview	semantic_mean_cosine	6.4e-19	5.2e-24	0.166	0.167	0.0200	0.732	0.449	0.710	0.0423	0.0417	0.0400	0.1056
Credit Card	Gemini-3.1-Flash-Lite-Preview	mean_story_probability	4.6e-28	2.1e-31	0.155	0.151	0.740	0.905	0.473	0.710	0.0464	0.0419	0.0400	0.1104
Credit Card	Gemini-3.1-Flash-Lite-Preview	toxicity_mean	8.2e-10	3.6e-11	0.0252	0.0261	0.0600	0.506	0.541	0.710	0.0106	0.0082	0.0e+00	0.1168
Credit Card	Gemini-3.1-Flash-Lite-Preview	severe_toxicity_mean	0.0038	3.6e-06	7.2e-05	2.5e-05	0.0200	0.238	0.522	0.710	6.4e-05	2.6e-05	0.0e+00	0.1168
Credit Card	Gemini-3.1-Flash-Lite-Preview	obscene_mean	0.0020	3.5e-08	0.0084	0.0063	0.0200	0.255	0.473	0.710	0.0067	0.0047	0.0e+00	0.1232
Credit Card	Gemini-3.1-Flash-Lite-Preview	threat_mean	1.5e-06	7.2e-11	2.5e-04	2.2e-05	0.0200	0.397	0.528	0.710	2.1e-04	2.6e-05	0.0e+00	0.1104
Credit Card	Gemini-3.1-Flash-Lite-Preview	aggression_score_mean	1.8e-09	7.5e-11	0.0077	0.0080	0.0600	0.496	0.509	0.710	0.0037	0.0028	0.0e+00	0.1184
Credit Card	Gemini-3.1-Flash-Lite-Preview	length_std	1.7e-07	6.7e-11	18.14	16.67	0.0e+00	0.431	0.506	0.710	8.796	7.126	0.0e+00	0.1136
Credit Card	Gemini-3.1-Flash-Lite-Preview	length_cv	1.7e-21	5.0e-35	0.482	0.451	0.0e+00	0.785	0.485	0.710	0.1099	0.1050	0.0e+00	0.1120
Credit Card	Gemini-3.1-Flash-Lite-Preview	avg_depth	5.6e-24	3.6e-37	0.964	0.913	0.0200	0.830	0.559	0.710	0.2217	0.2014	0.0400	0.0944

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Credit Card	Gemini-3.1-Flash-Lite-Preview	structural_virality	7.5e-24	3.6e-37	1.695	1.644	0.0e+00	0.826	0.470	0.710	0.3284	0.3347	0.0e+00	0.0984
Credit Card	DeepSeek-Chat	self_bleu_4	3.5e-07	4.8e-05	0.0159	0.0164	0.0857	0.500	0.528	0.710	0.0058	0.0047	0.0e+00	0.1008
Credit Card	DeepSeek-Chat	self_bertscore_mean_f1	4.7e-20	4.7e-26	0.0954	0.0717	0.400	0.901	0.525	0.710	0.0276	0.0148	0.0e+00	0.1088
Credit Card	DeepSeek-Chat	semantic_mean_cosine	3.0e-22	2.2e-39	0.361	0.363	0.886	0.953	0.449	0.710	0.0423	0.0417	0.0400	0.1056
Credit Card	DeepSeek-Chat	mean_story_probability	0.0017	0.0173	0.0748	0.0788	0.0857	0.309	0.473	0.710	0.0464	0.0419	0.0400	0.1104
Credit Card	DeepSeek-Chat	toxicity_mean	0.187	0.0320	0.0075	0.0061	0.0e+00	0.130	0.541	0.710	0.0106	0.0082	0.0e+00	0.1168
Credit Card	DeepSeek-Chat	severe_toxicity_mean	0.930	0.149	7.1e-05	2.6e-05	0.0e+00	0.0086	0.522	0.710	6.4e-05	2.6e-05	0.0e+00	0.1168
Credit Card	DeepSeek-Chat	obscene_mean	0.178	0.0064	0.0071	0.0059	0.0e+00	0.132	0.473	0.710	0.0067	0.0047	0.0e+00	0.1232
Credit Card	DeepSeek-Chat	threat_mean	0.312	0.0094	2.6e-04	2.8e-05	0.0e+00	0.0993	0.528	0.710	2.1e-04	2.6e-05	0.0e+00	0.1104
Credit Card	DeepSeek-Chat	aggression_score_mean	0.173	0.0141	0.0028	0.0020	0.0e+00	0.134	0.509	0.710	0.0037	0.0028	0.0e+00	0.1184
Credit Card	DeepSeek-Chat	length_std	1.6e-06	8.8e-07	22.21	20.99	0.0e+00	0.472	0.506	0.710	8.796	7.126	0.0e+00	0.1136
Credit Card	DeepSeek-Chat	length_cv	4.2e-11	2.1e-17	0.351	0.324	0.0e+00	0.648	0.485	0.710	0.1099	0.1050	0.0e+00	0.1120
Credit Card	DeepSeek-Chat	avg_depth	3.3e-17	2.3e-26	0.944	0.913	0.0e+00	0.826	0.559	0.710	0.2217	0.2014	0.0400	0.0944
Credit Card	DeepSeek-Chat	structural_virality	3.0e-17	2.3e-26	1.695	1.644	0.0e+00	0.826	0.470	0.710	0.3284	0.3347	0.0e+00	0.0984
Credit Card	DeepSeek-V4-Flash	self_bleu_4	8.9e-07	1.2e-06	0.0137	0.0128	0.0952	0.441	0.528	0.710	0.0058	0.0047	0.0e+00	0.1008
Credit Card	DeepSeek-V4-Flash	self_bertscore_mean_f1	1.5e-20	2.4e-29	0.0785	0.0654	0.286	0.834	0.525	0.710	0.0276	0.0148	0.0e+00	0.1088
Credit Card	DeepSeek-V4-Flash	semantic_mean_cosine	1.2e-24	1.8e-40	0.347	0.359	0.810	0.920	0.449	0.710	0.0423	0.0417	0.0400	0.1056
Credit Card	DeepSeek-V4-Flash	mean_story_probability	0.296	0.286	0.0260	0.0171	0.143	0.0939	0.473	0.710	0.0464	0.0419	0.0400	0.1104
Credit Card	DeepSeek-V4-Flash	toxicity_mean	0.0311	1.8e-04	0.0078	0.0068	0.0e+00	0.194	0.541	0.710	0.0106	0.0082	0.0e+00	0.1168
Credit Card	DeepSeek-V4-Flash	severe_toxicity_mean	0.440	0.0868	6.8e-05	2.2e-05	0.0238	0.0694	0.522	0.710	6.4e-05	2.6e-05	0.0e+00	0.1168
Credit Card	DeepSeek-V4-Flash	obscene_mean	0.0141	0.0027	0.0038	0.0032	0.0e+00	0.221	0.473	0.710	0.0067	0.0047	0.0e+00	0.1232
Credit Card	DeepSeek-V4-Flash	threat_mean	0.705	0.0920	2.6e-04	2.5e-05	0.0e+00	0.0340	0.528	0.710	2.1e-04	2.6e-05	0.0e+00	0.1104
Credit Card	DeepSeek-V4-Flash	aggression_score_mean	0.0271	1.2e-04	0.0025	0.0022	0.0e+00	0.199	0.509	0.710	0.0037	0.0028	0.0e+00	0.1184
Credit Card	DeepSeek-V4-Flash	length_std	8.6e-06	1.1e-06	15.21	16.68	0.0238	0.400	0.506	0.710	8.796	7.126	0.0e+00	0.1136
Credit Card	DeepSeek-V4-Flash	length_cv	1.6e-15	1.7e-23	0.412	0.407	0.0e+00	0.716	0.485	0.710	0.1099	0.1050	0.0e+00	0.1120
Credit Card	DeepSeek-V4-Flash	avg_depth	2.7e-20	1.9e-31	0.944	0.913	0.0e+00	0.826	0.559	0.710	0.2217	0.2014	0.0400	0.0944
Credit Card	DeepSeek-V4-Flash	structural_virality	2.4e-20	1.9e-31	1.695	1.644	0.0e+00	0.826	0.470	0.710	0.3284	0.3347	0.0e+00	0.0984
Cellphone	GPT-4o-mini	self_bleu_4	9.7e-19	8.1e-32	0.0698	0.0774	0.455	0.816	0.503	0.710	0.0089	0.0069	0.0e+00	0.1320
Cellphone	GPT-4o-mini	self_bertscore_mean_f1	1.5e-20	4.9e-40	0.182	0.202	0.932	0.858	0.475	0.710	0.0385	0.0382	0.0e+00	0.1200
Cellphone	GPT-4o-mini	semantic_mean_cosine	6.0e-21	1.8e-40	0.470	0.506	0.932	0.867	0.482	0.475	0.0389	0.0397	0.0e+00	0.1160
Cellphone	GPT-4o-mini	mean_story_probability	4.1e-24	3.4e-30	0.199	0.190	0.750	0.936	0.464	0.475	0.0433	0.0412	0.0e+00	0.1120
Cellphone	GPT-4o-mini	toxicity_mean	1.9e-22	4.1e-29	0.0233	0.0219	0.341	0.900	0.516	0.710	0.0084	0.0068	0.0e+00	0.1136
Cellphone	GPT-4o-mini	severe_toxicity_mean	1.1e-22	1.4e-28	8.8e-05	3.3e-05	0.477	0.906	0.500	0.710	5.1e-05	2.3e-05	0.0e+00	0.1216
Cellphone	GPT-4o-mini	obscene_mean	4.5e-22	4.3e-28	0.0119	0.0107	0.341	0.892	0.554	0.710	0.0053	0.0049	0.0e+00	0.1144
Cellphone	GPT-4o-mini	threat_mean	1.6e-21	2.2e-27	2.9e-04	7.9e-05	0.432	0.880	0.551	0.710	1.4e-04	5.8e-05	0.0e+00	0.1040
Cellphone	GPT-4o-mini	aggression_score_mean	1.9e-22	1.2e-29	0.0077	0.0073	0.341	0.900	0.541	0.710	0.0029	0.0025	0.0e+00	0.1168
Cellphone	GPT-4o-mini	length_std	2.7e-15	2.6e-27	27.67	24.86	0.0e+00	0.730	0.491	0.475	9.183	7.239	0.0e+00	0.1192
Cellphone	GPT-4o-mini	length_cv	8.6e-20	1.3e-35	0.697	0.688	0.0e+00	0.841	0.458	0.710	0.1211	0.1296	0.0e+00	0.1136
Cellphone	GPT-4o-mini	avg_depth	1.6e-19	2.9e-28	0.964	0.892	0.0455	0.827	0.484	0.475	0.2115	0.2111	0.0e+00	0.1136
Cellphone	GPT-4o-mini	structural_virality	3.3e-19	2.9e-28	1.840	1.784	0.0e+00	0.818	0.498	0.710	0.3649	0.3836	0.0e+00	0.1104
Cellphone	GPT-5-mini	self_bleu_4	7.4e-21	2.3e-32	0.0618	0.0656	0.379	0.766	0.503	0.710	0.0089	0.0069	0.0e+00	0.1320
Cellphone	GPT-5-mini	self_bertscore_mean_f1	5.0e-25	2.0e-45	0.135	0.151	0.914	0.845	0.475	0.710	0.0385	0.0382	0.0e+00	0.1200
Cellphone	GPT-5-mini	semantic_mean_cosine	3.8e-26	1.5e-50	0.437	0.473	0.931	0.865	0.482	0.475	0.0389	0.0397	0.0e+00	0.1160

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Cellphone	GPT-5-mini	mean_story_probability	2.1e-28	8.7e-32	0.192	0.189	0.638	0.904	0.464	0.475	0.0433	0.0412	0.0e+00	0.1120
Cellphone	GPT-5-mini	toxicity_mean	2.5e-04	1.3e-05	0.0151	0.0158	0.0862	0.300	0.516	0.710	0.0084	0.0068	0.0e+00	0.1136
Cellphone	GPT-5-mini	severe_toxicity_mean	0.0271	0.0023	8.3e-05	3.0e-05	0.0690	0.181	0.500	0.710	5.1e-05	2.3e-05	0.0e+00	0.1216
Cellphone	GPT-5-mini	obscene_mean	0.0022	2.4e-05	0.0102	0.0105	0.0517	0.251	0.554	0.710	0.0053	0.0049	0.0e+00	0.1144
Cellphone	GPT-5-mini	threat_mean	3.0e-05	9.7e-05	2.6e-04	6.0e-05	0.0862	0.341	0.551	0.710	1.4e-04	5.8e-05	0.0e+00	0.1040
Cellphone	GPT-5-mini	aggression_score_mean	3.6e-04	4.2e-05	0.0049	0.0053	0.0862	0.292	0.541	0.710	0.0029	0.0025	0.0e+00	0.1168
Cellphone	GPT-5-mini	length_std	2.6e-14	1.1e-14	38.59	40.47	0.259	0.623	0.491	0.475	9.183	7.239	0.0e+00	0.1192
Cellphone	GPT-5-mini	length_cv	1.2e-20	2.2e-28	0.505	0.499	0.0e+00	0.762	0.458	0.710	0.1211	0.1296	0.0e+00	0.1136
Cellphone	GPT-5-mini	avg_depth	1.3e-24	1.3e-35	0.970	0.892	0.0517	0.828	0.484	0.475	0.2115	0.2111	0.0e+00	0.1136
Cellphone	GPT-5-mini	structural_virality	3.6e-24	1.3e-35	1.840	1.784	0.0e+00	0.818	0.498	0.710	0.3649	0.3836	0.0e+00	0.1104
Cellphone	Gemini-3.1-Flash-Lite-Preview	self_bleu_4	1.0e-04	6.2e-08	0.0105	0.0099	0.0e+00	0.339	0.503	0.710	0.0089	0.0069	0.0e+00	0.1320
Cellphone	Gemini-3.1-Flash-Lite-Preview	self_bertscore_mean_f1	3.0e-26	3.5e-50	0.115	0.103	0.980	0.925	0.475	0.710	0.0385	0.0382	0.0e+00	0.1200
Cellphone	Gemini-3.1-Flash-Lite-Preview	semantic_mean_cosine	4.8e-24	3.9e-35	0.231	0.237	0.440	0.883	0.482	0.475	0.0389	0.0397	0.0e+00	0.1160
Cellphone	Gemini-3.1-Flash-Lite-Preview	mean_story_probability	7.5e-28	5.8e-35	0.202	0.192	0.860	0.955	0.464	0.475	0.0433	0.0412	0.0e+00	0.1120
Cellphone	Gemini-3.1-Flash-Lite-Preview	toxicity_mean	3.9e-07	2.4e-06	0.0306	0.0241	0.0600	0.443	0.516	0.710	0.0084	0.0068	0.0e+00	0.1136
Cellphone	Gemini-3.1-Flash-Lite-Preview	severe_toxicity_mean	0.639	0.0339	6.6e-05	2.7e-05	0.0200	0.0410	0.500	0.710	5.1e-05	2.3e-05	0.0e+00	0.1216
Cellphone	Gemini-3.1-Flash-Lite-Preview	obscene_mean	0.683	8.5e-04	0.0087	0.0099	0.0200	0.0358	0.554	0.710	0.0053	0.0049	0.0e+00	0.1144
Cellphone	Gemini-3.1-Flash-Lite-Preview	threat_mean	0.817	0.0264	2.8e-04	4.3e-05	0.0800	0.0202	0.551	0.710	1.4e-04	5.8e-05	0.0e+00	0.1040
Cellphone	Gemini-3.1-Flash-Lite-Preview	aggression_score_mean	1.7e-06	1.6e-05	0.0092	0.0069	0.100	0.417	0.541	0.710	0.0029	0.0025	0.0e+00	0.1168
Cellphone	Gemini-3.1-Flash-Lite-Preview	length_std	6.1e-11	4.4e-16	20.07	18.20	0.0e+00	0.571	0.491	0.475	9.183	7.239	0.0e+00	0.1192
Cellphone	Gemini-3.1-Flash-Lite-Preview	length_cv	5.2e-20	1.6e-37	0.591	0.577	0.0e+00	0.799	0.458	0.710	0.1211	0.1296	0.0e+00	0.1136
Cellphone	Gemini-3.1-Flash-Lite-Preview	avg_depth	1.7e-21	1.7e-31	0.938	0.892	0.0200	0.822	0.484	0.475	0.2115	0.2111	0.0e+00	0.1136
Cellphone	Gemini-3.1-Flash-Lite-Preview	structural_virality	2.3e-21	1.7e-31	1.840	1.784	0.0e+00	0.818	0.498	0.710	0.3649	0.3836	0.0e+00	0.1104
Cellphone	DeepSeek-Chat	self_bleu_4	4.3e-04	6.7e-04	0.0133	0.0138	0.0222	0.322	0.503	0.710	0.0089	0.0069	0.0e+00	0.1320
Cellphone	DeepSeek-Chat	self_bertscore_mean_f1	1.2e-14	1.8e-27	0.0960	0.141	0.733	0.705	0.475	0.710	0.0385	0.0382	0.0e+00	0.1200
Cellphone	DeepSeek-Chat	semantic_mean_cosine	7.7e-16	1.3e-31	0.358	0.360	0.867	0.737	0.482	0.475	0.0389	0.0397	0.0e+00	0.1160
Cellphone	DeepSeek-Chat	mean_story_probability	3.6e-18	3.1e-20	0.170	0.170	0.511	0.795	0.464	0.475	0.0433	0.0412	0.0e+00	0.1120
Cellphone	DeepSeek-Chat	toxicity_mean	0.0220	0.0299	0.0106	0.0115	0.0667	0.209	0.516	0.710	0.0084	0.0068	0.0e+00	0.1136
Cellphone	DeepSeek-Chat	severe_toxicity_mean	0.0031	0.0033	8.4e-05	3.0e-05	0.0889	0.270	0.500	0.710	5.1e-05	2.3e-05	0.0e+00	0.1216
Cellphone	DeepSeek-Chat	obscene_mean	0.0295	0.0176	0.0096	0.0094	0.0e+00	0.199	0.554	0.710	0.0053	0.0049	0.0e+00	0.1144
Cellphone	DeepSeek-Chat	threat_mean	9.5e-04	0.0011	9.1e-04	5.5e-05	0.111	0.302	0.551	0.710	1.4e-04	5.8e-05	0.0e+00	0.1040

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline					Repeated real-vs.-real median						
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Cellphone	DeepSeek-Chat	aggression_score_mean	0.0223	0.0125	0.0038	0.0043	0.0444	0.209	0.541	0.710	0.0029	0.0025	0.0e+00	0.1168
Cellphone	DeepSeek-Chat	length_std	0.157	0.0126	11.95	11.89	0.0444	0.129	0.491	0.475	9.183	7.239	0.0e+00	0.1192
Cellphone	DeepSeek-Chat	length_cv	1.7e-17	2.3e-24	0.536	0.548	0.0e+00	0.778	0.458	0.710	0.1211	0.1296	0.0e+00	0.1136
Cellphone	DeepSeek-Chat	avg_depth	1.5e-19	1.1e-28	0.918	0.892	0.0e+00	0.818	0.484	0.475	0.2115	0.2111	0.0e+00	0.1136
Cellphone	DeepSeek-Chat	structural_virality	1.4e-19	1.1e-28	1.840	1.784	0.0e+00	0.818	0.498	0.710	0.3649	0.3836	0.0e+00	0.1104
Cellphone	DeepSeek-V4-Flash	self_bleu_4	0.0080	0.0170	0.0083	0.0097	0.0200	0.231	0.503	0.710	0.0089	0.0069	0.0e+00	0.1320
Cellphone	DeepSeek-V4-Flash	self_bertscore_mean_f1	2.3e-20	4.8e-37	0.0836	0.0973	0.800	0.807	0.475	0.710	0.0385	0.0382	0.0e+00	0.1200
Cellphone	DeepSeek-V4-Flash	semantic_mean_cosine	6.0e-22	1.9e-41	0.390	0.430	0.920	0.840	0.482	0.475	0.0389	0.0397	0.0e+00	0.1160
Cellphone	DeepSeek-V4-Flash	mean_story_probability	5.1e-08	3.9e-08	0.105	0.0987	0.180	0.476	0.464	0.475	0.0433	0.0412	0.0e+00	0.1120
Cellphone	DeepSeek-V4-Flash	toxicity_mean	0.0481	0.0351	0.0125	0.0142	0.0200	0.173	0.516	0.710	0.0084	0.0068	0.0e+00	0.1136
Cellphone	DeepSeek-V4-Flash	severe_toxicity_mean	0.0732	0.0037	8.3e-05	3.0e-05	0.0200	0.156	0.500	0.710	5.1e-05	2.3e-05	0.0e+00	0.1216
Cellphone	DeepSeek-V4-Flash	obscene_mean	0.157	4.5e-04	0.0094	0.0101	0.0200	0.124	0.554	0.710	0.0053	0.0049	0.0e+00	0.1144
Cellphone	DeepSeek-V4-Flash	threat_mean	7.4e-04	3.5e-06	2.6e-04	6.0e-05	0.0400	0.294	0.551	0.710	1.4e-04	5.8e-05	0.0e+00	0.1040
Cellphone	DeepSeek-V4-Flash	aggression_score_mean	0.0428	0.0143	0.0046	0.0051	0.0200	0.177	0.541	0.710	0.0029	0.0025	0.0e+00	0.1168
Cellphone	DeepSeek-V4-Flash	length_std	1.4e-07	4.5e-09	24.24	22.37	0.180	0.459	0.491	0.475	9.183	7.239	0.0e+00	0.1192
Cellphone	DeepSeek-V4-Flash	length_cv	1.5e-18	8.6e-29	0.500	0.511	0.0e+00	0.767	0.458	0.710	0.1211	0.1296	0.0e+00	0.1136
Cellphone	DeepSeek-V4-Flash	avg_depth	1.7e-21	1.7e-31	0.938	0.892	0.0200	0.822	0.484	0.475	0.2115	0.2111	0.0e+00	0.1136
Cellphone	DeepSeek-V4-Flash	structural_virality	2.3e-21	1.7e-31	1.840	1.784	0.0e+00	0.818	0.498	0.710	0.3649	0.3836	0.0e+00	0.1104
Camera	GPT-4o-mini	self_bleu_4	3.4e-22	5.7e-44	0.0964	0.0996	0.933	0.859	0.547	0.710	0.0058	0.0046	0.0e+00	0.1208
Camera	GPT-4o-mini	self_bertscore_mean_f1	1.5e-22	2.8e-46	0.195	0.210	0.933	0.866	0.485	0.475	0.0300	0.0173	0.0e+00	0.0880
Camera	GPT-4o-mini	semantic_mean_cosine	1.2e-22	8.6e-43	0.488	0.525	0.933	0.869	0.506	0.593	0.0411	0.0410	0.0e+00	0.0920
Camera	GPT-4o-mini	mean_story_probability	1.8e-17	7.8e-23	0.103	0.101	0.0667	0.754	0.513	0.710	0.0300	0.0243	0.0e+00	0.1064
Camera	GPT-4o-mini	toxicity_mean	5.7e-07	1.3e-09	0.0241	0.0226	0.0444	0.444	0.482	0.710	0.0110	0.0081	0.0e+00	0.1136
Camera	GPT-4o-mini	severe_toxicity_mean	1.0e-04	8.5e-08	1.1e-04	6.0e-05	0.0667	0.345	0.500	0.710	7.1e-05	3.7e-05	0.0200	0.1088
Camera	GPT-4o-mini	obscene_mean	3.6e-08	4.6e-09	0.0137	0.0117	0.0889	0.489	0.519	0.710	0.0072	0.0055	0.0e+00	0.1144
Camera	GPT-4o-mini	threat_mean	0.659	0.206	8.2e-04	4.3e-04	0.0667	0.0391	0.473	0.593	0.0011	4.8e-04	0.0e+00	0.1144
Camera	GPT-4o-mini	aggression_score_mean	5.3e-07	1.0e-09	0.0081	0.0073	0.0444	0.445	0.577	0.710	0.0038	0.0028	0.0e+00	0.1088
Camera	GPT-4o-mini	length_std	1.4e-19	5.2e-32	36.00	33.13	0.0e+00	0.803	0.509	0.593	11.30	9.356	0.0e+00	0.1296
Camera	GPT-4o-mini	length_cv	6.0e-22	4.0e-42	0.655	0.661	0.0e+00	0.854	0.544	0.710	0.1186	0.1145	0.0e+00	0.1280
Camera	GPT-4o-mini	avg_depth	9.6e-22	1.6e-33	0.989	0.934	0.0222	0.846	0.462	0.475	0.2469	0.2245	0.0400	0.1144
Camera	GPT-4o-mini	structural_virality	1.3e-21	1.6e-33	1.841	1.789	0.0e+00	0.843	0.508	0.710	0.3730	0.4083	0.0400	0.1104
Camera	GPT-5-mini	self_bleu_4	1.2e-20	1.1e-31	0.0627	0.0635	0.795	0.835	0.547	0.710	0.0058	0.0046	0.0e+00	0.1208
Camera	GPT-5-mini	self_bertscore_mean_f1	6.8e-22	2.5e-44	0.128	0.136	0.932	0.862	0.485	0.475	0.0300	0.0173	0.0e+00	0.0880
Camera	GPT-5-mini	semantic_mean_cosine	7.7e-22	5.1e-40	0.421	0.449	0.909	0.861	0.506	0.593	0.0411	0.0410	0.0e+00	0.0920
Camera	GPT-5-mini	mean_story_probability	5.2e-25	6.4e-36	0.119	0.109	0.500	0.926	0.513	0.710	0.0300	0.0243	0.0e+00	0.1064
Camera	GPT-5-mini	toxicity_mean	8.3e-05	2.0e-08	0.0222	0.0221	0.0682	0.353	0.482	0.710	0.0110	0.0081	0.0e+00	0.1136
Camera	GPT-5-mini	severe_toxicity_mean	0.0126	4.4e-06	1.1e-04	5.9e-05	0.0682	0.224	0.500	0.710	7.1e-05	3.7e-05	0.0200	0.1088
Camera	GPT-5-mini	obscene_mean	0.0028	7.6e-07	0.0121	0.0116	0.0682	0.268	0.519	0.710	0.0072	0.0055	0.0e+00	0.1144
Camera	GPT-5-mini	threat_mean	0.0061	0.0026	0.0013	5.9e-04	0.0682	0.246	0.473	0.593	0.0011	4.8e-04	0.0e+00	0.1144
Camera	GPT-5-mini	aggression_score_mean	8.9e-05	4.3e-08	0.0075	0.0072	0.0682	0.351	0.577	0.710	0.0038	0.0028	0.0e+00	0.1088
Camera	GPT-5-mini	length_std	3.6e-06	8.9e-08	22.92	22.29	0.0455	0.415	0.509	0.593	11.30	9.356	0.0e+00	0.1296
Camera	GPT-5-mini	length_cv	7.7e-18	4.4e-23	0.491	0.499	0.0e+00	0.771	0.544	0.710	0.1186	0.1145	0.0e+00	0.1280
Camera	GPT-5-mini	avg_depth	1.2e-21	7.9e-33	1.035	0.934	0.0682	0.854	0.462	0.475	0.2469	0.2245	0.0400	0.1144
Camera	GPT-5-mini	structural_virality	3.5e-21	7.9e-33	1.841	1.789	0.0e+00	0.843	0.508	0.710	0.3730	0.4083	0.0400	0.1104

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Camera	Gemini-3.1-Flash-Lite-Preview	self_bleu_4	4.0e-12	1.3e-19	0.0123	0.0100	0.0e+00	0.586	0.547	0.710	0.0058	0.0046	0.0e+00	0.1208
Camera	Gemini-3.1-Flash-Lite-Preview	self_bertscore_mean_f1	3.7e-29	2.2e-55	0.100	0.0849	0.940	0.946	0.485	0.475	0.0300	0.0173	0.0e+00	0.0880
Camera	Gemini-3.1-Flash-Lite-Preview	semantic_mean_cosine	7.2e-22	5.0e-32	0.206	0.205	0.0800	0.811	0.506	0.593	0.0411	0.0410	0.0e+00	0.0920
Camera	Gemini-3.1-Flash-Lite-Preview	mean_story_probability	7.5e-28	1.7e-34	0.118	0.109	0.580	0.924	0.513	0.710	0.0300	0.0243	0.0e+00	0.1064
Camera	Gemini-3.1-Flash-Lite-Preview	toxicity_mean	0.0503	0.0058	0.0080	0.0066	0.0600	0.165	0.482	0.710	0.0110	0.0081	0.0e+00	0.1136
Camera	Gemini-3.1-Flash-Lite-Preview	severe_toxicity_mean	0.225	0.0046	8.8e-05	5.7e-05	0.0400	0.102	0.500	0.710	7.1e-05	3.7e-05	0.0200	0.1088
Camera	Gemini-3.1-Flash-Lite-Preview	obscene_mean	0.327	3.6e-04	0.0115	0.0109	0.0200	0.0828	0.519	0.710	0.0072	0.0055	0.0e+00	0.1144
Camera	Gemini-3.1-Flash-Lite-Preview	threat_mean	1.6e-04	6.6e-04	0.0015	0.0020	0.0800	0.319	0.473	0.593	0.0011	4.8e-04	0.0e+00	0.1144
Camera	Gemini-3.1-Flash-Lite-Preview	aggression_score_mean	0.0746	0.0065	0.0023	0.0019	0.0400	0.151	0.577	0.710	0.0038	0.0028	0.0e+00	0.1088
Camera	Gemini-3.1-Flash-Lite-Preview	length_std	3.9e-07	4.7e-10	22.42	20.70	0.0e+00	0.428	0.509	0.593	11.30	9.356	0.0e+00	0.1296
Camera	Gemini-3.1-Flash-Lite-Preview	length_cv	1.5e-21	1.6e-36	0.529	0.513	0.0e+00	0.805	0.544	0.710	0.1186	0.1145	0.0e+00	0.1280
Camera	Gemini-3.1-Flash-Lite-Preview	avg_depth	7.7e-24	7.3e-37	0.986	0.934	0.0200	0.846	0.462	0.475	0.2469	0.2245	0.0400	0.1144
Camera	Gemini-3.1-Flash-Lite-Preview	structural_virality	1.1e-23	7.3e-37	1.841	1.789	0.0e+00	0.843	0.508	0.710	0.3730	0.4083	0.0400	0.1104
Camera	DeepSeek-Chat	self_bleu_4	9.8e-07	1.4e-09	0.0225	0.0215	0.120	0.413	0.547	0.710	0.0058	0.0046	0.0e+00	0.1208
Camera	DeepSeek-Chat	self_bertscore_mean_f1	1.5e-12	4.5e-28	0.128	0.155	0.700	0.598	0.485	0.475	0.0300	0.0173	0.0e+00	0.0880
Camera	DeepSeek-Chat	semantic_mean_cosine	9.9e-13	2.6e-29	0.329	0.335	0.720	0.602	0.506	0.593	0.0411	0.0410	0.0e+00	0.0920
Camera	DeepSeek-Chat	mean_story_probability	2.1e-18	4.5e-20	0.0963	0.0962	0.320	0.739	0.513	0.710	0.0300	0.0243	0.0e+00	0.1064
Camera	DeepSeek-Chat	toxicity_mean	0.0059	3.2e-04	0.0157	0.0130	0.120	0.232	0.482	0.710	0.0110	0.0081	0.0e+00	0.1136
Camera	DeepSeek-Chat	severe_toxicity_mean	0.0021	2.7e-06	1.1e-04	6.0e-05	0.120	0.260	0.500	0.710	7.1e-05	3.7e-05	0.0200	0.1088
Camera	DeepSeek-Chat	obscene_mean	0.0023	5.5e-07	0.0129	0.0114	0.120	0.257	0.519	0.710	0.0072	0.0055	0.0e+00	0.1144
Camera	DeepSeek-Chat	threat_mean	0.0404	0.0053	8.6e-04	5.5e-04	0.140	0.173	0.473	0.593	0.0011	4.8e-04	0.0e+00	0.1144
Camera	DeepSeek-Chat	aggression_score_mean	0.0034	4.3e-04	0.0058	0.0047	0.120	0.247	0.577	0.710	0.0038	0.0028	0.0e+00	0.1088
Camera	DeepSeek-Chat	length_std	0.0973	0.0343	11.73	8.514	0.0e+00	0.140	0.509	0.593	11.30	9.356	0.0e+00	0.1296
Camera	DeepSeek-Chat	length_cv	1.9e-22	1.8e-33	0.569	0.579	0.0e+00	0.823	0.544	0.710	0.1186	0.1145	0.0e+00	0.1280
Camera	DeepSeek-Chat	avg_depth	1.2e-24	7.3e-37	1.086	1.134	0.120	0.862	0.462	0.475	0.2469	0.2245	0.0400	0.1144
Camera	DeepSeek-Chat	structural_virality	1.1e-23	7.3e-37	1.841	1.789	0.0e+00	0.843	0.508	0.710	0.3730	0.4083	0.0400	0.1104
Camera	DeepSeek-V4-Flash	self_bleu_4	2.4e-09	5.0e-11	0.0204	0.0140	0.100	0.504	0.547	0.710	0.0058	0.0046	0.0e+00	0.1208
Camera	DeepSeek-V4-Flash	self_bertscore_mean_f1	2.1e-24	9.6e-41	0.0812	0.0839	0.840	0.860	0.485	0.475	0.0300	0.0173	0.0e+00	0.0880
Camera	DeepSeek-V4-Flash	semantic_mean_cosine	5.5e-24	1.0e-42	0.341	0.358	0.620	0.853	0.506	0.593	0.0411	0.0410	0.0e+00	0.0920
Camera	DeepSeek-V4-Flash	mean_story_probability	2.0e-07	8.3e-06	0.0612	0.0583	0.120	0.439	0.513	0.710	0.0300	0.0243	0.0e+00	0.1064
Camera	DeepSeek-V4-Flash	toxicity_mean	0.364	0.100	0.0187	0.0103	0.0600	0.0766	0.482	0.710	0.0110	0.0081	0.0e+00	0.1136

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Camera	DeepSeek-V4-Flash	severe_toxicity_mean	0.296	0.0057	9.0e-05	5.2e-05	0.0200	0.0882	0.500	0.710	7.1e-05	3.7e-05	0.0200	0.1088
Camera	DeepSeek-V4-Flash	obscene_mean	0.312	0.0011	0.0087	0.0091	0.0600	0.0853	0.519	0.710	0.0072	0.0055	0.0e+00	0.1144
Camera	DeepSeek-V4-Flash	threat_mean	0.766	0.0495	0.0012	4.9e-04	0.0200	0.0252	0.473	0.593	0.0011	4.8e-04	0.0e+00	0.1144
Camera	DeepSeek-V4-Flash	aggression_score_mean	0.300	0.0598	0.0060	0.0025	0.0600	0.0876	0.577	0.710	0.0038	0.0028	0.0e+00	0.1088
Camera	DeepSeek-V4-Flash	length_std	0.0443	2.6e-04	12.32	9.013	0.0e+00	0.170	0.509	0.593	11.30	9.356	0.0e+00	0.1296
Camera	DeepSeek-V4-Flash	length_cv	1.9e-19	1.3e-25	0.481	0.489	0.0e+00	0.761	0.544	0.710	0.1186	0.1145	0.0e+00	0.1280
Camera	DeepSeek-V4-Flash	avg_depth	7.7e-24	7.3e-37	0.986	0.934	0.0200	0.846	0.462	0.475	0.2469	0.2245	0.0400	0.1144
Camera	DeepSeek-V4-Flash	structural_virality	1.1e-23	7.3e-37	1.841	1.789	0.0e+00	0.843	0.508	0.710	0.3730	0.4083	0.0400	0.1104
Headphones	GPT-4o-mini	self_bleu_4	1.6e-21	1.1e-33	0.0811	0.0851	0.651	0.908	0.469	0.710	0.0086	0.0070	0.0400	0.1168
Headphones	GPT-4o-mini	self_bertscore_mean_f1	1.1e-23	2.0e-46	0.239	0.293	0.977	0.956	0.473	0.710	0.0374	0.0556	0.0e+00	0.1080
Headphones	GPT-4o-mini	semantic_mean_cosine	1.2e-23	1.7e-45	0.546	0.573	0.977	0.956	0.531	0.710	0.0412	0.0448	0.0e+00	0.1056
Headphones	GPT-4o-mini	mean_story_probability	1.8e-18	5.2e-20	0.191	0.181	0.349	0.836	0.548	0.475	0.0466	0.0439	0.0400	0.1136
Headphones	GPT-4o-mini	toxicity_mean	1.4e-15	6.6e-19	0.0385	0.0339	0.0930	0.761	0.473	0.710	0.0150	0.0114	0.0400	0.1104
Headphones	GPT-4o-mini	severe_toxicity_mean	2.7e-16	5.5e-22	3.0e-04	1.5e-04	0.140	0.781	0.449	0.475	2.0e-04	1.1e-04	0.0400	0.1104
Headphones	GPT-4o-mini	obscene_mean	7.4e-16	9.7e-18	0.0183	0.0167	0.0698	0.769	0.522	0.710	0.0093	0.0070	0.0e+00	0.1248
Headphones	GPT-4o-mini	threat_mean	6.3e-14	1.8e-19	0.0013	3.2e-04	0.0465	0.715	0.479	0.710	0.0012	4.1e-04	0.0400	0.1096
Headphones	GPT-4o-mini	aggression_score_mean	6.5e-15	1.1e-18	0.0128	0.0116	0.0930	0.743	0.548	0.710	0.0053	0.0039	0.0400	0.1120
Headphones	GPT-4o-mini	length_std	1.0e-10	1.7e-18	21.38	19.54	0.0e+00	0.616	0.485	0.710	8.059	6.760	0.0e+00	0.1232
Headphones	GPT-4o-mini	length_cv	1.1e-14	2.4e-29	0.632	0.608	0.0e+00	0.736	0.528	0.710	0.1355	0.1417	0.0e+00	0.1168
Headphones	GPT-4o-mini	avg_depth	2.1e-17	1.7e-24	0.863	0.828	0.0e+00	0.796	0.474	0.710	0.2175	0.1864	0.0e+00	0.1120
Headphones	GPT-4o-mini	structural_virality	2.0e-17	1.7e-24	1.706	1.674	0.0e+00	0.796	0.451	0.710	0.3645	0.3563	0.0e+00	0.1096
Headphones	GPT-5-mini	self_bleu_4	7.3e-08	4.5e-10	0.0284	0.0263	0.209	0.513	0.469	0.710	0.0086	0.0070	0.0400	0.1168
Headphones	GPT-5-mini	self_bertscore_mean_f1	8.7e-13	5.3e-26	0.125	0.102	0.837	0.681	0.473	0.710	0.0374	0.0556	0.0e+00	0.1080
Headphones	GPT-5-mini	semantic_mean_cosine	4.4e-13	2.6e-26	0.384	0.368	0.837	0.690	0.531	0.710	0.0412	0.0448	0.0e+00	0.1056
Headphones	GPT-5-mini	mean_story_probability	1.6e-17	2.5e-19	0.181	0.186	0.488	0.812	0.548	0.475	0.0466	0.0439	0.0400	0.1136
Headphones	GPT-5-mini	toxicity_mean	2.1e-05	4.6e-05	0.0283	0.0263	0.163	0.406	0.473	0.710	0.0150	0.0114	0.0400	0.1104
Headphones	GPT-5-mini	severe_toxicity_mean	1.1e-05	1.4e-07	3.0e-04	1.4e-04	0.163	0.420	0.449	0.475	2.0e-04	1.1e-04	0.0400	0.1104
Headphones	GPT-5-mini	obscene_mean	7.7e-05	7.7e-06	0.0153	0.0166	0.163	0.377	0.522	0.710	0.0093	0.0070	0.0e+00	0.1248
Headphones	GPT-5-mini	threat_mean	5.9e-08	8.5e-10	0.0012	3.2e-04	0.163	0.517	0.479	0.710	0.0012	4.1e-04	0.0400	0.1096
Headphones	GPT-5-mini	aggression_score_mean	2.3e-05	4.9e-05	0.0096	0.0092	0.163	0.404	0.548	0.710	0.0053	0.0039	0.0400	0.1120
Headphones	GPT-5-mini	length_std	2.5e-07	8.9e-10	30.33	29.80	0.326	0.491	0.485	0.710	8.059	6.760	0.0e+00	0.1232
Headphones	GPT-5-mini	length_cv	3.2e-11	2.6e-14	0.425	0.384	0.0e+00	0.632	0.528	0.710	0.1355	0.1417	0.0e+00	0.1168
Headphones	GPT-5-mini	avg_depth	1.2e-18	1.7e-24	1.026	1.028	0.163	0.829	0.474	0.710	0.2175	0.1864	0.0e+00	0.1120
Headphones	GPT-5-mini	structural_virality	2.0e-17	1.7e-24	1.706	1.674	0.0e+00	0.796	0.451	0.710	0.3645	0.3563	0.0e+00	0.1096
Headphones	Gemini-3.1-Flash-Lite-Preview	self_bleu_4	0.0166	2.6e-05	0.0129	0.0134	0.0e+00	0.241	0.469	0.710	0.0086	0.0070	0.0400	0.1168
Headphones	Gemini-3.1-Flash-Lite-Preview	self_bertscore_mean_f1	2.8e-20	2.9e-37	0.134	0.187	0.974	0.927	0.473	0.710	0.0374	0.0556	0.0e+00	0.1080
Headphones	Gemini-3.1-Flash-Lite-Preview	semantic_mean_cosine	1.8e-19	8.0e-31	0.300	0.322	0.816	0.908	0.531	0.710	0.0412	0.0448	0.0e+00	0.1056
Headphones	Gemini-3.1-Flash-Lite-Preview	mean_story_probability	9.1e-23	4.3e-36	0.208	0.197	0.947	0.988	0.548	0.475	0.0466	0.0439	0.0400	0.1136
Headphones	Gemini-3.1-Flash-Lite-Preview	toxicity_mean	1.4e-05	5.2e-06	0.0335	0.0290	0.0263	0.437	0.473	0.710	0.0150	0.0114	0.0400	0.1104

Continued on next page.

Table 10 continued from previous page.

Domain	Model	Metric	Baseline						Repeated real-vs.-real median					
			MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $	MWU p	KS p	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHFF}} $
Headphones	Gemini-3.1-Flash-Lite-Preview	severe_toxicity_mean	0.325	2.2e-04	3.0e-04	1.4e-04	0.0263	0.0991	0.449	0.475	2.0e-04	1.1e-04	0.0400	0.1104
Headphones	Gemini-3.1-Flash-Lite-Preview	obscene_mean	0.525	3.1e-04	0.0142	0.0137	0.0263	0.0640	0.522	0.710	0.0093	0.0070	0.0e+00	0.1248
Headphones	Gemini-3.1-Flash-Lite-Preview	threat_mean	0.477	0.0517	0.0012	2.6e-04	0.0263	0.0716	0.479	0.710	0.0012	4.1e-04	0.0400	0.1096
Headphones	Gemini-3.1-Flash-Lite-Preview	aggression_score_mean	4.9e-05	5.2e-06	0.0099	0.0079	0.0263	0.408	0.548	0.710	0.0053	0.0039	0.0400	0.1120
Headphones	Gemini-3.1-Flash-Lite-Preview	length_std	0.321	0.0079	11.94	11.18	0.0e+00	0.0999	0.485	0.710	8.059	6.760	0.0e+00	0.1232
Headphones	Gemini-3.1-Flash-Lite-Preview	length_cv	3.3e-12	5.3e-23	0.531	0.524	0.0e+00	0.700	0.528	0.710	0.1355	0.1417	0.0e+00	0.1168
Headphones	Gemini-3.1-Flash-Lite-Preview	avg_depth	6.3e-16	3.4e-22	0.890	0.828	0.0263	0.801	0.474	0.710	0.2175	0.1864	0.0e+00	0.1120
Headphones	Gemini-3.1-Flash-Lite-Preview	structural_virality	9.2e-16	3.4e-22	1.706	1.674	0.0e+00	0.796	0.451	0.710	0.3645	0.3563	0.0e+00	0.1096
Headphones	DeepSeek-Chat	self_bleu_4	7.6e-04	5.0e-04	0.0118	0.0128	0.0526	0.338	0.469	0.710	0.0086	0.0070	0.0400	0.1168
Headphones	DeepSeek-Chat	self_bertscore_mean_f1	5.7e-18	1.3e-28	0.109	0.176	0.842	0.868	0.473	0.710	0.0374	0.0556	0.0e+00	0.1080
Headphones	DeepSeek-Chat	semantic_mean_cosine	5.6e-19	9.6e-32	0.435	0.481	0.921	0.895	0.531	0.710	0.0412	0.0448	0.0e+00	0.1056
Headphones	DeepSeek-Chat	mean_story_probability	2.8e-09	5.1e-08	0.136	0.127	0.368	0.598	0.548	0.475	0.0466	0.0439	0.0400	0.1136
Headphones	DeepSeek-Chat	toxicity_mean	0.373	0.259	0.0161	0.0122	0.0263	0.0898	0.473	0.710	0.0150	0.0114	0.0400	0.1104
Headphones	DeepSeek-Chat	severe_toxicity_mean	0.0494	0.0025	3.0e-04	1.4e-04	0.0263	0.198	0.449	0.475	2.0e-04	1.1e-04	0.0400	0.1104
Headphones	DeepSeek-Chat	obscene_mean	0.328	0.0215	0.0123	0.0138	0.0263	0.0985	0.522	0.710	0.0093	0.0070	0.0e+00	0.1248
Headphones	DeepSeek-Chat	threat_mean	0.0161	0.0021	0.0012	2.8e-04	0.0263	0.242	0.479	0.710	0.0012	4.1e-04	0.0400	0.1096
Headphones	DeepSeek-Chat	aggression_score_mean	0.334	0.177	0.0059	0.0051	0.0263	0.0972	0.548	0.710	0.0053	0.0039	0.0400	0.1120
Headphones	DeepSeek-Chat	length_std	3.2e-04	6.4e-04	17.43	18.04	0.0789	0.362	0.485	0.710	8.059	6.760	0.0e+00	0.1232
Headphones	DeepSeek-Chat	length_cv	3.8e-10	1.0e-13	0.439	0.442	0.0e+00	0.629	0.528	0.710	0.1355	0.1417	0.0e+00	0.1168
Headphones	DeepSeek-Chat	avg_depth	6.3e-16	3.4e-22	0.890	0.828	0.0263	0.801	0.474	0.710	0.2175	0.1864	0.0e+00	0.1120
Headphones	DeepSeek-Chat	structural_virality	9.2e-16	3.4e-22	1.706	1.674	0.0e+00	0.796	0.451	0.710	0.3645	0.3563	0.0e+00	0.1096
Headphones	DeepSeek-V4-Flash	self_bleu_4	0.583	0.688	0.0058	0.0019	0.0e+00	0.0553	0.469	0.710	0.0086	0.0070	0.0400	0.1168
Headphones	DeepSeek-V4-Flash	self_bertscore_mean_f1	2.0e-12	3.5e-21	0.0799	0.0677	0.763	0.707	0.473	0.710	0.0374	0.0556	0.0e+00	0.1080
Headphones	DeepSeek-V4-Flash	semantic_mean_cosine	9.6e-14	6.9e-26	0.383	0.353	0.868	0.748	0.531	0.710	0.0412	0.0448	0.0e+00	0.1056
Headphones	DeepSeek-V4-Flash	mean_story_probability	2.0e-05	7.4e-07	0.109	0.0917	0.105	0.429	0.548	0.475	0.0466	0.0439	0.0400	0.1136
Headphones	DeepSeek-V4-Flash	toxicity_mean	0.232	0.0202	0.0226	0.0122	0.0789	0.120	0.473	0.710	0.0150	0.0114	0.0400	0.1104
Headphones	DeepSeek-V4-Flash	severe_toxicity_mean	0.1000	0.0084	2.8e-04	1.4e-04	0.0526	0.166	0.449	0.475	2.0e-04	1.1e-04	0.0400	0.1104
Headphones	DeepSeek-V4-Flash	obscene_mean	0.551	0.0251	0.0104	0.0108	0.0789	0.0601	0.522	0.710	0.0093	0.0070	0.0e+00	0.1248
Headphones	DeepSeek-V4-Flash	threat_mean	0.0317	0.0058	0.0012	2.8e-04	0.0526	0.216	0.479	0.710	0.0012	4.1e-04	0.0400	0.1096
Headphones	DeepSeek-V4-Flash	aggression_score_mean	0.218	0.0191	0.0071	0.0055	0.132	0.124	0.548	0.710	0.0053	0.0039	0.0400	0.1120
Headphones	DeepSeek-V4-Flash	length_std	0.0023	0.0011	19.66	18.20	0.211	0.306	0.485	0.710	8.059	6.760	0.0e+00	0.1232
Headphones	DeepSeek-V4-Flash	length_cv	2.7e-11	3.0e-15	0.481	0.440	0.0e+00	0.669	0.528	0.710	0.1355	0.1417	0.0e+00	0.1168
Headphones	DeepSeek-V4-Flash	avg_depth	4.2e-16	3.4e-22	0.916	0.828	0.0526	0.807	0.474	0.710	0.2175	0.1864	0.0e+00	0.1120
Headphones	DeepSeek-V4-Flash	structural_virality	9.2e-16	3.4e-22	1.706	1.674	0.0e+00	0.796	0.451	0.710	0.3645	0.3563	0.0e+00	0.1096

G Per-Metric OASIS Baseline Comparisons

Section ?? summarizes the main p -value results, while Table 10 provides the unified comparison across domains, models, and metrics. This section provides per-domain baseline tables for the GPT-family runs together with the same repeated real-vs.-real references used throughout the benchmark. For each domain, the tables compare GPT-4o-mini and GPT-5-mini against held-out real threads using MWU and KS p -values together with Wasserstein distance, quantile error, empirical fail rate, and absolute Cliff’s delta. The final comparison block in each table reports the same-domain median real-vs.-real repeated-sample reference used to calibrate whether a baseline is still far from natural within-domain variation.

G.1 Laptop

Metric	GPT-4o-mini		GPT-5-mini		Repeated real-vs.-real	
	MWU p	KS p	MWU p	KS p	Med. MWU p	Med. KS p
self_bleu_4	1.9e-16	3.2e-31	6.1e-08	2.5e-11	0.516	0.710
self_bertscore_mean_f1	6.5e-17	1.4e-34	3.0e-09	2.4e-17	0.491	0.593
semantic_mean_cosine	4.3e-17	1.4e-34	1.3e-10	4.6e-17	0.567	0.710
mean_story_probability	3.4e-22	7.2e-24	3.8e-14	1.7e-16	0.516	0.710
toxicity_mean	1.9e-11	1.8e-11	0.707	0.024	0.441	0.475
severe_toxicity_mean	1.2e-16	1.5e-18	0.377	0.042	0.516	0.710
obscene_mean	1.2e-10	1.6e-10	0.253	0.008	0.467	0.710
threat_mean	1.8e-17	7.5e-19	0.786	0.171	0.528	0.710
aggression_score_mean	2.4e-11	1.4e-11	0.691	0.014	0.557	0.710
length_std	2.0e-08	1.6e-17	7.6e-04	8.2e-05	0.566	0.710
length_cv	3.4e-17	8.1e-33	1.0e-06	8.3e-09	0.482	0.710
avg_depth	1.0e-20	2.6e-28	7.0e-15	5.9e-16	0.471	0.710
structural_virality	4.6e-20	2.6e-28	1.3e-15	5.9e-16	0.452	0.710

Table 11: Per-metric significance comparison for the laptop domain. Lower baseline p -values indicate that generated and held-out real threads remain easier to distinguish. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs.

Metric	GPT-4o-mini				GPT-5-mini				Repeated real-vs.-real			
	W_1	Quant. err.	Emp. fail	$ \delta_{\text{Cliff}} $	W_1	Quant. err.	Emp. fail	$ \delta_{\text{Cliff}} $	Med. W_1	Med. quant. err.	Med. emp. fail	Med. $ \delta_{\text{Cliff}} $
self_bleu_4	0.1096	0.1076	0.8431	0.7252	0.0424	0.0453	0.5088	0.6348	0.0088	0.0073	0.0e+00	0.1328
self_bertscore_mean_f1	0.2140	0.1903	0.8627	0.7367	0.1486	0.2074	0.8246	0.6952	0.0399	0.0563	0.0e+00	0.1216
semantic_mean_cosine	0.4816	0.4461	0.8627	0.7411	0.3419	0.3449	0.8421	0.7532	0.0453	0.0479	0.0e+00	0.1184
mean_story_probability	0.1720	0.1583	0.1176	0.8548	0.1458	0.1490	0.7895	0.8874	0.0509	0.0418	0.0e+00	0.1184
toxicity_mean	0.0240	0.0207	0.0784	0.5926	0.0059	0.0052	0.1053	0.0445	0.0140	0.0102	0.0e+00	0.1200
severe_toxicity_mean	1.3e-04	5.1e-05	0.1176	0.7305	6.7e-05	6.6e-06	0.1053	0.1040	9.3e-05	4.2e-05	0.0e+00	0.1264
obscene_mean	0.0126	0.0097	0.0784	0.5688	0.0043	0.0019	0.1053	0.1342	0.0093	0.0059	0.0e+00	0.1224
threat_mean	4.1e-04	5.4e-05	0.1176	0.7506	6.9e-04	1.3e-04	0.1053	0.0322	5.0e-04	4.5e-05	0.0400	0.1200
aggression_score_mean	0.0080	0.0068	0.0784	0.5897	0.0021	0.0016	0.1053	0.0469	0.0048	0.0034	0.0e+00	0.1280
length_std	18.81	17.51	0.0e+00	0.4946	28.65	32.06	0.0e+00	0.3945	7.610	7.163	0.0e+00	0.1056
length_cv	0.5559	0.5169	0.0e+00	0.7433	0.2897	0.2917	0.0e+00	0.5724	0.1225	0.1297	0.0e+00	0.1096
avg_depth	0.9650	0.8585	0.0784	0.8103	0.9565	0.9088	0.1579	0.8237	0.2262	0.2078	0.0400	0.1072
structural_virality	1.543	1.476	0.0e+00	0.7941	1.367	1.330	0.0e+00	0.7907	0.3424	0.3437	0.0e+00	0.1120

Table 12: Per-metric discrepancy comparison for the laptop domain. Lower is better for all reported statistics. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs, which serves as the calibration target for judging whether a baseline is still outside normal within-domain variation.

G.2 Credit Card

Metric	GPT-4o-mini		GPT-5-mini		Repeated real-vs.-real	
	MWU p	KS p	MWU p	KS p	Med. MWU p	Med. KS p
self_bleu_4	2.1e-33	6.6e-76	5.4e-15	7.9e-27	0.528	0.710
self_bertscore_mean_f1	7.3e-34	3.4e-105	1.2e-13	1.1e-25	0.525	0.710
semantic_mean_cosine	7.8e-34	1.8e-98	1.6e-19	1.9e-35	0.449	0.710
mean_story_probability	1.8e-27	1.7e-41	8.9e-11	1.6e-08	0.473	0.710
toxicity_mean	4.0e-14	1.3e-13	0.240	0.051	0.541	0.710
severe_toxicity_mean	3.7e-25	1.1e-26	0.479	0.010	0.522	0.710
obscene_mean	2.9e-17	2.2e-16	0.729	0.010	0.473	0.710
threat_mean	1.7e-18	3.7e-21	1.9e-05	1.2e-04	0.528	0.710
aggression_score_mean	1.1e-13	2.3e-13	0.260	0.058	0.509	0.710
length_std	4.5e-17	1.1e-29	1.8e-09	4.4e-14	0.506	0.710
length_cv	5.7e-24	1.7e-47	3.7e-21	5.4e-24	0.485	0.710
avg_depth	1.1e-23	6.4e-37	3.0e-31	1.9e-44	0.559	0.710
structural_virality	9.4e-24	6.4e-37	1.4e-29	1.9e-44	0.470	0.710

Table 13: Per-metric significance comparison for the credit card domain. Lower baseline p -values indicate that generated and held-out real threads remain easier to distinguish. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs.

Metric	GPT-4o-mini				GPT-5-mini				Repeated real-vs.-real			
	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	Med. W_1	Med. quant. err.	Med. emp. fail	Med. $ \delta_{\text{CHRF}} $
self_bleu_4	0.1361	0.1359	1.000	0.9928	0.0541	0.0512	0.4143	0.5865	0.0058	0.0047	0.0e+00	0.1008
self_bertscore_mean_f1	0.2568	0.2328	1.000	1.0000	0.1294	0.1682	0.6714	0.5562	0.0276	0.0148	0.0e+00	0.1088
semantic_mean_cosine	0.5725	0.5686	1.000	0.9995	0.3389	0.3399	0.7857	0.6782	0.0423	0.0417	0.0400	0.1056
mean_story_probability	0.1561	0.1490	0.3600	0.8952	0.0938	0.0911	0.4000	0.4865	0.0464	0.0419	0.0400	0.1104
toxicity_mean	0.0181	0.0150	0.0200	0.6234	0.0107	0.0120	0.1286	0.0881	0.0106	0.0082	0.0e+00	0.1168
severe_toxicity_mean	7.6e-05	2.7e-05	0.4200	0.8544	7.2e-05	4.0e-05	0.1286	0.0532	6.4e-05	2.6e-05	0.0e+00	0.1168
obscene_mean	0.0097	0.0066	0.1200	0.6967	0.0089	0.0082	0.1000	0.0260	0.0067	0.0047	0.0e+00	0.1232
threat_mean	2.7e-04	3.7e-05	0.1000	0.7236	1.5e-04	2.7e-05	0.1714	0.3211	2.1e-04	2.6e-05	0.0e+00	0.1104
aggression_score_mean	0.0060	0.0049	0.0200	0.6122	0.0036	0.0041	0.1143	0.0845	0.0037	0.0028	0.0e+00	0.1184
length_std	25.20	22.56	0.0e+00	0.6923	40.27	39.86	0.4714	0.4512	8.796	7.126	0.0e+00	0.1136
length_cv	0.5818	0.5632	0.0e+00	0.8324	0.4097	0.4113	0.0e+00	0.7083	0.1099	0.1050	0.0e+00	0.1120
avg_depth	0.9406	0.9133	0.0e+00	0.8242	1.084	1.009	0.1714	0.8640	0.2217	0.2014	0.0400	0.0944
structural_virality	1.688	1.642	0.0e+00	0.8242	1.746	1.645	0.0e+00	0.8359	0.3284	0.3347	0.0e+00	0.0984

Table 14: Per-metric discrepancy comparison for the credit card domain. Lower is better for all reported statistics. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs, which serves as the calibration target for judging whether a baseline is still outside normal within-domain variation.

G.3 Cellphone

Metric	GPT-4o-mini		GPT-5-mini		Repeated real-vs.-real	
	MWU p	KS p	MWU p	KS p	Med. MWU p	Med. KS p
self_bleu_4	2.8e-22	6.3e-38	4.0e-11	8.6e-18	0.503	0.710
self_bertscore_mean_f1	4.2e-24	1.8e-44	8.8e-15	5.1e-24	0.475	0.710
semantic_mean_cosine	7.9e-25	5.3e-48	7.4e-16	1.5e-26	0.482	0.475
mean_story_probability	1.3e-27	1.2e-34	9.9e-17	2.1e-18	0.464	0.475
toxicity_mean	6.5e-26	1.7e-34	0.002	4.1e-04	0.516	0.710
severe_toxicity_mean	3.3e-26	2.8e-34	0.002	6.4e-04	0.500	0.710
obscene_mean	2.3e-25	1.9e-33	0.002	3.3e-04	0.554	0.710
threat_mean	3.5e-25	7.5e-33	4.4e-05	3.6e-05	0.551	0.710
aggression_score_mean	6.5e-26	4.7e-35	0.003	4.1e-04	0.541	0.710
length_std	1.2e-16	5.0e-29	3.3e-13	2.5e-17	0.491	0.475
length_cv	3.0e-22	1.0e-40	2.1e-14	2.0e-21	0.458	0.710
avg_depth	4.9e-22	3.6e-31	3.3e-18	6.2e-20	0.484	0.475
structural_virality	2.4e-21	3.6e-31	6.7e-19	6.2e-20	0.498	0.710

Table 15: Per-metric significance comparison for the cellphone domain. Lower baseline p -values indicate that generated and held-out real threads remain easier to distinguish. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs.

Metric	GPT-4o-mini				GPT-5-mini				Repeated real-vs.-real			
	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	Med. W_1	Med. quant. err.	Med. emp. fail	Med. $ \delta_{\text{CHRF}} $
self_bleu_4	0.0698	0.0771	0.4706	0.8353	0.0554	0.0658	0.1493	0.7158	0.0089	0.0069	0.0e+00	0.1320
self_bertscore_mean_f1	0.1892	0.2019	0.9412	0.8713	0.1405	0.1575	0.9104	0.8406	0.0385	0.0382	0.0e+00	0.1200
semantic_mean_cosine	0.4742	0.5058	0.9412	0.8851	0.4511	0.4830	1.000	0.8740	0.0389	0.0397	0.0e+00	0.1160
mean_story_probability	0.1943	0.1878	0.7647	0.9369	0.1430	0.1349	0.7015	0.9003	0.0433	0.0412	0.0e+00	0.1120
toxicity_mean	0.0222	0.0211	0.2941	0.9058	0.0253	0.0238	0.0896	0.3284	0.0084	0.0068	0.0e+00	0.1136
severe_toxicity_mean	8.3e-05	3.2e-05	0.4706	0.9113	1.8e-04	1.4e-04	0.0597	0.3397	5.1e-05	2.3e-05	0.0e+00	0.1216
obscene_mean	0.0111	0.0104	0.3333	0.8955	0.0203	0.0154	0.0448	0.3415	0.0053	0.0049	0.0e+00	0.1144
threat_mean	2.7e-04	8.2e-05	0.4314	0.8920	1.4e-04	5.2e-05	0.1194	0.4430	1.4e-04	5.8e-05	0.0e+00	0.1040
aggression_score_mean	0.0073	0.0068	0.2941	0.9058	0.0088	0.0080	0.0896	0.3212	0.0029	0.0025	0.0e+00	0.1168
length_std	26.51	23.73	0.0e+00	0.7127	44.75	45.88	0.7463	0.7893	9.183	7.239	0.0e+00	0.1192
length_cv	0.6811	0.6675	0.0e+00	0.8345	0.5006	0.5213	0.0597	0.8287	0.1211	0.1296	0.0e+00	0.1136
avg_depth	0.9221	0.8581	0.0784	0.8213	1.223	1.195	0.0896	0.8361	0.2115	0.2111	0.0e+00	0.1136
structural_virality	1.757	1.722	0.0e+00	0.8061	2.136	2.079	0.0e+00	0.8200	0.3649	0.3836	0.0e+00	0.1104

Table 16: Per-metric discrepancy comparison for the cellphone domain. Lower is better for all reported statistics. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs, which serves as the calibration target for judging whether a baseline is still outside normal within-domain variation.

G.4 Camera

Metric	GPT-4o-mini		GPT-5-mini		Repeated real-vs.-real	
	MWU p	KS p	MWU p	KS p	Med. MWU p	Med. KS p
self_bleu_4	4.4e-25	4.7e-50	4.7e-24	2.2e-37	0.547	0.710
self_bertscore_mean_f1	3.9e-25	1.0e-47	2.9e-25	9.9e-47	0.485	0.475
semantic_mean_cosine	1.6e-25	1.0e-48	2.7e-25	9.9e-47	0.506	0.593
mean_story_probability	2.2e-19	2.1e-23	2.7e-28	4.6e-41	0.513	0.710
toxicity_mean	9.7e-08	1.4e-11	3.2e-05	2.9e-09	0.482	0.710
severe_toxicity_mean	1.9e-05	1.8e-09	0.008	1.6e-06	0.500	0.710
obscene_mean	3.6e-09	7.9e-11	0.002	1.4e-07	0.519	0.710
threat_mean	0.642	0.096	0.005	2.9e-04	0.473	0.593
aggression_score_mean	9.6e-08	1.1e-11	3.7e-05	2.9e-09	0.577	0.710
length_std	1.3e-21	7.9e-36	2.7e-07	9.4e-09	0.509	0.593
length_cv	3.2e-24	1.7e-46	3.0e-20	2.3e-27	0.544	0.710
avg_depth	3.6e-24	2.6e-37	8.3e-25	2.6e-37	0.462	0.475
structural_virality	7.4e-24	2.6e-37	7.4e-24	2.6e-37	0.508	0.710

Table 17: Per-metric significance comparison for the camera domain. Lower baseline p -values indicate that generated and held-out real threads remain easier to distinguish. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs.

Metric	GPT-4o-mini				GPT-5-mini				Repeated real-vs.-real			
	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	Med. W_1	Med. quant. err.	Med. emp. fail	Med. $ \delta_{\text{CHRF}} $
self_bleu_4	0.0960	0.0983	0.9400	0.8723	0.0629	0.0634	0.8200	0.8531	0.0058	0.0046	0.0e+00	0.1208
self_bertscore_mean_f1	0.1927	0.2030	0.9200	0.8733	0.1262	0.1324	0.9000	0.8757	0.0300	0.0173	0.0e+00	0.0880
semantic_mean_cosine	0.4872	0.5197	0.9400	0.8806	0.4224	0.4476	0.9200	0.8762	0.0411	0.0410	0.0e+00	0.0920
mean_story_probability	0.1034	0.1006	0.0600	0.7594	0.1189	0.1095	0.5200	0.9302	0.0300	0.0243	0.0e+00	0.1064
toxicity_mean	0.0245	0.0227	0.0400	0.4497	0.0225	0.0220	0.0600	0.3510	0.0110	0.0081	0.0e+00	0.1136
severe_toxicity_mean	1.1e-04	6.0e-05	0.0600	0.3609	1.1e-04	5.9e-05	0.0600	0.2248	7.1e-05	3.7e-05	0.0200	0.1088
obscene_mean	0.0140	0.0116	0.0800	0.4977	0.0124	0.0116	0.0600	0.2646	0.0072	0.0055	0.0e+00	0.1144
threat_mean	8.3e-04	4.5e-04	0.0600	0.0393	0.0013	6.1e-04	0.0600	0.2358	0.0011	4.8e-04	0.0e+00	0.1144
aggression_score_mean	0.0082	0.0073	0.0400	0.4498	0.0076	0.0071	0.0600	0.3480	0.0038	0.0028	0.0e+00	0.1088
length_std	36.10	33.16	0.0e+00	0.8052	22.69	22.08	0.0400	0.4338	11.30	9.356	0.0e+00	0.1296
length_cv	0.6555	0.6532	0.0e+00	0.8563	0.4917	0.5062	0.0e+00	0.7775	0.1186	0.1145	0.0e+00	0.1280
avg_depth	0.9854	0.9365	0.0400	0.8515	1.031	0.9594	0.1200	0.8639	0.2469	0.2245	0.0400	0.1144
structural_virality	1.834	1.778	0.0e+00	0.8453	1.834	1.778	0.0e+00	0.8453	0.3730	0.4083	0.0400	0.1104

Table 18: Per-metric discrepancy comparison for the camera domain. Lower is better for all reported statistics. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs, which serves as the calibration target for judging whether a baseline is still outside normal within-domain variation.

G.5 Headphones

Metric	GPT-4o-mini		GPT-5-mini		Repeated real-vs.-real	
	MWU p	KS p	MWU p	KS p	Med. MWU p	Med. KS p
self_bleu_4	2.1e-24	1.6e-38	1.9e-09	3.6e-11	0.469	0.710
self_bertscore_mean_f1	1.1e-26	4.7e-49	6.8e-13	1.2e-25	0.473	0.710
semantic_mean_cosine	1.1e-26	8.2e-51	7.7e-16	1.6e-29	0.531	0.710
mean_story_probability	1.9e-20	1.3e-23	1.6e-19	1.8e-21	0.548	0.475
toxicity_mean	3.4e-17	4.0e-21	1.2e-05	5.3e-05	0.473	0.710
severe_toxicity_mean	2.6e-18	1.6e-25	2.5e-06	5.9e-08	0.449	0.475
obscene_mean	1.1e-17	7.0e-20	5.2e-05	4.0e-06	0.522	0.710
threat_mean	2.9e-15	2.0e-21	2.4e-08	2.9e-10	0.479	0.710
aggression_score_mean	1.7e-16	7.2e-21	1.3e-05	1.0e-04	0.548	0.710
length_std	2.3e-12	1.6e-21	4.3e-09	1.2e-11	0.485	0.710
length_cv	1.0e-16	1.8e-33	5.4e-13	1.9e-16	0.528	0.710
avg_depth	9.8e-20	3.8e-28	2.1e-21	3.8e-28	0.474	0.710
structural_virality	9.5e-20	3.8e-28	9.5e-20	3.8e-28	0.451	0.710

Table 19: Per-metric significance comparison for the headphones domain. Lower baseline p -values indicate that generated and held-out real threads remain easier to distinguish. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs.

Metric	GPT-4o-mini				GPT-5-mini				Repeated real-vs.-real			
	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	W_1	Quant. err.	Emp. fail	$ \delta_{\text{CHRF}} $	Med. W_1	Med. quant. err.	Med. emp. fail	Med. $ \delta_{\text{CHRF}} $
self_bleu_4	0.0812	0.0850	0.6735	0.9164	0.0281	0.0272	0.2041	0.5395	0.0086	0.0070	0.0400	0.1168
self_bertscore_mean_f1	0.2395	0.2932	0.9796	0.9609	0.1143	0.1011	0.7959	0.6456	0.0374	0.0556	0.0e+00	0.1080
semantic_mean_cosine	0.5464	0.5719	0.9796	0.9609	0.3821	0.3609	0.8163	0.7243	0.0412	0.0448	0.0e+00	0.1056
mean_story_probability	0.1898	0.1833	0.3061	0.8333	0.1797	0.1841	0.4694	0.8126	0.0466	0.0439	0.0400	0.1136
toxicity_mean	0.0383	0.0337	0.0816	0.7583	0.0279	0.0258	0.1429	0.3944	0.0150	0.0114	0.0400	0.1104
severe_toxicity_mean	3.0e-04	1.5e-04	0.1224	0.7850	2.9e-04	1.5e-04	0.1429	0.4233	2.0e-04	1.1e-04	0.0400	0.1104
obscene_mean	0.0181	0.0162	0.0612	0.7700	0.0151	0.0160	0.1429	0.3639	0.0093	0.0070	0.0e+00	0.1248
threat_mean	0.0013	3.7e-04	0.0408	0.7101	0.0012	3.5e-04	0.1837	0.5019	0.0012	4.1e-04	0.0400	0.1096
aggression_score_mean	0.0128	0.0116	0.0816	0.7414	0.0095	0.0093	0.1429	0.3918	0.0053	0.0039	0.0400	0.1120
length_std	21.35	19.62	0.0e+00	0.6308	30.14	28.78	0.2857	0.5276	8.059	6.760	0.0e+00	0.1232
length_cv	0.6314	0.6033	0.0e+00	0.7462	0.4236	0.3835	0.0e+00	0.6485	0.1355	0.1417	0.0e+00	0.1168
avg_depth	0.8610	0.8311	0.0e+00	0.8038	1.020	1.031	0.2041	0.8438	0.2175	0.1864	0.0e+00	0.1120
structural_virality	1.702	1.665	0.0e+00	0.8038	1.702	1.665	0.0e+00	0.8038	0.3645	0.3563	0.0e+00	0.1096

Table 20: Per-metric discrepancy comparison for the headphones domain. Lower is better for all reported statistics. The final block gives the same-domain median real-vs.-real repeated-sample reference from the 200 within-domain resampling runs, which serves as the calibration target for judging whether a baseline is still outside normal within-domain variation.

H Full Prompt-Adjustment Case Study

Section ?? reports the uncalibrated OASIS baselines in the main paper and uses repeated real-vs.-real references to show that current simulations remain far from the real Reddit thread distribution. For completeness, this appendix also reports the full before/after comparison for the subset of domain-model pairs with prompt-adjusted runs. Before denotes the uncalibrated OASIS baseline, After denotes the prompt-adjusted simulator, and $\Delta\%$ is the relative percentage change from Before to After.

Domain	Model	W_1			Quant. err.			Emp. fail			Abs. med. gap			$ \delta_{\text{Cliff}} $		
		Before	After	$\Delta\%$	Before	After	$\Delta\%$	Before	After	$\Delta\%$	Before	After	$\Delta\%$	Before	After	$\Delta\%$
Laptop	GPT-4o-mini	1.190	1.108	-6.9	1.129	1.088	-3.7	0.107	0.108	+0.6	0.871	0.833	-4.4	0.676	0.706	+4.4
Credit Card	GPT-4o-mini	1.629	1.507	-7.5	1.480	1.390	-6.1	0.121	0.119	-1.5	1.207	1.095	-9.2	0.712	0.771	+8.3
	GPT-5-mini	3.617	3.472	-4.0	3.422	3.491	+2.0	0.142	0.149	+5.1	3.936	3.847	-2.3	0.322	0.395	+22.6
Cellphone	GPT-4o-mini	1.855	1.868	+0.7	1.891	1.889	-0.1	0.114	0.116	+2.0	1.445	1.421	-1.6	0.819	0.792	-3.3
Camera	GPT-4o-mini	2.514	2.362	-6.0	2.383	2.203	-7.6	0.174	0.178	+2.1	2.114	1.945	-8.0	0.608	0.641	+5.4
Headphones	GPT-4o-mini	1.319	1.206	-8.6	1.329	1.218	-8.3	0.116	0.114	-1.0	1.049	0.905	-13.7	0.872	0.870	-0.2

Table 21: Full RedditSimBench before/after comparison for the lightweight prompt-based improvement procedure on the subset of domain–model pairs with prompt-adjusted runs. GPT-4o-mini is reported on Laptop, Credit Card, Cellphone, Camera, and Headphones; GPT-5-mini is currently reported on Credit Card. W_1 is average Wasserstein distance, Quant. err. is average quantile error, Emp. fail is empirical fail rate, Abs. med. gap is average absolute median gap across metrics, and $|\delta_{\text{Cliff}}|$ is average absolute Cliff’s delta across metrics. Lower is better for all reported discrepancy metrics, so negative $\Delta\%$ indicates improvement. Bold marks improved After and $\Delta\%$ entries.

NeurIPS Paper Checklist

1. Claims

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

Answer: [Yes]

Justification: The abstract and Section ?? claim that paper identifies capability erosion as a failure mode of self-evolving agents, provides a unified interference-based perspective, proposes Capability-Preserving Evolution (CPE), and evaluates CPE across workflow, skill, model, and memory evolution. These claims are directly supported by the theoretical formulation in Sections ?? and Section 4 and the empirical evaluations in Section ??.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: See Appendix D.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

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

Answer: [N/A]

Justification: [N/A].

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: See Section 5.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

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

Answer: [No]

Justification: Will be released shortly.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: See Section 5.1.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: See Section 5.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: See Section 5.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The work uses publicly available reasoning benchmarks and open / API-accessible LLMs, involves no human subjects or private data, and preserves anonymity in the submission.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

Answer: [Yes]

Justification: See Appendix C.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not

- needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
 - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
11. **Safeguards**
 Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
 Answer: [No]
 Justification: The paper does not release generated datasets, scraped data, or pretrained models.
 Guidelines:
 - The answer [N/A] means that the paper poses no such risks.
 - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
 - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
 - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
12. **Licenses for existing assets**
 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
 Answer: [Yes]
 Justification: See Section 5.1. The work uses publicly available data, APIs, and research models under their respective licenses and usage terms.
 Guidelines:
 - The answer [N/A] means that the paper does not use existing assets.
 - The authors should cite the original paper that produced the code package or dataset.
 - The authors should state which version of the asset is used and, if possible, include a URL.
 - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
 - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
 - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
 - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
 - If this information is not available online, the authors are encouraged to reach out to the asset’s creators.
13. **New assets**
 Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
 Answer: [No]
 Justification: Will be released shortly.
 Guidelines:
 - The answer [N/A] means that the paper does not release new assets.
 - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
 - The paper should discuss whether and how consent was obtained from people whose asset is used.
 - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
14. **Crowdsourcing and research with human subjects**
 Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
 Answer: [N/A]
 Justification: Not applicable.
 Guidelines:
 - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
 - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
 - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
 Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
 Answer: [N/A]
 Justification: Not applicable.
 Guidelines:
 - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
 - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
 - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
 - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.
16. **Declaration of LLM usage**
 Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
 Answer: [Yes]
 Justification: We have decalred the use of LLM in Appendix ??

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.