On the Effects of Reasoning Effort and Prompt-Based Diversification on Scientific Ideation Diversity

Yu Chinen [1] **Haruka Ozaki** [1]

Abstract

Frontier large language models (LLMs) performing extended chain-of-thought reasoning have advanced closed-ended task performance, motivating interest in AI Scientist systems that automate stages of the research pipeline. In such systems, scientific ideation matters as the most upstream stage, where the diversity of generated research ideas bounds the downstream search space. While reasoning effort improves closed-ended task accuracy, its effect on open-ended scientific ideation diversity has not been systematically measured. We generate over 300,000 ideas across three frontier LLMs (Claude Sonnet 4.6, GPT-5.4, Gemini 3.1 Pro), three reasoning-effort levels (`low`, `medium`, `high`), and the LiveIdeaBench keyword set. We evaluate diversity with lexical metrics, embedding-based metrics across three embedders, and a pairwise LLM-as-a-Judge rubric across two judge models, yielding over 1,500,000 pairwise judgments. For comparison, we additionally evaluate two prompt-based diversification methods—Verbalized Sampling and String Seed of Thought—at the `low` and `high` reasoning-effort levels. Across these analyses, three main findings emerge. (1) Increasing reasoning effort raises within-keyword embedding pair distance by 13–36% from `low` to `high`, with no detectable shift in LLM-judged originality, feasibility, and clarity ratings. (2) Verbalized Sampling at `low` effort matches or exceeds default-prompt `high`-effort embedding diversity on quartile-defined keyword subsets, using 80–100% fewer reasoning tokens per idea, with no detectable decline in LLM-judged quality ratings. (3) In embedding space, idea distributions produced by varying reasoning effort and by varying prompt are nearest-neighbor distinguishable across all model–embedder–keyword-subset combinations. These findings are consistent across embedders and judges, providing a large-scale empirical map of how reasoning effort shifts open-ended scientific-ideation diversity, and surface concrete directions for future work—most centrally, downstream evaluation of whether the observed shifts correspond to scientifically meaningful differences—toward the design of AI Scientist systems.

[1]Laboratory for AI Biology, RIKEN Center for Biosystems Dynamics, Japan. Correspondence to: Yu Chinen <yu.chinen@riken.jp>, Haruka Ozaki <haruka.ozaki@riken.jp>.

Accepted at ICML 2026 AI for Science Workshop.

1. Introduction

Frontier LLMs trained with extended chain-of-thought reasoning have demonstrated notable success in enhancing reasoning performance on mathematics and programming. Inference-scaling-law analyses now formalize how added test-time compute relates to accuracy on verifiable tasks (Wu et al., 2025), and reinforcement-learning-trained reasoning models achieve superior performance on mathematics and coding through extended chain-of-thought (Guo et al., 2025). In parallel, AI Scientist systems that automate stages of the scientific research pipeline have moved from speculation to deployable artifacts, including end-to-end research-paper generation that has passed first-round workshop peer review (Lu et al., 2026) and multi-agent hypothesis discovery in biomedical domains (Gottweis et al., 2025).

To build AI Scientist systems, scientific ideation matters as the most upstream stage of the research pipeline, where the diversity of generated research ideas bounds the downstream search space. Benchmarks for ideation are now beginning to probe this stage explicitly with single-keyword prompts and divergent-thinking rubrics (Ruan et al., 2026), and large-scale human studies report that LLM-generated ideas can be rated as *more novel* than expert-written ideas while LLMs are observed to lack diversity when ideation is scaled up (Si et al., 2024). This pattern is not unique to scientific ideation: more broadly, in open-ended generation, LLMs exhibit a pronounced "Artificial Hivemind" of intra-model repetition and inter-model homogeneity across 26,000 open-ended queries (Jiang et al., 2025).

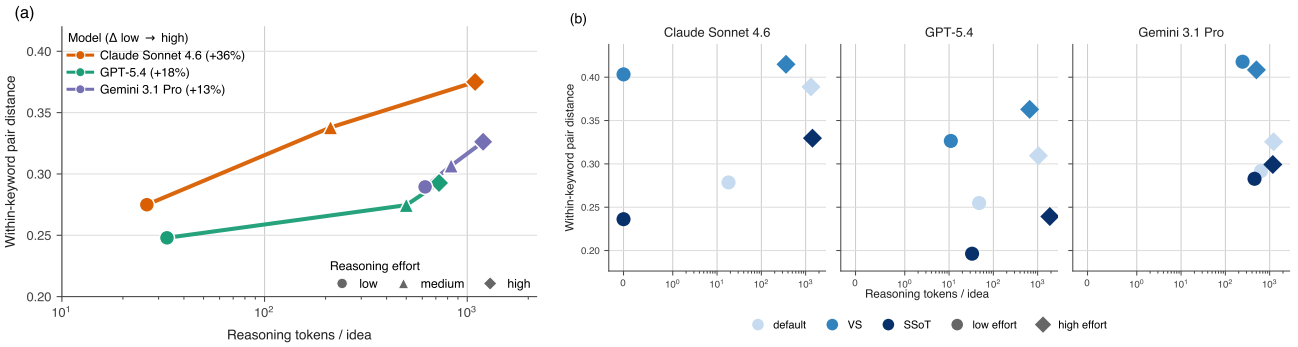

Figure 1. Two test-time axes for scientific-ideation diversity: reasoning effort (a) and prompt-based methods (b). *(a) Reasoning-effort axis* (default structured-facet prompt of Section 3.1; full LiveIdeaBench, 1,160–1,180 keywords per model). Within-keyword embedding pair distance rises by 13–36% from `low` to `high` effort in all three models. *(b) Reasoning-effort × prompt-based axis* (99–100 keywords per model: 50 keywords each drawn uniformly at random from the bottom and top quartiles of the within-keyword pair-distance distribution under default-prompt `low`-effort generation; symlog x-axis so that zero-reasoning-token conditions sit alongside the log-scale ticks). On these quartile-defined keyword subsets, Verbalized Sampling at `low` effort matches or exceeds default-prompt `high`-effort embedding pair distance, using 80–100% fewer reasoning tokens per idea on all three models. Both panels use text-embedding-3-large.

These observations motivate a question that has not been systematically evaluated: reasoning effort has been shown to improve accuracy on closed-ended tasks, but whether it shifts the diversity of open-ended scientific ideation, and how it interacts with existing diversification methods, remains open. Existing diversification methods divide broadly into train-time and test-time approaches; we survey both in Section 2. Train-time methods (e.g., diversity-aware post-training) require model-weight modification and are unavailable on the proprietary, API-only frontier models that expose the strongest reasoning capabilities, so we focus on test-time methods. Within test-time methods, decoding parameters of major reasoning models cannot be modified via their provider APIs, with `temperature` and `top_p` fixed or restricted to narrow ranges. Prompt-based methods therefore remain the principal available test-time lever for these models, and are in practice the first methods developers reach for.

In this work, we focus on how reasoning effort shifts the diversity of generated scientific ideas. We use LiveIdeaBench (Ruan et al., 2026) as the experimental substrate because its single-keyword prompts impose minimal task context, minimizing confounds from prompt scaffolding when isolating the reasoning-effort signal. We repeatedly sample 30 ideas per keyword across three reasoning-effort levels (`low`, `medium`, and `high`) for three frontier LLMs—Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro—totaling over 300,000 generations across 1,160–1,180 scientific keywords. We additionally evaluate two prompt-based methods—Verbalized Sampling (Zhang et al., 2025a) as a representative example and the more recent String Seed of Thought (Misaki & Akiba, 2026)—at the `low` and `high` reasoning-effort levels on the bottom and top quartiles of each model's diversity distribution. We evaluate diversity along three complementary dimensions—lexical, embedding-based, and a four-level LLM-as-a-Judge pairwise rubric—and additionally collect per-idea quality ratings. Our contributions are:

- We provide a large-scale measurement of how reasoning effort shifts scientific-ideation diversity, with cross-judge, cross-embedder, and cross-metric agreement across over 300,000 generations and 1,500,000 pairwise LLM-as-a-Judge judgments supporting the robustness of the observed effects.
- Reasoning effort: increasing reasoning effort raises within-keyword embedding pair distance by 13–36% from `low` to `high`, without a detectable decline in LLM-judged originality, feasibility, and clarity ratings.
- Reasoning-token efficiency: Verbalized Sampling at `low` effort matches or exceeds `high`-effort default-prompt embedding diversity on quartile-defined keyword subsets, using 80–100% fewer reasoning tokens per idea across the three models.
- Geometric distinguishability: the embedding distributions of reasoning-effort and prompt-based generations are nearest-neighbor distinguishable across all 18 combinations of 3 idea-generation models, 3 embedders, and 2 quartile-defined keyword subsets.
- Open questions: the present empirical map identifies and structures three follow-up directions: downstream evaluation of the observed diversity shifts, generalization across other reasoning-capable models and ideation tasks, and the mechanism by which test-time compute scaling transfers from closed-ended accuracy to open-ended ideation diversity.

2. Related Work

Scientific ideation benchmarks and systems. A growing line of work studies whether LLMs can produce expert-level research ideas, with benchmarks differing primarily in the amount of context provided and the breadth of scientific domains covered. IdeaBench (Guo et al., 2024) grounds LLMs in a target paper's title, abstract, and referenced works, and ranks generated ideas via an LLM judge with user-specified quality indicators. AI Idea Bench 2025 (Qiu et al., 2025) restricts ideation to AI research, with a dataset of 3,495 AI papers and their inspired works. LiveIdeaBench (Ruan et al., 2026), in contrast, uses single-keyword prompts to impose minimal task context and evaluates over 40 LLMs across 1,180 keywords spanning 22 scientific domains using an LLM-as-a-Judge rubric for divergent thinking. Scideator (Radensky et al., 2026) introduces a {`purpose, mechanism, evaluation`} facet schema for research-paper recombination. Si et al. (2024) run a human study with 100+ NLP researchers and find LLM-generated ideas are rated as more novel than expert-written ideas while flagging idea-diversity limitations. A follow-up by Si et al. (2025) reveals an ideation-execution gap: LLM-idea review scores decline after expert researchers execute the ideas, despite their initial ideation-stage advantage. Our work uses LiveIdeaBench's keyword set—chosen for its minimal-context, broad-domain design—together with Scideator's facet schema as the experimental substrate, and investigates a question not yet posed in this literature: how does reasoning effort affect the diversity of generated scientific ideas?

Open-ended LLM diversity and mode collapse. Recent work has shown pronounced mode collapse in open-ended generation: Jiang et al. (2025), evaluating 70+ LLMs with 50 samples per query across 26,000 open-ended queries, identify an "Artificial Hivemind" effect characterized by intra-model repetition and inter-model homogeneity, and NoveltyBench (Zhang et al., 2025c) reports a related pattern across 20 leading models, showing that within the Llama 3 (1B–405B) and Gemma 2 (2B–27B) families, larger variants often exhibit less diversity than smaller ones. The diversity-measurement landscape spans form-based metrics such as Self-BLEU (Zhu et al., 2018) and Distinct-N (Li et al., 2016); embedding-based metrics ranging from average pairwise similarity (used in the Hivemind study above) to kernel-based effective-number scores such as the Vendi Score (Friedman & Dieng, 2023); and out-of-set novelty signals computed as a harmonic mean of training-data n-gram unseenness and quality (Padmakumar et al., 2025). Recent meta-evaluation shows, however, that form-based metrics can assign near-maximum scores even to randomly-shuffled or syntactically incoherent outputs (Zhang et al., 2025b). Our work measures within-set diversity in this same open-ended-generation regime but combines lexical, embedding-based (mean pairwise cosine distance and Vendi Score), and LLM-as-a-Judge evaluations rather than relying on a single metric, and asks how reasoning effort shifts these signals.

Train-time methods for improving diversity. DARLING (Li et al., 2025) jointly optimizes for quality and a learned partition-function diversity reward in RL-style post-training, and Chung et al. (2025) add a *deviation-from-the-set* term to Direct Preference Optimization (DPO) and Odds-Ratio Preference Optimization (ORPO), reaching human-dataset-level diversity at 8B scale.

Test-time methods for improving diversity. Test-time scaling literature has studied both sampling-based scaling on verifiable tasks, where coverage rises log-linearly with sample count (Brown et al., 2024), and reasoning-based scaling via extended chain-of-thought (Guo et al., 2025; Wu et al., 2025), predominantly for accuracy rather than diversity; recent work also identifies limits of the reasoning-based direction, with longer chain-of-thought sometimes impairing reasoning when over-extended (Yang et al., 2025) and RL-trained reasoning models attaining higher pass@1 but lower pass@k at large k than their base models (Yue et al., 2025). For prompt-based diversification, Verbalized Sampling (Zhang et al., 2025a) prompts the model to verbalize a probability distribution over candidate responses and raises creative-writing diversity by 1.6–2.1×, and String Seed of Thought (Misaki & Akiba, 2026) instructs the model to first emit a random seed string and derive its response from that seed, raising response diversity on NoveltyBench. Meincke et al. (2024) earlier showed that chain-of-thought prompting raises idea-pool diversity on GPT-4, but elicited chain-of-thought via in-context instruction in a non-reasoning model—a distinct mechanism from the trained reasoning behavior of modern reasoning models that emit explicit reasoning tokens. This work jointly maps reasoning effort and prompt-based methods for open-ended scientific ideation, asking how the two compare and combine on a design space not yet examined in this literature.

3. Methods

3.1. Idea generation

We use the LiveIdeaBench keyword set (Ruan et al., 2026), which provides 1,180 scientific keywords spanning 22 scientific domains; we denote this set K_{full}. For the idea-generation prompt, we adapt LiveIdeaBench's single-keyword prompt template by inserting the {`purpose, mechanism, evaluation`} facet schema introduced by Scideator (Radensky et al., 2026), so each generated idea is structured as three short facet fields rather than free-form

prose; the default prompt is shown in Figure 5.

We parameterize each generation run by the 4-tuple $(model, K, prompt, effort)$, where K is a set of keywords. For each $k \in K$ we generate 30 ideas via repeated sampling; we call this collection an *idea set*.

We compare three frontier LLMs—Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro—and vary each provider's reasoning-effort API parameter across three levels (`low`, `medium`, `high`) while leaving all other sampling parameters (e.g., temperature, top-p) at the vendor default. In this section we use $K = K_{\text{full}}$ for GPT-5.4 and Gemini 3.1 Pro; Claude Sonnet 4.6 omits 20 keywords to safety refusals, so $|K| = 1{,}160$ for that model. Per-vendor model identifiers, the per-vendor reasoning-token definition used throughout this paper, and other configuration details are listed in Section A.1. The full corpus comprises approximately 300,000 generations (3 models $\times$ 1,160–1,180 keywords $\times$ 3 reasoning-effort levels $\times$ 30 samples).

3.2. Lexical diversity metrics

For each idea set, we measure lexical diversity over the 30 ideas using two within-set metrics: Self-BLEU-4 (Zhu et al., 2018) (within-set pairwise BLEU using NLTK `sentence_bleu` with smoothing method 1), and Distinct-N for $N \in \{1, 2\}$ (Li et al., 2016) (unique n-grams divided by total n-grams in the set). Lower Self-BLEU and higher Distinct-N both indicate greater lexical spread.

3.3. Embedding-based diversity metrics

Each generated idea is embedded with three independent models—OpenAI text-embedding-3-large (primary), Amazon titan-embed-text-v2:0 (Titan v2 hereafter), and SPECTER2 (Singh et al., 2023) with the `adhoc_query` adapter (configuration in Section A.4)—producing per idea both a combined-text embedding (concatenation of the `purpose`, `mechanism`, and `evaluation` facet strings) and one embedding per individual facet. On the combined-text embeddings we report two diversity metrics. The primary metric is mean pairwise cosine distance over the $\binom{30}{2}$ within-keyword pairs, adopted for interpretability and direct comparability with prior open-ended diversity work (Jiang et al., 2025). As a complementary metric we report the Vendi Score (Friedman & Dieng, 2023), $\exp(-\sum_i \lambda_i \log \lambda_i)$ for eigenvalues λ_i of K/n where K is the $n \times n$ cosine similarity matrix and $n = 30$ is the size of the idea set; this yields the effective number of unique ideas under the cosine kernel and serves as a validation metric, since the same trends should replicate under this non-pairwise-mean aggregation (Section B.4).

3.4. LLM-as-a-Judge pairwise protocol

To complement the lexical and embedding-based metrics with a content-level signal, we additionally use an LLM-as-a-Judge pairwise rubric. We adopt the LiveIdeaBench `fluency_critic_prompt`, which classifies each idea pair into one of four ordered labels ranging from A (completely different ideas) to D (academically identical). For each idea set we partition the $\binom{30}{2} = 435$ within-keyword pairs by combined-text embedding distance into top-10 / middle-10 / bottom-10 *distance bins*, and judge each selected pair in both AB and BA orders to mitigate position bias (Wang et al., 2024). This yields 60 judgments per idea set and approximately 774,000 judgments per judge model across the three idea-generation runs.

We use two judge models, GPT-4.1 and Claude Haiku 4.5, selected as high-performing non-reasoning judges feasible at our evaluation scale. The pairwise judge prompt is shown in Figure 8; other configuration details are in Section A.5.

3.5. Per-idea quality critic

Since increasing diversity could in principle reduce per-idea quality, we additionally collect per-idea quality ratings using the LiveIdeaBench `critic_prompt` along three axes—originality, feasibility, and clarity—each on a 1–10 scale and applied independently to each generated idea. We use GPT-4.1 and Claude Haiku 4.5 as judges, and additionally use Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro as judges on a smaller keyword subset (Section 3.6). The quality judge prompt is shown in Figure 9; other configuration details are in Section A.5. Per-model coverage on the quality rubric is reported in Section 4.3.

3.6. Prompt-based methods

We complement the reasoning-effort axis of Section 3.1 with two prompt-based methods. Verbalized Sampling (`VS`) (Zhang et al., 2025a) requests $k = 3$ ideas per call instead of one. String Seed of Thought (`SSoT`) (Misaki & Akiba, 2026) conditions each call on a distinct random string drawn at the start of the call. We compare both prompts against the `default` structured-facet prompt of Section 3.1 under the same three idea-generation models and the `low` and `high` reasoning-effort levels. We refer to each $(model, prompt, effort)$ tuple as an experimental *condition*, e.g., $(\texttt{default}, \texttt{low})$ or $(\texttt{VS}, \texttt{low})$ with the model implicit when context fixes it; per-keyword diversity metrics are aggregated across $k \in K$ to obtain a per-condition value. The `VS` and `SSoT` prompts are shown in Section A.2.

For each model we compute the within-keyword pair-distance distribution at $(\texttt{default}, \texttt{low})$ over K_{full} and define two quartile pools: the bottom 25% (lowest pair distances) and the top 25% (highest). From each pool—290

keywords for Claude Sonnet 4.6, 295 for GPT-5.4 and Gemini 3.1 Pro—we draw 50 keywords uniformly at random without replacement to form Q_1 and Q_4 respectively. This sub-sampling avoids over-weighting the absolute extreme tails of K_{full} that would result from selecting the lowest or highest 50 keywords directly. Q_1 thus targets keywords with low (`default`, `low`) baseline pair distance, where prompt-based and reasoning-effort methods could plausibly raise diversity; Q_4 targets keywords near the (`default`, `low`) diversity ceiling. GPT-5.4 has $|Q_1| = 49$ instead of 50 because the keyword *bioterrorism* elicits a safety refusal under `SSoT` and is excluded from that model's Q_1; Claude Sonnet 4.6 and Gemini 3.1 Pro retain $|Q_1| = 50$. Per-model Q_1 and Q_4 keyword lists are given in Section A.3.

We measure separation between two conditions under the same model by leave-one-out 1-nearest-neighbor two-sample classification accuracy (Schilling, 1986; Lopez-Paz & Oquab, 2018). For each keyword k we pool the 30 ideas from each of the two idea sets into a union of 60 vectors, label each vector by its source condition, and classify each vector by the label of its nearest neighbor in the union (cosine distance, self excluded). The reported metric is the fraction correctly classified, averaged across the 50 keywords in K (49 for GPT-5.4 Q_1). A value of 0.5 indicates that the nearest-neighbor structure cannot distinguish the two conditions, and 1.0 indicates perfect separation. Equivalently, this is the $k = 1$ case of the nearest-neighbor two-sample statistic of Schilling (1986) and corresponds to a $k = 1$, leave-one-out instance of the classifier two-sample test framework of Lopez-Paz & Oquab (2018).

All Wilcoxon and Mann–Whitney p-values reported in Section 4 are Benjamini–Hochberg FDR-corrected within per-model-run families (Section A.7).

4. Results

4.1. Reasoning-effort axis

Lexical. All three models show monotonic decreases in Self-BLEU-4 (Zhu et al., 2018) and increases in Distinct-2 (Li et al., 2016) from `low` to `high` effort on combined-text inputs (Table 1). The absolute shift is largest on Claude Sonnet 4.6, which also has the highest `low`-effort Self-BLEU-4 of the three models. Per-facet breakdowns are in Section B.1.

Embedding pair distance. Combined-text within-keyword pair distance shifts positively from `low` to `high` effort across all three models (Table 2): +36% for Claude Sonnet 4.6, +18% for GPT-5.4, and +13% for Gemini 3.1 Pro. Within each model the shift correlates positively with median reasoning tokens per sample (Pearson $r = +0.47$ / $+0.32$ / $+0.21$ for the three models). All 9 (model × embedder) shifts are positive under text-embedding-3-large, Titan v2, and SPECTER2, ranging from +8% to +36% (Section B.2).

Table 1. Combined-text lexical diversity across the three reasoning-effort levels. Arrows mark the direction of greater lexical diversity.

Model	Metric	low	medium	high
Claude Sonnet 4.6	Self-BLEU-4 (↓)	0.411	0.238	0.155
	Distinct-2 (↑)	0.638	0.760	0.815
GPT-5.4	Self-BLEU-4 (↓)	0.253	0.195	0.178
	Distinct-2 (↑)	0.736	0.777	0.793
Gemini 3.1 Pro	Self-BLEU-4 (↓)	0.316	0.294	0.265
	Distinct-2 (↑)	0.658	0.674	0.698

Table 2. Per-facet (`purpose`, `mechanism`, `evaluation`) and combined-text within-keyword pair distance shift from `low` to `high` effort. Computed as the average over all 504,600–513,300 pooled pairs per condition using text-embedding-3-large.

Model	Facet	low	high	Δ rel.
Claude Sonnet 4.6	`purpose`	0.325	0.468	+44%
	`mechanism`	0.364	0.462	+27%
	`evaluation`	0.406	0.439	+ 8%
	combined-text	0.275	0.375	**+36%**
GPT-5.4	`purpose`	0.347	0.426	+23%
	`mechanism`	0.328	0.387	+18%
	`evaluation`	0.371	0.450	+21%
	combined-text	0.248	0.293	**+18%**
Gemini 3.1 Pro	`purpose`	0.374	0.419	+12%
	`mechanism`	0.380	0.406	+ 7%
	`evaluation`	0.366	0.387	+ 6%
	combined-text	0.289	0.326	**+13%**

Vendi Score validation. Vendi Score (Section 3.3) replicates the effort-axis trend on the same combined-text embeddings: per-keyword Pearson r between Vendi and pair distance is ≥ 0.94 across all 9 (model, embedder) combinations, and the `low` → `high` lift in mean Vendi is positive in every combination, ranging from +6% to +76% (Table 11).

Pairwise judge classification. Under the GPT-4.1 judge, the top-distance-bin A-rate (fraction of within-keyword pairs labeled "completely different ideas"; coverage 100%) rises monotonically with effort across all three models: +17 pp for Claude Sonnet 4.6, +14 pp for GPT-5.4, and +8 pp for Gemini 3.1 Pro. The Claude Haiku 4.5 judge preserves the same ranking at smaller absolute rates; full per-distance-bin distributions and the effort-axis trend under both judges are in Section D.1. Aggregating across all distance bins via the LiveIdeaBench fluency score (Ruan et al., 2026) reproduces the same monotonic effort-axis trend and Claude Sonnet 4.6 > Gemini 3.1 Pro > GPT-5.4 ordering under both judges: pairs are judged as more different on average

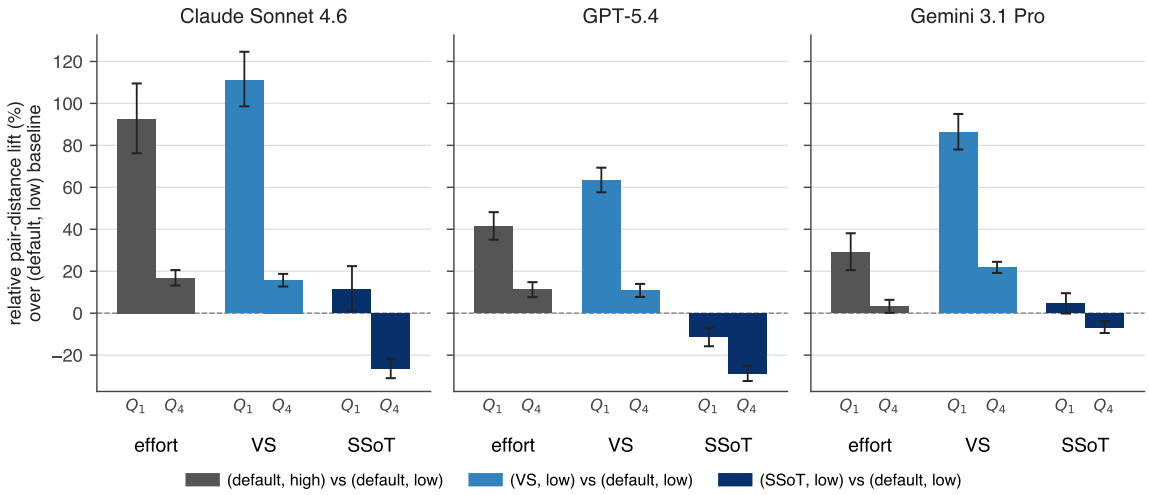

Figure 2. Per-K relative pair-distance lift of each (*prompt*, *effort*) condition over the (`default`, `low`) baseline, computed per keyword and averaged within $K = Q_1$ (left bars) and $K = Q_4$ (right bars). Error bars: bootstrap 95% CI on the per-K mean across keywords ($n_{\text{boot}} = 10{,}000$). At `low` effort, `default` → `VS` lifts Q_1 pair distance substantially; `default` → `SSoT` does not.

as effort rises (Table 17).

Cross-judge agreement. The two judges substantially agree on the pairwise rubric: quadratic-weighted Cohen's $\kappa = 0.875$ (Claude Sonnet 4.6), 0.838 (GPT-5.4), and 0.850 (Gemini 3.1 Pro) over 212,000 aligned pair judgments per generation model. Table 3 illustrates the rubric on keyword *reinforcement learning* (Claude Sonnet 4.6, `high` effort): the reference idea paired with three comparison ideas drawn from top/middle/bottom distance bins receives the A → B → D label gradient from both judges.

Facet localization. The `purpose` facet carries the largest `low` → `high` pair-distance shift in all three models (Table 2: Claude Sonnet 4.6 +44%, GPT-5.4 +23%, Gemini 3.1 Pro +12%); `mechanism` and `evaluation` shift less. The same `purpose`-heavy pattern holds under lexical metrics (Section B.1) and under the additional embedders (Section B.2), indicating the localization is not specific to the embedding space.

4.2. Prompt-based methods vs reasoning-effort scaling

We compare the `default` structured-facet prompt against the two prompt-based methods, `VS` and `SSoT`, under `low` and `high` reasoning effort, on the $K \in \{Q_1, Q_4\}$ keyword subsets defined in Section 3.6.

Reasoning-token efficiency. On the reasoning-token axis, (`VS`, `low`) achieves equal or higher pair distance than (`default`, `high`) in all three models (Figure 1, panel (b)) at substantially lower reasoning-token usage on both Q_1 and Q_4: Claude Sonnet 4.6's (`VS`, `low`) uses 100% fewer reasoning tokens per idea than (`default`, `high`) (no reasoning block), GPT-5.4 99% fewer, and Gemini 3.1 Pro 80% fewer. This arises because VS already attains (`default`, `high`)'s pair distance at `low` effort: within-prompt effort scaling (Q_1, `low` → `high`) adds only 3% under `VS` and at most 11% across models, while `default`'s effort scaling adds 29–93%.

Effect decomposition. Figure 2 reports per-K relative pair-distance lifts over the (`default`, `low`) baseline. At `low` effort the prompt-axis lift from `default` to `VS` is large on Q_1 (+111% / +63% / +86% for Claude Sonnet 4.6 / GPT-5.4 / Gemini 3.1 Pro, BH-FDR-adjusted Wilcoxon $p < 10^{-13}$ for all three) and shrinks at `high` effort (+17% / +32% / +44% in the same model order) because (`default`, `high`) already saturates the embedding pair distance that `VS` reaches at `low` effort. `SSoT` does not improve over `default` at either effort: its Q_1 pair distance is at or below (`default`, `low`) (Wilcoxon $p > 0.35$ in two of three models) and decreases relative to (`default`, `high`) at `high` effort. Effort scaling itself shows distinct shapes per prompt: large lifts under `default`, saturation under `VS`, and a depressed-baseline lift under `SSoT`. We attribute the depressed `SSoT` baseline to a mismatch between the seed-conditioning mechanism of Misaki & Akiba (2026) and open-ended structured-facet generation; Appendix E analyses the emitted seeds and the original method's evaluation regime.

Vendi Score validation. Vendi Score reproduces the prompt-axis pattern: at `low` effort on Q_1, `default` → `VS` raises mean Vendi in all 9 (model, embedder) combinations (range +13% to +154%), in the same direction as

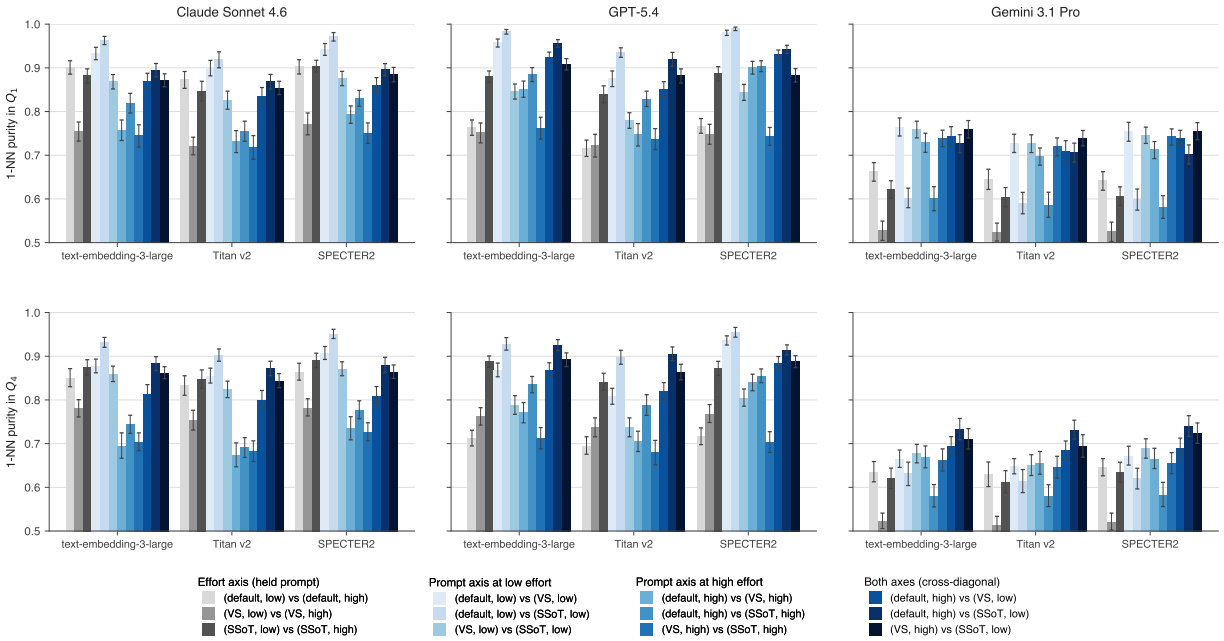

Figure 3. 1-NN two-sample classification accuracy (Section 3.6) for 12 comparisons of $(\texttt{prompt}, \texttt{effort})$ conditions $\times$ 3 embedders, on 6 panels (3 idea-generation models $\times$ $K \in \{Q_1, Q_4\}$). 0.5 means the two conditions are indistinguishable in embedding space; 1.0 means perfectly separable. Legend groups partition the 12 comparisons by which axes differ between the two conditions: effort only (held prompt), prompt only (held effort), or both (cross-diagonal). The headline comparison $(\texttt{default}, \texttt{high})$ vs. $(\texttt{VS}, \texttt{low})$ stays at or above 0.68 in every panel, and reaches 0.84 / 0.81 / 0.84 at the median across panels for text-embedding-3-large / Titan v2 / SPECTER2.

the corresponding pair-distance lifts; $\texttt{default} \rightarrow \texttt{SSoT}$ shows minimal or negative shifts across most combinations, mirroring its pair-distance behaviour (Table 12).

1-NN two-sample classification. The headline comparison $(\texttt{default}, \texttt{high})$ vs. $(\texttt{VS}, \texttt{low})$ has 1-NN accuracy 0.84 / 0.81 / 0.84 at the median across the six $(\textit{model}, K)$ combinations under text-embedding-3-large / Titan v2 / SPECTER2; all 18 $(\textit{model}, \textit{embedder}, K)$ combinations are above 0.68 (Figure 3). The same ranking holds under LLM-judge pairwise classification (top distance bin, $K = Q_1$, GPT-4.1): $(\texttt{VS}, \texttt{low})$ outputs are classified as "completely different ideas" 95.9% / 88.7% / 96.5% of the time on the three models, exceeding $(\texttt{default}, \texttt{high})$ by 18.9 / 25.5 / 25.4 percentage points.

4.3. Per-idea quality

Coverage was $\geq 99.96\%$ under both judges (156 skips from repetition-loop degeneration, $\geq 90\%$ per idea set), so every $(\textit{model}, \textit{prompt}, \textit{effort})$ condition is comparable. Figure 4 plots per-idea quality (mean of originality, feasibility, and clarity on a 1–10 scale) per condition under both judges: curves are nearly flat—within the resolution of this LLM-judge quality rubric—across $\texttt{low} \rightarrow \texttt{medium} \rightarrow \texttt{high}$ effort. Across the 18 (model, judge model, quality axis) combinations, the mean absolute $\texttt{low} \rightarrow \texttt{high}$ change is 0.13 pt (max 0.50 pt); the same flatness holds under prompt-based variation across $\{\texttt{default}, \texttt{VS}, \texttt{SSoT}\}$ on $K \in \{Q_1, Q_4\}$, with mean absolute prompt-axis change 0.21 pt (max 0.75 pt). Both judges reproduce the same condition-mean flatness despite an absolute-level offset (Section D.2), so the flatness is not a judge-specific artifact. Condition-mean flatness can in principle mask a per-keyword trade-off in which individual keywords gain diversity by sacrificing LLM-judged quality. We therefore compare per-keyword Δ(diversity) against Δ(quality); the resulting Pearson correlations are small across all 9 (model $\times$ quality axis) combinations ($|r| \leq 0.18$, $R^2 \leq 3.5\%$; Section B.6), giving no evidence for a systematic linear diversity–quality trade-off under this judge rubric. The same flatness holds when the three idea-generation models are used as judges (Section D.2).

5. Conclusion

Our work jointly measures reasoning-effort scaling and prompt-based methods for open-ended scientific ideation. Reasoning-effort scaling lifts within-keyword pair distance by 13–36%. Verbalized Sampling at `low` effort matches or exceeds `high`-effort default-prompt embedding diversity using 80–100% fewer reasoning tokens per idea, with no detectable decline in LLM-judged originality, feasibility, and clarity. The two axes' outputs are nearest-neighbor dis-

Table 3. Reference-grounded grid for the keyword *reinforcement learning*, `high` effort, generated by Claude Sonnet 4.6. The reference (Idea 1) is paired with three comparison ideas drawn from the top, middle, and bottom distance bins. Both judges (GPT-4.1, Claude Haiku 4.5) assign the labels shown, and they agree in both AB and BA orderings (12/12). Bold spans are shared verbatim between the reference and Idea 4: the mode-collapse signature that the bottom distance bin surfaces.

Role	d	Labels	Idea
Reference (Idea 1)	—	—	Sparse-reward RL exploration via **persistent homology** on the **visited-state manifold**: **Betti numbers** identify **topological holes** marking **unexplored regions**, which generate **intrinsic exploration rewards**. Benchmark against curiosity-driven and RND baselines on sparse-reward maze and robotic-navigation tasks.
Top (Idea 2)	0.544	A / A	Continual RL with a dual-memory architecture: a learned world model replays high-TD-error transitions while a differentiable symbolic rule extractor distils policy knowledge into logical constraints that regularise future updates. Benchmark on MiniGrid and Procgen continual transfer (forward / backward transfer, sample efficiency) versus EWC, PackNet, and experience replay.
Middle (Idea 3)	0.380	B / B	Sparse-reward RL exploration with intrinsic rewards derived from causal-discovery (PC algorithm) on offline trajectories: causally relevant state-variable transitions become shaped intrinsic-reward subgoals. Benchmark against HER and RND on MiniGrid and Montezuma's Revenge.
Bottom (Idea 4)	0.103	D / D	Sparse-reward RL exploration with geometry-aware intrinsic motivation: **persistent homology** (TDA) on the **visited-state manifold**; incremental **Betti numbers** flag **topological holes** as **unexplored regions**; assign **intrinsic** rewards for transitions that reduce topological complexity. Benchmark on Montezuma's Revenge and MiniGrid against count-based and curiosity-driven baselines.

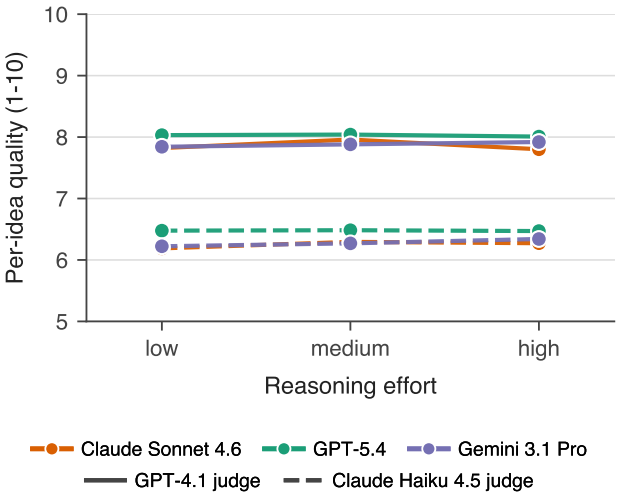

Figure 4. Per-idea quality (mean of originality, feasibility, and clarity on the 1–10 scale) stays flat across reasoning effort under both judges. The y-axis is held to the upper half of the scale so the two judges remain on the same axis despite Claude Haiku 4.5's systematically lower ratings.

tinguishable across all 18 combinations of 3 idea-generation models, 3 embedders, and $K \in \{Q_1, Q_4\}$. By mapping the joint (reasoning tokens, pair distance) plane with cross-validated metrics, this work provides an empirical starting point for downstream evaluation, compute-allocation studies, and mechanistic interpretation. We hope this catalyzes joint study of these axes for open-ended ideation, leaving downstream evaluation, generalization, and mechanism as open questions in Section 6.

6. Open Questions

Downstream evaluation. Our diversity metrics detect statistically significant differences across conditions, and qualitative spot-checks of additional samples beyond Table 3 confirm these differences correspond to qualitatively distinct ideas. Whether the differences carry scientific meaning has not been established at scale; large-scale expert evaluation remains open.

Generalization. Whether our findings generalize across (i) other reasoning-capable models, (ii) other generation methods (e.g., prompt-based methods beyond `default`/VS/SSoT, facet structures beyond Scideator), (iii) other quality dimensions beyond LiveIdeaBench's axes (e.g., factuality, novelty against prior literature, executability), and (iv) other ideation tasks (e.g., benchmarks beyond LiveIdeaBench, more concrete, practical research ideation) is open.

Mechanism. Whether the test-time-compute scaling established for closed-ended tasks (Section 2) extends to open-ended ideation diversity is unknown. Within each model, per-keyword pair distance correlates positively with realized reasoning tokens (Pearson $r = +0.47$ (Claude Sonnet 4.6), $+0.32$ (GPT-5.4), $+0.21$ (Gemini 3.1 Pro); Section 4.1), but the three models' patterns on the (reasoning tokens, pair distance) plane do not coincide.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This work is supported by JST CREST (JPMJCR2551) and the RIKEN TRIP Advanced General Intelligence for Science Program (AGIS).

References

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, 2024. URL http://arxiv.org/abs/2407.21787.

Chung, J. J. Y., Padmakumar, V., Roemmele, M., Sun, Y., and Kreminski, M. Modifying Large Language Model Post-Training for Diverse Creative Writing, 2025. URL http://arxiv.org/abs/2503.17126.

Friedman, D. and Dieng, A. B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning, July 2023.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. Towards an AI co-scientist, 2025. URL http://arxiv.org/abs/2502.18864.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z.

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E., Bekiranov, S., and Zhang, A. IdeaBench: Benchmarking Large Language Models for Research Idea Generation, October 2024.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., and Choi, Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond), 2025. URL http://arxiv.org/abs/2510.22954.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models, 2016. URL http://arxiv.org/abs/1510.03055.

Li, T., Zhang, Y., Yu, P., Saha, S., Khashabi, D., Weston, J., Lanchantin, J., and Wang, T. Jointly Reinforcing Diversity and Quality in Language Model Generations, 2025. URL http://arxiv.org/abs/2509.02534.

Lopez-Paz, D. and Oquab, M. Revisiting Classifier Two-Sample Tests, March 2018.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., and Clune, J. Towards end-to-end automation of AI research. 651(8107):914–919, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10265-5. URL https://www.nature.com/articles/s41586-026-10265-5.

Meincke, L., Mollick, E. R., and Terwiesch, C. Prompting Diverse Ideas: Increasing AI Idea Variance, 2024. URL http://arxiv.org/abs/2402.01727.

Misaki, K. and Akiba, T. String Seed of Thought: Prompting LLMs for Distribution-Faithful and Diverse Generation, February 2026.

Padmakumar, V., Chen, Y.-H., Pan, J., Chen, V., and He, H. Measuring LLM Novelty As The Frontier Of Original And High-Quality Output, 2025. URL http://arxiv.org/abs/2504.09389.

Qiu, Y., Zhang, H., Xu, Z., Li, M., Song, D., Wang, Z., and Zhang, K. AI Idea Bench 2025: AI Research Idea Generation Benchmark, May 2025.

Radensky, M., Shahid, S., Fok, R., Siangliulue, P., Hope, T., and Weld, D. S. Human-LLM Compound System for Scientific Ideation through Facet Recombination and Novelty Evaluation, 2026. URL http://arxiv.org/abs/2409.14634.

Ruan, K., Wang, X., Hong, J., Wang, P., Liu, Y., and Sun, H. Evaluating LLMs' Divergent Thinking Capabilities for Scientific Idea Generation with Minimal Context, 2026. URL http://arxiv.org/abs/2412.17596.

Schilling, M. F. Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986. ISSN 0162-1459. doi: 10.2307/2289012.

Si, C., Yang, D., and Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, 2024. URL http://arxiv.org/abs/2409.04109.

Si, C., Hashimoto, T., and Yang, D. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas, 2025. URL http://arxiv.org/abs/2506.20803.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations, 2023. URL http://arxiv.org/abs/2211.13308.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. Large Language Models are not Fair Evaluators. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models, 2025. URL http://arxiv.org/abs/2408.00724.

Yang, W., Ma, S., Lin, Y., and Wei, F. Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning, 2025. URL http://arxiv.org/abs/2502.18080.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., and Huang, G. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?, 2025. URL http://arxiv.org/abs/2504.13837.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity, 2025a. URL https://arxiv.org/abs/2510.01171v3.

Zhang, T., Peng, B., and Bollegala, D. Evaluating the Evaluation of Diversity in Commonsense Generation, 2025b. URL http://arxiv.org/abs/2506.00514.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. NoveltyBench: Evaluating Language Models for Humanlike Diversity, 2025c. URL http://arxiv.org/abs/2504.05228.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A Benchmarking Platform for Text Generation Models, 2018. URL http://arxiv.org/abs/1802.01886.

A. Methods supplement

This appendix collects per-component methodological detail extending Section 3.

A.1. Generation API and parameters

All three models are sampled with each vendor's default sampling parameters (temperature, top-p, etc.); only the reasoning-effort level is varied across runs. The per-vendor API endpoint, model identifier, reasoning-effort level values, and schema-constrained structured-output field used at generation time are listed below. Table 4 maps the corresponding reasoning-token field in each vendor's API response, which we use as the x-axis variable in Figure 1.

OpenAI Responses (GPT-5.4). Model identifier: `gpt-5.4-2026-03-05`. Reasoning-effort levels: `low`, `medium`, `high`. Structured output: `text.format = {type: json_schema, strict: true}`.

Bedrock Converse (Claude Sonnet 4.6). Model identifier: `jp.anthropic.claude-sonnet-4-6` (cross-region inference profile). Reasoning-effort levels: `low`, `medium`, `high`. Structured output: `outputConfig.textFormat = {type: json_schema, ...}`.

Vertex AI (Gemini 3.1 Pro). Model identifier: `gemini-3.1-pro-preview`. Reasoning-effort levels: `low`, `medium`, `high`. Structured output: `generationConfig.responseMimeType = "application/json"` with `generationConfig.responseSchema`.

Table 4. Per-vendor reasoning-token field used as the x-axis in Figure 1. OpenAI and Vertex AI report the reasoning-token count directly under the listed field. The Bedrock Converse `usage` block does not break out a reasoning subset, so for Claude Sonnet 4.6 we derive it post-hoc by subtracting the visible-answer token count from `outputTokens`; the visible-answer count is obtained by passing the response text to Anthropic's `count_tokens` endpoint, which uses the official Claude tokenizer and reports an estimate that may include constant system overhead. The estimate caveat does not affect the direction-of-claim in this paper: the possible overhead is small and constant at the call level, and VS token counts are consistently reported per idea by dividing the call-level count by k (Section A.2).

Vendor (model)	Reasoning-token field
OpenAI Responses (GPT-5.4)	`usage.output_tokens_details.reasoning_tokens`
Bedrock Converse (Claude Sonnet 4.6)	derived: `outputTokens − count_tokens(idea_text)`
Vertex AI (Gemini 3.1 Pro)	`usage_metadata.thoughts_token_count`

Retry policy. Each idea-generation call is retried up to five times on JSON schema validation failure or transient API error before being recorded as missing.

Claude Sonnet 4.6 safety refusals. Claude Sonnet 4.6's K_{full} has 1,160 keywords (versus 1,180 for GPT-5.4 and Gemini 3.1 Pro) because the following 20 LiveIdeaBench keywords elicit a safety refusal at the idea-generation step under the `default` prompt across all 30 sampled calls and are excluded from this model's K_{full}: *aerobiology*, *aerosol science*, *bacteriology*, *bioengineering*, *bioreactors*, *biosecurity*, *bioterrorism*, *fluidization*, *genetic engineering*, *genetically modified organisms*, *mutation*, *pandemic preparedness*, *protein engineering*, *protein synthesis*, *superbugs*, *vaccine development*, *vectorborne diseases*, *virology*, *viruses*, *zoonotic diseases*. The list is dominated by biosecurity-adjacent topics (pathogens, biocontainment, dual-use bioengineering) consistent with Claude's published safety policy. GPT-5.4 and Gemini 3.1 Pro return schema-valid responses on all 1,180 keywords under the `default` prompt; only `SSoT` on GPT-5.4 elicits one additional refusal (*bioterrorism*, dropping $|Q_1| = 50 \rightarrow 49$ for that model only; Section A.3).

A.2. Generation prompts

Figure 5 shows the `default` idea-generation prompt of Section 3.1; Figures 6 and 7 show the Verbalized Sampling (`VS`) and String Seed of Thought (`SSoT`) prompts of Section 3.6 that extend it. The placeholder {{`keywords`}} is replaced with the assigned keyword at call time.

Per-idea token counts. The `default` and `SSoT` prompts return one idea per call, so the call-level and per-idea token counts coincide. Verbalized Sampling (Section 3.6) requests $k = 3$ ideas per call; for `VS` we report per-idea token usage as the call-level token count divided by k.

I'll be submitting your next responses to a "Good Scientific Idea" expert review panel. If they consider your idea to be a good one, you'll receive a reward. Your assigned keyword is: "{{keywords}}". Generate a novel scientific research idea with the following facets, each in 2-3 sentences:
- purpose: the problem being addressed
- mechanism: the proposed solution
- evaluation: the method for determining whether the solution works
(Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).

Figure 5. Default idea-generation prompt (Section 3.1).

I'll be submitting your next responses to a "Good Scientific Idea" expert review panel. If they consider your idea to be a good one, you'll receive a reward. Your assigned keyword is: "{{keywords}}". Generate a novel scientific research idea with the following facets, each in 2-3 sentences:
- purpose: the problem being addressed
- mechanism: the proposed solution
- evaluation: the method for determining whether the solution works
(Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).

Generate 3 such ideas. For each idea, also include an estimated probability from 0.0 to 1.0 of this idea given the prompt (relative to the full distribution).

Figure 6. Verbalized Sampling (VS) idea-generation prompt (Zhang et al., 2025a).

Retry policy. Each `VS` or `SSoT` call is retried up to five times on JSON schema validation failure or transient API error before being recorded as missing.

A.3. Q_1 / Q_4 keyword selection

We list the Q_1 and Q_4 keyword sets used in the analysis of prompt-based methods (Section 3.6) in Table 5 (Claude Sonnet 4.6), Table 6 (GPT-5.4), and Table 7 (Gemini 3.1 Pro). For each idea-generation model, the within-keyword pair-distance distribution at (`default`, `low`) is computed across the LiveIdeaBench keywords described in Section 3.1; from the bottom 25% of this distribution we randomly sample 50 keywords to form Q_1, and similarly 50 keywords from the top 25% to form Q_4. GPT-5.4 Q_1 is reduced to 49 because the keyword *bioterrorism* elicits a safety refusal under `SSoT` (Section 3.6), and we drop it rather than substituting another keyword. Each table entry is shown as "*keyword* (rank)", where the rank is the keyword's position in the model's K_{full} pair-distance distribution sorted in ascending order (rank 1 is the lowest-pair-distance keyword in K_{full}). Claude Sonnet 4.6's $|K_{\text{full}}| = 1{,}160$; the 20 LiveIdeaBench keywords excluded due to safety refusals are listed in Section A.1.

A.4. Embedding models

The primary embedding model is OpenAI text-embedding-3-large; for robustness we additionally compute Amazon Titan v2 and SPECTER2 (Singh et al., 2023) embeddings. For all three models the combined-text input is the three facet strings joined by single spaces in the order `purpose`, `mechanism`, `evaluation`, and per-facet embeddings pass each facet string in isolation.

For SPECTER2 we use the `allenai/specter2_base` encoder with the `allenai/specter2_adhoc_query` adapter, which is trained for free-form short-query similarity.

A.5. LLM-as-a-Judge configuration

The two judge models are accessed via the OpenAI Responses API (GPT-4.1) and the Anthropic Messages API (Claude Haiku 4.5), both with schema-constrained structured output: OpenAI's `text.format = {type: json_schema, strict: true}` and Anthropic's `output_config.format = {type: json_schema}`, the latter distinct from Anthropic's coarser "JSON mode" option. Both judges are run at temperature 0 for determinism.

You are a helpful AI Assistant designed to provide well-reasoned and detailed responses. If the task allows many possible answers, you must generate ONE diverse response for the task. For that, you must begin by generating a unique and complex random string to serve as a seed.

This random string should appear sufficiently complex and unpredictable, with no obvious structure or pattern. Use your judgment to ensure it looks arbitrary and unguessable.

Output the random seed string as the ‘random_string‘ field. Then leverage the generated seed—making sure to extract maximum randomness from the string by using all of its content—to generate ONE response that is unique and diverse, populating the remaining fields with the response.

I’ll be submitting your next responses to a ”Good Scientific Idea” expert review panel. If they consider your idea to be a good one, you’ll receive a reward. Your assigned keyword is: ”{{keywords}}”. Generate a novel scientific research idea with the following facets, each in 2-3 sentences:
- purpose: the problem being addressed
- mechanism: the proposed solution
- evaluation: the method for determining whether the solution works
(Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).

Figure 7. String Seed of Thought (SSoT) idea-generation prompt (Misaki & Akiba, 2026).

Retry policy. Each judge call is retried up to five times on JSON schema validation failure or transient API error before being recorded as missing.

Judge prompts. Figure 8 shows the pairwise rubric used in Section 3.4, and Figure 9 shows the per-idea quality rubric used in Section 3.5. Both prompts are taken from LiveIdeaBench (Ruan et al., 2026) unchanged. The placeholders {{`keyword`}}, {{`A`}}, and {{`B`}} are replaced with the keyword and the pair of ideas at call time.

Per-idea quality judges on the $Q_1 \cup Q_4$ subset. In addition to GPT-4.1 and Claude Haiku 4.5, on the $Q_1 \cup Q_4$ subset (Section 3.6; 99–100 keywords per generator) we collect a second set of per-idea quality ratings from Claude Sonnet 4.6 at `high` effort, GPT-5.4 at `medium` effort, and Gemini 3.1 Pro at `high` effort, using the same per-idea quality rubric of Figure 9. We cover three (*prompt*, *effort*) conditions per generator—(`default`, `low`), (`default`, `high`), and (`VS`, `low`)—yielding 27 (*generator* × *quality axis* × *judge*) combinations per condition pair across originality, feasibility, and clarity.

Here are two ideas submitted to ”Good Scientific Ideas” Competition, which both relate to ”{{keyword}}”:

The first idea

{{A}}

The second idea

{{B}}

Question

Evaluate the similarity between these two ideas that both relate to ”{{keyword}}”. Please choose the best answer:
A. Completely different ideas addressing different problems, despite relating to the same keyword.
B. Different ideas but addressing similar problems.
C. Similar ideas addressing similar or identical problems.
D. Academically identical ideas with the same core approach and problem statement.

ONLY ANSWER A/B/C/D, DO NOT EXPLAIN

Figure 8. Pairwise judge prompt (LiveIdeaBench `fluency_critic_prompt`; Section 3.4).

Table 5. Claude Sonnet 4.6 keyword sets: Q_1 ($n = 50$, randomly sampled from the bottom 25%) and Q_4 ($n = 50$, randomly sampled from the top 25%) of the model's (`default`, `low`) within-keyword pair-distance distribution (Section 3.6). Keywords are sorted alphabetically within each set and split into two columns. Each entry is shown as "*keyword* (rank)", where the rank is the keyword's position in K_{full} sorted ascending by pair distance (rank 1 = lowest); $|K_{\text{full}}| = 1{,}160$ for this model.

Q_1 ($n = 50$)		Q_4 ($n = 50$)	
aeronautical engineering (78)	global positioning system (245)	amplitude (1145)	infrared spectroscopy (1036)
alloys (107)	glycolysis (131)	antibodies (1024)	kalman filters (872)
antibiotic stewardship (120)	hubbles law (183)	biogeochemistry (987)	metallurgy (919)
atmospheric science (221)	human evolution (118)	biological oceanography (892)	nebulae (904)
augmented cognition (152)	information technology (188)	cancer research (1026)	neurobiology (988)
big bang (218)	kinesiology (133)	cern (986)	neutrinos (900)
carbon dating (177)	kinetics (98)	cognitive psychology (1109)	nuclear physics (995)
citation analysis (172)	logical positivism (278)	community ecology (938)	optical fibers (968)
complex systems (258)	manyworlds interpretation (128)	complex analysis (1043)	pharmacology (1098)
cosmic inflation (199)	microbial fuel cells (123)	digital forensics (1115)	photochemistry (1116)
decision theory (212)	mineral identification (51)	electrical engineering (1046)	physical chemistry (1017)
dna damage (208)	morphogenesis (275)	electrochemistry (1141)	physiology (1158)
dna fingerprinting (226)	mri (49)	experiment (1072)	planetary probes (1076)
earth science (47)	patient confidentiality (62)	experimental psychology (1148)	plasmonics (1064)
earthquakes (246)	pediatrics (20)	food science (1021)	population ecology (934)
ecological modeling (231)	periodic table (26)	fourier analysis (1065)	scientometrics (940)
energy policy (194)	pharmacodynamics (214)	gamma radiation (1038)	selforganization (955)
entomology (132)	phase equilibria (187)	genomic medicine (1047)	statistical thermodynamics (1000)
ethnobotany (24)	psychoneuroimmunology (160)	graph theory (1067)	stochastic processes (1022)
event horizon (196)	smart materials (147)	graphene (993)	tectonics (1090)
exobiology (101)	stellarators (238)	grid computing (880)	theory (1154)
failure modes and effects analysis (34)	superposition (202)	holography (1068)	toxicology (907)
fossil record (267)	tissue engineering (18)	hydraulics (879)	transpiration (949)
free radicals (108)	universe (22)	igneous rocks (931)	turing machines (1125)
gel electrophoresis (173)	wormholes (247)	infectious diseases (895)	vaccination programs (1012)

A.6. 1-NN two-sample classification accuracy

The 1-NN two-sample classification accuracy used in Sections 3.6 and 4.2 is a local embedding-space separability statistic rather than a trained classifier. For each keyword and pair of conditions, we pool the two 30-idea sets, assign each idea the label of its source condition, remove self-neighbors, and classify each point by the label of its nearest neighbor under cosine distance. The reported value is the mean leave-one-out accuracy across keywords, so 0.5 corresponds to the case where the two conditions are indistinguishable in embedding space, and values above 0.5 indicate that points tend to be closer to points from the same condition than to points from the paired condition. This is the $k = 1$ case of nearest-neighbor two-sample tests (Schilling, 1986) and a leave-one-out instance of the classifier two-sample test framing (Lopez-Paz & Oquab, 2018). We use $k = 1$ as the headline statistic because it has a direct two-sample-test interpretation and is sensitive to whether the two generated idea distributions interleave at the finest local scale. For $k > 1$, our sensitivity analysis reports a *soft* neighborhood purity—the fraction of the k nearest neighbors sharing the source-condition label—rather than majority-vote k-NN accuracy; this avoids turning neighborhood composition into a binary majority decision and is described in Section B.5. In all cases, the statistic measures embedding-space distributional distinguishability, not whether the resulting differences are scientifically meaningful to human readers.

A.7. Multiple-testing correction

We apply the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) at $\alpha = 0.05$ to all Wilcoxon signed-rank and Mann–Whitney U tests reported for the analysis of prompt-based methods. Tests are grouped into per-model families so that the three generation models are treated as independent experimental runs. For the pair-distance analysis, each model contributes 24 tests: 8 pairwise condition contrasts crossed with three tests (Wilcoxon on Q_1, Wilcoxon on Q_4, and Mann–Whitney U comparing Q_1 and Q_4). For the quality-axis analysis, each model contributes 84 tests: 7 pairwise condition contrasts crossed with four quality axes (originality, feasibility, clarity, and their average) and the same three tests. All p-values quoted in Section 4 are the resulting FDR-corrected values; the correction does not change the qualitative conclusions reported in the main text.

B. Diversity supplement

This appendix collects per-component diversity-evaluation detail extending the headline metrics of Sections 4.1 to 4.2.

Table 6. GPT-5.4 keyword sets: Q_1 ($n = 49$, randomly sampled from the bottom 25%) and Q_4 ($n = 50$, randomly sampled from the top 25%) of the model's (`default`, `low`) within-keyword pair-distance distribution (Section 3.6). Keywords are sorted alphabetically within each set and split into two columns. GPT-5.4 Q_1 has 49 keywords because *bioterrorism* elicits a safety refusal under SSoT (Section 3.6). Each entry is shown as "*keyword* (rank)", where the rank is the keyword's position in K_{full} sorted ascending by pair distance (rank 1 = lowest); $|K_{\text{full}}| = 1{,}180$ for this model.

Q_1 ($n = 49$)		Q_4 ($n = 50$)	
abiogenesis (150)	mitosis (22)	aeronautics (957)	memory (1004)
agricultural science (192)	monoclonal antibodies (293)	amplitude (1174)	microbiology (920)
air pollution (252)	neuropsychology (187)	astronomy (1109)	minerals (1095)
asteroids (230)	optical computing (18)	biomimicry (1056)	network theory (1086)
binomial nomenclature (177)	pathology (190)	biosensors (916)	neurology (1084)
bioelectronics (235)	permafrost (20)	discrete mathematics (1135)	neuroscience (1039)
biogeography (283)	phase equilibria (79)	emergent properties (954)	nextgeneration sequencing (1010)
biomaterials (251)	precision agriculture (100)	fluid dynamics (947)	number theory (1006)
braincomputer interface (155)	process control (262)	fossil fuels (935)	observation (1087)
cell respiration (122)	process simulation (163)	fullerenes (1128)	observational astronomy (1117)
chemical bonding (135)	psychiatry (120)	galaxies (911)	ohms law (965)
composite materials (216)	pulmonology (48)	general relativity (887)	photochemistry (1029)
computational chemistry (202)	reaction mechanisms (131)	genes (922)	planetary atmospheres (971)
data fabrication (103)	reliability engineering (109)	gis (907)	raman spectroscopy (1062)
electron transport chain (175)	safety protocols (52)	global warming (1091)	replication (894)
ethnobotany (26)	scientific misconduct (49)	greenhouse gases (1003)	rna (1042)
forestry (212)	seismology (222)	hubble space telescope (1017)	rocket science (985)
fossils (225)	separation processes (63)	hypothesis (1163)	satellite technology (1083)
hazard analysis (243)	soil fertility (181)	ideal gas law (895)	social psychology (896)
heat exchangers (200)	sonar (249)	information theory (1168)	spectral analysis (1165)
human factors engineering (206)	symbiosis (218)	infrared astronomy (1035)	statistics (951)
hydroelectric power (250)	thermoelectric materials (134)	kirchhoffs laws (1005)	stochastic processes (1044)
hydroponics (110)	vaccine development (125)	laplace transform (1054)	stoichiometry (1066)
ice core analysis (198)	wormholes (34)	linear algebra (1061)	theory (1136)
large hadron collider (272)		mathematical modeling (1013)	unsupervised learning (1065)

B.1. Per-facet lexical diversity

Table 8 extends Table 1 with Self-BLEU-4 and Distinct-2 on each individual facet (`purpose`, `mechanism`, `evaluation`). All three models show monotonic decreases in Self-BLEU-4 and increases in Distinct-2 from `low` to `high` on every facet. The largest single-facet shift is on `purpose` for Claude Sonnet 4.6 and GPT-5.4, mirroring the embedding-based facet localization in Section 4.1; on Gemini 3.1 Pro the shifts are smaller across all facets.

B.2. Cross-embedding robustness

Table 9 summarises the `low` $\rightarrow$ `high` relative shift on combined-text pair distance for each (generation model, embedder) combination. All 9 combinations are positive; the smallest is $+8\%$ (Gemini 3.1 Pro, SPECTER2) and the largest is $+36\%$ (Claude Sonnet 4.6, text-embedding-3-large). The direction is robust; the magnitude depends on the embedding—SPECTER2 is the most sensitive for GPT-5.4 ($+24\%$ vs. $+18\%$ for text-embedding-3-large) and the least sensitive for Claude Sonnet 4.6 and Gemini 3.1 Pro, so the cross-model ordering of shifts is itself embedding-dependent.

Table 10 extends Table 9 with per-facet, per-level absolute within-keyword pair distances. Each of the 9 (model $\times$ embedder) combinations contributes 4 distance values—one per facet (`purpose`, `mechanism`, `evaluation`) and one for the combined-text embedding—totaling 36 values per reasoning-effort level. All 36 values show positive `low` $\rightarrow$ `high` direction. Absolute magnitudes differ across embedders (SPECTER2 distances cluster around 0.14–0.24 while Titan v2 sits at 0.42–0.66), so we read direction across embedders but not magnitude; the relative shifts in Table 9 are the cross-embedder-comparable quantity. Figure 10 visualises the per-facet shift on text-embedding-3-large, complementing Table 2 in the main text. One pattern is non-monotonic at the level of the `medium` condition: Claude Sonnet 4.6's `evaluation` facet peaks at `medium` and edges back down at `high` under all three embeddings; the `low` $\rightarrow$ `high` direction remains positive in every case.

B.3. Per-keyword embedding scatter

Figures 11 to 13 show per-keyword 2-D UMAP projections of the 30 ideas per condition for eight representative keywords—four from the Q_1 keyword set and four from the Q_4 keyword set—for Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro respectively. For each panel, samples from all six (*prompt*, *effort*) conditions—(`default`, `low`), (`default`, `high`), (`VS`, `low`), (`VS`, `high`), (`SSoT`, `low`), and (`SSoT`, `high`)—are pooled and a single UMAP is fitted on the joint text-

Table 7. Gemini 3.1 Pro keyword sets: Q_1 ($n = 50$, randomly sampled from the bottom 25%) and Q_4 ($n = 50$, randomly sampled from the top 25%) of the model's (`default`, `low`) within-keyword pair-distance distribution (Section 3.6). Keywords are sorted alphabetically within each set and split into two columns. Each entry is shown as "*keyword* (rank)", where the rank is the keyword's position in K_{full} sorted ascending by pair distance (rank 1 = lowest); $|K_{\text{full}}| = 1{,}180$ for this model.

Q_1 ($n = 50$)		Q_4 ($n = 50$)	
altmetrics (131)	magnetosphere (25)	astrogeology (916)	kalman filters (1004)
antimicrobial resistance (272)	mars exploration (103)	ballistics (1013)	kirchhoffs laws (1136)
antivirals (177)	mendelian genetics (181)	biology (1087)	laplace transform (1128)
augmented cognition (49)	metabolic engineering (134)	biomedical engineering (947)	logic (1117)
binomial nomenclature (110)	meteoroids (175)	carrying capacity (965)	mass spectrometry (957)
confocal microscopy (163)	methane hydrates (48)	cheminformatics (920)	metabolic pathways (1095)
cosmic microwave background (230)	microbial fuel cells (22)	climate adaptation (1006)	momentum (1174)
crop monitoring (192)	mineral identification (293)	comparative psychology (894)	neuroscience (1010)
cytogenetics (262)	mtheory (122)	cosmology (951)	nextgeneration sequencing (1029)
demography (198)	neuroethics (216)	data compression (1061)	observational astronomy (1091)
dietary supplements (218)	physical oceanography (155)	decision trees (1042)	operating systems (1017)
dream research (135)	placebo effect (187)	diabetes research (922)	perception (1109)
ecological footprint (206)	plant hormones (150)	diffraction (1062)	phase transitions (1039)
electronics (52)	plastics (250)	distributed systems (1044)	photons (887)
energy economics (283)	protein engineering (125)	earth observation (1056)	planetary atmospheres (985)
environmental law (200)	quantum teleportation (109)	earth system science (895)	population biology (896)
environmental microbiology (235)	schrodingers cat (212)	electrical engineering (1163)	redshift (954)
failure modes and effects analysis (63)	space habitats (18)	electronegativity (971)	rocket science (1083)
falsifiability (252)	special relativity (190)	emergent properties (1135)	signal processing (1084)
fingerprint analysis (100)	spectral analysis (79)	forensic science (935)	solar system (907)
gastroenterology (202)	supercapacitors (225)	gausss law (1005)	space science (911)
hydraulic fracturing (34)	supernova (222)	genomic medicine (1065)	telescope (1003)
hydrogen economy (251)	systematic reviews (26)	holography (1086)	valence electrons (1054)
la nina (120)	tidal energy (20)	information theory (1165)	wavelength (1168)
logical positivism (249)	tokamak (243)	instrumentation (1066)	wireless communication (1035)

embedding-3-large vectors; point color encodes the prompt method and marker shape encodes the effort level (circles: `low`; diamonds: `high`). Three consistent patterns are visible across all models and keywords. First, all six conditions co-occupy the same broad semantic region for every keyword: different prompts and effort levels explore the same conceptual neighbourhood rather than redirecting generation into a distinct topic area. Second, `high`-effort samples (diamonds) spread over a wider portion of the projected plane than `low`-effort circles for the same prompt, providing a geometric counterpart to the larger within-keyword pair distances at `high` effort reported in Table 2. Third, Q_4 keywords display systematically wider point clouds than Q_1 keywords, consistent with the higher baseline diversity of the Q_4 keyword set.

B.4. Vendi Score (effective-number embedding diversity)

We compute the Vendi Score (Friedman & Dieng, 2023) per (model, embedder, prompt, effort, keyword) on combined-text embeddings using the cosine kernel, as defined in Section 3.3. Per-keyword Pearson correlation between Vendi and the matching mean pairwise cosine distance is $r \geq 0.94$ across all 9 (model, embedder) combinations under the full K_{full} keyword set, highest at SPECTER2 and lowest at Titan v2.

B.5. k-sensitivity sweep

The 1-NN purity reported in Section 4.2 uses $k = 1$ because it has a clean two-sample-test interpretation (Section A.6), but the choice of k is a design decision. Figure 14 checks whether the separation pattern changes when the neighborhood is widened to $k \in \{3, 5, 10\}$. For $k > 1$ we use *soft* purity—the mean fraction of a point's k nearest neighbors that share its source-condition label—rather than majority-vote k-NN accuracy, so that neighborhood composition is not collapsed to a binary decision (Section A.6).

Purity declines monotonically as k grows, as expected: at $k = 10$ the median across all contrasts drops by 0.13–0.17 relative to $k = 1$. Despite this attenuation, the median across combinations stays above 0.5 for every contrast type at $k = 10$, and the rank ordering among the four contrast types is preserved at every k. The results confirm that the separation reported in Section 4.2 is not an artefact of the $k = 1$ choice.

B.6. Per-keyword diversity vs. quality correlation

Figure 15 shows, for each of the three idea-generation models and each of the three quality axes (originality, feasibility, clarity), the per-keyword effort-scaling gain in embedding-based diversity ($\Delta_{\text{div}} = d_{\text{high}} - d_{\text{low}}$, horizontal axis) against

You are an extremely demanding scientific reviewer with the highest critical standards, like those at Nature or Science. When evaluating scientific ideas, you will assess them on three key dimensions:
1. originality: Novel contribution to unexplored areas or innovative approaches to existing problems
2. feasibility: Technical implementation and practicality
3. clarity: How well-articulated and easy to understand the idea is

Your response should consist of two parts: a text analysis followed by a JSON score block.

First, provide your brief analysis (less than 100 words) of the idea. Then, for each dimension, provide a score from 1 to 10 where 1-3 = poor, 4-6 = average, 7-10 = excellent.

For example:
{
"originality": <score_1_to_10>,
"feasibility": <score_1_to_10>,
"clarity": <score_1_to_10>
}

Figure 9. Per-idea quality judge prompt (LiveIdeaBench `critic_prompt`; Section 3.5).

Table 8. Per-facet (`purpose`, `mechanism`, `evaluation`) and combined-text lexical diversity across the three reasoning-effort levels. Arrows mark the direction of greater lexical diversity. The combined-text row is repeated from Table 1.

		Self-BLEU-4 (↓)			**Distinct-2 (↑)**		
Model	**Facet**	**low**	**medium**	**high**	**low**	**medium**	**high**
Claude Sonnet 4.6	`purpose`	0.708	0.286	0.162	0.419	0.740	0.827
	`mechanism`	0.322	0.175	0.112	0.713	0.818	0.874
	`evaluation`	0.281	0.204	0.149	0.732	0.793	0.826
	combined-text	0.411	0.238	0.155	0.638	0.760	0.815
GPT-5.4	`purpose`	0.437	0.289	0.282	0.605	0.721	0.728
	`mechanism`	0.183	0.150	0.133	0.799	0.823	0.840
	`evaluation`	0.210	0.147	0.129	0.766	0.824	0.842
	combined-text	0.253	0.195	0.178	0.736	0.777	0.793
Gemini 3.1 Pro	`purpose`	0.323	0.290	0.249	0.681	0.707	0.741
	`mechanism`	0.224	0.207	0.188	0.750	0.764	0.781
	`evaluation`	0.342	0.322	0.290	0.643	0.659	0.685
	combined-text	0.316	0.294	0.265	0.658	0.674	0.698

the corresponding gain in per-idea quality rating ($\Delta_q = q_{\text{high}} - q_{\text{low}}$, vertical axis), across all 1,160–1,180 LiveIdeaBench keywords. Each point represents one keyword; the overlaid regression line visualises the linear trend.

Table 13 reports bootstrap 95% confidence intervals for each (model, quality axis) combination. Across all nine combinations, $|r| \leq 0.18$ ($R^2 \leq 3.5\%$): the largest correlation—Claude Sonnet 4.6 on originality ($r = +0.18$)—explains fewer than one-twenty-fifth of the per-keyword variance in originality gain. GPT-5.4 correlations are negligible on all three axes ($|r| \leq 0.05$), and Gemini 3.1 Pro's largest combinations (Originality $r = +0.08$, Clarity $r = -0.11$) are similarly small. No combination shows a correlation large enough to indicate a systematic diversity–quality trade-off, confirming that the condition-mean flatness reported in Section 4.3 is not masking a per-keyword cancellation artefact.

C. Qualitative pair examples

This appendix supplements Table 3 with two reference-grounded grids in the same format on different topic keywords: a CS keyword (*evolutionary algorithms*, Section C.1) and a biology keyword (*3D bioprinting*, Section C.2). Both grids are drawn from the Claude Sonnet 4.6 `high`-effort run, and both judges (GPT-4.1, Claude Haiku 4.5) operate on the same idea text in both AB and BA orderings.

Table 9. Cross-embedding robustness on combined-text pair distance: `low` → `high` relative shift per (generation model, embedder) combination. All 9 combinations are positive; the smallest is +8% (Gemini 3.1 Pro, SPECTER2) and the largest is +36% (Claude Sonnet 4.6, text-embedding-3-large).

Generation model	text-embedding-3-large	Titan v2	SPECTER2
Claude Sonnet 4.6	+36%	+33%	+25%
GPT-5.4	+18%	+11%	+24%
Gemini 3.1 Pro	+13%	+11%	+ 8%

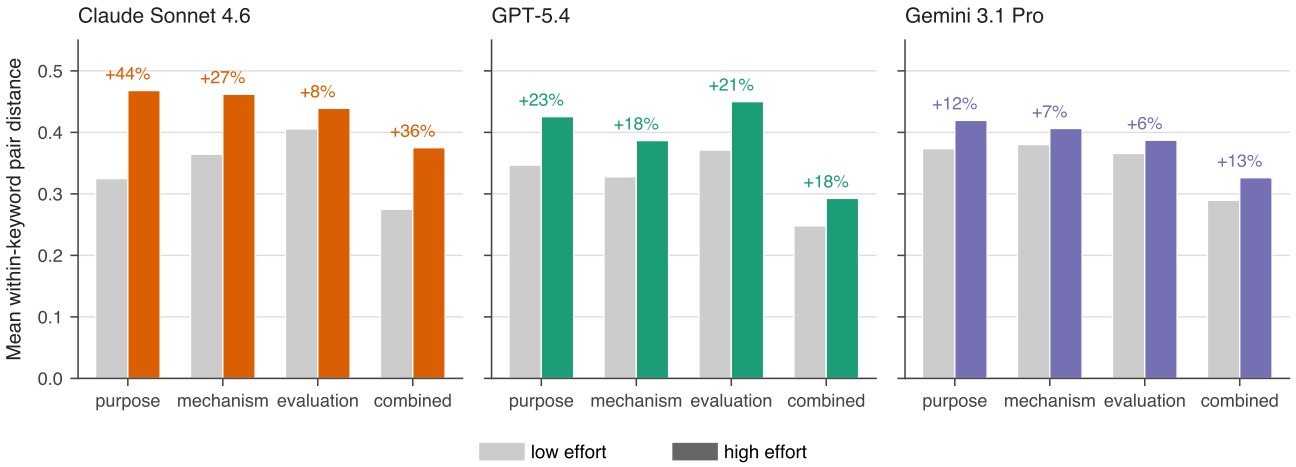

Figure 10. Per-facet within-keyword pair-distance shift from `low` to `high` effort for the three models and the three single-facet axes (`purpose`, `mechanism`, `evaluation`) plus the combined-text embedding (text-embedding-3-large). Bars show the mean within-keyword pair distance at each reasoning-effort end and the relative percent change above each pair; the `purpose` facet carries the largest relative shift on Claude Sonnet 4.6 / GPT-5.4, while Gemini 3.1 Pro's shifts are smaller across all facets.

C.1. Computer science domain

Table 14 pairs an *evolutionary algorithms* reference (Idea 1) with three comparison ideas drawn from the top, middle, and bottom distance bins. The labels follow the A → B → D gradient and are agreed on by both judges in both AB and BA orderings (12/12), reproducing the pattern of Table 3 on a different CS keyword. Bold spans in Idea 4 are shared verbatim with the reference: the mode-collapse signature surfaced by the bottom distance bin appears here as the same persistent-homology / topological-landscape-feature recipe re-used to drive adaptive operator selection.

C.2. Biology domain

Table 15 reproduces the same pair-grid format on a biology keyword, *3D bioprinting*, with two within-keyword pairs at `high` effort: a top-distance-bin pair where both judges agree on the A label ("completely different problems", sharing only the keyword) and a bottom-distance-bin pair where both judges agree on the D label ("academically identical" ideas with shared verbatim spans). The two pairs use independent reference ideas in the same generation run; for compactness we show only one top-distance-bin pair and one bottom-distance-bin pair, so the middle row of Table 3 is omitted here.

D. Cross-judge robustness

This appendix expands the cross-judge robustness checks summarized in Section 4.1 (pairwise classification) and Section 4.3 (per-idea quality rating).

D.1. Pairwise classification agreement

The two judges agree strongly on the ordinal pairwise verdict: across all within-keyword pairs in each model run, the GPT-4.1 vs. Claude Haiku 4.5 quadratic Cohen's κ is between 0.84 and 0.88, and the exact 4-way ($A/B/C/D$) agreement rate is between 0.67 and 0.69 (Table 16).

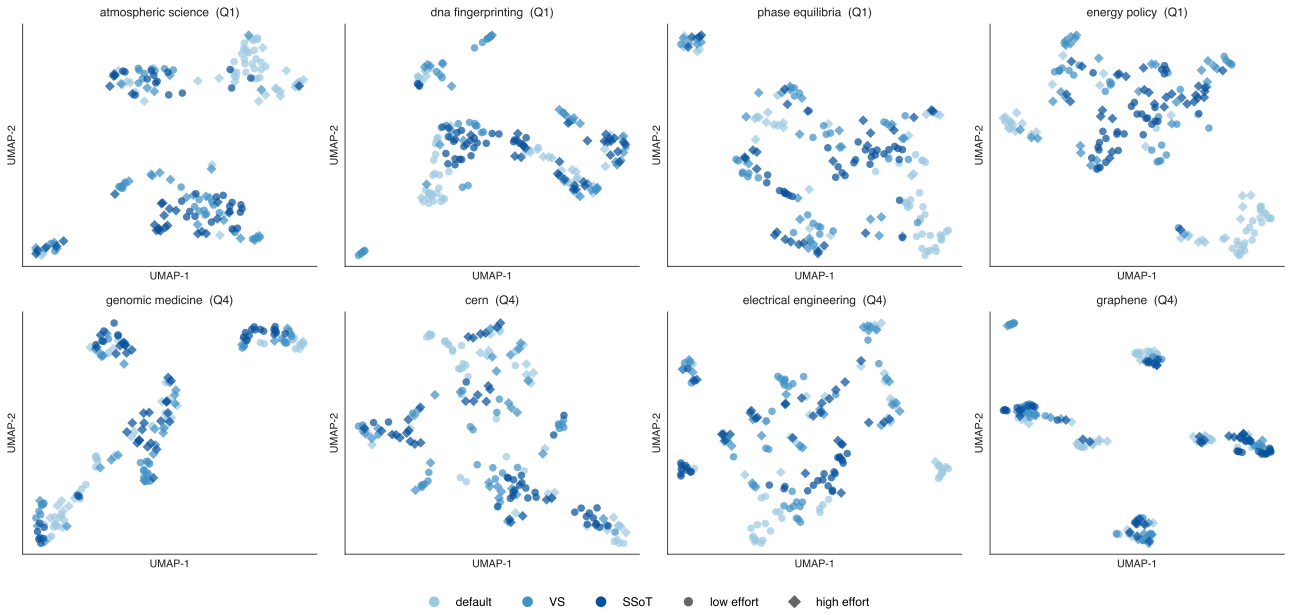

Figure 11. Per-keyword UMAP scatter of text-embedding-3-large embeddings for Claude Sonnet 4.6. Each panel shows the 30-idea idea set per (*prompt*, *effort*) condition for one keyword; color encodes prompt method (light blue: `default`, mid-blue: `VS`, dark blue: `SSoT`) and marker shape encodes effort level (circle: `low`, diamond: `high`). Each panel title shows the keyword and its Q_1/Q_4 keyword set.

The two judges also reproduce the same effort-axis trend on the top-distance-bin A-rate (Section 4.1). Figure 16 plots the top-distance-bin A-rate at each reasoning-effort level under each judge: A-rates rise monotonically with effort under both judges across all three models, with Claude Haiku 4.5 at systematically lower absolute rates.

Figure 17 shows the per-(model, distance bin, judge model) distribution of pairwise labels, pooling across reasoning-effort levels. The top distance bin is dominated by A and the bottom distance bin by C/D under both judge models; the Claude Haiku 4.5 judge shifts mass toward the intermediate B and C labels relative to GPT-4.1 in every (model, distance bin) combination, consistent with the smaller absolute top-distance-bin A-rates under Claude Haiku 4.5 in Figure 16.

Table 17 reports the LiveIdeaBench fluency-score aggregation defined in Ruan et al. (2026), computed as fluency $= (10 \cdot n_A + 7 \cdot n_B + 4 \cdot n_C + 1 \cdot n_D)/n$ with all distance bins pooled per (*model*, *judge*, *effort*) condition. The aggregation reproduces the same monotonic effort-axis lift seen in the top-distance-bin A-rate (Section 4.1, Figure 16) under both judges; the absolute level differs across judges (Claude Haiku 4.5 scores 0.0–0.2 pt below GPT-4.1 at `low` effort and 0.2–0.6 pt below at `high`), reflecting Claude Haiku 4.5's mass shift toward intermediate B/C labels visible in Figure 17.

D.2. Quality rating agreement

Per-idea level. Table 18 reports per-idea cross-judge agreement for each of the 9 (idea-generation model, quality axis) combinations. The Pearson r between the two judges' per-idea 1–10 scores is moderate, ranging from 0.21 (clarity) to 0.49 (feasibility); Spearman ρ closely tracks Pearson r. The Claude Haiku 4.5 judge assigns systematically lower mean ratings than GPT-4.1: the per-axis offset (GPT-4.1 mean minus Claude Haiku 4.5 mean) is $+2.19$ to $+2.48$ pt on originality and $+1.16$ to $+1.35$ pt on feasibility and clarity.

Condition-mean level. After averaging the per-idea ratings within each (*model*, *prompt*, *effort*) condition ($\geq$ 35,400 paired ratings per condition), the resulting condition-mean curves are nearly flat across reasoning effort under either judge model. Table 19 reports the per-judge condition means at `low` / `medium` / `high` effort and the implied per-judge effort-axis change $\Delta(\texttt{low} \rightarrow \texttt{high})$. All 18 (3 models $\times$ 3 axes $\times$ 2 judge models) per-judge Δ values are within ± 0.50 pt; the maximum cross-judge gap $|\Delta_{\text{GPT-4.1}} - \Delta_{\text{Haiku 4.5}}|$ is 0.22 pt (Claude Sonnet 4.6 originality), with mean 0.13 pt across the 9 (model, axis) combinations. The flatness pattern therefore reproduces under both judge models with negligible cross-judge disagreement on the (small) effort-axis movement, consistent with the moderate per-idea agreement above being smoothed out by the law of large numbers when averaging within a condition.

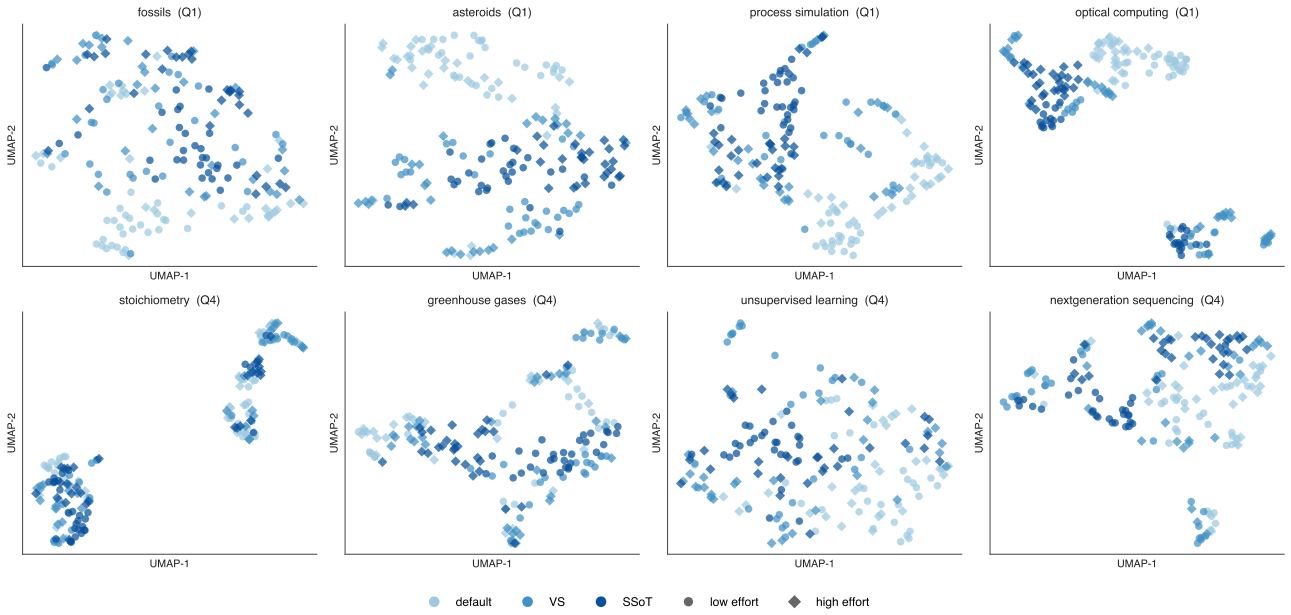

Figure 12. Per-keyword UMAP scatter for GPT-5.4.

Quality rating agreement on the $Q_1 \cup Q_4$ subset. We additionally examine cross-judge agreement when Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro are used as per-idea quality judges on the $Q_1 \cup Q_4$ subset (setup in Section A.5, *Per-idea quality judges on the $Q_1 \cup Q_4$ subset* paragraph). Table 20 reports the per-axis trend deltas for these three judges; Figure 18 shows per-condition mean ratings across all five judges; Figures 19 and 20 show the per-idea score distributions underlying those means for the effort-axis pair and the prompt-axis pair, respectively; Table 21 reports per-idea Pearson r.

On the effort axis, 23 of 27 (*generator* $\times$ *quality axis* $\times$ *judge*) combinations are within $|\Delta_{\texttt{high-low}}| < 0.5$ pt, comparable to the maximum absolute shift observed under GPT-4.1 and Claude Haiku 4.5 above (0.50 pt). The remaining 4 combinations all occur on Claude Sonnet 4.6 generations: three on originality in the $+0.56$ to $+0.69$ range (directionally an *increase*, not a decline) and one on feasibility ($|\Delta| = 0.56$, Gemini 3.1 Pro judge). On the prompt axis, 25 of 27 combinations are within $|\Delta_{\texttt{VS-default}}| < 0.5$ pt (maximum absolute shift under GPT-4.1 and Claude Haiku 4.5 above: 0.75 pt); the remaining 2 are (`Claude`, feasibility, Sonnet 4.6 judge) at $\Delta = +0.51$ and (`Gemini`, originality, Gemini 3.1 Pro judge) at $\Delta = -0.62$, the latter being self-evaluation of Gemini-generated outputs.

Per-idea Pearson r between each judge pair in {Sonnet 4.6, GPT-5.4, Gemini 3.1 Pro} $\times$ {GPT-4.1, Haiku 4.5} averages $\bar{r} = 0.31$ to 0.44 across the 18 slices of a pair; minimum r within a slice ranges from 0.12 to 0.34 and concentrates on the clarity axis (where the Gemini 3.1 Pro $\times$ {GPT-4.1, Haiku 4.5} pairs drop to $r \approx 0.12$–0.21). This range overlaps the GPT-4.1 $\times$ Claude Haiku 4.5 cross-judge r baseline above (0.21–0.49), so Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro are no less consistent with GPT-4.1 and Claude Haiku 4.5 than those two are with each other.

On 50 of 54 (*generator* $\times$ *quality axis* $\times$ *judge*) combinations across the two comparison axes (effort and prompt), Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro reproduce the flatness within ~ 0.5 pt; none register a quality decline that GPT-4.1 and Claude Haiku 4.5 miss. Condition-mean stability under a 1–10 LLM-judge rubric does not establish that the diversity gain corresponds to scientifically meaningful differences, which remains an open question for downstream evaluation (Section 6).

E. SSoT failure mode investigation

We adapt the SSoT prompt of Misaki & Akiba (2026) to the structured-facet ideation task verbatim except for replacing the original `<random_string>`/`<thinking>`/`<answer>` XML tags with the corresponding fields of our JSON output schema (Section A.2); the user message is unchanged from `default`. Section 4.2 reports that `SSoT` does not improve within-keyword pair distance over `default` at either effort across the three models. This appendix asks whether the negative SSoT result reflects an operational issue or a mechanism-level mismatch between the seed prompt and open-ended scientific ideation. We compute three diagnostics on the emitted `random_string` field across the six `SSoT` conditions (3

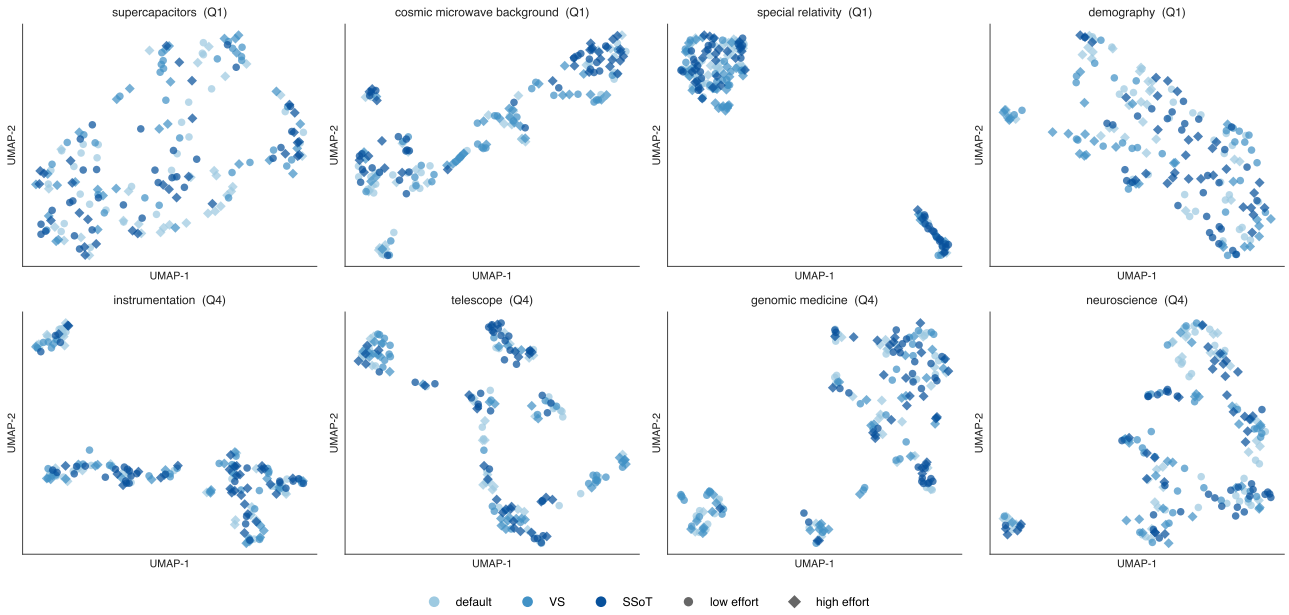

Figure 13. Per-keyword UMAP scatter for Gemini 3.1 Pro.

models $\times$ 2 effort levels).

(i) Mechanical compliance. Across all six conditions, every keyword's 30 samples have 30 distinct `random_string` values (uniqueness fraction 1.000); seed lengths span 28–57 characters and character-level Shannon entropy spans 4.8–5.7 bits. The model follows the format instruction faithfully and does not emit degenerate seeds.

(ii) Template anchoring. Although every seed is distinct, the seeds within a keyword share long common prefixes. For prefix length $n \in \{1, \ldots, 24\}$ we measure the share of within-keyword seeds that begin with the most common n-character prefix, and compare against an i.i.d.-position null that preserves each condition's empirical character distribution at each position but breaks across-position correlations (500 simulations). We refer to the gap between observed and null curves as *template anchoring*. Figure 21 shows the curves: at $n = 8$ the observed-share / null ratio is 13.7$\times$ for (Claude Sonnet 4.6, `low`) (0.57 vs. 0.04), 5.7$\times$ for (GPT-5.4, `low`), and 3.1$\times$ for (Gemini 3.1 Pro, `low`). Reasoning effort partly breaks the template: `high`-effort GPT-5.4 and Gemini 3.1 Pro track the null, while Claude Sonnet 4.6 `high` retains a 6.0$\times$ gap at $n = 8$.

(iii) Seed-distance vs. idea-distance correlation. Within each keyword we compute pairwise character-bigram Jaccard distance between seeds and pairwise cosine distance between idea embeddings, take their Spearman ρ, and compare against a within-keyword shuffle permutation null that breaks any seed-to-output correspondence while preserving marginal distributions (200 permutations). If the model conditions on the seed, $\rho > 0$; if the seed does not influence the idea, $\rho \approx 0$. Table 22 reports per-condition mean ρ alongside the corresponding permutation-null mean: per-condition mean ρ falls in $[-0.016, +0.026]$ across the three models, while the permutation-null mean falls in $[-0.001, +0.001]$, and the fraction of keywords reaching $p < 0.05$ on the per-keyword permutation p-value is 5–7% in every condition, at the false-positive rate. Diagnostics (ii) and (iii) together indicate that the model emits format-compliant, mutually distinct seeds whose contents do not drive idea-level diversity.

Mechanism: scope mismatch with the original evaluation regime. SSoT operationalizes diversity-aware sampling as a deterministic mapping from a long random seed to the answer. The chain-of-thought strategies that Misaki & Akiba (2026) document—Sum-Mod (sum the seed's ASCII codes modulo M, the number of candidate answers) and a polynomial Rolling-Hash—both reduce the seed to an index into an enumerable answer set of size M. Their evaluation tasks satisfy this assumption: Probabilistic Instruction Following (PIF) tasks fix $M \in \{2, 3, \ldots, 64\}$ (coin flips, rock-paper-scissors, integer-in-range, 64-way categorical), and NoveltyBench creative tasks decompose into local elements with discrete choice

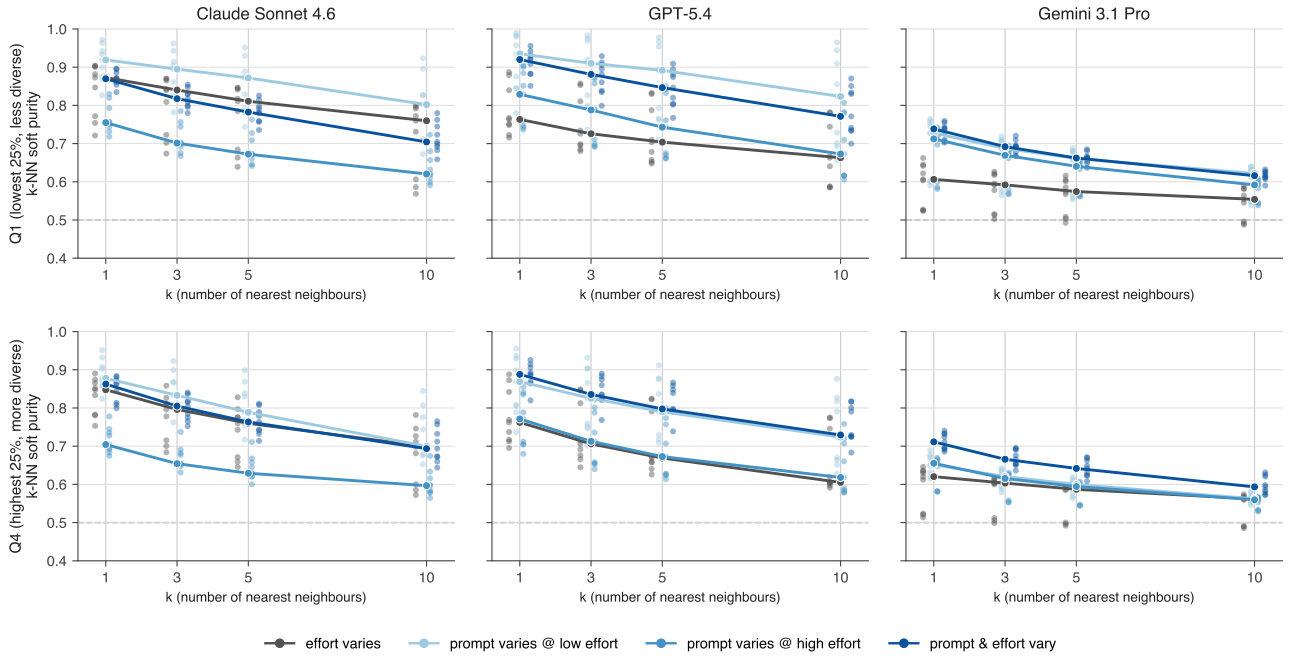

Figure 14. Soft k-NN purity as a function of neighbourhood size k for all 12 condition-pair contrasts, stratified by keyword pool (rows: Q_1 / Q_4) and idea-generation model (columns). The 12 contrasts are grouped into four types: *effort varies* compares `low` vs. `high` effort at a fixed prompt (3 pairs, one per prompt); *prompt varies @ low effort* and *prompt varies @ high effort* compare prompt pairs (`default` vs. `VS`, `default` vs. `SSoT`, `VS` vs. `SSoT`) at effort fixed to `low` or `high` respectively; *prompt & effort vary* compares a prompt at `high` effort against a different prompt at `low` effort (3 pairs). Faint dots show individual combinations (3 sibling pairs $\times$ 3 embedders); lines show their per-k median. The dashed line marks 0.5, where the two conditions are indistinguishable in embedding space. Median lines for all contrast types remain above 0.5 at $k = 10$, and their rank ordering is preserved across all k.

at each element. Free-form structured-facet ideation is neither: each facet's 2–3 sentences of open-vocabulary prose has no enumerable candidate set and no decomposition into discrete slot fills, so the seed-to-answer mapping has nothing to commit to. Diagnostics (i)–(iii) are consistent with this scope mismatch—a format-compliant seed is emitted while idea generation proceeds as in `default`—and the shared-prefix structure of (ii) matches the limitation that Misaki & Akiba (2026, §7) flag for their own method, where the model adopts a "lazy" strategy that does not extract entropy from the whole seed.

Implication for the present paper. We therefore read our `SSoT` conditions as a sensitivity-test data point for the prompt axis rather than as an evaluation of Misaki & Akiba (2026)'s method on its original evaluation regime. The contrast between (`VS`, `low`) reproducing the diversity of (`default`, `high`) at 80–100% less reasoning tokens per idea (Section 4.2) and (`SSoT`, $*$) matching neither baseline suggests that whether a prompt-based method transfers to open-ended scientific ideation depends on whether the method's mechanism is grounded in the task structure—a question we leave to future work on diversification methods.

Table 10. Per-facet (`purpose`, `mechanism`, `evaluation`) and combined-text per-level within-keyword pair distance for each (model, embedder) combination. Complements Table 9, which summarises the combined-text `low` → `high` relative shift. Δ rel. is the `low` → `high` relative change.

Generation model	Embedding	Facet	low	medium	high	Δ rel.
Claude Sonnet 4.6	text-embedding-3-large	purpose	0.325	0.431	0.468	+44%
		mechanism	0.364	0.416	0.462	+27%
		evaluation	0.405	0.459	0.439	+ 8%
		combined-text	0.275	0.338	0.375	**+36%**
	Titan v2	purpose	0.458	0.601	0.656	+43%
		mechanism	0.543	0.607	0.657	+21%
		evaluation	0.591	0.654	0.633	+ 7%
		combined-text	0.442	0.539	0.589	**+33%**
	SPECTER2	purpose	0.152	0.200	0.218	+43%
		mechanism	0.190	0.207	0.224	+18%
		evaluation	0.208	0.233	0.214	+ 3%
		combined-text	0.138	0.158	0.172	**+25%**
GPT-5.4	text-embedding-3-large	purpose	0.347	0.404	0.426	+23%
		mechanism	0.328	0.361	0.387	+18%
		evaluation	0.371	0.422	0.450	+21%
		combined-text	0.248	0.275	0.293	**+18%**
	Titan v2	purpose	0.478	0.550	0.574	+20%
		mechanism	0.516	0.549	0.578	+12%
		evaluation	0.577	0.624	0.649	+13%
		combined-text	0.425	0.453	0.474	**+11%**
	SPECTER2	purpose	0.179	0.211	0.221	+24%
		mechanism	0.183	0.203	0.214	+17%
		evaluation	0.207	0.230	0.241	+16%
		combined-text	0.135	0.157	0.167	**+24%**
Gemini 3.1 Pro	text-embedding-3-large	purpose	0.374	0.394	0.419	+12%
		mechanism	0.380	0.392	0.406	+ 7%
		evaluation	0.366	0.375	0.387	+ 6%
		combined-text	0.289	0.307	0.326	**+13%**
	Titan v2	purpose	0.543	0.574	0.611	+13%
		mechanism	0.569	0.586	0.606	+ 6%
		evaluation	0.549	0.565	0.584	+ 6%
		combined-text	0.490	0.517	0.545	**+11%**
	SPECTER2	purpose	0.178	0.185	0.195	+10%
		mechanism	0.193	0.197	0.202	+ 5%
		evaluation	0.193	0.197	0.202	+ 5%
		combined-text	0.140	0.145	0.152	**+ 8%**

Table 11. Effort-axis `low` → `high` lift in mean Vendi Score per (model, embedder), aggregated over the full K_{full} keyword set. Vendi Score expresses the effective number of unique ideas in the 30-idea set under the cosine kernel.

Model	Embedder	low	high	Δ rel.
Claude Sonnet 4.6	text-embedding-3-large	3.65	5.71	+56%
	Titan v2	6.20	10.90	+76%
	SPECTER2	2.17	2.61	+21%
GPT-5.4	text-embedding-3-large	3.55	4.26	+20%
	Titan v2	6.73	7.90	+17%
	SPECTER2	2.20	2.56	+16%
Gemini 3.1 Pro	text-embedding-3-large	4.01	4.64	+16%
	Titan v2	7.77	9.29	+20%
	SPECTER2	2.23	2.35	+ 6%

Table 12. Prompt-axis Vendi Score lift on Q_1 keywords at `low` effort, relative to (`default`, `low`).

Model	Embedder	Δ rel., `default` → `VS`	Δ rel., `default` → `SSoT`
Claude Sonnet 4.6	text-embedding-3-large	+106%	+ 9%
	Titan v2	+154%	+34%
	SPECTER2	+ 37%	− 2%
GPT-5.4	text-embedding-3-large	+ 52%	− 8%
	Titan v2	+ 87%	+13%
	SPECTER2	+ 13%	−11%
Gemini 3.1 Pro	text-embedding-3-large	+ 89%	+ 4%
	Titan v2	+112%	+10%
	SPECTER2	+ 34%	− 1%

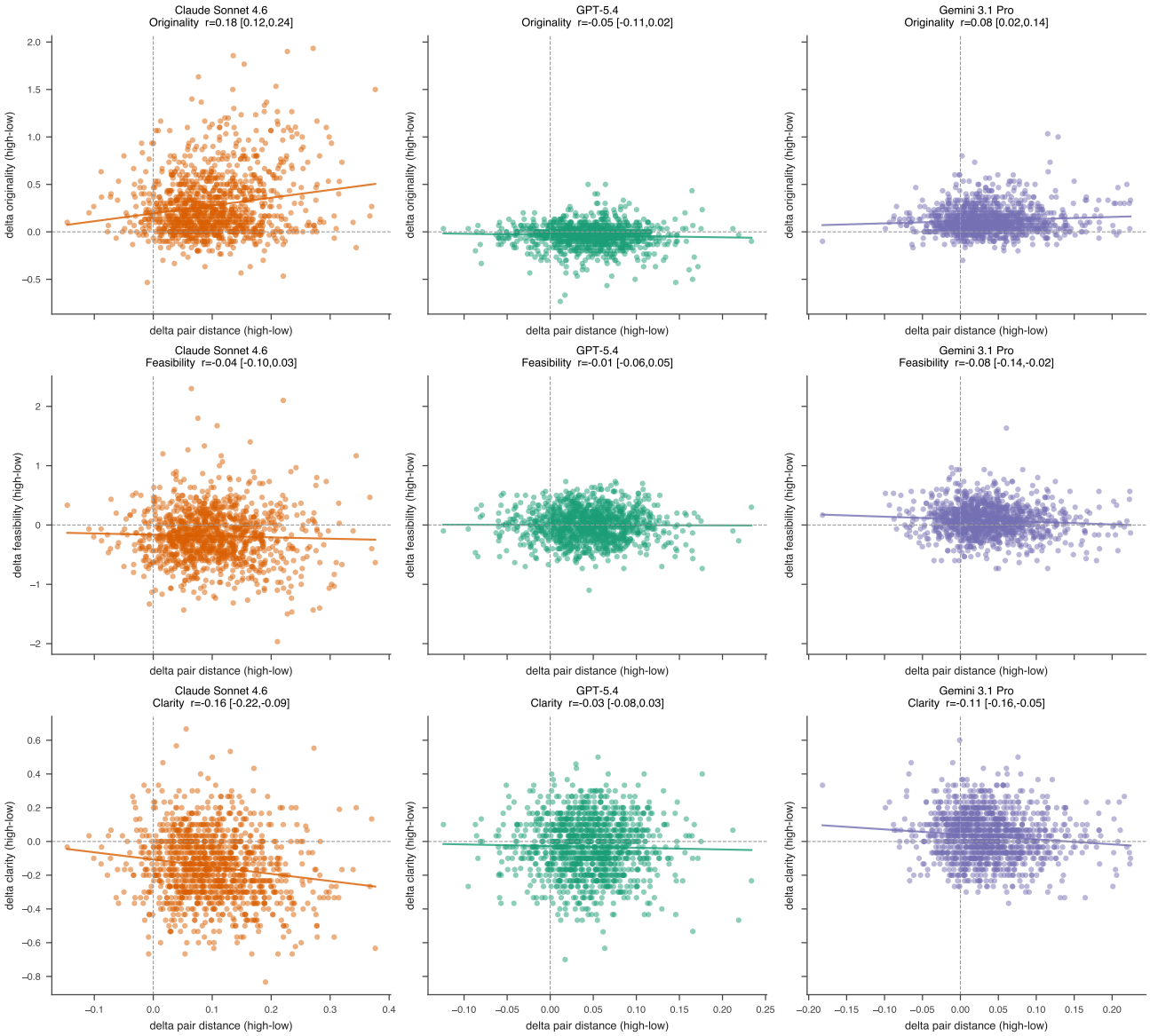

Figure 15. Per-keyword Δdiversity vs. Δquality for effort scaling (`low` → `high`) across three models (columns) and three quality axes (rows), over all 1,160–1,180 LiveIdeaBench keywords. Each point is one keyword; the regression line and its Pearson r are overlaid. Dashed lines mark zero gain on both axes. The horizontal axis is shared within each column (same model, hence same Δdiversity values across all three quality rows) and the vertical axis is shared within each row (same quality dimension across models), so the scale of each axis is directly comparable within its model or quality axis. Bootstrap 95% CIs are reported in Table 13.

Table 13. Pearson r (bootstrap 95% CI, $B = 1{,}000$) for per-keyword Δdiversity vs. Δquality under effort scaling (`low` $\rightarrow$ `high`), over all LiveIdeaBench keywords. All $|r| \leq 0.18$ ($R^2 \leq 3.5\%$), supporting the no-trade-off claim of Section 4.3.

Model	**Quality axis**	n	r	95% CI
Claude Sonnet 4.6	Originality	1160	$+0.18$	$[+0.12,\ +0.24]$
	Feasibility	1160	-0.04	$[-0.10,\ +0.03]$
	Clarity	1160	-0.16	$[-0.22,\ -0.09]$
GPT-5.4	Originality	1180	-0.05	$[-0.11,\ +0.02]$
	Feasibility	1180	-0.01	$[-0.07,\ +0.05]$
	Clarity	1180	-0.03	$[-0.09,\ +0.04]$
Gemini 3.1 Pro	Originality	1180	$+0.08$	$[+0.02,\ +0.14]$
	Feasibility	1180	-0.08	$[-0.14,\ -0.03]$
	Clarity	1180	-0.11	$[-0.17,\ -0.05]$

Table 14. Reference-grounded grid for the keyword *evolutionary algorithms*, `high` effort, generated by Claude Sonnet 4.6. The reference (Idea 1) is paired with three comparison ideas drawn from the top, middle, and bottom distance bins. Both judges (GPT-4.1, Claude Haiku 4.5) assign the labels shown, and they agree in both AB and BA orderings (12/12). Bold spans are shared verbatim between the reference and Idea 4: the mode-collapse signature that the bottom distance bin surfaces.

Role	d	**Labels**	**Idea**
Reference (Idea 1)	—	—	Evolutionary algorithms (EAs) treat all fitness landscapes uniformly, limiting performance on heterogeneous search spaces. Apply **persistent homology** to dynamically characterize **fitness landscape topology** during evolution; use real-time **topological** signatures to modulate **operator** probabilities, favoring crossover in funnel-shaped regions and mutation in rugged multimodal areas. Benchmark against standard EA variants (DE, GA, CMA-ES) on the BBOB test suite and real-world engineering problems, measuring **convergence speed**, **solution quality**, and **robustness across diverse landscape types**.
Top (Idea 2)	0.591	A / A	Automatically discover unknown conservation laws from observational data of dynamical systems. Co-evolve symbolic expressions (via genetic programming) representing candidate conserved quantities alongside numerical evaluators assessing time-invariance across simulated trajectories; a multi-objective fitness rewards conservation accuracy, generalization across diverse initial conditions, and symbolic parsimony. Validate on classical-mechanics benchmarks (pendulum, Kepler orbits, N-body) against SINDy and AI-Feynman baselines under increasing observational noise.
Middle (Idea 3)	0.355	B / B	Evolutionary algorithms suffer premature convergence on multimodal landscapes. Adopt epigenetic-inspired dual-layer encoding: each individual carries a genotype plus a heritable binary methylation mask that silences gene segments contextually; masks mutate and recombine independently, allowing phenotypic variation without altering underlying genetic material. Benchmark against standard EAs on CEC multimodal test suites and real-world hyperparameter-optimization tasks, measuring convergence speed, solution quality, and population-diversity metrics.
Bottom (Idea 4)	0.093	D / D	**Evolutionary** algorithm performance degrades when **operator**s are mismatched to **fitness landscape topology**. During evolution, **apply persistent homology** to current population samples to compute **topological** landscape features (Betti numbers, persistence diagrams), then dynamically select mutation and crossover operators matched to detected ruggedness, modality, and neutrality in real time. Benchmark against state-of-the-art adaptive EAs on CEC-2017 test suites and real-world protein-folding tasks, measuring **solution quality**, **convergence speed**, and **robustness across diverse landscape types**.

Table 15. Reference-grounded grid for the keyword *3D bioprinting*, `high` effort, generated by Claude Sonnet 4.6. Two independent within-keyword pairs are shown: a top-distance-bin pair (Ideas 5, 6; embedding distance 0.489; both judges A/A in both AB and BA orderings) and a bottom-distance-bin pair (Ideas 7, 8; embedding distance 0.068; both judges D/D in both AB and BA orderings). Bold spans in the bottom-distance-bin pair are shared verbatim between Ideas 7 and 8: the mode-collapse signature persists at `high` effort in the bottom distance bin.

Role	d	Labels	Idea
Reference (Idea 5)	—	—	Engineered cardiac tissue lacks the anisotropic electrical conduction of native myocardium, limiting functional integration post-transplantation. Embed graphene-oxide nanosheets into cardiac bioink and apply a weak DC electric field perpendicular to the print nozzle during extrusion: this simultaneously aligns graphene flakes and orients cardiomyocytes along the field axis as each layer deposits, creating directional conductivity matching native ventricular architecture. Assess cardiomyocyte alignment via α-actinin immunostaining, measure anisotropic conductance ratios using multielectrode arrays, and compare action-potential propagation velocities against native myocardium.
Top (Idea 6)	0.489	A / A	Address geometric inaccuracy in soft-tissue 3D bioprinting caused by post-print viscoelastic relaxation of hydrogel bioinks. Train a physics-informed neural network (PINN) on bioink rheological characterization data to generate pre-distorted CAD geometries; upon printing, viscoelastic relaxation compensates the distortion, yielding the intended shape without requiring new biomaterials. Validate by printing cartilage-mimicking constructs in alginate-gelatin bioinks, quantifying geometric fidelity via micro-CT across multiple bioink compositions and print resolutions.
Reference (Idea 7)	—	—	**3D bioprinting** struggles to replicate **native tissue anisotropy** critical for functional outcomes in **muscle and cartilage**. Integrate **miniaturized ultrasonic transducers** directly into the extrusion **nozzle** to generate **acoustic standing waves** within the extruded **bioink**, **aligning cells along programmable axes** immediately before photocrosslinking locks in orientation. Compare acoustically-aligned versus unaligned constructs via immunofluorescence (cytoskeletal orientation quantification), **tensile testing (mechanical anisotropy** ratios), and functional assays including myotube formation efficiency and **contractile force** generation.
Bottom (Idea 8)	0.068	D / D	**3D bioprinted tissue constructs** lack the anisotropic cellular alignment characteristic of **native tissues like skeletal muscle and cartilage**, severely limiting their mechanical function and biological integration. Integrate **miniaturized focused ultrasonic transducers** directly into the extrusion printhead to generate localized **acoustic standing waves** within the **bioink** during deposition; acoustic radiation forces passively **align** suspended **cells along programmable axes** in real time without chemical modification. Quantify alignment via confocal microscopy and F-actin immunostaining; measure **mechanical anisotropy** by **tensile testing**; assess functional output through **contractile force** measurements in muscle constructs versus unaligned controls.

Table 16. Pairwise classification cross-judge agreement, GPT-4.1 vs. Claude Haiku 4.5, pooling all within-keyword pairs in each model run. Quadratic Cohen's κ treats $A/B/C/D$ as ordered. Exact agreement rate is the fraction of pairs whose two judges output the identical $A/B/C/D$ label.

Model	n pairs	Quadratic κ	Exact agreement
Claude Sonnet 4.6	278,400	0.875	0.692
GPT-5.4	283,200	0.838	0.684
Gemini 3.1 Pro	212,400	0.850	0.670

Table 17. LiveIdeaBench full fluency-score aggregation per $(model, judge, effort)$ condition, $K = K_{\text{full}}$, default prompt. Score = $(10 \cdot n_A + 7 \cdot n_B + 4 \cdot n_C + 1 \cdot n_D)/n$ with all three distance bins (top / middle / bottom) pooled. n is the per-condition pair count. "Lift" is the `low` $\rightarrow$ `high` fluency-score change.

Model	Judge	n	low	medium	high	lift
Claude Sonnet 4.6	GPT-4.1	69,600	5.66	6.41	6.78	+1.12
	Claude Haiku 4.5	69,600	5.73	6.23	6.62	+0.89
GPT-5.4	GPT-4.1	70,800	5.65	5.82	6.04	+0.39
	Claude Haiku 4.5	70,800	5.65	5.72	5.88	+0.23
Gemini 3.1 Pro	GPT-4.1	70,800	5.98	6.24	6.52	+0.54
	Claude Haiku 4.5	70,800	6.00	6.16	6.36	+0.36

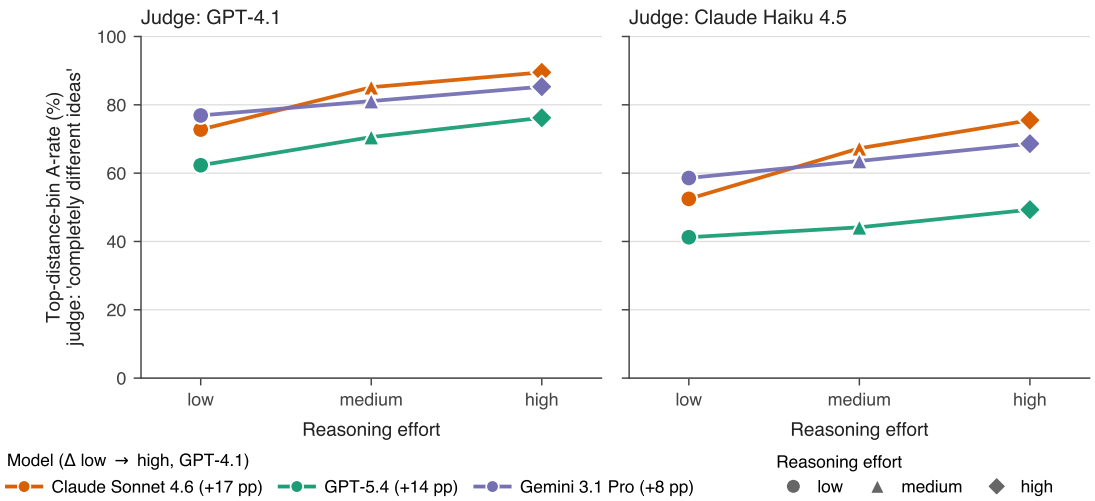

Figure 16. Top-distance-bin A-rate under both judge models (GPT-4.1, left; Claude Haiku 4.5, right) across the three models and the three reasoning-effort levels. The y-axis is the fraction of within-keyword pairs in the top distance bin that the judge labels as completely different ideas. The Δ values in the legend are `low` $\rightarrow$ `high` percentage-point increases under GPT-4.1.

Table 18. Per-idea cross-judge agreement on the per-idea quality rubric. Each row pools 106,200–141,537 aligned per-idea ratings under each judge. Pearson r and Spearman ρ are between the GPT-4.1 and Claude Haiku 4.5 per-idea scores. "Mean $|\Delta|$" is the average absolute per-idea disagreement on the 1–10 scale. GPT-4.1 mean and Claude Haiku 4.5 mean are the per-judge averages over the same paired sample.

Model	Quality axis	n ratings	Pearson r	Spearman ρ	Mean $\lvert\Delta\rvert$	GPT-4.1 mean	Claude Haiku 4.5 mean
Claude Sonnet 4.6	originality	139,107	0.376	0.373	2.48	8.74	6.26
	feasibility	139,107	0.488	0.448	1.26	6.33	5.16
	clarity	139,107	0.225	0.211	1.21	8.46	7.28
GPT-5.4	originality	141,537	0.259	0.255	2.19	8.89	6.70
	feasibility	141,537	0.392	0.379	1.30	6.73	5.49
	clarity	141,537	0.208	0.193	1.23	8.44	7.23
Gemini 3.1 Pro	originality	106,200	0.311	0.300	2.30	8.89	6.59
	feasibility	106,200	0.493	0.464	1.22	6.13	4.98
	clarity	106,200	0.248	0.231	1.36	8.62	7.27

Table 19. Condition-mean per-idea quality per (model, axis, judge model, reasoning-effort), with per-judge Δ(`low` $\rightarrow$ `high`). Columns under "GPT-4.1" are GPT-4.1 condition means at `low` / `medium` / `high` reasoning effort; columns under "Claude Haiku 4.5" are Claude Haiku 4.5 condition means at the same efforts. $\Delta_{\text{GPT-4.1}}$ and $\Delta_{\text{Haiku 4.5}}$ are the per-judge `low` $\rightarrow$ `high` change. All 18 Δ values are within ± 0.50 pt.

Model	Axis	GPT-4.1			Claude Haiku 4.5			$\Delta_{\text{GPT-4.1}}$	$\Delta_{\text{Haiku 4.5}}$
		low	medium	high	low	medium	high		
Claude Sonnet 4.6	originality	8.60	8.87	8.88	6.02	6.50	6.53	+0.28	+0.50
	feasibility	6.37	6.50	6.18	5.28	5.09	5.01	−0.19	−0.28
	clarity	8.50	8.51	8.35	7.26	7.30	7.28	−0.15	+0.03
GPT-5.4	originality	8.90	8.87	8.86	6.65	6.74	6.75	−0.04	+0.10
	feasibility	6.74	6.78	6.74	5.52	5.47	5.45	0.00	−0.07
	clarity	8.46	8.48	8.43	7.26	7.24	7.21	−0.03	−0.05
Gemini 3.1 Pro	originality	8.83	8.89	8.95	6.44	6.57	6.75	+0.12	+0.31
	feasibility	6.09	6.14	6.17	5.01	4.97	4.95	+0.08	−0.06
	clarity	8.60	8.62	8.63	7.22	7.27	7.32	+0.03	+0.10

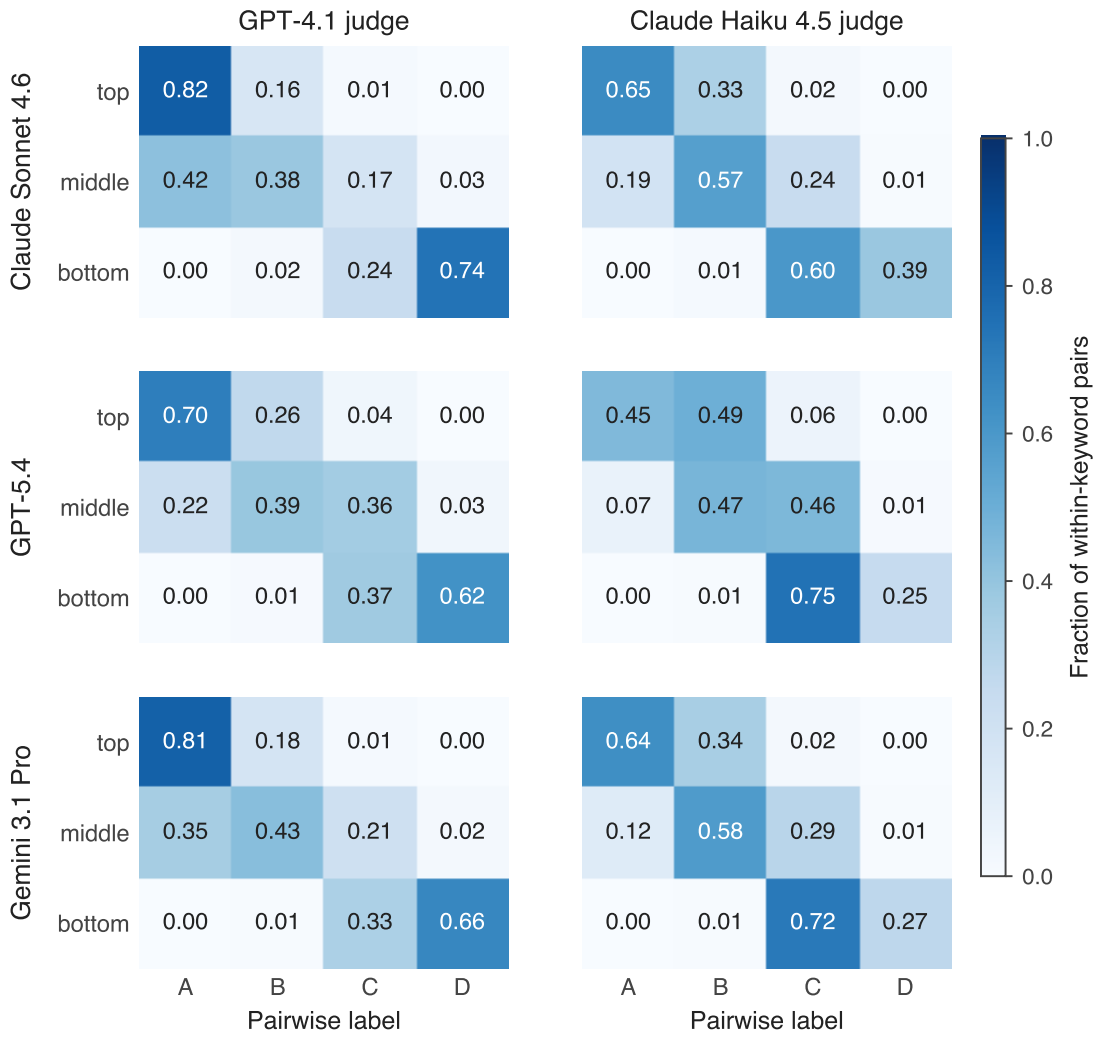

Figure 17. Per-(model, distance bin, judge model) distribution of pairwise classification labels. Rows: idea-generation model. Columns: judge model. Each 3×4 panel shows the fraction of within-keyword pairs in each of the three distance bins (top / middle / bottom) that was assigned each $A/B/C/D$ label, pooling across reasoning-effort levels. Both judge models show the same top-A, bottom-C/D pattern; the Claude Haiku 4.5 judge shifts mass toward the intermediate B and C labels in every panel.

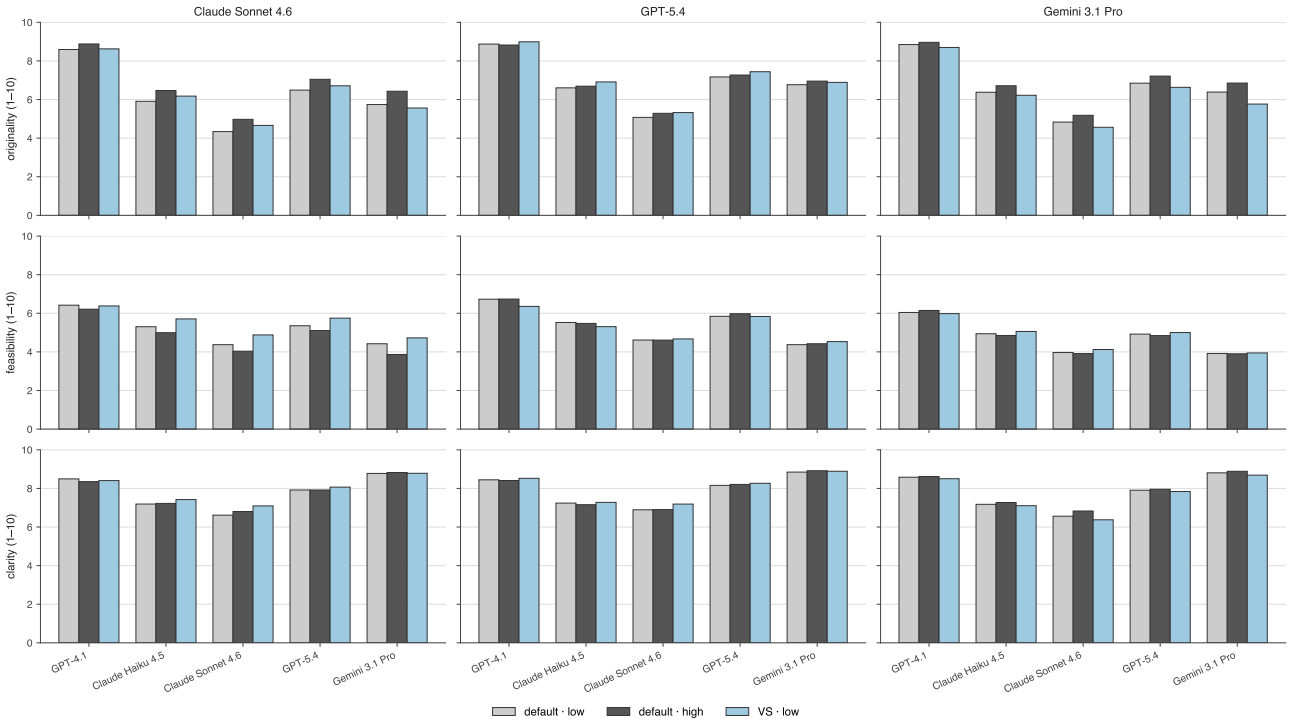

Figure 18. Per-condition mean quality ratings (1–10 scale) on the $Q_1 \cup Q_4$ keyword subset. Rows: idea-generation model. Columns: quality axis. Within each panel, bars $(\texttt{default}, \texttt{low})$ / $(\texttt{default}, \texttt{high})$ / $(\texttt{VS}, \texttt{low})$ are grouped under the five judges of Section A.5 (x-axis). Means track closely; the largest condition-mean differences, on Claude generator $\times$ originality, point upward rather than downward.

Table 20. Condition-mean trend deltas on the $Q_1 \cup Q_4$ subset under Claude Sonnet 4.6, GPT-5.4, and Gemini 3.1 Pro used as judges (columns). Each entry is $\Delta_{\texttt{high}-\texttt{low}}$ (effort axis, top block) or $\Delta_{\texttt{VS}-\texttt{default}}$ at `low` effort (prompt axis, bottom block).

Generator	Axis	Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro
Effort axis: $(\texttt{default}, \texttt{high}) - (\texttt{default}, \texttt{low})$				
Claude Sonnet 4.6	originality	+0.64	+0.56	+0.69
	feasibility	−0.34	−0.25	−0.56
	clarity	+0.19	0.00	+0.05
GPT-5.4	originality	+0.21	+0.10	+0.19
	feasibility	0.00	+0.13	+0.05
	clarity	+0.01	+0.05	+0.07
Gemini 3.1 Pro	originality	+0.35	+0.37	+0.47
	feasibility	−0.06	−0.08	−0.02
	clarity	+0.27	+0.05	+0.08
Prompt axis: $(\texttt{VS}, \texttt{low}) - (\texttt{default}, \texttt{low})$				
Claude Sonnet 4.6	originality	+0.32	+0.22	−0.18
	feasibility	+0.51	+0.39	+0.30
	clarity	+0.48	+0.15	+0.01
GPT-5.4	originality	+0.25	+0.27	+0.12
	feasibility	+0.06	−0.01	+0.16
	clarity	+0.30	+0.11	+0.04
Gemini 3.1 Pro	originality	−0.27	−0.21	−0.62
	feasibility	+0.15	+0.08	+0.03
	clarity	−0.19	−0.07	−0.12

Table 21. Per-idea Pearson r between each judge pair in $\{\text{Sonnet 4.6}, \text{GPT-5.4}, \text{Gemini 3.1 Pro}\} \times \{\text{GPT-4.1}, \text{Haiku 4.5}\}$, summarized across the 18 (condition $\times$ generator $\times$ quality axis) slices of each pair. $n \in [2{,}997, 3{,}000]$ paired ideas per slice.

Judge pair		**mean** r	**min** r	**max** r
Sonnet 4.6	GPT-4.1	0.336	0.159	0.535
Sonnet 4.6	Haiku 4.5	0.437	0.338	0.602
GPT-5.4	GPT-4.1	0.375	0.138	0.581
GPT-5.4	Haiku 4.5	0.404	0.196	0.633
Gemini 3.1 Pro	GPT-4.1	0.312	0.120	0.510
Gemini 3.1 Pro	Haiku 4.5	0.321	0.143	0.512

Table 22. Diagnostic (iii) per condition: keyword-mean Spearman ρ between within-keyword pairwise seed distance and idea distance, the corresponding permutation-null mean ($\bar{\rho}_{\text{null}}$), and the count of keywords whose per-keyword permutation p-value falls below 0.05. The shuffle null breaks the seed–idea correspondence within each keyword while preserving marginal distributions (200 permutations).

Model	**Effort**	$\bar{\rho}$	$\bar{\rho}_{\text{null}}$	**Keywords with** $p < 0.05$
Claude Sonnet 4.6	`low`	$+0.026$	$+0.000$	$6/100$
	`high`	$+0.002$	$+0.001$	$7/100$
GPT-5.4	`low`	$+0.002$	-0.001	$7/99$
	`high`	$+0.004$	$+0.000$	$5/99$
Gemini 3.1 Pro	`low`	$+0.019$	$+0.000$	$5/100$
	`high`	-0.016	$+0.000$	$5/100$

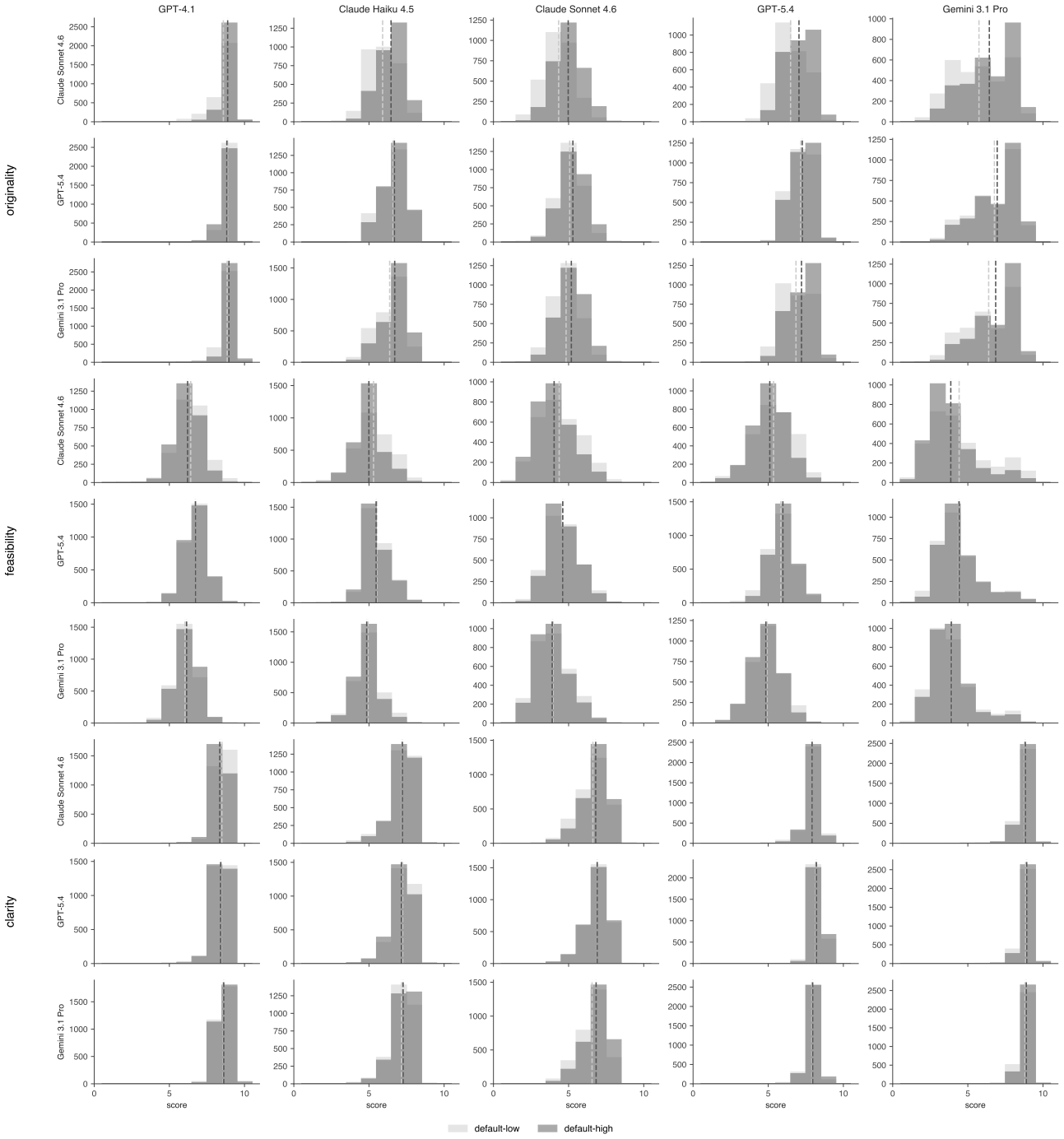

Figure 19. Per-idea quality score distributions for the effort-axis pair (`default`, `low`) vs (`default`, `high`) on the $Q_1 \cup Q_4$ subset. Rows (outer): quality axis (originality / feasibility / clarity); rows (inner per axis): Claude Sonnet 4.6, GPT-5.4, Gemini 3.1 Pro. Columns: five judges in the same left-to-right order as Figure 18. Dashed vertical lines mark the per-condition mean.

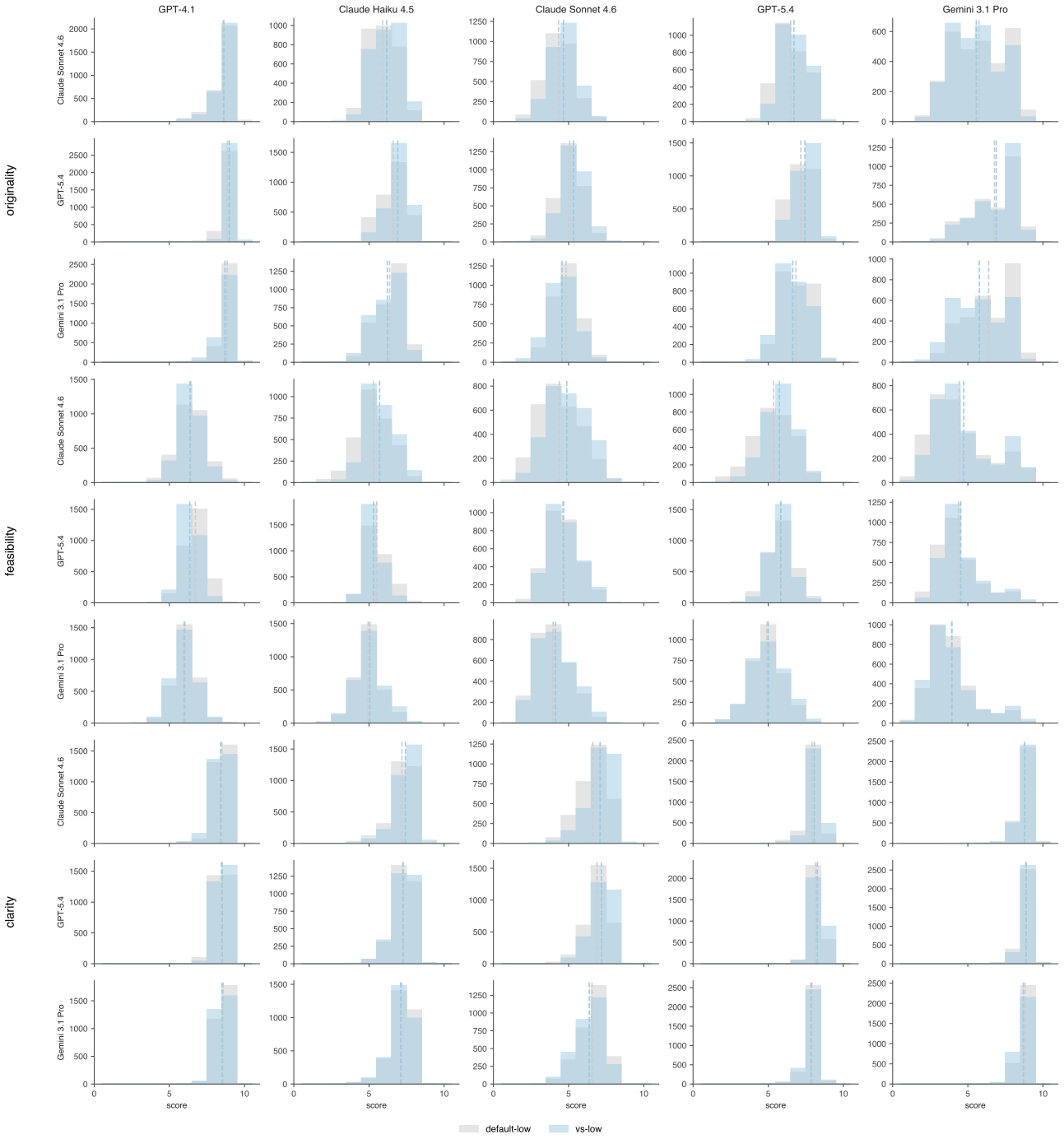

Figure 20. Per-idea quality score distributions for the prompt-axis pair (`default`, `low`) vs (`VS`, `low`) on the $Q_1 \cup Q_4$ subset (reasoning effort held at `low` on both sides). Layout and judges match Figure 19.

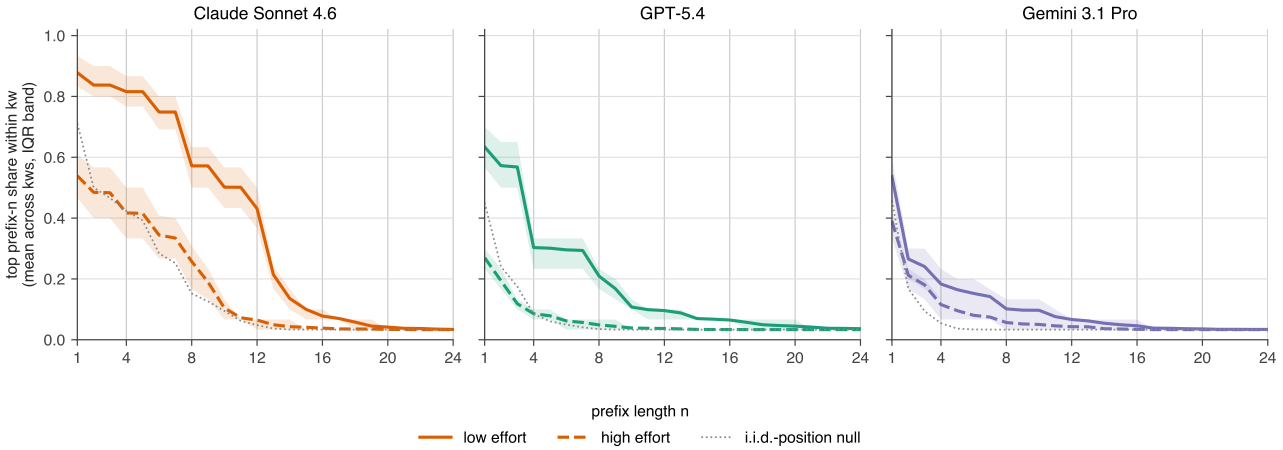

Figure 21. SSoT `random_string` prefix-sharing curves per (*model*, *effort*) condition. The y-axis is the share of within-keyword seeds that begin with the most common n-character prefix (mean across keywords; shaded band: $[25\%, 75\%]$ across keywords). The dotted grey curve is the i.i.d.-position null (per-position character draws preserving each condition's empirical alphabet at that position; 500 simulations) which captures alphabet bias without across-position correlations. Observed curves stay far above the null at $n \geq 4$ in every `low`-effort condition and in (Claude Sonnet 4.6, `high`), indicating that the seeds within a keyword share a long common prefix (*template anchoring*).