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

Ziliang Qiu
University of Illinois, Urbana-Champaign
Beijing Normal University
ziliang6@illinois.edu

Renfen Hu ✉
Beijing Normal University
irishu@mail.bnu.edu.cn

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 **A**ssociation **C**hains 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 Chatbot Arena Creative Writing rankings (Spearman's $\rho = 0.739$, $p < 0.001$) across 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, professional 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]

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). Recent studies evaluate the creative capabilities of large language models (LLMs) through diverse tasks, aiming to understand their strengths and limitations (Tian et al., 2023; Atmakuru et al., 2024; Si et al., 2024).

However, data contamination, a prominent issue in current LLM 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) and the subjective and time-consuming process of human scoring (Olson et al., 2021; Organisciak et al., 2023; Lu et al., 2024b).

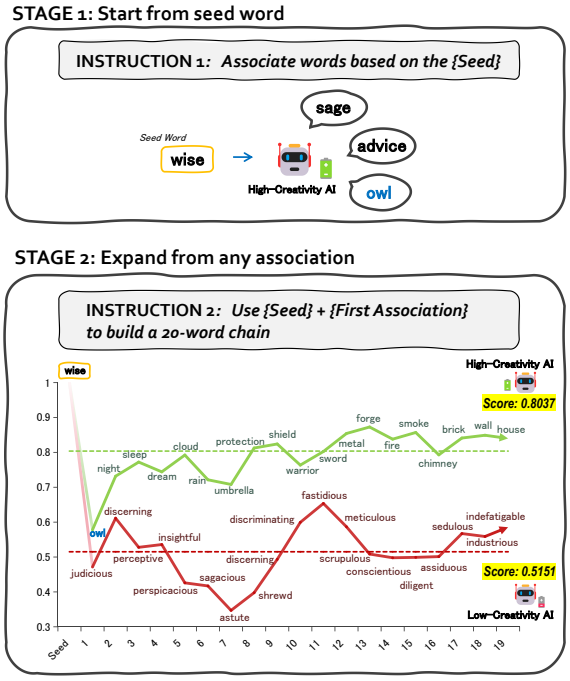

Figure 1: PACE evaluation process: For each seed word, three 20-word association chains are generated and their association distances are averaged to obtain the seed score. The model's creativity score is calculated by averaging all seed scores.

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

[1]Our code and data are publicly available at https://github.com/ziliang6/PACE

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10871–10884
November 4-9, 2025

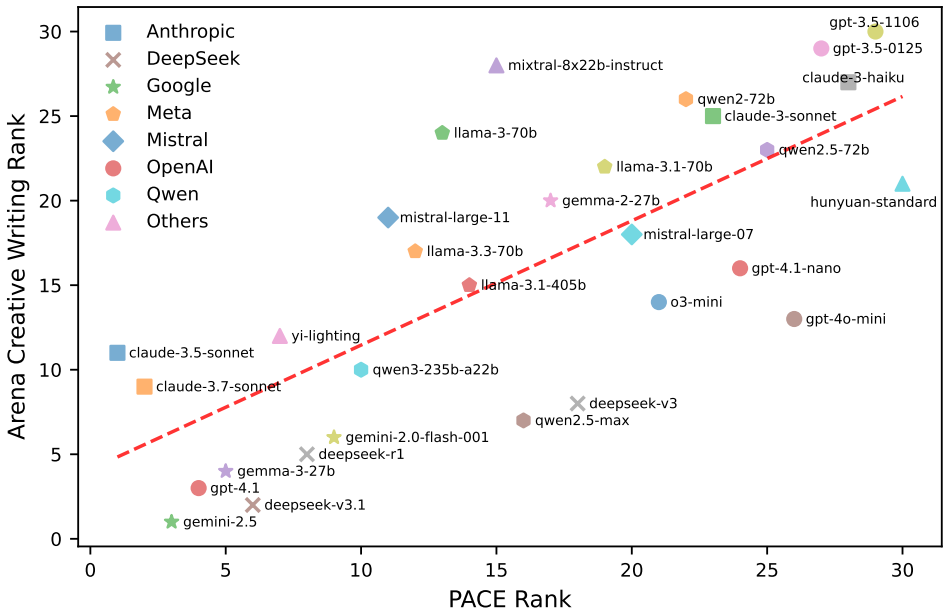

Figure 2: Comparison of model rankings according to PACE and Arena Creative Writing. Each point represents a language model, where different release versions of the same model are treated as separate variants, with the x-axis showing the PACE rank (based on association distance) and the y-axis showing the Arena Creative Writing rank. The red dashed line indicates the Spearman rank correlation fit ($\rho = 0.739$, $p < 0.001$). Claude-3.5-Sonnet achieves the highest PACE ranking among the evaluated models.

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

2 Related Work

2.1 Evaluating LLMs' Creativity

LLMs have demonstrated remarkable capabilities in diverse creative tasks, leading researchers to design increasingly complex evaluations that explore their potential for creative writing (Doshi and Hauser, 2024; Tian et al., 2024; Walsh et al., 2024), scientific hypothesis generation (Si et al., 2024; Tong et al., 2024), and co-creativity with human (Dell'Acqua et al., 2023; Ashkinaze et al., 2024; Boussioux et al., 2024).

In creativity evaluations with open-ended questions, human assessment offers reliable preference data but poses significant challenges for implementation and reproducibility. Tian et al. (2023) introduced constrained real-world questions to stimulate unconventional thinking and found that human evaluators often disagreed in their judgments, largely due to varying levels of question comprehension. To overcome these challenges, researchers have increasingly adopted LLM-as-judge approaches (Organisciak et al., 2023; Raz et al., 2024). However, concerns remain regarding their reliability and fairness (Stureborg et al., 2024; Thakur et al., 2024).

2.2 Adapting Human Creativity Assessments for LLM Evaluation

Creativity is central to human intelligence, making its assessment a fundamental topic in psychological research. Recently, numerous psychological assessments have been applied to evaluate the creativity of LLMs, including the Alternative Uses Task (Koivisto and Grassini, 2023; Hubert et al., 2024), the Remote Associates Test (Alavi Naeini et al., 2023), and the Torrance Tests of Creative Thinking (Guzik et al., 2023; Hubert et al., 2024).

Most existing studies employ LLMs to complete psychological assessments and compare their performance to that of human participants. However, since these creativity tests are widely available, they risk training data contamination. For example, GPT-3 could produce responses that directly replicated content from psychological journals and test manuals (Stevenson et al., 2022). Moreover, psychological assessments designed for human cognition lack empirical validation when applied to machine creativity, thereby undermining the correlation between model scores and real-world performance — a concern that current research has yet to adequately address.

3 Method

3.1 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.[2]

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 apply a chain-of-thought prompting strategy to guide the model's word associations[3], ensuring a structured yet flexible generation process. Prompts can be found in Appendix B.3.

3.2 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, quantity, and motion. From each chapter, five seed words are chosen based on their frequencies in the COCA corpus (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 per seed yield 6,270 associated words. The complete list of seed words is provided in Appendix A.

3.3 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 B.2. We use FastText (crawl-300d-2m; Mikolov et al., 2018) for computing cosine distance. Table 4 also reports results using alternative word embedding models.

4 Experiments and Results

4.1 Models and Parameters

Thirty models are selected from the Chatbot Arena Leaderboard, representing a balanced coverage of different performance ranks and license types (commercial and open-source). The selection spans from rank 1 (Gemini-2.5-Pro) to rank 184 (GPT-3.5-turbo-1106) out of 234 models as of May 2025. To enable robust correlation analysis with existing benchmark — which typically evaluate fewer models — at least 18 models were included in each evaluation (Bonett and Wright, 2000). Additionally, we compared Qwen models of varying versions and sizes on PACE. The complete list of models is provided in Table 3. Model responses were obtained via APIs with a temperature setting of 0, except for o3-mini (temperature fixed at 1). All other parameters were default.

[2]We choose a length of 20 words to align with the human participant data collected by Gray et al. (2019), allowing for fair comparisons between models and humans in Section 5. Furthermore, we compare different chain lengths. As shown in Table 7, PACE scores at this length (20 words) exhibit the strongest correlations with existing leaderboards.

[3]While multi-turn dialogue could also be used to elicit associations, both approaches have limitations: generating without conversational history often leads to redundant outputs, while providing the full conversational history introduces confounds such as long-context memory and coherence constraints inherent to multi-turn setups. Therefore, we adopt independent prompts to obtain interpretable and controlled measurements of creative associative capacity. Although this does not exactly mirror human reasoning processes, it helps avoid the issues associated with multi-turn association setups and also facilitates more straightforward computational evaluation.

Table 1: Spearman rank correlation between model rankings of PACE 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

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

4.2 Correlation with Existing Benchmarks

We select several representative benchmarks to validate our results, including the Chatbot Arena leaderboard (Arena Overall ranking and Arena Creative Writing ranking, hereafter Arena All and Arena CW, which rank models based on human voting preferences for 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). For each leaderboard, we calculate the Spearman correlation between the models' ranks in PACE and their ranks in the respective leaderboard.

4.3 Results

As illustrated in Table 1, **the Spearman rank correlations between PACE and various leaderboards are consistently moderate to strong**. Notably, PACE exhibits its highest correlation with Arena CW (0.739***), which is substantially higher than with Arena All (0.660***), indicating that PACE better captures creative capabilities than general performance.

In addition to its strong correlations with existing benchmarks, **PACE effectively differentiates between model variants** within the same series. As shown in Figure 2, DeepSeek-V3.1 scored 0.763 (rank 6), DeepSeek-R1 scored 0.759 (rank 8), and DeepSeek-V3 scored 0.748 (rank 19). We also compare different Qwen models to investigate the effects of model version and size on association distances. Newer model generations consistently achieve higher scores (e.g., Qwen-3 > Qwen-2.5 > Qwen-2), while within the same generation, larger models tend to perform better. These results demonstrate that PACE is sensitive to subtle differences among models (see Appendix C.2).

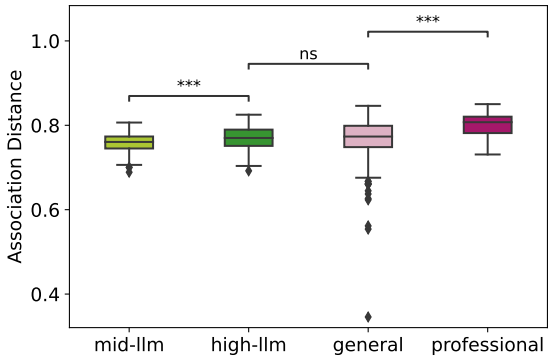

Figure 3: Comparison of association distances between humans and LLMs. Using human data from Gray et al. (2019), results show that **high-performing LLMs match average human performance**, but fall short of professional humans.

5 Comparison between Humans and Models

5.1 Associative Creativity

We compare human and LLM performance on associative creativity tests across four distinct groups. Human data from Gray et al. (2019) includes a demographically representative sample of American adults (the *general* group) and professional actors with higher creative abilities (the *professional* group). For LLMs, we evaluate *high-llm* models (top 20 on Arena leaderboard) and *mid-llm* models (ranked around position 75 of 234 total). All models are tested using identical seed word prompts as those applied in the human studies. Details of experimental settings are presented in Appendix B.4.

Current leading LLMs match average human creativity. As shown in Figure 3, high-performing models demonstrate comparable performance to *general* human groups, with no significant difference observed (Welch's t-test: $t = 0.644$, $p = 0.52$). This contrast with previous studies that reported significantly lower model performance compared to human participants (Wenger and Kenett, 2025). Furthermore, high-performing models demonstrate significantly superior performance compared to mid-performing models ($t = 3.781$, $p < 0.001$).

Best-performing human still outperforms LLMs. Both the overall group scores ($t = 6.152$, $p < 0.001$) and the maximum values ($\text{Human}_{\text{max}}$=0.8501, $\text{Model}_{\text{max}}$=0.8251) indicate that the top-performing humans still surpass the best LLMs, consistent with previous findings (Koivisto and Grassini, 2023). Moreover, a significant difference is observed between the pro-

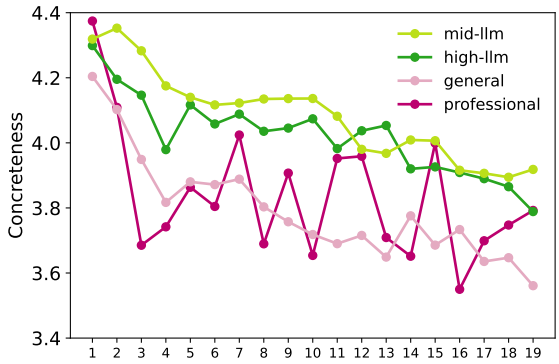

Figure 4: Average concreteness scores across chain positions for different groups. All groups show declining concreteness, with models exhibiting higher concreteness than humans.

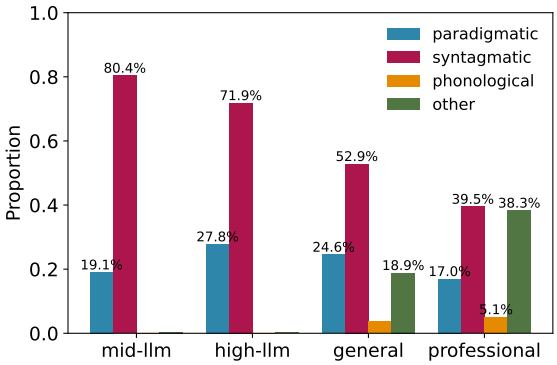

Figure 5: Types of associations within chains, categorized according to the association type framework described by Nissen and Henriksen (2006). Details are provided in Appendix D.2.

fessional group and all other groups, underscoring the unique value of human creativity (Rafner et al., 2023; Lee and Chung, 2024; Boussioux et al., 2024). In contrast, LLMs exhibit greater consistency in minimum performance ($\text{Human}_{\text{min}}$ = 0.3457; $\text{Model}_{\text{min}}$ = 0.6888), highlighting 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).

5.2 Associative Patterns

We further compare the patterns of association between humans and LLMs from two perspectives: the overall trends in associations and the types of associations observed.

Trends of associations. As shown in Figure 4, both humans and LLMs exhibit a decreasing trend in concreteness as the chain develops. However, the models consistently demonstrate higher average concreteness scores compared to humans at each step. This suggests that LLMs tend to rely more on concrete concepts rather than abstract ones, whereas humans are more inclined toward abstract cognition as they progress through the association chain. Furthermore, while both LLMs and the general human population show a relatively steady decline in concreteness, professionals exhibit greater variability, suggesting more frequent transitions between concrete and abstract associations (Kenett et al., 2014; Zhang et al., 2023). Fixed effects regression analysis confirmed significant declining trends in concreteness across all groups (all $p < 0.01$, see Appendix D.2), with LLMs showing higher baseline concreteness than humans.

Associations types. Like humans, LLMs show a stronger tendency to produce syntagmatic associations (e.g., *dog* $\rightarrow$ *bark*) than paradigmatic associations (e.g., *dog* $\rightarrow$ *cat*). However, humans demonstrate greater diversity in their associations, often generating non-semantic links such as phonological connections. Notably, professionals are more likely to produce *other* types of associations, suggesting that creative individuals often draw on personal experiences rather than relying solely on common linguistic patterns.

6 Conclusions

We propose PACE, a benchmark for evaluating the creative potential of LLMs based on parallel association chains. Compared to existing methods, PACE avoids training data contamination and offers a simple, scalable framework that greatly reduces manual evaluation costs. Experimental results demonstrate a strong and significant Spearman's rank correlation between PACE and several established leaderboards (e.g., $\rho = 0.739$ with Arena CW). Our findings show that measuring associative distance offers a highly effective way to assess LLMs' creativity, capturing their ability to move beyond surface-level co-occurrence and tap into deeper, less common semantic connections that underlie genuine creativity.

Further analysis show that while high-performing LLMs match general human scores, professionals consistently outperform models and display more diverse associative patterns. These findings underscore both advances and limitations in LLM creative association, and highlight PACE as an effective tool for benchmarking and advancing model creativity.

Limitations

Limited focus on English. Since we use English seed words and rely primarily on leaderboards with English-based rankings (such as Arena CW), 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 indirectly validate robustness, we rely on rankings from external leaderboards; however, this approach inherently constrains model selection due to the finite number of models represented across these platforms. Furthermore, to maintain comparability across disparate leaderboards, we restrict our analysis to models that consistently appear across all evaluated platforms, thereby further limiting the scope of our analytical sample. Following the guidelines established by Bonett and Wright (2000), Spearman correlations within the moderate range of $|\rho| \approx 0.5 - 0.7$ necessitate a minimum sample size of 20-30 observations to establish statistically reliable confidence intervals. While our primary analyses on Arena All and Arena CW include a sufficient number of models to ensure statistical reliability, some other leaderboard comparisons are closer to the minimum threshold, which may introduce minor limitations to the robustness and generalizability of our findings.

Acknowledgments

The authors would like to thank Dr. Wang Yin and his lab members for their helpful discussions and suggestions. This research was partially supported by the Tencent Basic Platform Technology Rhino-Bird Focused Research Program.

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*.

Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Advances in Neural Information Processing Systems*, 36:5631–5652.

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).

Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290.

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.

Erik E Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065.

Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14(1):3440.

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.

Tuval Raz, Roni Reiter-Palmon, and Yoed N Kenett. 2024. Open and closed-ended problem solving in humans and ai: the influence of question asking complexity. *Thinking Skills and Creativity*, 53:101598.

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*.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.

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*.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating psychological hypothesis generation with ai: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1–14.

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 Dataset Details

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

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 2: Chapters and their associated seed words

B Experimental Setup Details

B.1 Selected Models

Full list of selected models can be found in Table 3. PACE evaluation contains a comprehensive selection of LLMs, featuring both leading open-source models (such as DeepSeek, Gemma, LLaMA, and Qwen series) and prominent closed-source commercial models (including various versions of Claude, Gemini, and GPT series). This balanced selection represents the current state-of-the-art across commercial and open-source domains, with a total of 34 models evaluated. Among these 34 models, 30 have corresponding Chatbot Arena Leaderboard scores and rankings (based on the early May 2025 scoring version, Chiang et al., 2024), while the remaining four models (Command-R-Plus-08-2024, DeepSeek-R1-Distill-LLaMA-70b, DeepSeek-R1-Distill-Qwen-32b, and Hunyuan-Turbos-20250313) are included to ensure comprehensive coverage across different leaderboards, despite lacking Arena recordings.

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 3: Selected Models with Arena CW Scores (Cutoff: Early May 2025) and Their Association Distances

B.2 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:

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

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.

B.3 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 articulate task requirements clearly. 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 0 to observe their intrinsic associative patterns (with the exception of o3-mini, which has a fixed temperature setting of 1).

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": ""}
  ]
}
```

B.4 Settings for Comparison Between Human and LLMs

In Section 5, we compare LLM and human performance using data from Gray et al. (2019). Specifically, we use two groups from the original study: "general" (representative American adults, Group 2 in the original paper) and "professional" (professional actors, Group 4 with actor label in the original paper). The professional group achieved the highest scores in both the original association task and the traditional psychological validation tests. We use all human data without additional cleaning.

For LLM analysis, we select two parallel groups based on their Arena All 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 on the leaderboard, representing the standard performance of current models.

For seed words, we use the same set from human studies: bear, table, candle, snow, paper, and toaster. To match human sample sizes, we vary LLM temperature between 0 and 1. We generate three independent association chains per seed word at each temperature setting, calculating metrics separately for each chain rather than averaging, thereby simulating multiple participants. Each model generates 6 chains per seed word (3 chains × 2 temperatures).

C Additional Experimental Results

C.1 Robustness Analyses

Correlation with 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 4 demonstrate a consistently significant correlation between PACE and Arena CW, with MUSE achieving the highest correlation coefficient ($\rho = 0.757$). To ensure consistency with the concreteness predictions, we choose FastText as the evaluation method.

Table 4: Spearman Correlation Results Across Different Word Embedding Models

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 of correlation analysis. To validate the robustness of the correlation coefficient, we use a bootstrap method to resample the results of seed words and compute the Spearman correlation. Table 5 shows a stable and significant correlation with PACE rankings across all leaderboards (with a significance ratio of 1.000), with the exception of MMLU-Pro (which had a significance ratio of 0.962). Among these, Arena CW shows the strongest relationship, achieving the highest correlation with PACE, with Spearman correlation values ranging from 0.678 to 0.769.

Table 5: Bootstrap Results for Spearman Correlation Across Different Leaderboards

Leaderboard	Mean Corr.	SE	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

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

Impact of reduced elements on correlation. To enhance evaluation efficiency, we explore the impact of two parameters: the **number of seed words** (See Table 6) and the **chain length** (See Table 7). We use random sampling with 500 iterations to select various subsets of seed words. we also analyze the effect of different chain length by truncating the original chains and computing the correlation coefficients.

Table 6: 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)

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

Table 7: 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.

C.2 Extended Results on Models and Human Performance

Model comparisons. We compare different Qwen models to investigate how model versions and sizes influence association distances. As shown in Figure 6, association scores consistently follow a hierarchical pattern by version: Qwen-3 outperforms Qwen-2.5, which in turn outperforms Qwen-2. While smaller models from various versions (e.g., Qwen-2-7b, Qwen-2.5-3b) generally fall into lower-performing groups, larger models from older versions can still achieve performance comparable to that of more recent releases (e.g., Qwen-2-72b vs. Qwen-2.5-14b).

Performance across semantic categories. We further examine how semantic categories influence performance across different model versions. While newer models generally outperform older ones, the gap varies considerably by category. Strong performance is observed in objective categories such as **spatial relations, time, and quantity**, where even the earliest, smallest model (qwen-2-7b) achieves scores above 0.71. In contrast, subjective and abstract categories (e.g., **emotions and values**, 0.63–0.72; **kinship**, 0.65–0.78) exhibit substantially larger gaps, with newer models achieving up to 0.10 higher scores compared to prior versions.

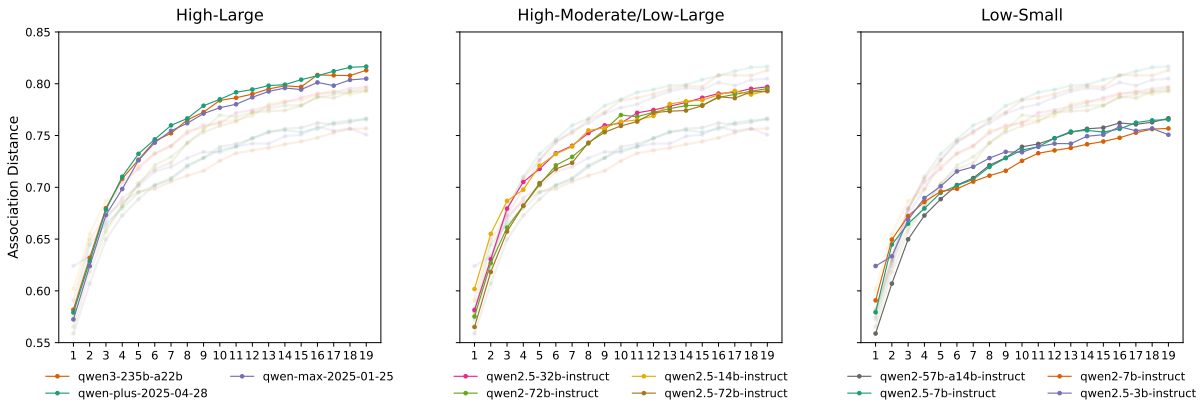

Figure 6: **Association distance comparison across versions and sizes of Qwen models.** This figure represents the association distance calculated at each position within the associative chains across different models and versions. Results reveal three performance clusters at different chain positions: (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.

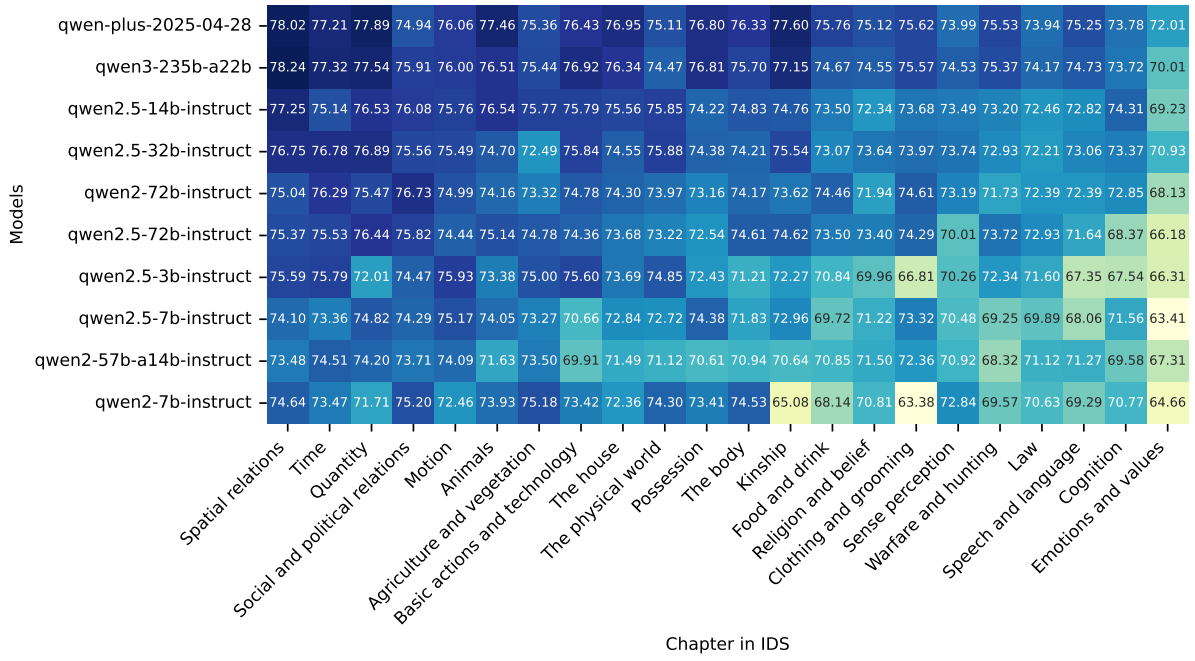

Figure 7: Association Distance Sorted by Chapters in IDS. The heatmap presents the association distance scores of different Qwen model versions across 22 semantic chapters in the IDS dataset. Each cell represents the performance score of a model (rows) on a particular semantic category (columns), with darker shades indicating higher scores. Objective categories, such as **Spatial relations**, **Time**, and **Quantity**, show consistently high performance across models, whereas subjective and abstract categories, such as **Kinship** and **Emotions and values**, display larger performance gaps, highlighting improvements in newer models.

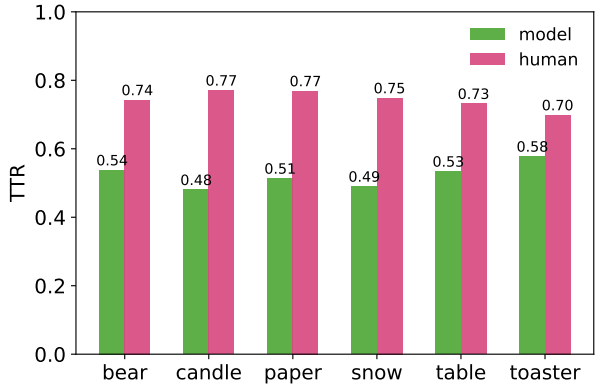

Figure 8: Type–Token Ratio (TTR) of responses generated by humans and models

Human–LLM lexical diversity. We combine responses from humans and LLMs and standardize sample sizes for each seed word to eliminate potential biases from differing data amounts. Type-Token Ratio (TTR) analysis reveal clear differences in lexical diversity: even when prompted and configured with temperatures of 0 and 1 to enhance diversity (see Sections 3.1 and B.4), LLMs consistently exhibit lower TTR values than humans across all seed words. This suggests that LLMs produce more homogeneous responses, highlighting their limitations as substitutes for human creative output (Walsh et al., 2024; Wenger and Kenett, 2025).

Examples of association chains across the score spectrum. Table 8 shows responses from humans and LLMs given the same seed word **candle**. Human responses tend to exhibit more jumping associations (e.g., lucky → irish → friend → wedding), while model-generated responses are generally more uniform and sequential (e.g., liquid → water → rain → storm).

D Additional Discussion

D.1 Association Type Classification

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 followed the association type framework described by Nissen and Henriksen (2006), which distinguishes four categories: **paradigmatic** (same word class with semantic relations like synonymy/antonymy, e.g., *love-heart*), **syntagmatic** (sequential/syntactic connections across word classes, e.g., *local–politician*), **phonological** (sound-based similarity without semantic connection, e.g., *quote-vote*), and **other** (personal associations, morphological variations, or unclassifiable connections, e.g., *desperate-rhino*). Table 11 presents examples of classified association types.

D.2 Concreteness Prediction via Embedding Models

Word embeddings have been shown to effectively predict semantic concreteness and other psychological dimensions of words (Charbonnier and Wartena, 2019; Flor, 2024; Hussain et al., 2024). We use the concreteness dataset (40,000+ words) developed by Brysbaert et al. (2014), one of the largest human-labeled concreteness databases, to train concreteness prediction models using three different word embedding approaches: FastText (English), GloVe (6B-300d), and MUSE (English). Model performance is evaluated using Pearson's correlation coefficient, root mean square error (RMSE), and Kendall's rank correlation. Table 9 indicates that FastText achieves the highest Pearson correlation and Kendall coefficient, as well as the lowest RMSE. Therefore, we use FastText to assign concreteness ratings to association responses. Table 11 shows examples of concreteness ratings assigned using this approach.

Table 9: Comparison of word embedding models for concreteness prediction

Model	Pearson r	Kendall τ	RMSE
Training Set			
FastText	**0.931 ± 0.000**	**0.760 ± 0.001**	**0.371 ± 0.001**
GloVe	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	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 10: Fixed Effects Model Results Across Groups

Group	Intercept (β_0)	Slope (β_1)	R^2
mid-llm	4.295	-0.023***	0.249
high-llm	4.197	-0.020***	0.279
general	4.022	-0.025***	0.350
professional	3.987	-0.017**	0.275

Note: *** $p < 0.001$, ** $p < 0.01$. 95% confidence intervals in brackets. Responses shorter than 20 words were excluded from analysis.

Table 8: Examples of Association Chains with Different Semantic Distances

Source	Association Chain	Score
professional	candle, flame, basketball, dad, lucky, irish, friend, wedding, ring, run, air, cold, ski, sister, proud, lion, documentary, satire, podcast, subway	0.8300
general	candle, fire, water, swim, kids, family, love, marriage, commitment, honor, life, decisions, problems, solutions, work, play, fun, joy, pain, death	0.8060
gemini-2.5-pro	candle, blackout, darkness, night, sleep, dream, fantasy, story, book, paper, tree, forest, wildlife, nature, growth, plant, seed, potential, energy, force	0.7947
llama-3.3-70b	candle, flame, heat, burn, fire, danger, warning, sign, symbol, language, communication, message, letter, paper, tree, forest, wildlife, habitat, ecosystem, balance	0.7651
mixtral-large-2411	candle, wick, flame, heat, melt, liquid, water, rain, storm, lightning, thunder, noise, silence, calm, serene, peaceful, tranquil, relax, sleep, dream	0.7112

Table 11: Examples of Concreteness and Association Types

(a) High-LLM (GPT-4.1)

Pos	**Seed**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Word	toaster	appliance	kitchen	cooking	heat	fire	wood	tree	forest	wildlife
Type	–	para	syn	syn	syn	syn	syn	para	para	syn
Conc.	–	4.31	4.86	3.96	3.75	4.54	5.00	4.92	4.76	4.17

Pos	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Word	animal	mammal	fur	coat	winter	snow	flake	crystal	glass	window
Type	para	para	syn	syn	syn	syn	para	para	syn	syn
Conc.	4.63	4.87	4.63	4.97	3.87	5.00	4.18	4.39	5.00	4.68

(b) Professional group

Pos	**Seed**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Word	toaster	bread	money	struggle	tussle	tout	flout	flounce	bounce	bunny
Type	–	syn	other	other	para	pho	pho	pho	pho	pho
Conc.	–	4.92	3.46	2.36	3.12	2.62	2.41	3.39	3.79	4.77

Pos	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Word	funny	laughter	song	dance	jig	pig	fortune	fame	frame	lame
Type	pho	para	syn	syn	para	pho	other	other	pho	pho
Conc.	2.57	3.72	3.96	4.26	4.27	5.00	3.04	2.39	4.19	2.43

Note: Pos = Position; Conc. = Concreteness; syn = syntagmatic; para = paradigmatic; pho = phonological; other = other types. Association types are labeled by DeepSeek-V3.1 and concreteness scores are predicted by FastText based on data from Brysbaert et al. (2014). The original concreteness scores range from 0 to 5; predictions exceeding this range are clipped to the nearest boundary value.