Deep Associations, High Creativity: A Simple yet Effective Metric for Evaluating Large Language Models

Anonymous ACL submission

Abstract

The evaluation of LLMs' creativity represents a crucial research domain, though challenges such as data contamination and costly human assessments often impede progress. Drawing inspiration from human creativity assessment, we propose PACE, asking LLMs to generate **P**arallel **C**hains of **A**ssociations to **E**valuate their creativity. PACE minimizes the risk of data contamination and offers a straightforward, highly efficient evaluation, as evidenced by its strong correlation with Arena Creative Writing (Spearman's $\rho = 0.739$, $p < 0.001$) on various proprietary and open-source models. A comparative analysis of associative creativity between LLMs and humans reveals that while high-performing LLMs achieve scores comparable to average human performance, top-performing humans consistently outperform LLMs. Furthermore, linguistic analysis reveals that both humans and LLMs exhibit a trend of decreasing concreteness in their associations, and humans demonstrating a greater diversity of associative patterns.

1 Introduction

Developing creative artificial intelligence and boosting co-creativity remain central goals in AI research (Rafner et al., 2023; Franceschelli and Musolesi, 2024; Lee and Chung, 2024). Current research conduct diverse creativity-based tasks to evaluate the creative capabilities of Large Language Models (LLMs), aiming to understand their potential and limitations (Tian et al., 2023; Atmakuru et al., 2024; Si et al., 2024).

However, data contamination, a prominent issue in current LLMs evaluations, may compromise the reliability of conclusions (Sainz et al., 2023; Xu et al., 2024; Lu et al., 2024a). Moreover, unlike tasks with definitive answers, establishing frameworks to evaluate creativity poses unique challenges, particularly due to its complex nature (Rafner et al., 2023; Ivcevic and Grandinetti, 2024)

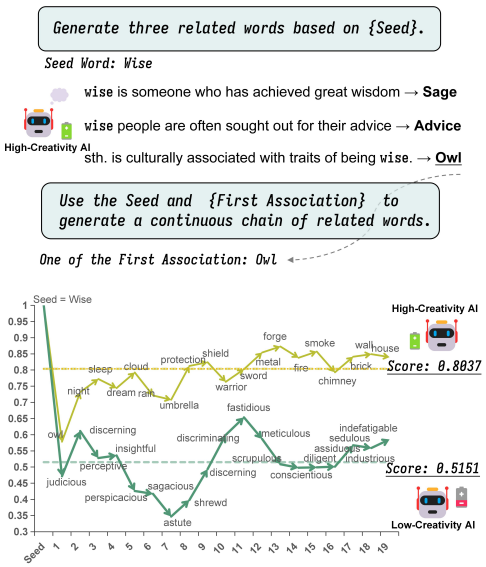

Figure 1: Structure of PACE: Three 20-word chains are generated for each seed. The average association distance of each chain is calculated to represent its score.

and the subjective and time-consuming process of human scoring (Olson et al., 2021; Organisciak et al., 2023; Lu et al., 2024b).

In light of these issues, this study draws inspiration from established psycholinguistic measures of human creativity and introduces PACE (Parallel Association Chain Evaluation), a highly efficient framework to evaluate LLMs. As shown in Figure 1, This approach requires no human-annotated data and enables automatic and reliable scoring. Associative evaluation lies at the core of human creativity research (Mednick and Halpern, 1968; Olson et al., 2021; Beaty and Kenett, 2023). The theory of associative creativity posits that individuals with higher creative capacity are more likely to generate unconventional connections, enabling them to link disparate concepts and produce original ideas (Mednick, 1962; Merseal et al., 2023). As for LLMs, measuring associative distance efficiently assesses their capacity for creative association, reflecting their ability to move beyond surface

co-occurrence patterns and tap into deeper, less common semantic links that underlie genuine creativity (Yao et al., 2022; Abramski et al., 2024).

Our results demonstrate a strong correlation between PACE and Arena creative writing ($\rho = 0.739$, $p < 0.001$), as well as other LLM leaderboards, through testing a series of open-source and closed-source models of varying capabilities. We further compare associative creativity between humans and LLMs, finding that state-of-the-art models perform comparably to general human groups, but still fall short of professional humans. Linguistic analysis reveals that both models and humans tend to produce associations with decreasing concreteness; however, human associations are generally more abstract and exhibit greater diversity in association types.

2 Method

2.1 Parallel Word Association Chains

The ability to generate distant associations is a key indicator of creativity, as it reveals unconventional connections between concepts and ideas (Mednick, 1962; Kenett et al., 2014; Zhang et al., 2023). Similarly, advanced models are expected to capture multi-level semantics and identify deeper connections, enabling them to foster novel insights.

To systematically evaluate this capability, we present a two-phase approach inspired by human participant studies from Gray et al. (2019). The approach consists of: (1) eliciting three distinct associations from LLMs as secondary seed words, and (2) generating 20-word association chains that contain both primary and secondary seeds.

Each association chain is generated independently to minimize mutual influence among the chains. Compared to single-chain association, this parallel approach improves the diversity of associative pathways, allowing a broader sampling of the model's creative potential. For each independent chain, we employ a chain-of-thought prompting strategy to guide the model's word associations [1], ensuring a structured yet flexible generation process. Prompts can be found in Appendix A.3.

[1]While multi-turn dialogue could also be used to elicit associations, generating without conversational history often results in redundant outputs. Conversely, providing full conversational history introduces confounds such as long-context memory and coherence constraints inherent to multi-turn or recursive setups. As a result, the two-step approach yields interpretable and controlled measurements of creative associative capacity, aligning with both human experimental paradigms and computational evaluation scenarios.

2.2 Seed Words

110 seed words are selected from the Intercontinental Dictionary Series (IDS, Key and Comrie, 2023), a multilingual project representing universal concepts across languages. The IDS consists of 22 chapters, each corresponding to a distinct semantic domain, such as time, quality, and motion. From each chapter, five seed words are chosen based on their frequency distribution in the Corpus of Contemporary American English (COCA; Davies, 2008), using five equally spaced frequency intervals to ensure balanced representation. This selection process combines semantic diversity and frequency variation to enable a comprehensive evaluation. For each model, three chains are generated for each seed word, resulting in 6,270 associated words. The complete list of seed words is provided in Appendix A.3.

2.3 Association Distance Metric

We measure the creativity score using the mean association distance. Each seed's score is derived by averaging the association distances of three chains, and the model's overall associative creativity is determined by averaging the scores of 110 seeds. See details in Appendix A.3. We use FastText (crawl-300d-2m; Mikolov et al., 2018) for computing cosine distance. Table 2 also reports results using alternative word embedding models.

3 Experiments and Results

3.1 Models and Parameters

Thirty models are selected from the Chatbot Arena Leaderboard, covering a diverse range of ranks and licenses (commercial and open-source). When comparing with other benchmarks that have relatively few models on their leaderboards, we included at least 18 models to ensure a robust correlation analysis (Bonett and Wright, 2000). Multiple versions and sizes of Qwen models were added to examine the relationship between scale and performance. The full list of models is in Table 6. Model responses are obtained via APIs with a temperature of 0, except for o3-mini (fixed at 1), while other parameters remained default.

3.2 Correlation with Existing Benchmarks

We select several representative benchmarks to validate our results, including the Chatbot Arena leaderboard (Arena All and Arena CW, which ranks models based on human voting preferences for

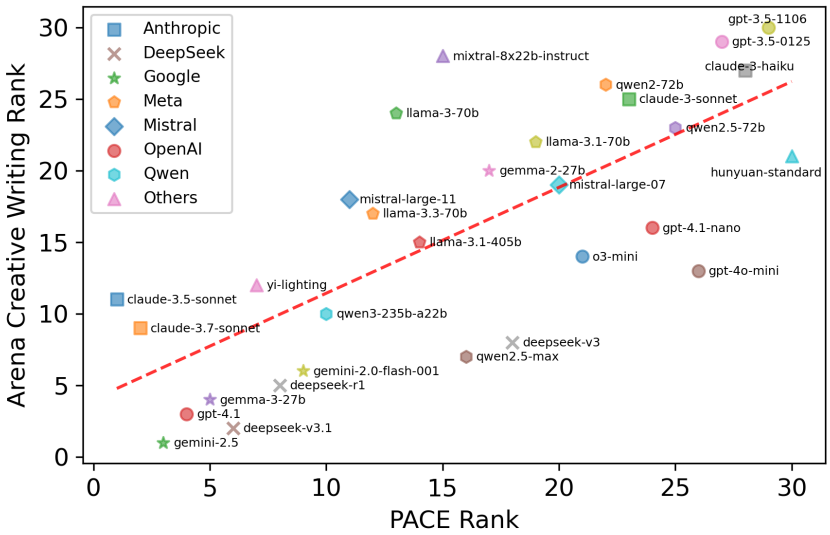

Figure 2: Spearman Rank Correlation Between Model Rankings Based on Association Distance and Arena Creative Rankings ($r = 0.739$, $p < 0.001$). Claude-3.5-Sonnet demonstrates the largest association distance.

Table 1: Spearman Rank Correlation between Model Rankings based on Association Distance and Different Benchmarks

Leaderboard	Corr.	P-value	Models
Arena All	0.660***	< 0.001	30
Arena CW	0.739***	< 0.001	30
MMLU-Pro	0.505*	< 0.05	23
LiveBench	0.691**	< 0.01	19
EQ-Bench	0.637**	< 0.01	18

* p < 0.05, ** p < 0.01, *** p < 0.001

anonymous models, Chiang et al., 2024), MMLU-Pro (a more complex and challenging version of Massive Multitask Language Understanding, Wang et al., 2024), livebench (releasing new questions regularly, White et al., 2024), EQ-Bench (specifically its creative writing leaderboard, scored by LLMs, Paech, 2023). We then calculate the ranks of the models in each leaderboard and their association scores.

3.3 Results

Association Distance Shows Significant Correlations with LLM Creative Ranks. As illustrated in Figure 2 and Table 1, the correlation between PACE and various benchmarks ranges from moderate to strong. Bootstrap analysis confirms the robustness of these correlations, with detailed results presented in Table 3.

As Figure 2 shows, models from the same organization with similar structures can exhibit different PACE rankings, e.g., DeepSeek-V3.1 scored 0.763 (rank 6), DeepSeek-R1 scored 0.759 (rank 8), and DeepSeek V3 scored 0.748 (rank 19), demonstrating PACE's effectiveness in differentiation. Additionally, we analyze various versions and sizes of the Qwen model, which provides diverse open-source variants for comparison (see Appendix A.2).

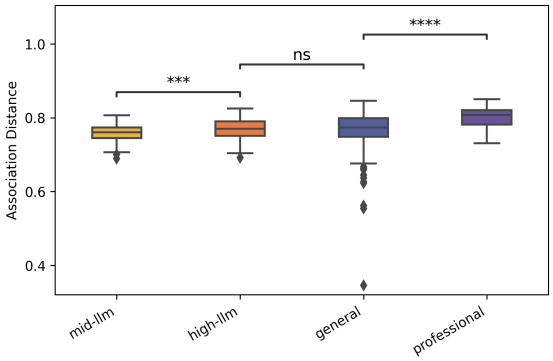

Figure 3: Comparison of Association Distances Between Humans and LLMs. Using human data from Gray et al. (2019), results show that **high-performing LLMs match average human performance**, but fall short of professional humans.

4 Comparison between Humans and Models

4.1 Associative Creativity

We compare human and LLM performance on associative creativity tests. For humans, we use data from Gray et al. (2019), including general American participants and professional performers. For LLMs, we evaluate top-20 models and those ranked around 75, using the same seed words for both groups. Details on seeds, model groups, and prompts are in Appendix A.3.

Current Leading LLMs Match Average Human Creativity. In terms of overall performance,

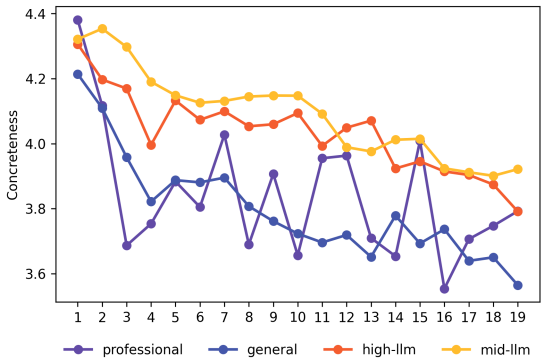

Figure 4: Concreteness scores for human and model responses in the association chain task decline across chain positions, with models consistently showing higher concreteness than humans. Details are provided in Appendix A.4.

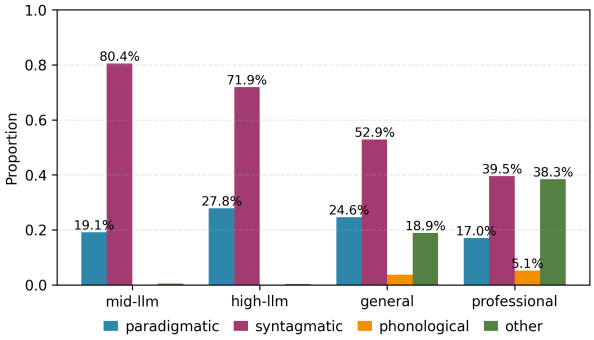

Figure 5: Types of associations within chains, categorized according to the four-class taxonomy developed by Nissen and Henriksen (2006). Details are provided in Appendix A.4.

high-performing models have achieved statistical parity with human control groups, as evidenced by the results of Welch's t-test ($t = 0.644$, $p = 0.52$), surpassing previous studies that reported lower model performance compared to human participants (Wenger and Kenett, 2025). Furthermore, significant performance differences are observed between high-performing models and mid-performing models ($t = 3.781$, $p < 0.001$).

Best-performing Human Still Outperforms LLMs. Both overall group scores ($t = 6.152$, $p < 0.001$) and the highest values from the human group ($\text{Human}_{\text{max}}$=0.8501, $\text{Model}_{\text{max}}$=0.8251) show that the best-performing humans still outperform the best LLMs in agreement with previous research (Koivisto and Grassini, 2023). Besides, a significant difference between the professional group and other groups can also be observed. This result demonstrates the irreplaceable role of human creativity (Rafner et al., 2023; Lee and Chung, 2024; Boussioux et al., 2024). However, LLMs demonstrate greater consistency in minimum performance ($\text{Human}_{\text{min}}$=0.3457, $\text{Model}_{\text{min}}$=0.6888), suggesting their potential as reliable co-creativity tools for generating consistent solutions (Dell'Acqua et al., 2023; Jia et al., 2024; Lee and Chung, 2024; Ashkinaze et al., 2024).

4.2 Associative Patterns

We further compare the patterns of association between human and LLMs from two aspects: trend of associations and type between associations.

Trend of Associations. As shown in Figure 4, both humans and LLMs exhibit a decreasing trend in concreteness. However, the model consistently demonstrates higher levels of concreteness compared to humans. This suggests that the model tends to rely more on concrete concepts rather than abstract ones, whereas humans are more inclined toward abstract cognition. Additionally, while both LLMs and the general human population display a relatively steady decline in concreteness, professionals exhibit greater variability, suggesting more frequent transitions in their associations (Kenett et al., 2014; Zhang et al., 2023).

Type of Associations. Similar to humans, LLMs exhibit a stronger tendency to generate syntagmatic associations (words that co-occur in sequences, like "dog" $\rightarrow$ "bark") compared to paradigmatic associations (words that can substitute for each other, like "dog" $\rightarrow$ "cat"). However, human associate more diversely, generating non-semantic relationships such as phonological connections. Moreover, association patterns among professionals show a tendency in "other" type of association, suggesting that creative individuals tend to form associations based on personal experiences rather than common linguistic patterns.

5 Conclusions

We propose PACE as a benchmark to evaluate LLMs. Our findings demonstrate a strong and significant correlation between PACE scores and several established benchmarks, e.g. $\rho = 0.739$ with Arena CW. Our results prove that measuring associative distance provides an efficient way to assess a LLMs' capacity for creative association, reflecting its ability to move beyond surface co-occurrence patterns and tap into deeper, less common semantic links that underlie genuine creativity.

Limitations

Limited Focus on English. Since we use English seed words, rankings in Arena Creative Writing (with English prompts), and English word embeddings, the evaluation of PACE is conducted in English, focusing on its correlation with creativity performance. Consequently, our results are limited to the assessment of English creative ability.

Limited Sample Model Sizes. To validate robustness indirectly, we rely on rankings from other leaderboards, which restrict the selection of models due to the limited number of models available in those leaderboards. Additionally, to ensure comparability across different leaderboards, we select models that are commonly present in all leaderboards, further narrowing the range of models available for analysis. Based on Bonett and Wright (2000), Spearman correlations in the range of $|\rho| \approx 0.5 - 0.7$, typically require a sample size of 20-30 to achieve reliable confidence intervals. For the main results related to Arena CW, we report a sufficient number of models, although in other leaderboards, the number of models is close to the expected threshold, which may slightly affect the robustness of the conclusions.

References

Katherine Abramski, Riccardo Improta, Giulio Rossetti, and Massimo Stella. 2024. The" llm world of words" english free association norms generated by large language models. *arXiv preprint arXiv:2412.01330*.

Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How ai ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. *arXiv preprint arXiv:2401.13481*.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv preprint arXiv:2410.04197*.

Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683.

Douglas G Bonett and Thomas A Wright. 2000. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65:23–28.

Léonard Boussioux, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R Lakhani. 2024. The crowdless future? generative ai and creative problem-solving. *Organization Science*, 35(5):1589–1607.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA). Online corpus. Available online at https://www.english-corpora.org/coca/.

Simon De Deyne, Chunhua Liu, and Lea Frermann. 2024. Can gpt-4 recover latent semantic relational information from word associations? a detailed analysis of agreement with human-annotated semantic ontologies. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024*, pages 68–78.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Michael M Flor. 2024. Three studies on predicting word concreteness with embedding vectors. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024*, pages 140–150.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, pages 1–11.

Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S Christian, John Patrick, Laura Huang, Yoed N Kenett, and Kevin Lewis. 2019. "forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist*, 74(5):539.

Zak Hussain, Rui Mata, Ben R Newell, and Dirk U Wulff. 2024. Probing the contents of semantic representations from text, behavior, and brain data using the psychnorms metabase. *arXiv preprint arXiv:2412.04936*.

Zorana Ivcevic and Mike Grandinetti. 2024. Artificial intelligence as a tool for creativity. *Journal of Creativity*, 34(2):100079.

Nan Jia, Xueming Luo, Zheng Fang, and Chengcheng Liao. 2024. When and how artificial intelligence augments employee creativity. *Academy of Management Journal*, 67(1):5–32.

Yoed N Kenett, David Anaki, and Miriam Faust. 2014. Investigating the structure of semantic networks in low and high creative persons. *Frontiers in human neuroscience*, 8:407.

Mary Ritchie Key and Bernard Comrie, editors. 2023. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601.

Byung Cheol Lee and Jaeyeon Chung. 2024. An empirical investigation of the impact of chatgpt on creativity. *Nature Human Behaviour*, 8(10):1906–1914.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, et al. 2024a. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2024b. Benchmarking language model creativity: A case study on code generation. *arXiv preprint arXiv:2407.09007*.

Martha T Mednick and Sharon Halpern. 1968. Remote associates test. *Psychological Review*.

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

Hannah M Merseal, Simone Luchini, Yoed N Kenett, Kendra Knudsen, Robert M Bilder, and Roger E Beaty. 2023. Free association ability distinguishes highly creative artists from scientists: Findings from the big-c project. *Psychology of Aesthetics, Creativity, and the Arts*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Henriette Bagger Nissen and Birgit Henriksen. 2006. Word class influence on word association test results 1. *International Journal of Applied Linguistics*, 16(3):389–408.

Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Janet Rafner, Roger E Beaty, James C Kaufman, Todd Lubart, and Jacob Sherson. 2023. Creativity in the age of generative ai. *Nature Human Behaviour*, 7(11):1836–1838.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Melanie Walsh, Anna Preus, and Elizabeth Gronski. 2024. Does chatgpt have a poetic style? *arXiv preprint arXiv:2410.15299*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

Peiran Yao, Tobias Renwick, and Denilson Barbosa. 2022. Wordties: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970.

Jingyi Zhang, Kaixiang Zhuang, Jiangzhou Sun, Cheng Liu, Li Fan, Xueyang Wang, Jing Gu, and Jiang Qiu. 2023. Retrieval flexibility links to creativity: evidence from computational linguistic measure. *Cerebral cortex*, 33(8):4964–4976.

A Appendix

A.1 Full Results

Validation of Different Embedding Models. To validate the correlation, we employ three widely-used English word embeddings to compute association distances: GloVe (GloVe-6B-300D; Pennington et al., 2014), MUSE (English; Conneau et al., 2017), and FastText (crawl-300d-2m; Mikolov et al., 2018). Results presented in Table **??** demonstrate a consistently significant correlation between PACE rankings and Arena Creative Writing (Arena CW) scores , with MUSE achieving the highest correlation coefficient ($\rho = 0.76$). To ensure alignment with the concreteness prediction (Details can be found in Appendix A.4), we use FastText to show our results.

Table 2: Spearman Correlation Results Across Different Word Embedding Methods

Leaderboard	Glove	Muse	FastText	Models
Arena CW	0.529**	0.757***	0.739***	30
Arena All	0.488**	0.675***	0.660***	30
MMLU-Pro	0.383	0.555**	0.505*	23
Livebench	0.490*	0.651***	0.691***	19
EQ-Bench	0.304	0.796***	0.637**	18

* p < 0.05, ** p < 0.01, *** p < 0.001

Bootstrap Results for Correlation Analysis. To validate the robustness of the correlation coefficient, we employ a bootstrap method to randomly select the results of seed words and compute Spearman correlation. Except for MMLU-Pro (with a significant ratio of 0.96), other leaderboards demonstrate a stable and significant correlation (with a significant ratio of 1.00) with PACE rankings. Among these, Arena-CW achieve the highest correlation with PACE, with Spearman correlation values ranging from 0.67 to 0.77, indicating a strong relationship.

Table 3: Bootstrap Results for Spearman Correlation Across Different Leaderboard

Leaderboard	Mean Corr.	Std. Corr.	95% CI	Sig. Ratio
Arena CW	0.726***	0.023	[0.678, 0.769]	1.000
Arena All	0.650***	0.023	[0.602, 0.695]	1.000
MMLU-Pro	0.489*	0.045	[0.405, 0.578]	0.962
LiveBench	0.669***	0.031	[0.607, 0.725]	1.000
EQ-Bench	0.624**	0.043	[0.537, 0.714]	1.000

* p < 0.05, ** p < 0.01, *** p < 0.001

Reducing Elements Cause Lower Correlation But Stay Significant. We explore methods to optimize evaluation efficiency by modifying two key parameters: the number of seed words and chain length. Using random sampling with 500 iterations, we select various subsets of seed words. Additionally, we analyze the impact of different chain lengths by truncating the original chains and computing Spearman's rank correlation coefficients.

Table 4: Impact of Reducing Seed Nums

Leaderboard	Num-1	Num-2	Num-3	Num-4
Arena CW	0.587 (0.048)	0.609 (0.034)	0.613 (0.025)	0.617 (0.017)
Arena All	0.598 (0.050)	0.621 (0.035)	0.626 (0.025)	0.630 (0.019)
MMLU-Pro	0.439 (0.084)	0.453 (0.056)	0.465 (0.045)	0.471 (0.033)
LiveBench	0.589 (0.071)	0.604 (0.051)	0.613 (0.034)	0.612 (0.027)
EQ-Bench	0.649 (0.080)	0.673 (0.056)	0.681 (0.040)	0.686 (0.027)

Values in parentheses indicate Standard Deviations.

Table 5: Impact of Reducing Chain Length

Leaderboard	Length-5	Length-10	Length-15	Length-20
Arena CW	0.582***	0.698***	0.717***	0.739***
Arena All	0.502**	0.618***	0.637***	0.660***
MMLU-Pro	0.249	0.479*	0.461*	0.505*
LiveBench	0.558*	0.632**	0.633**	0.691**
EQ-Bench	0.370	0.554*	0.562*	0.637**

* p < 0.05, ** p < 0.01, *** p < 0.001

The results demonstrate that larger sample sizes yield higher correlation coefficients, indicating enhanced performance stability.

A.2 Supplementary Analysis of Results

PACE Captures Subtle Difference Between Models. We merge results from different Qwen model to examine how model versions and parameter sizes would affect association distances. As Figure 6shows, there are clear distinctions in association abilities across different model series, with Qwen models demonstrating a consistent version-based hierarchy: Qwen-3, Qwen-2.5, and Qwen-2. In addition, while lower-performing groups typically comprise smaller models of different versions (e.g., Qwen-2-7b, Qwen-2.5-3b), larger models with older versions can match the performance of newer versions (e.g., Qwen-2-72b). This result highlights how architectural improvements and increased parameter counts represent two distinct but complementary paths for advancing model performance.

Subjective and Abstract Semantic Categories Differentiate Models' Performance. As Figure 7 shows, while newer models generally outperform older ones, the performance gap varies significantly across semantic categories. All model show great performance in some subjective category, e.g., spatial relationship, time, quantitiy. Even earlier version with small sizes achieve relatively high performance (>0.71). However, for subjective and ab-

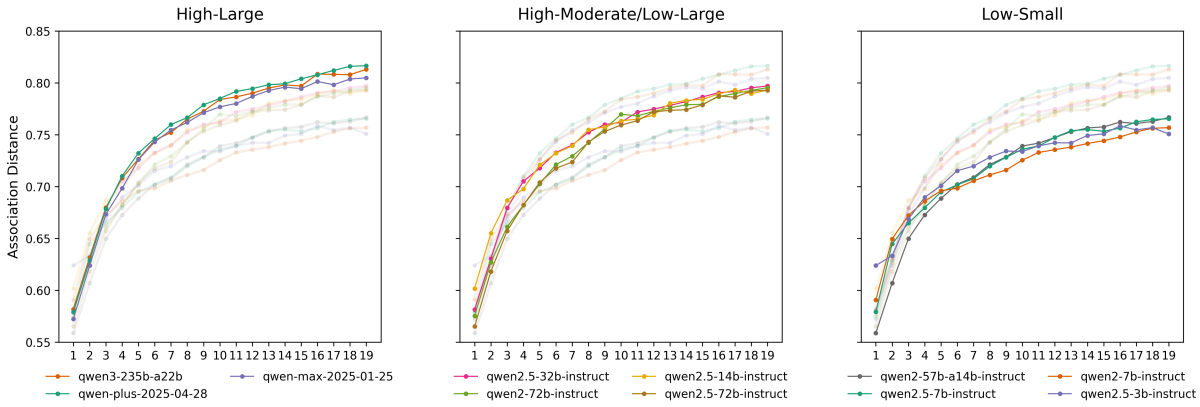

Figure 6: **Association distance comparison across versions and sizes of Qwen models.** This figure represents the association distance calculated at each position within the associative chains across different models and versions. Results reveal three performance clusters at different chain positions: (1) **high-large models** (new architectures, larger parameters), (2) **high-moderate and low-large models** (mixed newer models with moderate parameters and older models with larger parameters), and (3) **low-small models** (smaller architectures, fewer parameters). These findings highlight the combined effect of model version and parameter size and validate PACE as an effective evaluation framework.

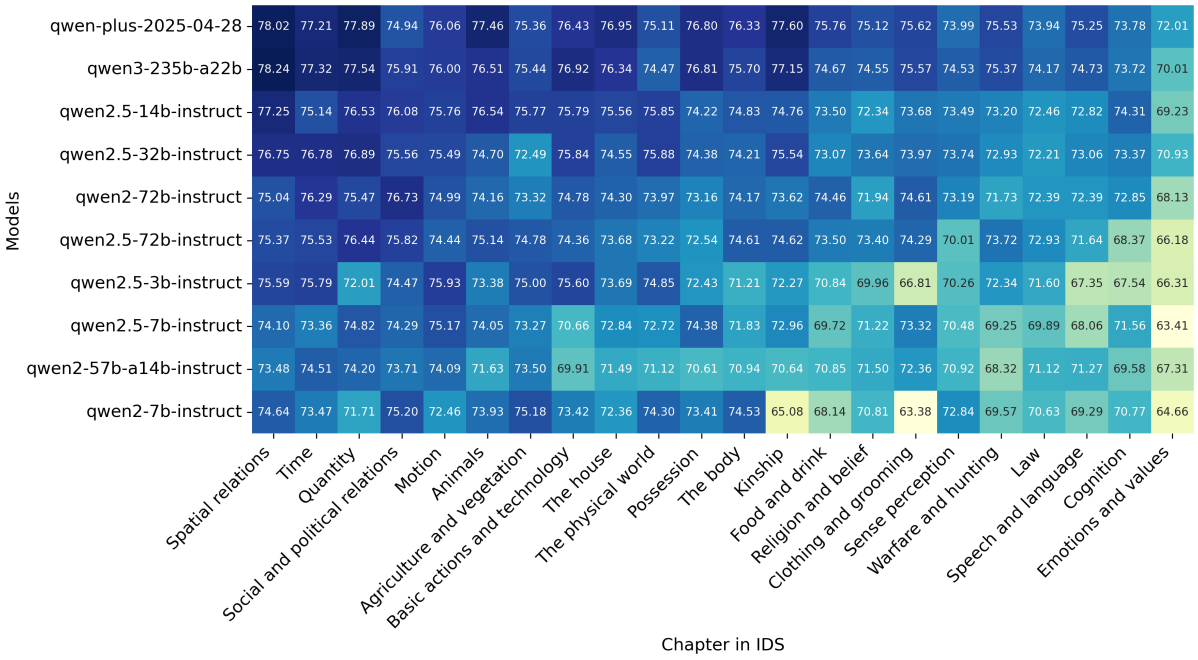

Figure 7: Association Distance Sorted by Chapters in IDS

stract categories, e.g., Emotions and values (0.63-0.72), kinship (0.65-0.78), the performance gap between model generations widens substantially, with newer models demonstrating up to 10 percentage point improvements over their previous versions. Notably, these subjective and abstract elements often constitute the core components of creative writing.

Humans Generate More Diverse Associations Than LLMs. We combine the responses generated by the model and humans, respectively, and standardize the sample sizes for each seed word to eliminate potential biases arising from varying data sizes. The analysis of the responses using the Type-Token Ratio (TTR) revealed distinct patterns in lexical diversity between LLMs and human participants. Despite the use of prompts specifically designed to enhance response diversity in LLMs, their TTR values consistently remained lower than those of human participants across all seed words. This finding suggests that LLMs produce more homogeneous responses compared to humans, underscoring their limitations as substitutes for human creative output (Walsh et al., 2024; Wenger and Kenett, 2025).

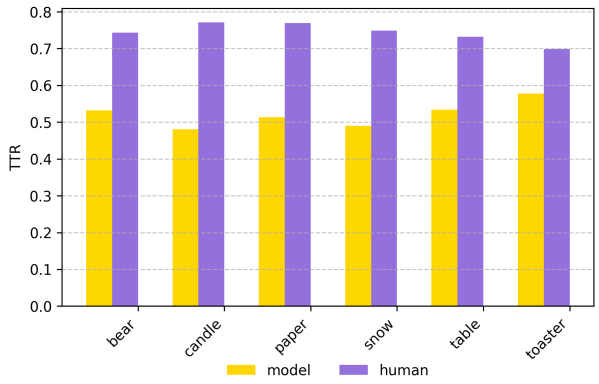

Figure 8: TTR of Responses from Models and Human

A.3 Experimental Details

Selected Models. Full list of selected models can be found in Table 6. PACE evaluation contains a comprehensive selection of LLMs, featuring both prominent closed-source commercial models (including various versions of Gemini, GPT, and Claude series) and leading open-source models (such as DeepSeek, Qwen, Gemma, and LLaMA series). This selection provides a balanced view of the current state-of-the-art in both commercial and open-source models, with 34 models in total being evaluated. Among the 34 total models evaluated, 30 models have Chatbot Arena scores (early May 2025 scoring version, Chiang et al., 2024), while four additional models (command-r-plus-08-2024, deepseek-r1-distill-llama-70b, deepseek-r1-distill-qwen-32b, and hunyuan-turbos-20250313) are included to ensure minimum model coverage across other leaderboards despite lacking Arena CW scores.

Model	License	Arena CW	Association Distance
gemini-2.5-pro-preview-03-25	-	1450	0.7757
deepseek-chat-v3-0324	✓	1376	0.7628
gpt-4.1-2025-04-14	-	1364	0.7728
deepseek-r1	✓	1356	0.7588
gemini-2.0-flash-001	-	1348	0.7576
qwen3-235b-a22b	✓	1314	0.7553
gemma-3-27b-it	✓	1358	0.7673
qwen-max-2025-01-25	-	1334	0.7505
deepseek-v3	✓	1331	0.7480
o3-mini-2025-01-31	-	1270	0.7388
claude-3.7-sonnet	-	1316	0.7817
yi-lightning	-	1282	0.7614
claude-3.5-sonnet	-	1289	0.7885
gpt-4o-mini-2024-07-18	-	1270	0.7297
gpt-4.1-nano	-	1256	0.7340
hunyuan-standard	-	1244	0.7171
llama-3.1-405b-instruct	✓	1264	0.7521
llama-3.3-70b-instruct	✓	1255	0.7542
qwen2.5-72b-instruct	✓	1228	0.7339
mistral-large-2407	✓	1246	0.7429
mistral-large-2411	✓	1246	0.7548
llama-3.1-70b-instruct	✓	1239	0.7476
gemma-2-27b-it	✓	1245	0.7488
llama-3-70b-instruct	✓	1214	0.7532
claude-3-sonnet	-	1188	0.7345
qwen2-72b-instruct	✓	1184	0.7371
claude-3-haiku	-	1163	0.7236
mixtral-8x22b-instruct	✓	1147	0.7515
gpt-3.5-turbo-0125	-	1099	0.7283
gpt-3.5-turbo-1106	-	1044	0.7226
command-r-plus-08-2024	✓	-	0.7397
deepseek-r1-distill-llama-70b	✓	-	0.7461
deepseek-r1-distill-qwen-32b	✓	-	0.7437
hunyuan-turbos-20250313	-	-	0.7260

Table 6: Selected Models with Arena CW Scores (Cut-off: Early May 2025) and Their Association Distances

Selected Seed Words. The final set of 110 seed words was selected through a two-step process. First, using NLTK part-of-speech tagger, we identified nouns by filtering for words with the "NN" prefix, as nouns frequently serve as stimuli in association experiments. While our initial focus was on nouns, we included all identified words in our dataset since words from different syntactic categories can effectively trigger associations. Second, we ranked these words based on their frequency in COCA2020 (Davies, 2008), divided the corpus into five equal segments, and selected the final words based on this stratification.

Formula for Association Distance. Our association distance measurement builds upon Gray et al. (2019). For each position n in an association chain, we calculate the association distance as the average semantic distance from the current position to all preceding positions:

Chapter	Seed
The physical world	rock, wood, dust, rainbow, headland
Kinship	son, female, widow, son-in-law, stepdaughter
Animals	eagle, worm, dove, firefly, midge
The body	sick, toe, blink, eyelid, earwax
Food and drink	meal, pepper, crush, ripe, unripe
Clothing and grooming	spin, soap, bracelet, braid, awl
The house	bed, pole, ladder, chimney, cookhouse
Agriculture and vegetation	grass, mushroom, bamboo, sickle, banyan
Basic actions and technology	strike, broken, cord, glue, adze
Motion	push, lift, swim, dive, outrigger
Possession	seek, hire, possess, lend, stingy
Spatial relations	center, ball, collect, round, fathom
Quantity	piece, count, pair, twelve, multitude
Time	month, summer, yesterday, cease, timepiece
Sense perception	dark, dry, rough, sour, brackish
Emotions and values	pain, correct, anxiety, sadness, deceit
Cognition	seem, explain, reflect, wise, imitate
Speech and language	speak, refuse, confess, howl, rebuke
Social and political relations	subject, neighbor, plot, ruler, chieftain
Warfare and hunting	peace, defeat, bow, fortress, fishhook
Law	murder, judgment, punishment, plaintiff, arson
Religion and belief	pray, temple, fairy, phantom, portent

Table 7: Chapters and their associated seed words

$$A_n = \frac{\sum_{i=1}^{n-1} D_{n,i}}{n-1}, \tag{1}$$

where $D_{n,i}$ represents the semantic distance between positions n and i, capturing the conceptual relatedness between thoughts at these positions.

The association distance of an entire sequence is then calculated by averaging the association distances across all positions:

$$A_{\text{chain}} = \frac{\sum_{i=2}^{n} \left(\frac{\sum_{j=1}^{i-1} D_{i,j}}{i-1} \right)}{n-1}, \tag{2}$$

where n is the total number of positions in the association chain.

To enhance diversity of LLMs' responses, we generate three association chains for each seed. The association distance for each seed is computed by averaging the three chain scores:

$$A_{\text{seed}} = \frac{1}{3} \sum_{c=1}^{3} A_{\text{chain},c}, \tag{3}$$

Finally, the overall association distance metric for a model is derived by averaging across all seeds:

$$A_{\text{model}} = \frac{1}{S} \sum_{s=1}^{S} A_{\text{seed},s}, \tag{4}$$

where S represents the total number of seeds evaluated.

Prompts. We use a two-step approach to construct parallel association chains. First, we generate prompts based on the methodology proposed by Gray et al. (2019), incorporating more detailed instructions to clearly articulate task requirements. This modification addresses our observation that certain lower-tier language models tend to generate associations consistently based on the seed word rather than the immediately preceding word. Additionally, we require models to provide reasoning for each association between consecutive words, which serves two purposes: ensuring adherence to task specifications and enhancing label accuracy in association type classification.

To compare different LLMs, we set the temperature parameter to zero to observe their intrinsic associative patterns (with the exception of o3-mini, which has a fixed temperature setting of 1). For comparisons between LLMs and human responses, we use both zero and one temperature settings to obtain a broader range of responses.

First Stage Prompt

Starting with the word "{seed}", generate three different words that directly associate with this initial word only (not with each other). Please put down only single words, and do not use proper nouns (such as names, brands, etc.). For each word, provide a brief explanation of its connection to "{seed}". Return in JSON format:

```
{
  "results": [
    {"word": "", "reason": ""},
    {"word": "", "reason": ""},
    {"word": "", "reason": ""}
  ]
}
```

Second Stage Prompt

Starting with the word pair "{seed}" → "{second_word}", generate a chain of 20 words where each new word should be associated with ONLY the word immediately before it. Generate the third word based on "{second_word}", then generate the fourth word based on your third word, and so on. Please put down only single words, and do not use proper nouns (such as names, brands, etc.). For each word, provide a brief explanation of its connection to the previous word. Return in JSON format with exactly 20 entries:

```
{
  "results": [
    {"word": "{second_word}",
    "reason": "{second_word_reason}"},
    {"word": "", "reason": ""},
    {"word": "", "reason": ""},
    ...
    {"word": "", "reason": ""}
  ]
}
```

Settings for Comparison Between Human and LLMs. In Section 4, we compare the performance between LLMs and humans, using human participant data from Gray et al. (2019). Specifically, we

analyze data from two groups: Group 2 (representative American samples) serving as the general population group, and Group 4 (professional actors) representing the professional expertise group. The professional actors' group demonstrate superior performance, achieving the highest ratings in both the original evaluations and original validation tests.

For LLM analysis, we select two parallel groups based on their Chatbot Arena Rankings. The high-performing group comprises four LLMs ranked within the top 20: DeepSeek-Chat-v3.1, Gemini-2.5-Pro-03-25-preview, Qwen3-235b-a22b, and GPT-4.1. The mid-performing group includes Yi-Lightning, Gemma-2-27b-it, LlaMA-3.3-70b-Instruct, and Mistral-Large-2411, with an average ranking of 75 in the Arena leaderboard, representing the standard performance of current models. This selection includes models from different organizations to ensure fair TTR calculations (details can be found in Figure 8).

For seed words, we select the same set used in human participant trials to ensure valid comparisons: bear, table, candle, snow, paper, and toaster. To achieve a comparable sample size with human responses, we generated LLMs'responses by varying the temperature parameter between 0 and 1. In this section, three association chains independently, rather than using averaged values for each seed word in the section of correlation analysis, thereby better simulating abundant LLM participants. Consequently, each model generated six chains (three chains plus two temperature settings) per seed word.

A.4 Supplementary Experiment Details.

Labelling Association Type. Given that LLMs have demonstrated the ability to identify various types of associations (De Deyne et al., 2024), we use DeepSeek-V3.1 to classify the semantic relationships between consecutive word pairs in each association chain. The classification adhered to the association type framework established by Nissen and Henriksen (2006), which categorizes relationships as syntagmatic, paradigmatic, phonological, or other.

Prediction of Concretness Using Embedding Model. Word embeddings can effectively predict various psychological dimensions of lexicon (Charbonnier and Wartena, 2019; Flor, 2024; Hussain et al., 2024).

We used the concreteness dataset developed by Brysbaert et al. (2014), one of the largest human-labeled concreteness databases, to train three embedding-based prediction models: FastText (English), GloVe (6B-300d), and MUSE (English). Model performance are evaluated using Pearson's correlation coefficient, root mean square error (RMSE), and Kendall's rank correlation. Table 9 indicate that FastText achieved the highest Pearson correlation and Kendall coefficient, as well as the lowest RMSE. Consequently, we finally use FastText to assign concreteness ratings to association responses.

Table 8: Fixed Effects Model Results Across Groups

Group	Intercept (β_0)	Slope (β_1)	p-value	R^2
Professional	3.994	-0.017	1.04e-3	0.275
	[3.823, 4.166]	[-0.027, -0.007]		
General	4.030	-0.026	1.32e-53	0.350
	[3.969, 4.091]	[-0.029, -0.022]		
High LLM	4.212	-0.020	4.32e-16	0.277
	[4.132, 4.292]	[-0.025, -0.015]		
Mid LLM	4.305	-0.024	3.67e-23	0.249
	[4.231, 4.378]	[-0.028, -0.019]		

Note: 95% confidence intervals in brackets.
Degrees of freedom: Professional = 626, General = 5,642, High LLM = 2,697, Mid LLM = 2,640. Since some model responses do not meet the required length of 20, these instances are considered missing values. Consequently, we exclude them from the calculations as their absence may impact the overall results.

Table 9: Comparison of word embedding models for concreteness prediction

Model	Pearson r	Kendall	RMSE
Training Set			
FastText	**0.931 ± 0.000**	**0.760 ± 0.001**	**0.371 ± 0.001**
GloVe 6B	0.902 ± 0.001	0.728 ± 0.001	0.442 ± 0.001
MUSE	0.848 ± 0.001	0.658 ± 0.001	0.541 ± 0.001
Test Set			
FastText	**0.910 ± 0.002**	**0.722 ± 0.003**	**0.421 ± 0.004**
GloVe 6B	0.837 ± 0.004	0.638 ± 0.004	0.556 ± 0.006
MUSE	0.845 ± 0.004	0.654 ± 0.004	0.545 ± 0.005

Note: Values shown as mean ± standard deviation. **Bold** indicates best performance. Valid words: FastText (35,424), GloVe (31,617), MUSE (27,101).

Table 8 provides detailed information on the fixed effects, with group differences modeled as fixed effects. In this analysis, position within the association trend serves as the independent variable (X), and concreteness is treated as the dependent variable (Y).