GRAPH ENHANCED CREATIVE COGNITION FOR ALTERNATE USES TASK

Pronita Mehrotra*
MindAntix

ABSTRACT

Large language models can generate fluent responses to creative tasks, but repeated generations often converge on familiar or semantically related ideas. We propose a graph-guided associative prompting framework for improving creative generation on the Alternate Uses Task (AUT). Our method uses ConceptNet to identify two-hop associative triggers that are moderately remote to the source object. We evaluate three open-weight LLMs across six AUT objects and compare standard prompting against graph-guided prompting. To better capture distributional creativity, we introduce a frequency-based semantic clustering evaluation that measures both ideational diversity and corpus-relative originality. Under this analysis, graph-guided prompting increases the number of distinct idea clusters across all three models, with gains of 147%, 70%, and 33% for Gemma-4B, Mistral-7B, and OLMo-7B, respectively. We further train a relational graph neural network to rank ConceptNet paths by their downstream creative utility and show that GNN-selected paths produce 16–20% more semantic clusters than randomly selected paths and more highly original ideas. These results suggest that structured conceptual graphs can serve not only as knowledge sources for reasoning, but also as controllable and learnable search spaces for expanding the creative output of LLMs.

Keywords Creativity · Divergent Thinking · Alternate Uses Task · Graph Neural Networks

1 Introduction

> The Brain, within its Groove / Runs evenly — and true —
> But let a Splinter swerve — / 'Twere easier for You —
> To put a Current back — / When Floods have slit the Hills —
> And scooped a Turnpike for Themselves — / And trodden out the Mills —
> — *Emily Dickinson, "The Brain, within its Groove (556)"*

The evocative imagery of Emily Dickinson's poem reveals a key insight into creative thought. Creativity, in this view, starts with a small deviation (the "Splinter") that redirects cognition from the probable path toward a path that did not previously exist. The resulting creative process can be thought of as a "blend of spontaneity and deliberate effort" to discover new meaningful paths through a conceptual landscape [27].

A long tradition in cognitive psychology models thought as an associative search over a dense interconnected network of concepts through which activation, retrieval, and inference propagate [9, 4]. Creativity research extends this view by suggesting that creative ideas often emerge when cognition goes beyond dominant associations towards more remote but still meaningful concepts [16, 7]. The description of the brain as an "associative engine" [12] provides a useful summary of this tradition where one idea activates another, and thought proceeds through cascades of related concepts.

Contemporary autoregressive LLMs can produce fluent poems, stories or product ideas, often from only a prompt or a handful of examples. GPT-3 demonstrated that scaling such models yields strong few-shot performance across diverse language tasks and can generate long-form text that is hard to distinguish from human writing [8]. The first and most direct way to make models more creative was by controlling the token-level stochasticity through temperature. Higher temperature values flatten the token probability distribution making lower-probability tokens more likely.

*Corresponding Email: pronita@mindantix.com

However, recent empirical work shows that temperature is only weakly related to novelty and moderately correlated with incoherence [20]. This limitation has motivated higher-order approaches including prompting, iterative refinement, planning, and search over intermediate ideas. However, what remains missing is an explicit representation of the conceptual path by which a model moves from one idea to another. Without such a representation, higher-order methods can improve outputs while still leaving the creative trajectory largely implicit.

In this paper, we show that with the aid of conceptual graphs that identify new pathways, we can improve the originality and diversity of LLM responses to the Alternate Uses Task (AUT), a foundational divergent thinking task. With this approach, associative triggers are generated by navigating structured conceptual graphs and used to "nudge" LLMs towards more novel outputs.

This graph-guided approach offers three advantages. First, it provides a closer analogue to human associative creativity, potentially allowing LLMs to better mimic human creative thinking. Second, it enables more structured exploration of the creative search space: one can explicitly control distance, bridge concepts and domain crossings. Finally, it improves interpretability by providing a "cognitive lineage". An output can be traced back to a path through the concept graph exposing which concepts were combined and why a particular idea emerged. This is especially important in domains such as scientific hypothesis generation or education, where the value of an idea depends not only on its novelty but also on the reasoning path.

This paper makes the following contributions:

1. We propose a graph-guided associative prompting framework that uses two-hop ConceptNet paths to generate moderately remote but semantically linked triggers for creative LLM generation.
2. We show that these graph-guided triggers consistently increase ideational diversity across three open-weight LLMs, producing more distinct semantic clusters and more responses in rare, high-originality clusters than standard prompting.
3. We introduce a frequency-based semantic clustering evaluation that adapts traditional divergent-thinking originality scoring to LLM outputs, enabling joint analysis of corpus-relative originality and model-level diversity.
4. We train a relational GNN to identify preferred ConceptNet paths and show that GNN-ranked paths produce higher ideational diversity than randomly selected paths, establishing conceptual graphs as a learnable search space for creative prompting.

We argue that creative generation can be improved by shaping the associative path that guides generation through structured conceptual space.

2 Background

2.1 Creativity, divergent thinking, and originality assessment

Creativity is commonly defined as the production of ideas that are both novel and useful [22]. Although Rhodes' 4P framework characterizes creativity through the person, process, press, and product [21], empirical evaluation is often product-centered because novelty and usefulness must ultimately be judged from observable outputs. We follow this tradition by evaluating the creative artifacts generated by LLMs.

The psychometric study of creative potential is closely tied to divergent thinking. Guilford argued that creative thought differs from conventional intelligence in its ability to generate multiple possible responses rather than converge on a single correct answer [11]. This framing introduced dimensions such as fluency, flexibility, originality, and elaboration, which Torrance later incorporated and extended in the Torrance Tests of Creative Thinking [24, 1]. The Alternate Uses Task (AUT), which asks participants to generate unusual uses for common objects, remains one of the most widely used divergent-thinking tasks because it provides a simple way to study whether a system can move beyond obvious ideas.

Originality in AUT responses has traditionally been scored relative to a reference population: common ideas receive lower scores, while rare ideas receive higher scores [24, 6, 2]. More recent work has automated originality scoring using semantic distance, showing that the distance between an AUT prompt and a response can approximate human originality judgments [5]. However, semantic distance is not appropriate in our case, as we control the associative distance by selecting only two-hop ConceptNet triggers. In the spirit of Guilford's definition of originality and TTCT scoring, we use a frequency-based semantic clustering as a measure of originality. Along with ideational diversity, it paints a better picture of the creative potential of a model.

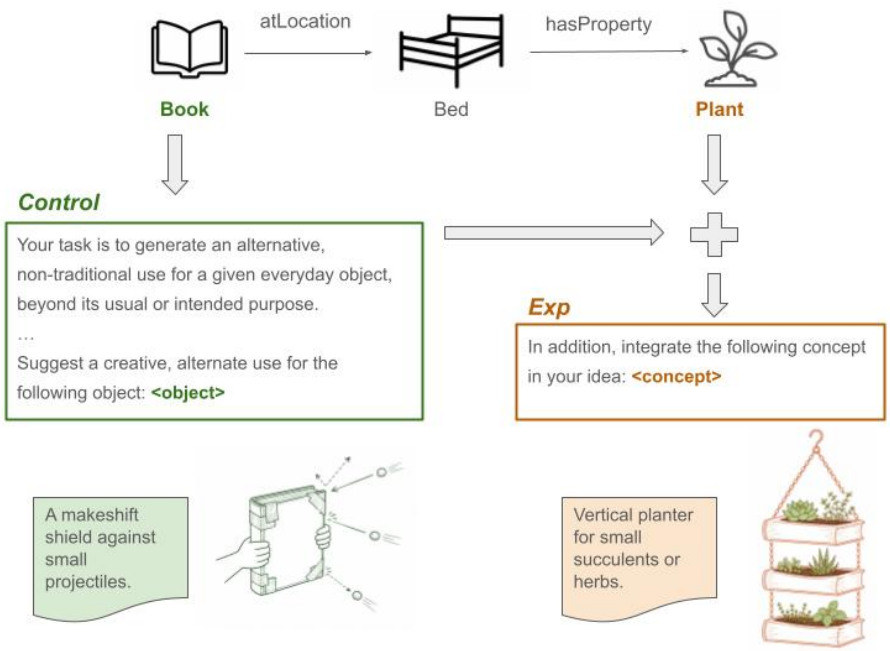

Figure 1: Prompts used in the control and graph conditions, along with sample Gemma-3 (4B) outputs. The top part of the image shows a sample 2-hop traversal of the ConceptNet graph from the root node, "book."

2.2 Associative creativity and LLM generation

Associative theories of creativity argue that novel ideas often emerge by forming useful connections among remote concepts [16, 7]. This perspective is relevant to LLMs because autoregressive generation tends to favor likely continuations unless prompts, decoding strategies, or external structures expand the search space.

Recent work has begun to adapt associative mechanisms to LLM generation. Associative prompting asks models to connect seemingly unrelated concepts before producing an output, improving originality in open-ended creative tasks [17]. Benchmarks such as CREATE evaluate whether models can construct meaningful associative paths between concepts [25]. Other approaches intervene at inference time or training time to increase sustained diversity and reduce generic responses [15]. These methods suggest that, much like humans, LLMs must move away from obvious ideas while preserving coherence and task relevance. What remains less explored is how to control the conceptual path that drives this movement.

2.3 ConceptNet and graph-guided associative search

Structured commonsense graphs provide a natural mechanism for making associative paths explicit. ConceptNet represents everyday concepts as nodes connected by labeled relations, making it possible to trace interpretable paths between objects and associated concepts [23]. Prior work has used ConceptNet and related knowledge graphs to improve reasoning in language models, including graph-based commonsense QA systems such as KagNet [13], QA-GNN [26], and GreaseLM [28]. These systems primarily use graphs to improve correctness-oriented reasoning.

We use ConceptNet differently: as a structured search space for creative association. Rather than grounding a question-answer pair in a graph to identify the correct answer, we select intermediate-distance concepts that can act as associative triggers for LLM generation. We then train a graph neural network to identify which ConceptNet paths are most likely to produce diverse and rare AUT responses. This reframes conceptual graphs not only as repositories of commonsense knowledge, but as controllable search spaces for creative generation.

3 Methodology

3.1 Graph-guided associative prompting

Our goal is to improve the originality of LLM-generated ideas on the Alternate Uses Task (AUT), a standard divergent-thinking task in which a subject is asked to generate unusual uses for a common object. The AUT is well suited to our task because creative performance can be evaluated at the level of individual generated products, while also allowing repeated sampling to measure diversity across responses. Our method is motivated by associative theories of creativity, which suggest that novel ideas often emerge by connecting remote concepts. In the context of LLM generation, we provide an additional associative trigger that nudges the model toward a less conventional region of semantic space. However, originality and usefulness are often in tension. A trigger that is too close to the source object may produce predictable responses, while a trigger that is too distant may produce ideas that are incoherent or impractical. We therefore use graph distance as a structured way to select triggers that are remote enough to promote novelty but close enough to appear relatable. Research with human participants suggests that a 2-hop distance results in ideas that are perceived as surprising or delightful, without appearing incongruent [14].

To construct these triggers, we use ConceptNet, a large multilingual commonsense knowledge graph in which natural-language concepts are represented as nodes and commonsense relations are represented as labeled, weighted edges. ConceptNet aggregates knowledge from different sources and is designed to represent the kinds of everyday relational knowledge needed for language understanding [23]. For each AUT object, we extract concepts that are two hops away from the object in ConceptNet. We also prune the subgraph more by removing some of the relation categories (e.g. "relatedTo" as these don't have a clear and concrete relationship) and nodes consisting of phrases (which are harder to use as a trigger). We then use these two-hop concepts as associative triggers, with the intuition that they occupy a useful middle ground: they are not direct associations of the object, but they remain reachable through interpretable paths.

3.2 Models and experimental conditions

We evaluate LLM responses using three open-weight small language models: Gemma-3 4B [10], Mistral 7B [18], and OLMo 7B [3]. All models are quantized to 4-bit precision to reduce memory requirements and allow inference on NVIDIA T4 GPUs. We use the same inference settings across models and conditions, including the same decoding parameters, prompt format, and output constraints. Prompts and model parameters used are provided in the Appendix A.

We compare two prompting conditions. In the control condition ("Base"), the model is asked to generate an alternate use for a common object. In the experimental condition ('Exp"), the model receives the same AUT prompt but is additionally instructed to incorporate a ConceptNet-derived associative trigger. Figure 1 summarizes the experimental flow. In Base, the model generates a response directly from the object. In Exp, the object is first mapped to a two-hop ConceptNet concept, and the model is then prompted to generate an alternate use that meaningfully incorporates that trigger.

We use six common AUT objects: *book, table, fork, pants, bottle, and coin*. For each object, model, and condition, we run 100 independent generations, yielding 3,600 possible observations before filtering. We remove empty responses, malformed outputs, and responses that do not contain a valid alternate use. After filtering, the final dataset contains 3,552 valid generations.

3.3 Evaluating Creativity

We assess creativity by evaluating the originality and diversity of ideas. We use two methods for assessing originality: independent assessment of individual responses and a frequency-based approach based on cosine similarity.

3.3.1 Originality scoring with OCSAI

OCSAI is an automated creativity-scoring system based on large language models fine-tuned on human-rated divergent-thinking responses [19]. OCSAI is designed to score the originality of short creative responses, including AUT-style responses, and prior work has shown that LLM-based automated scoring can substantially improve agreement with human originality ratings compared with earlier semantic-distance-only methods. OCSAI returns an originality score on a five-point scale, where higher scores indicate more original responses.

For each valid generation, we submit the object-response pair to OCSAI and record the resulting originality score. We then compare Base and Exp conditions within each model and object. This allows us to test whether ConceptNet-derived associative triggers increase the perceived originality of individual responses.

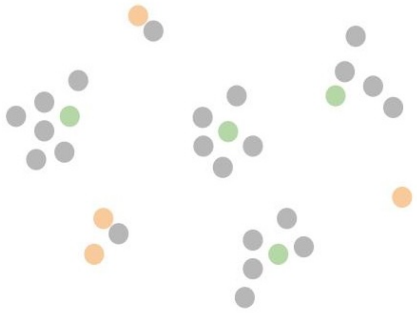

Figure 2: Diversity vs. Originality. Green dots show higher diversity while Orange ones show higher originality

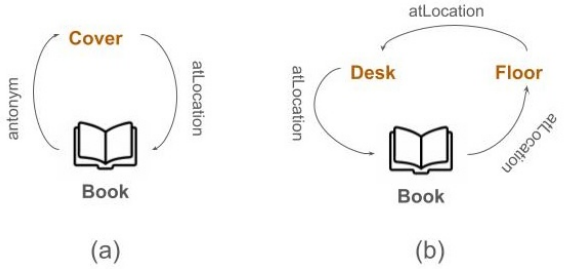

Figure 3: Semantically close 2-hop concepts. (a) shows a scenario where the 2-hop concept loops back to the original, Book->Cover->Book. (b) shows a 2-hop scenario that is effectively 1-hop, Book->Floor->Desk but Book also connects to Desk directly.

3.3.2 Semantic Diversity and Frequency-based originality

Although OCSAI provides a useful estimate of the originality of individual responses, we observed several limitations that we discuss in the Results section. We therefore complement OCSAI with a second measure inspired by traditional originality scoring in divergent-thinking research. In human AUT scoring, originality is often assessed by comparing an idea against the distribution of ideas produced by other participants: common responses receive lower originality scores, whereas statistically infrequent responses receive higher originality scores. This approach treats originality not as an absolute property of a response, but as a function of how rarely that idea appears within a reference population.

We adapt this logic to LLM-generated responses by pooling all valid responses across models and experimental conditions that acts as a reference corpus for each object. We embed each response using a sentence-embedding model, and compute pairwise cosine similarities. We then cluster responses based on semantic similarity so that closely related ideas are grouped into the same response category.

We use the resulting clusters in two ways.

First, we use the number of clusters generated by each model-condition pair as a measure of ideational diversity. A model-condition pair that produces responses spread across many clusters is interpreted as exploring a broader region of the idea space, whereas one whose responses concentrate in a small number of clusters is interpreted as producing more semantically similar outputs.

Second, cluster size provides a frequency-based estimate of originality. Responses assigned to large clusters correspond to ideas that many generations converged on and are therefore treated as less original. Responses assigned to small clusters correspond to rarer idea types and are therefore treated as more original. We bucket clusters into five bands based on cluster frequency, with the most frequent clusters assigned to the lowest originality bucket and the rarest clusters assigned to the highest originality bucket.

The conceptual difference between diversity and originality of ideas is shown in Figure 2. The green dots are spread across 4 larger clusters (higher diversity, lower originality), while the orange dots span 3 smaller clusters in comparison (lower diversity, higher originality). Although diversity and originality are often found to be correlated, a model could produce more diverse but less original ideas or vice versa.

3.4 Graph neural network for path selection

The Exp condition uses two-hop ConceptNet concepts as associative triggers, but not all two-hop paths are equally useful. Some paths lead back to the root concept or to another concept in its 1-hop neighborhood as shown in 3. An earlier study found that prompting an LLM to self-identify concepts at a 2-hop distance doesn't work as the model preferentially picks closer concepts [17]. ConceptNet subgraphs can exhibit a similar limitation, because several two-hop paths form local loops or return to concepts that are semantically close to the root. In addition, some paths may simply not result in original ideas due to the underlying quality of data.

To address this, we train a 2-Layer Relational Graph Convolutional Network to predict which ConceptNet paths are more likely to produce diverse and original generations. We focus this proof-of-concept experiment on the object "book". We first construct the two-hop ConceptNet subgraph rooted at book and enumerate candidate paths from the

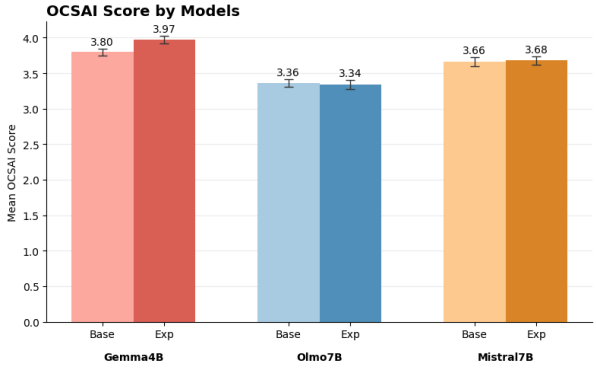

Figure 4: Average OCSAI originality scores for different models. Error bars correspond to 95% CI.

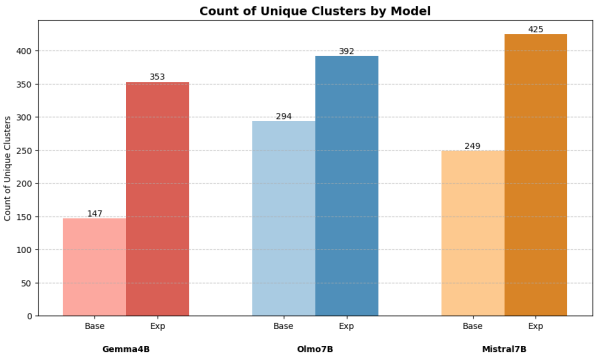

Figure 5: Number of unique semantic clusters for different models.

root object to potential trigger concepts. To create supervision, we run additional Exp generations covering all the candidate paths in the book subgraph. For each path, we test each model three times and collect the resulting AUT responses.

We then apply the same semantic clustering procedure described above and group paths into five buckets. Buckets 1–3 correspond to lower-diversity paths and account for approximately half of the responses; these are labeled not preferred. Buckets 4–5 correspond to higher-diversity paths and are labeled preferred. This approach creates a roughly balanced dataset for training a path-selection model.

The GNN uses the Adam optimizer, a learning rate of 1×10^{-3} and is trained for 200 epochs using the dataset. After training, we rank all paths in the subgraph and use the top 100 paths as associative triggers to compare against an equal number of paths picked randomly.

4 Results

We evaluate two research questions:

- **RQ1:** Can graph-guided associative triggers improve the originality and diversity of LLM responses on the Alternate Uses Task?
- **RQ2:** Can a graph neural network identify preferred ConceptNet paths that lead to more diverse creative outputs?

4.1 RQ1: Graph-guided associative triggers improve ideational diversity

We first compare graph-guided associative prompting against the Base condition across three small language models. We evaluate six objects across three models, using 100 generations per object, model, and condition. Figure 4 reports the average OCSAI originality score across all objects for each model and condition, with object-level results provided in Appendix B. On this metric, only Gemma-4B shows a statistically significant improvement under graph-guided prompting.

A closer inspection of OCSAI scores revealed two limitations. First, the score distribution was compressed toward the higher end: more than 66% of valid responses received a score of 3.5 or higher. This ceiling effect made it difficult to distinguish moderately original responses from highly original ones. Second, OCSAI sometimes assigned substantially different scores to semantically similar responses. For example, in the Base condition for the object book, 15% responses proposed using the book as some form of shield, yet these responses received scores ranging from 3.0 to 4.5. These examples suggest that OCSAI captures perceived originality at the level of individual responses, but not corpus-level originality: whether a model repeatedly returns to the same common ideas across generations or suggests responses that are common across models.

To address this limitation, we performed a frequency-based semantic clustering analysis described in the Methodology section. This approach more closely follows traditional AUT originality scoring, where ideas that occur frequently in a reference population are treated as less original, while rare ideas are treated as more original. For each object, we

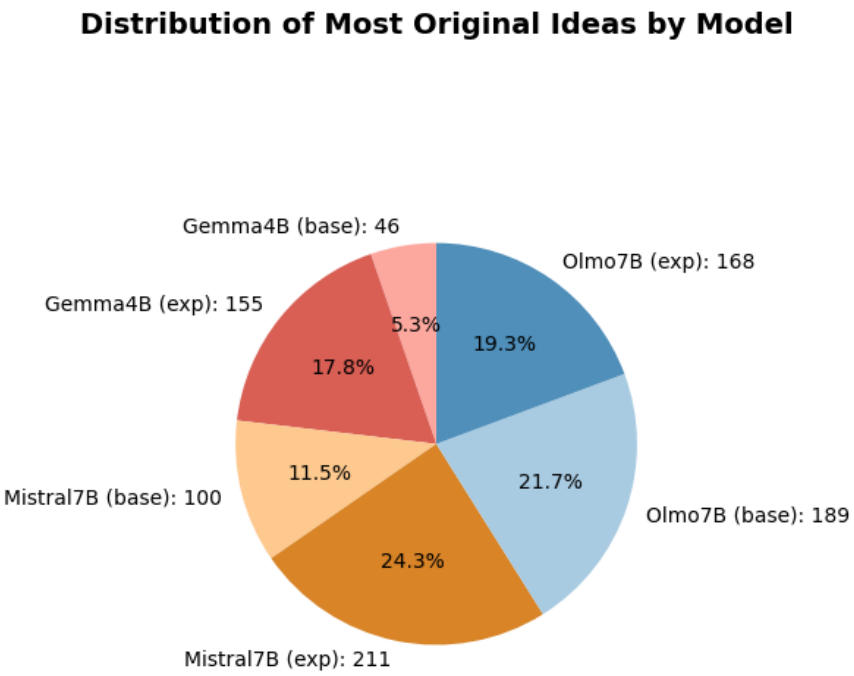

Figure 6: Number of highly original ideas by each model.

pool responses across all models and conditions, cluster semantically similar responses, and use cluster frequency as a corpus-relative originality signal.

Figure 5 shows that graph-guided associative prompting increases ideational diversity for all three models. Compared with the Base condition, the graph-guided condition produces a larger number of distinct semantic clusters, indicating that the models explore a broader range of idea types. The relative increase is largest for Gemma-4B, which shows a 147% increase in cluster count, followed by Mistral-7B with a 70% increase and OlMo-7B with a 33% increase. These results indicate that the ConceptNet triggers do not merely alter surface wording; they expand the range of semantic categories represented in the generated responses. Appendix C shows the unique clusters for each object.

We next examine the distribution of responses for the most original bucket. Figure 6 focuses on the highest-originality bucket, which contains responses assigned to the rarest semantic clusters. For Gemma-4B and Mistral-7B, the graph guided condition contributed more responses to the highly original bucket compared to the Base condition, while Olmo-7B showed a small drop despite having increased diversity.

Together, these results suggest that graph-guided associative prompting is quite effective at inducing diversity and originality of responses.

4.2 RQ2: GNN-based path selection improves over random ConceptNet triggers

The graph-guided condition in RQ1 samples ConceptNet paths without learning which paths are most likely to produce diverse outputs. To test whether a GNN-based approach might be useful or not, we train a two-layer relational graph convolutional network to classify ConceptNet paths as preferred or not preferred, using the path-labeling procedure described in the Methodology section. After 200 epochs, the model achieves a training loss of 0.5169, test accuracy of 0.6413, and ROC AUC of 0.6822. Table 1 reports the confusion matrix and classification metrics.

These results indicate that the GNN learns a meaningful but imperfect signal, with recall for preferred class being the strongest. Qualitative inspection confirms that the GNN assigns low scores to paths that appear to loop back toward the

Table 1: Held-out performance of the GNN path classifier.

(a) Confusion matrix

	Predicted	
True label	**Not Pref.**	**Pref.**
Not Preferred	331	239
Preferred	166	393

(b) Classification metrics

Class	**Precision**	**Recall**	**F1**	**Support**
Not Preferred	0.67	0.58	0.62	570
Preferred	0.62	0.70	0.66	559
Macro avg.	0.64	0.64	0.64	1129
Weighted avg.	0.64	0.64	0.64	1129

Overall: Accuracy = 0.641; ROC AUC = 0.682.

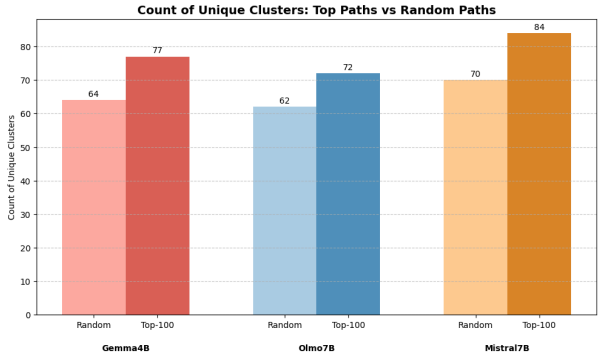

Figure 7: Number of unique semantic clusters for each model for random and top-100 paths.

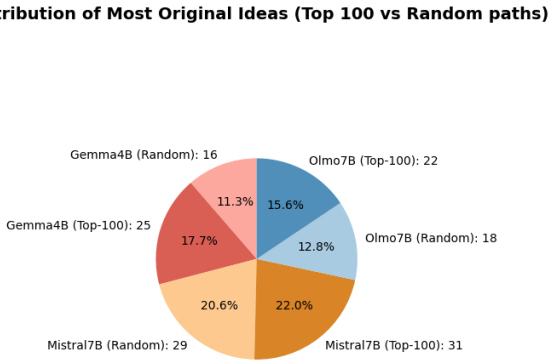

Figure 8: Number of highly original ideas by each model for random vs top-100 paths

source concept. For example, the path book → cover → book receives a score of 0.007, while book → floor → desk receives a score of 0.0003. These examples suggest that the model is learning to penalize at least some paths that may be less useful for generating original responses.

Finally, we test whether GNN-ranked paths improve generation quality relative to random path selection. We select the top 100 paths ranked by the GNN and compare them against 100 randomly sampled ConceptNet paths. The top-ranked paths produce 16–20% more semantic clusters than random paths as shown in Figure 7, indicating higher ideational diversity. In addition, all the ideas coming from the top 100 paths show higher originality for all three models as shown in Figure 8. These results suggests that the GNN can improve the graph-guided prompting pipeline by filtering ConceptNet paths before generation. Overall, RQ2 shows that a GNN trained on response diversity can identify paths that are more likely to expand the model's ideational range and originality, providing an additional layer of control over graph-guided creative generation.

5 Limitations

Our study has several limitations. First, like many other studies in this space, we focus primarily on originality and ideational diversity, which represent only part of the broader construct of creativity. Creativity is typically understood as requiring both novelty and usefulness; however, usefulness is difficult to assess reliably without human judgment, particularly in open-ended AUT responses where practical value may depend on context. Although our frequency-based clustering approach provides a useful proxy for corpus-relative originality, it does not determine whether a response is genuinely useful or even feasible. Future work should therefore incorporate human evaluation, or hybrid human-AI evaluation protocols, to assess usefulness alongside originality. Second, our graph-guided prompting strategy uses two-hop ConceptNet associations. We selected this distance because it offers a middle ground that works well with humans: concepts that are too close to the source object often produce obvious responses, while concepts that are too

distant may produce ideas that feel arbitrary or incongruent. However, this "Goldilocks" assumption may not transfer directly to LLMs. Models may benefit from different graph distances depending on their size, training data, decoding parameters, and the structure of the task. Future studies should systematically compare different relational paths to identify the distances that best balances novelty and usefulness for different models and domains. Third, our GNN path-selection experiment is currently limited to a single object, *book*. While the results suggest that GNNs can learn to identify ConceptNet paths that produce more diverse outputs than randomly selected paths, this finding should be interpreted as a feasibility result rather than a general claim. The learned path preferences may not transfer equally well to other objects, semantic neighborhoods, or creative tasks. Future work should train and evaluate GNNs over larger portions of ConceptNet, across many AUT objects, and with more diverse forms of supervision, including human-rated originality and usefulness. Finally, our experiments are limited to a small set of open-weight models and a constrained AUT setting. The observed effects may differ for larger models, or creative tasks beyond alternate uses. Similarly, ConceptNet provides interpretable commonsense structure, but it is incomplete and may reflect biases or gaps in its source data.

6 Conclusion

Improving the creative output of LLMs requires more than asking models to "be creative." Creativity depends on the ability to move beyond obvious associations while preserving coherence, usefulness, and task relevance. In this paper, we explored whether structured conceptual knowledge can support this process by guiding LLMs toward remote yet related associations.

We introduced a graph-guided associative prompting framework that uses two-hop ConceptNet paths to provide LLMs with additional conceptual triggers during the Alternate Uses Task. Across three open-weight models, we found that these graph-guided triggers consistently increased ideational diversity, producing more distinct semantic clusters and more responses in rare, high-originality clusters than standard prompting.

We also introduced a frequency-based semantic clustering analysis that adapts traditional divergent-thinking originality scoring to LLM outputs. By comparing each response to a generated reference corpus, this approach captures both corpus-relative originality and model-level diversity.

Finally, we demonstrated the feasibility of using graph neural networks to improve associative path selection. A relational GNN trained on path-level diversity signals was able to identify ConceptNet paths that produced greater ideational diversity than randomly selected paths. While this experiment was limited to a single object, it suggests that conceptual graphs can serve not only as static knowledge sources, but also as learnable search spaces for creative generation.

Taken together, our results suggest that structured conceptual knowledge offers a promising path toward more controllable and interpretable creativity in LLMs. Rather than relying solely on larger models, broader sampling, or generic creativity prompts, graph-guided methods can explicitly shape the conceptual space from which models generate ideas. Future work should extend this approach to broader ConceptNet subgraphs, additional creative tasks, larger models, and human evaluations of usefulness. More broadly, our findings point toward a view of machine creativity as guided exploration in a larger conceptual space.

References

[1] A. M. A. Alabbasi, S. H. Paek, D. Kim, and B. Cramond. What do educators need to know about the torrance tests of creative thinking: A comprehensive review. *Frontiers in psychology*, 13:1000385, 2022.

[2] A. G. Alhashim, A. M. Marshall, K. McManus, B. R. Shapiro, A. N. Katz, and S. K. Reed. Assessing creativity of alternative uses task responses: A detailed procedure. In *Proceedings of the 2020 ASEE Virtual Annual Conference*, 2020.

[3] Allen Institute for AI. allenai/OLMo-2-1124-7B-Instruct. `https://huggingface.co/allenai/OLMo-2-1124-7B-Instruct`, 2024.

[4] J. R. Anderson. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–295, 1983.

[5] R. E. Beaty, D. R. Johnson, D. C. Zeitlen, and B. Forthmann. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, 34(3):245–260, 2022.

[6] M. Benedek, C. Mühlmann, E. Jauk, and A. C. Neubauer. Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, 7(4):341–349, 2013.

[7] M. Benedek and A. C. Neubauer. Revisiting mednick's model on creativity-related differences in associative hierarchies. evidence for a common path to uncommon thought. *The Journal of creative behavior*, 47(4):273–289, 2013.

[8] T. B. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[9] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

[10] Google. google/gemma-3-4b-it. `https://huggingface.co/google/gemma-3-4b-it`, 2025.

[11] J. P. Guilford. Creativity. *American Psychologist*, 5(9):444–454, 1950.

[12] D. Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[13] B. Y. Lin, X. Chen, J. Chen, and X. Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. 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)*, pages 2829–2839, 2019.

[14] G. D. Ludden, B. M. Kudrowitz, H. N. Schifferstein, and P. Hekkert. Surprise and humor in product design: Designing sensory metaphors in multiple modalities. *Humor*, 25(3):285–309, 2012.

[15] Q. Luo, G. King, M. Puett, and M. D. Smith. Inducing sustained creativity and diversity in large language models. *arXiv preprint arXiv:2603.19519*, 2026.

[16] S. Mednick. The associative basis of the creative process. *Psychological review*, 69(3):220, 1962.

[17] P. Mehrotra, A. Parab, and S. Gulwani. Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*, 2024.

[18] Mistral AI. mistralai/Mistral-7B-Instruct-v0.3. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`, 2024.

[19] P. Organisciak, S. Acar, D. Dumas, and K. Berthiaume. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356, 2023.

[20] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous. Is temperature the creativity parameter of large language models? In *International Conference on Computational Creativity*, 2024.

[21] M. Rhodes. An analysis of creativity. *The Phi delta kappan*, 42(7):305–310, 1961.

[22] M. A. Runco and G. J. Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

[23] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

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

[25] M. Wadhwa, T. S. Roy, H. Lederman, J. J. Li, and G. Durrett. Create: Testing llms for associative creativity. *URL https://arxiv. org/abs/2603.09970*, 2026.

[26] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 535–546, 2021.

[27] M. J. Zhai. *The Mind as Interior Space in the Poetry of John Keats, Emily Dickinson, and Elizabeth Bishop*. PhD thesis, Wellesley College, 2024.

[28] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*, 2022.

A Prompts

Table 2 defines the variables used in the prompt templates. The Base condition uses only the target object, while the graph-guided condition additionally includes a two-hop ConceptNet trigger. In addition, all models were called with the following parameters: $max\_new\_tokens = 100, top\_p = 0.95, temperature = 0.9$.

Table 2: Variables used in the prompt templates.

Variable	Description
`{object}`	Target AUT object, e.g., *book*, *fork*, or *bottle*.
`{trigger}`	Two-hop ConceptNet concept used as an associative trigger.

A.1 Base condition

```
Your task is to generate an alternative, non-traditional use for a given everyday
object, beyond its usual or intended purpose.

You will return the alternate use of the given object along with a short, 1-sentence
description of how to use it.  Your output should be exactly one JSON object:
{
"Alternate Use":  ‘‘string’’,
"Description":  ‘‘string’’
}

Suggest a creative, alternate use for the following object:  {object}.
```

A.2 Graph-guided associative condition

```
Your task is to generate an alternative, non-traditional use for a given everyday
object, beyond its usual or intended purpose.

You will return the alternate use of the given object along with a short, 1-sentence
description of how to use it.  Your output should be exactly one JSON object:
{
"Alternate Use":  ‘‘string’’,
"Description":  ‘‘string’’
}

Suggest a creative, alternate use for the following object:  {object}.

In addition, integrate the following concept in your idea:  {trigger}.
```

B OCSAI Scores for Different Objects

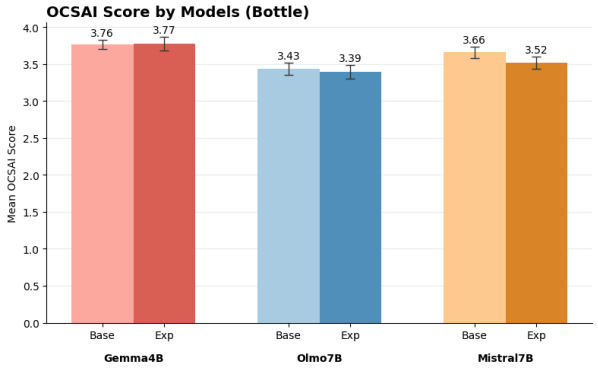

Figure 9: Average OCSAI originality scores for Bottle. Error bars correspond to 95% CI.

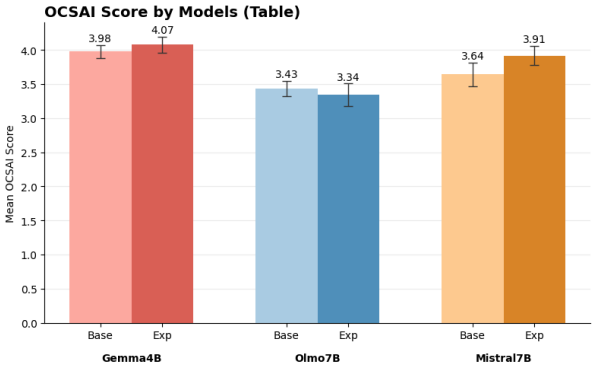

Figure 10: Average OCSAI originality scores for Table. Error bars correspond to 95% CI.

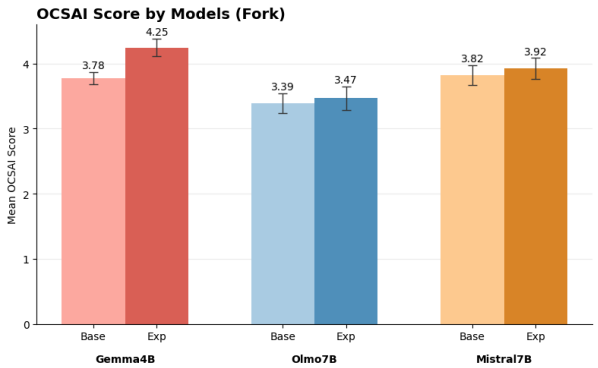

Figure 11: Average OCSAI originality scores for Fork. Error bars correspond to 95% CI.

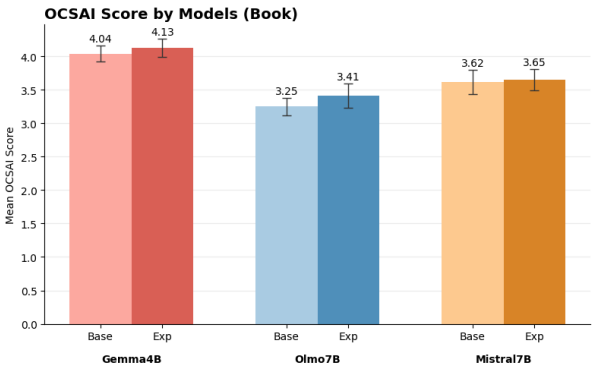

Figure 12: Average OCSAI originality scores for Book. Error bars correspond to 95% CI.

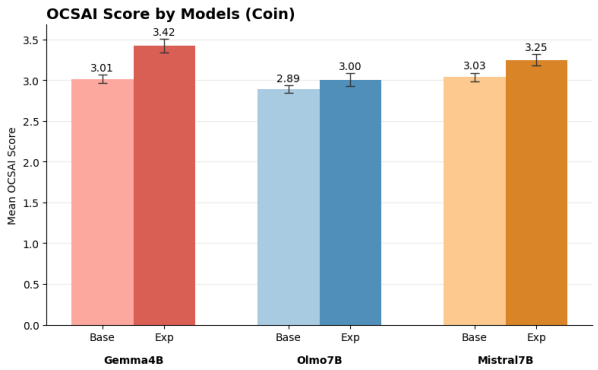

Figure 13: Average OCSAI originality scores for Coin. Error bars correspond to 95% CI.

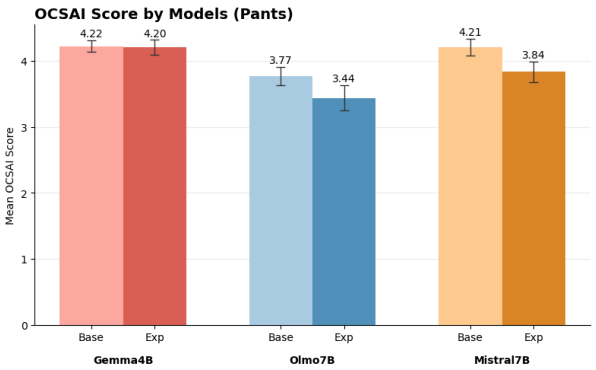

Figure 14: Average OCSAI originality scores for Pants. Error bars correspond to 95% CI.

C Count of Unique Clusters for Different Objects

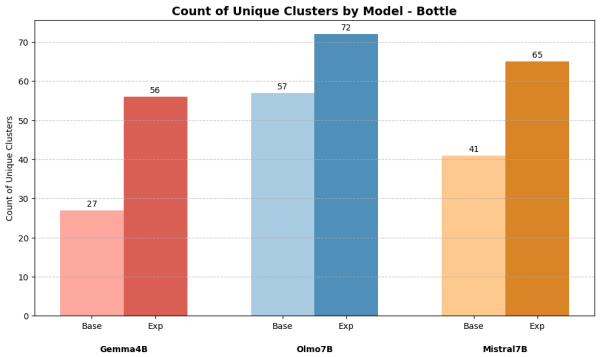

Figure 15: Count of unique clusters for Bottle.

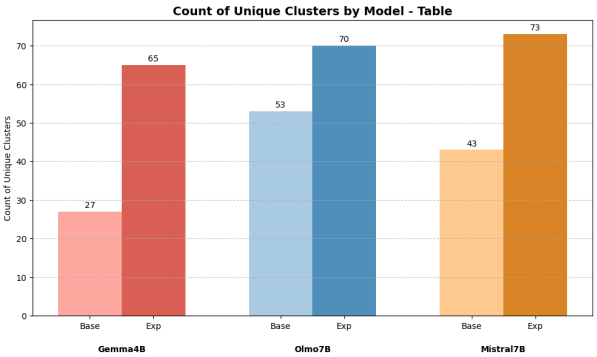

Figure 16: Count of unique clusters for Table.

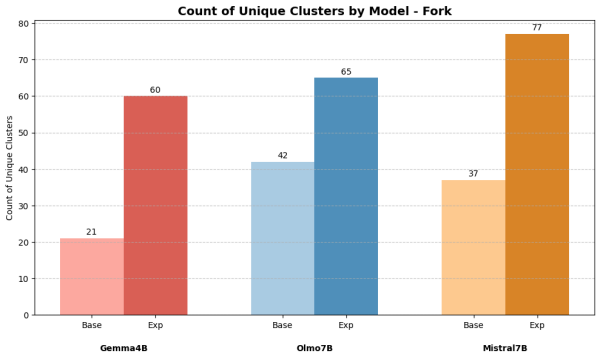

Figure 17: Count of unique clusters for Fork.

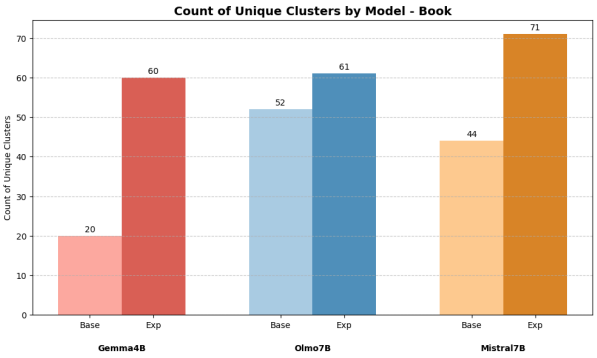

Figure 18: Count of unique clusters for Book.

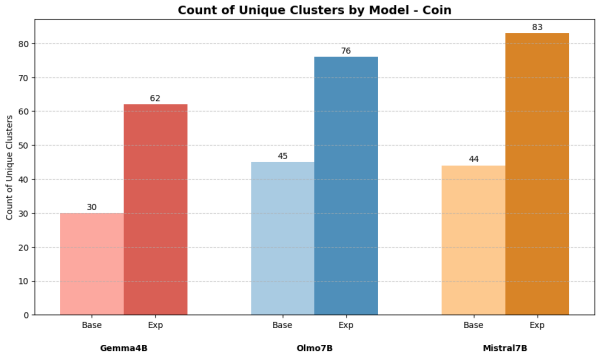

Figure 19: Count of unique clusters for Coin.

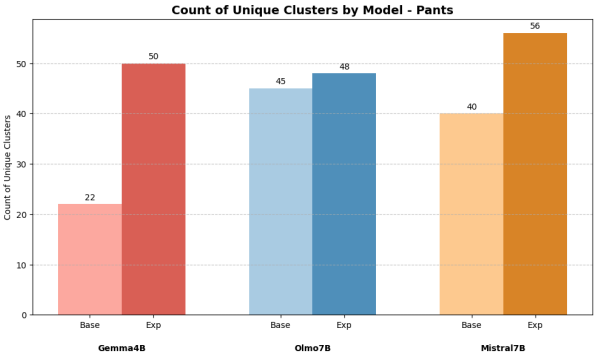

Figure 20: Count of unique clusters for Pants.