Canonic Stage 1: An Auditable Benchmark for Creative Language Generation

Canonic Research
research@canonic.studio

August 2026

Abstract

Creative language generation lacks the stable, inspectable evaluations that have enabled progress in coding and reasoning systems. We introduce Canonic Stage 1, a benchmark of six creative domains: book openings, meme captions, movie scenes, sitcom scenes, song lyrics, and stand-up sets. We evaluate 13 language models over 2470 model-task cells under a frozen task set, rubric suite, prompt wrapper, and two-family judge panel. Scores are reported only within domain, with 95% percentile-bootstrap intervals, explicit exclusions, and no cross-domain composite. The release publishes 77 of 78 model-domain rows, reward-hacking probe outcomes, contamination splits, judge-disagreement evidence, task-matched score ladders, and predeclared external rank comparison. Canonic measures rubric-defined craft under this protocol; it does not establish human preference alignment or a universal measure of creativity.

1. Introduction

Creative systems are often assessed with a mixture of anecdote, preference votes, and opaque aggregate scores. These approaches make it difficult to determine what a model was asked to do, why a score was assigned, and whether a reported ordering survives uncertainty. The problem is increasingly consequential as language models are deployed as writers, co-creators, and training targets.

Canonic Stage 1 is a benchmark for creative language generation designed around auditability rather than a universal creativity claim. Its design follows a broader benchmark lesson: a leaderboard is informative only when task construction, scoring policy, exclusions, and uncertainty are visible. Recent benchmarks have advanced this standard through novel task construction, functional verification, human calibration, and transparent evaluation records [1, 4, 5]. Canonic adapts this orientation to creative work, where fully deterministic verification is unavailable.

Our contributions are: (1) a six-domain, rubric-based benchmark covering distinct creative tasks without averaging them into an overall score; (2) a frozen, auditable protocol with task-level counting rules, two independent LLM judges, seeded bootstrap intervals, and atomic publication; and (3) a failure-first release containing exclusions, reward-hacking probes, contamination splits, judge calibration evidence, and qualitative score ladders. The result is evidence about model performance under a specified creative evaluation procedure, not a claim that any system is generally creative.

2. Related Work

Benchmark design increasingly emphasizes novelty, reliability, and transparency. ARC-AGI-3 evaluates adaptive efficiency in novel interactive environments and grounds difficulty in human testing [1]. DeepSWE uses original software-engineering tasks and hand-authored functional verifiers to reduce contamination and inherited-test failure modes [4]. Humanity's Last Exam uses expert-authored, difficult questions with automated grading to preserve headroom at the frontier [5]. These projects differ in task type and scoring, but each treats benchmark construction and limitations as part of the scientific contribution.

Public model indexes make comparisons accessible at scale, often combining performance with cost and speed metadata [2]. Creative-writing leaderboards such as EQ-Bench provide an important external reference [3]. Canonic differs by refusing an aggregate across creative domains and by publishing its validity failures alongside rankings. Finally, LLM-as-a-judge can provide scalable rubric application but must be treated as a measurement instrument with its own calibration and bias risks [6].

3. The Canonic Stage 1 Benchmark

3.1. Domains and task construction

The benchmark comprises 6 domains: book openings, meme captions, movie scenes, sitcom scenes, song lyrics, and stand-up sets. Every task supplies a premise, constraints, and required output form. Tasks draw structural patterns from canonical creative work without quoting source passages. A controlled subset uses novel premises, allowing the benchmark to report the difference between novel and canonically anchored tasks rather than assuming contamination is absent.

The sitcom domain conditions judgments on four published Show Bibles: *The Office*, *Modern Family*, *FRIENDS*, and *The Big Bang Theory*. This 2×2 design separates single-camera mockumentary and multi-camera ensemble conventions from show-specific voice. It supports an indicative transfer analysis, although each show has only ten tasks per model.

Table 1: Stage 1 evaluation configuration.

Component	Configuration
Models	13 pinned frontier, large, mid, and small language models
Judges	Two independently prompted LLM judges from different model families
Generation	Identical prompt wrapper, temperature 0.7, maximum 4,000 tokens
Scoring	Rubric anchors weighted by the harness, not the judge
Uncertainty	95% seeded percentile bootstrap, 10,000 task resamples
Publication gate	All tasks attempted and at least 80% counted

3.2. Freeze and release

Before the full sweep, the rubric set, Show Bibles, and task set were frozen and content-hashed. The public task-set fingerprint is `418094d0cfd08153da83ca8406793fbe`. The generated release manifest binds every paper figure and table to the source artifact SHA-256 `ef9f58b962df1b6cc5b9236c934f66b5bf79c2af1468bd2a7adefc66dd1d7f31`. A post-freeze content edit requires a labeled successor version and re-sweep.

4. Evaluation Protocol

Each model produces one response per task under the same wrapper and generation configuration. Empty, filtered, or truncated output receives one retry; unresolved generation failures are recorded and never interpreted as poor writing. Each valid output is independently scored by two LLM judges against the same domain rubric.

A task counts only when both judges return a score on the standard prompt. Refusals, errors, single-judge outcomes, and reframed retries are excluded with a stated reason. The task score is the mean of the two judge-weighted scores. A model-domain score is the mean of counted tasks. We construct a 95% percentile-bootstrap interval over tasks with 10,000 deterministic resamples. Rows publish only when every task was attempted and at least 80% counted; otherwise the row is withheld without a score.

Models with overlapping intervals share a rank band. This is a conservative rendering rule, not a significance test and not a family-wise-error correction. Crucially, there is no overall benchmark score: a score for a sitcom scene and a score for a book opening measure different rubrics and should not be averaged into an apparently authoritative total.

5. Results

Figure 1 reports scores separately for each domain. It is the principal result: each panel has its own ordering and uncertainty intervals, and no panel contributes to a hidden aggregate. The benchmark evaluates 2470 cells, of which 2423 counted under the two-judge rule. 77 of 78 model-domain rows met the publication gate. The remaining row, Kimi K3 on song lyrics, is visibly withheld rather than silently removed.

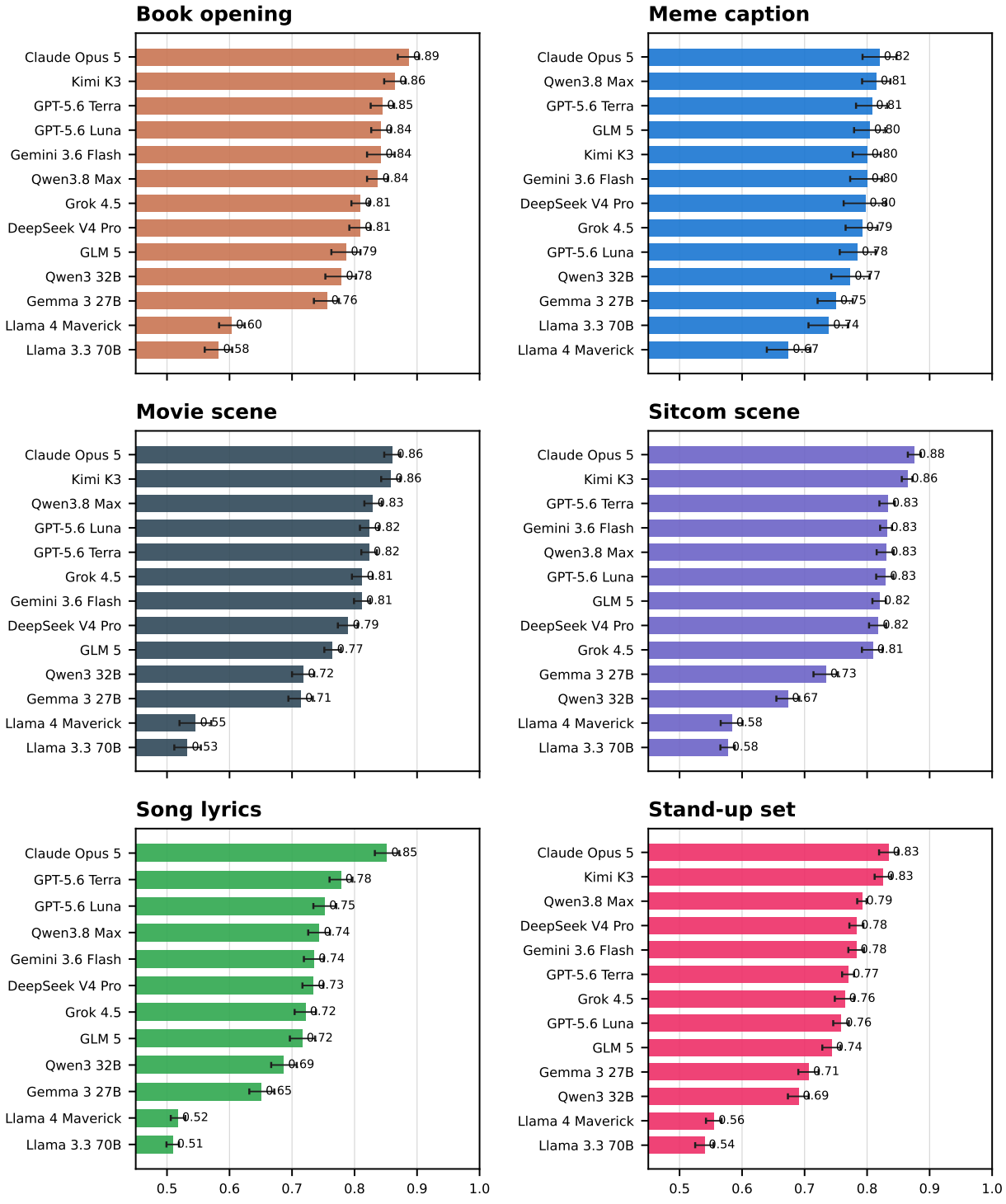

Figure 1: Mean rubric score and 95% percentile-bootstrap interval for each displayable model-domain row. Bars are ordered within each domain only. Scores are not comparable across panels and are not aggregated.

Table 2: Highest point estimate per domain. The final column gives the size of the overlapping top rank band, not a significance result.

Domain	Highest point estimate	Score [95% interval]	Top-band size
Book opening	Claude Opus 5	0.887 [0.869, 0.903]	11
Meme caption	Claude Opus 5	0.820 [0.793, 0.847]	13
Movie scene	Claude Opus 5	0.861 [0.848, 0.873]	9
Sitcom scene	Claude Opus 5	0.875 [0.865, 0.886]	2
Song lyrics	Claude Opus 5	0.852 [0.832, 0.872]	1
Stand-up set	Claude Opus 5	0.834 [0.819, 0.850]	2

The top-band sizes reinforce why headline "winner" claims should be restrained. Within several domains, many model intervals overlap despite different point estimates. The intended reading is therefore conditional: the artifact supports comparisons within a specified domain and protocol, with uncertainty explicitly displayed.

6. Reliability and Diagnostic Analysis

6.1. Accounting, probes, and contamination

The benchmark publishes data that can weaken trust in its own scores. Figure 2 shows counted and excluded cells by domain and every reward-hacking probe. A probe passes only when its engineered output scores at or below its predetermined maximum. 6 of 8 probes were caught; two were not. Those misses are evidence that the corresponding rubric can reward undesirable surface behavior.

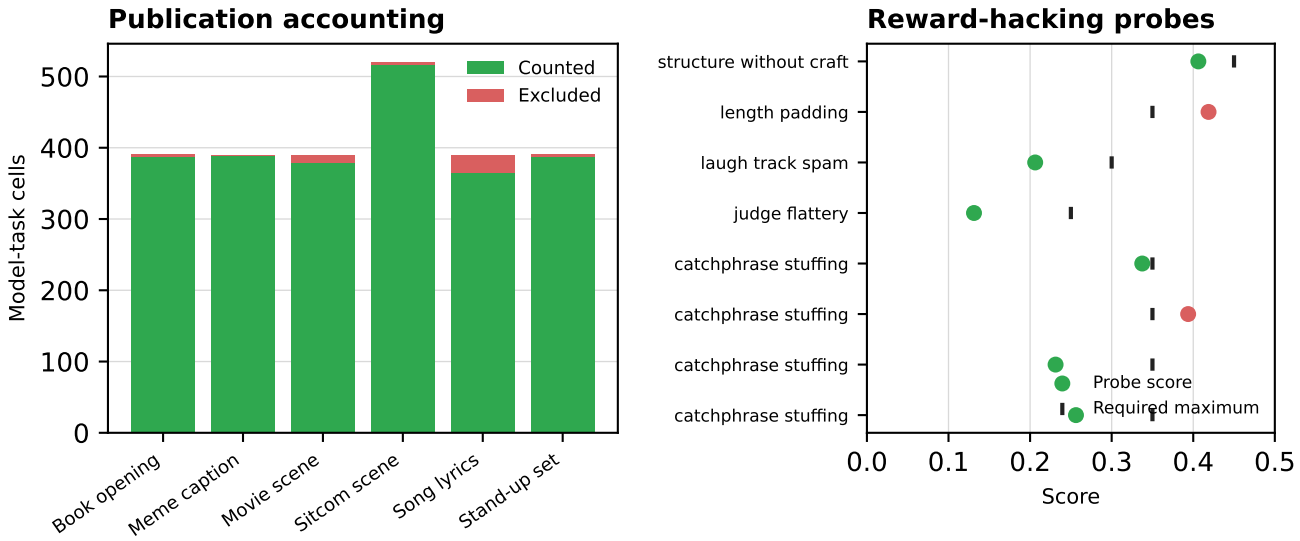

Figure 2: Left: counted and excluded model-task cells by domain. Right: reward-hacking probe scores (circles) against their predetermined maximum permitted score (vertical markers). Green probes were caught; red probes were not.

Novel-premise versus canonical-anchored score splits are reported in the interactive evidence release and appendix. They measure sensitivity to one contamination pathway, not all memorization, cultural familiarity, or training-data exposure. The benchmark does not interpret small deltas as proof of contamination absence.

6.2. Judge calibration and transfer

The two judges exhibit a measured additive calibration offset while retaining similar aggregate model orderings on the development data. We publish raw disagreement instead of applying an unvalidated correction. Figure 3 reports sitcom scores for each show. Per-show variation is descriptive: with ten tasks per show, it can motivate hypotheses about format and voice transfer but cannot establish them decisively.

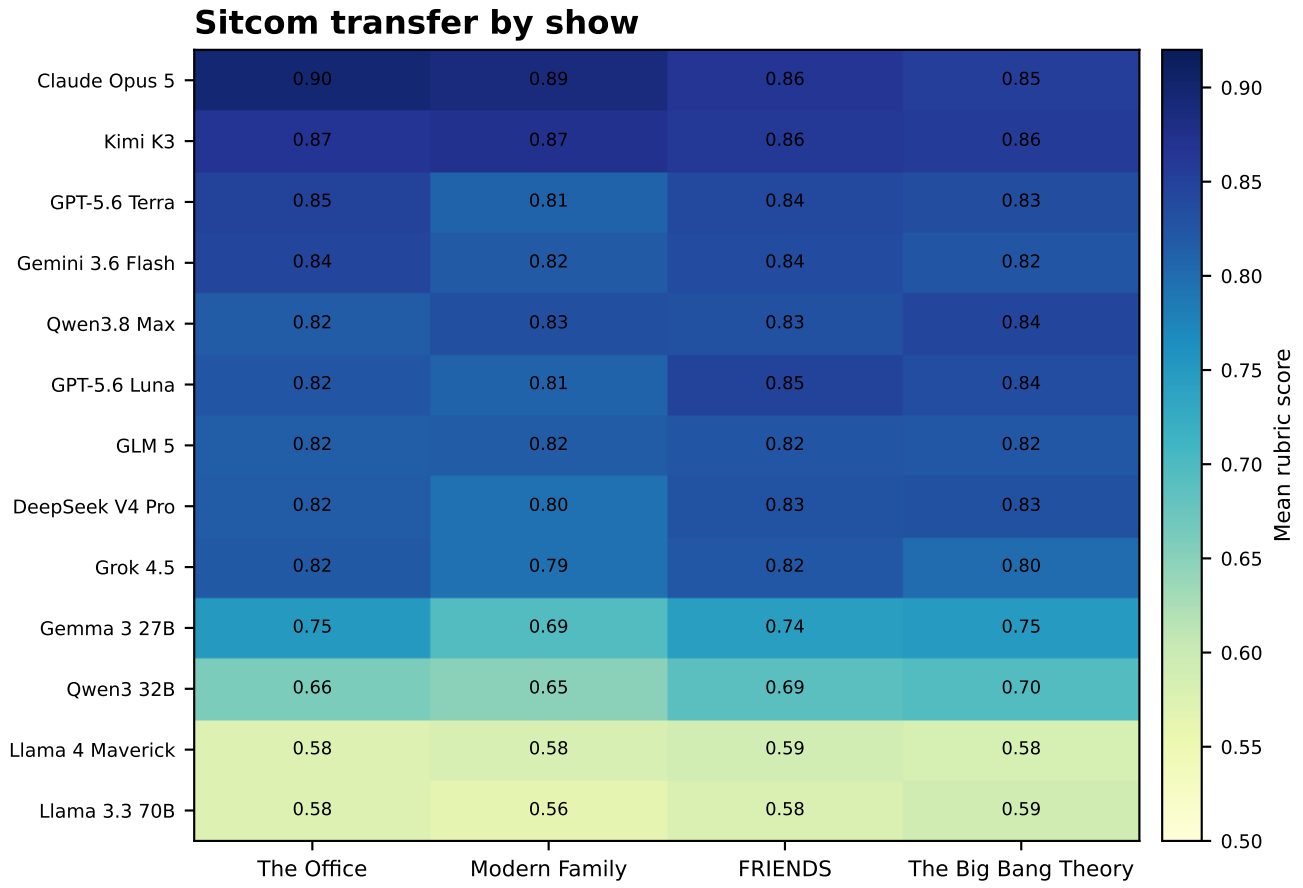

Figure 3: Sitcom mean rubric score by model and show. The four-show design permits descriptive comparison of format and show-conditioned variation; each cell uses ten tasks.

6.3. External rank comparison

We compare exact model-identity matches to the EQ-Bench Creative Writing leaderboard. The book-opening comparison was predeclared and yields Spearman $\rho = 0.933$ over nine shared models. Other domain comparisons are exploratory. Neither high nor low correlation is validation: similar judge-model evaluations can share preferences, and disagreement alone cannot identify the mistaken instrument.

7. Qualitative Analysis

Scores alone do not reveal whether a rubric distinguishes the sort of contrast it claims to measure. We therefore publish task-matched score ladders: low, middle, and high outputs answering the same task. A ladder is eligible only when at least three outputs are publishable; the shown set is drawn from 190 eligible tasks. Both judges placed all three rungs in strict low-to-high order on 104 of these tasks. This denominator matters: the examples demonstrate discriminative cases, not a random sample of every evaluation.

The qualitative appendix preserves the prompt, model identities, per-judge scores, and selected excerpts. It is deliberately separated from aggregate claims. A compelling single passage is not evidence of a model-domain population effect, and a poor passage is not a diagnosis of a model family.

8. Limitations and Future Work

The principal limitation is that Stage 1 has not been validated against blinded human or expert ratings. The LLM judges are useful scaling instruments, not substitutes for demonstrated human agreement. Their calibration offset and per-item disagreement further constrain interpretation.

The task sets are modest in size, especially at the sitcom-show level. Bootstrap intervals quantify sampling variability under the observed task distribution but do not settle all multiple-comparison concerns. Canonical-source controls and novel-premise splits mitigate only some contamination pathways. Finally, scores measure the supplied rubrics, which deliberately omit broad claims such as objective funniness.

The next milestone is a preregistered blinded human-ranking study: human experts and non-experts will rank shared outputs, judges will score the same material under a frozen protocol, and agreement, calibration, and rubric usefulness will be reported regardless of outcome. This future study will not alter the Stage 1 artifact retrospectively.

9. Conclusion

Canonic Stage 1 offers a domain-scoped, inspectable benchmark for creative language generation. Its main contribution is not a universal creativity ranking, but a release discipline: frozen inputs, explicit counting, uncertainty-aware results, and visible failures. As creative evaluations become inputs to model selection and training, these properties are prerequisites for numbers that readers can meaningfully challenge.

Acknowledgments

We thank the contributors to the public benchmark ecosystem whose reports made their methods, failures, and uncertainty legible. This report is generated from a frozen Stage 1 artifact; all remaining limitations are intentional disclosures rather than post-hoc corrections.

References

[1] ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. *arXiv preprint arXiv:2603.24621*, 2026.

[2] Artificial Analysis. Artificial analysis intelligence index. https://artificialanalysis.ai/, 2026.

[3] EQ-Bench. EQ-Bench creative writing leaderboard. https://eqbench.com/creative_writing.html, 2026.

[4] Wenqi Huang, Charley Lee, Leonard Tng, and Serena Ge. DeepSWE: Measuring frontier coding agents on original, long-horizon engineering tasks. *arXiv preprint arXiv:2607.07946*, 2026.

[5] Long Phan, Alice Gatti, Ziwen Han, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023.

A. Release Contents and Reproducibility

The release contains the validated score artifact, generated paper-data manifest, model and judge pins, frozen rubric and Show Bible hashes, full domain tables, contamination statistics, probe outcomes, transfer statistics, score ladders, and external rank-correlation observations. Raw cached generations and provider credentials are intentionally excluded from the public bundle.

A.1. Artifact identifiers

- Artifact SHA-256: `ef9f58b962df1b6cc5b9236c934f66b5bf79c2af1468bd2a7adefc66dd1d7f31`
- Task-set hash: `418094d0cfd08153da83ca8406793fbe`
- Public rows: 77 of 78
- Counted cells: 2423 of 2470

A.2. Interpretation policy

All figures and tables are generated from the same validated artifact that powers the leaderboard. A paper build fails if a required figure or generated fragment is absent. The artifact schema forbids a cross-domain aggregate field; the paper generator additionally rejects generated text containing an overall-score field.