How Context Injection Shapes Creativity in LLM-Generated Alternative Uses
Paper type: Short Paper

Theo Wong, Fabricio Goes, Marco Volpe, Paulo Mann
School of Computing and Mathematical Sciences
University of Leicester, Federal University of Rio de Janeiro
tthw1@leicester.ac.uk, fabricio.goes@leicester.ac.uk, marco.volpe@leicester.ac.uk, paulomann@ic.ufrj.br

Abstract

Context injection influences large language model (LLM) outputs across many tasks, but its effect on divergent-ideation measures is underexplored. We systematically inject four types of context (object-focused, semantically related, unrelated, and random) into Alternative Uses Test (AUT) prompts and evaluate responses from four open-weight LLMs using both LLM-judged proxy metrics (creativity, novelty, value) and automated text metrics, including clustering-based flexibility. The main finding is metric divergence: no single context type optimises all evaluated dimensions at once. Instead, different context types steer outputs toward distinct evaluated profiles; for instance, random context maximises judged novelty while unrelated context uniquely increases categorical flexibility. These results indicate that context type determines the direction of measured output change rather than providing a uniform creativity boost, offering practical guidance for matching prompting strategy to target outcome.

Introduction

Understanding how to steer LLM creativity is a growing concern in computational creativity research. Large Language Models have shown competitive performance on creative tasks such as the Alternative Uses Test (AUT), in some cases matching or exceeding human baselines (Haase, Hanel, and Pokutta 2025). Several strategies have been proposed to enhance creative output, including prompt design (Meincke, Mollick, and Terwiesch 2024), sampling methods (Zhang et al. 2025), temperature tuning (Peeperkorn et al. 2024), multi-agent collaboration (Nguyen and Singla 2025), and preference optimisation (Ismayilzada et al. 2025). However, these approaches primarily vary the structure or parameters of generation rather than the semantic content supplied to the model. Context injection, the practice of providing supplementary factual information alongside a prompt, is known to shape LLM outputs in knowledge-intensive tasks. Retrieval-augmented generation has demonstrated that grounding outputs in external information improves answer relevance and accuracy (Lewis et al. 2020), and research on in-context learning has shown that the type, quality, and relevance of injected information significantly affect model behaviour (Dong et al. 2024). In creative tasks, semantically coherent context may encourage convergence toward expected solutions, while less related or random context may trigger more divergent exploration. Yet the application of context injection to creative tasks has received little systematic attention. This raises an empirical question: do different types of injected context shape creativity in the same way, or do they push it in different directions? We investigate this question using the AUT as a controlled experimental setting. The AUT asks participants to generate novel uses for everyday objects and is a standard benchmark for divergent thinking (Guilford 1967), requiring both fluency and flexibility and making it well suited for isolating the effect of context type on generated alternative uses. It does not, however, measure creativity as a whole: it focuses on divergent ideation rather than on the full social, cognitive, and evaluative process through which creative products are recognised. We develop four context-injection techniques (object-focused, semantically related, unrelated, and random) and compare them against a zero-shot baseline and a Verbalized Sampling prompt (Zhang et al. 2025) across four open-weight LLMs and six objects. Responses are evaluated with both LLM-judged proxy metrics (creativity, novelty, value) and automated text metrics. The main finding is that context type steers outputs toward distinct evaluated profiles rather than providing a uniform boost.

Related Work

Creativity research has long recognised that creative products must be both novel and valuable (Rhodes 1961; Gruszka and Tang 2017). In computational creativity, Cropley applies this duality to LLM outputs and argues that the generative mechanism constrains creativity to a ceiling near the little-c/Pro-c boundary (Cropley 2025), implying that prompting strategies operate within a bounded creative space. Our experiment does not test whether LLMs are "truly" creative; it asks a narrower question: how contextual input changes scores assigned to generated products under a fixed AUT-style evaluation pipeline. The Alternative Uses Test (AUT), designed to assess Guilford's divergent thinking dimensions of fluency, flexibility, originality, and elaboration (Guilford 1967), has become a standard benchmark for evaluating creative ideation. This makes the AUT useful for controlled comparisons of generated idea sets, but it also limits the scope of the claim: divergent think-

ing is one component of creativity, not a complete model of creative cognition or social valuation. Stevenson et al. showed that GPT-3 achieves human-comparable AUT performance (Stevenson et al. 2022), and subsequent work has used LLMs both as generators and evaluators of alternative uses (Al Rabeyah et al. 2025). Hadas et al. validated LLM-based AUT scoring against human expert judgments (Hadas, Avital-Lev, and Hershkovitz 2026), which is relevant here because we rely on LLM judges for creativity, novelty, and value ratings. Broader surveys have synthesised evidence on generative AI creativity while highlighting open methodological challenges around evaluation validity (Holzner, Maier, and Feuerriegel 2025; Haase, Hanel, and Pokutta 2025). Several strategies have been proposed to enhance LLM creative performance. Peeperkorn et al. showed that the temperature parameter is only weakly correlated with creativity dimensions (Peeperkorn et al. 2024), and Morain and Ventura demonstrated that prompt structure acts as a lever for creative output quality (Morain and Ventura 2025). Other approaches include multi-agent collaboration (Nguyen and Singla 2025), verbalized sampling (Zhang et al. 2025), and preference optimisation (Ismayilzada et al. 2025). These works vary the structure or parameters of generation, but do not systematically compare how different types of injected semantic content affect distinct creativity dimensions. Meincke et al. examined prompt design for increasing idea variance (Meincke, Mollick, and Terwiesch 2024), which is the closest precedent, though their focus is on prompt structure rather than on the semantic relationship between injected content and the creative task. Our study addresses this gap by systematically varying context type and measuring its effect on multiple creativity dimensions simultaneously.

Methodology

Experimental Setup and Prompt Design

The experiment follows the AUT format: each LLM generates 20 alternative uses for each of six everyday objects under each prompting condition. The objects are "bamboo", "chopsticks", "fork", "shoe", "soap", and "teapot", selected to vary in cultural specificity and familiarity; all are presented explicitly as physical household items to avoid polysemous readings.

To ensure that results reflect general patterns rather than single-model idiosyncrasies, we use four open-weight LLMs of comparable scale: Google/gemma-3-27b-it, Qwen/Qwen3-30B-A3B-Instruct-2507, and Mistralai/Mistral-Small-3.1-24B-Instruct-2503 as standard instruction-tuned models, and Qwen/QwQ-32B-AWQ as a reasoning-oriented model. Each model produces responses under six conditions, yielding 2880 responses in total (4 models $\times$ 6 objects $\times$ 6 conditions $\times$ 20 responses).

The six conditions are drawn from three prompt families, inspired by prior work (Meincke, Mollick, and Terwiesch 2024; Zhang et al. 2025). The zero-shot prompt asks the model to generate a single alternative use with no additional context and serves as the baseline. The context-injection prompts append a factual snippet to the same instruction; four variants differ only in the type of context provided (object-focused, semantically related, unrelated, and random). The Verbalized Sampling (VS) prompt asks the model to produce multiple candidate responses in a single call, each tagged with an explicit probability estimate, encouraging sampling from the full output distribution (Zhang et al. 2025). All prompts share a common structure with variables for the test object, the output format, and a list of previously generated answers to prevent duplicates within each 20-response set. Prompt templates are provided in the Appendix. For each object-condition pair, the same pre-generated list of 20 context snippets is reused across all four generator models so that between-model comparisons reflect model differences rather than different injected content.

Evaluation Metrics

Responses are evaluated on two complementary layers. Three LLM-judged scores (creativity, novelty, and value) provide product-level proxy assessments: "creativity" captures a composite judgment, while "novelty" and "value" probe the two theoretically central dimensions individually (Rhodes 1961). Each response is rated on a 1–100 scale by the same four models used for generation; the judge prompt is provided in the Appendix. We interpret these ratings as operational measures of how the evaluation models respond to generated products, not as direct evidence of human-perceived creativity. This distinction is important because creative value is socially situated, and because model judges may share training distributions, stylistic preferences, or failure modes with the generators.

Four automated text metrics provide a second, independent evaluation layer. **Flexibility** counts the number of distinct categories in a response set, estimated via K-means clustering on TF-IDF vectors with the silhouette-optimal k as the score (Nath, Dayan, and Stevenson 2024; Guilford 1967; Rousseeuw 1987). Because TF-IDF clustering is lexical, this measure may split semantically similar ideas expressed with different words; a semantic-embedding variant is therefore a natural robustness check. **METEOR** (Banerjee and Lavie 2005) and **Mutual Information (MI)** (Church and Hanks 1990) are pairwise overlap metrics computed across all response pairs (lower = more diverse). **MTLD** estimates vocabulary richness with reduced sensitivity to text length (McCarthy and Jarvis 2010); higher values indicate greater lexical diversity, though gains may partly reflect verbosity rather than idea-level novelty. We include Relative Standard Deviation (RSD) as a measure of cross-model consistency. Together, these automated metrics assess diversity-related properties of the generated text rather than creativity itself.

Results

To ground the quantitative results, Table 1 shows illustrative responses for the object "fork" under each prompting condition. Figures 1–7 show average scores across all models. All comparisons against the baseline were tested for statistical significance (t-tests (Semenick 1990), Mann–Whitney U (Collingridge 2013), permutation tests (Cheung, Pesigan,

Table 1: Illustrative examples for the object "fork", showing possible injected information and output for each condition.

Condition	Injected information	Example output
Baseline	N/A	"Loosening soil in a small potted plant"
VS	N/A	"When gardening, a fork is excellent for loosening soil in small pots or around delicate plants. The tines allow for aeration without causing significant root disturbance, especially useful for seedlings."
Object-focused	"Info about fork: The fork,s four tines maximize grip and minimize tearing; early two tine versions were for carving. Once derided as vanity in medieval Europe, it normalized by the 18th century, enabling blunter table knives and influencing pasta shapes designed to wind neatly around its tines."	"Carving meat, as early forks possessed only two tines specifically for this purpose"
Semantically related	"Info about pitchfork: A pitchfork typically has 3 spring-steel tines set on a 25 offset socket, optimizing lift with minimal wrist torque. Tine spacing is crop-tuned (wide for hay, narrow for manure). Curved, tapered tines reduce drag through plant mats; historical variants doubled as improvised polearms. Not a digging fork."	"Historically, a pitchforks tines could be utilized as an improvised polearm in self-defense or conflict."
Unrelated	"Info about friesian: Friesian horses, from Friesland, were rescued from near extinction in 1913 when only three approved stallions remained. The breed demands an all-black coat (a small star allowed). Their high-knee, trot and abundant feathering suit carriage work; they also carry recessive dwarfism and hydrocephalus alleles."	"A fork could be used to carefully tease out tangles in a Friesian horses feathering, though a specialized comb is preferred to avoid discomfort."
Random	"Info about virginal: Rectangular, plucked keyboard instrument of the harpsichord family; strings run parallel to keys. Favored in 16th and 17th-century England and Flanders, often bearing painted lids and Latin mottos. Core repertoire appears in the Fitzwilliam Virginal Book. Variants include muselar and spinet virginals, with contrasting timbres."	"A fork could be used to carefully pluck the strings of a virginal, though it wouldn't replicate the instrument's intended sound or playing experience."

and Vong 2023)) and effect size (Cohen's d (Cohen 1977), Hedges' g (Taylor and Alanazi 2023), bootstrap confidence intervals (Cheung, Pesigan, and Vong 2023)); full statistical tables are provided in the Appendix.

LLMs as Judges

Figures 1–3 show the average LLM-judged scores. Three patterns emerge. First, random context receives the highest judged creativity and novelty scores (88.145 vs. 69.567 and 85.433 vs. 69.009 baseline), with statistically significant gains. Second, object-focused and semantically related contexts receive the largest value scores (82.296 and 81.472 vs. 71.864 baseline), while unrelated context's value effect is inconclusive (t-test $p = 0.077$ vs. Mann–Whitney U $p = 0.010$). This discrepancy is expected when mean differences and rank-based distributional comparisons respond differently to skew, outliers, or non-uniform shifts across samples. We therefore treat such cases as distribution-sensitive rather than decisive.

Automated Metrics

Figures 4–7 show the automated text metrics, which reveal a different structure. Flexibility increases significantly only for unrelated context (12.375 vs. 9.417 baseline, $p = 0.002$). MTLD shows the most consistent gains: all conditions outperform baseline (all $p < 0.001$), with unrelated context

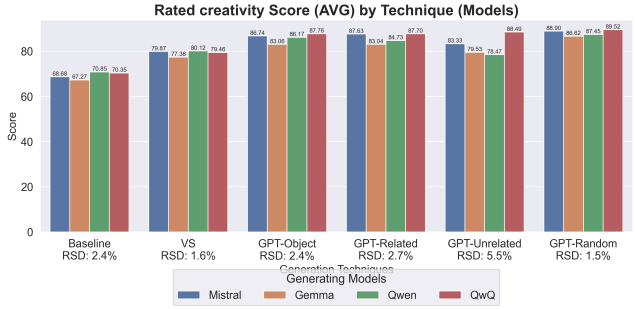

Figure 1: Creativity Scores

producing the largest increase (10.117 vs. 3.889). The overlap metrics diverge further: METEOR is higher (less diverse) for object-focused and unrelated context, whereas mutual information is lower (more diverse) for random context and higher (less diverse) for unrelated context. These patterns show that broader category spread, richer vocabulary, and lower lexical overlap are not interchangeable textual outcomes, and they caution against reading any single automated metric as creativity itself.

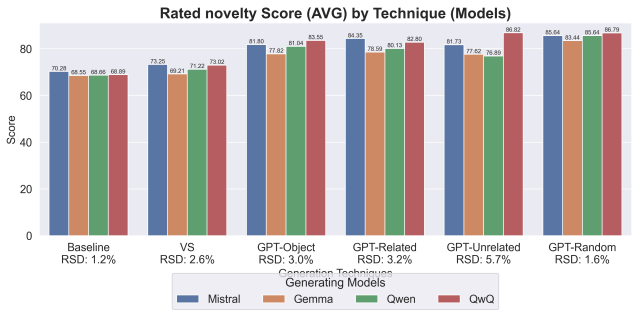

Figure 2: Novelty Scores

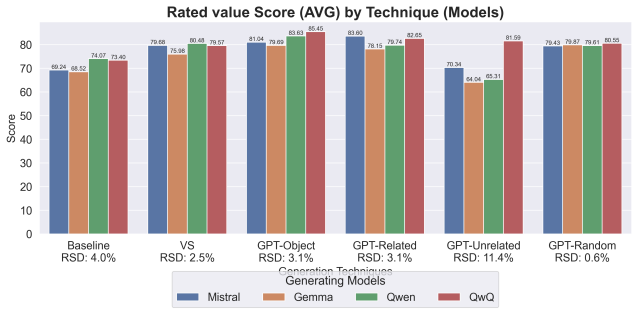

Figure 3: Value Scores

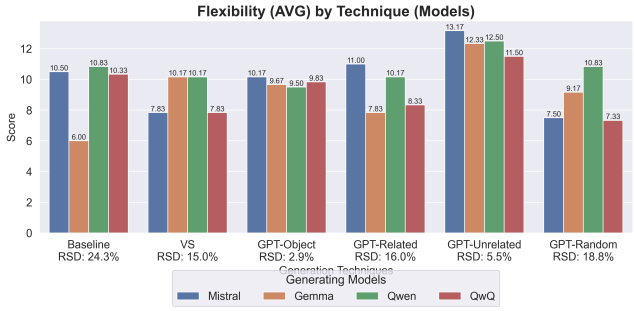

Figure 4: Flexibility Scores

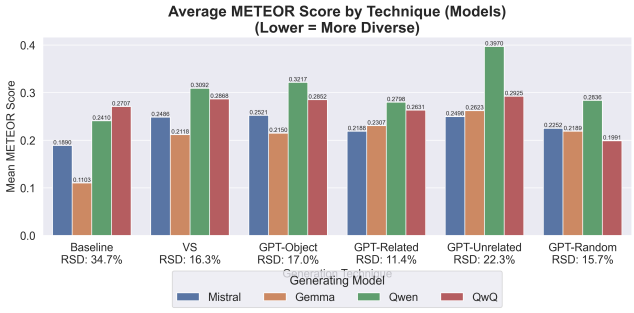

Figure 5: METEOR Scores

Discussion and Conclusion

The main finding of this study is that context type determines the measured direction of LLM outputs rather than providing a uniform creativity boost. All four context types improved over the baseline on multiple proxy metrics, but they did so along different axes, and these patterns were consistent across all six objects.

For practice, the results support a cautious decision rule: use random context when the goal is to maximize judge-perceived originality, use unrelated context when the goal is to broaden the lexical or categorical spread of ideas, and use object-focused or semantically related context when usefulness and groundedness matter most. The answer to which prompting strategy is "best" depends on the specific downstream task and evaluation method.

Future work should prioritise adding human judges and at least one independent LLM judge not used for generation to test whether the metric divergence observed here reflects genuine differences in perceived creative quality. Extending the evaluation to additional creative tasks beyond the AUT would strengthen the generalisability of the findings. On the prompting side, richer context families (e.g., cross-domain analogical or constraint-based context) and adaptive strategies that switch context type across stages of idea generation are promising directions.

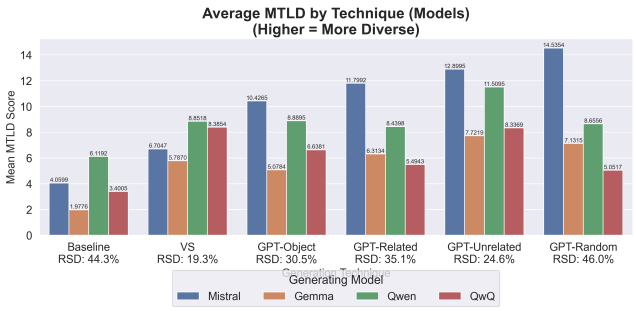

Figure 6: MTLD Scores

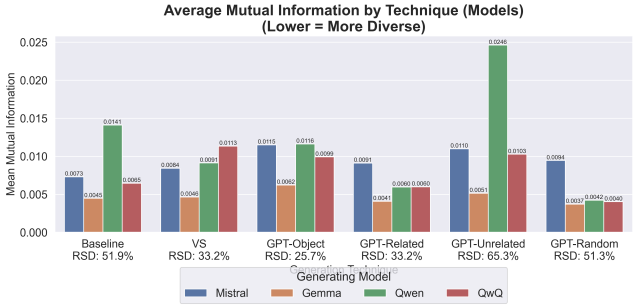

Figure 7: Mutual Information Scores

Appendix

This appendix provides prompt templates, context construction details, and full statistical tables. A repository with all source code, generated responses, and evaluation scores is available at: `https://anonymous.4open.science/r/ICCC-2026-66EA/README.md`.

Prompt Templates

All prompts share three variables: **problem** (the test object), **format** (always "`<answer>answer`

here</answer>"), and **ans_list** (previously generated answers, to prevent duplicates). The zero-shot prompt (Figure 8) provides the baseline. The context-injection prompt (Figure 9) is shared across all four variants; only the **info** field changes. The VS prompt (Figure 10) generates five candidates per iteration. Context snippets are produced by GPT-5 (OpenAI 2025) using the prompt in Figure 11.

```
{
   "role": "system",
   "content": f"You are an assistant
       generating a single answer for {
       problem}.",
},
{
   "role": "user",
   "content": f"Generate an answer for
       this problem:{problem}. Output one
        single answer in this exact
       format: {format}. Answers already
       used: {ans_list}",
},
```

Figure 8: Zero-shot prompt for baseline generation.

```
{
   "role": "system",
   "content": f"You are an assistant
       generating a single answer for {
       problem}.",
},
{
   "role": "user",
   "content": f"Generate a new answer
       about {problem} using the
       information here: {info[len(
       ans_list)]}. Output one single
       answer in this exact format: {
       format}. Answers already used: {
       ans_list}",
},
```

Figure 9: Context-injection prompt.

Figure 12 shows the judge prompt. Its variables are: **problem** (the test object), **metric** (creativity, novelty, or value), **ans_list** (answers to be rated), **rating_list** (previously rated answers), and **format** (always "<score>score here</score>").

Context Construction

The semantically related and unrelated context types select terms using NLTK WordNet's `path_similarity` function (NLTK 2025). For a given test object T, we compute the similarity of each WordNet lemma k against every synset of T, retaining the maximum non-null value as $S(T, k)$. The random condition draws terms from WordNet uniformly, bypassing the scoring pipeline (see Figures 15–16).

```
{
   "role": "system",
   "content": f"""You are a helpful
       assistant. For each query, please
       generate a set of {set_num}
       possible responses, \
each within a separate <response> tag.
    Responses should each include a <text
    > and a numeric <probability>. \
Please sample at random from the full
    distribution.""",
},
{
   "role": "user",
   "content": f"Generate an answer for
       this problem:{problem}. Answers
       already used: {ans_list}",
},
```

Figure 10: VS prompt, adapted from (Zhang et al. 2025).

```
{
   "role": "system",
   "content": "You are providing unique
       information of an object.",
},
{
   "role": "user",
   "content": f"Please provide unique
       information regarding {keyword}
       within 50 words.",
},
```

Figure 11: Context generation prompt.

The **keyword** variable changes across context types: for object-focused prompts it is the test object itself, while for the other three types it is a term selected through WordNet. Figure 13 shows the full retrieval and generation pipeline. Once scored, terms are sorted by $S(T, k)$ and selected according to the rules of each context type. Each selected term is then passed as the **keyword** variable to the context generation prompt (Figure 11), producing one snippet per term. The four context types are defined below.

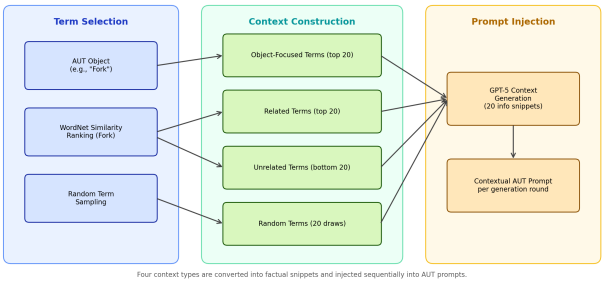

Figure 13: Context retrieval and generation pipeline.

```
{
   "role": "system",
   "content": f"You are an expert rating
       {problem} in terms of {metric}.",
},
{
   "role": "user",
   "content": f"Rate this {problem} from
       1 to 100 as an expert: {ans_list[
       len(rating_list)]}, give a single
       final score in this format: {
       format}",
},
```

Figure 12: LLM judge prompt.

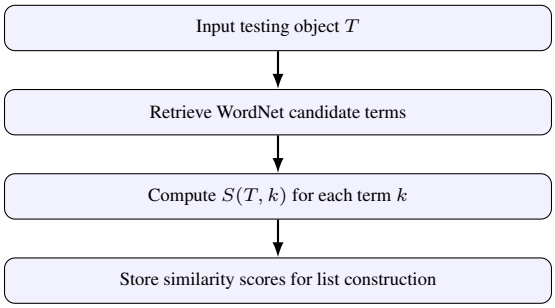

Figure 14: Process graph for term-similarity scores.

Object-Focused The **keyword** is the test object itself (e.g., "bamboo"). The list contains 20 context snippets about the object, with a duplicate-handling function that removes repeated snippets so that all 20 are unique.

Semantically Related The 20 highest-scoring terms form the related-term list, each used as **keyword** to generate one snippet (Figure 15).

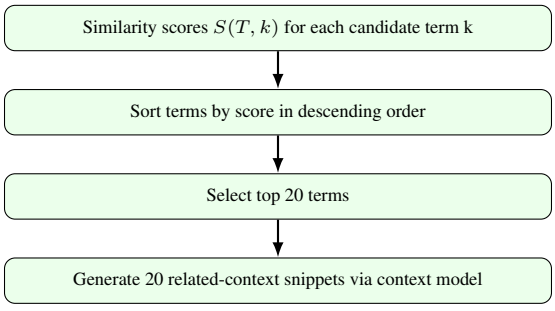

Figure 15: Process graph for semantically related keywords.

Unrelated The 20 lowest-scoring terms form the unrelated-term list, mirroring the related condition but sorted in ascending order (Figure 16).

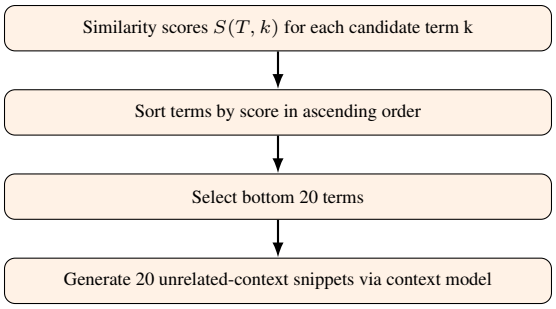

Figure 16: Process graph for the unrelated keyword list.

Random Twenty terms are sampled uniformly at random from the full WordNet lemma pool, with no reference to the similarity scores. Each sampled term is used as the **keyword** to generate one context snippet (Figure 17).

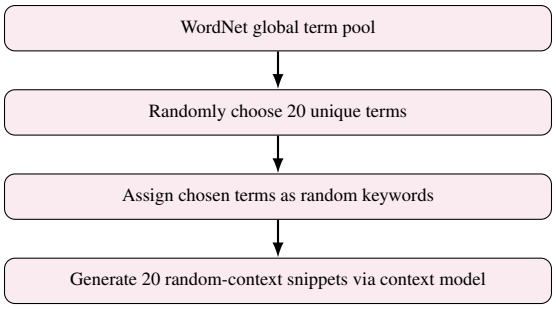

Figure 17: Process graph for the random keyword list.

Statistical Tables

Tables 2–8 report full statistical results for all metrics.

Table 2: Average Creativity score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	69.567	22.400	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	85.862	11.104	28.559	< 0.001	0.922	0.922	16.292	< 0.001	< 0.001
avg-related	86.011	11.423	28.656	< 0.001	0.925	0.925	16.441	< 0.001	< 0.001
avg-unrelated	83.038	18.070	20.509	< 0.001	0.662	0.662	13.466	< 0.001	< 0.001
avg-random	88.145	9.747	33.324	< 0.001	1.076	1.075	18.576	< 0.001	< 0.001
avg-VS	79.248	12.821	16.435	< 0.001	0.530	0.530	9.680	< 0.001	< 0.001

Table 3: Average Novelty score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	69.009	21.624	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	80.757	15.911	19.173	< 0.001	0.619	0.619	11.747	< 0.001	< 0.001
avg-related	81.645	14.917	21.076	< 0.001	0.680	0.680	12.632	< 0.001	< 0.001
avg-unrelated	81.100	17.894	18.876	< 0.001	0.609	0.609	12.084	< 0.001	< 0.001
avg-random	85.433	11.749	29.243	< 0.001	0.944	0.944	16.421	< 0.001	< 0.001
avg-VS	71.624	18.593	4.018	< 0.001	0.130	0.130	2.615	< 0.001	0.010

Table 4: Average Value score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	71.864	21.729	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	82.296	15.698	17.052	< 0.001	0.550	0.550	10.431	< 0.001	< 0.001
avg-related	81.472	16.404	15.463	< 0.001	0.499	0.499	9.605	< 0.001	< 0.001
avg-unrelated	70.475	26.627	1.771	0.077	0.057	0.057	1.392	0.073	0.010
avg-random	79.922	18.633	12.335	< 0.001	0.398	0.398	8.061	< 0.001	< 0.001
avg-VS	78.986	14.397	11.972	< 0.001	0.386	0.386	7.123	< 0.001	< 0.001

Table 5: Average Flexibility score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	9.417	3.400	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	9.792	3.162	0.396	0.694	0.114	0.112	0.384	0.723	1.000
avg-related	9.333	4.040	0.077	0.939	0.022	0.022	0.094	0.969	1.000
avg-unrelated	12.375	2.667	3.354	0.002	0.968	0.952	2.974	0.002	0.004
avg-random	8.708	2.971	0.769	0.446	0.222	0.218	0.708	0.470	0.339
avg-VS	9.000	2.993	0.451	0.654	0.130	0.128	0.407	0.683	0.507

Table 6: Average METEOR score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	0.203	0.111	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	0.268	0.064	2.514	0.016	0.726	0.714	0.066	0.016	0.038
avg-related	0.248	0.043	1.867	0.072	0.539	0.530	0.045	0.066	0.110
avg-unrelated	0.300	0.077	3.536	0.001	1.021	1.004	0.098	0.001	0.005
avg-random	0.232	0.049	1.166	0.252	0.337	0.331	0.029	0.254	0.348
avg-VS	0.264	0.056	2.419	0.021	0.698	0.687	0.061	0.019	0.036

Table 7: Average Mutual Information score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	0.008	0.005	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	0.010	0.005	1.196	0.238	0.345	0.340	0.002	0.232	0.158
avg-related	0.006	0.003	1.442	0.158	0.416	0.409	0.002	0.150	0.529
avg-unrelated	0.013	0.009	2.208	0.033	0.638	0.627	0.005	0.031	0.035
avg-random	0.005	0.003	2.113	0.041	0.610	0.600	0.003	0.040	0.054
avg-VS	0.008	0.005	0.198	0.844	0.057	0.056	0.000	0.841	0.477

Table 8: Average MTLD score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	3.889	3.250	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	7.758	2.821	4.404	0.000	1.271	1.251	3.874	0.000	0.000
avg-related	8.012	2.912	4.628	0.000	1.336	1.314	4.129	0.000	0.000
avg-unrelated	10.117	3.750	6.148	0.000	1.775	1.746	6.241	0.000	0.000
avg-random	8.844	4.331	4.482	0.000	1.294	1.273	4.974	0.000	0.000
avg-VS	7.432	2.327	4.342	0.000	1.253	1.233	3.542	0.000	0.000

References

Al Rabeyah, A.; Góes, F.; Volpe, M.; and Medeiros, T. 2025. Do LLMs agree on the creativity evaluation of alternative uses? In *Proceedings of the 16th International Conference on Computational Creativity*, 217–227. Campinas, Brazil: Association for Computational Creativity.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Cheung, S. F.; Pesigan, I. J. A.; and Vong, W. N. 2023. DIY bootstrapping: Getting the nonparametric bootstrap confidence interval in spss for any statistics or function of statistics (when this bootstrapping is appropriate). *Behavior research methods* 55(2):474–490.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cohen, J. 1977. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Collingridge, D. S. 2013. A primer on quantitized data analysis and permutation testing. *Journal of mixed methods research* 7(1):81–97.

Cropley, D. H. 2025. "The cat sat on the . . . ?" Why generative AI has limited creativity. *The Journal of creative behavior* 59(4).

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. Miami, Florida, USA: Association for Computational Linguistics.

Gruszka, A., and Tang, M. 2017. The 4p's creativity model and its application in different fields. In Tang, M., and Werner, C. H., eds., *Handbook of the Management of Creativity and Innovation: Theory and Practice*. World Scientific. chapter 3, 51–71.

Guilford, J. P. 1967. *The nature of human intelligence*. New York: McGraw-Hill.

Haase, J.; Hanel, P. H. P.; and Pokutta, S. 2025. Has the creativity of large-language models peaked?: An analysis of inter- and intra-LLM variability. *Journal of Creativity* 35(3):100113.

Hadas, E.; Avital-Lev, B.; and Hershkovitz, A. 2026. Validating LLM-based alternative uses test scoring across ages. *Thinking Skills and Creativity* 60:102066.

Holzner, N.; Maier, S.; and Feuerriegel, S. 2025. Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis. *arXiv e-prints* arXiv:2505.17241.

Ismayilzada, M.; Laverghetta, Jr., A.; Luchini, S. A.; Patel, R.; Bosselut, A.; van der Plas, L.; and Beaty, R. 2025. Creative Preference Optimization. *arXiv e-prints* arXiv:2505.14442.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.-F.; and Lin, H.-T., eds., *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 9459–9474. Curran Associates, Inc.

McCarthy, P. M., and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42(2):381–392.

Meincke, L.; Mollick, E. R.; and Terwiesch, C. 2024. Prompting Diverse Ideas: Increasing AI Idea Variance. *arXiv e-prints* arXiv:2402.01727.

Morain, R., and Ventura, D. 2025. Is prompt engineering the creativity knob for large language models? In *Proceedings of the 16th International Conference on Computational Creativity*, 30–40. Campinas, Brazil: Association for Computational Creativity.

Nath, S. S.; Dayan, P.; and Stevenson, C. 2024. Characterising the creative process in humans and large language models. In *Proceedings of the 15th International Conference on Computational Creativity*, 325–330. Jönköping, Sweden: Association for Computational Creativity.

Nguyen, M. H., and Singla, A. 2025. Divergent-Convergent Thinking in Large Language Models for Creative Problem Generation. *arXiv e-prints* arXiv:2512.23601.

NLTK. 2025. NLTK: Sample usage for wordnet — nltk.org. [Accessed 02-12-2025].

OpenAI. 2025. GPT-5 is here — openai.com. [Accessed 02-12-2025].

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity*, 226–235. Jönköping, Sweden: Association for Computational Creativity.

Rhodes, M. 1961. An analysis of creativity. *The Phi Delta Kappan* 42(7):305–310.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65.

Semenick, D. 1990. Tests and measurements: The T-test. *Strength & Conditioning Journal* 12(1):36–37.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, 392–398. Association for Computational Creativity.

Taylor, J. M., and Alanazi, S. 2023. Cohen's and hedges' g. *The Journal of nursing education* 62(5):316–317.

Zhang, J.; Yu, S.; Chong, D.; Sicilia, A.; Tomz, M. R.; Manning, C. D.; and Shi, W. 2025. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. *arXiv e-prints* arXiv:2510.01171.