HIGH SCORES, NARROW VOCABULARIES: MEASURING CROSS-TRIAL DIVERSITY IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used in creative industries as sources of novel ideas, treating output variation as evidence of genuine creative exploration. The Divergent Association Task (DAT), a validated creativity benchmark, has repeatedly produced at- or above-human level scores for LLMs, but a single-trial score cannot distinguish broad semantic exploration from repeated draws on a narrow, high-scoring vocabulary. We demonstrate this dissociation directly: LLMs can attain human-level or above-human DAT scores while showing substantially narrower cross-trial vocabularies than humans. We test five models, three open-source and two frontier, with 100 trials per condition under varied prompts, reasoning instructions, and decoding parameters, assessed using five metrics: DAT score, word overlap, vocabulary breadth, entropy, and semantic-region overlap. Prompt wording and reasoning instructions are the strongest predictors of DAT score, but they do not reliably increase lexical diversity. By contrast, widening the sampling pool reliably increases diversity without a corresponding gain in score for most models. Therefore, neither lever, within the standard decoding settings, closes the gap between benchmark performance and genuine exploration. Frontier models show the most extreme version of convergence, showing the highest cross-trial overlap and the narrowest vocabulary of any model tested. These findings show that human-like DAT performance does not imply human-like diversity. We propose multiple metrics of cross-trial diversity as a behavioral diagnostic of generative convergence that single-trial scores cannot reveal. As LLMs become routine in creative work, such convergence may homogenize otherwise independent outputs, warranting caution in treating the properties of individual outputs as evidence of genuine creative diversity.

1 INTRODUCTION

Large language models are increasingly used in creative industries such as marketing, advertising, and content production. Users treat their output as a source of novel ideas. This use rests on an implicit assumption: that the diversity of an LLM's response reflects genuine creative exploration rather than pattern-matching to training data.

Divergent thinking is the capacity to generate many conceptually distinct responses to an open-ended problem. It is a well-established indicator of creative ability (Guilford, 1950). The Divergent Association Task (DAT) operationalizes this: respondents name ten maximally dissimilar nouns. The response is scored automatically as the mean pairwise cosine distance between GloVe embeddings of the words produced (Olson et al., 2021). The measure correlates with established creativity tests and has been validated across samples exceeding 100,000 participants (Olson et al., 2021; Bellemare-Pepin et al., 2026), making the DAT a rare large-scale automated creativity benchmark. Other widely used divergent thinking tests such as Guilford's Alternative Uses Task (Guilford, 1967) and Torrance's Tests of Creative Thinking (Torrance, 1966) require human judgment to score critical dimensions like flexibility, originality, and cleverness (Silvia et al., 2008; Reiter-Palmon et al., 2019), making evaluation at scale prohibitively costly. Using LLMs as judges does not resolve this problem because LLM judges exhibit systematic biases such as preference towards their own generations (Zheng et al., 2023; Panickssery et al., 2024), which are amplified at scale. DAT avoids this problem

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

through a simpler design of the task around embedding-based metrics of semantic diversity. Related semantic-diversity measures also correlate with human creativity ratings in narrative generation (Orwig et al., 2024; Fan et al., 2023).

LLMs have repeatedly achieved human-level or above-human DAT scores. Under greedy decoding, GPT-4 placed above 96% of human respondents, and GPT-3.5-Turbo exceeded the human average as well (Chen & Ding, 2023). Bellemare-Pepin et al. (2026) similarly report LLMs exceeding average human DAT performance. Both studies rely on raw DAT scores; however, a high DAT score does not tell us whether the process behind it was genuine exploration. A model could instead have learned to draw from a narrow set of high-scoring words, potentially including benchmark-specific information (Xu et al., 2024), or on simpler heuristics such as repeatedly opening with the same word. Our multi-metric approach addresses this gap.

Indeed, substantial repetition has been observed across 102 DAT responses from both GPT-3.5 and GPT-4, with recurring words and semantic categories identified despite their high mean DAT scores (Cropley, 2023). However, these observations stop at descriptive evidence of repetition: word-frequency counts and manually identified recurring categories. It does not quantify cross-trial convergence as a systematic outcome, nor compare multiple dimensions of diversity. We extend this observation by measuring cross-trial lexical and semantic diversity across multiple models.

A related failure has also been documented in frontier models outside semantic generation: models can produce biased empirical distributions even when prompted to sample from known target distributions (Gu et al., 2026). This bias persists across temperature, top-p, and top-k manipulations and is not corrected by chain-of-thought (CoT) prompting. Although their analysis concerns numerical tasks with known target distributions, it suggests that apparent sampling diversity may not faithfully reflect the underlying distribution a model is instructed to produce.

The analogous question in semantic generation is harder because there is no known target distribution: "correct" output is instead a judgment of creative divergence. We address this gap directly through repeated sampling with varied generation conditions and multi-metric assessment to measure convergence, and examine whether the pattern survives different prompting or decoding conditions. **We test to what extent LLMs can achieve high DAT scores while exhibiting greater cross-trial lexical and semantic convergence than the human baseline.**

Five models, spanning frontier and open-source systems, each completed the DAT one hundred times per condition. We varied decoding parameters and prompt phrasing systematically. Model behavior was assessed through a multi-metric suite designed to make cross-trial convergence visible in ways a single benchmark score cannot. Our results support three claims.

First, benchmark (DAT) score does not indicate how broadly a model explores the response space. Models can score at or above the human level while using a substantially narrower vocabulary than the human baseline. This holds even when a model's score sits at the human level rather than above it: the vocabulary constraint appears independently of whether it happens to pay off in score.

Second, prompt rewording and reasoning instructions do not reliably increase diversity. These interventions consistently shift DAT score, but their effect on vocabulary breadth is inconsistent: in both frontier models, a reasoning step significantly changes score and word overlap without significantly changing type-token ratio, so the shift reflects a redistribution among existing word choices rather than genuine exploration. Where vocabulary breadth does shift, it does not move consistently toward greater diversity, and the same formulation of an instruction can produce opposite relationships between score and diversity depending on the model. Thus, these interventions alter model output, but they do not provide a reliable means of increasing diversity.

Third, broader sampling increases the vocabulary a model draws on, but that growth does not reliably carry over to either more even word usage or higher benchmark scores. Widening the sampling pool reliably increases vocabulary diversity in every model we test, but often fails to produce a significant gain in DAT score. Where score does increase with higher temperature, the same setting also raises the number of invalid responses, which points to degraded performance instead of broader exploration; we had to exclude a condition for one of the models as it produced too many invalid responses under broader sampling settings.

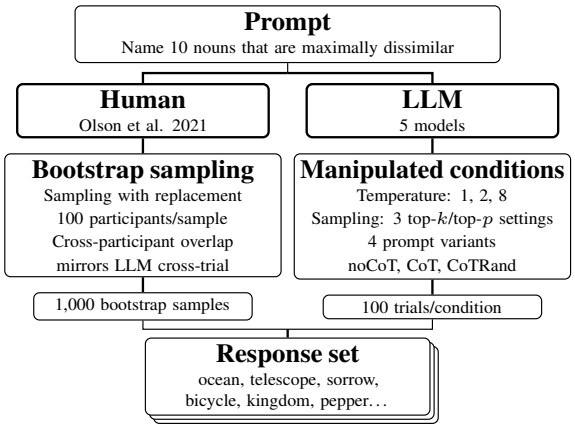


Figure 1: Schematic of the DAT task flow. Human and LLM branches each produce a response set. The displayed prompt is a shortened, synonymous instruction meant to illustrate the task.

2 METHOD

Experimental design. This study employs an empirical behavioral design in which each model repeatedly completes the DAT: In each trial, the model is prompted to generate ten words that are as semantically dissimilar from one another as possible, following the standard DAT procedure. Prompt formulation and generation settings are systematically varied across conditions. Each trial is initialized independently. Locally run models are fully unloaded from memory between runs to prevent persistent application state from influencing subsequent output, and no random seeds are specified so that sampling remains stochastic rather than being constrained to a reproducible pseudorandom sequence.

Models. Five models span a range of scales and deployment contexts: three open-source models run locally via Ollama (Llama3-8B, Mistral-7B, Phi4-14B) on a MacBook Pro with an Apple M1 chip (8 cores; 4 performance, 4 efficiency) and 16GB unified memory. These hardware constraints limited the practical model size and context length, motivating the use of $7-14$B-parameter models. Two frontier models were accessed via the OpenAI API (GPT-5.4, GPT-5.4-mini) and required no local compute. This selection spans small-to-mid open-source models testable at extreme decoding settings alongside frontier systems constrained to realistic API parameters.

Manipulations. The primary point of interest was cross-trial convergence, but the experimental design additionally varied prompt formulation, reasoning instructions, and decoding parameters to test whether any observed convergence was robust to changes in generation conditions. Four synonymous prompt variants (P1−P4) of the standard DAT instruction test sensitivity to instruction wording (Table 14). Each is paired with three reasoning conditions: no reasoning instruction (noCoT), a step-by-step reasoning prefix (CoT), and a silent random-number-generation prefix before the reasoning instruction (CoTrand). A regular expression strips common preamble openers (e.g., "Here", "Sure", "Certainly") from all responses.

Three decoding parameters govern token selection: temperature, which rescales the output distribution toward or away from the most probable tokens; top-k, which restricts selection to the k most probable candidates; and top-p, which restricts selection to the smallest candidate set whose combined probability exceeds p. Temperature is varied across three values (1, 2, 8); the OpenAI API caps temperature at 2, so the temperature $= 8$ condition is not run for either GPT model. Top-k (40, 100, 1000) is applied to Ollama models only. GPT models are tested across top-p values instead (0.4, 0.7, 1.0). Initial pilot testing produced largely non-significant effects. Because the API does not expose top-k, GPT models were tested with top-p instead. The two are not interchangeable, but both control the size of the eligible token pool. We restricted decoding manipulations to temperature, top-k, and top-p because these are the default controls available to users of both local inference tools and commercial APIs, and therefore reflect how LLMs are used in practice in creative workflows.

Data collection. Data collection uses a custom Python script with timeout and retry logic: each Ollama query has a 120-second timeout and up to three attempts. All raw responses are logged alongside their parsed status (ok, bad_N_words, skipped). Word-level cleaning strips non-ASCII and non-alphabetic characters and lowercases text; tokens under 2 or over 20 characters are rejected as likely hallucinated output. Conditions were excluded when invalid responses would require substantial resampling; one condition met this criterion (GPT-5.4-mini, temperature $= 2$, top-$p = 1.0$); all 12 corresponding prompt and reasoning conditions were therefore excluded.

Metrics. Five metrics are computed per condition.

- **1. DAT score** is the mean pairwise cosine distance between GloVe embeddings (glove.840B.300d) (Pennington et al., 2014) of the first seven valid words per trial, multiplied by 100 (Olson et al., 2021).
- **2. Average pairwise word overlap** quantifies cross-trial lexical convergence: for each trial pair in a condition, the number of shared words is divided by 10 and averaged across all pairs, yielding a value between 0 (no shared words) and 1 (identical trials).
- **3. Type-token ratio (TTR)** is the proportion of unique word types among all word slots in a condition, measuring vocabulary breadth independently of pairwise overlap.
- **4. Shannon entropy** is the concentration of the word-frequency distribution; the same calculation applied separately at each of the ten response positions yields **positional entropy**, used to detect whether word choice is more constrained at particular positions.
- **5. Semantic region overlap** measures convergence at a coarser level. The 20,000 most frequent GloVe words are L2-normalized and clustered into 100 MiniBatchKMeans regions; trial words are assigned to their nearest region, and region overlap is computed the same as word overlap.

No single metric captures convergence alone. Word overlap measures pairwise repetition but not total vocabulary size. TTR captures that missing dimension, measuring the overall pool across all trials. Entropy captures how evenly word choices are distributed across the aggregate vocabulary. A condition can have a large vocabulary, but low entropy if it is dominated by a small number of frequently recurring words. Region overlap tests convergence at a coarser semantic level than word overlap. It is able to detect synonym substitution, where a model varies surface vocabulary while returning to the same semantic territory, a pattern none of the lexical metrics or the DAT score itself can see.

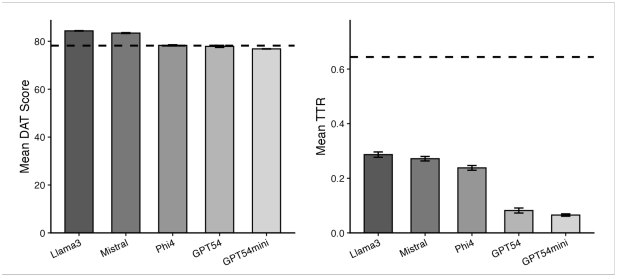
Human baseline. Human data is drawn from Olson et al. (2021). We repeatedly sampled groups of 100 participants and applied the same cross-response overlap procedure used for the 100 model trials in each condition (total of 1,000 bootstrap samples). This establishes comparability of the computed metrics between LLMs and humans.

3 RESULTS

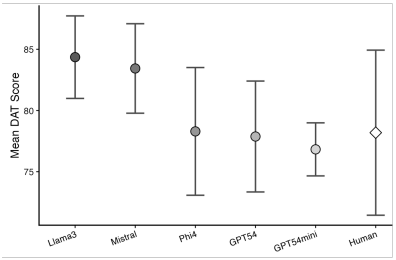
Benchmark performance against vocabulary breadth. GPT-5.4 ($M = 77.88$), GPT-5.4-mini ($M = 76.83$), and Phi4 ($M = 78.29$) scored near the human mean (78.19), while Mistral ($M = 83.43$) and Llama3 ($M = 84.36$) exceeded it (Figure 2a; Table 1). Additionally, all models show a narrower trial-level DAT score spread than humans (Figure 2b; Table 1). Type-token ratio was below the human value (0.644) for every model, ranging from 0.065 (GPT-5.4-mini) to 0.287 (Llama3) (Figure 2a; Table 4). Mean word overlap exceeded the human bootstrap baseline ($M = 0.026$) in every model, from 4.4 times (Llama3) to 21.1 times (GPT-5.4) the human value (Table 2; Table 8).

Normalized entropy showed a similar pattern: every model fell below the human value (0.956), ranging from 0.814 (GPT-5.4-mini) to 0.893 (Llama3) (Table 5). Taken alone, entropy would suggest models distribute their word choices nearly as evenly as humans do; type-token ratio shows that this evenness applies to a pool two to ten times smaller. This illustrates the central dissociation: all models achieve a DAT score near or above the human level with a substantially narrower vocabulary.

To examine the influence of the manipulated conditions, we plot the relationship between DAT score and word overlap for each model (Figure 3; by each manipulation in Figures 8 to 11). A flat

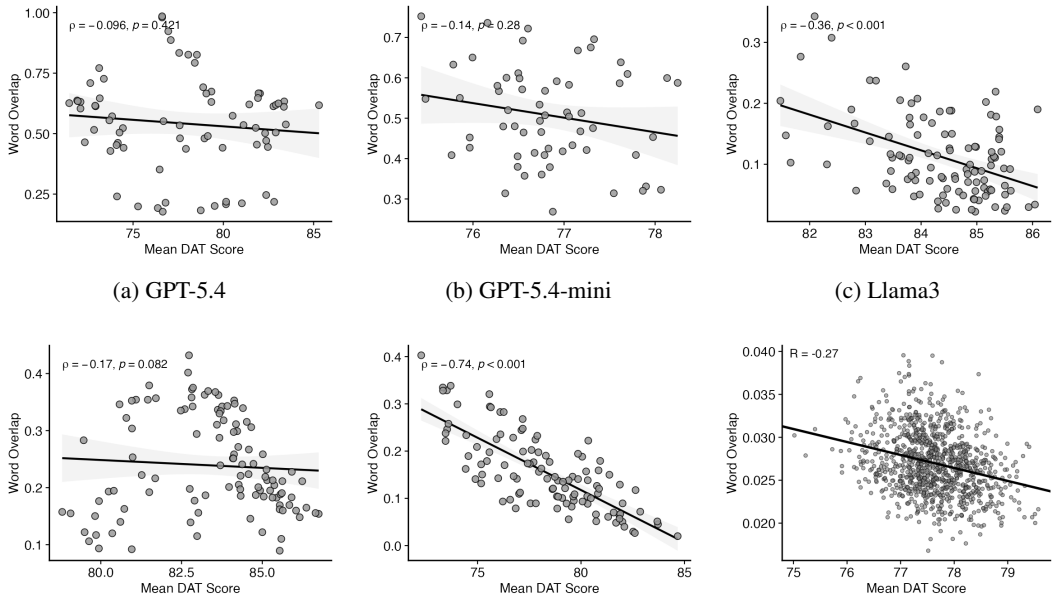


(a) DAT Score and TTR



(b) Score stability

Figure 2: Benchmark performance and its stability across conditions. Left: Mean DAT score and TTR per model, human baseline for each shown as a dotted line. *Note*. A DAT score on par with the human baseline does not mean a humanlike vocabulary was used. Right: Mean DAT score and trial-level standard deviation per model (Table 1). Note. Every model's scores spread less than the human baseline.



(a) GPT-5.4 (b) GPT-5.4-mini (c) Llama3

(d) Mistral (e) Phi4 (f) Human

Figure 3: DAT score vs. average pairwise word overlap per model across all conditions. Human panel: 1,000 bootstrap samples. Each point is one experimental condition; a negative slope indicates that manipulations which raise a model's score also tend to reduce its repetition.

correlation indicates that score gains across conditions are not accompanied by reduced repetition, so a manipulation that raises the score does not thereby broaden the vocabulary. A negative correlation suggests that higher scores coincide with reduced repetition, although its interpretation depends on the specific manipulation.

Robustness to task framing. Chain-of-thought prompting significantly altered word overlap in four of five models (all except Mistral, which was ns; Table 11) and significantly altered DAT score in all five, where it was the dominant predictor in Llama3 ($F(2, 68) = 96.67$, $p < .001$; Table 10). Within Llama3, the DAT-overlap correlation was flat under noCoT ($r = -0.080, ns$) but became significant with reasoning (CoT: $r = -0.599$, $p < .001$; CoTrand: $r = -0.570$, $p < .001$) (Table 3). However, adding reasoning was associated with lower mean DAT scores and higher word overlap, showing that the correlation alone does not capture the manipulation's effect (Figure 3).

Prompt formulation was a strong predictor of DAT score in every model (Table 10), but its relationship to overlap was unstable. For Mistral, three of four prompts produced strong negative correlations (P1: $r = -0.498$, $p < .01$; P2: $r = -0.963$, $p < .001$; P4: $r = -0.808$, $p < .001$), while P3 reversed this relationship ($r = 0.614$, $p < .001$) (Table 3). This sign reversal under a same-meaning reword suggests prompt-specific rather than stable model-level behavior.

Across every CoT level, the DAT-overlap correlation remained non-significant for both frontier models (GPT-5.4: overall $r = -0.096$; GPT-5.4-mini: $r = -0.141$). At the prompt level, however, both models show significant correlations (GPT-5.4 at P2 and P3; GPT-5.4-mini at P1, P2, and P4; Table 3); for GPT-5.4-mini the sign reverses with wording (P1: $r = -0.689$, $p < .01$; P2: $r = 0.539$, $p < .05$), as it does for Mistral.

Robustness to sampling. Expanding the sampling pool (top-*k*/*p*) significantly increased type-token ratio in every model tested (all $p < .001$; Table 12), the most consistent effect of any manipulation on vocabulary breadth in the study. Its effect on normalized entropy was weaker and less consistent, reaching significance in three of five models (Table 13). It had no significant effect on DAT score in four of five models, the exception being GPT-5.4-mini ($F(2, 29) = 4.76$, $p < .05$; Table 10).

Temperature showed a parallel pattern for overlap: a significant reducer of overlap in all three open-source models (Mistral: $F(2, 68) = 98.88$; Phi4: $F(2, 68) = 135.72$; Llama3: $F(2, 68) = 211.56$, all $p < .001$) but not in either frontier model (Table 11). This overlap reduction did not carry over to score consistently: temperature's effect on DAT score reached significance for Mistral ($F(2, 68) = 8.31$, $p < .001$) and Phi4 ($F(2, 68) = 19.30$, $p < .001$), but not for Llama3 ($F(2, 68) = 1.38$, ns), despite Llama3 showing the largest overlap reduction of the three (Table 10). Both frontier models showed no significant temperature effect on DAT score (GPT-5.4: $F(1, 40) = 1.93$, ns; GPT-5.4-mini: $F(1, 29) = 3.15$, $p < .10$).

Phi4 showed the strongest temperature effect on DAT score of any model ($F(2, 68) = 19.30$, $p < .001$; Table 10), and is the one model whose DAT-overlap correlation was negative and significant across all thirteen predictor-level subsets ($r = -0.708$ to -0.904, all $p < .001$; Table 3), including the noCoT and temperature = 1 subsets.

Secondary results. Every model's normalized entropy fell below the human value (Table 5). However, this aggregate score pools all ten response positions into one distribution, which can mask a bias concentrated in a single position; positional entropy revealed a pronounced first-word bias in GPT-5.4, GPT-5.4-mini, and Mistral (gaps of 0.29, 0.39, and 0.45 respectively, against a human baseline gap of 0.11), while Phi4 (0.14) and Llama3 (0.11) showed no such concentration (Figure 4; Table 6). Repeated first-word choices indicate a strong positional constraint.

Semantic region overlap exceeded word overlap in every model, but the size of that gap differs per model class (Table 7). The human gap is 0.170, indicating some variation despite lexical semantic similarity of the chosen words. The open-source models show wider gaps: Mistral (0.186), Phi4 (0.188), and Llama3 (0.250). This indicates that when these models diverge across trials, they rely more on synonym or other related word substitution than humans. The frontier models show the opposite pattern. Their gap is a fraction of the human value: GPT-5.4 (0.077) and GPT-5.4-mini (0.038), meaning the models are rigid in their choices within a semantic region.

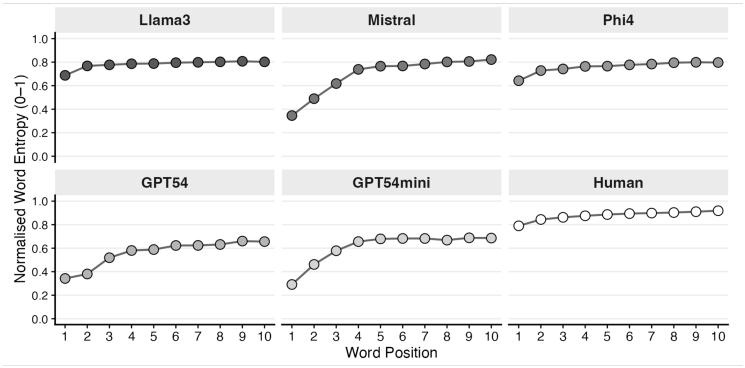


Figure 4: Normalized Shannon entropy of the word distribution at each output position (1 = first word, 10 = last). *Note*. The first-word bias for GPT-5.4, GPT-5.4-mini, and Mistral.

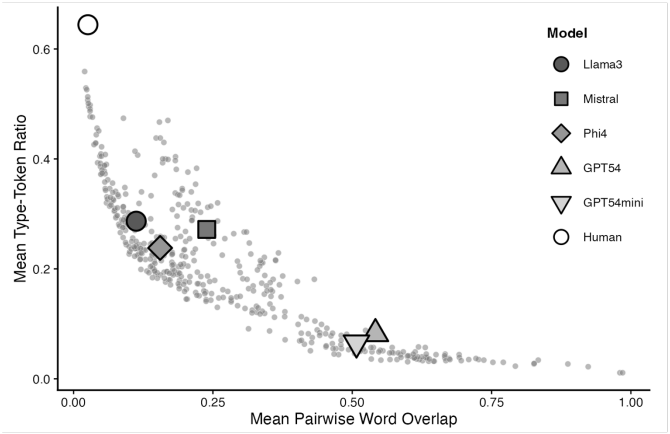


Figure 5: Correlation between TTR and mean word overlap. Each small dot represents one experimental condition; large dots represent each model's centroid. *Note*. Frontier models cluster near the high-overlap / low-TTR end because their vocabulary is extremely concentrated. Humans sit at the opposite low-overlap / high-TTR end. The open-source models sit in between.

4 Discussion

While every model performs near or above the human level on the DAT, they all fall short of the human baseline in vocabulary breadth (Table 9). When plotted, these average vocabulary sizes are grouped in three distinct clusters: humans, open-source models, and frontier models (Figure 5). However, the form of cross-trial convergence differs between models.

Llama3. Llama3's negative DAT-overlap correlation is closely associated with the reasoning manipulation. It appears only once a reasoning step is introduced. Under noCoT alone, the relationship is flat (Figure 14b). One possibility is that chain-of-thought commits the model to a narrower semantic trajectory before word selection. If that trajectory narrows early, the final words are drawn from a smaller region than noCoT would reach. This is not a new stochasticity. It is a narrower search, not a broader one.

Mistral. Mistral's instability points to a different mechanism. Three of the four prompt variants produce a strong negative correlation, but the fourth reverses it entirely, with markedly lower scores (Figure 15d). A sign reversal under a same-meaning rewording suggests that the relationship can be prompt-specific rather than a stable property of the model.

Phi4. Phi4 is the only model where our two diversity signals move together as the DAT is meant to detect: across all thirteen predictor-level subsets, lower cross-trial word overlap is significantly and

consistently paired with a higher DAT score ($r = -0.708$ to -0.904, all $p < .001$; Table 3). At face value, this looks like a clear sign of emergent diversity. However, its comparative performance among the open-source models complicates this reading: Phi4's mean DAT score and vocabulary breadth are the lowest of the three open-source models. It is also the least stable of the three (Figure 2b), raising questions about Phi4's reliability. Now, it could merely be gains from its own narrower initial response distribution, making it difficult to interpret it as useful divergent exploration.

Frontier models. The frontier models sit at the extreme end of nearly every metric in this study: highest overlap, lowest TTR, smallest region-word gap, and flattest overall DAT-overlap correlation. One plausible explanation is that the frontier models' training and deployment objectives favor consistent, high-probability responses; Kirk et al. (2024) found that RLHF, a techique used extensively in post-training of frontier models, while improving generalization across tasks, significantly reduces output diversity. Our results suggest that this may indeed come at the cost of real creativity.

Limitations. First, the DAT and the accompanying diversity metrics are behavioral measures of lexical and semantic value, not direct measures of the mechanisms underlying generation. Second, the study covers only five models and is therefore not intended to establish that the observed pattern is representative of LLMs as a whole. Third, the diversity measures depend on GloVe representations and, for semantic region overlap, an arbitrary clustering of the embedding space. Different representations or partitions may yield different estimates of convergence. Finally, the decoding conditions were not fully matched across model classes: open-source models were tested with temperature and top-k, while GPT models were constrained by the API and tested with top-p instead. The observed differences between model classes should therefore be interpreted as behavioral differences under the tested conditions rather than as a controlled comparison of model architectures.

Future work should also test whether adaptive decoding strategies or other generation-time interventions can broaden cross-trial diversity without sacrificing coherence; the evaluation of sampling levers was bounded by standard static truncation methods, and distribution-adaptive sampling algorithms designed to mitigate mode collapse, such as min-sampling, typical decoding, or contrastive decoding, are viable directions for future exploration.

5 Conclusion

Across five models and hundreds of experimental conditions, every model performed near or above human parity on the DAT, but produced substantially more concentrated response distributions, with every model's mean far outside the human bootstrap distribution: higher word overlap, lower type-token ratio, lower entropy, and, for the open-source models, greater reliance on semantic substitution at the region level; the frontier models showed the opposite pattern, converging so tightly at the lexical level that little room remained for regional substitution at all.

Prompt, reasoning, and sampling manipulations failed to remedy this consistently. This suggests that DAT performance alone may overstate the breadth of semantic exploration represented by repeated model generations. Phi4 provides the clearest apparent exception to this erratic effect: its DAT-overlap correlation is consistently negative, meaning that parameter manipulations tend to raise DAT score and reduce word overlap. However, its overall response distribution remains far more concentrated than the human baseline, so this trajectory does not exempt it from the paper's broader conclusion.

The frontier models show the opposite pattern, combining the narrowest vocabularies and highest overlap with DAT score insensitive to the tested manipulations. This is the paradox of creativity in current language models: the properties that make a model dependable are the same properties that foreclose genuine divergent thinking. This converges with findings on other tasks, where LLMs likewise struggle to generate diverse outputs (Wang et al., 2026; Jiang et al., 2025; Zhang et al., 2025), suggesting the lack of true divergence which we observe is not an artifact of DAT.

Together, these patterns demonstrate why single-trial creativity scores are insufficient to characterize generative diversity, as done in previous work (Bellemare-Pepin et al., 2026). To address this, we proposed a cross-trial, multi-metric diversity assessment able to detect both the lexical and semantic convergence across trials, which any metric, such as the DAT, alone cannot. The broader concern extends beyond individual models: if these narrow distributions characterize models used repeatedly in creative workflows, independent outputs may increasingly draw from the same underlying vocabulary.

This risks undermining the diversity such sectors are meant to create, where many ideas would feed into a shared pool of ideas. Therefore, we advise to take considerable caution when using modern LLMs as sources of creativity.

AI USE STATEMENT

In this work, generative AI tools were used to generate the response data analyzed in this paper; this use is central to the paper's methodology and is described in full in Section 2 (Method) and the Appendix. We additionally used generative AI tools to assist with LaTeX formatting and template adaptation, and for light prose polishing and translation. Generative AI tools were also used to refactor the analysis pipeline into modular, documented components; we verified that the refactored code reproduces all reported results exactly (see Reproducibility Statement). We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide an anonymized package containing code and data as supplementary materials. This package features a metrics-calculation pipeline developed in Python, which processes raw model outputs to produce that dataset, along with fixed dependency specifications for both components. Additionally, it includes a complete analysis workflow written in R, capable of generating all tables and figures present in the publication directly from an embedded dataset. The dataset itself contains raw responses from the model across various conditions, calculated metrics, and human-baseline outputs, with documentation of its structure and origins available in the supplementary README. Licenses and access terms for all third-party assets used (pretrained word embeddings, the open-weight language models, and the human baseline data of Olson et al., 2021) are documented in the LICENSE file. Recreating the paper's tables and figures solely from the supplementary package does not require API keys, GPU hardware, or external downloads; however, reconstructing a dataset from raw responses does, due to reasons explained in the supplementary README.

REFERENCES

A. Bellemare-Pepin, F. Lespinasse, P. Thölke, Y. Harel, K. Mathewson, J. A. Olson, Y. Bengio, and K. Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1): 1279, 2026.

H. Chen and N. Ding. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12881–12888, 2023.

D. Cropley. Is artificial intelligence more creative than humans?: ChatGPT and the divergent association task. *Learning Letters*, 2:13, 2023.

L. Fan, K. Zhuang, X. Wang, J. Zhang, C. Liu, J. Gu, and J. Qiu. Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology*, 60:e14239, 2023.

X. Gu, S. De, M. Titsias, L. Markeeva, P. Veličković, and R. Pascanu. The illusion of stochasticity in LLMs. *arXiv preprint arXiv:2604.06543*, 2026.

J. P. Guilford. Creativity. *American Psychologist*, 5:444–454, 1950.

J. P. Guilford. *The Nature of Human Intelligence*. McGraw-Hill, 1967.

L. Jiang, Y. Chai, M. Li, M. Liu, R. Fok, N. Dziri, Y. Tsvetkov, M. Sap, and Y. Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, volume 38, 2025.

R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, and R. Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, pp. 20620–20653, 2024.

J. A. Olson, J. Nahas, D. Chmoulevitch, S. J. Cropper, and M. E. Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021.

W. Orwig, E. R. Edenbaum, J. D. Greene, and D. L. Schacter. The language of creativity: Evidence from humans and large language models. *Journal of Creative Behavior*, 2024.

A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

R. Reiter-Palmon, B. Forthmann, and B. Barbot. Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):144–152, 2019.

P. J. Silvia, B. P. Winterstein, J. T. Willse, C. M. Barona, J. T. Cram, K. I. Hess, J. L. Martinez, and C. A. Richard. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2 (2):68–85, 2008.

E. P. Torrance. *The Torrance Tests of Creative Thinking: Norms-Technical Manual*. Personnel Press, Princeton, NJ, 1966.

D. Wang, D. Huang, H. Shen, and B. Uzzi. A large-scale comparison of divergent creativity in humans and large language models. *Nature Human Behaviour*, 10(3):531–540, 2026.

C. Xu, S. Guan, D. Greene, and M. Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Y. Zhang, H. Diddee, S. Holm, H. Liu, X. Liu, V. Samuel, B. Wang, and D. Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.05228.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, and H. Zhang. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A Appendix

A.1 Tables

Table 1: Descriptive statistics for DAT scores across all individual trials.

Model	N_trials	Mean DAT	SD	Min	Max	Range
GPT-5.4	7198	77.876	4.529	66.992	89.551	22.56
GPT-5.4-mini	6000	76.825	2.166	66.94	86.353	19.413
Mistral	10800	83.433	3.655	63.594	95.433	31.839
Phi4	10800	78.293	5.216	56.551	94.428	37.877
Llama3	10798	84.356	3.371	59.424	94.584	35.16
Human	8857	78.188	6.742	23.846	95.735	71.889

Note. Statistics computed over all individual trials. Human row = raw statistics over all individual participant trials (not bootstrap). Model SD reflects variability across trials; human SD reflects variability across participants.

Table 2: Descriptive statistics for average pairwise word overlap per condition.

Model	Mean Overlap	SD	Min	Max
GPT-5.4	0.5416	0.1971	0.1772	0.9851
GPT-5.4-mini	0.5076	0.1189	0.2684	0.7524
Mistral	0.239	0.0853	0.0895	0.432
Phi4	0.1553	0.0808	0.0199	0.4028
Llama3	0.1124	0.0659	0.0221	0.3428
Human (bootstrap)	0.0257	0.0035	0.0159	0.0405

Note. Human row = bootstrap condition-level statistics (1,000 samples of $n = 100$ participants).

Table 3: Spearman correlations (r) between average pairwise word overlap and mean DAT score, within each decoding predictor level.

Level	GPT-5.4	GPT-5.4-mini	Mistral	Phi4	Llama3
Overall r	-0.096 ns	-0.141 ns	-0.168 .	-0.739 ***	-0.362 ***
Temperature 1	-0.006 ns	-0.291 .	-0.247 ns	-0.895 ***	-0.465 **
Temperature 2	-0.125 ns	0.104 ns	0.203 ns	-0.843 ***	-0.438 **
Temperature 8	—	—	0.230 ns	-0.779 ***	-0.448 **
noCoT	0.007 ns	-0.230 ns	-0.263 ns	-0.841 ***	-0.080 ns
CoT	-0.082 ns	-0.168 ns	0.076 ns	-0.729 ***	-0.599 ***
CoTrand	-0.059 ns	-0.114 ns	-0.413 *	-0.751 ***	-0.570 ***
Top-k 1000	—	—	-0.139 ns	-0.779 ***	-0.389 *
Top-k 100	—	—	-0.140 ns	-0.737 ***	-0.314 .
Top-k 40	—	—	-0.167 ns	-0.708 ***	-0.298 .
Top-p p04	-0.115 ns	-0.126 ns	—	—	—
Top-p p07	-0.117 ns	0.062 ns	—	—	—
Top-p p10	0.054 ns	-0.497 ns	—	—	—
Prompt P1	-0.007 ns	-0.689 **	-0.498 **	-0.728 ***	-0.325 .
Prompt P2	-0.730 ***	0.539 *	-0.963 ***	-0.904 ***	-0.216 ns
Prompt P3	-0.781 ***	0.461 .	0.614 ***	-0.844 ***	-0.528 **
Prompt P4	0.348 ns	-0.604 *	-0.808 ***	-0.753 ***	-0.213 ns

Note. *** $p < .001$, ** $p < .01$, * $p < .05$, . $p < .10$, ns not significant.

Table 4: Descriptive statistics for type-token ratio (TTR) per condition.

Model	Mean TTR	SD
GPT-5.4	0.0819	0.0781
GPT-5.4-mini	0.0654	0.0338
Mistral	0.2717	0.0855
Phi4	0.2381	0.0896
Llama3	0.2867	0.1013
Human (bootstrap)	0.6441	0.0247

Note. Human row = bootstrap condition-level statistics (1,000 samples, $n = 100$).

Table 5: Descriptive statistics for normalized Shannon entropy per condition.

Model	Mean entropy	SD
GPT-5.4	0.819	0.038
GPT-5.4-mini	0.814	0.024
Mistral	0.831	0.035
Phi4	0.873	0.036
Llama3	0.893	0.034
Human (bootstrap)	0.956	0.005

Note. Entropy is normalized Shannon entropy (raw entropy divided by $\log_2(n_{\text{unique}})$), in $[0, 1]$, for both model and human rows. Human row = bootstrap condition-level statistics (1,000 samples, $n = 100$).

Table 6: First-word positional bias: normalized entropy at position 1 versus plateau (positions 5–10).

Model	Entropy at Word 1	Plateau (5-10)	First-word gap
GPT-5.4	0.3425	0.6304	0.2879
GPT-5.4-mini	0.2911	0.681	0.3899
Mistral	0.346	0.7912	0.4452
Phi4	0.6422	0.7864	0.1442
Llama3	0.6877	0.7991	0.1114
Human	0.7898	0.9017	0.1119

Note. Plateau = mean normalized entropy across positions 5–10. First-word gap = Plateau − Entropy at Word 1. Human values computed from Olson et al. (2021) pooled data.

Table 7: Semantic region overlap descriptive statistics and correlations.

Model	Mean Region Overlap	SD	Gap	r(region, DAT)	r(region, word)
GPT-5.4	0.6182	0.1352	0.0767	-0.006 ns	0.940 ***
GPT-5.4-mini	0.5456	0.0643	0.0381	-0.115 ns	0.924 ***
Mistral	0.4248	0.0713	0.1858	-0.146 ns	0.949 ***
Phi4	0.3435	0.0402	0.1882	-0.575 ***	0.924 ***
Llama3	0.3626	0.0608	0.2502	-0.329 ***	0.932 ***
Human (bootstrap)	0.1953	0.0111	0.1696	0.090 **	0.517 ***

Note. Gap = Mean Region Overlap − Mean Word Overlap. Human row = bootstrap condition-level statistics. *** $p < .001$, ** $p < .01$, * $p < .05$, ns not significant.

Table 8: Model word and region overlap as multiples of the human bootstrap baseline.

Model	Word Overlap	Region Overlap	Word Ratio	Region Ratio
GPT-5.4	0.5416	0.6182	21.1×	3.2×
GPT-5.4-mini	0.5076	0.5456	19.7×	2.8×
Mistral	0.239	0.4248	9.3×	2.2×
Phi4	0.1553	0.3435	6.0×	1.8×
Llama3	0.1124	0.3626	4.4×	1.9×

Note. Ratios relative to human bootstrap baseline (word overlap: $M = 0.0257$; region overlap: $M = 0.1953$).

Table 9: Cross-comparison of human baseline and all five LLMs across key metrics.

Model	N	DAT Mean (SD)	Word Overlap	Region Overlap	TTR	r(overlap,DAT)
GPT-5.4	7198	77.876 (3.808)	0.5416	0.6182	0.0819	-0.096 ns
GPT-5.4-mini	6000	76.825 (0.661)	0.5076	0.5456	0.0654	-0.141 ns
Mistral	10800	83.433 (1.981)	0.239	0.4248	0.2717	-0.168 .
Phi4	10800	78.293 (2.794)	0.1553	0.3435	0.2381	-0.739 ***
Llama3	10798	84.356 (1.018)	0.1124	0.3626	0.2867	-0.362 ***
Human (bootstrap)	8857	78.188 (6.742)	0.0257	0.1953	0.6441	-0.257 ***

Note. Human N and DAT Mean/SD reflect the full valid-DAT population ($n = 8{,}857$); human DAT SD here is trial-level, matching the human row in Table 1. Word Overlap and Region Overlap for the human row are resampled from that same $n = 8{,}857$ population (matching how these metrics are computed for the models: any response with at least one valid word contributes). TTR is resampled from the same draws but, within each bootstrap sample, only participants with all 10 words present count toward it ($n = 7{,}091$ have all 10), again matching how TTR is computed for the models. All human statistics here are 1,000 bootstrap condition-level samples. For all five models, every statistic in this table (including DAT SD) is a condition-level statistic and does NOT match the models' trial-level DAT SD in Table 1 – see Table 1 for that instead. *** $p < .001$, ** $p < .01$, * $p < .05$, ns not significant.

Table 10: Factorial ANOVA results for DAT score by manipulated factor (Type III sums of squares; prompt, CoT, temperature, and top-k/p plus their two-way interactions).

Model	effect	df_effect	df_resid	F_val	p_val	eta2p	sig
GPT-5.4	prompt	3	40	103.035	0.0	0.8854	***
GPT-5.4	cot	2	40	71.58	0.0	0.7816	***
GPT-5.4	temp	1	40	1.929	0.17258	0.046	ns
GPT-5.4	topp	2	40	1.528	0.22939	0.071	ns
GPT-5.4	prompt:cot	6	40	19.093	0.0	0.7412	***
GPT-5.4	prompt:temp	3	40	0.592	0.62373	0.0425	ns
GPT-5.4	prompt:topp	6	40	5.544	0.0003	0.454	***
GPT-5.4	cot:temp	2	40	0.81	0.45216	0.0389	ns
GPT-5.4	cot:topp	4	40	1.235	0.3114	0.1099	ns
GPT-5.4	temp:topp	2	40	3.164	0.053	0.1366	.
GPT-5.4-mini	prompt	3	29	27.811	0.0	0.7421	***
GPT-5.4-mini	cot	2	29	3.666	0.03807	0.2018	*
GPT-5.4-mini	temp	1	29	3.146	0.08663	0.0979	.
GPT-5.4-mini	topp	2	29	4.755	0.01637	0.2469	*
GPT-5.4-mini	prompt:cot	6	29	13.306	0.0	0.7335	***
GPT-5.4-mini	prompt:temp	3	29	1.065	0.37911	0.0992	ns
GPT-5.4-mini	prompt:topp	6	29	9.519	1e-05	0.6632	***
GPT-5.4-mini	cot:temp	2	29	1.746	0.19232	0.1075	ns
GPT-5.4-mini	cot:topp	4	29	2.032	0.11603	0.2189	ns
GPT-5.4-mini	temp:topp	1	29	0.36	0.55337	0.0122	ns
Mistral	prompt	3	68	31.939	0.0	0.5849	***
Mistral	cot	2	68	10.851	8e-05	0.2419	***
Mistral	temp	2	68	8.308	0.00059	0.1964	***
Mistral	topk	2	68	0.162	0.85046	0.0048	ns
Mistral	prompt:cot	6	68	25.086	0.0	0.6888	***
Mistral	prompt:temp	6	68	19.602	0.0	0.6336	***
Mistral	prompt:topk	6	68	0.164	0.98551	0.0142	ns
Mistral	cot:temp	4	68	0.507	0.73085	0.0289	ns
Mistral	cot:topk	4	68	0.562	0.69091	0.032	ns
Mistral	temp:topk	4	68	0.876	0.48305	0.049	ns
Phi4	prompt	3	68	114.848	0.0	0.8352	***
Phi4	cot	2	68	32.001	0.0	0.4849	***
Phi4	temp	2	68	19.298	0.0	0.3621	***
Phi4	topk	2	68	2.667	0.07672	0.0727	.
Phi4	prompt:cot	6	68	42.465	0.0	0.7893	***
Phi4	prompt:temp	6	68	1.783	0.11567	0.1359	ns
Phi4	prompt:topk	6	68	1.629	0.15265	0.1256	ns
Phi4	cot:temp	4	68	0.81	0.52291	0.0455	ns
Phi4	cot:topk	4	68	0.638	0.63742	0.0362	ns
Phi4	temp:topk	4	68	4.462	0.00292	0.2079	**
Llama3	prompt	3	68	6.99	0.00036	0.2357	***
Llama3	cot	2	68	96.669	0.0	0.7398	***
Llama3	temp	2	68	1.375	0.25966	0.0389	ns
Llama3	topk	2	68	1.003	0.37222	0.0286	ns
Llama3	prompt:cot	6	68	16.351	0.0	0.5906	***
Llama3	prompt:temp	6	68	2.604	0.02485	0.1868	*
Llama3	prompt:topk	6	68	1.09	0.37724	0.0877	ns
Llama3	cot:temp	4	68	1.384	0.24865	0.0753	ns
Llama3	cot:topk	4	68	2.216	0.07631	0.1153	.
Llama3	temp:topk	4	68	1.779	0.1431	0.0947	ns

Note. *** $p < .001$, ** $p < .01$, * $p < .05$, . $p < .10$, ns not significant. Top-k/p reported for open-source / GPT models respectively.

Table 11: Factorial ANOVA results for average pairwise word overlap by manipulated factor (Type III sums of squares; prompt, CoT, temperature, and top-k/p plus their two-way interactions).

Model	effect	df_effect	df_resid	F_val	p_val	eta2p	sig
GPT-5.4	prompt	3	40	31.115	0.0	0.7	***
GPT-5.4	cot	2	40	11.974	8e-05	0.3745	***
GPT-5.4	temp	1	40	1.072	0.30671	0.0261	ns
GPT-5.4	topp	2	40	21.216	0.0	0.5147	***
GPT-5.4	prompt:cot	6	40	2.983	0.01672	0.3091	*
GPT-5.4	prompt:temp	3	40	0.521	0.67019	0.0376	ns
GPT-5.4	prompt:topp	6	40	15.526	0.0	0.6996	***
GPT-5.4	cot:temp	2	40	2.565	0.0895	0.1137	.
GPT-5.4	cot:topp	4	40	4.116	0.00693	0.2916	**
GPT-5.4	temp:topp	2	40	73.313	0.0	0.7857	***
GPT-5.4-mini	prompt	3	29	7.998	0.00049	0.4528	***
GPT-5.4-mini	cot	2	29	10.757	0.00032	0.4259	***
GPT-5.4-mini	temp	1	29	0.012	0.91256	0.0004	ns
GPT-5.4-mini	topp	2	29	45.126	0.0	0.7568	***
GPT-5.4-mini	prompt:cot	6	29	13.656	0.0	0.7386	***
GPT-5.4-mini	prompt:temp	3	29	0.479	0.69961	0.0472	ns
GPT-5.4-mini	prompt:topp	6	29	0.999	0.44469	0.1713	ns
GPT-5.4-mini	cot:temp	2	29	0.212	0.81045	0.0144	ns
GPT-5.4-mini	cot:topp	4	29	0.766	0.55622	0.0955	ns
GPT-5.4-mini	temp:topp	1	29	6.201	0.01874	0.1762	*
Mistral	prompt	3	68	5.077	0.00313	0.183	**
Mistral	cot	2	68	2.198	0.11885	0.0607	ns
Mistral	temp	2	68	98.878	0.0	0.7441	***
Mistral	topk	2	68	1.961	0.14856	0.0545	ns
Mistral	prompt:cot	6	68	5.39	0.00013	0.3223	***
Mistral	prompt:temp	6	68	4.292	0.001	0.2747	**
Mistral	prompt:topk	6	68	0.219	0.96934	0.019	ns
Mistral	cot:temp	4	68	1.149	0.34131	0.0633	ns
Mistral	cot:topk	4	68	0.2	0.93772	0.0116	ns
Mistral	temp:topk	4	68	1.856	0.12819	0.0985	ns
Phi4	prompt	3	68	119.591	0.0	0.8407	***
Phi4	cot	2	68	8.935	0.00036	0.2081	***
Phi4	temp	2	68	135.724	0.0	0.7997	***
Phi4	topk	2	68	14.468	1e-05	0.2985	***
Phi4	prompt:cot	6	68	35.412	0.0	0.7576	***
Phi4	prompt:temp	6	68	21.681	0.0	0.6567	***
Phi4	prompt:topk	6	68	0.636	0.70134	0.0531	ns
Phi4	cot:temp	4	68	3.592	0.01023	0.1745	*
Phi4	cot:topk	4	68	0.432	0.78507	0.0248	ns
Phi4	temp:topk	4	68	6.357	0.00021	0.2722	***
Llama3	prompt	3	68	47.755	0.0	0.6781	***
Llama3	cot	2	68	71.184	0.0	0.6768	***
Llama3	temp	2	68	211.561	0.0	0.8615	***
Llama3	topk	2	68	36.741	0.0	0.5194	***
Llama3	prompt:cot	6	68	18.686	0.0	0.6225	***
Llama3	prompt:temp	6	68	22.808	0.0	0.668	***
Llama3	prompt:topk	6	68	2.214	0.05208	0.1634	.
Llama3	cot:temp	4	68	9.051	1e-05	0.3474	***
Llama3	cot:topk	4	68	1.196	0.32065	0.0657	ns
Llama3	temp:topk	4	68	1.529	0.20363	0.0825	ns

Note. *** $p < .001$, ** $p < .01$, * $p < .05$, . $p < .10$, ns not significant. Top-k/p reported for open-source / GPT models respectively.

Table 12: Factorial ANOVA results for type-token ratio by manipulated factor (Type III sums of squares; prompt, CoT, temperature, and top-k/p plus their two-way interactions).

Model	effect	df_effect	df_resid	F_val	p_val	eta2p	sig
GPT-5.4	prompt	3	40	2.231	0.09943	0.1433	.
GPT-5.4	cot	2	40	0.104	0.90189	0.0051	ns
GPT-5.4	temp	1	40	0.646	0.42642	0.0159	ns
GPT-5.4	topp	2	40	10.126	0.00028	0.3361	***
GPT-5.4	prompt:cot	6	40	0.37	0.89371	0.0526	ns
GPT-5.4	prompt:temp	3	40	3.892	0.01568	0.2259	*
GPT-5.4	prompt:topp	6	40	5.569	0.00028	0.4552	***
GPT-5.4	cot:temp	2	40	0.428	0.65448	0.021	ns
GPT-5.4	cot:topp	4	40	0.349	0.84333	0.0337	ns
GPT-5.4	temp:topp	2	40	361.242	0.0	0.9475	***
GPT-5.4-mini	prompt	3	29	6.123	0.00234	0.3878	**
GPT-5.4-mini	cot	2	29	0.251	0.77997	0.017	ns
GPT-5.4-mini	temp	1	29	0.144	0.70714	0.0049	ns
GPT-5.4-mini	topp	2	29	141.011	0.0	0.9068	***
GPT-5.4-mini	prompt:cot	6	29	4.734	0.00179	0.4948	**
GPT-5.4-mini	prompt:temp	3	29	1.252	0.30917	0.1147	ns
GPT-5.4-mini	prompt:topp	6	29	11.15	0.0	0.6976	***
GPT-5.4-mini	cot:temp	2	29	0.107	0.89864	0.0073	ns
GPT-5.4-mini	cot:topp	4	29	8.363	0.00013	0.5356	***
GPT-5.4-mini	temp:topp	1	29	4.822	0.03624	0.1426	*
Mistral	prompt	3	68	3.928	0.01202	0.1477	*
Mistral	cot	2	68	6.242	0.00325	0.1551	**
Mistral	temp	2	68	159.547	0.0	0.8243	***
Mistral	topk	2	68	13.625	1e-05	0.2861	***
Mistral	prompt:cot	6	68	23.806	0.0	0.6775	***
Mistral	prompt:temp	6	68	9.409	0.0	0.4536	***
Mistral	prompt:topk	6	68	1.27	0.2827	0.1008	ns
Mistral	cot:temp	4	68	1.369	0.25374	0.0745	ns
Mistral	cot:topk	4	68	0.947	0.44227	0.0528	ns
Mistral	temp:topk	4	68	18.395	0.0	0.5197	***
Phi4	prompt	3	68	40.953	0.0	0.6437	***
Phi4	cot	2	68	14.473	1e-05	0.2986	***
Phi4	temp	2	68	45.943	0.0	0.5747	***
Phi4	topk	2	68	18.438	0.0	0.3516	***
Phi4	prompt:cot	6	68	29.244	0.0	0.7207	***
Phi4	prompt:temp	6	68	3.149	0.00875	0.2175	**
Phi4	prompt:topk	6	68	25.967	0.0	0.6962	***
Phi4	cot:temp	4	68	0.96	0.4350299999999999	0.0535	ns
Phi4	cot:topk	4	68	0.56	0.69213	0.0319	ns
Phi4	temp:topk	4	68	62.428	0.0	0.786	***
Llama3	prompt	3	68	17.262	0.0	0.4323	***
Llama3	cot	2	68	13.573	1e-05	0.2853	***
Llama3	temp	2	68	67.618	0.0	0.6654	***
Llama3	topk	2	68	13.116	2e-05	0.2784	***
Llama3	prompt:cot	6	68	8.094	0.0	0.4166	***
Llama3	prompt:temp	6	68	2.939	0.0131	0.2059	*
Llama3	prompt:topk	6	68	3.344	0.00603	0.2278	**
Llama3	cot:temp	4	68	0.203	0.93582	0.0118	ns
Llama3	cot:topk	4	68	0.944	0.44403	0.0526	ns
Llama3	temp:topk	4	68	53.211	0.0	0.7579	***

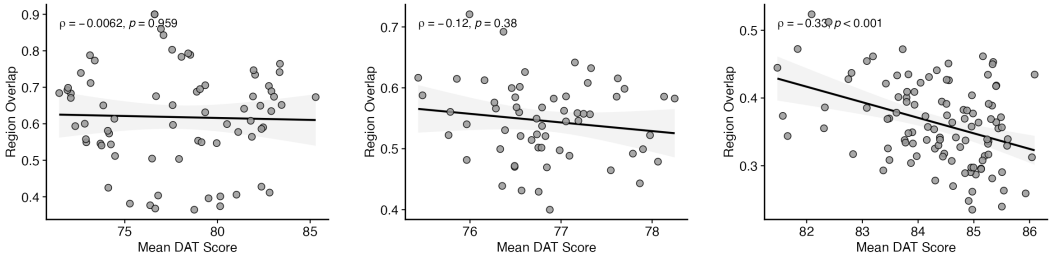
Note. *** $p < .001$, ** $p < .01$, * $p < .05$, . $p < .10$, ns not significant. Top-k/p reported for open-source / GPT models respectively.

Table 13: Factorial ANOVA results for normalized entropy by manipulated factor (Type III sums of squares; prompt, CoT, temperature, and top-k/p plus their two-way interactions).

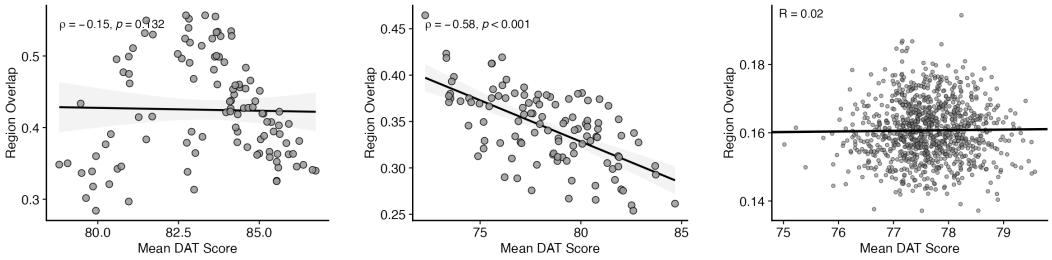
Model	effect	df_effect	df_resid	F_val	p_val	eta2p	sig
GPT-5.4	prompt	3	40	0.985	0.40951	0.0688	ns
GPT-5.4	cot	2	40	0.772	0.46892	0.0372	ns
GPT-5.4	temp	1	40	0.332	0.56744	0.0082	ns
GPT-5.4	topp	2	40	1.184	0.31655	0.0559	ns
GPT-5.4	prompt:cot	6	40	3.108	0.01356	0.318	*
GPT-5.4	prompt:temp	3	40	0.165	0.91948	0.0122	ns
GPT-5.4	prompt:topp	6	40	2.112	0.07313	0.2406	.
GPT-5.4	cot:temp	2	40	0.376	0.68897	0.0185	ns
GPT-5.4	cot:topp	4	40	1.202	0.32488	0.1073	ns
GPT-5.4	temp:topp	2	40	5.722	0.00652	0.2225	**
GPT-5.4-mini	prompt	3	29	4.716	0.00844	0.3279	**
GPT-5.4-mini	cot	2	29	11.942	0.00016	0.4516	***
GPT-5.4-mini	temp	1	29	0.07	0.79356	0.0024	ns
GPT-5.4-mini	topp	2	29	7.432	0.00248	0.3389	**
GPT-5.4-mini	prompt:cot	6	29	10.299	0.0	0.6806	***
GPT-5.4-mini	prompt:temp	3	29	0.826	0.4902	0.0787	ns
GPT-5.4-mini	prompt:topp	6	29	0.951	0.4746	0.1645	ns
GPT-5.4-mini	cot:temp	2	29	0.897	0.41864	0.0583	ns
GPT-5.4-mini	cot:topp	4	29	3.763	0.01386	0.3417	*
GPT-5.4-mini	temp:topp	1	29	3.158	0.08603	0.0982	.
Mistral	prompt	3	68	6.17	0.0009	0.214	***
Mistral	cot	2	68	0.626	0.53775	0.0181	ns
Mistral	temp	2	68	96.539	0.0	0.7395	***
Mistral	topk	2	68	0.464	0.63098	0.0135	ns
Mistral	prompt:cot	6	68	7.906	0.0	0.4109	***
Mistral	prompt:temp	6	68	3.444	0.00498	0.2331	**
Mistral	prompt:topk	6	68	0.153	0.98774	0.0134	ns
Mistral	cot:temp	4	68	1.731	0.15323	0.0924	ns
Mistral	cot:topk	4	68	0.041	0.99671	0.0024	ns
Mistral	temp:topk	4	68	8.401	1e-05	0.3307	***
Phi4	prompt	3	68	60.395	0.0	0.7271	***
Phi4	cot	2	68	4.257	0.01811	0.1113	*
Phi4	temp	2	68	154.674	0.0	0.8198	***
Phi4	topk	2	68	3.463	0.03696	0.0924	*
Phi4	prompt:cot	6	68	22.885	0.0	0.6688	***
Phi4	prompt:temp	6	68	1.741	0.12474	0.1332	ns
Phi4	prompt:topk	6	68	6.905	1e-05	0.3786	***
Phi4	cot:temp	4	68	2.501	0.05043	0.1282	.
Phi4	cot:topk	4	68	0.751	0.56082	0.0423	ns
Phi4	temp:topk	4	68	24.819	0.0	0.5935	***
Llama3	prompt	3	68	34.12	0.0	0.6008	***
Llama3	cot	2	68	48.437	0.0	0.5876	***
Llama3	temp	2	68	181.827	0.0	0.8425	***
Llama3	topk	2	68	4.624	0.0131	0.1197	*
Llama3	prompt:cot	6	68	18.88	0.0	0.6249	***
Llama3	prompt:temp	6	68	6.32	3e-05	0.358	***
Llama3	prompt:topk	6	68	1.759	0.12082	0.1343	ns
Llama3	cot:temp	4	68	7.816	3e-05	0.3149	***
Llama3	cot:topk	4	68	0.964	0.43278	0.0537	ns
Llama3	temp:topk	4	68	44.241	0.0	0.7224	***

Note. *** $p < .001$, ** $p < .01$, * $p < .05$, . $p < .10$, ns not significant. Top-k/p reported for open-source / GPT models respectively.

A.2 GRAPHS

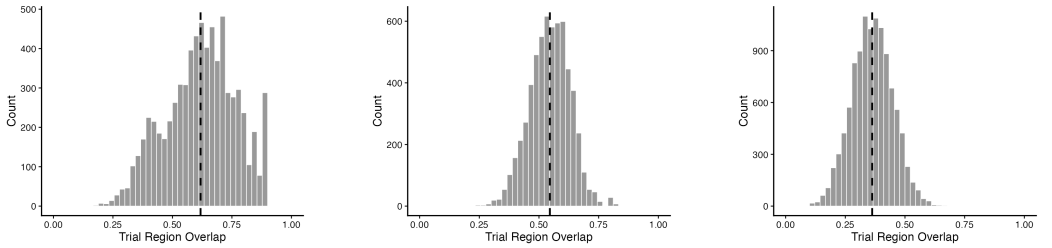


(a) GPT-5.4 (b) GPT-5.4-mini (c) Llama3

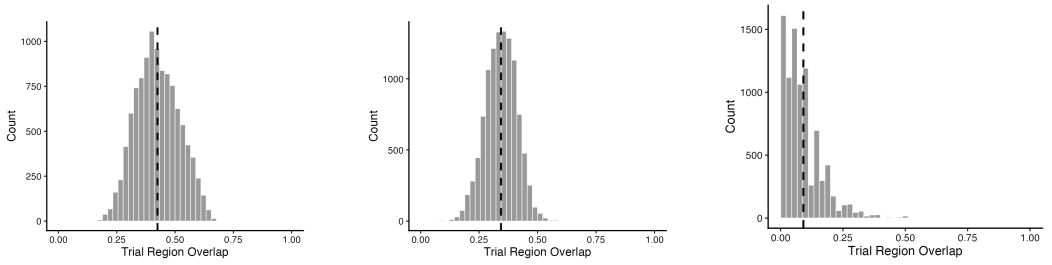


(d) Mistral (e) Phi4 (f) Human

Figure 6: DAT score vs. semantic region overlap across all conditions. Regions defined by k-means clustering ($k = 100$) of GloVe vectors. Human panel: 1,000 bootstrap samples.

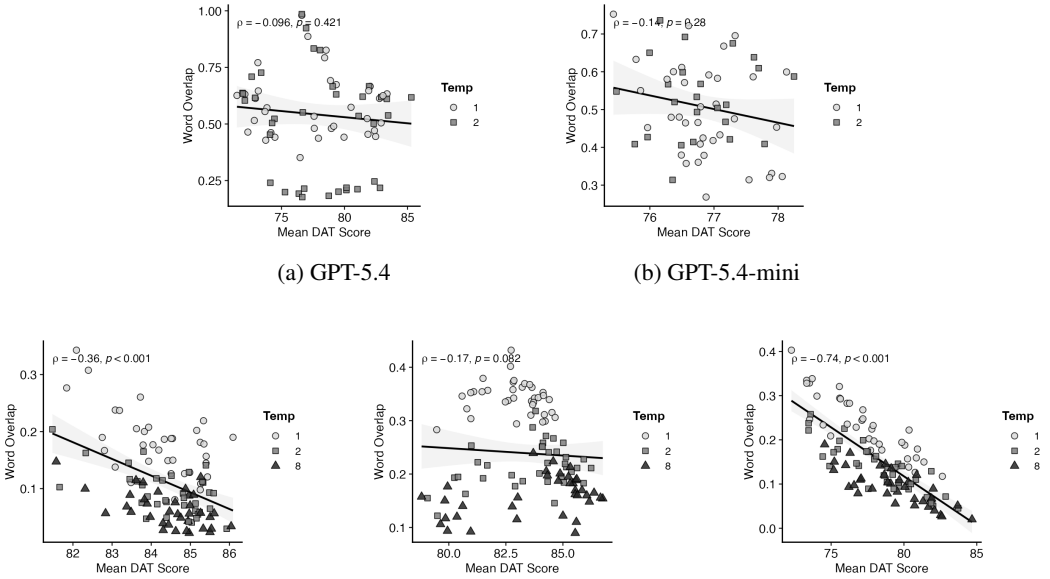


(a) GPT-5.4 (b) GPT-5.4-mini (c) Llama3



(d) Mistral (e) Phi4 (f) Human

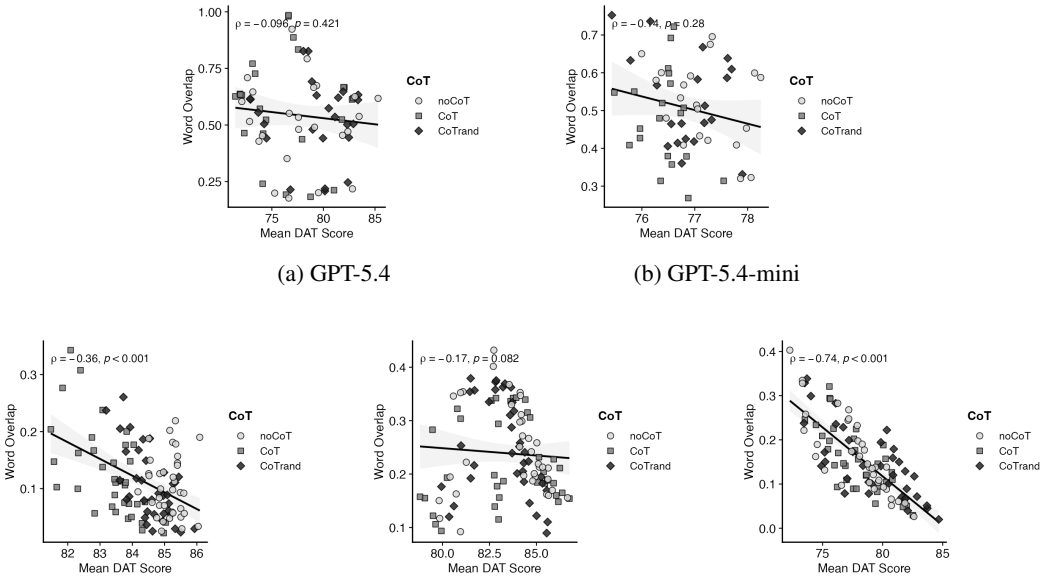
Figure 7: Distribution of per-trial semantic region overlap for each model. Dashed line = mean. Human panel = individual human trials (non-bootstrapped).



(a) GPT-5.4 (b) GPT-5.4-mini

(c) Llama3 (d) Mistral (e) Phi4

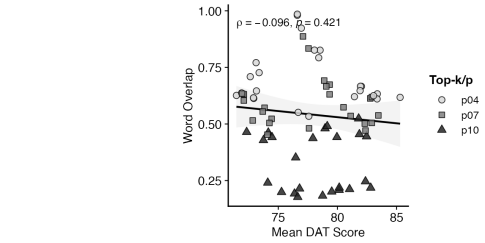
Figure 8: DAT score vs. word overlap, denoted by temperature.



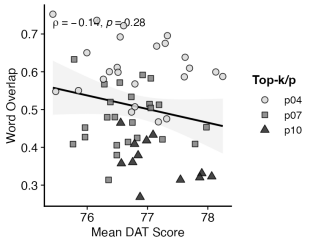
(a) GPT-5.4 (b) GPT-5.4-mini

(c) Llama3 (d) Mistral (e) Phi4

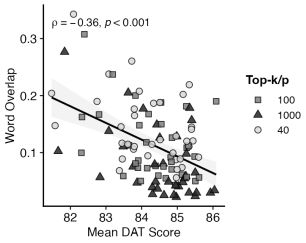
Figure 9: DAT score vs. word overlap, denoted by CoT condition.



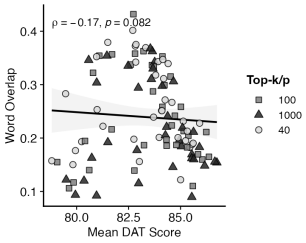
(a) GPT-5.4



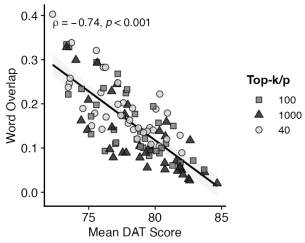
(b) GPT-5.4-mini



(c) Llama3

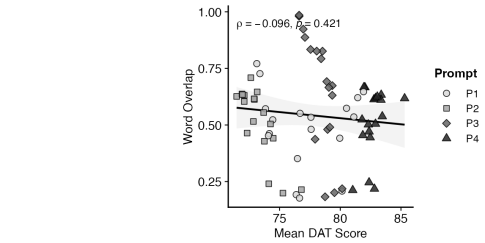


(d) Mistral

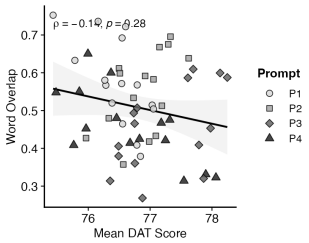


(e) Phi4

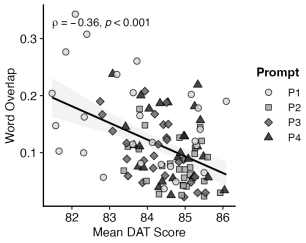
Figure 10: DAT score vs. word overlap, denoted by top-k/top-p.



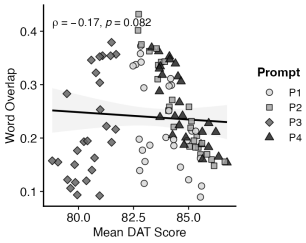
(a) GPT-5.4



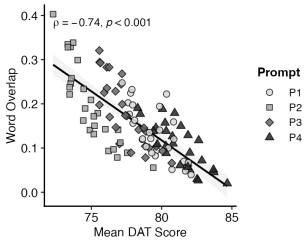
(b) GPT-5.4-mini



(c) Llama3

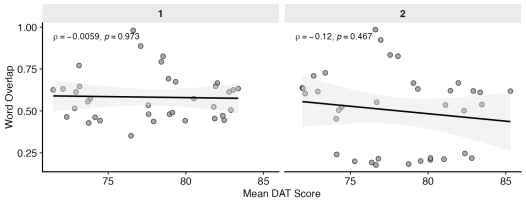


(d) Mistral

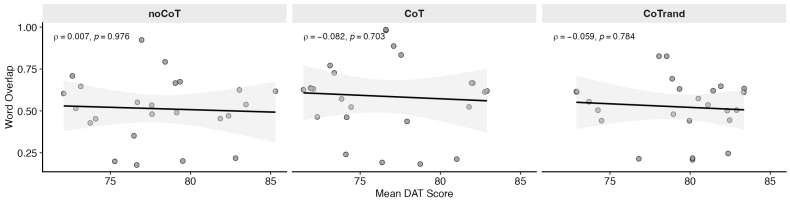


(e) Phi4

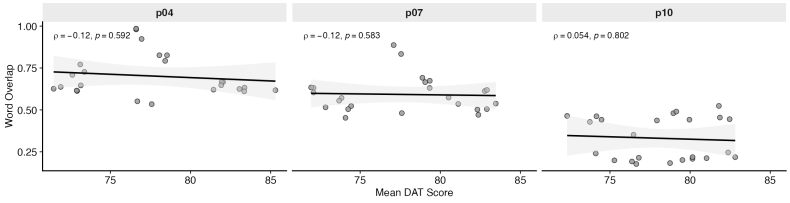
Figure 11: DAT score vs. word overlap, denoted by prompt variant.



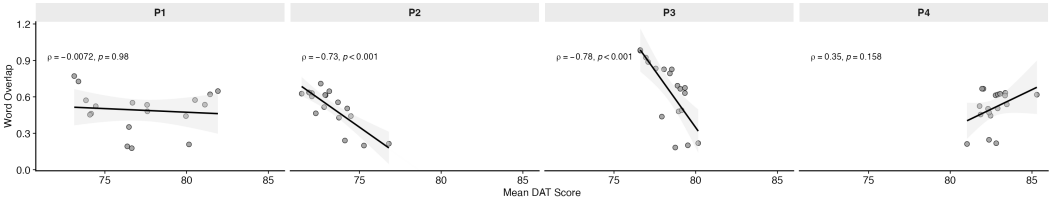
(a) By temperature



(b) By CoT condition

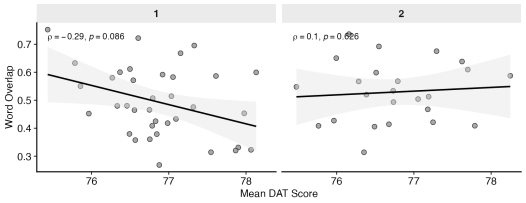


(c) By top-p

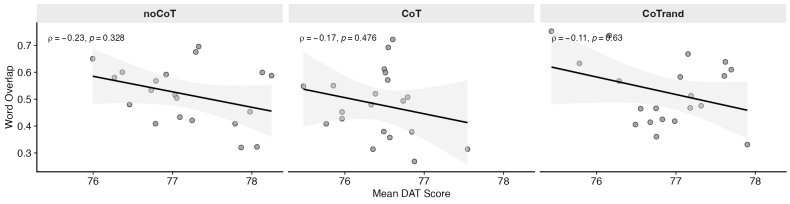


(d) By prompt variant

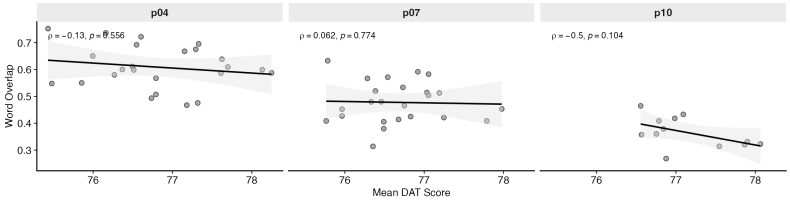
Figure 12: GPT-5.4: DAT score vs. word overlap, faceted by each predictor.



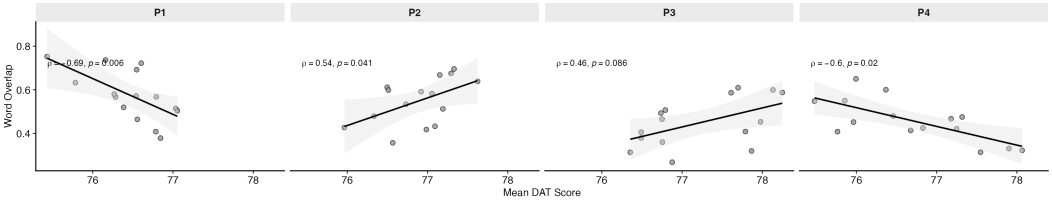
(a) By temperature



(b) By CoT condition

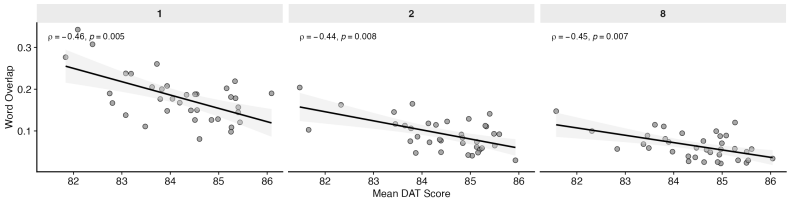


(c) By top-p

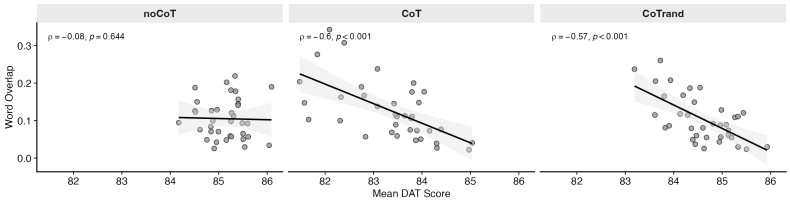


(d) By prompt variant

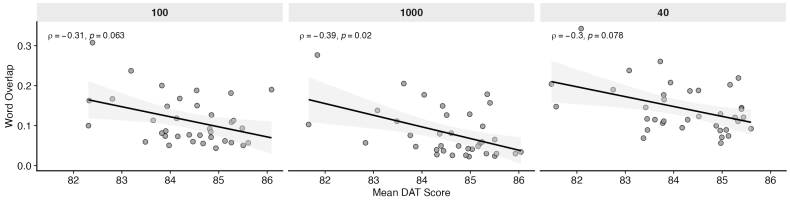
Figure 13: GPT-5.4-mini: DAT score vs. word overlap, faceted by each predictor.



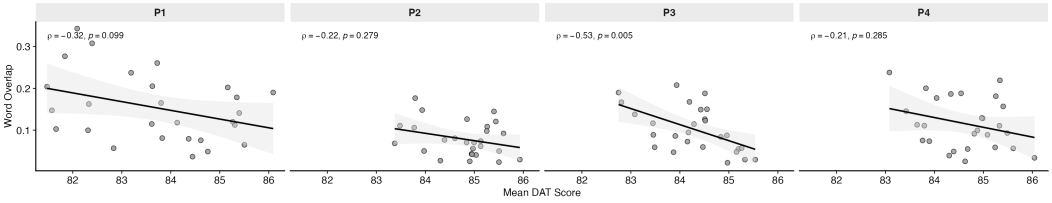
(a) By temperature



(b) By CoT condition

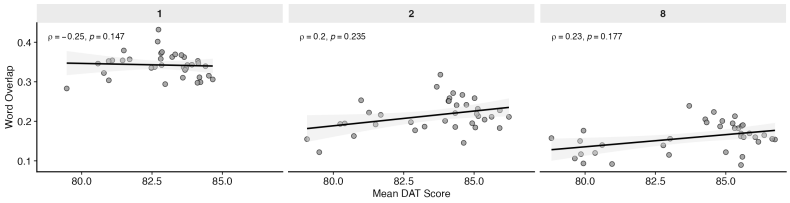


(c) By top-k

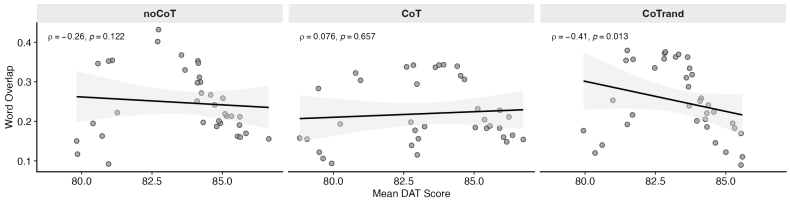


(d) By prompt variant

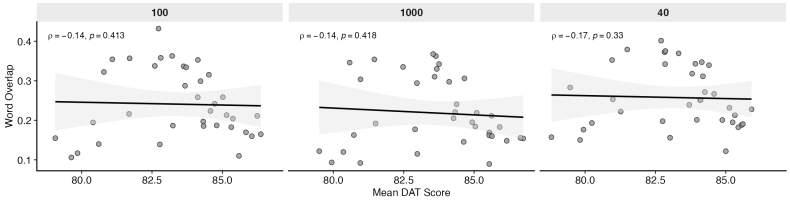
Figure 14: Llama3: DAT score vs. word overlap, faceted by each predictor.



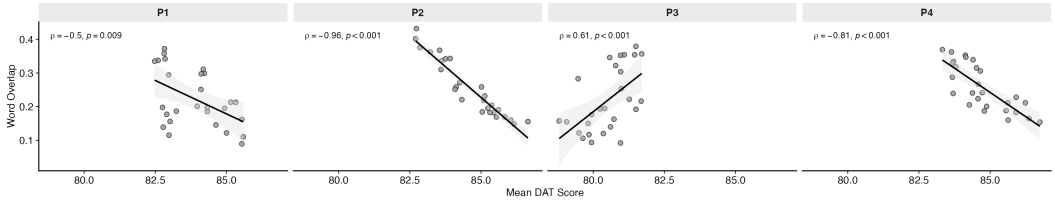
(a) By temperature



(b) By CoT condition

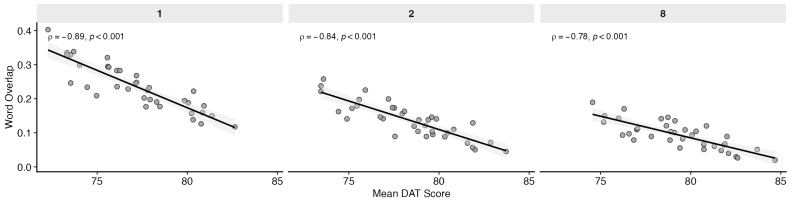


(c) By top-k

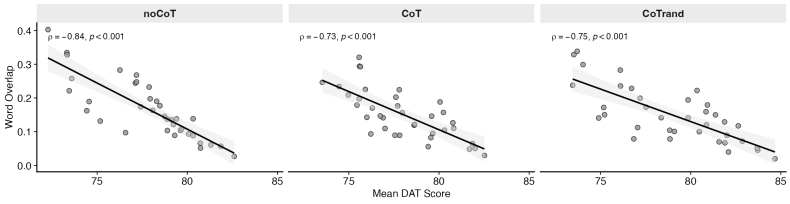


(d) By prompt variant

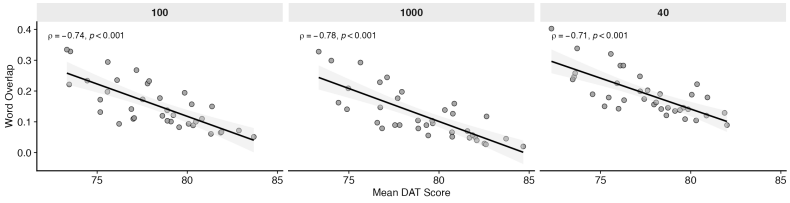
Figure 15: Mistral: DAT score vs. word overlap, faceted by each predictor.



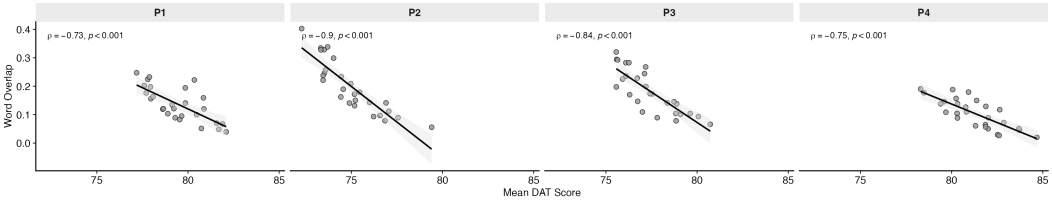
(a) By temperature



(b) By CoT condition



(c) By top-k



(d) By prompt variant

Figure 16: Phi4: DAT score vs. word overlap, faceted by each predictor.

A.3 PROMPT AND COT CONDITIONS

Table 14: Prompt variants (P1−P4) and chain-of-thought (CoT) prefixes used across all experimental conditions. CoT prefixes were prepended to the core prompt text; the NOCOT condition used no prefix.

Condition	Text
Core prompts	
P1	Please enter 10 words that are as different from each other as possible, in all meanings and uses of the words. Only single words in English. Only nouns. No proper nouns. No specialised vocabulary. Reply with exactly 10 words, one per line, numbered 1 to 10. No extra explanation.
P2	Provide a list of 10 English nouns that are maximally distinct from one another across all possible meanings and uses. Single words only. No proper nouns. No technical terms. Reply with exactly 10 words, one per line, numbered 1 to 10. No extra explanation.
P3	Select 10 nouns in English that differ from each other as much as possible in meaning and use. Each entry must be a single word. Avoid proper nouns and technical vocabulary. Reply with exactly 10 words, one per line, numbered 1 to 10. No extra explanation.
P4	Give 10 English nouns that are as semantically distant from each other as possible. Single words only. No proper nouns or technical terms. Reply with exactly 10 words, one per line, numbered 1 to 10. No extra explanation.
Chain-of-thought prefixes (prepended to P1–P4)	
NOCOT	*(no prefix)*
COT	Think step by step about what makes words semantically different, then…
COTRAND	First, generate a random number between 1 and 1000 (do not report it). Use it to guide your thinking, then…

A.4 LICENSES FOR EXISTING ASSETS

GloVe pretrained vectors (glove.840B.300d) (Pennington et al., 2014) are distributed under the Public Domain Dedication and License (PDDL) 1.0.

Llama3-8B is distributed under the Meta Llama 3 Community License Agreement (April 18, 2024).

Mistral-7B is distributed under the Apache License 2.0.

Phi4-14B is distributed under the MIT License.

GPT-5.4 and GPT-5.4-mini were accessed via the OpenAI API under OpenAI's API Terms of Use; no model weights were obtained or redistributed.

Human baseline data (Olson et al., 2021) are hosted on the Open Science Framework (`https://osf.io/vjazn/`) without a formal attached license and are used for reanalysis and comparison as reported in the original publication.