Characterizing Positional Bias in Large Language Models: A Multi-Model Evaluation of Prompt Order Effects

Patrick Schilcher[1], Dominik Karasin[1], Michael Schöpf[1], Maisam Saleh[1], Antonela Tommasel[1,3], Markus Schedl[1,2]
[1]Johannes Kepler University Linz, Austria
[2]Linz Institute of Technology, Linz, Austria
[3]ISISTAN, CONICET-UNICEN, Argentina
{antonela.tommasel, markus.schedl}@jku.at

Abstract

Large Language Models (LLMs) are widely used for a variety of tasks such as text generation, ranking, and decision-making. However, their outputs can be influenced by various forms of biases. One such bias is positional bias, where models prioritize items based on their position within a given prompt rather than their content or quality, impacting on how LLMs interpret and weigh information, potentially compromising fairness, reliability, and robustness. To assess positional bias, we prompt a range of LLMs to generate descriptions for a list of topics, systematically permuting their order and analyzing variations in the responses. Our analysis shows that ranking position affects structural features and coherence, with some LLMs also reordering or omitting topics. Nonetheless, the impact of positional bias varies across different LLMs and topics, indicating an interplay with other related biases.

1 Introduction

Large Language Models (LLMs) are widely used for text generation, ranking, and evaluation tasks. However, they exhibit distinct biases that can impact their reliability, fairness, and robustness (Shi et al., 2024). One such bias, positional bias, occurs when the placement of information within a prompt influences the model's output (Shi et al., 2024). This bias can lead to systematic prioritization of inputs based on order rather than content.

Specifically, an item's position within a list given as part of a prompt may influence how LLMs describe, emphasize, or interpret its importance, leading to inconsistent judgments. For example, in content generation tasks like summarization, question answering, and text completion, LLMs may disproportionately prioritize information appearing earlier in a prompt, affecting coherence and informativeness (Ko et al., 2020; Tian et al., 2024). In ranking and scoring applications, such as machine translation assessment or essay grading, bias toward the first or second option can result in unfair rankings (Wang et al., 2023). Similarly, in structured tasks (like chain-of-thought reasoning), positional bias may distort logical consistency by prioritizing certain steps (Liu et al., 2024).

Positional bias not only undermines the fairness of LLM-generated assessments but also raises concerns about their robustness across contexts and domains. Moreover, understanding biases in LLMs is critical given their increasing role in generating educational (Moore et al., 2023; Tan et al., 2024), informative (Muñoz-Ortiz et al., 2024; Fang et al., 2024), and decision-support content. When models systematically favor items based on position rather than content, they may subtly shape human preferences, judgments, and beliefs in ways that go unnoticed. For instance, biased ordering in generated summaries, comparisons, or lists could mislead users about what is most relevant or important (e.g., as shown for decision making tasks (Rey et al., 2020)). While some studies have revealed ranking biases in LLMs, the study of positional effects in list-based prompts has received less attention. To address this gap, we conduct a systematic evaluation of positional bias in LLM-generated lists, aiming to answer the following research questions:

- **RQ1: How do positional effects manifest in terms of structural and coherence features in LLM-generated descriptions of listed items?**
- **RQ2: To what extent do the observed positional effects stem from the LLM's biases rather than the inherent characteristics of topics?**

Our study contributes to addressing this gap by providing a systematic, cross-model analysis of how topic order affects generation outcomes across different families of LLMs. To this end, we prompt LLMs to generate descriptions for a list of topics, systematically varying their order to analyze how the generated descriptions change as a function of position, aiming at uncovering patterns of topic pri-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20643–20664
November 4-9, 2025

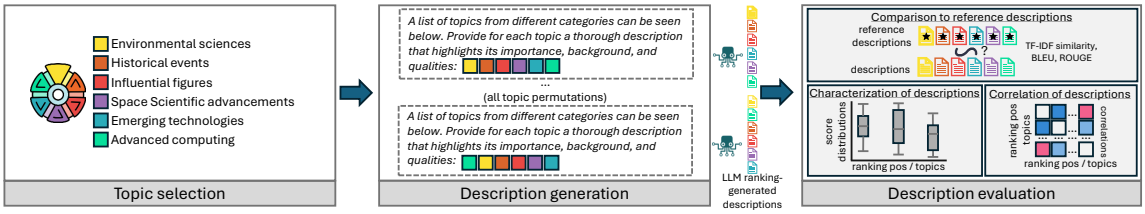

Figure 1: Schematic steps of the applied method

oritization, structure and coherence shifts linked to position, or the specific characteristics of the selected topics themselves. We not only quantify structural and content-level variations but also surface qualitative differences in how models treat position-sensitive inputs. Our findings reveal that positional bias affects not only the structure and coherence of outputs but also the relative attention and completeness of item descriptions, with variations across models and topics.

2 Related Work

While LLMs have gained popularity, research on positional bias is still limited. Prior studies have demonstrated that LLMs exhibit positional bias in decision-making tasks, often favoring options based on their placement rather than intrinsic quality (Li et al., 2024; Shi et al., 2024; Zheng et al., 2023; Wang et al., 2023; Koo et al., 2023; Panickssery et al., 2024), for example, by prioritizing early and, to a lesser extent, late information, while overlooking the middle (Liu et al., 2023; Ravaut et al., 2024).

This bias has been shown to be not random, and to vary across tasks, LLMs, and candidate similarity, with inconsistencies in how LLMs assess options (Shi et al., 2024). Also, while prompt length has been shown to have small effects on bias, bias direction has been inconsistent across tasks (Shi et al., 2024). Beyond positional bias, research has also identified verbosity bias (favoring longer responses) and self-preference bias (Zheng et al., 2023; Panickssery et al., 2024), while Koo et al. (2023) introduced a benchmark identifying six types of biases, including selection preferences based on response length, author and order.

Our study extends prior work on positional bias, which has largely focused on selection tasks and pairwise comparisons with direct/reverse orderings (Wei et al., 2024; Pezeshkpour and Hruschka, 2023). Rather than selecting items from a list, in our study, LLMs generate responses for each item in a list, enabling analysis of both selection tendencies and quality-related metrics. We systematically evaluate all possible permutations of items, offering a more comprehensive view of positional effects. Additionally, our analysis spans a broad set of LLM families, allowing us to assess whether these effects generalize across architectures.

3 Task and Methodology

Figure 1 provides a schematic overview of the applied method for evaluating positional biases in LLMs' responses, illustrating topic selection, description generation (along with the prompt used for each LLM) and description evaluation[1].

Topic selection. We selected six diverse topics, *Albert Einstein*, *World War II*, *Quantum Computing*, *Climate Change*, *Blockchain*, and the *James Webb Telescope*, covering historical, scientific, and technological domains or categories. These topics, commonly found in trivia games, represent general knowledge and public interest while varying in complexity and controversy. Selection criteria included relevance in scientific discourse, presence in widely available training data, and potential to reveal variations in response quality. This approach follows prior studies on LLM biases (Hu et al., 2024; Tao et al., 2024), which highlighted the role of topic diversity in assessing LLM behavior, as they may exhibit varying biases across subject areas, which might reflect underlying cultural values.

Description generation. To study LLMs in their natural setting, we prompted each model with the list of selected topics and asked it to generate a short description for each[2]. We did not enforce a structured output (e.g., numbered lists or bullet points), allowing the models to generate responses in their own natural style. This choice follows recent findings that enforcing specific output formats can significantly impact LLM performance (Long et al., 2024; Tam et al., 2024) biasing model behaviour, and may degrade reasoning (Tam et al., 2024). While our approach better captures the model's natural decision-making, also introduced variability in how descriptions were formatted and

[1]Additional details can be found in Appendix A.

[2]We refer to these descriptions as "list-based descriptions".

separated, creating challenges for downstream extraction. Given this variability in structure and length, we opted for a cosine similarity–based approach using sentence and topic embeddings rather than fixed criteria to reliably identify and extract individual topic descriptions[3].

We evaluated 12 LLMs from 5 different providers, including open-source and commercial, ensuring a diverse range of architectures and application domains. The evaluation included[4] *OpenAI* (`GPT-3.5 Turbo`, `GPT-4o`, `GPT-4o mini`), *Meta* (`LLaMA 3.2 1B`, `LLaMA 3.2 3B`), *Google* (`Gemma 2 2B`, `Gemma 2 27B`, `Gemini 1.5 Flash`, `Gemini 1.5 Flash-8B`, `Gemini 2.0 Flash`), *UpstageAI* (`Solar 10B`), and *Mistral AI* (`Mistral`).

Description evaluation. As part of our evaluation, we considered two temperature settings: each model's default value (ranging between 0.5 and 1) and a fixed temperature of 0.5. To account for response variability due to inherent randomness, we ran each model on all topic permutations 3 times, generating 2,160 topic descriptions per model.

Comparison to reference descriptions. We defined *reference descriptions* as those generated by each LLM when prompted with a single topic in isolation rather than in a list. These serve as baselines to assess how descriptions generated from topic lists differ due to positional bias. Evaluation is based on *percentile ranks*, *TF-IDF cosine similarity*, *BLEU* (Papineni et al., 2002) and *ROUGE* (Lin, 2004) to measure lexical and structural alignment. *TF-IDF* is preferred over embeddings (e.g., BERT) as it better captures phrasing and word choices, while embeddings focus on semantics, meaning descriptions could be highly similar even if using entirely different words or structures.

Characterization of descriptions. We examined LLM-generated descriptions' positional effects by analyzing the distribution of key structural and coherence metrics across ranking positions. These included[5] readability (*Flesch Reading Ease* (FRE) (Flesch, 1948)), information content (*Kolmogorov Complexity - compression ratio* (KC) (Ming Li, 2019)) and lexical properties (*String Length* and *Named Entity Density*). We also assessed whether LLMs omitted or reordered topics. For evaluating topic reordering we computed the Kendall's Tau corelation. We tested for significant differences across ranking positions using unpaired sample tests (with $\alpha = 0.01$), to identify whether specific positions consistently resulted in lower or higher-quality descriptions, potentially revealing positional biases in LLM outputs.

Correlation of description characteristics. We assessed description stability using Spearman correlations across ranking positions and topics based on their input positions across multiple runs, assessing the model's consistency in handling information order. High correlations suggest consistent description characteristics regardless of position. Low correlations suggest that positions influence descriptions unevenly, meaning some topics are more sensitive to positional changes than others, hinting at a topic-dependent rather than uniform positional bias.

4 Experimental Analysis

RQ1: Positional effects in descriptions.

Characterization of descriptions per ranking position. Figure 2 shows score distribution for a selection of metrics and LLMs across ranking positions and topics[6]. FRE scores varied significantly across ranking positions, particularly for `LLaMA`, `GPT`, and `Gemini 1.5 Flash-8B`, with higher scores at the top positions. Despite these variations, topics maintained a consistent relative readability order across positions, indicating that while position influenced readability, relative differences across topics remained stable, suggesting that LLMs' responses were primarily affected by position. `LLaMA` and `Gemma` showed the most significant differences across all positions, whereas `GPT` models exhibited differences mainly at the top and last positions. Topics followed a consistent ranking pattern for KC, even more pronounced for FRE. Named Entity Density was less affected by ranking position,

[3]We validated our automated splitting strategy (Appendix A.1.2) against a manual segmentation approach, finding high similarity. This supports the method's reliability and scalability. Further details in Appendix B.2.

[4]We also considered 3 additional models that were excluded during evaluation: `WizardLM-2` (answered in languages other than English), `Qwen 2.5` (inconsistent answers), and `Phi-3` (incoherent answers).

[5]A description of these metrics and others included in the analysis can be found in Appendix A.2.2.

[6]We report results obtained with the default temperature setting because they might better reflect realistic usage scenarios in which some degree of randomness is expected and desirable to avoid overly deterministic outputs. While temperature 0.5 led to slightly lower variability and fewer outliers, the overall trends, correlations, and distribution shapes were consistent across both configurations. A more detailed analysis, a comparison with the results observed for temperature 0.5 and additional charts can be found in Appendix C. Code and more charts can be found in https://github.com/hcai-mms/positional-bias-llms

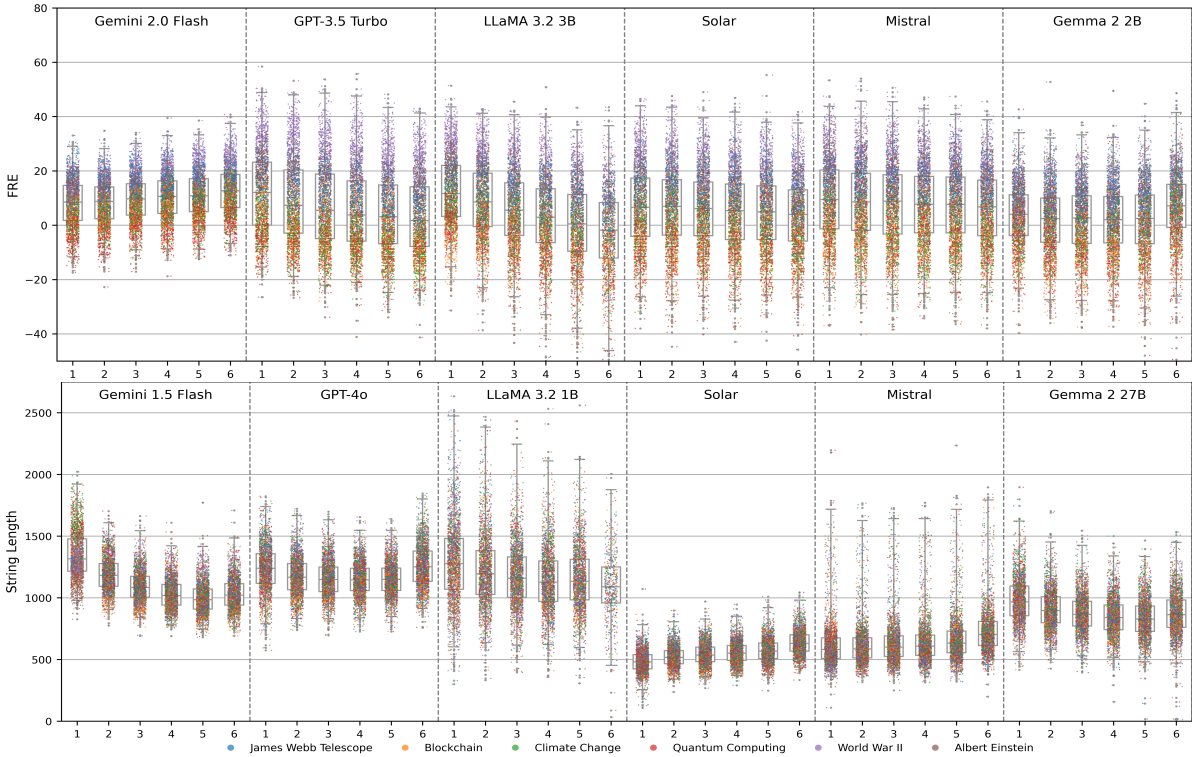

Figure 2: Score distribution per position across topics

with variations mainly at the last position, particularly for LLaMA 3.2, GPT-4o, and Gemma 2 2B. GPT-4o mini was the only model displaying significant differences across all positions. In contrast, String Length followed a U-shaped pattern, with longer responses at the first and last positions and shorter ones in between. Solar produced notably shorter responses than the other LLMs while maintaining Named Entity Density, suggesting a distinct summarization approach[7]. Unlike other metrics, for String Length, the lack of clear topic-based separation hints that ranking position, rather than topic, plays a dominant role in determining response length.

Topic ordering. Most models preserved input order, with Kendall's Tau values close to 1, except for LLaMA 3.2 1B, Gemma, and Mistral. Gemma 2 2B displayed the most reordering, often inverting ranks and clustering related topics (e.g., science). Mistral and Gemma 2 27B made occasional swaps, while LLaMA 3.2 1B showed moderate reordering, particularly in middle to last positions.

Missing topics per position. Figure 3 shows the missing topics per position[8]. LLMs primarily omit-

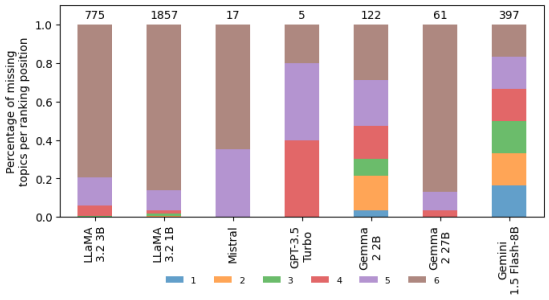

Figure 3: Percentage of missing positions per LLM.

ted topics at the last position, with Gemini 1.5 Flash-8B and Gemma 2 2B missing topics across all positions, and LLaMA 3.2 1B also missing topics at the top. Gemini 1.5 Flash-8B exhibited a uniform omission pattern (i.e., it consistently missed topics at all positions, even missing all topics altogether), while LLaMA 3.2 3B showed varied last-position omissions, with certain topics omitted more frequently than others.

Key findings. Ranking position affected readability, complexity, and response length, with earlier positions being generally more readable and more complex (higher KC), and later positions more prone to truncation or omission. While readability shifts reflected presentation style rather than changes in topic difficulty, response complexity and length showed strong position-based patterns, often resembling primacy/recency effects. The strength and form of these effects varied across

[7]More qualitative examples can be found in Appendix C.4.

[8]Although the metrics results for the three runs of each permutation were averaged, for this analysis they were not summarized, hence the maximum number of missing topics is $720 \times 3 \times 6$. The Figure includes models missing positions.

models, with some showing clearer neglect of middle positions, pointing to architecture (or training) specific biases rather than a universal behavior.

RQ2: LLM bias vs. topic influence in positional effects.

Comparison with reference descriptions. The analysis showed misalignments, with the reference descriptions scores often being statistical outliers when compared to the list-based output, and rarely aligning with median position scores. While list-based descriptions showed moderate/high TF-IDF similarity to reference ones, they primarily rephrased content rather than replicating it, as indicated by low BLEU and moderate ROUGE scores[9]. Position influenced word choice and phrasing without altering core content.

Reference descriptions were consistently longer reflecting models' preference for brevity in multi-description outputs. Topic treatment varied, with some models showing position-dependent alignment (e.g., `Mistral` aligned inconsistently across positions, while `Gemini 2.0 Flash` showed no alignment). Most models favored first and last positions over middle ones. The stronger alignment of reference descriptions with certain positions suggested that LLMs adjusted linguistic features differently for single versus list-based outputs, highlighting the complexity and variability of these biases.

Characterization of topic descriptions across ranking positions. Both ranking position and topic influenced scoring, though effects varied across LLMs. While general ranking scoring effects appeared in some models (e.g., `GPT-3.5 Turbo` assigned lower FRE scores to later positions, while `LLaMA 3.2` did so for top positions), other models displayed topic-specific trends, such as `Gemini 1.5 Flash` assigning lower KC scores to top positions for Climate Change and Quantum Computing. Correlations were inconsistent, with some models maintaining stability while others fluctuated by topic. FRE score dispersions were generally similar, except for `LLaMA 3.2`, showing greater variation for Quantum Computing, Albert Einstein, and World War II, indicating uneven sensitivity to ranking position.

Named Entity Density preserved relative topic relationships but showed little ranking scoring influence, aside from some position-specific trends in `Gemini` (e.g., `Gemini 1.5 Flash-8B` assigned lower scores to top positions and higher to middle ones). String Length exhibited the strongest ranking scoring effects, with `Solar` showing near-perfect correlations and `Gemini` displaying strong but slightly varied patterns. These results reinforce how positions and topics shape LLM behavior, with biases differing across models and metrics.

Key findings. Positional bias was confirmed, but its effects were not uniform and interacted with topic-related priors. Certain topics showed stronger variation across positions, while others remained stable, indicating that training data or preconceptions shape how LLMs encode content. Some topics were also more likely to be simplified or omitted at later positions, suggesting implicit prioritization. While core content was usually preserved, as TF-IDF similarity suggested, surface-level phrasing and ordering varied, showing that positional bias often affects how information is framed rather than its semantics. Moreover, no topic behaved consistently across models, highlighting that both architecture and data influenced positional sensitivity.

5 Conclusions

This study shows that positional bias in LLM-generated responses is real and multifaceted. Across metrics and models, we found that ranking position systematically influenced structural and lexical features of outputs, including readability, complexity, and length. While some models also reordered or omitted topics, others produced consistent outputs, highlighting that positional effects are not universal but instead vary across architectures and families.

Moreover, our study shows that positional bias does not operate in isolation. It interacts with model-specific training patterns and topic-specific priors, meaning that variation arises not only from list position but also from how models encode and prioritize different kinds of content. Some models exhibited clear position-based scoring trends, while others showed topic-dependent shifts in framing or emphasis. This variability suggests that mitigation strategies may need to be tailored to models rather than assuming a single correction method.

Overall, understanding and mitigating positional bias is critical for fairness, transparency, and reliability in LLM applications. Left unaddressed, these biases could distort information presentation, subtly shaping user perceptions and decisions. By surfacing both systematic patterns and model- or topic-specific nuances, our work contributes empirical evidence to guide the evaluation and mitigation of positional bias in large language models.

[9] Charts can be found in the Appendix C.2.

Limitations

We acknowledge several limitations that should be addressed in future works. First, we focused on lists of length six, which, while reasonable, may not capture how positional biases manifest in shorter or longer lists. Second, our topic selection was limited primarily to technical and natural-scientific subjects, potentially limiting the generalizability of our findings. Future work could explore a broader range of topics, including mainstream and obscure subjects and assessing how similar topics (e.g., Blockchain vs. distributed ledger technology) are treated in different positions in the same list. Additionally, mixing real and fake topics could provide further insights into LLM behavior. Third, we did not test prompt variations, such as explicitly instructing the model to follow the input order, prioritize or focus on specific positions, or include unrelated instructions to examine potential confusion effects (e.g., ask the LLM about the relationship between the topics in the second and last position). Fourth, we relied solely on automated metrics. Nonetheless, human evaluations could offer deeper insights into the perceived quality and fairness of responses. Fifth, we did not account for temporal or iterative effects, though real-world interactions of users and LLMs often involve users refining their inputs or reordering lists, potentially influencing positional biases over multiple turns. Exploring these aspects could offer a more comprehensive understanding of positional bias in LLMs.

Finally, while our study faced some threats to validity, we explicitly addressed them through additional analyses. First, although the temperature of LLMs was not fully standardized across all models, potentially introducing variation in response randomness, we mitigated this limitation by conducting an evaluation at two different temperature settings, and providing an analysis of the variability of responses across multiple runs. Our analysis showed that trends and relative differences remained consistent across configurations, suggesting that our findings are robust to temperature-induced variability. Nonetheless, some models still exhibited greater variability, which should be considered when interpreting individual results. Second, we allowed LLMs to structure responses freely, which may have introduced variability in formatting and made it harder to isolate positional bias. To address this, we conducted an additional evaluation using a more manual segmentation approach based on lexical splitting. We then compared these partitioned descriptions to the similarity-based partitions using Jaccard and Ratcliff-Obershelp similarity. Overall, the results supported the reliability of the similarity-based strategy, which provided a more consistent and automated method of response extraction.

Ethical considerations. Understanding and mitigating biases in LLMs is crucial for fairness, reliability, and trust. If LLMs systematically favor certain inputs, they may reinforce unintended biases in decision-making tasks, such as content moderation, automated grading, or hiring. Moreover, biases in response generation could influence public perception, particularly in applications like news summarization or recommendation systems, where misrepresenting information could distort narratives. Addressing these biases is essential to prevent misleading or unfair outcomes and to ensure that LLMs support ethical and transparent applications.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): https://doi.org/10.55776/COE12.

References

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations. *Preprint*, arXiv:2407.01085.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in llm-based evaluators. *Preprint*, arXiv:2310.01432.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A dynamic llm-powered agent network for task-oriented agent collaboration. *Preprint*, arXiv:2310.02170.

Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2024. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *Preprint*, arXiv:2408.08656.

Paul Vitányi Ming Li. 2019. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Cham.

Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. 2023. Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. *Preprint*, arXiv:2310.10570.

Arnaud Rey, Kévin Le Goff, Marlène Abadie, and Pierre Courrieu. 2020. The primacy order effect in complex decision making. *Psychological Research*, 84(6):1739–1748.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *Preprint*, arXiv:2406.07791.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *Preprint*, arXiv:2408.02442.

Kehui Tan, Jiayang Yao, tianqi pang, Chenyou Fan, and Yu Song. 2024. Elf: Educational llm framework of improving and evaluating ai generated content for classroom teaching. *ACM Journal of Data and Information Quality*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Runchu Tian, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang, and Xiaojiang Liu. 2024. Distance between relevant information pieces causes bias in long-context llms. *Preprint*, arXiv:2410.14641.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. *arXiv preprint arXiv:2406.03009*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Task and methodology

A.1 LLM description generation

A.1.1 Model Selection

We originally selected 15 LLMs of 7 different providers, ensuring a wide variety of employed architectures and application domains.

LLaMA 3.2[10]. We selected LLaMA 3.2 1B and LLaMA 3.2 3B of the LLaMA model family with one and three billion parameters, respectively. This models are adequate for application domains with limited computational resources.

Qwen25 7B and Qwen25 32B[11]. The 7B variant represents a mid-sized model and the 32B variant offers significantly greater capacity. The larger model demonstrates enhanced capability for handling complex language tasks and provides more precise and detailed outputs. These models were discarded as they tended to provide inconsistent answers.

WizardLM-2 7B[12]. This is a model specifically optimized for reasoning and logical inference tasks. With 7 billion parameters, it is particularly effective in multi-turn dialogue scenarios and applications requiring detailed, step-by-step explanations. This model was discarded due to its tendency to answer in languages other than English, skewing results.

Gemma 2 2B and Gemma 2 27B[13]. The Gemma 2 series includes models with 2 and 27 billion parameters, offering scalability to meet different performance requirements. The smaller model is lightweight and resource-efficient, while the larger variant excels in processing complex and nuanced language tasks.

Solar[14]. Solar is a versatile general-purpose language model with 10 billion parameters, balancing efficiency and complexity. It is designed to handle diverse text-processing tasks with a strong focus on accuracy and contextual relevance.

Gemini 1.5 Flash, Gemini 1.5 Flash-8B and Gemini 2.0 Flash[15]. These models are designed for high-speed processing, as indicated by the "flash" designation. The 8B version provides additional capacity, enabling it to handle larger and more demanding tasks while maintaining the model's optimized speed. Gemini 2.0 Flash incorporates advanced architectural features that aim to improve both efficiency and accuracy. It represents ongoing innovation in language model development.

GPT-3.5 Turbo, GPT-4o and GPT-4o mini[16]. GPT-3.5 Turbo is a widely used model, valued for its balance between cost efficiency and performance. It is known for its versatility, making it suitable for tasks such as text generation, summarization, and conversational AI. GPT-4o and GPT-4o mini are optimized iterations of the GPT-4 architecture, with enhancements for performance and resource efficiency. The "mini" variant is a compact version.

Mistral[17]. Mistral is a model with 7.3 Billion parameters engineered for superior performance and efficiency. It offers great performance for its size.

A.1.2 LLM output splitting

We chose not to enforce a structured output format in our study to allow LLMs to generate responses in their natural style, minimizing potential biases introduced by format constraints. Recent research has shown that requiring LLMs to follow specific output formats can significantly impact their performance. For instance, format restrictions can introduce biases due to the models' varying familiarity with different structures, as some formats (e.g., Python lists) are more commonly encountered during training than others (Long et al., 2024). Furthermore, strict adherence to structured formats has been found to degrade LLMs' reasoning abilities, as it constrains their generation space and can interfere with their ability to produce high-quality, contextually appropriate responses (Tam et al., 2024). By allowing free-form responses, we aimed to capture LLMs' natural decision-making process without introducing artificial constraints that could skew our analysis of positional bias. However, this choice also introduced challenges in ensuring comparability across responses, which we acknowledge as a limitation.

There was no consensus on how LLMs structured their responses. For example, some provided numbered lists, while others introduced each

[10] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[11] https://qwenlm.github.io/blog/468qwen2.5/.469

[12] https://huggingface.co/WizardLM

[13] https://ai.google.dev/gemma

[14] https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0

[15] https://deepmind.google/technologies/gemini/flash/

[16] https://openai.com/news/

[17] https://mistral.ai/en

description with the topic name or used a more free-flowing format. Given this variability in structure and length, we opted for a cosine similarity-based approach using sentence and topic embeddings rather than fixed criteria. We used Sentence-BERT[18] to generate the individual sentence (E_S) and topic (E_T) embeddings. Similarity was computed as shown in Eq. 1.

$$\cos(\theta) = \frac{E_S \cdot E_T}{\|E_S\| \cdot \|E_T\|} \quad (1)$$

To refine similarity scores, we applied an unweighted moving average, incorporating one neighboring value on each side (i.e., the similarity with the previous and following sentences). Each sentence was then assigned to the topic with the highest average similarity, and the longest consecutive sequence of sentences for each topic was selected as its corresponding description.

A.2 Description evaluation

A.2.1 Comparison to reference descriptions

To compare the reference descriptions with the list-based ones, we used percentile rank to determine where the reference description fell within the distribution of list-based descriptions for each metric. A high or low percentile rank indicates whether the reference description aligns more with extreme or median-ranked outputs, revealing potential biases in how LLMs prioritize information. We also considered statistical analysis such as t/Wilcoxon tests to assess whether differences between reference and list-based descriptions were statistically significant.

For the similarity analysis, we chose *TF-IDF* over semantic embedding (e.g., *BERT*) representations because *TF-IDF* directly captures lexical similarity, allowing us to analyze structural similarities between reference and list-based descriptions. Since *TF-IDF* relies on exact word matches weighted by importance, it reflects how closely the wording is preserved. This is crucial for our study, as we are interested in whether list-based descriptions maintain the same structural and lexical choices as the reference ones. In contrast, *BERT*-based embeddings primarily capture semantic similarity, meaning two descriptions could have high similarity even if they use entirely different words or structures. While high lexical overlap (as measured by *TF-IDF*) often implies high semantic similarity, the reverse is not necessarily true, two semantically similar descriptions can have different structures, making it harder to assess whether LLMs maintain original phrasing or reorder information. Thus, *TF-IDF* is better suited for analyzing how positional effects influence structural consistency in LLM-generated descriptions.

A.2.2 Characterization of descriptions

To investigate positional bias in a multifaceted way, we selected a diverse set of metrics. These metrics cover various aspects, such as readability, information content and lexical metrics. In all cases, text processing was implemented using NLTK[19].

Flesh Reading Ease (FRE). FRE (Flesch, 1948) aims to measure the readability of text by considering the total number of words, sentences, and syllables. Higher FRE scores indicate easier readability, while lower scores indicate a more complex text. Given that this score is defined based on constants, it can be negative when sentence length and word complexity are extremely high. Since the formula subtracts both factors from 206.835, long sentences and multisyllabic words can drive the score below zero. This typically occurs in dense academic, legal, or technical texts with highly complex language.

Kolmogorov Complexity (KC). Kolmogorov Complexity (Ming Li, 2019) estimates the complexity of a string by determining the length of its shortest possible description, often approximated using data compression. An approximation of KC can be calculated using zlib[20] by obtaining the length of the compressed data in bytes. Input strings with higher complexity result in a longer compressed data length, while strings with lower complexity or redundant structure result in a shorter estimation. We considered KC in the form of its compression ratio, i.e., the compressed length divided by the original string length.

Named Entity Density (NED). It tries to explain the density of all named entities (proper nouns in singular or plural form). A higher proportion of proper nouns may indicate that the text contains more specific information. This metric can be calculated as follows:

$$\text{NED} = \frac{\text{total proper nouns}}{\text{total words}} \quad (2)$$

[18] We used the model 'all-MiniLM-L6-v2' https://sbert.net/

[19] https://www.nltk.org/

[20] https://www.zlib.net/

String length. It is measured as the total number of characters, including punctuation and special symbols. It serves as an approximation for the amount of context in a given LLM response, since longer texts usually contain more detailed explanations.

Missing Topics. This metric evaluates if all topics, as defined in the input prompt, are present in the LLM response. It is computed as the number of missing topics per position.

A.2.3 Correlation of descriptions

We assessed description stability using Spearman correlations between topics based on their input positions across multiple runs, assessing the model's consistency in handling information order. High correlations might indicate stable description characteristics across positions, suggesting minimal positional influence. In contrast, low correlations imply that positional effects vary by topic, meaning some topics are more sensitive to ordering changes than others. This variability suggests an interaction between positional and topic-specific biases rather than a uniform positional bias.

B Methodological robustness validations

B.1 Analysis of variability of results due to LLM temperature Settings

To assess the consistency of LLM-generated descriptions across multiple runs, we computed the coefficient of variation (CV) for the 3 runs of each prompt, for each position and metric. The CV provides a standardized measure of dispersion by normalizing the standard deviation relative to the mean, with higher values indicating greater variability and lower values indicating more consistency. This approach allows to compare the variability of metrics across different ranking positions and prompts while accounting for differences in scale, enabling a fair comparison of the consistency of model responses regardless of the specific score (unlike the standard deviation).

B.1.1 Default LLM temperature setting

`Gemini 1.5 Flash` demonstrated the most consistent results, exhibiting low variation across KC, Named Entity Density, and String Length. `Gemini 2.0 Flash` also showed strong stability, with slight fluctuations but generally stable performance. `GPT-4o mini` and `GPT-4o` maintained moderate stability across the metrics, with some fluctuations, especially in complexity measures, but still offering relatively consistent outputs. `Gemma 2 27b` provided stable results across positions, though with slightly more variation than the `Gemini` models. `GPT 3.5-Turbo` exhibited moderate stability, with noticeable variability in certain metrics, particularly in KC and String Length. `LLaMa 3.2` was the least stable, showing significant variation, especially in complexity and length measures. `Mistral` and `Solar` showed mixed stability, with `Solar` exhibiting more inconsistency in Named Entity Density, but both still maintained reasonable performance in the other metrics. Based on the observed trends, the stability of the metrics can generally be trusted, particularly for the `Gemini` models, which exhibited consistent results across various metrics. However, variability in models like `LLaMa 3.2` and `Mistral` suggested that careful consideration is needed when making conclusions for these less stable models.

B.1.2 LLM temperature set to 0.5

At a temperature setting of 0.5, overall variability across models was slightly reduced compared to the default setting. While minimum CVs were generally higher, maximum CVs were consistently lower, indicating a more compact distribution of variation. On average, for the 0.5 temperature, CVs were lower, reflecting increased overall stability. For up to 25% of the permutations, CVs across positions remained largely similar across both temperature settings, with differences under one absolute unit for up to 75% of the samples in several cases, suggesting that most variation was concentrated in a small portion of the samples. Mean scores were also comparable between settings, with slightly lower values (by 1–2 units) for the 0.5 temperature, indicating slightly more stable outputs while still capturing natural variability. Notably, the most consistent behavior was again observed for the `Gemini` family and `GPT-4o`. In contrast, `Mistral` and `Solar` showed more pronounced differences across several metrics, though `Solar` in particular became significantly more stable at the 0.5 temperature, especially in Named Entity Density and complexity-based metrics. These trends were consistent across positions.

B.2 Validation of the response extraction method

To verify the reliability of the similarity-based partitioning method used, we conducted an additional

evaluation using a more manual approach. Specifically, we segmented the LLM-generated responses based on newline characters (\n) and markdown-style markers (e.g., *, -, **), which indicated the structural boundaries in unformatted text for the selected LLMs. We then compared these manually derived splits to the similarity-based partitions by computing Jaccard similarity, which measures set overlap, and Ratcliff-Obershelp similarity, which detects shared subsequences. In some cases, responses exhibited structures that warranted further adjustments to improve segmentation.

Across all LLM and topic combinations, over 90% of prompts yielded similarities above 0.87. For Solar and Mistral, more than 95% of prompts had similarities exceeding 0.99. In all cases, the CV remained below 10%, and the mean similarity across prompts for each topic was consistently above 0.94, with standard deviations below 0.08. A few outliers with lower similarities (minimum similarities ranged from 0.14 to 0.72) were manually reviewed. These cases resulted from incorrect lexical splitting caused by inconsistent separations between descriptions or non-relevant phrases inserted by some LLMs (e.g., "Each individual's story offers valuable insights into their life, work, and contributions to society."). After adjusting these descriptions, similarities exceeded 0.98. Minor discrepancies remained due to variations in extra spaces and markdown characters. Overall, these results demonstrate the reliability of the similarity-based split strategy, which allowed for a more robust and automated extraction than the lexical-based extraction.

C Experimental analysis

As previously mentioned, we considered two temperature settings for the LLMs: the default configuration and 0.5. We chose to report results obtained with the default temperature setting because they might better reflect realistic usage scenarios in which some degree of randomness is expected and desirable to avoid overly deterministic outputs. While temperature 0.5 led to slightly lower variability and fewer outliers, the overall trends, correlations, and distribution shapes were consistent across both configurations. Therefore, the default setting offers a valid and representative baseline without compromising reliability. Nonetheless, we also provide a summary of the results for the 0.5 setting to support transparency and facilitate comparison.

C.1 RQ1. Positional effects in descriptions

Figure 4[21] shows the score distributions per each ranking position across all topics for the evaluated metrics.

Flesch Reading Ease (FRE). Several LLMs exhibited statistically significant differences in readability depending on topic's position. Specifically, the Dunn test identified significant differences for the LLaMA and GPT families and Gemini 1.5 Flash-8B. The most common pattern was higher readability scores for topics in the first two positions, compared to later positions. LLaMA 3.2 3B and Gemma 2 2B showed the largest number of significant differences, including variations in middle to last positions, while GPT-4o mini exhibited differences only for the top position.

One interesting finding is the relative consistency of topic rankings across models. Regardless of position, certain topics consistently received the highest readability scores, followed by a middle tier of topics, while others were typically assigned the lowest scores. This suggests that while ranking position affects readability scores, the relative difficulty of topics is largely preserved across models, which hints that the observed effects are likely tied to how each LLM processes and structures responses rather than fundamental changes in topic complexity.

Across model families, LLaMA models showed the largest number of statistically significant differences, particularly with the last positions. Gemini models varied, with some (e.g., Gemini 1.5 Flash-8B) showing differences between the top and last positions, while others (e.g., Gemini 1.5 Flash) had none. GPT models exhibited differences for the top position with respect of the others, with additional variations in lower-ranked positions for smaller variants. Gemma models followed a similar pattern, with Gemma 2 2B displaying differences for more positions than Gemma 2 27B.

Kolmogorov Complexity - Compression ratio (KC). This analysis revealed statistically significant differences for almost all evaluated LLMs, with most differences occurring between the top position and the others. Additionally, Gemini 1.5 Flash-8B, GPT-4o mini, Solar, GPT-4o,

[21] In all cases, full size charts can be found in the companion repository.

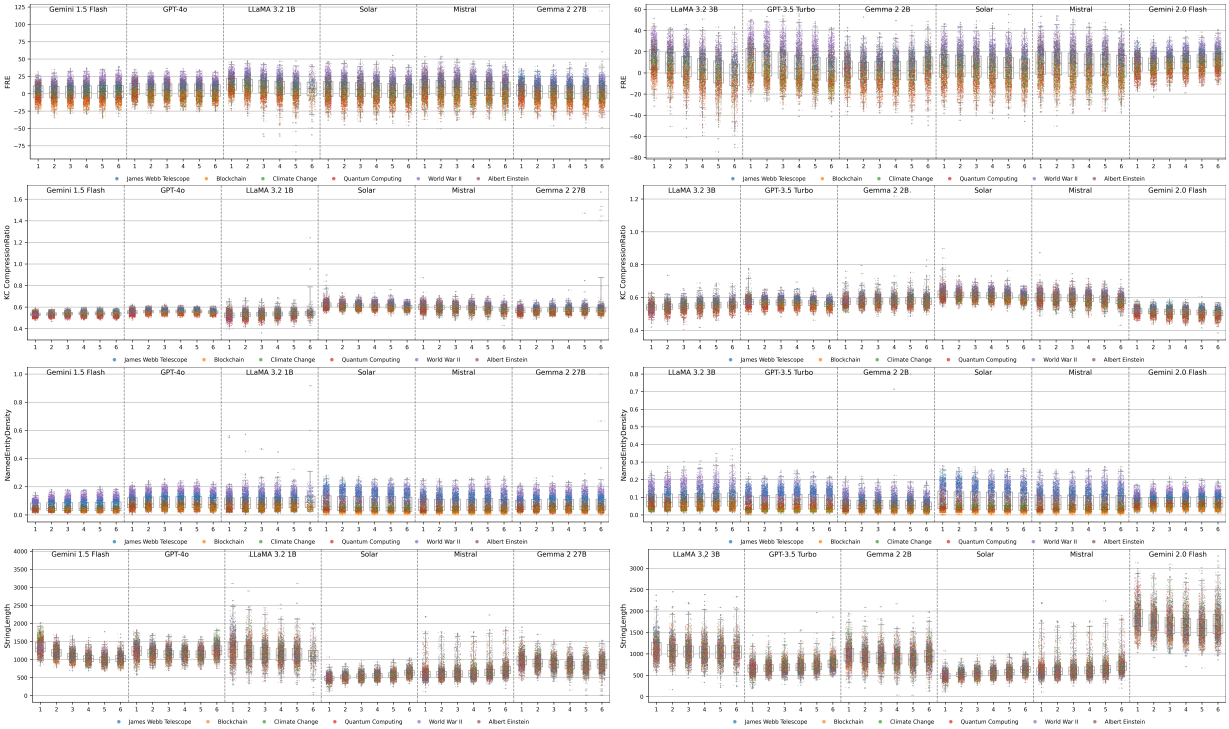

Figure 4: Score distribution per ranking position across topics

GPT-3.5 Turbo, and Mistral showed significant differences between all positions and the last, suggesting that responses at the last position tended to have notably different complexity patterns compared to earlier positions.

While topics again appeared to be sorted in terms of complexity, differences were less pronounced compared to FRE. This suggests that while certain topics consistently yielded more or less complex outputs, the contrast is subtler. Overall, these results suggest that complexity patterns differ across LLMs, with some showing broader distributions and others generating more compressed outputs. The findings also highlight that the last position frequently deviated from the rest, potentially indicating that responses generated for lower-ranked topics tend to be structurally simpler or less varied.

LLaMA and Gemini models followed similar trends. GPT models showed some variations, particularly in GPT-3.5 Turbo, which exhibited significant differences also between the middle and second-to-last positions. Also GPT-4o mini showed significant differences between the two top positions. Gemma models differed in scope, with Gemma 2 27B exhibiting a larger number of significant differences than Gemma 2 2B across all positions.

Named Entity Density (NED). For this metric, there were fewer significant differences across ranking positions compared to the previous metrics, suggesting that the presence of named entities in responses remained stable across ranking position, with fewer models displaying significant differences compared to complexity-based metrics. The most common pattern of variation was observed between the last position and the others.

While some variations were still observed at the lower-ranked positions (particularly last position), the general consistency suggests that named entity usage is less affected by ranking position than other text properties such as readability or complexity.

String length. Clear scoring patterns across ranking positions were observed for this metric, with Solar standing out as the model generating significantly shorter responses compared to all others, suggesting that this LLM may apply a more aggressive summarization strategy. Nonetheless, despite these shorter responses, Solar's Named Entity Density remained in line with other models, indicating that it maintained entity inclusion even with fewer characters.

A notable visual trend in most LLMs was the formation of a U-shape across ranking positions. Responses tended to be longer for the first position, decrease for intermediate positions, and then in-

crease again for the last position. In GPT-4o mini and Solar, however, the trend appeared different, as the length mainly increased at the last position without an initial drop.

Most models showed significant differences between every position and the last one, particularly in the Gemini, GPT, Solar, and Mistral families. GPT-4o and GPT-3.5 Turbo shared similar patterns, while GPT-4o mini exhibited fewer differences for the top position. Gemma and Gemini models exhibited consistent differences within their families, and LLaMA 3.2 1B showed fewer overall significant differences.

In summary, String Length hinted at the existence of bias, as LLMs tended to generate shorter responses in intermediate positions and longer responses for the top and last ones. The exceptionally short outputs from Solar stood out as an outlier, suggesting potential qualitative differences in response generation. The lack of clear topic-based separation in this metric further emphasizes that ranking position, rather than topic, plays a dominant role in determining response length.

Topic ordering. To assess whether LLMs preserve the input ordering of topics in their generated descriptions, we computed Kendall's tau correlation between the input and output rankings, considering only runs where all topics were present. Overall, most models maintained the original order, with only four models exhibiting noticeable deviations: LLaMA 3.2 1B, Gemma 2 2B, Gemma 2 27B, and Mistral. Among these, Gemma 2 2B showed the most significant changes, often reordering topics in an inverse manner. A qualitative examination revealed that this behavior was caused by the model's tendency to group topics belonging to similar domains or categories, such as science and technology, regardless of their original positions.

For Mistral, and Gemma 2 27B, the lower correlations appeared to be isolated outliers, rather than a systematic effect. This suggests that, in general, LLMs tended to respect the input order of topics. Examining the specific positions that were exchanged, Gemma 2 2B displayed the highest number of swaps, particularly in middle to last positions, whereas LLaMA 3.2 1B exhibited a moderate level of reordering. Interestingly, some models also swapped the first position. These findings indicate that while most LLMs adhere to the given order, a few models exhibited systematic biases in how they organize descriptions, with Gemma 2 2B standing out as the most prone to reordering topics.

Missing topics per position. In most cases, missing topics most often occurred at the last position, suggesting that LLMs tended to truncate topics rather than skip topics from the middle or top positions. Only two models (Gemini 1.5 Flash-8B and Gemma 2 2B) missed topics at all positions, with Gemini 1.5 Flash-8B displaying a relatively uniform distribution of missing topics across positions. LLaMA 3.2 1B also occasionally missed a topic in the first position, while other models only began missing topics from middle positions onward.

Gemini 1.5 Flash-8B and Gemma 2 2B exhibited relatively low standard deviations in their omission patterns, meaning they missed topics more evenly across different positions. Other models, however, followed a tailed distribution, with a pronounced tendency to drop topics at the end. Among the models that omitted topics most frequently, LLaMA 3.2 3B and LLaMA 3.2 1B stood out. While LLaMA 3.2 1B showed a relatively consistent omission frequency in the last position (ranging between 252 and 289 times per topic), LLaMA 3.2 3B displayed a wider dispersion (from 60 to 150 times per topic), suggesting that certain topics were disproportionately omitted.

C.1.1 Comparison with results observed for LLM temperature of 0.5

We further evaluated the impact of temperature on the consistency of our findings by comparing the distribution of results across models, metrics, and positions using the Wilcoxon signed-rank test and Spearman correlation. This analysis aimed to verify whether setting the temperature to 0.5 altered the tendencies observed with the default setting.

Across most models, we found moderate to strong Spearman correlations (often between 0.5 and 0.9) indicating that the relative ranking of values was largely preserved. For several LLM–metric combinations, particularly involving Named Entity Density and String Length, the Wilcoxon test revealed no statistically significant differences, suggesting that the central tendencies of the distributions remained stable. While a few metrics (e.g., KC and String Length for some models) showed weaker correlations, these were often due to the distributions being tightly clustered around zero differences, making correlations less informative despite overall consistency. Taken together, these

results suggest that the tendencies observed in the default temperature configuration are largely preserved when lowering the temperature, supporting the robustness and validity of our findings across different sampling conditions.

When comparing the boxplots of metric scores across positions[22], we observed that the overall shapes of the distributions were largely preserved between the default and 0.5 temperature settings, indicating that position-based tendencies remained stable across configurations. Similarly, in those cases in which a topic order was observed in the boxplots (e.g., the scores of a certain topic were concentrated on the top of the boxplot, whereas the scores of another topic were concentrated on the bottom of the boxplot) for the default temperature setting, such relative ordering of topics was also observed for the 0.5 temperature setting. A notable difference, however, was the presence of outliers: at the default temperature, outliers were more frequent in some metrics, such as Named Entity Density for `LLaMa 3.2`, whereas for the 0.5 temperature, outliers were considerably reduced. When analyzing metric behavior per topic across positions, only one deviation from general trends was observed, again with `LLaMa 3.2` and Named Entity Density, where the pattern across topics diverged from those at the default temperature.

Regarding missing topics, the same general trend was found, with position 6 leading the occurrence of missing topics for most models. `Gemma 2 27B` and `Gemini 1.5 Flash` continued to exhibit missing outputs in a broader set of positions, though the overall number of missing topics was reduced for the 0.5 temperature.

C.2 RQ2. LLM bias vs Topic Influence in Positional Effects

Comparison with reference descriptions. Figure 5[23] shows the score distributions of the comparisons between the reference descriptions and the list-based descriptions in terms of *BLUE*, ROUGE and *TD-IDF* cosine similarity. This comparison revealed significant differences, suggesting that the ranking process of LLMs does not fully align with the characteristics of the reference descriptions (i.e., individually generated descriptions). Across different models and topics, the scores observed for the reference descriptions were often statistical outliers, as indicated by t/Wilcoxon tests. The percentile ranks further showed that the reference descriptions rarely fell within the median range of ranked scores.

As regards lexical similarity metrics, BLEU scores were generally low across all models, indicating minimal exact word or phrase overlap between ranked and reference descriptions. This suggests that LLMs rephrased content rather than replicating specific sequences. The penalty for text length may have further lowered scores, as reference descriptions tended to be longer. No consistent relative relations among topics emerged across models. However, some ranking scoring effects were observed. For example, `Gemini 1.5 Flash` and `Gemma 2 27b` showed a tendency for the first position to achieve higher scores, particularly for Blockchain, Climate Change, Quantum Computing, and James Webb Telescope. Dispersion patterns also varied, reinforcing the inconsistency in how LLMs handle topics across rankings.

ROUGE scores were higher than BLEU. This suggests that while list-based descriptions retained some words and phrases from the reference descriptions, their overall structure differed. As with BLEU, relative relations among topics varied across LLMs, indicating differences in topic treatment. Some ranking scoring effects were observed for `Gemini 1.5 Flash`, for which the highest scores were observed for the top ranking position for James Webb Telescope, Blockchain, Climate Change and Quantum Computing (at a lesser extent). On the other hand, `Gemini 1.5 Flash-8B` showed a more consistent ranking effect across all topics, while `GPT-4o mini` tended to exhibit lower scores for World War II and Albert Einstein, while no clear patterns emerged for the other topics.

Finally, TF-IDF similarities were moderate, implying that while exact phrasing varied, core concepts remained consistent, suggesting that list-based descriptions retained similar lexical elements to the reference descriptions but with different structure or emphasis/frequency. Across LLMs, Quantum Computing consistently showed the highest similarities, with stable relative differences across topics. Dispersion was generally uniform, except for `Gemini 1.5 Flash-8B`, which exhibited some outliers. Unlike other metrics, clear ranking scoring effects were not observed, except for `Gemini 1.5 Flash` in Climate Change, where the top position showed the highest scores.

[22] The full set of boxplots can be found in the companion repository.

[23] In all cases, full size charts can be found in the companion repository.

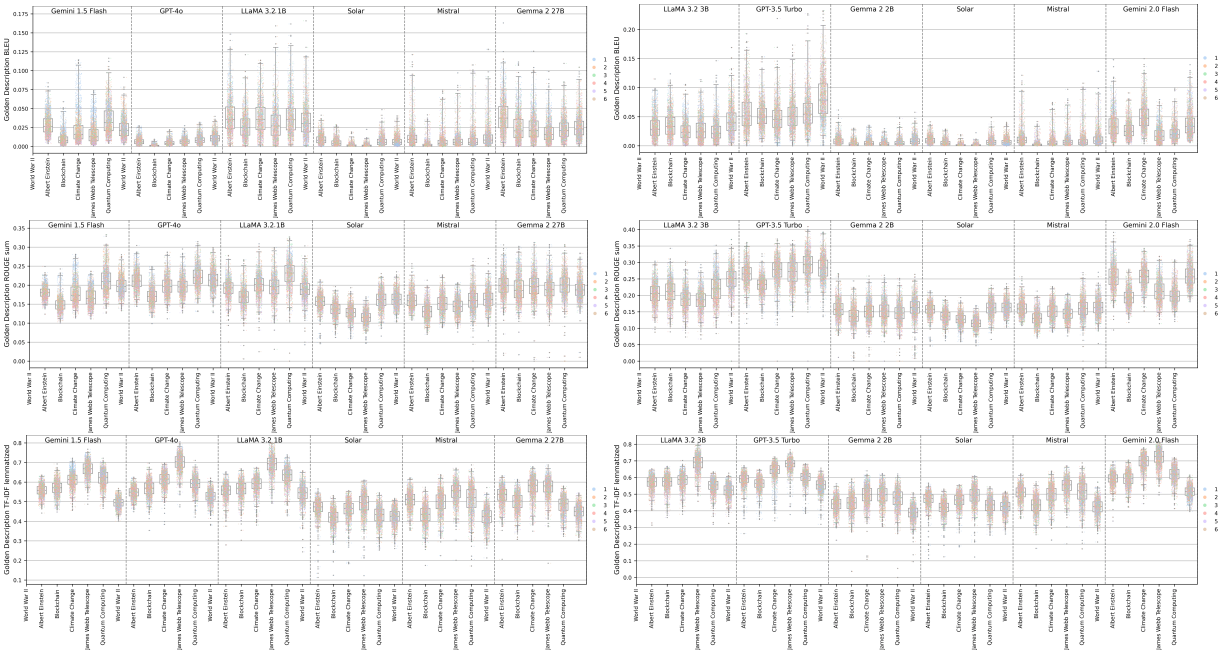

Figure 5: Score distribution per topic across ranking positions of the comparison with reference descriptions

A closer look at individual metrics highlighted different trends. For KC, reference descriptions consistently had lower scores than ranked ones across all models and topics, suggesting that list-based descriptions tended to be more structurally compact or predictable in their information distribution. Similarly, String Length showed a clear discrepancy, reference descriptions were always longer than list-based descriptions, indicating a possible preference for brevity when models are asked to generate multiple descriptions at the same time. For NED, some models showed more alignment with reference descriptions than others, but even in these cases, alignments were not uniform across topics or positions. For example, for `LLaMA 3.2 1B` James Webb Telescope was aligned for all positions but the third one, while Albert Einstein was aligned for the top and last positions, and Blockchain and Climate Change only showed alignment for the last position. On the other hand, `GPT` models only showed alignment for the top position.

While some LLMs displayed topic preferences (with certain models consistently aligning with reference descriptions for specific topics and metrics) this alignment was inconsistent across different positions. For example, for `Mistral` the reference description for Quantum Computing was aligned with the descriptions of all positions, but the last, while Climate Change was aligned with all but the top position. On the other hand, reference descriptions for `Gemini 2.0 Flash` did not align with any position. Many models also showed a neglect of middle-ranked positions, favoring extreme positions (either top or low), as showed in (Liu et al., 2023; Ravaut et al., 2024). The overall lack of agreement between reference and list-based descriptions suggested that the ranking biases were not purely random, but rather driven by specific characteristics that LLMs prioritize differently than the criteria underlying the reference descriptions. This misalignment hinted at the presence of bias in how LLMs evaluate and rank textual descriptions. The tendency for reference descriptions to be treated as outliers (often aligning more with lower-ranked positions rather than the top) suggests that models systematically favor certain linguistic features differently based on the task at hand. The variability across models and topics further emphasized that these biases are not uniform, meaning that different LLMs likely apply different implicit criteria when generating descriptions in a ranking.

C.2.1 Characterization of topic descriptions across ranking positions

Figure 6[24] shows the score distributions per each topic across all ranking positions for the evaluated metrics.

Flesch Reading Ease (FRE). Across LLMs, World War II and Albert Einstein consistently re-

[24] In all cases, full size charts can be found in the companion repository.

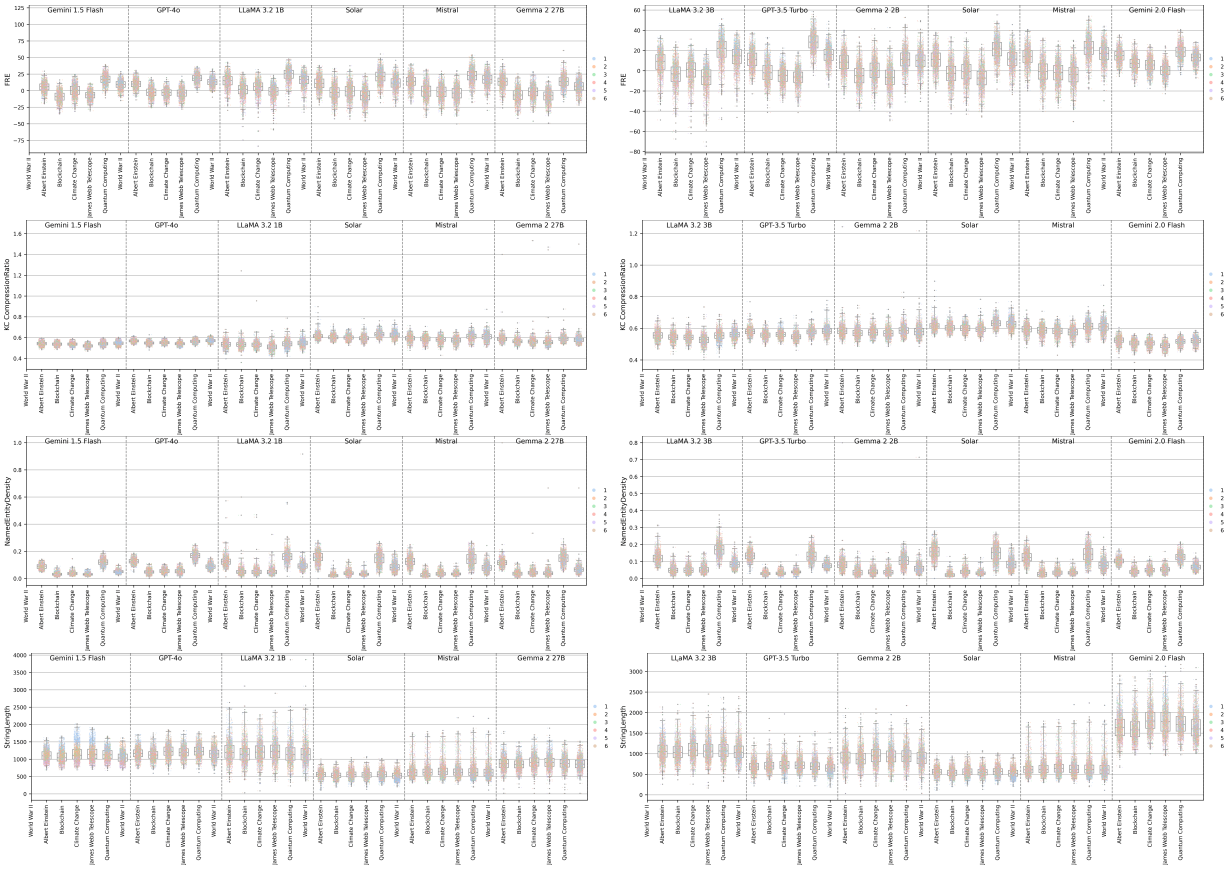

Figure 6: Score distribution per topic across ranking positions

ceived the highest FRE scores. Some models exhibited ranking scoring patterns. For example, `GPT-3.5 Turbo` showed lower scores for the last two positions, particularly for World War II and Albert Einstein, while `LLaMA 3.2` demonstrated a clearer trend where the highest scores appeared in the top-ranking positions. No consistent ranking pattern emerged for other models.

For `LLaMA 3.2` and `GPT-3.5 Turbo`, correlations were generally high across ranking positions, though Blockchain, Albert Einstein, and Quantum Computing introduced some variability. `Mistral` exhibited negative correlations for Blockchain and all other topics, while `Solar` showed negative correlations for Quantum Computing and low positive correlations for Blockchain. `GPT-4o mini` and `GPT-4o` had mixed results for World War II, with low positive and negative correlations, respectively. `Gemini` models displayed more diverse behaviours, with different topic preferences, `Gemini 1.5 Flash` showed stronger patterns for Quantum Computing and Blockchain, while `Gemini 2.0 Flash` showed two groups of topics, one with positive and high correlations (Albert Einstein, Blockchain, Climate Change) and other with low positive/negative correlations (James Webb Telescope, Quantum Computing and World War II). This suggests that both ranking position and topic influenced LLM behavior, with different families showing distinct topic biases.

Most LLMs displayed similar score dispersions (in terms of the standard deviation) across topics. However, `LLaMA 3.2` exhibited greater variation in scores for Quantum Computing, followed by Albert Einstein and World War II, indicating that ranking positions had a stronger effect for these topics. In contrast, `GPT-4o` showed low, consistent dispersions, meaning it treated all topics similarly in terms of FRE scores.

Kolmogorov Complexity - Compression ratio (KC). Some models, like `GPT-4o mini` and `GPT-4o`, showed a slight ranking scoring effect, where middle ranking positions consistently achieved high scores across topics. Other models displayed different tendencies: `LLaMA 3.2` placed the top position among the lower scores for World War II, while `Mistral` and `Gemini 2.0 Flash` consistently assigned lower scores to the last two positions across all topics. `Gemini 1.5 Flash` showed

a distinct ranking effect for Climate Change, where the first position showed low scores while the last ones showed higher ones. For Quantum Computing, the top position tended to show low scores, while middle positions showed higher ones. Notably, Gemini 1.5 Flash-8B produced outliers for Blockchain and World War II, suggesting irregular behavior for these topics.

Correlations varied across models and topics, reinforcing the idea that LLMs treated at least one topic differently, but not consistently across models. Solar showed strong positive correlations across rankings except for James Webb Telescope, which was not the case for FRE. LLaMA 3.2 showed strong correlations for all topics except for Albert Einstein. Similarly, Albert Einstein stood out as the uncorrelated topic for Gemini 1.5 Flash. GPT-4o and GPT-4o mini exhibited weaker correlations specifically for Quantum Computing, though the differences were less pronounced than for the other LLMs. Gemini 1.5 Flash-8B showed low or no correlations for Blockchain and World War II, while GPT-3.5 Turbo treated James Webb Telescope as the outlier. This suggests that while all models exhibited topic-specific biases, they did not consistently focus on the same topics, even within the same LLM family.

Unlike FRE, there were no significant differences in dispersion across ranking positions, with a relatively stable score range across topics. However, the variations in correlation patterns suggested that ranking positions still influenced topics differently, though not necessarily through greater or lower score variability.

Named Entity Density (NED). These scores exhibited similar topic-based patterns across all LLMs, indicating that relative relationships between topics were preserved even as the overall score ranges varied. Unlike the previous metrics, no strong ranking scoring effects were observed. However, two cases stood out. Gemini 1.5 Flash-8B, where the top position consistently appeared among the lower scores while middle positions achieved the highest scores across all topics; and Gemini 1.5 Flash, which showed contrasting topic-specific ranking score tendencies. For Blockchain, last positions had lower scores and the top position had the highest, while World War II showed the opposite behaviour.

Solar displayed strong negative correlations for World War II but had nearly perfect correlations across rankings for all other topics, except Climate Change, which had moderate correlations with all topics. LLaMA 3.2 3B exhibited high correlations among Blockchain, Climate Change, and James Webb Telescope, while the other three topics (World War II, Albert Einstein, Quantum Computing) showed consistently negative correlations with the other topics. Similarly, LLaMA 3.2 1B also showed high correlations between Quantum Computing and Blockchain). For GPT-3.5 Turbo, Albert Einstein was the outlier topic with different correlation behavior. GPT-4o showed slightly weaker correlations for Climate Change and Albert Einstein compared to other topics.

There was a noticeable dispersion in scores, particularly for World War II, which consistently showed the largest one, followed by James Webb Telescope. This suggests that certain topics inherently generated more variation in Entity Density across ranking positions.

String length. This metric exhibited the most noticeable ordering effects across ranking positions, with varying effects across models, and consistent trends across nearly all topics. Solar exhibited perfect correlations across all topics, indicating that its ranking-based length adjustments were highly consistent. Gemini 1.5 Flash and Gemini 1.5 Flash-8B also showed strong positive correlations, though with slight variations. This suggests that while Gemini models treated different topics slightly differently, the effect, for this metric, was subtle. GPT-3.5 Turbo revealed a cluster of four highly correlated topics (Albert Einstein, Blockchain, Climate Change, and Quantum Computing), which were weakly correlated with the remaining two. Mistral showed the lowest correlations for James Webb Telescope and Quantum Computing. Among all models, Solar showed the lowest dispersion, with a maximum of 8 characters. This suggests that Solar maintained more uniform string lengths across ranking positions compared to other models.

C.3 Summary of findings

Findings that varied across models (but not necessarily topics). These indicate architecture- or training-specific behaviors and help explain inconsistency in outcomes.

• Most models preserved topic order, but some reordered or omitted them based on position. Position affected not only content quality but also

content inclusion. Last topics risked being underdeveloped or dropped altogether.

• Middle positions were systematically disadvantaged.

Multiple models (particularly `Gemini` and `Mistral`) tended to neglect middle-ranked topics, both in alignment with reference descriptions and in complexity. This aligns with prior findings (e.g., (Ravaut et al., 2024; Liu et al., 2023)) and suggests that models tended to over-prioritize the start and end of the list, making middle items more susceptible to being different.

• Response complexity and length showed strong positional effects.

Models tended to compress mid-position items and elaborate on extremes, potentially mimicking cognitive primacy/recency effects.

• Differences across models pointed to divergent handling of position.

While positional effects were observed, they manifested differently across model families, hinting at architecture-specific biases rather than universal LLM behavior.

• Readability was impacted by position, but topic difficulty remained consistent.

Positional bias affected presentation style, not just content complexity.

Findings that varied across topics (but not necessarily models). These underscore the importance of topic-dependent priors in model outputs.

• Topic characteristics shaped how positional bias manifests.

Certain topics showed more variations across positions, especially in readability and complexity. This suggests that latent topic-related preconceptions or training-data patterns influenced how LLMs encode and prioritize content when generating multiple outputs. Some topics received consistent structural treatment regardless of position, while others (e.g., Climate Change) shifted in framing or lexical emphasis.

• Certain topics were more prone to omission or simplification.

Especially in lower-ranked positions, indicating that LLMs apply implicit prioritization depending on perceived topic importance or familiarity.

Findings that varied across both models and topics. These highlight nuanced interactions.

• Models diverged in their treatment of topics.

There was no universal topic that behaved the same across models. For example, Climate Change may be aligned with reference descriptions at all positions but the first for one model but not for another. This divergence highlights how both model-specific architectures and pretraining data can shape positional sensitivity differently.

• TF-IDF suggested content preservation, but BLEU/ROUGE differences in word ordering.

Although BLEU scores were low, TF-IDF similarity was moderate, indicating that while LLMs did not replicate phrasing, with some exceptions, they preserved core content across positions. This supports the claim that position influences surface-level features (like word choice) rather than semantic substance, except in edge cases where description length or structure may impair information delivery.

C.4 Qualitative Examples of Positional Bias

To complement the quantitative analysis, we examined selected outputs to qualitatively illustrate how positional bias manifests in practice. These examples show that ranking position not only influences structural features (e.g., length, readability) but also affects topical framing, emphasis, and stylistic tone. Such observations highlight that positional bias extends beyond surface-level metrics and can shape the way information is contextualized and presented.

First, we observe variation in how models allocate length and detail across positions. Table 1 presents two example descriptions of the Climate Change topic, highlighting differences in length and detail for two models. Unlike most models, which tend to elaborate more on earlier items, `Solar` produced longer descriptions for the last positions and noticeably compressed content in earlier ones. By contrast, `Gemini 1.5 Flash` followed the more common pattern of providing richer outputs at the top of the list. This contrast illustrates that positional bias can manifest in opposite ways across models

Further evidence comes from `LLaMA 3.2` on the World War II topic (Table 2). Descriptions at the top positions emphasized key events, prominent leaders, and general consequences (e.g., number of casualties). In contrast, middle positions shifted the focus toward geopolitical consequences and the restructuring of international order. By the last position, the model repeated some of the earlier information but with reduced detail and a lower density of named entities. A similar positional pattern was observed for the Climate Change topic

with LLaMA 3.2.

Finally, outputs from Gemma 2 2B on the Blockchain topic (Table 3) revealed shifts not only in topical framing, alternating between perspectives such as "Technology," "Finance," "Information Network," and "Cryptocurrency", but also in stylistic tone. For example, the description at position 3 adopted a more enthusiastic style, with repeated use of exclamation marks, whereas later positions presented a more neutral, formal tone.

`Gemini 1.5 Flash`: Climate change
(*Position 1: 1296 chars*) Importance: Climate change, driven primarily by human activities, is arguably the most significant environmental challenge facing humanity. Its effects include rising global temperatures, melting glaciers and ice sheets, sea-level rise, more frequent and intense extreme weather events (heatwaves, droughts, floods, hurricanes), biodiversity loss, and disruptions to ecosystems. These changes pose significant threats to human health, food security, water resources, and economic stability. Background: The scientific understanding of climate change emerged gradually over decades, building upon research in atmospheric physics, chemistry, and biology. The burning of fossil fuels (coal, oil, and natural gas) for energy, deforestation, and industrial processes have dramatically increased the concentration of greenhouse gases in the atmosphere, trapping heat and causing a warming effect. Qualities: Climate change is a complex, multifaceted problem with global implications. It requires international cooperation to mitigate its effects through emissions reductions and adaptation strategies. The challenge lies in balancing economic growth with environmental sustainability and ensuring equitable solutions that address the disproportionate impacts on vulnerable populations.
(*Position 6: 878 chars*) Importance: Climate change, driven primarily by human activities, represents a major existential threat to the planet. Its consequences include rising sea levels, more frequent and intense extreme weather events, biodiversity loss, and disruptions to global food systems. Background: The scientific consensus on climate change is that it is largely caused by the increase in greenhouse gas emissions due to the burning of fossil fuels, deforestation, and other human activities. This leads to a warming of the planet and significant changes in global climate patterns. Qualities: Climate change is characterized by its global scale, its long-term consequences, and its complex interplay of natural and human factors. Addressing it requires international cooperation, technological innovation, and significant changes in societal practices and energy systems.
`Solar`: Climate Change
(*Position 1: 411 chars*) Climate Change - Also known as global warming, climate change is a long-term alteration in Earth's climate patterns caused by several factors like greenhouse gas emissions from human activities, deforestation, land use changes, etc. Rising global temperatures have led to severe consequences such as melting ice caps and glaciers, rising sea levels, intensifying weather events, and shifts in wildlife habitats.
(*Position 6: 666 chars*) Climate Change: This term refers to long-term shifts in Earth's climate systems caused by changing weather patterns interwoven with human activities contributing to altered global temperatures (global warming) and ecosystem transformations. The primary triggers include greenhouse gas emissions, deforestation, industrial developments, and unsustainable agricultural practices leading to rising sea levels, increased frequency of extreme weather events (droughts, hurricanes), melting glaciers, and loss/displacement of flora-fauna species. Combating climate change requires concerted global efforts in promoting sustainable practices and environmental conservation.

Table 1: Qualitative response examples exemplifying the variations in String length across models and positions

Position 1 World War II (1939-1945) was a global conflict between the Axis powers (Germany, Italy, Japan) and the Allied powers (primarily the United States, Great Britain, France). It resulted from the complex web of alliances and rivalries that had formed in the aftermath of World War I. Key Events: Germany invaded Poland (1939) and France (1940). Japan launched surprise attacks on Pearl Harbor (1941), Australia, New Zealand, and the Philippines. The United States entered the war after the Japanese attack followed by Germany's invasion of France and Britain. The Allies won major battles at Midway, Stalingrad, D-Day, and Iwo Jima. Atomic bombings of Hiroshima and Nagasaki led to Japan's surrender (1945). Consequences: World War II was the deadliest conflict in human history, resulting in over 50 million deaths. It significantly reshaped the international order, redrew national borders, and marked a turning point in modern politics. Important leaders: Winston Churchill, Franklin D. Roosevelt, Adolf Hitler, Joseph Goebbels, Joachim von Ribbentrop Key outcomes: establishment of NATO, European Economic Community (precursor to the EU), rise of international organizations such as the United Nations, development of nuclear technology and atomic bombings.
Position 3 World War II was a global conflict that lasted from 1939 to 1945, involving most countries in the world at some point. It began with the invasion of Poland by Nazi Germany and Italy, which led to the United Kingdom, France, and the Soviet Union declaring war on Germany. The war had far-reaching consequences: Global devastation: Millions of people lost their lives, cities were destroyed, and economies suffered. New World Order: The international system was fundamentally transformed, with the emergence of the United States and the Soviet Union as superpowers. Rise of the United Nations: In response to the devastation of World War I and II, a new global organization aimed at promoting peace and security.
Position 6 World War II (1939-1945) was a global conflict that involved most of the world's nations. It began when Nazi Germany, under Adolf Hitler, invaded Poland, prompting Britain and France to declare war. Background: The war was fought between two main alliances: the Allies (including the United States, the Soviet Union, and the United Kingdom) and the Axis powers (Germany, Italy, and Japan). Qualities: - Massive devastation of entire nations and cities - Industrial-scale use of atomic bombs by both sides - Implementation of strategic bombing campaigns, including firebombing of civilian areas - Daring military maneuvers, such as Operation Overlord (D-Day Invasion) Importance: World War II had a profound impact on world history: - Led to the defeat of Nazi Germany and Imperial Japan - Shaped the modern global order with the establishment of the United Nations - Spurred technological innovations in computing, radar, and plastics - Had significant social and cultural implications due to mass destruction and genocide, particularly during the Holocaust.

Table 2: Qualitative response examples exemplifying changes of focus in the descriptions across positions for `LLaMa 3.2` and the topic *World War II*

Position 3
Blockchain (Technology) Importance: Blockchain technology, first developed as the ledger system used in Bitcoin, is rapidly changing the way finance and various industries operate! Background: The concept originated in 2008 with a white paper by Satoshi Nakamoto. A decentralised database allows for efficient record-keeping, transparency, and security when it comes to transactions like currency or other data! This decentralized ledger, while built on cryptography and math, can help eliminate fraud and theft of assets! Qualities: Secure transactions, transparent data, decentralization for power shift.
Position 6
Blockchain (Finance) Importance: A revolutionary technology that enables secure, transparent, and decentralized records of transactions. It has found widespread application in revolutionizing various sectors like finance, supply chain management, and governance. Background: Popularized by Bitcoin as a way to record transactions without the reliance on intermediaries, blockchain technology offers decentralized systems where information is recorded in "blocks" verified by a network consensus through cryptography and complex mathematical protocols. Qualities: The core principle of blockchain lies in its use of a public, decentralized ledger system (like a secure digital "notebook") that prevents tampering and allows for verifiable transparency. This technological advancement holds vast potential.

Table 3: Qualitative response examples exemplifying changes in the focus of descriptions and the usage of punctuation symbols for `Gemma 2 2B` and the topic *Blockchain*