ON THE DIVERSITY OF ANALOGY MAKING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable potential for analogy making, a core cognitive capability that drives novelty and creativity. While prior research has extensively investigated the applications and underlying mechanisms of LLM-based analogy making, its output diversity remains largely unexplored, despite being essential for broadening cross-domain connections and fostering scientific innovation. In this work, we present a comprehensive evaluation of analogy diversity across ten state-of-the-art open- and closed-source LLMs. Our findings highlight a concerning issue of domain homogeneity, a prevalent tendency for LLMs to generate analogies from a narrow set of target domains, limiting both inter-query and intra-model diversity. Furthermore, our analysis reveals a fundamental trade-off in existing LLM diversity-enhancement methods: increasing output diversity often comes at the expense of output quality. Finally, our causal analysis of LLM internals uncovers substantial differences in the model-sensitive regions governing analogy diversity across LLMs, suggesting a potential mechanistic basis for the observed diversity–quality trade-off. To our knowledge, this is among the first studies to systematically investigate output diversity in LLM-based analogy making. Our findings provide empirical insights and practical guidance for developing more diverse LLMs for analogy generation. Our code are available at `[url_placeholder]`.

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have demonstrated remarkable potential to accelerate scientific discovery, ushering in a new era of AI-for-Science (AI4S). Beyond their broad range of applications in domains such as protein design Fry et al. (2026); Liu et al. (2025), sustainable energy Hong et al. (2026); Zhang et al. (2026), materials science Park & Walsh (2026), and interdisciplinary knowledge discovery Shen et al. (2026b), recent research has begun to explore LLMs' emerging capacity to generate diverse and creative analogies across scientific domains, a cognitive capability central to human intelligence that sparks scientific innovation Shen et al. (2026a).

In AI4S, where scientific innovation is driven by interdisciplinary knowledge integration, analogy diversity strengthens LLMs' ability to uncover both explicit and implicit connections across domains and disciplines Shen et al. (2026b). A classic historical example is James Clerk Maxwell's development of electromagnetic theory Clerk Maxwell (1864). Maxwell employed fluid and mechanical systems as the **source domain** to reason about electromagnetic phenomena, the **target domain**. Although these domains are conceptually distant, their structural correspondence enabled a breakthrough in scientific understanding, highlighting the importance of diverse analogies in driving innovation, as illustrated in Figure 1.

Motivated by this observation, a fundamental question is whether current LLMs can generate the diverse analogies needed to foster cross-domain knowledge transfer and scientific innovation. Although LLM-based analogy making has received increasing attention, with existing research primarily focusing on analogy quality, applications, and underlying mechanisms Shen et al. (2026a), the diversity of generated analogies remains largely unexplored. Different from conventional generation tasks, analogy-making explicitly requires a departure from the source domain and a bold selection of the target domain, where both source and target domains share similar functional abstractions

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.

Gentner (1983); Turney (2012). The overlook of diversity in analogy making further hinders a comprehensive understanding in LLM analogy-making mechanisms, leaving a notable gap to this central research question: **Are LLMs capable of *generating* diverse and informative analogies?**

In this work, we provide a thorough evaluation of analogy-making diversity across multiple LLMs from two complementary perspectives. First, we evaluate how well mainstream LLMs are able to generate analogical sentences by introducing novel and multi-faceted metrics that focus on domain selection diversity. Our experimental results show that, LLMs tend to generate analogous sentences closely aligned with similar target domains, exhibiting a homogeneity trend across different model families. The excessive homogeneity in analogy generation poses a fundamental limitation when these models are applied to scientific discovery, a process that inherently relies on diversity and creativity.

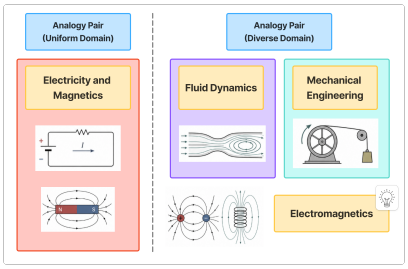

Figure 1: Maxwell's development of the theory of electromagnetism through analogical reasoning across diverse domains. The left subfigure illustrates analogies drawn from closely related domains, while the right subfigure shows more diverse cross-domain analogies. This example demonstrates how greater analogy diversity enables broader knowledge integration and can foster scientific discovery, motivating our study of analogy diversity in large language models.

Second, to better understand the effect of existing diversity-enhancing methods under an analogy making setting, we evaluate approaches that span three categories of inference-time perturbation, including (i) diversity prompting based methods, (ii) probability and logits steering, and (iii) entropy-gated steering. Our experimental results show that existing methods provide only limited improvements in analogy diversity and are often accompanied by the sacrifice of analogy quality. The trade-off between analogy diversity and output quality suggests that techniques originally designed for general-purpose generation may not readily address domain homogeneity in analogy-making tasks. Additional analysis on LLMs' internal layers further suggests that such inefficacy might be caused by a lack of a consistent region of analogy making across different models.

To the best of our knowledge, our work is among the first to systematically study output diversity under the context of LLM analogy generation. We summarize our main contributions as follows:

- **Comprehensive evaluation of LLM analogy making diversity.** Our work provides a thorough evaluation of diversity in LLMs' analogy making, using a novel and multi-faceted metrics that provide detailed insights into the output of LLM generations.
- **Domain homogeneity in LLM analogy making.** Our results reveal domain homogeneity in analogy making of these models based on both intra-query and intra-model perspective, highlighting an alarming limitation in the promotion of creativity and diversity when applying LLMs in scientific innovation.
- **Diversity-quality tradeoff in LLM analogy making.** Further analysis on the efficacy of applying diversity enhancing frameworks in LLM analogy making shows a trade-off between output diversity and quality. We also identify notable differences in perturbation sensitivity in different models, showing a divergence in models' internal behaviors in analogy making.

2 Related Works

Diversity in Generation. Generation diversity in Large Language Models (LLMs) has been well-studied, especially in the realm of applications in creative writing Chung et al. (2025). Prior work has demonstrated the importance of diversity in applications such as short story completion Tian et al. (2024) and creative writing Ding et al. (2026), where creative variation and exploration of alternative narratives are fundamental to high-quality generation. To improve diversity, early works aim to solve the mode collapse phenomenon in LLMs Jiang et al. (2025), where models often generate repetitive, bland, or generic outputs. Moreover, frameworks such as beam search Wu et al. (2016),

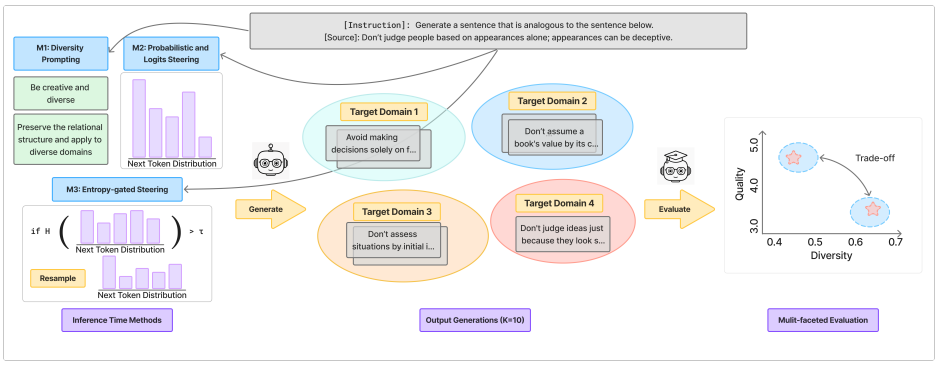

Figure 2: Overview of experiments and evaluation methods on LLM's analogy making diversity. The model is given the input of the format `[Instruction, source]` along with different inference time methods in the left column. We then cluster the output generations into different target domains and conduct multi-faceted evaluations on those generations.

Top-k and Nucleus (Top-p) sampling Fan et al. (2018); Holtzman et al. (2020) are proposed to mitigate the issue. Beyond inference-time strategies to promote diversity, metrics such as *Distinct-n* Li et al. (2016) and *Self-BLEU* Zhu et al. (2018) have become standard for measuring distributional breadth, while recent efforts have begun to emphasize "effective semantic diversity" to ensure that variety does not come at the cost of task-specific quality Wiher et al. (2022). However, these works often address surface-level or stylistic diversity in creative writing, leaving a sparse space for the exploration of outcome diversity in cognitive tasks such as analogical reasoning.

Analogical Reasoning in LLMs. Analogical reasoning is deemed a central faculty of human intelligence Minegishi et al. (2026); Hofstadter & Sander (2013). Modern methods that study human analogy often refer to the process of abstraction and structure mapping that make hops across different domains Gentner (2010). The recent burgeoning of mechanistic interpretability methods Huben et al. (2024); Belrose et al. (2023); Ghandeharioun et al. (2024) has fueled studies on LLM analogical reasoning abilities from an anatomical perspective, offering a shift towards a more comprehensive understanding within LLMs than of their human counterparts. In light of pioneering works that treat analogy-making as a dual interplay between domain and function Gentner (1983), recent works make remarkable progress ranging from model behavioral studies Webb et al. (2023) to efforts that explore underlying mechanisms, with proposals of certain hypotheses such as structure mapping Gentner (1983) and graph functor Minegishi et al. (2026). However, to our knowledge, all of these explorations are conducted under a classification setup, which often involves comparisons between analogous pairs, thus causing these efforts either to remain relatively superficial or to bear the risk of obliterating the expansive nature of the solution space in analogy-making, as argued in the book Hofstadter & Sander (2013). To bridge this gap, our work moves beyond comparing binary labels to the open-ended generation analysis in LLM analogical reasoning.

3 Experiment Setup

Datasets. To provide a comprehensive evaluation of diversity in LLM-based analogy, we conduct experiments on three recently released analogy generation datasets. AnaloBench (AB) Ye et al. (2024) is a benchmark dedicated to story-level analogy evaluation that contains 340 pairs of high-quality analogous stories from human annotators. Metaphoric Analogies (MA) Boisson et al. (2025) focuses on extracting structured analogical mappings from literary metaphors with 203 samples that build source and target concept pairs in a four-term analogy structure. Metaphor Understanding Challenge (MUNCH) Tong et al. (2024) evaluates whether LLMs can understand metaphors as cross-domain mappings through paraphrasing, providing over 10K paraphrases for sentences containing metaphor use. We use the full evaluation dataset for both AB and MA; for MUNCH, we extract 1K samples randomly.

Models and Devices. We include five closed-source and five open-source models in our evaluation. We use `GPT-5.2`, `Grok-4.5`, `gemini-2.5-pro`, `gemini-3.1-pro`, and `claude-4.6-sonnet` for generation diversity evaluation. We further test five open source models, including `llama-3.1-8B-instruct`, `gemma2-9B-it`, `qwen-3-8B`, `mistral-7B-Instruct`, and `phi-4-mini-instruct`. The maximum generation length is set to 100 output tokens for all open source models. We run our experiments using single `l40s` and `h100` GPUs.

Generation Perturbation Methods. We evaluate a spectrum of inference-time perturbation methods following the categories defined in Ostermann et al. (2026) to provide a comprehensive understanding of their efficacy in LLM analogy making.

- **Diversity Prompting.** In addition to base prompting, we run diversity prompting strategies, including Logic prompts and Diversity prompts to explicitly encourage analogies from creative and diverse domains.
- **Probabilistic and Logits Steering.** We run probabilistic methods including Top-k ($k = 20$), Top-p ($p = 0.9$), Top-$\eta\sigma$ ($\eta\sigma = 1.0$), and Min-p ($p = 0.1$) in our experiments that resample the output logits to improve output diversity. We also include G2 Ruan et al. (2025) in our experiments, where we set the number of iteration rounds to be 3.
- **Entropy-gated Steering.** We also include an adaptation from Li et al. (2026) that uses entropy as an indicator to control the output diversity. We set the resampling threshold τ to be larger than the 90th percentile of the whole generated token sequence.

Output Evaluations. We run our evaluation from $K = 10$ generations for each sample. Following previous studies that evaluate generation diversity by mapping natural language into a semantic space Reimers & Gurevych (2019), we adopt a similar paradigm. Specifically, we employ the `all-MiniLM-L6-v2` sentence transformer as an encoder to obtain semantic embeddings, calculating their mean pairwise cosine distance. Furthermore, we incorporate the MAUVE score Pillutla et al. (2022) to better capture distributional differences, providing a more comprehensive analysis of generation over quality and diversity. To evaluate the coherence and quality of the generated analogies, we use `GPT-4o-mini` under an existing LLM-as-a-judge setup Zheng et al. (2023).

4 Domain Homogeneity in Analogy Generation

We first conduct a study in diversity of models' analogy-generation capabilities, focusing on whether the generated analogous sentences fall into diverse target domains. We assess this from two perspectives: i) Intra-output analogy diversity, which measures the pairwise and clustering distances among generations within a single model, and ii) Inter-model analogy diversity measures, which assess whether analogy generation converges to the same target domain or commonalities in sentence structure across different models.

4.1 Intra-Output Diversity

To obtain intra-output diversity for each specific model, following similar evaluation for general diversity Ruan et al. (2025); Ding et al. (2026), we first measure the pairwise distances among the distinct generations to show how closely related each generation is to the others. Specifically, we assemble prompts in the form of a binary tuple `[Instruction, Source]`, where the `Instruction` field asks the model to generate analogical sentences under *base*, *diverse*, or *logical* prompting strategies, and the `Source` field contains a sentence where the model should identify its source domain and generate analogies thereafter. Under the *base* setting, the model is simply instructed to generate an analogy. In the *diverse* setting, the model is additionally encouraged to produce creative and diverse outputs. Finally, in the *logical* setting, the model is instructed to preserve the same relational structure while generating analogies in different target domains, without explicitly encouraging diversity.

Figure 3 (a) presents the pairwise cosine distances between generation embeddings from two perspectives: (i) solid lines (less saturated colors), where embeddings are computed from the complete generated analogies, and (ii) dashed lines, where embeddings are computed only from the

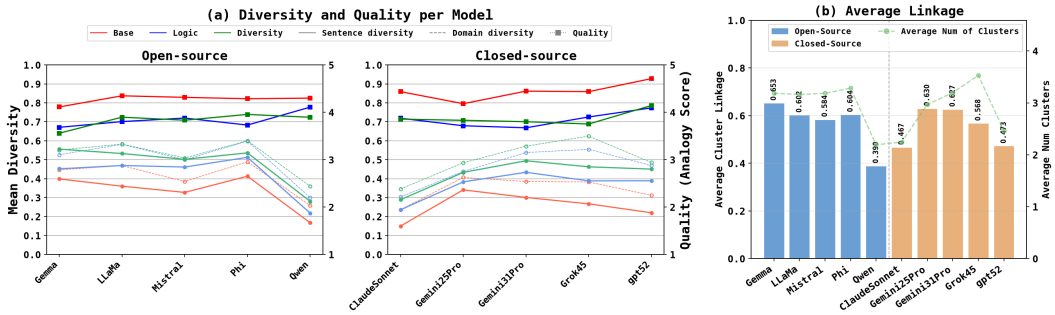

Figure 3: (a) Comparison of sentence diversity, domain diversity, and output quality across prompting strategies. Solid lines denote sentence diversity, dashed lines denote domain diversity, and square markers indicate quality scores for open-source and closed-source models. (b) Inter-domain diversity using average linkage. Bars denote average linkage; dashed line indicates mean clusters per sample.

extracted words describing the target domains. For the latter, target domains are extracted using `GPT-4o-mini`. To evaluate the quality of the generated analogies, following Zheng et al. (2023), we employ an LLM-as-a-judge approach, with the results also reported in the figure with solid lines and saturated colors.

Our results show intuitive behavior of the tested models: when explicitly instructed to be diverse, the models generate more creative outputs, while prompts emphasizing logical behavior also improve diversity compared to naive prompting strategies (less saturated colors in Figure 3 (a). However, the quality tends to be better for base prompts (red saturated color). Comparing closed-source and open-source models, there is no clear evidence of superiority between them, showing that more complex post-training may not necessarily improve diversity. Nevertheless, even among the mainstream closed-source models evaluated, the best-performing model for the diversity prompt (less saturated solid line in the figure), `Gemini3.1-Pro`, achieves only an average intra-output diversity of 0.49, indicating that overall diversity remains limited. Considering only the extracted target domains, where irrelevant lexical content is removed and the evaluation focuses solely on the target domains, all models exhibit improved diversity, showing a similar standard among the prompts. Nevertheless, the overall diversity remains limited, with the mean pairwise diversity not exceeding 0.63.

Except for the target domain extraction step, the evaluation shown in Figure 3 (a) follows a standard approach described in several works for assessing diversity in text generation. However, a key limitation of this evaluation is that it remains difficult to interpret whether an average similarity score of 0.6 truly reflects high or low diversity. To provide a complementary perspective, we employ `GPT-4o-mini` to identify the semantic target domains of the generated analogies and group the outputs into distinct clusters. These clusters enable us to assess diversity at the semantic level by considering both the number of discovered clusters and the hierarchical Average Linkage, i.e., the average similarity over all pairs of sentences belonging to different clusters [1], where a greater number of clusters and higher Average Linkage values indicate greater diversity. Similar clustering-based strategies have been adopted in other contexts to evaluate output diversity Pillutla et al. (2022).

Figure 3 (b) presents the clustering results obtained using the diverse prompt across all evaluated models. Despite explicitly instructing the models to generate 10 distinct and diverse analogies, we observe an average of only two semantic clusters per model. Among the evaluated models, `Grok-4.5` produces the highest diversity, with an average of 3.5 clusters, whereas `Qwen3-8B` generates fewer than two clusters on average. Regarding the linkage analysis, both closed-source and open-source models exhibit comparable Average Linkage values (0.39–0.65), indicating similar levels of semantic separation between target domains. With an average of three clusters, `Gemma-2-9B` achieves the highest Average Linkage ($\overline{\text{avg}} = 0.653$), followed by the Gemini family among the closed-source models ($\overline{\text{avg}} = 0.628$). In contrast, `Qwen3-8B` obtains the lowest score ($\overline{\text{avg}} = 0.390$). To assess the reliability of the results in Figure 3, we conducted a human evaluation

[1] We also evaluated Single Linkage and Complete Linkage and obtained similar results.

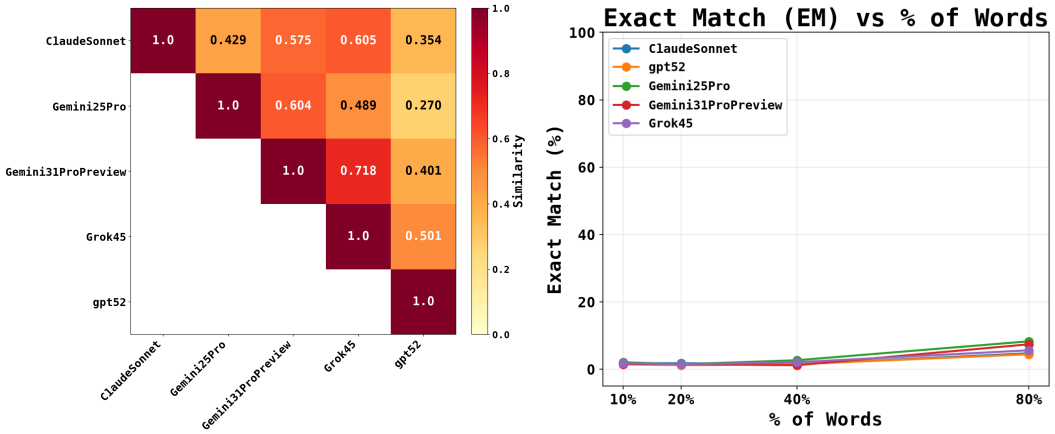

Figure 4: (a) Heatmap for Inter-Model MAUVE scores. MAUVE is originally designed to measure how closely machine-generated text matches human-written text. We adapt MAUVE to compare how closely related the generated texts are across different close source models. (b) Analysis on data pollution. We use the method from Golchin & Surdeanu (2024) that makes varied percentage of context and compute the recovery accuracy with the original sentence.

of the generated clusters. Based on a randomly sampled subset comprising 10% of the data, the human annotations agreed with the automatic clustering in 95% of the cases.

These findings reveal a notable tendency toward target-domain homogeneity in analogy generation, raising concerns about the capability of current LLMs to consistently produce diverse analogical mappings. For instance, when the model is provided with a `context` sentence starting with "Don't judge a person...", the model has strong preferences for repeating similar target domains such as "book" or "quiet river", with the variety not following a uniform distribution for K analogy generations.

4.2 Inter-Model Diversity

In addition to domain homogeneity within a single model, we further explore the diversity across different models. Using the diverse prompt settings of the previous section, Figure 4 (a) showcases the inter-model diversity comparison heatmap reported using adapted MAUVE scores Pillutla et al. (2022), where a higher score means more similar and less diverse. As described in Figure 4 (a), although some model pairs exhibit different domain selection preferences (`gemini-2.5-pro` v.s. both `gpt-5.2` and `claude-4.6-sonnet`), the inter-model MAUVE scores remain high (less diverse) for the remaining models in the comparison. Notably, when comparing the generation distribution between `grok-4.5` and `gemini-3.1-pro-preview`, we obtain a MAUVE score of 0.72, indicating an overwhelming overlap in the generation patterns between these models.

4.3 Data Pollution

One potential issue that poses a threat to the accuracy of our comparison is data pollution, i.e., the model has already memorized the corresponding analogous pairs given the `source`. To tackle this challenge, we apply the method introduced by Golchin & Surdeanu (2024) to ensure there is no data pollution issue. Specifically, for each sample in our dataset, we mask out words at different percentage levels and check whether the model is able to recover the complete sentence measured by token exact match. Figure 4 (b) reports the results in our analysis. It can be seen that all the models in our experiment are **not** suffering from a data pollution issue, as the highest exact match score of 14% is achieved by `gemini-3.1-pro-preview` at the 80% context level. This means that even the best-performing model only recovers a small portion of the original sentence when 80% of the context is provided. The analysis of data pollution further enhances the validity of our observation regarding domain homogeneity.

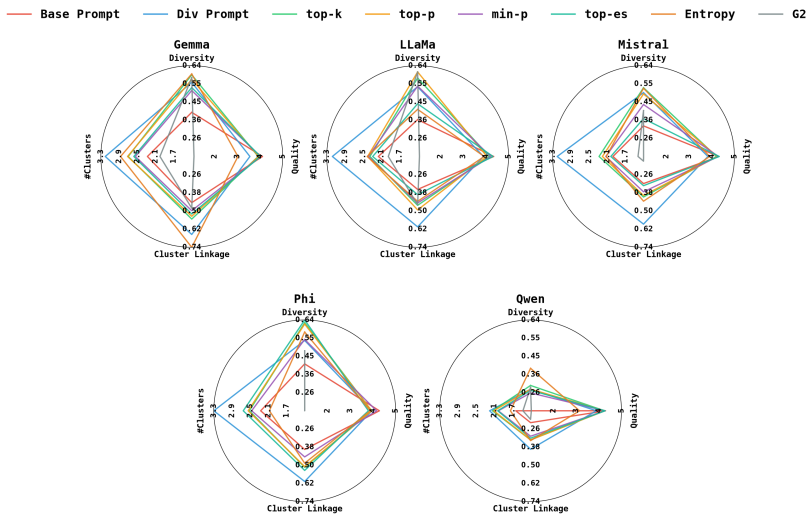

Figure 5: Radar chart comparing 8 decoding methods across 5 LLMs. Each model is shown in its own polar plot with 4 metrics (Diversity, Cluster Linkage, Quality, #Clusters) normalized to 0,1 via min-max scaling (quality capped at 1,5). Tick labels on each spoke indicate the original-scale values.

5 Evaluation of Diversity Enhancing Methods in Analogy Making

Given the domain homogeneity in LLM analogy making observed above, we further explore the efficacy of existing frameworks that aim to enhance the output diversity of LLMs. As is mentioned in Ostermann et al. (2026), existing methods that improve output diversity can be divided into three categories, including fine-tuning, prompting, and steering. Since the scope of our work lies at inference time, we mainly investigate the latter two methods of prompting and steering.

For prompting, we are using the best results from our previous section, *basic* and *diverse* settings, which are optimal in quality and diversity respectively. We also include approaches that fall into the steering category and tend to perturb the model's internal hidden states in order to modify and encourage output diversity. We include approaches that span two mainstream subcategories from this perspective in our evaluation, including probabilistic / logistic resampling and entropy-gated resampling.

Probabilistic and Logistic steering. We employ **Top-k** that restricts sampling to the k most probable tokens Fan et al. (2018); **Top-p** that samples from the smallest set of tokens whose cumulative probability reaches a threshold Holtzman et al. (2020); **Top-$\eta\sigma$** that retains tokens whose logits are within η standard deviations of the highest logit Tang et al. (2024), and **Min-p** that discards tokens whose probability falls below a minimum threshold relative to the most probable token Minh et al. (2025). We also include G2 Ruan et al. (2025) in our evaluation [2]. G2 modifies the output logits based on previous outputs from guide models to improve diversity.

Entropy-gated steering. We take the motivation and insights from Li et al. (2026) and implement an adapted version to accommodate our task of analogy making, where we set a threshold τ with respect to the entropy of the next token to be generated. The resampling operation is only triggered when the entropy is higher than τ and the next token to be generated is identified as the domain token. Following observations in Fan et al. (2026), we run an additional forward pass in the same model in order to modify the output token distribution for resampling. The left column in Figure 2 summarizes these methods in our evaluation.

[2]We include a reproduction of G2 following the algorithm in the github repository.

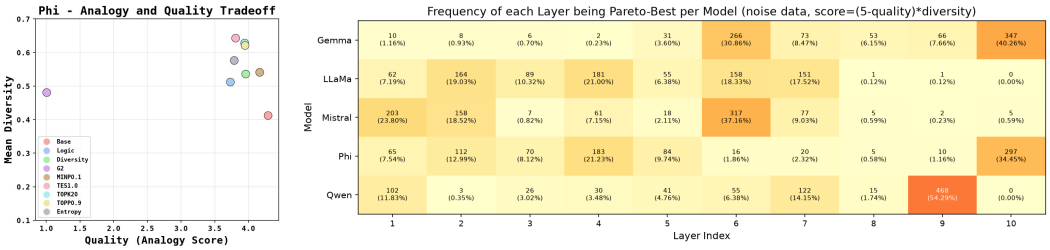

Figure 6: (a) Scatter plot of the Pareto frontier across all diversity enhancing frameworks using `Phi-4-mini-instruct`. (b) Heatmap of each layer group being selected as Pareto-best under layer-wise noise perturbation across different model families.

We showcase the performance comparison averaged across three datasets in a radar chart in Figure 5 [3] when applying these methods introduced above. Overall, there is no outstanding method that excels in all four evaluation dimensions across the five open-source models tested. For instance, despite the *base* prompting method achieving the highest output quality on all models (red color is close to 5 in all models), it falls short in terms of diversity metrics when compared to other methods. The shortcoming in diversity is particularly notable in the case of `Mistral-7B-instruct-v0.3`: with an average output quality rating of 4.32 in *base* prompting, it merely achieves an output diversity score of 0.33 and an average target domain cluster number of 1.97 on three datasets (Table 1 in the Appendix describes this numbers). In comparison, when applying the Top-k method (green color in Figure 5) on the `Mistral-7B-instruct-v0.3`, with an average analogy quality of 4.21, it achieves an average diversity score of 0.53 as well as the average number of target domain clusters of 2.27, indicating a compromise of diversity in the *base* prompting method when compared with probabilistic steering methods such as Top-k.

In addition to the lack of a dominating method that excels in all evaluation metrics, we also observe that different models exhibit unstable rankings of these approaches even when evaluated on a single metric. Taking the diversity score metric (indicated at the top of the circle in Figure 5) as an example, the best performing methods in enhancing diversity for `Llama-3.1-8B-instruct` are Top-p and G2 (which belong to the probabilistic and Logistic steering category), both achieving an average score of 0.61 on all three datasets. However, when the same two methods are applied to `Qwen-3-8B`, they only achieves a diversity score of 0.28 and 0.29, respectively. In comparison, the entropy-gated steering method achieves an average performance of 0.39 in diversity score.

6 Detailed Analysis

In this section, we provide an in-depth analysis, aiming to account for the observations of divergent performance across these diversity-enhancing methods elaborated in the previous section. Specifically, we show that (1) there exists a tradeoff between the output diversity and output quality, and (2) the misalignment between intervention position and the models' analogy-making region may account for the unstable performance in analogy making.

Diversity-quality tradeoff in analogy making. Following the observations in the radar chat (Figure 5) showing that the increase in analogy diversity is often accompanied by a decrease in output quality, we further investigate this tradeoff by plotting the Pareto frontier of each model. Specifically, we collect the average performance across three datasets on diversity and quality metrics and make a scatter plot of these methods. Due to space limitations, we only showcase the scatter plot for `Phi-4-mini-instruct` in Figure 6 (a). When compared with the base prompting (light red dot in Figure 6 (a)) that yields analogies with an average quality of 4.28, the most diverse method, Top-$\eta\sigma$, (light pink dot in Figure 6 (a)) only yields an average quality of 3.80, indicating a compromise in the diversity enhancing approaches in output quality. Additionally, we also observe a clustering effect of those diversity-enhancing approaches on the Pareto frontier. This further indicates an

[3] Complete results are also provided in Table 1 in the Appendix.

absence of a dominating approach in improving models' analogy diversity meanwhile maintaining high generation quality.

Misalignment between intervention position and the analogy-making region in LLMs. Given the observed tradeoff between diversity and quality, one might also be curious about whether the internal computations governing analogy generation are confined to particular regions of the model, regardless of the model family. To probe this question, we take inspiration from exiting works in causal perturbation analysis Todd et al. (2024) and introduce the following method that aims to identify the most fragile regions when perturbed with looped computation, a novel perturbation method suggested in Fan et al. (2026). Specifically, we blur the hidden states via looped inference and compute the indirect effect, a mainstream method introduced by Todd et al. (2024) to identify the most important layers within an LLM. In our case, we define the metric of indirect effect to be the combined score of $(5 - \text{quality}) \times \text{diversity}$. The most fragile regions are thus identified by maximizing the metric when replaced with perturbations, meaning that these layers are least informative yet diverse in analogy making. Then, we aggregate, for each model, the frequency with which each layer is selected across all evaluation samples. Figure 6 (b) visualizes these frequencies as a heatmap, revealing the layer positions that are most frequently identified as Pareto-optimal under our perturbation analysis.

From the resulting heatmap, we observe a fluid and divergent span of layer regions that are sensitive to analogy making across different models. For instance, `llama` and `mistral` models have a more widespread span of layers (the 20th percentile to the 70th percentile among layer depth) when perturbed with noise, while for `phi-4-mini` and `qwen-3`, the sensitive regions are concentrated between certain layer positions (the 90th percentile for `qwen-3` and the 100th percentile for `phi-4-mini`). The wide span along with the fluid regions across different models that are sensitive to analogy making also echos with the insights from Minegishi et al. (2026), where they argue that analogical reasoning is an ability emerging from the late phase of model internals. In contrast to this observation, however, all steering methods that influence output diversity operate on the logits that are obtained from the last layer. The misalignment between the sensitive layer regions and these perturbation methods thus offer a potential explanation for the unstable performance of methods when applied to different LLMs.

7 Conclusion and Outlook

We present a systematic study of output diversity in LLM-based analogy making across ten state-of-the-art open- and closed-source LLMs. Our results indicate the phenomenon of domain homogeneity from both intra-output and inter-model perspectives, where models repeatedly generate analogies grounded in a narrow set of target domains, a concerning issue that bears the risk of undermining AI4S creativity. Further evaluation on various inference-time diversity-enhancing approaches from three categories shows that their effectiveness is generally limited and strongly model-dependent, and there exists a trade-off in which gains in analogy making diversity are often accompanied by degradation in analogy making quality. Our further analysis also provides a potential explanation via layer perturbation, showing a misalignment between the intervention location of these methods and the model's sensitive region in analogy making.

To the best of our knowledge, this work is among the first to systematically investigate output diversity in open-ended LLM analogy generation. Our findings may help inform future LLM model designs that are more diversity-aware in analogy making.

AI use statement

In this work, we used generative AI tools for polishing in implementing research code elaborated in our method section, polish text grammar, and perform figure formatting. We have not used generative AI tools for the development of research ideas and the design of experiments, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *ArXiv*, abs/2303.08112, 2023. URL https://api.semanticscholar.org/CorpusID:257504984.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6692–6704, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.448/.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing, 2025. URL https://arxiv.org/abs/2503.17126.

J Clerk Maxwell. On faraday's lines of force. *Transactions of the Cambridge Philosophical Society*, 10:27, 1864.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14932–14948, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.681. URL https://aclanthology.org/2026.acl-long.681/.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Ying Fan, Anej Svete, and Kangwook Lee. Bridging the gap between latent and explicit reasoning with looped transformers, 2026. URL https://arxiv.org/abs/2606.31779.

Benjamin Fry, Kaia Slaw, and Nicholas F. Polizzi. Zero-shot design of drug-binding proteins via neural iterative selection–expansion. *Nature*, June 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10670-w. URL https://doi.org/10.1038/s41586-026-10670-w.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3. URL https://www.sciencedirect.com/science/article/pii/S0364021383800093.

Dedre Gentner. Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775, 2010. doi: https://doi.org/10.1111/j.1551-6709.2010.01114.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2010.01114.x.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.06102.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Douglas Hofstadter and Emmanuel Sander. *Surfaces and Essence: : Analogy as the Fuel and Fire of Thinking*. 04 2013. ISBN 978-0465018475.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR 2020)*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Xufeng Hong, Xizhe Wang, Stephen J. Harris, Hongbo Zhao, Jiashen Meng, Qingshan Jia, Qianchuan Zhao, Kang Xu, Quanquan Pang, and Benben Jiang. Deep active learning and knowledge transfer for rapid discovery of lithium metal battery electrolytes. *Nature Communications*, 17(1):5146, March 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-70973-4. URL https://doi.org/10.1038/s41467-026-70973-4.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL https://arxiv.org/abs/2510.22954.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Xianzhi Li, Ethan Callanan, Abdellah Ghassel, and Xiaodan Zhu. Entropy-gated branching for efficient test-time reasoning. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5054–5069, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.235. URL https://aclanthology.org/2026.eacl-long.235/.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *Nature Machine Intelligence*, 7(4):580–591, April 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01011-z. URL https://doi.org/10.1038/s42256-025-01011-z.

Gouki Minegishi, Jingyuan Feng, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Emergent analogical reasoning in transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GxTkgMBiz8.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. From weights to activations: Is steering the next frontier of adaptation? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29854–29879, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1377. URL https://aclanthology.org/2026.acl-long.1377/.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, pp. 1–13, 2026.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Zaid Harchaoui, et al. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in neural information processing systems*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Zhiwen Ruan, Yixia Li, Yefeng Liu, Yun Chen, Weihua Luo, Peng Li, Yang Liu, and Guanhua Chen. G2: Guided generation for enhanced output diversity in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14116–14134, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.713. URL https://aclanthology.org/2025.emnlp-main.713/.

Andrew Shen, Shaul Druckmann, and James Zou. Unlocking llm creativity in science through analogical reasoning. *arXiv preprint arXiv:2605.11258*, 2026a.

Yuanhao Shen, Daniel Xavier de Sousa, Ricardo de Andrade Nascimento, Ali Asad, Hongyu Guo, and Xiaodan Zhu. Understanding large language models' ability on interdisciplinary research. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Seoul, South Korea, July 2026b. PMLR. July 6–11, 2026.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. Top-$n\sigma$: Not all logits are you need. *arXiv preprint arXiv:2411.07641*, 2024.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17659–17681, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.978. URL https://aclanthology.org/2024.emnlp-main.978/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG. arXiv:2310.15213.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. Metaphor understanding challenge dataset for LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3517–3536, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.193. URL https://aclanthology.org/2024.acl-long.193/.

Peter D. Turney. Domain and function: a dual-space model of semantic relations and compositions. *J. Artif. Int. Res.*, 44(1):533–585, May 2012. ISSN 1076-9757.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13060–13082, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.725. URL https://aclanthology.org/2024.emnlp-main.725/.

Zirui Zhang, Zhihao Wang, Yiwen Liao, Yu Chang, Li Ding, Guangsheng Luo, Xiaonan Wang, and Haihui Wang. Machine learning–driven discovery of optimal designs for water electrolysis devices. *Science Advances*, 12(19):eadz1865, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A Appendix

A Experimental Setup

Datasets Experiments are conducted on three analogy datasets: **AB** (`AnaloBench_S1`), containing 340 samples; **MA** (`metaphoric_analogies`), containing 262 samples; and **MUNCH** (`munch_metaphors`), containing 300 samples. For each source sentence, every model generates $K = 10$ candidate analogies.

Evaluated Models We evaluate ten large language models (LLMs), comprising five proprietary and five open-weight models. The proprietary models include Claude Sonnet (`anthropic/claude-sonnet-4.5`), GPT-5.2 (`openai/gpt-5.2`), Gemini 2.5 Pro (`google/gemini-2.5-pro`), Gemini 3.1 Pro (`google/gemini-3.1-pro-preview`), and Grok 4.5 (`x-ai/grok-4.5`). The open-weight models consist of Phi-4 (`microsoft/Phi-4-mini-instruct`), Mistral (`mistralai/Mistral-7B-Instruct-v0.3`), Gemma 2 (`google/gemma-2-9b-it`), Qwen 3 (`Qwen/Qwen3-8B`), and Llama 3.1 (`meta-llama/Llama-3.1-8B-Instruct`). We access the proprietary models via the OpenRouter API, while the open-weight models are executed locally in FP16 precision using the Hugging Face `transformers` library

B Experiment 1 (a): Analogy Generation

The objective of this experiment is to evaluate both the diversity and analogy quality of model-generated sentences under different prompting and decoding strategies.

Unless otherwise specified, all models use standard sampling with a temperature of $T = 0.7$ and generate $K = 10$ candidate analogies for each input sentence. We evaluate three prompting strategies corresponding to the Base, Diversity, and Logic prompts described in the Prompts section.

To further investigate the diversity–quality trade-off, we additionally evaluate a higher temperature ($T = 1.5$) using Base Prompt combined with four decoding strategies: Top-k, Top-p, Top-$\eta\sigma$, and Min-p.

B.1 Evaluation Pipeline

Generation For every model and decoding configuration, a JSON file is produced containing the source sentences together with the K generated analogies.

Diversity Metrics For each input sample, a set of analogies is generated and evaluated using three complementary diversity metrics, which capture semantic, lexical, and vocabulary-level variation. The final diversity score is computed as a weighted combination of these metrics:

$$D_{\text{total}} = 0.5D_{\cos} + 0.25D_{\text{BLEU}} + 0.25D_{\text{EAD}}.$$

Average Cosine Distance ($D_{\cos}$). Each generated analogy is embedded using the SentenceTransformer model `all-MiniLM-L6-v2`. Pairwise cosine distances are computed between all valid analogy embeddings for each sample, and their average is used as the semantic diversity score. Higher values indicate greater semantic dispersion among the generated analogies. Samples containing fewer than two analogies receive a score of zero.

Div-BLEU (D_{BLEU}). Lexical diversity is measured using Self-BLEU. Each analogy is treated once as the candidate while all remaining analogies serve as references. Sentence-level BLEU is computed using SacreBLEU with exponential smoothing, and Self-BLEU is obtained by averaging across all candidates. The diversity score is then defined as

$$D_{\text{BLEU}} = 1 - \text{Self-BLEU},$$

such that higher values correspond to lower lexical overlap.

Expectation-Adjusted Distinctness (D_{EAD}). To measure vocabulary diversity while correcting the length bias of the original Distinct-n metric, we adopt Expectation-Adjusted Distinctness (EAD).

For each generation method, the reported diversity statistics correspond to the mean, median, and standard deviation of D_{total} across all evaluated samples.

Analogy Quality The quality of every generated analogy is independently assessed by a judge LLM using the Analogy Quality Scoring Prompt described in the prompt section in Appendix. The judge receives the original sentence, the generated analogy, and a formal definition of analogy based on domain mapping theory, and assigns an integer score from 1 (poor) to 5 (excellent) according to the validity of the mapping, semantic coherence, and overall meaningfulness. To ensure deterministic evaluation, inference is performed with temperature zero, and only the integer score is extracted from the model output. Responses outside the valid range are discarded.

For each input sample, the quality score is computed as the mean of the scores assigned to its generated analogies. The quality of a generation method (i.e., a specific combination of model, prompt, and decoding strategy) is then reported as the mean and standard deviation of these sample-level scores across the entire evaluation set.

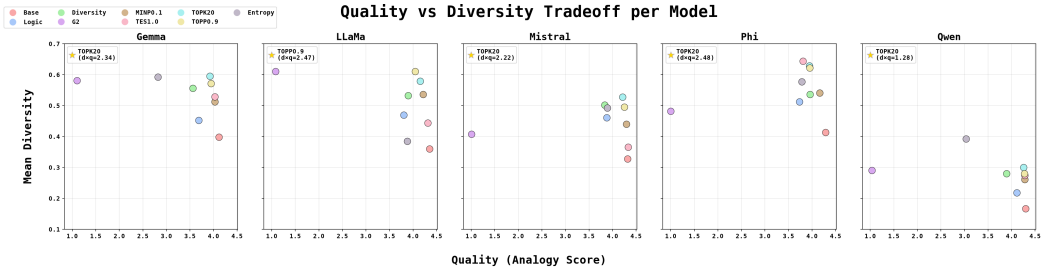

Figure 7: Quality–diversity tradeoff for each evaluated language model. Each point corresponds to a prompting or decoding strategy, positioned according to its mean analogy quality and mean diversity. The highlighted star denotes the selected operating point, determined by maximizing the combined objective $\text{Quality} \times \text{Diversity}$.

C Experiment 1 (b): Domain Extraction

This experiment measures the diversity of conceptual domains employed by language models during analogy generation. For each generated analogy, GPT-4o-mini ($T = 0.0$) extracts its source and target domains in JSON format using the Domain Extraction Prompt described in the Prompts section.

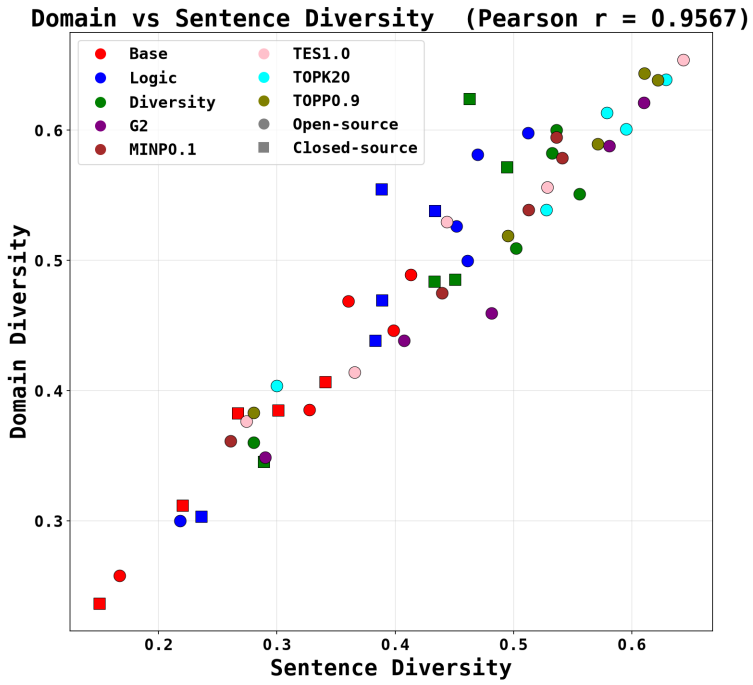

Figure 8: Relationship between sentence diversity and domain diversity across all evaluated generation methods. Each point corresponds to a model–method configuration, with colors indicating the prompting strategy and marker shapes distinguishing open-source and closed-source models.

Domain Diversity The extracted target-domain labels are treated as text samples and evaluated using the same diversity metrics described in Experiment 1(a). This allows us to quantify the semantic diversity of the conceptual domains independently of the linguistic variation in the generated analogies. Figure 8 compares domain diversity with sentence diversity across all evaluated generation methods. The strong positive correlation indicates that methods producing more diverse analogies also tend to explore a broader range of conceptual domains.

D Experiment 2: MAUVE Similarity

This experiment measures distributional similarity between proprietary language models using MAUVE.

Only generations produced using the Base prompt in the Prompts section with $T = 0.7$ are considered. Outputs from every model are pooled across all datasets before computing pairwise MAUVE scores. GPT-2 Large hidden representations are used for featurization, and the resulting pairwise similarities are presented as a heatmap.

E Experiment 3: Memorization Test

This experiment investigates whether proprietary models have memorized examples from the AnaloBench dataset.

For every prefix ratio

$$r \in \{0.1, 0.2, 0.4, 0.8\},$$

the first r fraction of the reference completion is provided to the model.

Using the Base Prompt described in Prompts section, models are instructed to reproduce the continuation from the provided prefix using few-shot examples. Performance is evaluated using Token Accuracy, Character Accuracy, Exact Match (EM), and Contains Accuracy. Higher Exact Match scores indicate stronger evidence of memorization.

F EXPERIMENT 4: LLM-BASED CLUSTERING

This experiment investigates the semantic organization of generated analogies through LLM-based clustering.

For every source sentence, the corresponding K generated analogies are submitted to GPT-4o-mini, which groups them into semantic categories represented as JSON objects using the Clustering prompt described in the Prompts section in Prompts section. The prompt instructs the model to cluster analogies according to their conceptual domains while avoiding lexical variations and synonym duplication.

Cluster-Linkage Analysis Cluster embeddings are computed using SentenceTransformer (`all-MiniLM-L6-v2`). For each pair of clusters within a sample, we compute the single, complete, and average linkage. The linkage score for a given method is obtained by averaging the corresponding linkage values across all samples. Samples containing only a single cluster are assigned a linkage score of 0. Additionally, we report the average number of semantic clusters per source sentence.

G EXPERIMENT 5: FRAGILE REGION IDENTIFICATION

Noise Blur This experiment investigates whether analogy generation is associated with specific regions of transformer models or is distributed across the network. For each input sample, the model generates K analogies by applying hidden-state perturbations independently at transformer layers located at 10% depth intervals, from 10% to 100% of the network. The generated analogies are evaluated using the same diversity and quality metrics described in Experiment 1(a), yielding a diversity score and an analogy quality score for every layer and sample. These scores are subsequently used to identify the layers most strongly associated with the analogy-making process through a perturbation-based analysis.

Layer Evaluation For each input sample, every transformer layer is represented as a point in the quality–diversity space, where quality is minimized and diversity is maximized. The Pareto frontier is then computed by retaining only the non-dominated layers. Among the Pareto-optimal candidates, a single layer is selected by maximizing the combined objective

$$(5 - \text{Quality}) \times \text{Diversity},$$

which favors layers achieving both low-quality and diverse analogy generation. This procedure is repeated independently for every sample. Finally, for each language model, we report the frequency with which each layer is selected as the optimal Pareto layer across the evaluation set, providing an estimate of the model regions that are most consistently sensitive to hidden-state perturbations during analogy generation.

Model Name	Method	AB				MA				MUNCH			
		Div. ↑	Qual. ↑	Clus. Linkage	#Clus.	Div. ↑	Qual. ↑	Clus. Linkage	#Clus.	Div. ↑	Qual. ↑	Clus. Linkage	#Clus.
Llama-3.1-8B-instruct	Std. Prompt	0.355	**4.66**	0.351	2.10	0.378	4.06	0.421	2.40	0.348	4.33	0.286	1.97
	Div. Prompt	0.564	4.18	0.639	**3.24**	0.495	3.71	0.577	2.96	0.538	3.80	0.590	3.28
	Top-k	0.584	4.46	0.432	2.18	0.582	3.83	0.550	2.68	0.571	4.18	0.395	2.16
	Top-p	0.613	4.40	0.450	2.24	0.614	3.70	0.571	2.70	0.604	4.04	0.446	2.23
	Top-$\eta\sigma$	0.446	4.64	0.359	2.09	0.455	4.01	0.459	2.57	0.430	4.30	0.347	2.18
	Min-p	0.542	4.53	0.435	2.31	0.543	3.89	0.502	2.63	0.523	4.22	0.394	2.18
	Entropy	0.393	4.36	0.376	2.17	0.465	3.54	0.526	2.70	0.384	3.88	0.395	2.33
	G2	0.601	1.07	0.474	2.03	0.620	1.10	0.503	2.06	0.610	1.07	0.385	1.73
Gemma2-9B-it	Std. Prompt	0.374	4.44	0.420	2.15	0.422	3.77	0.493	2.38	0.399	4.13	0.410	2.27
	Div. Prompt	0.564	3.85	0.690	3.09	0.514	3.46	0.582	2.93	0.590	3.37	0.686	**3.54**
	Top-k	0.584	4.26	0.536	2.53	0.609	3.57	0.606	2.91	0.593	3.93	0.509	2.62
	Top-p	0.549	4.29	0.511	2.51	0.592	3.59	0.581	2.87	0.572	3.95	0.503	2.72
	Top-$\eta\sigma$	0.506	4.36	0.496	2.32	0.556	3.66	0.561	2.71	0.523	4.06	0.493	2.66
	Min-p	0.493	4.38	0.471	2.38	0.534	3.67	0.560	2.61	0.511	4.03	0.454	2.60
	Entropy	0.589	3.40	**0.721**	2.82	0.619	2.88	**0.751**	2.97	0.592	2.82	**0.743**	2.77
	G2	0.570	1.08	0.461	2.05	0.573	1.12	0.506	2.03	0.599	1.09	0.466	1.90
Phi-4-mini-instruct	Std. Prompt	0.417	4.57	0.399	2.30	0.423	3.97	0.450	2.33	0.399	4.31	0.325	2.15
	Div. Prompt	0.543	4.31	0.596	3.09	0.515	3.71	0.607	**3.28**	0.550	3.85	0.610	3.48
	Top-k	0.641	4.19	0.520	2.48	0.628	3.64	0.556	2.70	0.616	4.00	0.453	2.34
	Top-p	0.629	4.24	0.498	2.43	0.628	3.60	0.570	2.75	0.609	4.00	0.448	2.42
	Top-$\eta\sigma$	**0.659**	4.03	0.523	2.57	**0.638**	3.49	0.574	2.86	**0.632**	3.89	0.494	2.49
	Min-p	0.545	4.45	0.426	2.34	0.552	3.83	0.517	2.74	0.526	4.20	0.379	2.29
	Entropy	0.603	4.24	0.493	2.01	0.560	3.62	0.477	2.16	0.577	3.79	0.477	2.04
	G2	0.487	1.00	0.128	1.27	0.481	1.00	0.167	1.42	0.476	1.00	0.105	1.21
Mistral-7B-instruct-v0.3	Std. Prompt	0.325	4.56	0.317	1.94	0.351	4.02	0.394	2.18	0.307	4.37	0.230	1.78
	Div. Prompt	0.518	4.16	0.604	3.06	0.468	3.68	0.542	2.95	0.520	3.66	0.606	**3.54**
	Top-k	0.538	4.49	0.404	2.24	0.545	3.90	0.505	2.57	0.500	4.23	0.302	2.00
	Top-p	0.497	4.53	0.382	2.13	0.518	3.93	0.462	2.41	0.470	4.29	0.336	2.05
	Top-$\eta\sigma$	0.368	4.58	0.336	1.99	0.386	**4.04**	0.398	2.22	0.343	**4.38**	0.251	1.83
	Min-p	0.441	4.55	0.347	2.06	0.460	4.01	0.454	2.40	0.417	4.32	0.306	1.91
	Entropy	0.515	4.20	0.387	1.88	0.565	3.65	0.549	2.49	0.492	3.89	0.359	1.97
	G2	0.389	1.00	0.148	1.38	0.416	1.01	0.215	1.58	0.418	1.00	0.131	1.30
Qwen3-8B	Std. Prompt	0.160	4.61	0.203	1.59	0.172	4.03	0.229	1.65	0.170	4.26	0.206	1.63
	Div. Prompt	0.282	4.21	0.407	2.24	0.263	3.72	0.363	2.03	0.296	3.75	0.398	2.31
	Top-k	0.285	4.56	0.311	2.06	0.314	3.98	0.361	2.19	0.300	4.22	0.309	2.14
	Top-p	0.266	4.60	0.305	2.01	0.294	4.01	0.372	2.23	0.281	4.21	0.304	2.07
	Top-$\eta\sigma$	0.263	4.58	0.290	1.89	0.285	4.02	0.353	2.11	0.274	4.24	0.305	2.05
	Min-p	0.252	4.60	0.286	1.93	0.269	4.00	0.317	2.00	0.261	4.24	0.306	2.08
	Entropy	0.377	3.47	0.334	1.81	0.398	2.95	0.362	1.78	0.394	3.04	0.300	1.63
	G2	0.287	1.03	0.177	1.42	0.292	1.04	0.222	1.61	0.290	1.03	0.174	1.37

Table 1: Evaluation results on three datasets using different generation perturbation methods. AB, MA stands for AnaloBench Ye et al. (2024) and Meterphoric Analogy Tong et al. (2024), respectively. The best results among the perturbation methods are underlined, and the best results across different models are **bolded**.

G.1 Complete Results

This section presents the complete experimental results in Table 1.

H Prompts

This section presents all prompt templates used throughout the experiments. In every template, text enclosed in {curly_braces} denotes a placeholder that is replaced with the corresponding input at runtime.

Generation Base Prompt

```
Generate a sentence that is analogous to the sentence below.  Directly
output the answer with no explanation.
Sentence:  {source_story}
Answer:
```

Generation Logic Prompt

```
Generate a sentence that follows the same relational structure of the
sentence below, but in a completely different domain.  Directly output
the sentence with no explanation.
Sentence:  {source_story}
Answer:
```

Generation Diversity Prompt

```
Generate a sentence that is analogous to the sentence below.  Be as
creative and diverse as possible in your choice of domain.  Directly
output the answer with no explanation.
Sentence:  {source_story}
Answer:
```

Analogy Quality Scoring Prompt

```
Read the sentence and the generated analogy, and give it a score
between 1 and 5 based on whether the generated analogy is a valid
analogy for the sentence.


The official definition of analogy can be found below:

"An analogy is composed of object mappings and shared relations.
Object mappings establish correspondences between equivalent objects
across different domains, while shared relations represent equivalent
relationships between these objects in both domains."


Think carefully before making your verdict.  When assigning the
score, also evaluate whether the generated analogy is coherent and
meaningful.


Sentence:  {sentence}
Generated Analogy:  {analogy}


Respond ONLY with a single integer from 1 to 5.
```

Domain Extraction Prompt

```
You are analyzing an analogy to identify its conceptual structure it
establishes.
An analogy maps objects from a source domain to corresponding objects
in a different target domain while preserving their relationships.


**Input sentence:**
{{input_sentence}}


**Provided analogy:**
{{input_analogy}}


**Your task:**
Analyze the provided analogy and identify the conceptual domains
involved.  Do NOT generate a new analogy or modify the provided one.
Only extract the conceptual structure already present.


**Rules for extracting domains:**
1.  Granularity:  Use broad, standard fields of activity, industries,
or areas of knowledge.
2.  No Specific Entities:  Do NOT use specific events, projects, or
historical names.
3.  Format:  Use lowercase, singular noun phrases.  Max 3 words per
```

```
domain.
4.  Language:  Output the domains in English.

**Return a JSON object with:**

1.  "source_domain":  The primary conceptual or real-world domain
represented in the original input sentence.
2.  "target_domain":  The conceptual or real-world domain represented
in the provided analogy.

Return ONLY the JSON object, with no additional text.
```

Clustering Prompt

```
You are given {K} sentences.  Your task is to group them into semantic
categories based on their underlying conceptual domain.

**Task**
Analyze the meaning of each sentence and assign each sentence to
exactly one category representing its primary conceptual domain.

**Category Guidelines**
* Categories must represent broad semantic concepts, not the exact
wording used in the sentences.
* Normalize synonyms and closely related concepts into the same
category.
* Do not create separate categories for lexical variations that refer
to the same underlying domain.
Examples:
- "recipe creation", "recipe invention", "recipe development" -> same
category
- "car repair" and "vehicle maintenance" -> same category
- "writing a book" and "book creation" -> same category
* Prefer abstract domain names that describe the shared concept.
* Category names should not simply copy phrases from individual
sentences.

**Classification Rules**
* Analyze the meaning of each sentence, not just keywords.
* Identify the main subject/domain of the sentence.
* Group sentences together when they concern the same conceptual area,
even if they use different wording or describe different scenarios.
* If a sentence is an analogy, classify it according to the domain of
the objects being compared, not according to the analogy structure.
* Do not group sentences from different domains just because they
share similar relationships, emotions, or logical patterns.

**Category Creation Rules**
* The number of categories is not predefined.
* Create a new category only when the underlying conceptual domain is
genuinely different.
* Before creating a new category, check whether an existing category
can describe the sentence after semantic normalization.
* Prefer fewer categories when categories would only differ by
wording.
* Prefer more categories when the real-world domains differ.

**Constraints**
* Every sentence must belong to exactly one category.
* Do not duplicate sentence indices.
```

```
* Do not create categories that differ only by synonyms.

**Input**
A numbered list of sentences:

{sentences}

**Output Requirements**
Return ONLY a valid JSON object.

* Keys must be concise category names representing semantic domains.
* Values must be arrays containing the sentence indices belonging to
that category.
* Do not include explanations or additional text.

Example output:
{ "Recipe Development":  [1, 2, 3], "Vehicle Maintenance":  [4, 5] }
```