Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Disrupting convergence in large language model generation

C. Blaine Horton

cbhorton@albany.edu

University at Albany, State University of New York

Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11354680/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Disrupting convergence in large language model generation

C. Blaine Horton, Jr.*

Department of Management, Massry School of Business, University at Albany, State University of New York, Albany, NY, USA

*Correspondence: cbhorton@albany.edu

Disrupting convergence in large language model generation

Abstract

Efforts to reduce homogenization in large language model (LLM) outputs increasingly use prompting interventions, but these interventions often change both the information supplied to models and the instructions governing its use. We experimentally separate this behavioral guidance from a second potential source of variation: the information available during generation. Across 16,200 generations spanning three preregistered studies and two cross-model robustness replications, supplying heterogeneous external cues substantially increased semantic diversity and generally did so more reliably than explicit instructions to produce different responses. Crucially, this effect did not increase monotonically with stronger instructions to use the information. Merely presenting varied cues produced substantial diversification, whereas instructing models to disregard the same information largely eliminated the effect. When cue identities and cue-use instructions were held constant, differentiation instructions increased collective diversity relative to the other experimental conditions and increased departure from models' default response spaces, whereas convergence instructions reduced such departure. These findings distinguish two sources of variation often conflated in prompt design: informational inputs can supply variation, while accompanying instructions can suppress their influence or further shape the resulting response distribution. Preserving diversity may therefore require designing information environments, not simply better prompts.

Large language models (LLMs) are being used to support brainstorming, problem solving, and concept development [1, 2]. Yet these benefits may come with a collective cost. Although LLMs are capable of generating an enormous range of possible responses, growing evidence from embedding-based measures, human evaluations, and analyses of linguistic variation suggests that AI-generated and AI-assisted outputs can become more similar across users, reducing collective diversity [3–8]. As generative systems become increasingly embedded in tasks for which variation itself is valuable, the practical question is therefore no longer only whether convergence occurs, but how diversity across repeated generations can be encouraged or preserved.

Although convergence can be valuable when users seek reliable or standardized responses, it presents a different problem when variation itself is the goal. Creative ideation, innovation, design, and exploratory problem-solving benefit from considering alternatives rather than repeatedly arriving at the same ones [9–11]. In these settings, individually useful assistance can coexist with collective homogenization: each user may receive a reasonable response even while similar responses across users narrow the final population of ideas being generated [4, 12]. An important challenge, then, is not simply to obtain a useful response from an individual model, but to broaden the distribution of responses produced across repeated generations. What kinds of interventions can reliably do so?

Nascent work has begun to address this challenge by seeking prompting and generation strategies that can increase output diversity [13, 14]. Current findings establish that convergence is not inevitable. Yet many successful interventions change more than one feature of the generation context at once. Interventions that assign different personas or functional constraints, for example, can both supply models with heterogeneous information and give that information different roles in shaping a response. Consequently, it remains unclear whether diversification arises from varying the information available to the model, from instructions that direct the model to use that information, or from their combination.

One route to diversification is to change how a model is instructed to respond. The most direct strategy might be simply to ask the model to produce something different. Such differentiation instructions are appealing because they are inexpensive, scalable, and require no information beyond the original task: users can ask a model to avoid obvious responses, generate something unusual, or produce an answer unlike those it would typically provide. Yet increasing collective diversity requires more than eliciting an unusual response on occasion. Repeated independent generations must spread across different possibilities rather than simply shifting together toward another shared response.

The difficulty of doing so becomes especially apparent when repeated generations are strongly concentrated. We use the term *response basin* descriptively to refer to a relatively narrow set of responses, or region of semantic space, that repeated generations disproportionately occupy for a

given task. A modest exploratory pilot illustrates how pronounced such concentration can be. We constructed 50 different formulations of the same simple question—asking a model to name an object that could be placed next to a red chair—and obtained 20 independent generations from GPT-5.6-terra for each formulation. Across the resulting 1,000 generations, “table” accounted for 89.9% of all responses and was the modal response for 47 of the 50 prompt formulations (see Supplementary Information). For 31 of the 50 formulations, all 20 generations produced exactly the same response. The result illustrates an important distinction between generative capacity and realized variation: a model may be capable of producing many plausible responses while repeated generations nevertheless concentrate on a much narrower subset.

An alternative route is therefore to introduce variation into the generation context itself. Rather than giving every generation the same task and asking for unique output, one can hold the task constant while varying the external information supplied alongside it. Different generations then receive different informational inputs, potentially shifting the distribution of responses they produce even when the underlying task remains unchanged. In essence, the relevant question is not whether a model is capable of producing alternatives, but whether, and to what degree, varying those inputs systematically broadens the distribution of alternatives it actually produces.

And yet, even if varied external information has the potential to broaden the distribution of responses, it remains unclear what determines whether (and how strongly) that information shapes generation. The presence of a cue and the instructions governing its use are often bundled within the same intervention, such that it is not clear whether they contribute in the same way. A cue may influence generation simply by being available, while explicit guidance may strengthen, redirect, or leave that influence largely unchanged. Moreover, the direction of guidance may matter in its own right. Instructions to use information and instructions to disregard it need not operate as simple opposites. These possibilities are difficult to distinguish when interventions simultaneously change both the information models receive and the instructions governing that information. Answering these questions therefore requires holding informational content constant while independently varying how models are instructed to treat it, allowing us to separate the availability of information from instructional control and examine whether and how each contributes to diversification.

The present research investigates these two potential sources of diversification across three preregistered studies, multiple language models, and three distinct ideation tasks. Study 1 compares an instruction-only differentiation strategy with varied external cues, testing whether introducing heterogeneous information broadens collective semantic diversity beyond both baseline generation and an explicit request to differentiate. Study 2 then examines whether external information must be behaviorally relevant to exert this influence, comparing cues presented without guidance to otherwise identical cues accompanied by instructions that encourage, require, or discourage their use. Study 3 isolates the role of behavioral guidance more directly by holding the external information constant while varying the direction of the

accompanying instruction, testing whether identical information produces different levels of diversity when models are encouraged to differentiate, given neutral guidance, or encouraged to converge. Together, the studies separate the diversification produced by varying the information available to a model from the behavioral conditions that determine how strongly that information shapes generation.

Study 1 Results

External cues substantially increased semantic diversity relative to baseline prompting

To examine whether introducing external information increases the diversity of language model outputs, we compared responses generated under three conditions: a baseline prompt, a differentiation request instructing the model to produce a response unlike those generated for other users, and an external-cue condition in which a different externally supplied cue word accompanied each generation. Responses were generated independently by Claude Sonnet 5 and GPT-5.6-terra across three ideation tasks: naming an object to place next to a red chair, generating an alternative use for a cardboard box, and proposing an improvement to a reusable water bottle. Semantic diversity was quantified as the mean pairwise cosine distance among response embeddings within each model × task × condition cell, with higher values indicating that responses occupied a broader region of semantic space.

Across all six model × task combinations, externally supplied cues substantially increased semantic diversity relative to the baseline prompt (Fig. 1). Mean pairwise semantic distance increased from baseline values ranging from 0.013 to 0.363 to values between 0.584 and 0.762 following the introduction of external cues. Bootstrap confidence intervals excluded zero for every comparison, indicating a consistent increase in semantic diversity across models and tasks.

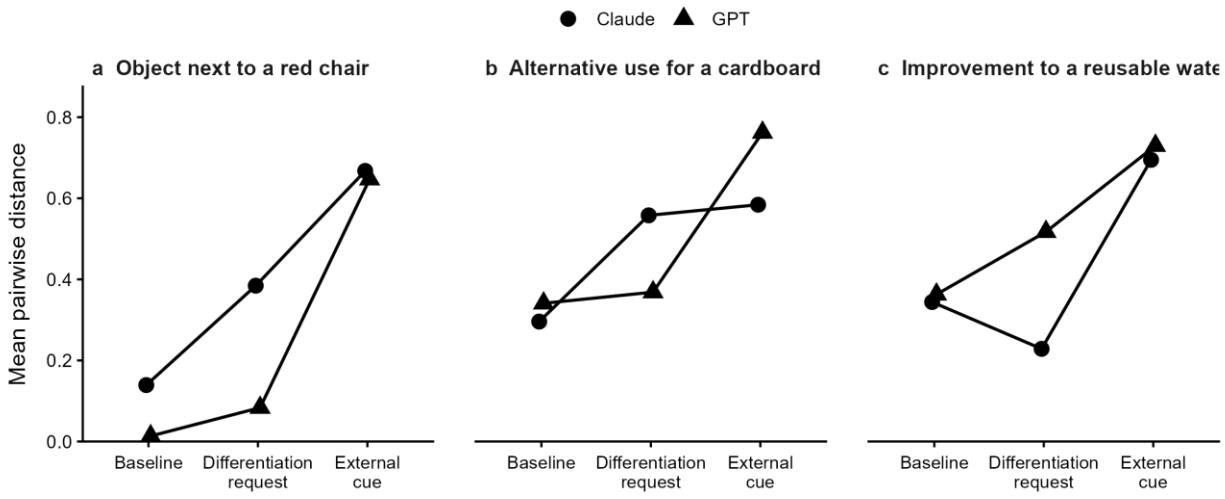


Figure 1. External cues increased semantic diversity more consistently than differentiation

requests. Mean pairwise cosine distance among responses generated by Claude Sonnet 5 and GPT-5.6-terra across three ideation tasks. Each point shows the mean pairwise cosine distance among 100 independently generated responses within one model × task × condition cell. Lines connect conditions within model and task for visualization; higher values indicate greater semantic diversity.

Asking language models to differentiate produced inconsistent improvements

In contrast, instructing language models to produce responses that differed from those generated for others produced less consistent benefits. Differentiation requests increased semantic diversity in four of the six model × task combinations but failed to do so reliably in the remaining two. The weakest performance occurred for Claude, where differentiation guidance on one task unexpectedly *reduced* semantic diversity relative to baseline. Together, these findings suggest that modifying a model's instructions alone may be a less reliable means of increasing semantic diversity than supplying external information during generation.

External cues generally outperformed differentiation requests

Direct comparisons between the two intervention strategies highlighted the advantage of supplying external information. Responses generated with external cues exhibited greater semantic diversity than responses generated under differentiation instructions in five of the six model × task combinations. The only exception was the Claude cardboard-box task, where the two conditions produced comparable levels of diversity. Across the remaining tasks and models, externally supplied cues consistently produced broader semantic dispersion than simply requesting different responses. Collectively, these findings indicate that an external-cue intervention produced more reliable increases in semantic diversity than prompting language models to differentiate from their expected responses.

Cue distance was associated with response displacement

We next examined whether an external cue’s semantic distance from a model’s baseline-response centroid predicted the resulting response’s distance from that same centroid. Cue-to-baseline distance positively predicted response-to-baseline distance (b = 0.336, 95% CI [0.098, 0.573], p = .006), such that cues farther from the centroid tended to elicit responses that were also farther from it. This association remained significant when standard errors were clustered by cue identity to account for the reuse of cues across model × task combinations (95% CI [0.064, 0.608], p = .016). Because cue distance was observed rather than experimentally manipulated, this analysis demonstrates an association between cue position and response displacement but does not establish a causal effect. A second preregistered analysis examining cue distinctiveness is reported in the Supplementary Information.

Robustness across language models

To further examine the generality of these findings, we repeated the study using Gemini 3.6 Flash. The principal pattern replicated. External cues again produced substantially greater semantic diversity than baseline prompting across all three ideation tasks.

Study 2 Results

Instructions to disregard external information largely eliminated its diversity advantage.

Study 2 examined whether the effect of external information depends on how language models are instructed to treat that information. The study included one baseline condition without an external cue and four cue-present conditions that differed only in how the model was instructed to treat the same type of cue. Models were either shown the cue without any instruction regarding its use (*cue only*), told they could use the cue if they wished (*optional use*), instructed to incorporate the cue (*required use*), or instructed not to use it (*disregard*). Responses were generated independently by Claude Sonnet 5 and GPT-5.6-terra across the same three ideation tasks used in Study 1.

The preregistered primary analysis tested whether the semantic diversity of output increased monotonically as prompts increasingly encouraged incorporation of the cue, from *cue only* to *optional use* to *required use*. This prediction was not supported ($\Delta = -0.002$, 95% bootstrap CI [−0.008, 0.004], one-sided $p = .770$; Table 1). Instead, the results revealed a pronounced non-monotonic pattern (Fig. 2). Simply presenting an external cue without any instruction regarding its use substantially increased semantic diversity relative to baseline generation. Averaged equally across the six model × task combinations, the cue-only condition increased mean pairwise semantic distance by 0.415 (95% bootstrap CI [0.398, 0.428], one-sided $p < .001$). This increase appeared across every model × task combination.

In contrast, explicitly instructing models to *disregard* the same external cue largely eliminated this diversity advantage. Relative to the *cue-only* condition, *disregard* instructions reduced mean pairwise semantic distance by 0.402 (95% bootstrap CI [−0.412, −0.388], one-sided $p < .001$). Semantic diversity in the *disregard* condition was correspondingly close to baseline levels ($\Delta = 0.013$, 95% bootstrap CI [−0.001, 0.028], one-sided $p = .039$).

In sum, external information substantially diversified model outputs even when models received no instruction to use it. Explicitly requiring cue incorporation produced a similarly strong diversity advantage, whereas instructing models to disregard the same information largely eliminated the effect. Behavioral guidance therefore exerted an asymmetric influence: encouragement was not necessary for external information to diversify generation, but instructions could nevertheless attenuate or suppress that influence.

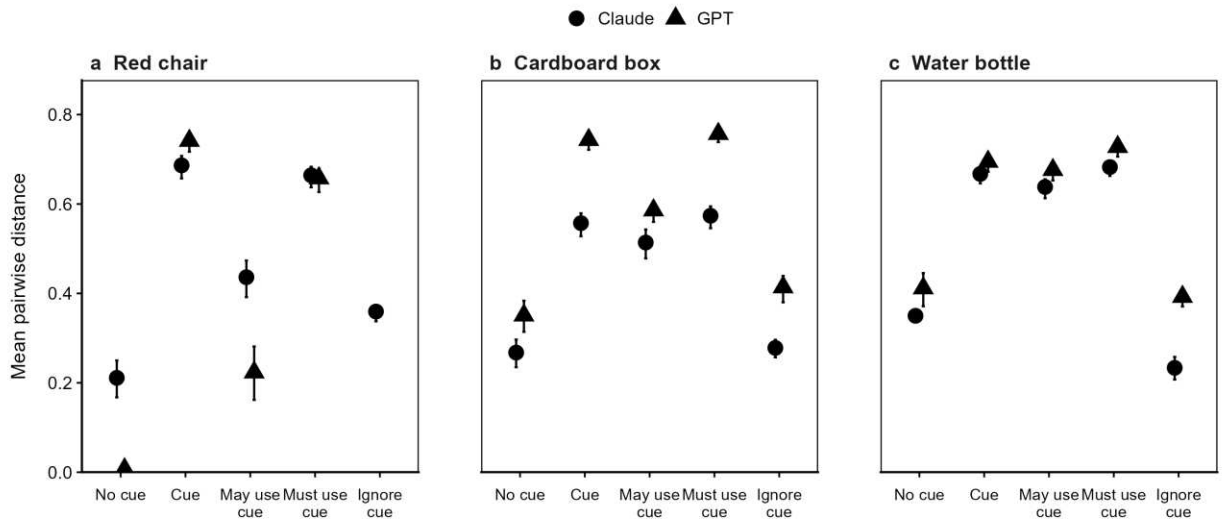


Figure 2 illustrates the semantic diversity observed across all five prompting conditions. Semantic diversity across prompting conditions in Study 2. Mean pairwise cosine distance across baseline, cue-only, optional-use, required-use, and disregard conditions for each model × task combination. Higher values indicate greater semantic diversity.

Table 1. Preregistered aggregate bootstrap tests of semantic diversity in Study 2.

Preregistered test	**Estimate**	**95% bootstrap CI**	**One-sided bootstrap p**
Linear trend across cue-only, optional-use, and required-use conditions	-0.002	[-0.008, 0.004]	.770
Cue − No cue	0.415	[0.398, 0.428]	< .001
May use cue − Cue	-0.169	[-0.186, -0.151]	1.000
Must use cue − May use cue	0.165	[0.147, 0.181]	< .001
Ignore cue − Cue	-0.402	[-0.412, -0.388]	< .001
Ignore cue − No cue	0.013	[-0.001, 0.028]	.039

Note. Estimates are differences in mean pairwise cosine distance. Effects were calculated within each of the six model × task strata and then averaged with equal weight. Confidence intervals are two-sided 95% percentile-bootstrap intervals; p values are one-sided because the preregistered predictions

were directional. Inference was based on 10,000 nonparametric bootstrap resamples within each cell. The primary linear contrast used equally spaced coding of the cue-only, optional-use, and required-use conditions and is equivalent to (Must use cue − Cue) / 2.

Encouraging cue use did not produce a monotonic increase in diversity

Framing cue use as optional reduced semantic diversity relative to cue-only presentation ($\Delta = -0.169$, 95% bootstrap CI [−0.186, −0.151]). Requiring cue use then increased semantic diversity relative to the optional-use condition ($\Delta = 0.165$, 95% bootstrap CI [0.147, 0.181], one-sided $p < .001$), restoring diversity to approximately the level observed when the cue was simply presented without guidance.

Accordingly, stronger encouragement to use the cue did not progressively increase semantic diversity. Instead, behavioral guidance produced a non-monotonic pattern: cue-only and required-use conditions yielded similarly high diversity, whereas framing cue use as merely optional substantially attenuated the effect. Combined with the near-elimination of the diversity advantage under disregard instructions, these results reveal an asymmetric role for behavioral guidance. Models did not need to be encouraged to use external information for it to shape their outputs, but the way that information was framed could weaken—and, when explicitly discouraged, nearly eliminate—its diversifying influence.

A preregistered cue-source robustness analysis yielded the same condition pattern for cues reused from Study 1 and newly sampled cues, with no evidence that cue source moderated any preregistered condition effect (all two-sided bootstrap $ps \geq .369$).

Robustness across language models

We repeated Study 2 using Gemini 3.6 Flash while holding the prompting procedure, external cues, and analytic pipeline constant. The principal pattern replicated across all three ideation tasks. Presenting an external cue substantially increased semantic diversity relative to baseline, whereas instructing the model to disregard that same cue largely eliminated the effect. Optional cue use again produced lower semantic diversity than cue-only presentation, and requiring cue incorporation restored diversity to levels comparable to the cue-only condition. The preregistered ordered trend was supported in the Gemini replication (estimate = 0.024, 95% bootstrap CI [0.015, 0.032], one-sided $p < .001$), although the overall pattern of condition means closely resembled that observed for Claude Sonnet 5 and GPT-5.6-terra.

Study 3 Results

Identical external information produced different levels of diversity under different response instructions

Study 3 examined whether different response instructions produce different patterns of generation when externally supplied information and its prescribed use are held constant. Responses were generated independently by Claude Sonnet 5, GPT-5.6-terra, and Gemini 3.6 Flash across the same three ideation tasks used in the preceding studies. The experiment included four cue-present conditions. The same set of 100 externally supplied cue words appeared once within every model × task × condition combination, and all four conditions included the same instruction to use the external cue as inspiration. The *cue-use-only* condition added no further instruction. In the remaining conditions, models were additionally instructed to produce responses that either differed from those likely to be generated for others (*differentiation*), were clear and easy to understand (*neutral*), or were similar to those likely to be generated for others (*convergence*). Separate no-cue reference responses were used only to estimate each model's default response space and were not included in the experimental condition comparisons.

The preregistered primary analysis compared mean pairwise semantic distance across the four experimental conditions using a stratified bootstrap that averaged condition contrasts equally across the nine model × task combinations. Because the task, cue identities, cue-use instruction, and response-format instructions were held constant across conditions, differences among conditions reflect the additional response instructions rather than differences in the external information provided.

Differentiation instructions produced greater semantic diversity than each of the other cue-present conditions. Relative to cue-use only, differentiation instructions increased semantic diversity by $\Delta = 0.046$ (95% bootstrap CI [0.033, 0.057], directional $p < .001$). Differentiation instructions likewise exceeded neutral instructions ($\Delta = 0.053$, 95% bootstrap CI [0.041, 0.064], directional $p < .001$) and convergence instructions ($\Delta = 0.074$, 95% bootstrap CI [0.061, 0.086], directional $p < .001$). Across the nine model × task combinations, differentiation instructions consistently produced the greatest or nearly greatest semantic diversity. Thus, holding external information and its prescribed use constant, instructing models to differentiate produced broader collective exploration of semantic space than providing no additional response instruction, a neutral instruction, or a convergence instruction.

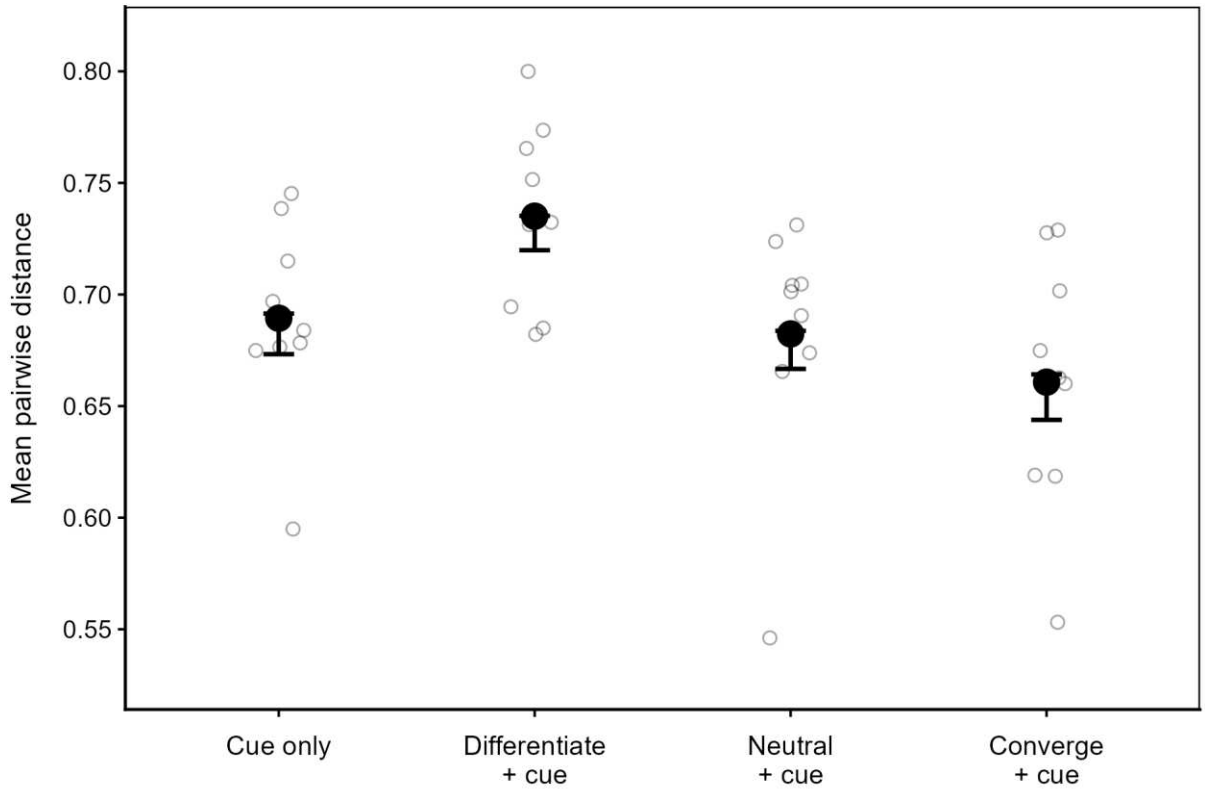


Figure 3. **Response instructions altered semantic diversity under otherwise identical conditions.** Open circles show mean pairwise cosine distance within each of the nine model × task combinations. Filled points show the equally weighted aggregate across strata; error bars show 95% percentile-bootstrap confidence intervals. All four experimental conditions used the same externally supplied cue words and the same cue-use instruction, differing only in the additional response instruction. Higher values indicate greater collective semantic diversity.

Response instructions also altered displacement from each model's default response space

We next examined whether response instructions changed how far responses departed from each model's independently estimated default response space. A separate reference centroid was calculated for each of the nine model × task combinations using 100 independently generated no-cue responses per combination (900 reference responses in total). Each experimental response was then assigned its cosine distance from the corresponding model × task reference centroid.

The preregistered regression analysis showed that, relative to cue-use only, differentiation instructions increased distance from the default response space ($\beta = 0.06895$, $p < .001$), whereas convergence instructions reduced distance from the default response space ($\beta = -0.04587$, $p < .001$). Neutral instructions did not differ from cue-use only ($\beta = -0.00070$, $p = .884$). Preregistered bootstrap contrasts yielded the same pattern. Differentiation instructions increased distance from the default response space relative to cue-use only by $\Delta = 0.069$ (95% bootstrap CI [0.057, 0.081], $p < .001$), exceeded neutral instructions by $\Delta = 0.070$ (95% bootstrap CI [0.057,

0.082], $p < .001$), and exceeded convergence instructions by $\Delta = 0.115$ (95% bootstrap CI [0.099, 0.130], $p < .001$).

Together with the semantic-diversity analyses, these findings show that response instructions affected not only how broadly responses were distributed, but also how far responses departed from each model's default response space, even though the external cues and instructions to use them were identical across conditions.

A preregistered secondary analysis showed that cue-to-baseline distance positively predicted response displacement in the cue-use-only reference condition ($b = .053$, $p < .001$), and this relationship did not differ significantly under differentiation, neutral, or convergence instructions (all ps ≥ .15; see Supplementary Information). Thus, cues farther from a model's default response space tended to produce more displaced responses, but this association did not differ reliably across response-instruction conditions.

Study 1 Methods

Design and preregistration

Study 1 was preregistered (AsPredicted #304,474) and examined whether supplying varied external information increases the collective semantic diversity of repeated language-model generations more effectively than either an unmodified prompt or an explicit request to generate a response different from those likely to be produced for other users. The preregistered design crossed three prompting conditions, three ideation tasks, and two language models. We collected 100 successful generations within each condition × task × model combination, yielding 1,800 responses.

The primary preregistered predictions were that external cues would increase semantic diversity relative to baseline prompting and would produce greater diversity than explicitly asking models to differentiate their responses. A secondary comparison examined whether differentiation requests increased diversity relative to baseline. Two additional preregistered analyses examined whether semantic properties of the external cues predicted the distinctiveness and displacement of the responses they elicited.

Ideation tasks

Models completed three ideation tasks selected to provide variation in the context and structure of generation. In an object-placement task, models were instructed: "Name one object that could be placed next to a red chair." In an alternative-use task, adapted from the widely used psychological measure of divergent thinking [15, 16], models were instructed: "Name one use for an empty cardboard box other than storing or transporting objects." In a product-

improvement task, models were instructed: “Suggest one improvement to a reusable water bottle.”

Response-format instructions were designed to preserve the natural response form of each task while limiting variation in verbosity and explanation unrelated to the idea being generated. The object-placement task therefore required only the name of the object, without explanation, whereas the alternative-use and product-improvement tasks permitted brief phrases of no more than 12 words. These constraints limited extraneous variation in response length and format while allowing sufficient language to express ideas in the two more open-ended tasks.

Prompting conditions

The three conditions differed in whether the baseline prompt was supplemented with an instruction to differentiate the response or with externally supplied information. In the *baseline condition*, models received only the task and response-format instructions. In the *differentiation-request condition*, the baseline prompt was supplemented with: “Your answer should be different from what other people are likely to receive.” In the *external-cue condition*, it was supplemented with: “Use the following word as inspiration when answering: ‘[cue]’."

The external-cue condition used a prespecified set of 100 cue words assembled before data collection using random-word generators. The set was intentionally heterogeneous, including multiple parts of speech and words varying in length and complexity. These different cues allowed us to test the effect of supplying varied external information rather than the effect of any particular word. The same 100 cues were used across both models and all three tasks, with each cue appearing exactly once within every model × task external-cue cell.

Language models and generation procedure

Responses were generated independently using OpenAI GPT-5.6-terra and Anthropic Claude Sonnet 5, with reasoning and thinking disabled, respectively. These settings were used to focus the study on variation arising directly from the experimental prompts, without additional model-generated deliberation preceding the response. Because the study examined repeated generation under tightly controlled prompting conditions, disabling these processes provided a more controlled test of whether external information altered the distribution of responses by limiting an additional source of variation across generations. This choice therefore prioritized experimental control over coverage of all available inference settings. Each generation used a fresh API context, with no conversational history carried across responses. The common system instruction was: “Follow the user's instruction exactly. Return only the requested answer with no explanation.”

Semantic embeddings and primary outcome

Responses were embedded using `sentence-transformers/all-mpnet-base-v2`, as specified in the preregistration. Response embeddings were normalized to unit length prior to computing semantic distances. Cue words were embedded using the same model.

The primary dependent variable was **mean pairwise semantic distance**, which captured how semantically similar or different repeated responses to the same prompt were from one another. For each combination of model, task, and prompting condition, we calculated the cosine distance between each of the 100 responses and every other response generated under that same combination, and then averaged these distances. Cosine distance was defined as one minus cosine similarity. Thus, higher values indicate that responses generated under the same model, task, and prompting condition were, on average, more semantically distinct from one another.

Confirmatory statistical analyses

To determine whether each prediction held across different models and tasks, we tested it separately for each of the six model–task pairings (two models × three tasks), rather than pooling responses across them. Within each pairing, we compared (1) differentiation with baseline, (2) external cues with baseline, and (3) external cues with differentiation. Our measure of semantic diversity yielded a single value for each set of 100 responses. We therefore used bootstrap resampling to estimate uncertainty in the difference between conditions. For each comparison, we resampled the responses within both conditions, recalculated their mean pairwise semantic distances, and recorded the difference. This procedure was repeated 10,000 times. The resulting distributions were used to construct 95% percentile-bootstrap confidence intervals. For each directional prediction, the *p* value was calculated as the proportion of resampled differences that were zero or opposite to the predicted direction, with a +1 correction.

Preregistered cue-level analyses

Two secondary preregistered analyses examined whether differences among the external cues predicted the responses they elicited. First, we examined whether more semantically distinctive cues produced more distinctive responses. Cue distinctiveness was defined as each cue's mean cosine distance from all other cues in the fixed cue set. Response distinctiveness was defined as each externally cued response's mean cosine distance from all other responses in the corresponding model × task external-cue condition.

Second, we examined whether cues located farther from a model's baseline responses produced greater response displacement. A baseline-response centroid was constructed separately for each model × task combination by averaging and normalizing the corresponding baseline-response embeddings. Cue-to-baseline distance and response-to-baseline distance were then calculated relative to this centroid. For each analysis, the response-level measure was regressed on the corresponding cue-level measure using ordinary least squares with task and model fixed effects.

Gemini robustness replication

To examine whether the primary pattern generalized beyond the two preregistered models, we repeated Study 1 using Gemini 3.6 Flash. The same tasks, prompting conditions, external cues, generation counts, semantic-diversity measure, and analysis procedure were retained, yielding 900 additional responses.

Study 2 Methods

Design and preregistration

Study 2 was preregistered before data collection (AsPredicted #304,614). The study examined whether the effect of externally supplied information on semantic diversity depends on how language models are instructed to treat that information. Whereas Study 1 compared external cues with prompting alone, Study 2 held the externally supplied information constant across cue-present conditions while varying the instructions governing its use.

The design crossed five prompting conditions, three ideation tasks, and two language models. We collected 200 successful generations within each condition × task × model combination, yielding 6,000 responses. The primary preregistered prediction was that semantic diversity would increase as prompts increasingly encouraged incorporation of an external cue. Additional preregistered comparisons tested whether simply presenting a cue increased diversity relative to baseline, whether progressively encouraging its use further increased diversity, and whether instructing models not to use the cue attenuated or eliminated the effect of cue presentation.

Prompting conditions

Study 2 included one baseline condition without an external cue and four conditions in which an external cue was presented. Within each task, the task and response-format instructions were held constant across conditions, and the same cue words appeared across all four cue-present conditions. These conditions differed in how the prompt instructed the model to treat the cue. In the *disregard* condition, models were told: "The following word is unrelated to the task. Do not use it when answering: '[cue]'." In the *cue only* condition, the cue was presented after the task, but without any additional instructions concerning its use: "Word: '[cue]'." In the *optional condition*, models were told: "You may use the following word as inspiration when answering, but you are not required to: '[cue]'." In the *required condition*, models were told: "Use the following word as inspiration when answering: '[cue]'." The baseline condition contained no external cue or cue-related instructions.

This design allowed us to begin distinguishing the effect of presenting external information from the effect of instructions governing its use. Comparisons among the four cue-present conditions held the external cue constant while varying whether its use was discouraged, left unspecified, permitted, or required.

Language models, tasks, and generation procedure

Responses were generated independently using OpenAI GPT-5.6-terra and Anthropic Claude Sonnet 5 under the same model settings and generation procedure used in Study 1. Models completed the same three ideation tasks: naming an object that could be placed next to a red chair, generating an alternative use for an empty cardboard box, and proposing an improvement to a reusable water bottle. The same task-specific response-format restrictions were employed. Each generation used a fresh API context with no conversational history carried across responses.

External cues

The four cue-present conditions used a fixed set of 200 external cue words. This set comprised the 100 cues used in Study 1 and 100 additional cues sampled before data collection using the same stimulus-construction procedure. The same 200 cues were used across both models, all three tasks, and all four cue-present conditions. Each cue appeared exactly once within every model × task × cue-present condition combination.

Consequently, differences among the cue-present conditions could not be attributed to different sets of external words appearing across conditions. Instead, the same external information was repeatedly paired with different instructions governing its use.

Semantic embeddings and primary outcome

Responses were embedded using `sentence-transformers/all-mpnet-base-v2` and analyzed using the same semantic-diversity procedure as in Study 1. The preregistered primary dependent variable was mean pairwise semantic distance, calculated separately within each model × task × condition combination. Higher values indicate that repeated generations within a condition were, on average, more semantically distinct from one another.

Confirmatory statistical analyses

The confirmatory analyses tested differences in semantic diversity between prompting conditions and used bootstrap resampling to estimate uncertainty around those differences. Unlike Study 1, where confirmatory comparisons were evaluated separately for each model and task, Study 2's preregistered confirmatory conclusions concerned effects averaged across the six model × task combinations. This approach tested whether prompting conditions produced systematic differences in semantic diversity across the experimental contexts rather than requiring the same effect to reach significance within every individual model × task combination.

Inference was based on 10,000 response-level bootstrap resamples. Within each bootstrap iteration, responses were resampled with replacement separately within each model × task × condition combination and mean pairwise semantic distance was recalculated. Condition contrasts were then calculated separately within each of the six model × task combinations and averaged with equal weight. The resulting bootstrap distributions were used to calculate 95% percentile-bootstrap confidence intervals and directional bootstrap *p* values for the preregistered directional predictions. Model- and task-specific results were additionally examined descriptively to characterize heterogeneity across contexts.

The preregistered primary test was an ordered contrast across the *cue-only*, *optional-use*, and *required-use* conditions, corresponding to increasing encouragement to incorporate the external cue. We additionally conducted five preregistered comparisons: cue only versus baseline, optional use versus cue only, required use versus optional use, disregard versus cue only, and disregard versus baseline. Together, these comparisons tested whether cue presentation alone increased semantic diversity, whether increasingly encouraging cue use further increased diversity, and whether explicitly discouraging cue use attenuated the effect of otherwise identical external information.

Cue-source robustness analysis

Because half of the external cues were reused from Study 1 and half were newly sampled for Study 2, the preregistration also specified a robustness analysis examining whether cue source affected the condition effects. We repeated the aggregate condition analyses separately for the

two cue sets and directly tested whether the corresponding condition effects differed between Study 1 and newly sampled cues.

Secondary and descriptive analyses

Additional descriptive analyses of exact-response uniqueness and cue–response semantic similarity are reported in the Supplementary Information; these analyses did not determine evaluation of the preregistered primary prediction.

Gemini robustness replication

To examine whether the primary pattern generalized beyond the two preregistered models, we repeated Study 2 using Gemini 3.6 Flash with the minimal thinking setting. The replication retained the same three ideation tasks, five prompting conditions, 200 external cues, and 200 independent generations per task × condition combination, yielding 3,000 additional responses. Gemini responses were analyzed using the same semantic-diversity measure and aggregate bootstrap procedure as the preregistered study.

Study 3 Methods

Design and preregistration

Study 3 was preregistered before data collection (AsPredicted #305,037). The study examined whether different response instructions alter collective semantic diversity and displacement from a model's default response space when externally supplied information and its prescribed use are held constant. The design crossed three language models, three ideation tasks, four experimental conditions, and a fixed set of 100 external cues. The experimental conditions differed only in whether and how an additional response instruction accompanied the common cue-use instruction. Each cue appeared exactly once within every model × task × condition combination, yielding 3,600 experimental responses. We additionally collected 100 no-cue reference responses for each model × task combination (900 responses), for a total of 4,500 generations.

The primary preregistered prediction was that differentiation instructions would increase collective semantic diversity relative to the cue-use-only condition and the neutral-instruction condition. A further preregistered directional-reversal comparison tested differentiation against convergence instructions. We additionally examined whether response instructions changed how far responses departed from each model's default response space and whether cue distance moderated these effects.

Prompting Conditions

Every experimental condition in Study 3 included the same external-cue instruction: "Use the following word as inspiration when answering: '[cue]'." Task and response-format instructions were also held constant across conditions. The four conditions therefore differed only in whether and how an additional response instruction accompanied the cue.

A *cue-use-only* condition provided no further instruction. A *differentiation* condition added, "Your answer should be different from what other people are likely to receive"; a *neutral* condition added, "Your answer should be clear and easy for other people to understand"; and a *convergence* condition added, "Your answer should be similar to what other people are likely to receive." The neutral condition was included specifically to control for the addition of a response instruction without directing responses toward either differentiation or convergence.

We refer to the first condition as *cue-use only* to distinguish it from the cue-only condition in Study 2. Whereas Study 2's cue-only condition presented the cue without instructing the model to use it, Study 3's cue-use-only condition explicitly instructed the model to use the cue as inspiration but provided no additional response instruction.

Because the task, response-format instructions, external-cue instruction, and cue identities were all held constant across conditions, the experiment was designed to isolate the effect of the additional response instructions accompanying otherwise identical external information.

Language models, tasks, and generation procedure

Responses were generated independently using OpenAI GPT-5.6-terra, Anthropic Claude Sonnet 5, and Google Gemini 3.6 Flash, with reasoning disabled, thinking disabled, and minimal thinking, respectively. Models completed the same three ideation tasks using the same task and response-format instructions as in the preceding studies. Each generation used a fresh API context with no conversational history. Trial order was randomized before generation using a prespecified random seed.

External cues and no-cue reference responses

The experiment used a fixed set of 100 external cue words assembled before preregistration to provide heterogeneous English words varying in content and word length. The same 100 cues were used across all three models, three tasks, and four experimental conditions, with each cue appearing exactly once within every model × task × condition combination.

Separately, we collected 100 responses without an external cue for each of the nine model × task combinations, yielding 900 no-cue reference responses. Reference prompts contained only the corresponding task and response-format instructions. These responses were not part of the four-

condition experimental manipulation; rather, they provided an independent estimate of each model's default response space for each task.

Semantic embeddings and outcome measures

Responses were embedded using sentence-transformers/all-mpnet-base-v2, and collective semantic diversity was quantified using the same mean pairwise cosine-distance measure as in the preceding studies.

We additionally preregistered response-to-baseline distance to measure how far individual experimental responses departed from the model's default response space. A reference centroid was constructed separately for each of the nine model × task combinations by averaging the corresponding 100 no-cue reference-response embeddings and normalizing the resulting vector. Each experimental response's cosine distance from the corresponding centroid was then calculated. The two measures captured complementary properties of generation: mean pairwise semantic distance measured how different responses were from one another, whereas response-to-baseline distance measured how far individual responses were from what the same model typically generated for that task without an external cue.

Confirmatory statistical analyses

The primary confirmatory analysis tested whether differentiation instructions increased collective semantic diversity relative to the cue-use-only and neutral-instruction conditions. A third preregistered directional-reversal test compared differentiation with convergence instructions.

Collective semantic diversity was analyzed using the same stratified-bootstrap approach as in Study 2. Responses were resampled with replacement separately within each model × task × condition combination, mean pairwise semantic distance was recalculated, and the three condition contrasts were calculated within each of the nine model × task strata before being averaged with equal weight across strata. This procedure was repeated across 10,000 bootstrap resamples. We report 95% percentile-bootstrap confidence intervals and directional bootstrap p values, calculated with a +1 correction, for the preregistered directional predictions ($\alpha = .05$).

A separate preregistered analysis tested whether response instructions affected departure from the default response space. Response-to-baseline distance was regressed on experimental condition, with cue-identity and model × task fixed effects; standard errors were clustered by cue. Planned linear contrasts compared differentiation instructions with cue-use only, neutral instructions, and convergence instructions.

Secondary cue-distance analysis

A secondary preregistered analysis examined whether the semantic distance of an external cue predicted response displacement and whether that relationship differed across response-instruction conditions. Cue-to-baseline distance was calculated as the cosine distance between each cue embedding and the corresponding model × task reference centroid and standardized within each model × task combination. Standardized cue-to-baseline distance and its interaction with experimental condition were then added to the response-displacement regression. This analysis tested whether more distant cues produced greater response displacement and whether cue distance moderated the effects of the response instructions.

During manuscript preparation, OpenAI ChatGPT was used to assist with proofing prose, code troubleshooting, and manuscript formatting. All AI-assisted output was reviewed and edited by the author, who takes responsibility for the final content.

General Discussion

Across three preregistered studies, we identified a reliable way to disrupt convergence in language-model generation: change not only how the model is asked to respond, but what it has available to draw on when responding. Supplying varied external information substantially broadened the distribution of responses and did so more reliably than simply asking models to differentiate. Crucially, this diversification did not require increasingly explicit instructions to use the information. Merely making varied information available was often sufficient to produce large shifts in generation, whereas instructing models to disregard that information largely suppressed the effect. Response instructions nevertheless remained consequential: even when external information and its prescribed use were held constant, different instructions produced systematically different response distributions. The non-monotonic pattern in Study 2 further showed that greater instructional emphasis did not progressively increase diversity, suggesting that the value of external information lies less in forcefully directing models to use it than in making distinct information available and allowing it to influence generation. Together, these findings suggest that external information supplies a powerful source of variation, while accompanying instructions can suppress its influence or further shape the resulting pattern of generation.

The effects of external information were also systematic rather than indiscriminate. In Study 1, cues located farther from a model's default response space tended to produce larger departures from that space, suggesting that the semantic position of the supplied information matters for where generation moves. Yet Study 2 showed that this influence cannot be reduced to the strength of an instruction to use the information. Contrary to our preregistered prediction, increasingly explicit encouragement to incorporate a cue did not progressively increase semantic diversity: incidental cues already produced substantial diversification, whereas explicitly treating the same information as irrelevant largely suppressed its effect. Together, these findings suggest that external information can redirect generation without needing to be strongly prioritized, but

that its influence depends both on what information is introduced and on whether the surrounding instructions permit that information to shape the response.

Behavioral guidance nevertheless remained consequential once varied information was available. In Study 3, the external cues and instructions to use them were held constant while only the additional response instructions changed. Under these conditions, differentiation instructions increased both collective semantic diversity and displacement from each model's default response space, whereas convergence instructions reduced displacement. Thus, identical informational inputs did not produce a fixed pattern of generation: the response instructions accompanying them systematically altered where and how broadly models generated. Considered alongside Studies 1 and 2, the studies reveal a consistent asymmetry between the two interventions. Varied information can broaden generation even without explicit encouragement, whereas accompanying instructions can suppress its influence, preserve the resulting diversity, or redirect the response distribution.

These findings extend work documenting homogenization in AI-assisted generation by identifying a distinct intervention target: the information environment surrounding generation. When many users approach similar generative systems with similar problems, simply instructing models to be creative or different may not reliably preserve variation across outputs. The present results suggest that a complementary strategy is to vary what models have available to draw on. Importantly, this does not appear to require forceful instructions to incorporate the information. In Study 2, incidental cues already produced substantial diversification, whereas explicitly treating the same cues as irrelevant largely suppressed their influence. This asymmetry suggests that preserving variation may depend less on compelling models to use novel information than on ensuring that potentially useful variation is available—and not prematurely excluded from consideration. Such a principle could extend beyond arbitrary cue words to examples, analogies, perspectives, documents, retrieved information, or ideas supplied by other people.

The experimental control that allowed these effects to be isolated also limits the scope of these conclusions. The studies used independent generations from three frontier language models across three relatively short ideation tasks, rather than naturally unfolding human–AI interactions. Models were also run with deliberative reasoning disabled or minimized. This design focused the experiments on how differences in the information and instructions contained in the prompt shaped generation under controlled inference settings. Whether the same patterns emerge when models engage in more extensive reasoning remains an open question, as additional deliberation may change how supplied information is interpreted, integrated, or disregarded. Real users may also revise prompts across turns, incorporate previous outputs, combine multiple sources of information, and adapt their strategies in response to the model. Nevertheless, the possibility that early informational inputs exert disproportionate influence is consistent with longstanding evidence from human creativity and judgment: initial content can anchor subsequent idea development, and exposure to examples can systematically redirect or

constrain the regions of a problem space that people explore [9, 17–19]. Our exploratory pilot suggests that variation in the wording and framing of the same underlying task does not necessarily eliminate strong response concentration, but that pilot was illustrative rather than a representative sample of naturally occurring prompts. An important next step is therefore to estimate how often strong response basins arise across a broader and defensibly sampled range of tasks, inference settings, prompting strategies, and interaction histories.

The informational intervention itself was also deliberately simple. We introduced variation through heterogeneous single-word cues, allowing the information supplied to the model to vary while the underlying task remained unchanged. Real information environments are considerably richer: users may encounter examples, analogies, documents, retrieved evidence, alternative perspectives, or ideas generated by other people. Whether these sources produce similar effects—and which kinds broaden generation without sacrificing relevance, usefulness, or accuracy—remain open questions. Our primary outcome likewise captures semantic diversity rather than creativity or quality. Greater dispersion is not inherently desirable; its value depends on whether the additional variation expands the set of useful possibilities available for the task. Future work should therefore examine when interventions that reduce convergence also improve downstream ideation, exploration, and decision making.

The present studies also distinguish what can broaden repeated model generations from what will preserve diversity across a population of users. Even an effective differentiation instruction may do little to preserve collective diversity if many users rely on similar strategies and are consequently directed toward similar alternatives. Likewise, informational variation can support population-level diversity only when different users or generations actually encounter meaningfully different information. The experiments establish that both informational inputs and behavioral guidance can systematically alter response distributions under controlled conditions; they do not establish how diverse the prompting strategies or information environments of real users are. Understanding how these processes aggregate across heterogeneous users, prompts, and interaction histories is therefore an important next step.

Generative capacity does not guarantee realized variation. Across the present studies, repeated generations were broadened more reliably by varying what models had available to draw on than by simply asking them to differentiate. That information could exert substantial influence even without explicit encouragement, yet its effects remained sensitive to behavioral guidance: the instructions surrounding that information could suppress its influence or systematically alter the resulting response distribution. Mitigating collective convergence may therefore require attention not only to how generative systems are prompted, but to the informational environments from which their responses are produced.

Data availability

The datasets generated and analyzed during the current study, along with the associated experimental materials and analysis code, are publicly available in the Zenodo repository at https://doi.org/10.5281/zenodo.23405091. Preregistrations for Studies 1–3 are available through AsPredicted (#304,474; #304,614; #305,037).

References

1. Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V. & Lakhani, K. R. The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* **35**, 1589–1607 (2024).
2. Chen, Z. & Chan, J. Large language model in creative work: The role of collaboration modality and user expertise. *Manag. Sci.* **70**, 9101–9117 (2024).
3. Anderson, B. R., Shah, J. H. & Kreminski, M. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* 413–425 (Association for Computing Machinery, 2024). https://doi.org/10.1145/3635636.3656204
4. Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
5. de Rooij, A. & Biskjaer, M. M. Generative AI makes creative output more homogeneous. In *Proceedings of the 2026 European Conference on Cognitive Ergonomics (ECCE '26)* 198–201 (Association for Computing Machinery, 2026). https://doi.org/10.1145/3822301.3822304
6. Moon, K., Green, A. E. & Kushlev, K. Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Comput. Hum. Behav.: Artif. Hum.* **6**, 100207 (2025). https://doi.org/10.1016/j.chbah.2025.100207
7. Sourati, Z. et al. The shrinking landscape of linguistic diversity in the age of large language models. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-026-02550-0 (2026).
8. Wenger, E. & Kenett, Y. N. Large language models are homogeneously creative. *PNAS Nexus* **5**, pgag042 (2026).
9. Berg, J. M. The primal mark: How the beginning shapes the end in the development of creative ideas. *Organ. Behav. Hum. Decis. Process.* **125**, 1–17 (2014).
10. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).

11. Nijstad, B. A., De Dreu, C. K., Rietzschel, E. F. & Baas, M. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *Eur. Rev. Soc. Psychol.* **21**, 34–77 (2010).
12. Holzner, N., Maier, S. & Feuerriegel, S. Generative AI and creativity: A systematic literature review and meta-analysis. Preprint at https://arxiv.org/abs/2505.17241 (2025).
13. Deng, Y., Brucks, M. & Toubia, O. Examining and addressing barriers to diversity in LLM-generated ideas. Preprint at https://arxiv.org/abs/2602.20408 (2026).
14. Terwiesch, C. et al. Artificial intelligence and its impact on creativity and diversity: An empirical study of large language model-generated product ideas. *Prod. Oper. Manag.* https://doi.org/10.1177/10591478261474243 (2026).
15. Guilford, J. P. *The Nature of Human Intelligence* (McGraw-Hill, 1967).
16. Gilhooly, K. J., Fioratou, E., Anthony, S. H. & Wynn, V. Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *Br. J. Psychol.* **98**, 611–625 (2007).
17. Leahy, K., Daly, S. R., McKilligan, S. & Seifert, C. M. Design fixation from initial examples: Provided versus self-generated ideas. *J. Mech. Des.* **142**, 101402 (2020).
18. Sio, U. N., Kotovsky, K. & Cagan, J. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Des. Stud.* **39**, 70–99 (2015).
19. Smith, S. M., Ward, T. B. & Schumacher, J. S. Constraining effects of examples in a creative generation task. *Mem. Cognit.* **21**, 837–845 (1993).

Author contributions

C.B.H. conceived and designed the research, developed the experimental materials, collected and analyzed the data, prepared the figures, and wrote the manuscript.

Additional Information

Competing interests

The author declares no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx