AGC-BENCH: MEASURING ARTIFICIAL GENERAL CREATIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Creativity research has long debated whether creativity is a general ability or specific to particular domains (e.g., creative writing, visual art, scientific discovery), and whether it is separable from general intelligence. Both questions now apply to LLMs, but a fragmented evaluation landscape across hundreds of heterogeneous creativity benchmarks has left them empirically intractable. We introduce **AGC-Bench**, a meta-benchmark for artificial general creativity built from a PRISMA-compliant systematic review of the AI creativity literature ($3,101$ papers screened, 497 unique benchmarks identified) paired with an agentic onboarding harness that converts source-paper benchmarks into runnable HELM-style scenarios. The first release covers 78 datasets (67 text-only, 11 multimodal) spanning brainstorming, problem solving, STEM, narrative, figurative language, and humor. To address bias in LLM-as-judge, we apply Judge Response Theory (JRT)—a psychometric calibration of judge leniency/severity—with three frontier LLM judges. We then fine-tune Qwen3-30B-A3B-Instruct-2507 on the resulting $48,299$ JRT-corrected ratings to produce **AGC-Judge**, an open-weight scoring model that matches the three-judge ensemble and predicts frontier-judge ratings with high accuracy on creativity benchmarks it was not trained on. Results reveal frontier models at the top of the leaderboard, with open-weight models close behind. However, different models show different creative strengths, ranking higher on some domains (e.g., creative writing) than others (e.g., scientific ideation). We conduct several experiments and report three main findings. First, applying factor analysis across 83 LLMs, we recover a single creativity factor 'c', analogous to the 'g' factor of general intelligence, that explains 82.4% of variance, related to but separable from general knowledge and reasoning. Second, in within-model comparisons, we show that prompting models to "be creative" boosts their performance far more than enabling their reasoning capability, evidence that the benchmark tracks creativity specifically. Third, on a human-matched subset of five tasks, the top human still leads the top LLM on the novelty and diversity composite. We release AGC-Bench with a public leaderboard, the onboarding harness, AGC-Judge weights, and human comparison data as open infrastructure for measuring AI creativity at scale.

1 INTRODUCTION

Creativity is widely regarded as a hallmark of human intelligence, and large language models (LLMs) are increasingly being asked to produce creative output—writing stories, generating research ideas, designing marketing copy, composing poetry (Peeperkorn et al., 2024; Bhat et al., 2025; Ruan et al., 2026). Evaluating how well they do it, however, remains challenging. Hundreds of creativity benchmarks have been published in recent years, each targeting a narrow slice of creative behavior, each using its own scoring protocol, and most run in isolation. Yet the field has no infrastructure for asking whether a model's creative ability generalizes across domains—whether a model that writes compelling short stories also generates clever metaphors, designs novel product taglines, or formulates original scientific hypotheses.

The same question is non-trivial in humans. Previous work suggests that creative performance reflects both general abilities that support many kinds of creative work and domain-specific abilities tied to particular tasks (Baer, 2010; Plucker, 2004; Kaufman & Beghetto, 2009). Someone who writes good

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

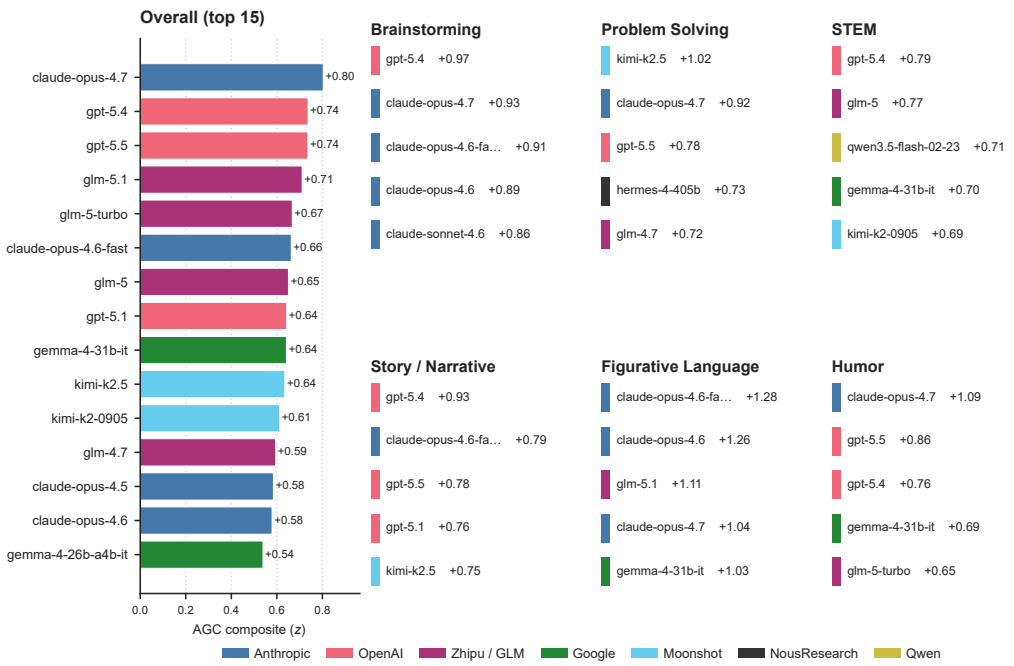


Figure 1: **Overall and per-domain rankings. Left:** top 15 of 83 models on the AGC-Bench composite. **Right:** top 5 within each of the six domains. Per-domain lineups overlap substantially with the overall ranking but reorder at the margins (e.g., `kimi-k2.5` on Problem Solving, `claude-opus-4.6-fast` on Figurative Language). Appendix F lists the top 30 models; the complete 83-model leaderboard is included in the released data.

poetry, for example, does not need to be equally inventive when sketching or solving engineering problems. Whether a similar pattern holds for LLMs is unknown, in part because the field lacks the evaluation infrastructure to answer it. We call this open question, by analogy with artificial general intelligence, the problem of *artificial general creativity*: does creative performance in LLMs reflect a single general capacity that transfers across domains, or a set of domain-specific abilities shaped by training data, architecture, and post-training objectives? To address this question, we take a psychometric approach. Creativity researchers have long used psychometric tools, such as factor analysis and item response theory, to test whether human creative ability is general or domain-specific. We apply the same tools to large-scale evaluation data from LLMs. To do so, we need a unified framework that measures creative ability across diverse domains.

We introduce AGC-Bench, a meta-benchmark built by aggregating the published literature on AI creativity evaluation. A systematic review (Page et al., 2021) of **3,101** candidate papers yielded **497** candidate benchmarks published between 2018 and 2025. We onboarded 78 of them into a standardized evaluation harness modeled on HELM (Liang et al., 2023), organized around six creativity domains drawn from human creativity research—brainstorming, problem solving, STEM, story / narrative creativity, figurative language, and humor—as well as a multimodal visual and design subset. We evaluated 83 models spanning diverse model families and scales. Each benchmark was scored using the metric from the original study or a documented adaptation. To address bias in LLM-as-judge scoring, we applied Judge Response Theory (JRT) (Myszkowski & Storme, 2019)—a psychometric calibration of each judge's leniency and severity in rating creativity—with a panel of frontier LLM judges from three vendors (Gemini-3-Flash, Grok-4.1-Fast, GPT-4.1-mini) under a planned-missing 2-of-3 design. We used the resulting calibrated estimates to train AGC-Judge, an open-weight creativity rater described below. To test whether creative ability in LLMs is general or domain-specific, we aggregated per-model scores across the 67 benchmarks into six domain composites and used principal component analysis to extract a general creativity component (C).

We report three main findings. First, across 83 LLMs, principal component analysis identifies a dominant creativity component, C, explaining 82.4% of the variance across the six domains

(standardized $\alpha = .96$); the component correlates with general knowledge and reasoning yet remains separable from each. Second, in within-model comparisons, prompting models to "be creative" produced larger performance gains than enabling reasoning. Third, on a human-matched subset of five tasks, the top human still leads the top LLM on the novelty and diversity composite.

This work contributes three primary artifacts:

- **AGC-Bench.** A benchmark for artificial general creativity built from 78 onboarded datasets (67 in the primary analysis, 11 additional multimodal scenarios). Each dataset runs as a HELM-compatible scenario scored using the metric from the original study or a documented adaptation. We release the harness, model responses, and a public leaderboard.
- **AGC-Judge.** An open-weight Qwen3-30B-A3B scoring model fine-tuned on 48,299 ratings from three LLM judges, calibrated with JRT. AGC-Judge reproduces model rankings from the three-judge ensemble (Spearman $\rho = .98$ in distribution) and predicts ratings from frontier LLM judges on creativity benchmarks it was not trained on ($\rho = .83$), at a fraction of the cost of using frontier LLM judges.
- **Meta-benchmarking pipeline.** A reproducible workflow that converts source-paper benchmarks into runnable scenarios, combining a systematic review with an agentic onboarding harness. The pipeline supports incremental coverage of the literature beyond this release, and extends naturally to building meta-benchmarks in other domains.

We also release the full benchmark inventory of 497 candidate benchmarks surveyed during onboarding—a superset of the 78 we run, annotated with modality, scoring protocol, and source paper, supporting future expansion of the harness. AGC-Bench constitutes a substantially broader creativity benchmark than prior work and an initial empirical test of the artificial general creativity question. All artifacts are available at `https://huggingface.co/agcbench-2026`.

2 Related Work

General and domain-specific creativity. A central question in human creativity research is whether creative ability reflects a single general capacity, analogous to Spearman's g for intelligence, or whether creativity is domain-specific. Assessments of creative behavior, including the Creative Achievement Questionnaire (Carson et al., 2005), the Kaufman Domains of Creativity Scale (Kaufman, 2012), and the Inventory of Creative Activities and Achievements (Diedrich et al., 2018), consistently partition creative performance across multiple content domains—visual arts, creative writing, humor, scientific and technical creativity, and everyday problem-solving—with moderate correlations across domains and substantial domain-specific variance (Baer, 2010). Creative performance is typically assessed using tasks that require the generation of ideas or products that are both novel and appropriate (Runco & Jaeger, 2012; Amabile, 1996). AGC-Bench adopts this framework directly. The six domains map onto content categories that recur across these taxonomies, paired with process-level traditions (divergent thinking and creative problem solving) that creativity researchers have long studied alongside them. Model performance can therefore be situated within the same structure used to characterize human creative ability.

Intelligence, creativity, and implications for LLMs. The relationship between intelligence and creativity has a long and contested history in psychology, with direct implications for the evaluation of LLMs. Within the Cattell-Horn-Carroll framework (McGrew, 2005), intelligence comprises a general factor (g) and broad abilities including fluid intelligence (G_f, the capacity for novel problem solving through inductive and deductive reasoning) and crystallized intelligence (G_c, accumulated declarative and procedural knowledge). In the context of creative thinking, G_f supports the generation and evaluation of candidate ideas, while G_c provides the knowledge base for developing them (Gerwig et al., 2021). Frontier LLMs exhibit performance analogous to both components, and many existing "creativity" benchmarks plausibly index G_f in disguise. Distinguishing creative ability from reasoning therefore requires a targeted G_f indicator that resists mere pattern-matching. Lewis & Mitchell (2025) construct counterfactual letter-string analogies, replacing the Latin alphabet with shuffled alphabets and symbol sequences to force rule abstraction over memorization—a logic paralleled by Chollet's Abstraction and Reasoning Corpus (Chollet, 2019). We adopt the counterfactual letter-string task as our pure-G_f probe alongside the AGC-Bench composite.

The LLM creativity benchmark landscape. Benchmarks evaluating creative capability in LLMs have proliferated rapidly across multiple lines of work, with each subfield developing its own conventions in relative isolation. The earliest and largest group adapts the classical psychometric tasks of divergent thinking—the Alternate Uses Task, the Torrance Tests, the Divergent Association Task—to LLMs, generally reporting that frontier models match or exceed average human performance while falling short of the most creative humans (Stevenson et al., 2022; Guzik et al., 2023; Hubert et al., 2024; Bellemare-Pepin et al., 2026). Creative writing has become the most active area, with benchmarks that differ in length, granularity, and evaluation philosophy and range from constrained short-story comparisons (Gómez-Rodríguez & Williams, 2023) to expert-rubric evaluation (Chakrabarty et al., 2024). Humor evaluation has consolidated around image captioning such as the New Yorker Caption Contest (Hessel et al., 2023), with extensions targeting humor reasoning (Narad et al., 2025) and expert-aligned ranking (Zhou et al., 2025). Other lines of work address figurative language and metaphor (Chakrabarty et al., 2023; Liu et al., 2022), scientific ideation (Ruan et al., 2026), visual creativity (Fang et al., 2025), human-versus-AI creative drawing (Nath et al., 2025), and "lateral thinking" through puzzle-solving tasks (Jiang et al., 2023). Each benchmark deepens understanding of one creative domain, but because they were built independently—different prompts, scoring protocols, and model rosters—their results cannot be combined to ask broader questions about a model's creative profile across domains.

Multi-domain evaluation and the present work. A few recent efforts have begun to bridge across domains. The closest prior consolidation is CreativityPrism (Hou et al., 2026), which evaluates 17 LLMs on eight tasks spanning divergent thinking, creative writing, and logical reasoning, using automated judges validated against human annotations. The authors report limited generalization across domains. AGC-Bench extends this work to 67 benchmarks across six domains and 83 models, with JRT calibration and principal component analysis. Other work has examined the validity of the field's evaluation practices and metrics. Lu et al. (2026) catalog systematic confounds in existing creativity benchmarks, e.g., prompt sensitivity, model inconsistency, and potential data contamination, arguing that even within a single benchmark, reported scores may not robustly reflect underlying creative ability. Bhat et al. (2025), the most rigorous single-domain study to date, found tightly clustered model performance and poor agreement between LLM judges and human experts on a large pairwise marketing-creativity comparison. Taken together, prior work indicates that creative performance is heterogeneous across domains and tasks, and that characterizing LLM creativity requires breadth that no single benchmark can provide. AGC-Bench addresses this gap.

3 METHODS

AGC-Bench contains 67 benchmarks in the primary analysis, onboarded from our systematic review. Each dataset is scored using the metric from the original study or a documented adaptation. The present submission is the first release of an ongoing onboarding effort; the harness is designed to extend to additional benchmarks as they are integrated. The release also includes a human subset, *AGC-Human*, of five tasks from the Creativity Assessment Platform (Patterson et al., 2025) that place LLMs and humans on directly comparable scales.

3.1 BENCHMARK CURATION

Our theory-driven inclusion criteria required an AI creativity task with multiple valid solutions or judgments of creative outputs, accessible data, and faithful reproduction of prompts and scoring through current model interfaces.

We assembled the benchmark inventory through a systematic literature review with six reproducible stages: literature harvesting, automated pre-screening, dataset verification, dual human review, benchmark extraction, and deduplication. The pipeline interleaves LLM automation for triage at scale with human review at every accept step. The flow of papers and benchmarks through the pipeline is summarized in a PRISMA diagram (Page et al., 2021) (Figure S1).

Pre-screening. From an initial pool of 3,101 candidate papers (January 2018–December 2025), `gpt-4.1` applied a structured inclusion rubric to each title and abstract. A paper was retained if it (a) targeted an AI or LLM system, (b) introduced or substantially modified a benchmark or

dataset, (c) focused on a creativity-related construct (e.g., ideation, originality, humor, metaphor, discovery), (d) included either a generation task with multiple valid solutions or a judgment task rating creative outputs, and (e) was published between 2018 and 2025. Survey papers without a new benchmark and papers where creativity was incidental were excluded. The rubric retained 683 papers. `gemini-2.5-flash` with Google Search grounding then verified that 431 had publicly accessible code or data. Four trained graduate students then reviewed the resulting candidate set under a rotating assignment protocol, with conflict resolution by the lead authors. See Appendices A and B for the curation flow and onboarding procedures.

Extraction and deduplication. Of the 431 candidates with verified data access, 283 passed the dual human review and were forwarded to extraction. Benchmark extraction with `gemini-2.5-pro` parsed the 276 papers with retrievable PDFs into structured records, producing 546 candidate benchmarks across 1,160 constituent tasks. Each record captured benchmark name, modality, task description, evaluation metric, and dataset URL. We standardized benchmark names, then used Gemini 2.5 Pro to identify duplicates across 10 runs, requiring agreement in at least eight. This identified 33 duplicate groups, leaving 497 candidate benchmarks.

Candidate selection. A relevance review excluded incorrectly included benchmarks in general NLP and computer vision, leaving 432 of the 497 candidates. Each candidate was then classified by modality and infrastructure requirements: `gemini-3-flash-preview` assigned a tier based on input/output type, with the lead authors reviewing tier-boundary cases. The classifier yielded 241 Tier-1a (text-to-text, ordinary model APIs without a custom interactive environment) and 86 Tier-1b (image-to-text, vision-capable APIs), leaving 105 higher-tier candidates. Tier-1a and selected Tier-1b candidates entered the onboarding pipeline (Section 3.2). Higher-tier candidates are retained in the catalog for future releases.

3.2 Benchmark Onboarding

The 327 candidate benchmarks in Tiers 1a and 1b use different data formats and scoring procedures. An agentic onboarding workflow (`benchmark-onboarder`) built with Claude Code on `claude-opus-4.6` coordinates a nine-state pipeline from paper parsing through scenario generation and pilot execution, with three design principles—verbatim prompt extraction, source-paper scoring metrics preserved, and regex/code-based extraction preceding LLM inference—and human review of prompts, dataset correspondence, and scoring procedures before acceptance. The pipeline produced 159 HELM-compatible scenarios; after a launch-readiness audit, 78 entered the present evaluation cycle (67 in the primary analysis and 11 additional multimodal scenarios), with the remaining scenarios queued for future evaluation. Failure modes and the onboarding workflow are described in Appendix B.

3.3 Benchmark Taxonomy

The 67 benchmarks in the primary analysis are organized into a taxonomy grounded in human creativity research, with six creativity domains mapped onto content categories shared across major self-report inventories of creative behavior (Carson et al., 2005; Kaufman, 2012; Diedrich et al., 2018): *Brainstorming* (3 benchmarks), *Problem Solving* (10), *STEM* (13), *Story / Narrative* (20), *Figurative Language* (12), and *Humor* (9). Domain assignments come from a three-model LLM panel (`gemini-3-flash-preview`, `grok-4.1-fast`, `gpt-4.1-mini`) on benchmark metadata, adjudicated by majority vote with lead-author resolution (mean pairwise Cohen's $\kappa \approx 0.85$, Fleiss' $\kappa \approx 0.85$ on the released taxonomy panel); per-dataset assignments, per-domain counts, and the partition rationale appear in Appendix D.

Scope of the present analysis. The first AGC-Bench release covers 78 datasets, including 67 in the primary analysis and 11 additional multimodal scenarios released as artifacts. The primary set includes 66 text-input benchmarks and IRFL, which uses image and text inputs. IRFL scores are available for 33 of the 83 models. The additional 11 multimodal scenarios are not analyzed here.

3.4 Evaluation Setup

Harness and sampling. Our evaluation harness extends HELM (Liang et al., 2023) and lm-evaluation-harness (Gao et al., 2024). Each benchmark is implemented as a self-contained scenario with a stable identifier. We evaluate 83 LLMs released between 2024 and 2026, drawn from the major proprietary and open-weight families and run via vendor APIs and OpenRouter with reasoning off by default. For each model and benchmark, we sampled up to 50 items using a fixed seed, evaluated smaller datasets in full, and preserved the original data splits. This cap prioritizes breadth across benchmarks and domains over depth on any single one.

Scoring and aggregation. Each dataset is scored using the metric from the original study or a documented adaptation. Implemented metrics, rubrics, and judge prompts are provided in the released harness (Appendix E). JRT calibration covered 24 benchmarks in the primary analysis. For calibration, each response was assigned to two of three judges—`gemini-3-flash-preview`, `grok-4.1-fast`, and `gpt-4.1-mini`—under a planned-missing design. The single-judge composite uses the original judge ratings before JRT calibration, with the other benchmark metrics unchanged. To correct systematic differences in severity and discrimination across judges, we fit a Bayesian graded response model from the Judge Response Theory framework (Myszkowski & Storme, 2019), which recovers ability θ for each response alongside discrimination α for each judge and score thresholds β, linked by shared responses. For example, ratings of 3 and 4 can be adjusted toward 3 if the judge assigning 4 is generally more lenient. The raw and adjusted response scores correlate $\rho = 0.97$ (Appendix H.4; full specification and comparisons of raw and JRT-adjusted rankings in Appendix H). For each model, the AGC-Bench composite averages standardized scores across the 67 benchmarks, using JRT-adjusted panel ratings where available and canonical scores otherwise. Multi-metric datasets are mean-aggregated, and domain composites are computed analogously for the principal component analysis (Section 4.3).

Data-quality audit and reproducibility. A separate audit using `grok-4.1-fast` assessed three randomly sampled responses per model–benchmark cell. The on-task rate was 95.1%, and agreement with a second judge was $\kappa = .67$ (Appendix G). The catalog, harness, responses, judge verdicts, and a public submission leaderboard are released at `https://huggingface.co/agcbench-2026`, with each scenario versioned and traceable to its source paper.

External validity. We assessed external validity using counterfactual letter-string analogies as a measure of fluid reasoning (Lewis & Mitchell, 2025) and two within-model interventions on AGC-Human: creativity instructions and enabling reasoning (Appendices I and K).

3.5 The AGC-Judge scorer

We fine-tuned `Qwen3-30B-A3B-Instruct-2507` with LoRA to produce **AGC-Judge**, an open-weight model trained to predict JRT-calibrated creativity ratings. The training corpus comprises 48,299 (model, item) ratings drawn from the JRT-corrected pool, with each row pairing a model response to a JRT-derived integer rating on the original benchmark's scale, and the adapter is trained with the standard causal-LM objective. We release the weights and the scoring harness so others can score new models against the AGC-Bench leaderboard without running a frontier-judge ensemble (full training configuration, hyperparameters, and length filtering in Appendix H.5; training comparison in Appendix H.6; held-out evaluation in Section 4.2). We also trained a version on mean raw ratings using the same examples and data splits (Appendix H.6).

3.6 Human comparison subset (AGC-Human)

AGC-Human places humans and LLMs on directly comparable scales across five tasks from the Creativity Assessment Platform (Patterson et al., 2025): three idea-generation tasks (Alternate Uses, Scientific Creative Thinking, Design Problems) and two creative writing tasks (Metaphor Completion, Short Story). The cohort comprises 224 humans and 125 LLM evaluations. Response validity is assessed by majority vote across three LLM judges. Excluding participants or models with available validity rates below 95% and restricting LLMs to the AGC-Bench cohort leaves 201 humans and 80 LLMs in the primary analysis sample. Each response is then scored on novelty (Johnson et al., 2023)

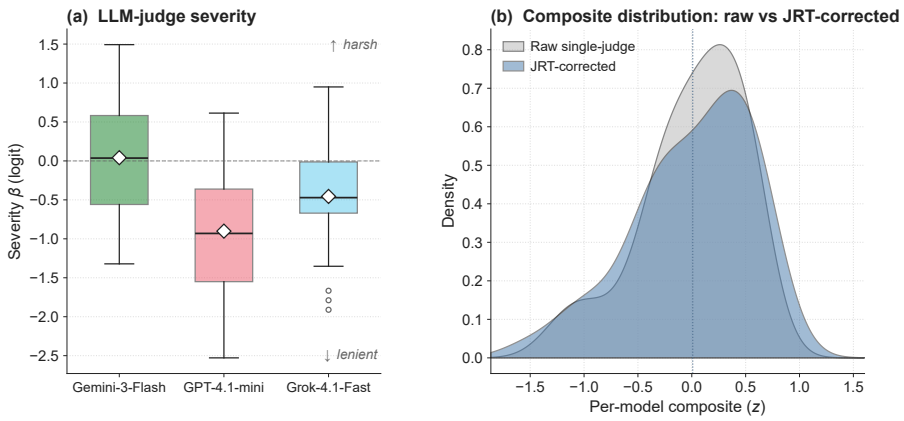


Figure 2: **Judge Response Theory (JRT) calibration of benchmarks scored by LLM judges. (a)** Per-judge severity β across 24 LLM-judge cells: `gpt-4.1-mini` runs substantially lenient (median $\beta = -0.93$); `gemini-3-flash` sits near zero. **(b)** Per-model composite density for the raw single-judge (gray) and JRT-corrected (blue) scores.

and diversity using embedding models, with scores adjusted for response length and standardized within each task and prompt. Validity criteria are provided in Appendix L. We use embedding-based novelty and diversity scores as the primary measures because they allow new models to be evaluated automatically against the existing human sample. These measures also align with expert ratings of short stories (Ismayilzada et al., 2025). AGC-Judge is not used to calculate them. To validate the automated scores, we collected originality ratings from 416 participants recruited through Prolific (Appendix L). Their ratings covered all 24,172 responses on a 1–5 scale using a planned-missing design. Responses were not identified as human or AI responses.

4 Results

4.1 Frontier models lead overall, but no model leads every domain

The left panel of Figure 1 shows the top 15 models on the overall composite. Proprietary frontier systems lead the field—`anthropic/claude-opus-4.7`, `openai/gpt-5.4`, `openai/gpt-5.5`—with open-weight `z-ai/glm-5.1` and `glm-5-turbo` close behind. The right panel of Figure 1 reorders the same models by per-domain composite and surfaces dissociations that diverge from the overall ranking. `kimi-k2.5` tops Problem Solving (+1.02) and also enters the top five in Story / Narrative. `claude-opus-4.6-fast` leads Figurative Language (+1.28). Open-weight `gemma-4-31b-it` ranks among the top five in three domains (STEM, Figurative Language, Humor) despite ranking ninth overall. Creative ability is therefore unevenly distributed even among top models, a pattern we examine quantitatively in Section 4.3.

4.2 AGC-Judge generalizes across held-out splits

We test whether the calibrated three-judge ensemble can be approximated by a single fine-tuned open-weight scorer. We find that AGC-Judge predicts JRT-corrected scores at pooled Spearman $\rho = 0.94$ on held-out items and frontier models and pooled $\rho = 0.83$ on three held-out benchmarks (Figure 3). On an out-of-distribution corpus of creative writing from humans and LLMs (Orwig et al., 2024), AGC-Judge tracks human ratings comparably to a frontier GPT judge ($\rho = .65$ versus .66; Appendix H.7). Training on raw ratings better reproduced the original scale, whereas AGC-Judge better preserved rankings within benchmarks and metrics on the test set and for models excluded from training. Results on new benchmarks varied by metric (Appendix H.6). AGC-Judge scores were strongly correlated with AGC-Human embedding scores across 80 LLMs ($\rho = .80$; Appendix L).

4.3 A GENERAL CREATIVITY COMPONENT (C) IN LLMS

We next ask whether creative ability in LLMs is general across the six domains or specific to each. To this end, we conducted a principal component analysis of the correlation matrix of the six domain composites across 83 LLMs. Results show that the structure is clearly unidimensional. The first eigenvalue was 4.94; the remaining five were 0.48 or lower. Parallel analysis using $1{,}000$ random matrices supported retaining one component (Horn, 1965). The first principal component, which we call *C* by analogy with *g* for general intelligence, explains 82.4% of variance, with all six component loadings between $+0.87$ and $+0.95$ (standardized Cronbach's $\alpha = .96$). Thus, knowing how a model performs in one creative domain reliably predicts its performance in the others.

Critically, the dominant component is preserved across three robustness checks. First, it holds after partialling out fluid reasoning ($\alpha = 0.94$ on the matrix after adjusting for letter-string analogy scores), so *C* does not appear to be a re-parameterization of G_f. Second, analyses restricted to benchmarks scored using formulas ($\alpha = .91$) or LLM judges ($\alpha = .96$) each yielded a single dominant component, so *C* is not an artifact of any single scoring method. Finally, *C* remained stable across model subsamples. Excluding Story/Narrative, the largest domain, also preserved the structure ($\alpha = .94$). Full robustness diagnostics appear in Appendix J, with contamination checks in Appendix N.

4.4 *C* CORRELATES WITH INTELLIGENCE YET REMAINS SEPARABLE

A unitary *C* would not be a meaningful psychometric finding if creativity in LLMs simply reduced to general capability. We therefore tested three competing reductions: *C* as fluid reasoning, *C* as crystallized knowledge, and *C* as raw model scale.

We find that *C* correlates substantially with both components of human intelligence, but separably. The AGC-Bench composite correlates with fluid reasoning, measured with counterfactual letter-string analogies, at $\rho = +0.56$ ($n = 82$), and with MMLU-Pro, a proxy for general knowledge, at $\rho = +0.63$ ($n = 57$). On the 57 models with both measures, the correlations are statistically indistinguishable (Williams' $t = +0.04$, $p = .96$), mirroring effects reported in human research on intelligence and creativity (Gerwig et al., 2021; Jauk et al., 2013).

Notably, the same separability holds for model size. Among the 42 open-weight models with disclosed parameter counts, total parameters correlate with the composite at $\rho = +0.69$, weakening to $\rho = +0.37$ when counting active parameters (Appendix K). Taken together, *C* tracks general intelligence and model scale at moderate-to-strong magnitudes, consistent with a creative ability that draws on both reasoning and knowledge yet is not reducible to either.

4.5 AGC-HUMAN IS MORE SENSITIVE TO CREATIVITY INSTRUCTIONS THAN TO ENABLING REASONING

A creativity benchmark should respond more to manipulations that target creativity than to manipulations that target reasoning. We tested sensitivity to creativity instructions with two within-model interventions on AGC-Human, run on separate model cohorts because not all models support a clean reasoning-mode toggle. Prompting 31 frontier models to "be creative" versus "be effective" produced a large upward shift ($d_z = +1.40$ after adjustment for response length). In contrast, enabling reasoning (vs. baseline) on 10 different reasoning-capable models produced a much smaller composite shift ($d_z = +0.34$ length-residualized), roughly one-quarter the magnitude. This comparison demonstrates sensitivity to creativity instructions, consistent with instruction effects in human divergent thinking (Acar et al., 2020); instruction following can contribute to the gain. Independent validity evidence comes from cross-domain prediction, scoring-family controls, and reasoning controls (Section 4.3; see Appendix I for the intervention designs).

4.6 HUMAN AND LLM CREATIVITY ON AGC-HUMAN

On the five AGC-Human tasks, the top human scored $+0.65$ and the top LLM scored $+0.53$ on the novelty and diversity composite, a difference of 0.12 standardized units. The comparison includes 201 humans and 80 LLMs. Consistency across tasks did not differ significantly between the groups (standardized $\alpha = .57$ for LLMs and $.64$ for humans; Appendix L). Across 80 LLMs, human

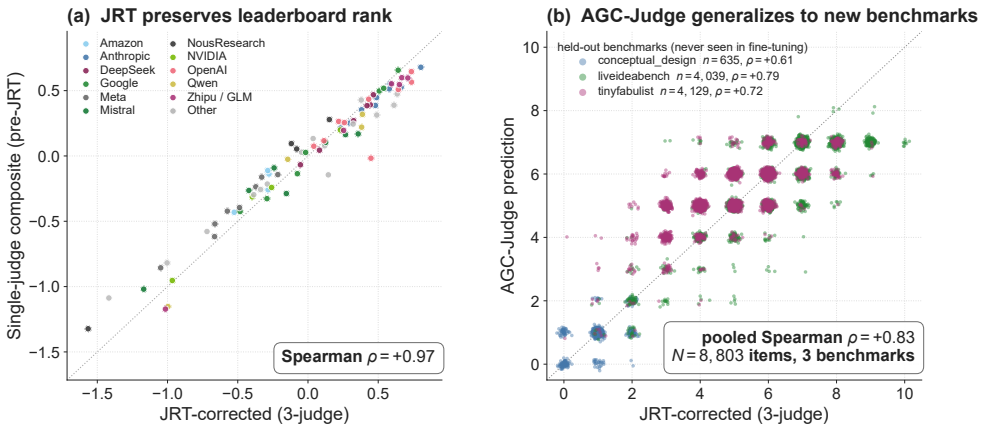


Figure 3: **JRT calibration validation and AGC-Judge generalization. (a)** JRT-corrected and single-judge canonical composites cluster on the diagonal at Spearman $\rho = +0.97$ across 83 models—calibration broadly preserves leaderboard rank. **(b)** AGC-Judge (open-weight Qwen3-30B LoRA fine-tuned on 48,299 JRT-corrected ratings) generalizes to three held-out creativity benchmarks at pooled Spearman $\rho = +0.83$ on 8,803 items against JRT-corrected panel scores.

ratings were strongly and positively correlated with AGC-Human embedding scores ($\rho = .653$) and AGC-Bench scores ($\rho = .646$), supporting the validity of the automated measures.

5 Conclusion

AGC-Bench provides a psychometric tool for LLM creativity evaluation, offering insight into the domain-specificity of creativity and how creative ability relates to intelligence in AI systems. The pipeline assembles 67 benchmarks under a systematic review, applies Judge Response Theory calibrated LLM-judge training for automated creativity scoring, and releases the open-weight AGC-Judge scorer that reproduces model rankings from the calibrated panel. Using principal component analysis, we find that verbal creativity is strongly unidimensional, related to but separable from knowledge and reasoning, with frontier models competitive with human performance on ideation and creative writing tasks. On the novelty and diversity composite, the highest-scoring human outperforms the highest-scoring LLM. We view AGC-Bench as scaffolding for the next phase of work on LLM creativity, with future efforts targeting shorter scales for cost-efficient evaluation (Polo et al., 2024; Kipnis et al., 2025), extension into different modalities such as audio and vision, and ultimately the use of these tools to improve creativity in current models.

Limitations. A few caveats are worth noting. The composite creativity score is specific to this cohort, so scores have no theoretical ceiling similar to Chatbot Arena ELO—a feature of any benchmark scored against frontier peers rather than a defect. AGC-Bench is currently English-dominant, with analyses restricted to the primary benchmark set. Extensions to other languages and modalities are natural targets for follow-on work, especially the visual tasks, which are slated for the next release of AGC-Bench. Our checks cannot establish that the evaluated set represents all 497 candidate benchmarks or rule out paraphrased exposure and familiarity with common task formats. Finally, although our benchmark covers many creative tasks, some aspects of creativity remain harder to assess, including transformational creativity, where a model reshapes the conceptual space itself rather than simply recombining within it (Schapiro et al., 2025). Nevertheless, our findings provide the first large-scale test of the structure of artificial creativity and how it may scale with increasing model capability.

AI use statement

We used generative AI tools to assist with benchmark implementation, data processing/analyses, and manuscript preparation. The authors take responsibility for the final manuscript, code, and results.

REFERENCES

Selcuk Acar, Mark A. Runco, and Hyeri Park. What should people be told when they take a divergent thinking test? a meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1):39–49, 2020. doi: 10.1037/aca0000256.

Teresa M. Amabile. *Creativity in Context: Update to the Social Psychology of Creativity*. Westview Press, Boulder, CO, 1996. ISBN 9780813330341.

John Baer. Is creativity domain specific? In James C. Kaufman and Robert J. Sternberg (eds.), *The Cambridge Handbook of Creativity*, pp. 321–341. Cambridge University Press, 2010. doi: 10.1017/CBO9780511763205.021.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026. doi: 10.1038/s41598-025-25157-3. URL https://www.nature.com/articles/s41598-025-25157-3.

Ninad Bhat, Kieran Browne, and Pip Bingemann. Creativity benchmark: A benchmark for marketing creativity for large language models. *arXiv preprint arXiv:2509.09702*, 2025. URL https://arxiv.org/abs/2509.09702.

Shelley H. Carson, Jordan B. Peterson, and Daniel M. Higgins. Reliability, validity, and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, 17(1):37–50, 2005. doi: 10.1207/s15326934crj1701_4.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7370–7388, 2023. doi: 10.18653/v1/2023.findings-acl.465. URL https://aclanthology.org/2023.findings-acl.465/.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024. doi: 10.1145/3613904.3642731.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. URL https://arxiv.org/abs/1911.01547.

Jennifer Diedrich, Emanuel Jauk, Paul J. Silvia, Jeffrey M. Gredlein, Aljoscha C. Neubauer, and Mathias Benedek. Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts*, 12(3):304–316, 2018. doi: 10.1037/aca0000137.

Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, and Dahua Lin. Creation-MMBench: Assessing context-aware creative intelligence in MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 447–456, October 2025. doi: 10.1109/ICCV51701.2025.00049. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fang_Creation-MMBench_Assessing_Context-Aware_Creative_Intelligence_in_MLLMs_ICCV_2025_paper.html.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. EleutherAI/lm-evaluation-harness: v0.4.3, July 2024. URL https://zenodo.org/records/12608602.

Anne Gerwig, Kirill Miroshnik, Boris Forthmann, Mathias Benedek, Maciej Karwowski, and Heinz Holling. The relationship between intelligence and divergent thinking—a meta-analytic update. *Journal of Intelligence*, 9(2):23, 2021. doi: 10.3390/jintelligence9020023.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830, 2025. doi: 10.1162/tacl.a.20. URL https://aclanthology.org/2025.tacl-1.37/.

Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14504–14528, 2023. doi: 10.18653/v1/2023.findings-emnlp.966. URL https://aclanthology.org/2023.findings-emnlp.966/.

Erik E. Guzik, Christian Byrge, and Christian Gilde. The originality of machines: AI takes the Torrance Test. *Journal of Creativity*, 33(3):100065, 2023. doi: 10.1016/j.yjoc.2023.100065.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, 2023. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41/. Best Paper Award.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965. doi: 10.1007/BF02289447.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. CreativityPrism: A cross-domain evaluation framework for large language model creativity. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2510.20091.

Kent F. Hubert, Kim N. Awa, and Darya L. Zabelina. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14(1):3440, 2024. doi: 10.1038/s41598-024-53303-w.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity*, 2025. URL https://arxiv.org/pdf/2411.02316.

Myeongjun Erik Jang and Fran Silavong. InstaJudge: Aligning judgment bias of LLM-as-judge with humans in industry applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1158–1172, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-industry.82. URL https://aclanthology.org/2025.emnlp-industry.82/.

Emanuel Jauk, Mathias Benedek, Beate Dunst, and Aljoscha C. Neubauer. The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, 41(4):212–221, 2013. doi: 10.1016/j.intell.2013.03.003.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14317–14332, 2023. doi: 10.18653/v1/2023.emnlp-main.885. URL https://aclanthology.org/2023.emnlp-main.885/.

Dan R. Johnson, James C. Kaufman, Brendan S. Baker, John D. Patterson, Baptiste Barbot, Adam E. Green, Janet van Hell, Evan Kennedy, Grace F. Sullivan, Christa L. Taylor, Thomas Ward, and Roger E. Beaty. Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55(7):3726–3759, 2023. doi: 10.3758/s13428-022-01986-2.

James C. Kaufman. Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4):298–308, 2012. doi: 10.1037/a0029751.

James C. Kaufman and Ronald A. Beghetto. Beyond big and little: The Four C model of creativity. *Review of General Psychology*, 13(1):1–12, 2009. doi: 10.1037/a0013688.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4T33izzFpK.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4-Abstract-Conference.html.

Martha Lewis and Melanie Mitchell. Evaluating the robustness of analogical reasoning in large language models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=t5cy5v9wph.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4437–4452, 2022. doi: 10.18653/v1/2022.naacl-main.330. URL https://aclanthology.org/2022.naacl-main.330/.

Zhuo Liu, Moxin Li, Xun Deng, Qifan Wang, and Fuli Feng. Assistant-guided mitigation of teacher preference bias in LLM-as-a-judge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9610–9631, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.510. URL https://aclanthology.org/2025.findings-emnlp.510/.

Li-Chun Lu, Miri Liu, Pin Chun Lu, Yufei Tian, Shao-Hua Sun, and Nanyun Peng. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6329–6352, 2026. doi: 10.18653/v1/2026.eacl-long.297. URL https://aclanthology.org/2026.eacl-long.297/.

Kevin S. McGrew. The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In Dawn P. Flanagan and Patti L. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, pp. 136–182. Guilford Press, New York, NY, 2 edition, 2005.

Nils Myszkowski and Martin Storme. Judge response theory? a call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):167–175, 2019. doi: 10.1037/aca0000225.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S. L. Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. Which LLMs get the joke? probing non-STEM reasoning abilities with HumorBench. *arXiv preprint arXiv:2507.21476*, 2025. URL https://arxiv.org/abs/2507.21476.

Surabhi S. Nath, Guiomar del Cuvillo y Schröder, and Claire E. Stevenson. Pencils to pixels: A systematic study of creative drawings across children, adults and AI. In *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, volume 47, pp. 1123–1130, San Francisco, CA, USA, 2025. Cognitive Science Society. URL https://escholarship.org/uc/item/1rq6h8xg.

William Orwig, Emma R. Edenbaum, Joshua D. Greene, and Daniel L. Schacter. The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior*, 58(1):128–136, 2024. doi: 10.1002/jocb.636.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71, 2021. doi: 10.1136/bmj.n71.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37*, pp. 68772–68802, 2024. doi: 10.52202/079017-2197. URL https://papers.nips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

John D. Patterson, Jimmy Pronchick, Ruchi Panchanadikar, Mark Fuge, Janet G. van Hell, Scarlett R. Miller, Dan R. Johnson, and Roger E. Beaty. CAP: The creativity assessment platform for online testing and automated scoring. *Behavior Research Methods*, 57(9):264, 2025. doi: 10.3758/s13428-025-02761-9.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models? In Kazjon Grace, Maria Teresa Llano, Pedro Martins, and Maria M. Hedblom (eds.), *Proceedings of the 15th International Conference on Computational Creativity*, pp. 226–235, Jönköping, Sweden, 2024. Association for Computational Creativity. URL https://computationalcreativity.net/iccc24/papers/ICCC24_paper_70.pdf.

Jonathan A. Plucker. Generalization of creativity across domains: Examination of the method effect hypothesis. *The Journal of Creative Behavior*, 38(1):1–12, 2004. doi: 10.1002/j.2162-6057.2004.tb01228.x.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/polo24a.html.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating LLMs' divergent thinking capabilities for scientific idea generation with minimal context. *Nature Communications*, 17(1):3625, 2026. doi: 10.1038/s41467-026-70245-1. URL https://www.nature.com/articles/s41467-026-70245-1.

Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

Sameh Said-Metwaly, Belén Fernández-Castilla, Eva Kyndt, and Wim Van den Noortgate. Testing conditions and creative performance: Meta-analyses of the impact of time limits and instructions. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1):15–38, 2020. doi: 10.1037/aca0000244.

Samuel Schapiro, Jonah Black, and Lav R. Varshney. Transformational creativity in science: A graphical theory. In Hugo Gonçalo Oliveira, Brad Spendlove, Pablo Gervás, and Dan Ventura (eds.), *Proceedings of the 16th International Conference on Computational Creativity*, pp. 258–263, Campinas, Brazil, 2025. Association for Computational Creativity. URL https://computationalcreativity.net/iccc25/papers/iccc25-schapiro2025transformational.pdf.

Claire Stevenson, Matthijs Baas, and Han L. J. van der Maas. A minimal theory of creative ability. *Journal of Intelligence*, 9(1):9, 2021. doi: 10.3390/jintelligence9010009.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity (ICCC'22)*, pp. 164–168, 2022. URL https://computationalcreativity.net/iccc22/wp-content/uploads/2022/08/ICCC22_proceedings.pdf.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3GTtZFiajM.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K. Jain, Robert D. Nowak, Bob Mankoff, and Jifan Zhang. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16273–16287, 2025. doi: 10.18653/v1/2025.findings-emnlp.884. URL https://aclanthology.org/2025.findings-emnlp.884/.

A PRISMA FLOW DIAGRAM

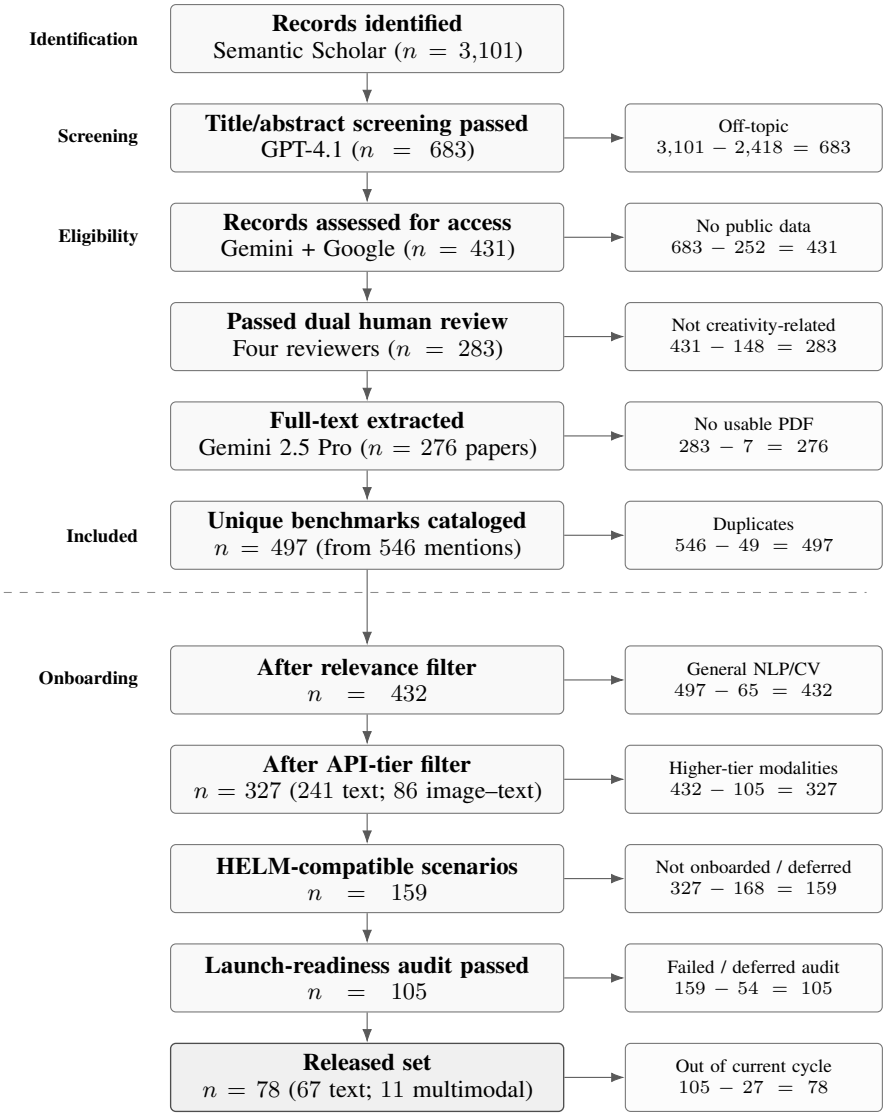


Figure S1: PRISMA-based flow of papers and benchmarks through the meta-benchmark pipeline. Top five tiers follow PRISMA 2020 reporting conventions; the dashed rule separates the systematic review from the benchmark onboarding stage, which filters the cataloged benchmarks to API-runnable text and image-text tiers and converts the released set into runnable evaluation scenarios. Right-hand boxes show the arithmetic for the transition into the corresponding left-hand box.

B DETAILED ONBOARDING METHODOLOGY

This appendix provides the detailed pipeline architecture, failure-mode analysis, and robustness additions that support the summary onboarding description in Section 3.2.

B.1 AGENTIC PIPELINE ARCHITECTURE

Manual onboarding of 327 candidate benchmarks (241 text-to-text plus 86 image-to-text) would have been prohibitively slow, so we developed an agentic workflow built on Claude Code. The

workflow is encapsulated in a reusable skill[1] (`benchmark-onboarder`) implemented as a deterministic nine-state machine that walks each candidate from PDF parsing through data loading, semantic validation, prompt and judge extraction, scenario code generation, code review against the paper's stated metric, and a pilot run on a small sample. State transitions are controlled by deterministic code rather than agent decisions, and LLM reasoning is bounded to specific states with well-defined inputs and outputs. Each onboarding run used `google/gemini-2.0-flash` for structured extraction and `anthropic/claude-sonnet-4` as the review and judge model. The full skill specification, per-state prompts, and transition logic are released in the artifact bundle (`skills/benchmark-onboarder/`).

B.2 Human-in-the-loop verification

Although the agentic workflow automated the bulk of scenario construction, each onboarded benchmark passed through a human review step before acceptance. Four team members were each assigned 60–71 candidate benchmarks (mean 68) and reviewed the agent's output for prompt fidelity, correctness of the dataset schema and appropriateness of the evaluation function. Reviewers verified URLs against the original paper, corrected errors in the scenario file, and either approved the benchmark for inclusion or documented the reason for rejection in a shared tracking file. A `learnings.md` document accumulated cross-team discoveries about common failure modes and was consulted before each onboarding session; entries were added automatically by the agent whenever it encountered a new dataset quirk (e.g., test splits without labels, fields requiring non-standard loading parameters, field-name mismatches between documentation and data).

B.3 Failure modes and mitigations

A review of 155 completed onboarding attempts surfaced recurring failure patterns at the agent stage (before human review and acceptance), each addressed by a specific validation step in the pipeline. The most common pattern was *hallucinated evaluation logic*, which arose when the source paper did not pin down a scoring function precisely; the VALIDATE-IMPL state now compares the agent's generated metric against the paper's stated metric before pilot execution, and the agent must cite a source file or paper section for each evaluation function. The second-most-common pattern was *missing LLM-judge configuration*, addressed by requiring judge-field extraction in the EXTRACT state. *Prompt hallucination* (the agent paraphrasing rather than extracting prompts) was reduced by prioritizing regex and code-based extraction over LLM paraphrasing and flagging any prompt that could not be traced to a specific source location. Lower-frequency failures (metric-name mismatches between paper and code, semantic drift from the paper's task, multimodal benchmarks silently downcast to text-only, task-direction inversions) are documented in the released learnings index alongside their corresponding pipeline checks.

C Example prompts by domain

We illustrate the prompt format with excerpts from one benchmark per domain. Full prompt templates are provided in the released harness (`https://huggingface.co/agcbench-2026`).

Brainstorming *(sdat)*.

```
Please enter 10 words that are as different from each other
as possible, in all meanings and uses of the words.  Rules:
Only single words in English.  Only nouns (e.g., things,
objects, concepts).  No proper nouns (e.g., no specific
people or places).  No specialised vocabulary (e.g., no
technical terms).  Think of the words on your own (e.g., do
not just look at objects in your surroundings).  Make a list
of these 10 words, a single word in each entry of the list.
```

[1] A Claude Code *skill* is a structured set of instructions, code, and reference materials that directs an agent through a multi-step task.

Problem Solving *(macgyver)*.

```
Your desk drawer has become misaligned and difficult to
open.  However, the desk is made of metal.  Tools available
include a bottle of water, a keyring, a pair of pajamas, a
hand towel, a coaster, a bottle opener, a shoebox, a pair of
scissors, a coffee mug, a makeup compact, a paint roller, an
ice cube tray.  How do you realign the drawer and open it
without damaging its contents?

Give a valid (feasible and efficient) solution very
concisely.  Use step1, step2, etc, and mention the tools
to achieve each step.  Use as few steps as possible and the
answer should ideally be less than 100 words....
```

STEM *(futuregen)*.

```
You are an AI trained to analyze scientific research and
suggest future directions based on the content of a paper.
Below, you will find sections from a scientific article
including the 'Abstract', 'Introduction', and 'Conclusion'.
Based on these details, please generate comprehensive and
plausible future work suggestions that could extend the
research findings, address limitations, and propose new
avenues for exploration.

{abstract / introduction / conclusion of the paper}...
```

Story / Narrative *(eqbench_creative_writing_v3)*.

```
Darkroom Revelations
Write a first-person narrative from the perspective of Dani,
a 25-year-old assertive art student known for her bold style
and take-charge attitude in relationships.  For the past
semester, she's been quietly fascinated by Elliot...Set
during one of these intimate darkroom sessions where the
red light creates a confessional atmosphere.  Explore Dani's
complex attraction...The narrative should capture the tender
awkwardness of their conversations...
```

Figurative Language *(metaphor_generation)*. The input sentence is preceded by the instruction: "Rewrite the literal sentence below as exactly one metaphorical sentence. Preserve the original meaning while making the main verb phrase more metaphorical. Output only the rewritten sentence."

```
I wrestled with this decision for years
```

Humor *(showerthoughts)*.

```
Please generate one Showerthought, which is inspired by
the Reddit community r/Showerthoughts.  Try to be clever,
creative, and funny.  The Showerthought should be relatable
and connected to things that people might encounter during
mundane tasks.  Return only the Showerthought text as one
concise standalone statement.

Showerthought:
```

D PER-DATASET DOMAIN ASSIGNMENTS

Each benchmark maps to one of the six domains in the creativity taxonomy (Section 3.3). The assignments come from a three-model LLM panel (Gemini 3 Flash, Grok-4.1-Fast, GPT-4.1-mini)

prompting on the benchmark's title, task description, example prompts, scoring method, and source-paper abstract, with disagreements resolved by majority vote and edge cases adjudicated by the lead authors. Mean pairwise Cohen's $\kappa \approx 0.85$ across the three LLM raters (Fleiss' $\kappa \approx 0.85$ on the released $n = 75$ taxonomy panel) supports the domain assignment. Per-dataset domain assignments, scoring protocol, source paper, license, canonical metric, and item-level statistics are released as machine-readable CSV.

E CANONICAL SCORING-METRIC HETEROGENEITY

The released harness includes formula-based metrics, embedding measures, and LLM-judge ratings. JRT calibration covered 24 benchmarks in the primary analysis. The released dataset metadata identify the benchmarks with calibrated scores.

F MODELS

The 83 models we evaluated span 22 providers and cover proprietary, open-weight, and reasoning-focused variants released between 2024 and 2026. Table S1 reports the top 30 models ranked by AGC-Bench composite. The full per-model roster (including provider, version string, parameter count where publicly documented, access method, reasoning-mode support, datasets covered, mean and median z, and rank) is released as machine-readable CSV.

Table S1: Top 30 models by AGC-Bench composite, computed by standardizing benchmark scores across the 83 models and averaging each model's available scores. n_{ds} is the number of datasets on which the model produced a parseable score. All entries shown are models that scored on at least 65 of 67 datasets.

Rank	Provider	Version	n_{ds}	Mean z	Median z
1	anthropic	`claude-opus-4.7`	67	+0.804	+0.884
2	openai	`gpt-5.4`	67	+0.737	+0.931
3	openai	`gpt-5.5`	67	+0.736	+0.796
4	z-ai	`glm-5.1`	66	+0.711	+0.697
5	z-ai	`glm-5-turbo`	66	+0.668	+0.783
6	anthropic	`claude-opus-4.6-fast`	66	+0.663	+0.822
7	z-ai	`glm-5`	66	+0.651	+0.700
8	openai	`gpt-5.1`	66	+0.643	+0.790
9	google	`gemma-4-31b-it`	67	+0.642	+0.670
10	moonshotai	`kimi-k2.5`	67	+0.635	+0.730
11	moonshotai	`kimi-k2-0905`	66	+0.612	+0.655
12	z-ai	`glm-4.7`	66	+0.595	+0.643
13	anthropic	`claude-opus-4.5`	67	+0.585	+0.772
14	anthropic	`claude-opus-4.6`	65	+0.579	+0.837
15	google	`gemma-4-26b-a4b-it`	67	+0.539	+0.610
16	google	`gemini-2.5-flash`	66	+0.508	+0.542
17	moonshotai	`kimi-k2.6`	67	+0.490	+0.618
18	anthropic	`claude-sonnet-4.5`	67	+0.489	+0.767
19	anthropic	`claude-sonnet-4.6`	67	+0.480	+0.707
20	deepseek	`deepseek-v3.2-exp`	66	+0.465	+0.589
21	openai	`gpt-5.4-mini`	66	+0.448	+0.525
22	deepseek	`deepseek-v3.1-terminus`	66	+0.444	+0.581
23	openai	`gpt-4.1`	66	+0.432	+0.414
24	deepseek	`deepseek-v3.2`	66	+0.420	+0.491
25	qwen	`qwen3.5-flash-02-23`	67	+0.388	+0.363
26	qwen	`qwen3-next-80b-a3b-instruct`	66	+0.382	+0.382
27	anthropic	`claude-sonnet-4`	66	+0.381	+0.603
28	baidu	`ernie-4.5-300b-a47b`	66	+0.379	+0.321
29	mistralai	`mistral-medium-3.1`	67	+0.357	+0.399
30	deepseek	`deepseek-chat-v3.1`	66	+0.325	+0.407

Additional model providers (ai21, allenai, arcee-ai, deepcogito, ibm-granite, minimax, morph, nex-agi, nousresearch, tencent) each contribute one to four models. The released roster CSV documents per-provider model counts for them alongside nvidia, meta-llama, amazon, baidu, mistralai, deepseek, moonshotai, qwen, google, z-ai, openai, and anthropic.

G DATA-QUALITY AUDIT DETAILS

The audit summarized in Section 3.4 assessed whether sampled model responses were on-task and interpretable.

Judge model and rubric. The audit judge was x-ai/grok-4.1-fast, chosen for parser robustness against verbose judge replies. The audit measures response validity using binary flags. A second judge outside the JRT panel assessed agreement on these validity judgments. The judge sampled three random instances per cell, saw the rendered prompt and the full completion, and returned three boolean flags per response: *on-task*, *garbled* (empty, refusal, gibberish, or cut off before producing meaningful content), and *uninterpretable* (rare). The cell-level composite is valid_response $= \neg$garbled $\wedge$ on_task. Multimodal cells were re-rated by the same judge under a vision-aware pass with the input image attached.

Coverage. The audit covered $6{,}635$ text-only model-dataset combinations across the full set of evaluated models (a superset of the 83 used in the main analyses), with zero judge errors. The on-task rate was 95.1% and the invalid rate was 5.5%. We additionally ran a vision-aware pass on the 730 multimodal model-dataset combinations, obtaining 94.4% on-task rate and 6.1% invalid rate.

Validation. A $2{,}725$-cell stratified subsample was re-rated by deepseek-v4-flash, drawn from a different model family than the primary audit judge and outside the JRT scoring panel. The binary valid_response composites agreed at $\kappa = 0.67$ (substantial agreement on the Landis–Koch scale). Per-instance Likert agreement was not computed for this overlap because the v4 audit pass recorded only the binary validity flags rather than a 1–5 quality score.

Truncation patches. Frontier 2026 models produce verbose outputs that exceeded the original source-paper token budgets on some open-ended generation datasets. We detected truncation by comparing generation length to the configured max_tokens and re-ran 16 datasets at raised token caps (typically $4{,}096 \rightarrow 8{,}192$). critics_story still exceeded budget at the higher cap and was excluded. fann_or_flop was retained despite verbose-tail truncation, with the per-dataset truncation rate documented as a caveat.

Pre-release data-quality sweep and cascade impact. A pre-release sweep cross-referenced the on-task audit above with heuristic flags on the raw scenario_state files (empty output, refusal, repetitive sequence collapse). 32 cells flagged by both signals were masked at the score level (dq_masked in the released long table) rather than scored as failed attempts. Pre-release cascade-impact analysis found little change in the creativity component and intelligence correlations under the mask versus an unmasked run. Per-cell flags and the cascade-impact log are released in the artifact bundle (audit/dq_sweep/).

H JUDGE CALIBRATION: GRM SPECIFICATION, AUDIT, AND AGC-JUDGE DETAILS

This appendix specifies the planned-missing 2-of-3 design (PM1-50), the Bayesian Graded Response Model used to fit per-judge severity and ability for each response, the audit of parsing failures that had produced zero scores, the per-benchmark inter-judge agreement table, and the AGC-Judge fine-tuning details summarized in Section 3.5.

H.1 PLANNED-MISSING 2-OF-3 DESIGN AND PARSER AUDIT

Each of the 92,080 (benchmark, model, item) units in the 24 LLM-judge canonical-metric cells was rated by two of three vendor judges (`google/gemini-3-flash-preview`, `x-ai/grok-4.1-fast`, `openai/gpt-4.1-mini`) under a planned-missing 2-of-3 design, with each unit assigned one of three judge pairs (seed 42). The design targets 184,160 (rater $\times$ unit) cells; 182,924 produced a non-sentinel rating (99.3% completion) and the residual missing observations are masked at the unit level in the joint-rater fit. A data-integrity patch fixed an existing parser that extracted the first integer in 23 of 35 LLM-judge annotators that returned silent zeros on parse failures and grabbed stray integers from verbose judge replies. The patched parser prefers labeled patterns ("Score: N"), falls back to the last integer within the permitted range, and emits the missing-data sentinel otherwise; it reduced contamination from 0.32% to 0%, with historical out-of-range ratings re-coded as missing for the JRT fit.

H.2 BAYESIAN GRADED RESPONSE MODEL

For each (benchmark, metric) cell with N units and R raters, we fit ability θ_p for each response, per-rater discrimination α_r, and per-rater ordered thresholds $\beta_{r,k}$ via a Graded Response Model:

$$P\big(\text{rating}_{r,p} \geq k+1\big) = \sigma(\alpha_r(\theta_p - \beta_{r,k}))\,, \tag{1}$$

with $\theta_p \sim \mathcal{N}(0,1)$, $\log \alpha_r \sim \mathcal{N}(0, 0.3)$, and ordered $\beta_{r,:}$ drawn from $\mathcal{N}(0,2)$. Inference uses NumPyro stochastic variational inference (AutoNormal guide, Adam at learning rate 0.05, 800 update steps per cell), fit cell-by-cell so per-cell rater identifiers and rubric scales remain isolated. All 24 cells reached converged ELBO; per-judge severity ranges from $\beta = -0.90$ (`gpt-4.1-mini`) to $\beta = +0.04$ (`gemini`), with `grok-4.1-fast` at $\beta = -0.46$. Full pipeline configuration and per-cell fit diagnostics are released in the artifact bundle.

H.3 PER-BENCHMARK INTER-JUDGE AGREEMENT

Pairwise Spearman ρ on co-rated units spans 0.30 (`showerthoughts`) to 0.88 (`eqbench_creative_writing_v3`); agreement is highest for benchmarks with concrete rubrics and lowest for short, subjective tasks. The pattern is rater-invariant (all three judge pairs track together across benchmarks), so low agreement reflects task ambiguity rather than a single outlier judge. The full per-(benchmark, judge-pair) table is released as machine-readable CSV in the artifact bundle.

H.4 DRIFT OF VALIDATION CORRELATIONS UNDER JRT CORRECTION

The JRT correction recalibrates ability θ for each response without destroying the rank signal raw means already carry. Per-unit raw and JRT-corrected scores correlate at $\rho = 0.97$, and the composite leaderboard recomputed against JRT-corrected supervision agrees with the single-judge leaderboard at Spearman $\rho = 0.97$ across 83 models, using the same 67 benchmarks and quality masks.

H.5 AGC-JUDGE TRAINING DETAILS

Converting latent scores to integer training targets. The GRM produces a continuous latent ability θ_p on the standard normal scale, but AGC-Judge fine-tuning targets the integer rubric range each source paper uses (1–5, 10–50, 0–100, etc.). For each unit the training target is recovered by inverting the GRM at the per-rater severity thresholds: the integer score $k \in \{1, \ldots, K\}$ chosen is the one for which the GRM's cumulative probability $P(\text{rating}_{r,p} \geq k)$ crosses 0.5 under the calibrated $(\alpha_r, \beta_{r,k})$ for the rater whose severity is closest to zero (the reference rater), then rounded to the nearest integer in the rubric range. The resulting target is on the source paper's native scale and is what AGC-Judge predicts directly.

Training data was filtered to drop the 4.5% of rows whose total chat-template length exceeded 7,500 approximate tokens (characters / 4), to fit within the Together fine-tune harness's 8,192-token sequence budget while leaving headroom for tokenizer drift. The filter dropped 2,298 of 50,597 training rows and 345 of 7,228 validation rows. Token outliers concentrated in benchmarks with

long instruction text (Arabic-language story benchmarks, multi-turn dialogue benchmarks) and contribute negligible learning signal at high cost in compute. The held-out test, held-out-models, and held-out-benchmarks splits retain outliers to evaluate generalization across the full input distribution. Other hyperparameters: LoRA $r = 16$, $\alpha = 32$, learning rate $1e{-}4$, 1 epoch, AdamW with default Together SFT settings (`train_on_inputs = "auto"`, masking system and user tokens during loss computation). The fine-tune ran in 20 minutes on $2\times$ H100-80GB SXM at a one-time cost of \$211. The evaluation splits contained $14{,}457$ test responses, $10{,}966$ responses from held-out models, and $8{,}803$ responses from held-out benchmarks. The reported correlations used $14{,}455$, $10{,}965$, and $8{,}803$ available predictions, respectively. The released held-out evaluation reports parse-failure rates and per-bench correlations, with the full held-out-bench scatter (pooled $\rho = 0.83$) released in the artifact bundle. AGC-Judge is trained exclusively on LLM-produced responses with frontier-LLM ratings; for cross-source (human vs. LLM) scoring on AGC-Human cells, we use a fairness-aware inference prompt that discloses the LLM self-preference bias documented in prior judge work (Panickssery et al., 2024; Ye et al., 2025), instructs the judge to focus on the underlying idea rather than stylistic register, and labels the response source before scoring (an instance of the prompt-based debiasing strategies validated by Liu et al., 2025 and Jang & Silavong, 2025). The prompt template, validation results, and a counterfactual style-flip experiment are documented in Appendix L.

H.6 Training on raw panel ratings

We trained a second AGC-Judge on the same $48{,}299$ examples, replacing JRT-derived targets with raw frontier-panel means rounded half up and clipped to each rubric's range. Prompts, example order, splits, and the Qwen3-30B-A3B-Instruct-2507 base model were unchanged. The raw run used rank-16 LoRA, one epoch, learning rate 10^{-4}, batch size 16, and seed 42, with no validation-set evaluation during training. We compared it with the original JRT adapter; the original training settings were not fully matched, so differences cannot be attributed exclusively to target transformation.

Table S2 evaluates both adapters on identical responses against continuous raw panel means, distinct from the JRT-corrected targets in Figure 3. Both returned valid numeric scores throughout. Normalized MAE divides absolute error by the rubric range; macro Spearman averages correlations within benchmark–metric combinations. These are descriptive point estimates.

Table S2: Raw-target versus JRT-target training, evaluated against raw frontier-panel means. Each metric reports raw / JRT; lower MAE is better.

Split	Responses	Pooled ρ	Normalized MAE	Macro ρ
Test	14,457	.9288 / .9289	.1381 / .1663	.5973 / .6916
Held-out models	10,966	.9381 / .9361	.1363 / .1602	.6033 / .6970
Held-out benchmarks	8,803	.7436 / .7588	.1164 / .1322	.7130 / .6980

Pooled test correlations are nearly identical. Raw targets reproduce the panel's absolute scale more closely across all splits; JRT preserves stronger average ranking agreement within each benchmark and metric on test and held-out models. Results on held-out benchmarks are mixed, supporting a tradeoff rather than uniform superiority.

H.7 External validation on the Orwig creative writing data

As an out-of-distribution check that does not appear in any AGC-Bench training or evaluation split, we ran AGC-Judge on the Orwig short-story corpus (Orwig et al., 2024): 718 stories on six 3-word cue triplets, with conditions split as human ($n = 300$), GPT-3 ($n = 298$), and GPT-4 ($n = 120$). Each story carries a human-rater creativity rating averaged across multiple raters, an existing GPT-judge creativity rating from the original Orwig pipeline, a divergent semantic integration (DSI) score, word count, and a perceptual-details measure. Critically for the source-bias question, the Orwig human raters do not show a strong source preference (humans 2.93 vs. GPT-4 2.94 mean creativity), so the data offer a fairer test of AGC-Judge alignment with humans than corpora in which raters strongly favored one source.

AGC-Judge under the human-source-attribution prompt (Section 3.5) recovered Spearman $\rho = +0.65$ against the human creativity ratings and $\rho = +0.80$ against the existing LLM ratings ($n = 718$). The two correlations were computed on the same data with the shared variable AGC-Judge, so we tested their difference with Williams' modification of Hotelling's t-test for dependent correlations: $t = -9.42$, $p < 0.0001$ on Pearson, $t = -7.98$, $p < 0.0001$ on Spearman. AGC-Judge is significantly more aligned with the LLM ratings than with the human ratings. Two pieces of context matter for interpreting that difference. First, AGC-Judge's correlation with human creativity ($\rho = +0.65$) is essentially identical to the original GPT-judge's correlation with humans on the same data ($\rho = +0.66$): AGC-Judge predicts human ratings about as well as a frontier GPT judge does, while also tracking the LLM-judge construct more tightly. Second, the source-preference gap on AGC-Judge (GPT-4 vs. human mean) is $+0.51$ on the 1–5 scale, compared with $+0.83$ for the original GPT-judge and $+0.01$ for human raters (Table S3). AGC-Judge inherits some of the LLM-stylistic preference of its supervision but at roughly 60% of the original LLM-judge magnitude.

Table S3: Per-condition mean creativity rating on the Orwig 718-story corpus, scored by AGC-Judge (this work, JRT-corrected supervision), human raters (Orwig original), and the GPT-judge in the original Orwig pipeline. Source-preference gap is the GPT-4 vs. human-condition mean. Smaller gap is better for source neutrality.

Condition	n	**AGC-Judge**	**Human raters**	**GPT-judge**
human	300	2.47	2.93	3.09
GPT-3	298	2.38	2.85	2.95
GPT-4	120	2.98	2.94	3.92
GPT-4 − human gap	—	+0.51	+0.01	+0.83

Within-condition correlations decompose the alignment further. On human-written stories, AGC-Judge is equally aligned with human and LLM ratings (both $\rho \approx +0.74$). On LLM-written stories, AGC-Judge tracks the LLM-judge much more tightly than the human raters (GPT-4 stories: $\rho = +0.86$ vs. $\rho = +0.65$). We treat the Orwig result as evidence that AGC-Judge generalizes its scoring construct beyond the AGC-Bench training distribution while inheriting a moderate, characterizable LLM-stylistic preference from its supervision.

I Intelligence assessments and intervention designs

This appendix details the letter-string analogy assessment and the two AGC-Human interventions summarized in Section 3.4.

Letter-string analogies as a targeted G_f indicator. The human creativity literature situates creative ability primarily against *fluid* intelligence (G_f), the component most directly implicated in creative idea generation (McGrew, 2005; Gerwig et al., 2021; Stevenson et al., 2021). Standard LLM intelligence batteries (GPQA, HLE, LiveCodeBench, SciCode, and the Artificial Analysis composite) blend G_f with substantial crystallized (G_c) and quantitative-knowledge (G_q) content, making them poorly suited as a clean creativity–G_f test. We therefore use the Lewis–Mitchell counterfactual letter-string analogy task (Lewis & Mitchell, 2025) as our primary G_f indicator: a Hofstadter-style rule-abstraction probe whose alphabets are permuted to block surface-pattern shortcuts. We administer 500 items spanning the task's transformation types, with reasoning off and temperature zero. As a robustness check, we also evaluate the cohort on each component of the AA battery (GPQA, HLE, LiveCodeBench, SciCode, and the AA composite); the AA-battery comparison appears in Appendix K.

Be-creative versus be-effective intervention. A be-creative prompt manipulation, robustly known in the human creativity literature to increase originality and diversity on divergent-thinking tasks (meta-analyses by Acar et al. (2020) and Said-Metwaly et al. (2020)), was applied to 31 frontier-cohort models on the five CAP tasks (AUT, Design, SCTT, Story, Metaphor—see Section 3.6) with three repetitions per cell. Each model was prompted twice for every item, once with a be-creative instruction and once with a be-effective instruction. Composite shifts are reported using paired differences within each model and task (152 pairs; $d_z = 1.40$) on the AGC-Human embedding

composite after adjustment for response length. An earlier validity audit covered 2,700 responses from 18 models; 99.4% met the validity criteria.

Reasoning on/off intervention. A second intervention toggled provider-supported thinking-mode prompts on a matched-pairs subset of 10 reasoning-capable frontier systems (`deepseek-v3.1-terminus`, `deepseek-v3.2`, `gemini-2.5-flash`, `gemini-3.1-flash-lite-preview`, `gpt-5.4-mini`, `qwen3-30b-a3b`, `grok-4.20`, `glm-4.5-air`, `glm-5-turbo`, `glm-5.1`). Anthropic Claude variants are excluded because the OpenRouter reasoning-disable flag was not reliably honoured at the time of the run. Composite shifts and per-sub-metric (diversity, DSI; surprise was dropped from the canonical embedding measures after the alignment-LM bias diagnostic) shifts are reported as within-model paired-difference effect sizes (50 model–task pairs; composite $d_z = .34$) to characterize whether reasoning produces uniform gains or splits across sub-metrics.

J Robustness of the creativity component

The creativity component reported in Section 4.3 is the first principal component of the correlation matrix of the six domain composites across 83 LLMs (eigenvalue 4.94, $\alpha = 0.96$, 82.4% variance, parallel-analysis $p_{95} = 1.53$ on 1,000 random matrices). Unless otherwise noted, α denotes standardized Cronbach's alpha. Variance percentages throughout this appendix follow the convention of dividing the eigenvalue by the number of domains, as in the body. We tested three threats to the result: capability confounding (whether c is a re-parameterization of general capability), scoring-class artifact (whether c reflects one scoring family rather than creative ability), and composition sensitivity (whether the structure depends on the specific set of 83 models). We additionally tested sensitivity to benchmark selection and prediction across domains using the same 67-benchmark matrix.

Capability control. Counterfactual letter-string analogy accuracy (Appendix I) was regressed out of each domain composite and principal component analysis was performed on the correlation matrix of the residuals. Result: eigenvalue 4.63, $\alpha = 0.94$, 77.1% variance ($n = 82$; one model lacks LSA coverage). The dominant component is preserved after partialling out fluid reasoning, indicating that c is not a re-parameterization of G_f.

Scoring-class invariance. We assigned benchmarks with any LLM-judge metric to the LLM-judge subset and benchmarks with formula-based metrics but no LLM-judge metric to the formula-based subset. We then extracted the first principal component for each subset. The formula-based subset (41 datasets, five domains) yielded eigenvalue 3.70, $\alpha = 0.91$, 74.1% variance. The LLM-judge subset (25 datasets, five domains) yielded eigenvalue 4.30, $\alpha = 0.96$, 86.0% variance. The model-based subset reduces to a single dataset and is omitted. The two scoring classes that recover meaningful subsets each emerge as unidimensional, indicating that c is not an artifact of any single scoring family.

Composition sensitivity. Two checks probe the creativity component's dependence on the specific composition. (a) *Dropping a domain.* Story / Narrative is the largest domain at 20 datasets. We re-ran the principal component analysis on the remaining five domains and recovered a similar structure: eigenvalue 4.09, $\alpha = 0.94$, 81.8% variance. Each remaining domain's variance share barely moves, indicating that the creativity component is not disproportionately driven by Story / Narrative. (b) *Subsampling models.* We repeatedly sampled 60 of the 83 models without replacement (500 draws; seed 42) and extracted the first principal component each time. Eigenvalue, α, and variance all stay tight: eigenvalue 4.95 (95% empirical percentile interval $[4.76, 5.09]$), $\alpha = 0.96$ $[0.95, 0.96]$, variance 82.5% $[79.4, 84.8]$. The component is stable under both perturbations.

The dominant creativity component is preserved across all of the above robustness checks.

Benchmark-selection sensitivity. We tested sensitivity to benchmark selection using the same 67-benchmark matrix. This 67-benchmark reference gives 82.37% first-component variance, raw $\alpha = .951$, standardized $\alpha = .957$, and component loadings .874–.947. We retained JRT-adjusted scores where available and canonical scores otherwise. Table S4 summarizes the domain correlation matrix and rank stability of the mean dataset-z composite. All table alphas are raw.

Table S4: Sensitivity to benchmark selection using the 67-benchmark reference. Half-benchmark brackets are 2.5th–97.5th percentiles across 1,000 domain-stratified subsets, not confidence intervals for a fixed parameter.

Selection	Variance (%)	α	Leaderboard ρ
Full set (67)	82.37	.951	1.000
Leave one out (67 runs)	80.06–82.99	.940–.953	.9987–.9999
Half per domain (median)	74.40	.923	.986
[95% empirical interval]	[68.95,79.05]	[.899,.941]	[.974,.992]
Sources dated 2024–2025 (43)	79.60	.943	.990

No single benchmark drives the structure, which remains stable across large subsets. These checks cannot establish representativeness of all 497 catalogued benchmarks: the other 430 were not evaluated on the same model cohort. Source dates reduce the time available for public circulation but do not establish that items postdate each model's training cutoff.

Prediction across domains. Using this matrix, repeated ten-fold cross-validation predicted each domain from the other five for held-out model IDs. Across the six domains, median R^2 ranged from .684 to .852 and Spearman ρ from .794 to .909. Leaving one provider out at a time yielded R^2 = .673–.849 and ρ = .787–.906. Providers were the prefixes of released model identifiers (22 groups); they are vendor labels, not audited architecture lineages. These results support prediction within the sampled domains and providers.

K Artificial Analysis intelligence battery: robustness check

The Artificial Analysis (AA) intelligence battery aggregates public benchmarks across knowledge, reasoning, coding, and agentic tasks. We report it here as a secondary check on the relationship between AGC-Bench and general capability. Our discriminant-validity analysis uses the pure-G_f letter-string analogy task instead (Appendix I) because the AA battery blends fluid reasoning with substantial G_q and G_c content, has saturated in the top tail, and covers only a subset of the cohort.

Table S5 reports per-indicator Spearman and Pearson correlations between the JRT-corrected AGC-Bench composite and each AA indicator, restricted to the 83 models we evaluated (intersected with AA coverage per indicator). The AA Intelligence Index correlates with the AGC-Bench composite at $\rho = +0.78$ ($n = 74$).

Parameter-count correlation on the open-weight subset. Parameter counts are not disclosed for the proprietary frontier (41 of 83 models), so an analysis of model size and creativity is feasible only on the open-weight subset. Across the 42 open-weight models with documented parameter counts, total parameters correlate with the JRT-corrected AGC-Bench composite at Spearman $\rho = +0.69$ ($p < 10^{-6}$, Pearson $r = +0.66$ on $\log_{10}$ total parameters, Figure S2). The relationship weakens under active-parameter accounting ($\rho = +0.37$, $p = .015$). Within the dense-only sub-cohort ($n = 22$) the correlation is $\rho = +0.54$ ($p = 0.009$); within the MoE-only sub-cohort ($n = 20$) total params yield $\rho = +0.33$ ($p = 0.16$, n.s.) and active params yield $\rho = +0.18$ ($p = 0.45$, n.s.). Per-model parameter assignments and architectural classification are released as machine-readable CSV in the artifact bundle (`model_size_correlation.csv`).

L AGC-Human comparisons and validation

AGC-Human and the broader AGC-Bench composite ask the same question with different instruments: one applies embedding measures to five paired-human CAP tasks, the other aggregates source-paper canonical metrics across 67 datasets. Across the 80 LLMs with coverage on both, the two rankings agree at Spearman $\rho = +.65$ ($p < 10^{-10}$). The top human reaches a composite of $+0.65$ and the top LLM (`moonshotai/kimi-k2-0905`) reaches $+0.53$, with the human ahead by 0.12 z on the canonical two-component panel (Figure S3). Across the five tasks, standardized Cronbach's α was .57 for LLMs and .64 for humans. The difference was not statistically significant (10,000 bootstrap

Table S5: Correlations between the JRT-corrected AGC-Bench composite (Section 3.5) and indicators on the Artificial Analysis intelligence battery (AA snapshot 2026-04), on the 83 models we evaluated, intersected with AA coverage per indicator. n varies because not every model has AA coverage on every indicator.

AA indicator	n	ρ	r
AA Intelligence Index (composite)	74	+0.778	+0.681
GPQA-Diamond	75	+0.770	+0.663
Humanity's Last Exam	75	+0.638	+0.589
LiveCodeBench (AA)	57	+0.538	+0.450
SciCode	75	+0.766	+0.685
AA Coding Index (composite)	69	+0.783	+0.688
AA Agentic Index (composite)	67	+0.748	+0.671
MMLU-Pro	57	+0.629	+0.536
AA Math Index (composite)	53	+0.518	+0.414
MATH-500	33	+0.599	+0.581
MMMU-Pro	35	+0.861	+0.779
τ^2-Bench	70	+0.696	+0.599
IFBench	72	+0.677	+0.538
CritPT	68	+0.680	+0.436

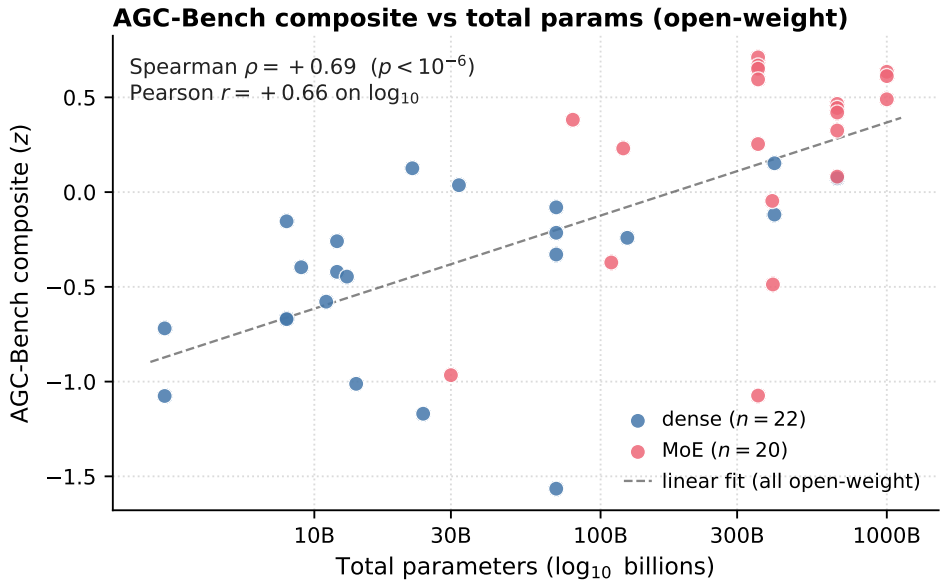


Figure S2: AGC-Bench composite vs total model parameters on the 42 open-weight models with documented parameter counts. Each point is one model, color-coded by architecture (dense vs. MoE). Proprietary frontier models are excluded because providers do not disclose parameter counts. Spearman $\rho = +0.69$ ($p < 10^{-6}$), Pearson $r = +0.66$ on $\log_{10}$ total parameters; the same relationship under active-parameter accounting weakens to $\rho = +0.37$.

resamples within each group; LLM minus human difference $= -.07$, 95% CI $[-.22, .12]$, $p = .529$). The earlier estimates (.64 for LLMs and .42 for humans) used a three-component score that included surprise and a larger sample of 125 LLMs and 224 humans. The current comparison uses the primary novelty and diversity scores and the same 201 humans and 80 LLMs as the leaderboard.

Independent human ratings. We collected complementary human originality ratings of all 24,172 original AGC-Human responses from 416 participants recruited through Prolific using a planned-

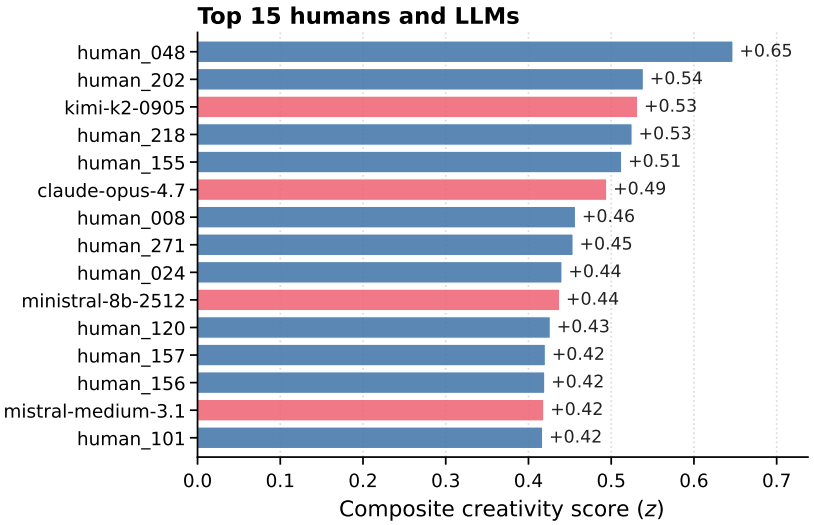


Figure S3: Top 15 humans and LLMs by composite creativity score (AGC-Human embedding composite, length-residualized z), drawn from 201 humans and 80 LLMs in the paired-task cohort. The top human (`human_048`) leads the top LLM (`moonshotai/kimi-k2-0905`) by 0.12 z (+0.65 vs. +0.53).

missing design. Participants were fluent in English, resided in the United States, United Kingdom, or Canada, and had a Prolific approval rate of at least 95%. Participants provided electronic informed consent and were offered $15 for completing the study. The study was deemed exempt from institutional review board review (institution and protocol number withheld for anonymous review). Raters used a 1–5 originality scale and were informed that the study contained human and AI responses; individual responses were not identified as human or AI responses. The dataset contains 89,983 judgments: 85,837 numeric ratings and 4,146 N/A judgments, excluded from ordinal estimation.

We estimated originality with a regularized cumulative-logit many-facet ordinal model, including response quality, global rater severity, rater-by-task deviations, and task-specific thresholds. The first three terms used zero-centered Gaussian penalties with standard deviations 1.0, .6, and .35, respectively. This human-rater model is separate from the LLM-judge GRM. Scores were averaged equally over units within prompts, prompts within tasks, and tasks within each respondent or model. Primary comparisons include 201 humans and 80 LLMs with complete five-task profiles. Judges were instructed to accept sensible, on-topic attempts even when responses were short or unoriginal. Empty responses, gibberish, refusals, repetitions of the prompt, and clearly off-topic responses were defined as invalid. Two models, Claude Opus 4.6 and Gemini 2.5 Flash, were retained with missing validity scores. Bootstrap intervals resample respondents and models separately, conditioning on fitted scores rather than propagating all measurement-model uncertainty. Five-task score reliability was .797 for humans and .754 for LLMs; individual raw-rating agreement was modest (macro ICC(1,1) = .292).

Across 80 LLMs, scores based on human ratings correlate with AGC-Human embedding scores at $\rho = .653$, 95% CI $[.501, .765]$, and with AGC-Bench scores at .646 $[.506, .751]$. For human respondents, the correlation with AGC-Human embedding scores is .466 $[.341, .576]$. After adjusting the original 1–5 ratings for rater severity and $\log(1 + \text{word count})$, LLMs scored .575 points higher than humans $[.524, .625]$, averaging equally across tasks. The difference was positive in all five tasks.

Mean cross-task correlations, averaged in Fisher-z space, are .422 for LLMs and .279 for humans using ordinal scores (difference .143, 95% CI $[.030, .259]$). A separate sensitivity analysis residualizes scores for log word count within source, task, and prompt; correlations become .262 and .241 (difference .021, 95% CI $[-.087, .142]$). Thus, greater LLM domain generality may partly reflect differences in model elaboration; the adjusted result provides no clear evidence of a group difference and does not establish equivalence.

We retain embedding scores as the primary AGC-Human measure because they apply the same procedure to human and LLM responses and remain scalable. Human ratings provide complementary evidence of convergent validity. Their convergence supports originality measurement but does not directly compare human and LLM preferences for different definitions of diversity, which remains future work.

Supplementary AGC-Judge diagnostics. As a separate diagnostic, we re-scored every cell with the open-weight AGC-Judge (Section 3.5). For each (entity, task, prompt) we sent the response to AGC-Judge with a task-specific creativity rubric (10–50 scale: "A creative response is novel and effective"). We ran two prompt variants on the same 7,647 cells:

- *Standard*: the canonical scoring template used at training time (rubric, prompt, response, request a single integer).
- *Fairness-aware*: a counter-bias prompt that discloses the self-preference bias in LLM judges documented in prior judge work (Panickssery et al., 2024; Ye et al., 2025), instructs the judge to focus on the underlying idea rather than stylistic register, and discloses the source (human study participant vs. AI language model) before scoring.

The standard prompt reproduces a strong LLM-favoring effect: LLMs averaged $33.4/50$ vs. humans $22.1/50$ (gap $+11.4$, roughly 1.4 SD). AGC-Judge's fine-tuning corpus contains exclusively LLM responses paired with frontier-LLM ratings ($48{,}299$ rows of JRT-corrected supervision spanning 24 canonical-metric cells; no human responses and no human ratings appear in training). This score gap alone does not establish bias: independent human raters also favor LLM responses. The fairness-aware prompt reduces this gap by $\approx 35\%$ to $+7.4$ (humans 24.1, LLMs 31.5) while remaining strongly correlated with scores obtained with the standard prompt ($r = +0.91$) and slightly improving convergent validity with the primary AGC-Human composite ($\rho = +0.51 \rightarrow +0.56$). This is consistent with prior findings that prompt-based debiasing attenuates self-preference bias without eliminating it (Liu et al., 2025; Jang & Silavong, 2025). The reduced gap does not establish that the prompt removes source bias.

To test whether the score gap acts on stylistic register or on content, we ran a counterfactual style-flip experiment on a sample of 50 cells (5 humans + 5 LLMs per task at the anchor prompt). For each cell we generated a paraphrase that preserved every idea verbatim while shifting register — humans rewritten in polished LLM-style, LLMs rewritten in casual human-style — via Grok-4.1-fast (following the adversarial paraphrase paradigm of Krishna et al., 2023). We then re-scored the transformed cell with the standard prompt. Mean score deltas were near zero in both directions ($\Delta_{\text{polish}} = +0.40$, $d = 0.29$; $\Delta_{\text{humanize}} = +0.52$, $d = 0.15$ on the 10–50 scale), indicating that holding the underlying idea constant and swapping stylistic register barely shifts AGC-Judge. This small probe cannot rule out style or source bias. The AGC-Judge fairness-aware composite correlates with the primary AGC-Human composite at $\rho = .56$ (80 LLMs: $\rho = .79$, compared with .80 under the standard prompt; 201 humans: $\rho = .56$), showing agreement within groups but different rankings across groups. These diagnostics do not establish unbiased AGC-Judge performance on human text; its direct human-rating validation remains limited to the external story dataset (Appendix H.7).

M AGC-HUMAN QUALITATIVE SAMPLES

To convey the flavor of top-tier creative responses, we reproduce verbatim responses to the Alternate Uses Task (BRICK) from two high-ranked humans (`human_048`, `human_218`) and the two top-ranked LLMs (`moonshotai/kimi-k2-0905`, `anthropic/claude-opus-4.7`) under the AGC-Human embedding composite. Responses are unedited. Selection follows the original embedding scores.

Alternate Uses Task — prompt: BRICK. **`human_048`**: Rustic paperweight; Boho door stop; Visual for class about updating your devices in a timely fashion; Visual when explaining to one's parents why they cannot activate their HTC hero.

`human_218`: Clean a pair of bricks and wrap them in fabric, leather, or decorative paper to keep heavy hardcovers upright on a shelf. Use a brick to hand-tamp soil or gravel for a small paver project

or a fence post hole. A single brick makes a heat-proof, rustic base for thick pillar candles or incense burners.

moonshotai/kimi-k2-0905: Pocket-sized "dumb-phone" for digital detox vacations—just a brick you talk to and then feel better about real life; official "interruption veto" gavel for Zoom meetings—slam it on the desk to mute everyone instantly; DIY home-planet kit: wrap it in aluminum foil, suspend it from the ceiling, and watch your very own asteroid orbit the living room; instant gravity-powered alarm clock—place it on a balanced shelf above your bed; when the snooze ends the shelf tilts; reusable dinner plate that guarantees no one asks for leftovers; pet rock's aggressive cousin—set it in the garden to intimidate actual rocks into growing flowers faster.

anthropic/claude-opus-4.7: Hollow it out and use it as a rustic planter for succulents; freeze it and place it in a cooler as a long-lasting ice block substitute; wrap it in festive paper as a gag gift inside a larger package to confuse recipients; use it as a pitching target in the backyard for baseball practice; grind it into powder to make natural red pigment for paint or chalk; strap one to each foot as makeshift resistance training for leg day.

Two stylistic registers are visible: human responses tend toward casual, idiomatic phrasing ("HTC hero"), while LLM responses tend toward literary or technical elaboration ("natural red pigment for paint or chalk"). Both registers are legible as upper-tail creative, but on different stylistic axes. These selected examples do not establish a group-level production strategy (Appendix L). Verbatim responses for the remaining four tasks (SCTT, Design Problems, Short Story, Metaphor Completion) are released in the artifact bundle (`agc_human_qualitative_samples.csv`).

N CHECKS FOR DATA CONTAMINATION

We examined benchmark-text exposure using recognition, continuation, and public-web searches. Proprietary training corpora prevent a complete exposure estimate or a confirmed uncontaminated subset.

Recognition. We adapted the Data Contamination Quiz (Golchin & Surdeanu, 2025) for eight models using 92 passages from nine benchmark sources, six newly authored controls, and five familiar public-domain controls. We first measured preferences for answer position using four versions of each passage with small wording changes and a `none` option; the original passage was absent. We then replaced a variant with the original in positions the model selected infrequently, testing all four positions if none met this criterion. There were 824 valid position-preference responses and $2{,}884$ recognition responses. At each model's highest benchmark-recognition position, original selection rates were 52.0% for benchmarks, 62.5% for authored controls, and 90.0% for familiar controls. The benchmark-minus-authored difference was -10.5 percentage points (descriptive two-way bootstrap interval $[-37.1, 17.5]$). Across all positions the rates were 37.3%, 44.6%, and 77.1%. Preferences for newly authored wording show why recognition rates cannot be interpreted as contamination percentages; the small control sets limit precision.

Continuation and web retrieval. The continuation tests covered 112 benchmark targets from ten tasks plus controls under neutral and benchmark-cued prompts. The initial cued MERMAID prompt inadvertently exposed target text; those trials were replaced with corrected prompts. Excluding 21 failed trials, exact reproduction occurred in $0/1{,}771$ benchmark trials, $0/160$ authored controls, and $38/80$ familiar controls. Manual review found one non-exact BrainTeaser near-match. Archived exact-phrase searches of the Bing index found no verified exact match among 92 benchmark passages; two candidate matches were rejected as generic word overlap.

These checks provide no clear evidence of benchmark-specific recognition beyond newly authored controls or exact reproduction of withheld benchmark text. Their limited sensitivity means they cannot rule out paraphrased exposure, memorization elsewhere in the benchmark set, or familiarity with task formats. The recent-source sensitivity in Appendix J likewise does not establish absence from training data.