Bias or Factual Recall? Understanding How LLMs Compare Entities.

Anonymous ACL submission

Abstract

We analyze the ability of LLMs to answer comparison questions (e.g., *"Which is longer, the Danube or the Nile?"*). Our central observation is that LLMs often make mistakes when answering such questions, even when they have the required knowledge (e.g. the length of the rivers involved). We furthermore find that their predictions are heavily influenced by superficial biases, such as the position of the entities in the question, their relative popularity, and shallow co-occurrence statistics. These findings suggest that simple prompting-based strategies may not leverage the ranking abilities of LLMs to their full potential, and that LLMs continue to struggle with even simple reasoning tasks.

1 Introduction

Ranking is at the core of many applications, and it is thus perhaps not surprising that LLMs are increasingly used for this purpose. For example, they are commonly used for (re-)ranking in document retrieval (Sun et al., 2023; Qin et al., 2024; Ma et al., 2024) and recommendation (Gao et al., 2025), and for evaluating models in LLM-as-a-judge settings (Zheng et al., 2023; Liusie et al., 2024). Different strategies can be used for ranking, including *pointwise* methods, which assign a score to each item, and *pairwise* methods, which compare two items.[1] Pointwise methods are easier to use, but in the learning-to-rank literature they are consistently found to underperform pairwise methods. In the context of LLMs, however, the relative merit of these approaches remains unclear; e.g., Qin et al. (2024) find the pairwise approach to be superior, but Tripathi et al. (2025) find that pairwise approaches are more susceptible to biases.

In this paper, we aim to increase our understanding of pairwise ranking with LLMs, by focusing on a simplified task where LLMs are asked to compare two entities according to some factual numerical attribute (e.g., "Which river is longer, the Danube or the Nile?"). This task has the advantage that there is a clear, unambiguous ground truth, which facilitates analysis. Moreover, it allows us to study whether LLMs follow a principled strategy (i.e., retrieve the attributes for the entities, then compare their values) or rather rely on heuristics.

In particular, we ask the question: **Do LLMs use numerical attributes for pairwise comparisons?** (see Section 3). We show that the pairwise predictions are often inconsistent with predicted attribute values, which suggests that LLMs do not consistently exploit their internal knowledge about these attributes. This is despite the fact that the pairwise approach underperforms a pointwise approach based on predicted attribute values. To better understand the underlying reasons, we ask our next question: **How susceptible is the pairwise approach to biases?** (see Section 4). We show that pairwise predictions are strongly biased by the position of an entity in the prompt, entity popularity, and shallow co-occurrence statistics. Given the observed strength of these biases, we finally ask: **To what extent can an LLM's pairwise predictions be *explained* by these surface cues?** (see Section 5). We find that the majority of model predictions can indeed be explained by these biases.

2 Experimental Setup

Datasets. We focus on a pairwise ranking task. We prompt the language model with direct comparison questions (e.g., "Which river is longer, the Danube or the Rhine?") and evaluate whether the model selected the correct item according to the ground truth. To obtain a sufficiently large set of test queries, we collected data on 10 different numerical attributes across 9 entity types from Wikidata (`https://www.wikidata.org`). For each at-

[1] Listwise approaches form a third category, but these will not be considered in this paper.

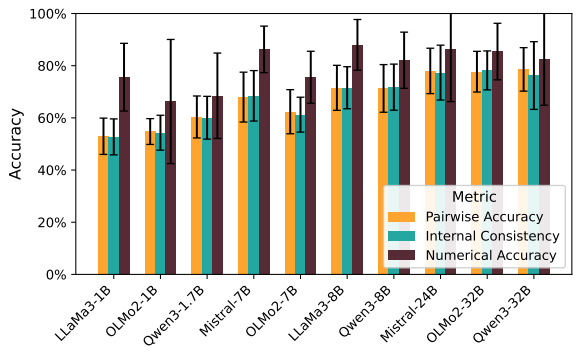


Figure 1: Overall pairwise-ranking performance of each language model. For every model, we report the mean and standard deviation across all datasets for three evaluation metrics.

tribute, we begin by selecting the most popular entities, based on their QRank score (https://qrank.toolforge.org). To obtain a set of entity pairs that span a range of difficulty levels, we divide the set of entities into two bins of the same size based on their ground-truth values. For every entity, we construct two comparison pairs by randomly selecting one partner from each bin. Details on the resulting dataset can be found in Appendix B.

Prompting Strategy. The performance of LLMs can be sensitive to the choice of prompt. For this reason, each entity pair is evaluated across six prompt templates. In the first three templates, we ask which of the two entities has the highest attribute value. In the remaining three templates, we ask for the entity with the lowest value. Furthermore, for each template, we prompt the model twice for every entity pair, i.e., once for either of the possible entity orderings (e.g., `(Danube, Nile)` and `(Nile, Danube)`). In total, we thus have $6 \times 2 = 12$ prompts per entity pair.

In addition, we also prompt the model to predict the numerical attribute values of the entities. To this end, we use three numerical extraction templates for each attribute and select the prediction with the lowest perplexity (i.e., we select the model's most confident numerical estimate). The full set of prompt templates is listed in Appendix D.

Evaluation Metrics. We assess model performance along three dimensions. First, we measure *pairwise accuracy*, defined as the proportion of pairwise predictions that are correct according to the ground truth. Second, we compute *internal consistency*, which we define as the proportion of pairwise predictions that are in agreement with the ranking implied by the model's own numerical predictions. Finally, we evaluate *numerical accuracy*, which evaluates the quality of the model's predicted attribute values. It is defined as the proportion of pairwise comparisons for which the ranking implied by the predicted numerical values agrees with the ground truth ranking. Note that this evaluates a pointwise approach. To ensure comparability, we remove all samples for which the model did not produce a valid answer, either in the pairwise or numerical setting. As a result, all metrics are computed over the same filtered set of samples.

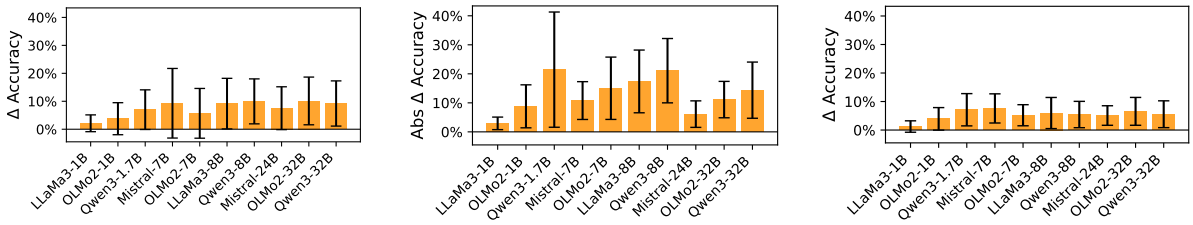
Models. We experiment with models of different families and sizes: Llama3-1B, Llama3-8B (Dubey et al., 2024), OLMo2-1B, OLMo2-7B, OLMo2-32B (OLMo et al., 2025), Qwen3-1.7B, Qwen3-8B, Qwen3-32B (Yang et al., 2025), Mistral-7B (Jiang et al., 2023a) and Mistral-24B. Full details on these models can be found in Appendix A.

3 Do LLMs Use Numerical Attributes for Pairwise Comparisons?

Figure 1 summarizes the performance of the different language models, averaged across all 10 attributes. A more detailed breakdown can be found in Appendix E. A number of important findings can be observed. First, numerical accuracy is consistently and substantially higher than pairwise accuracy, showing that pairwise ranking underperforms pointwise ranking on our task. For the smallest models, pairwise accuracy is barely above random chance. Pairwise accuracy increases with model size. For numerical accuracy, on the other hand, Mistral-7B and Llama3-8B both outperform much bigger models. For these models, the underperformance of the pairwise approach can thus not be explained by a lack of knowledge (cf. Section 5). This can also be clearly seen from the surprisingly low internal consistency values, which are even lower than pairwise accuracy in most cases. This means that (for smaller models) the pairwise predictions are inaccurate, both relative to the ground truth and relative to their own internal beliefs. Overall, the results suggest that LLMs rely on shortcuts when making pairwise predictions, which we will further analyze in the next section.

4 How Susceptible Is the Pairwise Approach to Biases?

We analyze the impact of three types of biases on the pairwise predictions. A breakdown of the



(a) Popularity bias (b) Position bias (c) Co-occurrence bias

Figure 2: Accuracy differences illustrating three types of bias in pairwise ranking decisions.

results can be found in Appendix E.

Popularity Bias. A heuristic that LLMs may exploit is that popular entities might have higher values (e.g. cities that are mentioned more often may have higher populations). To analyze this effect, we estimate the popularity of each Wikidata entity using its QRank score, which reflects the number of page views of the corresponding Wikipedia article, and some additional sources. The results are summarized in Figure 2a. The figure shows the difference in accuracy between comparisons where the most popular entity has the highest value, and comparisons where the opposite is true. As can be seen, this accuracy difference is positive for all models, which shows that LLMs are indeed biased by entity popularity. Interestingly, increasing model size does not seem to reduce this effect.

Position Bias. LLM evaluators have been found to suffer from position bias, favouring responses depending on the order in which they are presented (Wang et al., 2024). We analyze whether a similar bias is also present when comparing entities. To this end, we compare the accuracy across two sets of comparisons: those where the first or second entity has the higher value. Figure 2b summarizes the results. Note that position bias is not consistent across models: some models favour the first entity, while others favour the second entity. We therefore report the absolute value of the difference in accuracy. We find that all models are affected by position bias. There is no clear relationship between model size and the strength of this bias.

Co-occurrence Bias. LLM predictions can be affected by shallow co-occurrence statistics (Kang and Choi, 2023). To analyze this effect, we rely on the ConceptNet Numberbatch pre-trained word embeddings (Speer et al., 2017) as a model of distributional similarity. For each numerical attribute, we selected 5 adjectives that are indicative of high values (e.g. *longest* for river length) and averaged their embeddings, yielding a vector $\mathbf{v}^+$. We do the same for 5 adjectives that are indicative of low values (e.g. *shortest*) and obtain $\mathbf{v}^-$. We then score entity e as $\cos(\mathbf{e}, \mathbf{v}^+ - \mathbf{v}^-)$, where $\mathbf{e}$ is the Numberbatch embedding of e.[2] Figure 2c analyzes the co-occurrence bias, showing the difference in accuracy between comparisons where the entity with the highest score has the higher ground truth value and those where the opposite is true. We can see that all models suffer from co-occurrence bias, although the magnitude of this effect is smaller than for popularity and position bias.

5 Can an LLM's Pairwise Predictions Be *Explained?*

The fact that the pairwise predictions are biased in some ways is, in itself, not unexpected. However, the magnitude of these biases (0–20%) is more surprising, noting that the pairwise accuracy is only 0–20% above random guessing for most models (cf. Figure 1). We may thus wonder to what extent these biases are enough to explain the pairwise predictions. To analyze this, we first train a logistic regression model, which we call a *meta-predictor*, to predict whether the LLM will predict the first or second entity. It takes as input two binary features: whether the first entity is more popular than the second, and whether its embedding is more similar to the vector $\mathbf{v}^+ - \mathbf{v}^-$. Note that position bias is implicitly taken into account by design. The meta-predictor is evaluated using 5-fold cross validations. It is trained for a particular LLM, prompt and attribute. We average the results across different prompts and attributes.

We can now predict an LLM's pairwise judgments based on (i) its own numerical knowledge about an attribute (i.e. pointwise prediction); or

[2]Full details of how the scores are obtained can be found in Appendix C.

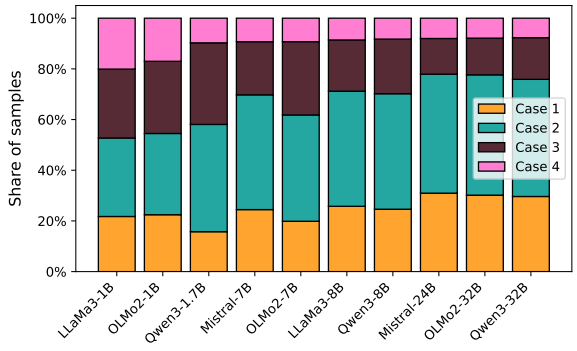


Figure 3: Share of the four cases for every model, aggregated over all datasets.

(ii) the meta-predictor. We distinguish between the following cases. **Case 1:** pairwise and pointwise predictions are *consistent*, but pairwise and meta predictions are *not consistent* (⇒ numerical knowledge is used); **Case 2:** pairwise, pointwise and meta predictions are *consistent* (⇒ numerical knowledge or bias); **Case 3:** pairwise and pointwise predictions are *inconsistent*, but pairwise and meta prediction are *consistent* (⇒ bias is used); **Case 4:** pairwise and pointwise predictions are *inconsistent*, and pairwise and meta predictions are *inconsistent* (⇒ noise or unexplained bias)

In Figure 3, we plot the relative frequency of the four cases for every model, across all datasets. We observe the following. Case 4 is infrequent, suggesting that the three biases and the pointwise predictions can almost completely explain the pairwise predictions, with an exception for the two smallest models. For LLMs with lower pairwise accuracy (i.e., LlaMa3-1B, OLMo2-1B, Qwen3-1.7B, OLMo2-7B), Case 3 occurs more often than Case 1.[3,4] Many of the predictions can be explained by either bias or numerical knowledge (Case 2), which may explain why the biases are prevalent in the first place (i.e. popularity and co-occurrence bias are somewhat predictive). Overall, the three biases are sufficient to explain the predictions in the majority of instances (i.e., Case 2 + Case 3).

[3]See also Figure 15 in Appendix F.

[4]Figure 3 also gives an explanation for our observation in Section 3, i.e., even though Mistral-7B and Llama3-8B perform *better* wrt. numerical accuracy than larger models (e.g., Mistral-24B) (see Figure 1), they perform *worse* wrt. pairwise accuracy: compared with larger models, Mistral-7B and Llama3-8B are more influenced by biases (i.e., Case 3) than numerical knowledge (i.e., Case 1) when doing pairwise comparisons (see also Figure 15 in Appendix F).

6 Related Work

Previous work has already found that LLM predictions can be influenced by various types of superficial features. Wang et al. (2024) identified a position bias in LLM evaluators, where the result is influenced by the order in which candidates are presented. McCoy et al. (2023) found how the accuracy of an LLM is influenced by the probability of the output, which aligns with our findings of popularity bias. The fact that shallow co-occurrence statistics can mislead LLMs, being the third bias that we study, has also been shown in several studies (Kang and Choi, 2023). While it is thus not surprising that these biases are present in our analysis, the significance of our finding stems from the extent to which these biases affect the result: these three biases together almost completely explain pairwise judgments for smaller models. The lack of internal consistency of LLMs with numerical features also aligns with various findings from the literature. In the context of ranking, the non-transitive nature of pairwise judgments by LLMs has been highlighted (Xu et al., 2025; Kumar et al., 2024). The reversal curse (Berglund et al., 2024), where models fail to answer inverse formulations of questions, also suggests a lack of internal consistency. Allen-Zhu and Li (2024) also find that LLMs sometimes memorize knowledge without being capable of reliably exploiting it for answering questions. The problem of ranking entities with LLMs was studied by Kumar et al. (2024), but their focus was on designing fine-tuning strategies.

7 Conclusion

We have analyzed how LLMs behave when asked to rank entities according to some well-defined numerical attribute. Intuitively, an LLM could extract the attribute values for the two given entities and simply compare these. We found that LLMs can generally approximate the numerical attributes sufficiently well for such a strategy to be successful. However, the actual performance of LLMs on pairwise comparisons dramatically underperforms this strategy. We then showed that these pairwise predictions are affected by at least three biases, namely popularity bias, position bias and co-occurrence bias. Finally, we showed that together, these three biases are highly predictive of model predictions, especially for smaller models, suggesting that these biases largely drown out more principled mechanisms that may be present in the models.

Limitations

Our study has been limited to an analysis of the outputs of LLMs, and we have not attempted to interpret these models mechanistically. For instance, it would be interesting to see whether (or under which conditions) updating the numerical knowledge inside models would alter their pairwise judgments. Furthermore, our analysis has been limited to zero-shot prompting. In preliminary experiments, we observed that few-shot prompting may help to partially overcome some of the biases that we studied, although not entirely. Similarly, it would be interesting to study whether the biases persist after fine-tuning models on pairwise ranking tasks.

References

Mistral AI. 2024. Mistral small 3. https://mistral.ai. Large Language Model.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2025. Llm4rerank: Llm-based auto-reranking framework for recommendations. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 228–239. ACM.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023b. Mistral 7b. *Preprint*, arXiv:2310.06825.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore. Association for Computational Linguistics.

Nitesh Kumar, Usashi Chatterjee, and Steven Schockaert. 2024. Ranking entities along conceptual space dimensions with LLMs: An analysis of fine-tuning strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7974–7989, Bangkok, Thailand. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2421–2425. ACM.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *CoRR*, abs/2309.13638.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *CoRR*, abs/2501.00656.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling

Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating non-transitivity in llm-as-a-judge. *CoRR*, abs/2502.14074.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Appendices

A Model Details

Unless specified otherwise, all models were run with greedy decoding and thinking was disabled, if applicable. Models with more than 10B parameters were run in 8 bit quantization. All other models were run with 16 bit floating point precision. An overview of all models used, along with citations and Hugging Face repository links, is provided in Table 1.

B Dataset Details

Table 2 summarizes the datasets used in our experiments, including the target attribute, number of entities, unique entity pairs, and total samples evaluated.

C Details on Numberbatch Embeddings

If the word is out-of-vocabulary, we apply a two-step back-off strategy. We first attempt *token-level averaging*: the word is split into individual words, and any tokens that are found in the vocabulary are embedded individually and averaged. If no token vector is obtained, we fall back to *prefix matching*, progressively trimming the word from the end until the longest prefix that *is* in the vocabulary is located and using its vector as a surrogate. Given the resulting entity embedding $\mathbf{e}$, we compute its scalar projection onto the "bigger–smaller" axis as $s = \langle \mathbf{e}, \mathbf{d} \rangle$, where more-positive scores correspond to stronger semantic alignment with "larger" and more-negative scores with "smaller". Table 3 provides qualitative examples in which cosine similarity to attribute-related keywords (e.g., "larger," "bigger," "more") suggests the wrong ranking, highlighting potential co-occurrence bias. The list of positive and negative keywords used to construct the "bigger–smaller" axis is shown in Table 4.

D Prompts

Table 5 to Table 26 list the prompt templates used in our experiments. Each attribute–dataset combination includes six pairwise prompts (three prompting for the "larger" entity and three for the "smaller" one) and three numerical extraction prompts.

E Detailed Accuracies

As explained in the paper, we use both prompts that ask for the entity with the highest value and prompts that ask for the entities with the lowest value. We refer to these as prompts with *positive polarity* and *negative polarity*, respectively. As there are some differences in the results between prompts with positive and negative polarity, we report results for these types of prompts separately.

Main Accuracy Metrics. Figures 4 and 5 report accuracy comparisons for ranking accuracy, internal consistency, and numerical accuracy under positive and negative polarity prompts, respectively. Each panel corresponds to a specific model, and each group of bars represents performance on one dataset.

Popularity Bias. We analyze whether models are more accurate when the more popular entity (based on QRank) also has a higher value. Figures 6 and 7 show the impact of popularity bias for positive and negative polarity prompts, respectively. Solid bars indicate accuracy when the popular entity has the higher value, and hatched bars when it has not. A large accuracy drop in the latter case suggests reliance on popularity as a heuristic.

Position Bias. Figures 8 and 9 examine whether the order in which entities are mentioned affects model predictions. Specifically, we compare accuracy when the entity with the higher value appears first vs. second in the prompt. Consistent differences indicate a surface-level position bias.

Co-occurrence Bias. Finally, we assess whether models perform better when the entity with the higher value is more semantically associated with "larger" or "smaller" concepts based on a keyword-derived embedding axis. Figures 10 and 11 show this effect for positive and negative polarity prompts, respectively. Accuracy gaps here suggest reliance on semantic co-occurrence cues even when they contradict ground-truth values.

F Detailed Meta-predictor Analysis

This section provides a deeper analysis of the two meta-predictors introduced in Section 5. The first is a numerical oracle that predicts pairwise outcomes based solely on the model's extracted numerical values. The second is a logistic-regression meta-predictor trained on surface-level cues: entity popularity (QRank), positional advantage (first vs. second entity), and semantic similarity to "larger" keywords (co-occurrence bias).

Figures 12 and 13 show the absolute prediction accuracy of both meta-predictors across all datasets and models, for positive and negative polarity prompts respectively. The comparison highlights the extent to which models' choices can be explained by shallow heuristics rather than grounded numerical reasoning. Figure 14 presents a breakdown of the four diagnostic cases discussed in Section 5. Each panel corresponds to a different model, and each bar to a dataset. This figure complements the main paper's analysis by revealing which types of errors are most prevalent in each domain, and whether failures to follow numerical predictions correlate with surface-level biases. Finally, Figure 15 summarizes how much better the bias-only meta-predictor performs compared to the numerical baseline for each model. Positive values indicate that surface-level features are more predictive of the model's behavior than its own internal numeric extractions—evidence of strong reliance on popularity, position, and co-occurrence cues.

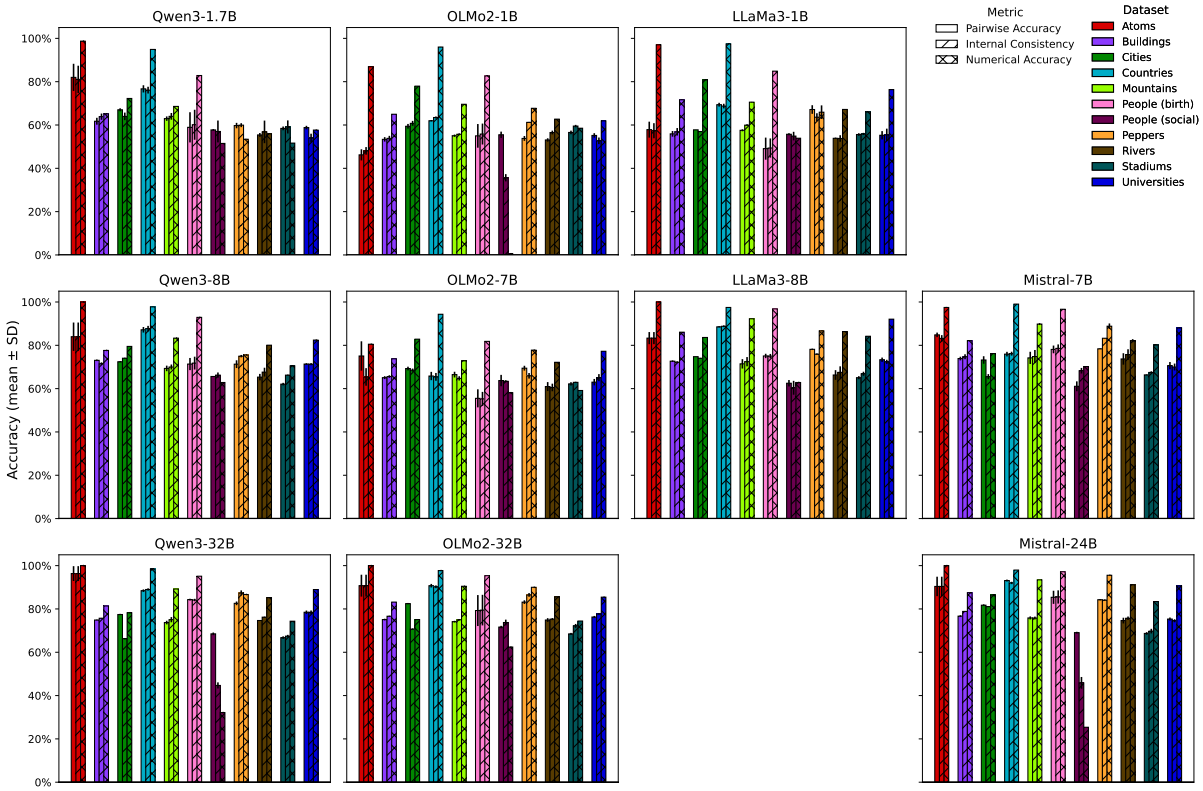


Figure 4: **Accuracy comparison for positive polarity prompts**. Each panel shows results for a single language model. For each dataset, we display three side-by-side bars: the solid bar represents the ranking accuracy, the hatched bar shows the internal consistency and the cross-hatched bars show the numerical accuracy. Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates.

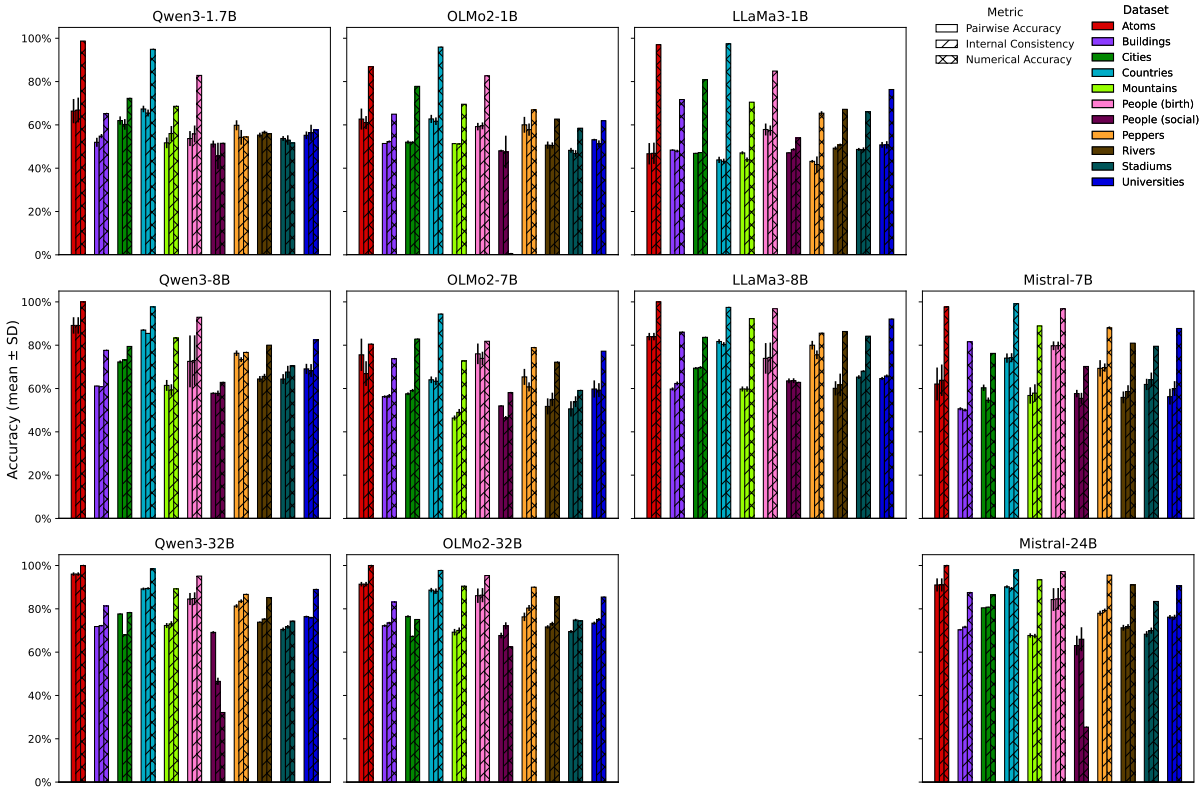


Figure 5: **Accuracy comparison for negative polarity prompts**. Same layout as Figure 4, but for negative polarity prompts.

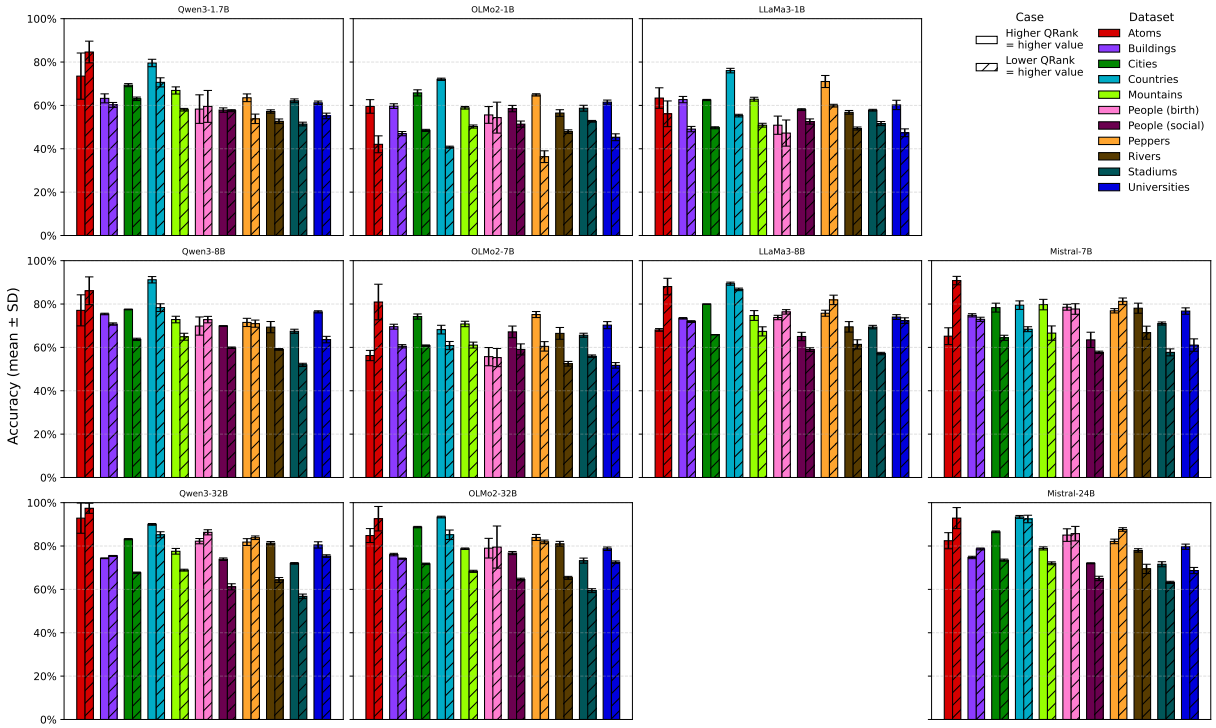


Figure 6: **Accuracy comparison in different popularity settings for positive polarity prompts.** For each dataset, we display two side-by-side bars: the solid bar represents the accuracy of the model in cases where the entity with the higher QRank, also had the higher value, the hatched bar shows the accuracy of the model for cases where the lower QRank entity had the higher value. Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates. Each panel shows results for a single language model.

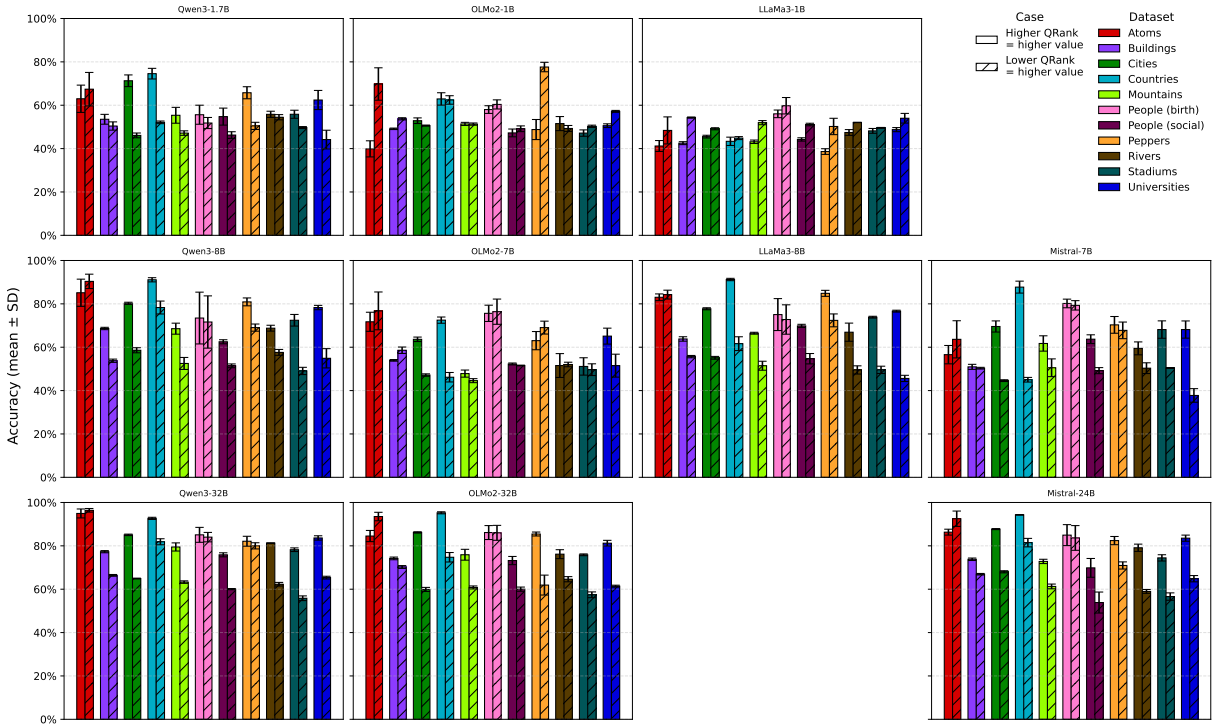


Figure 7: **Accuracy comparison in different popularity settings for positive negative prompts.** Same layout as in Figure 6, but for negative polarity prompts.

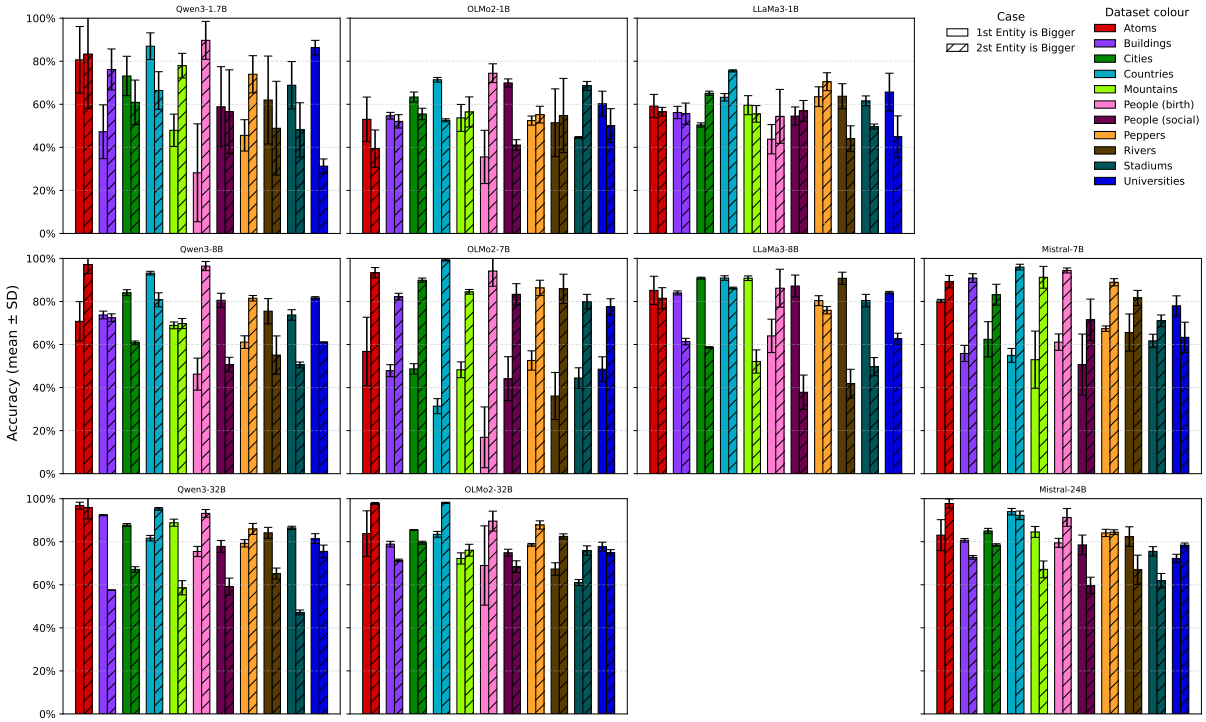


Figure 8: **Accuracy comparison depending on which position the bigger entity had for positive polarity prompts.** For each dataset, we display two side-by-side bars: the solid bar represents the accuracy of the model in cases where the first mentioned entity had the bigger value, the hatched bar shows the accuracy of the model in cases where the second mentioned entity had the bigger value. Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates. Each panel shows results for a single language model.

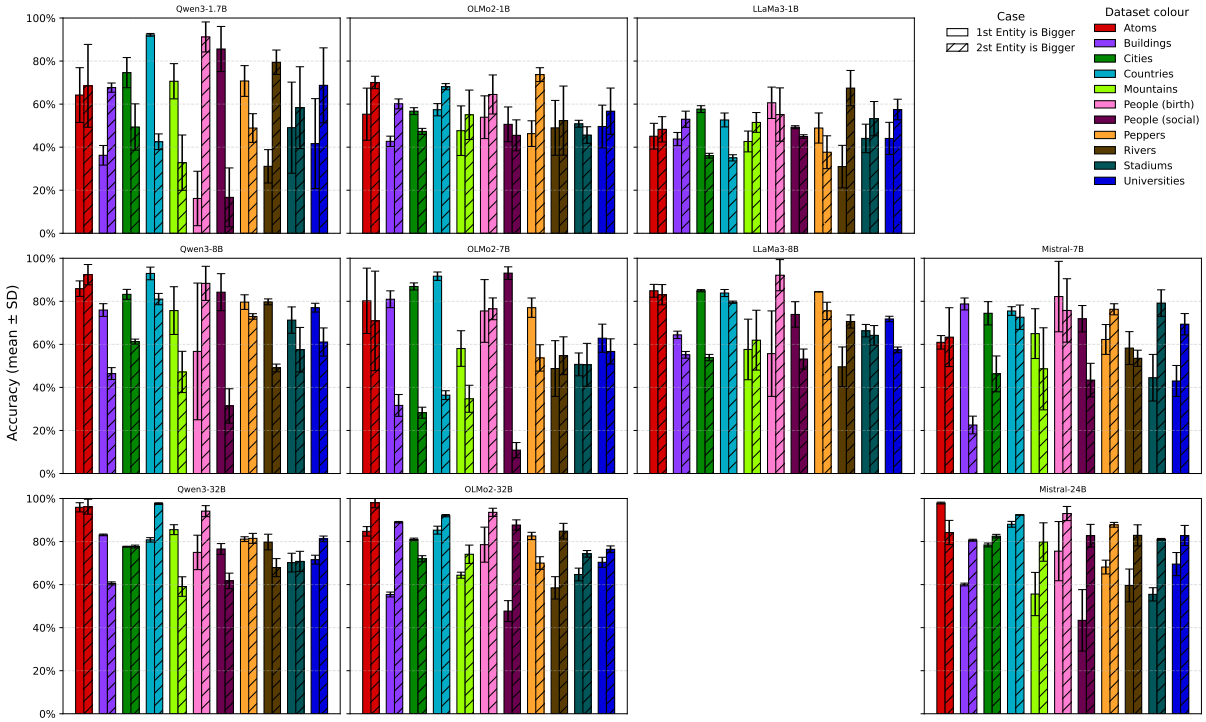


Figure 9: **Accuracy comparison depending on which position the bigger entity had for positive polarity prompts.** Same layout as in Figure 8.

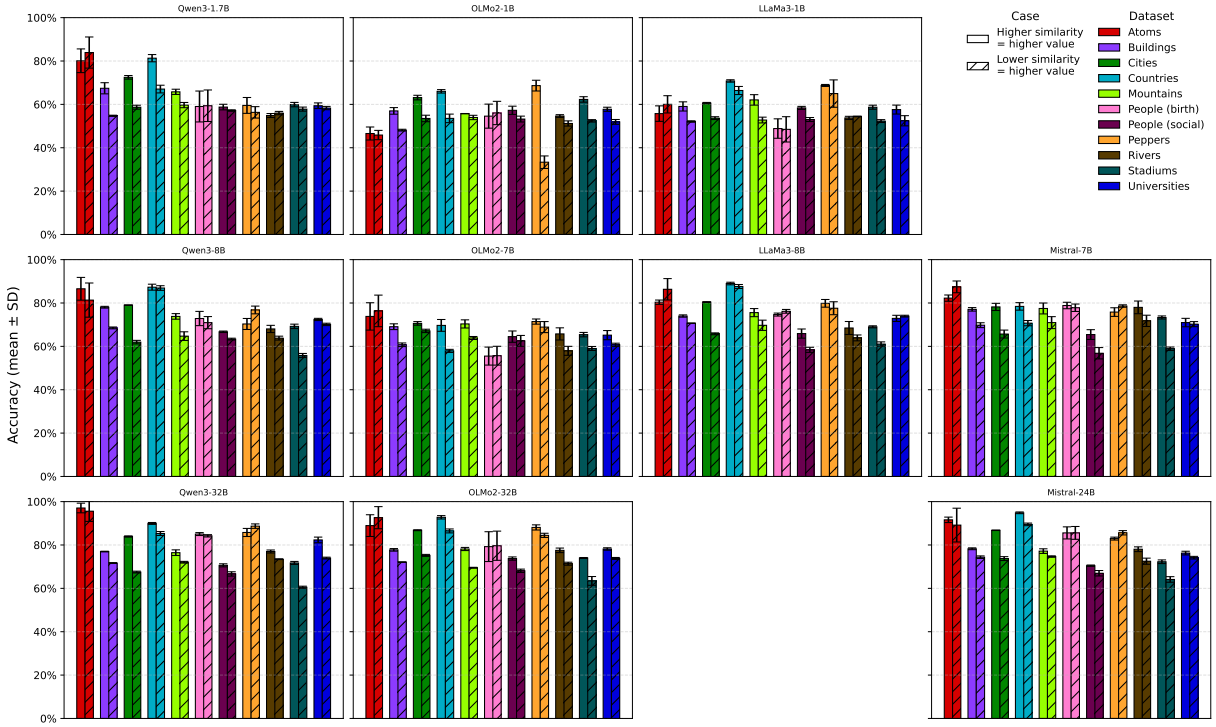


Figure 10: **Accuracy comparison depending on PMI proxy of the entities for positive polarity prompts.** For each dataset, we display two side-by-side bars: the solid bar represents the accuracy of the model ..., the hatched bar shows the accuracy of the model ... Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates. Each panel shows results for a single language model.

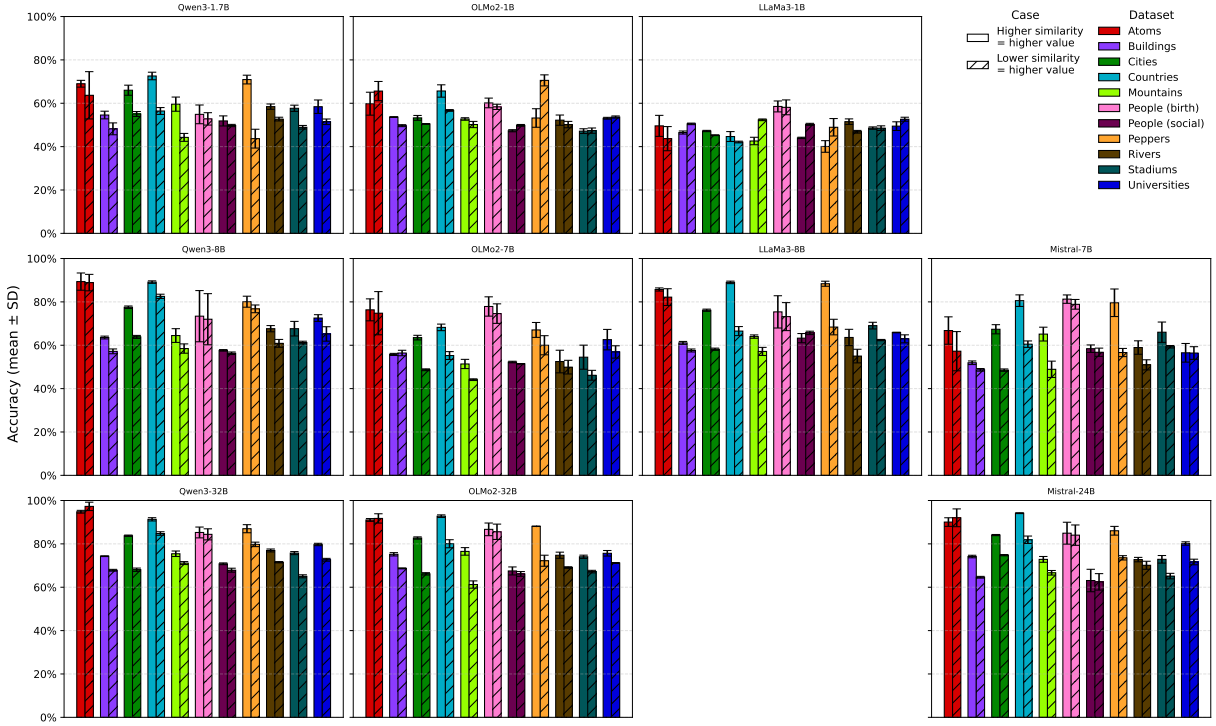


Figure 11: **Accuracy comparison depending on PMI proxy of the entities for positive polarity prompts.** Same layout as Figure 10.

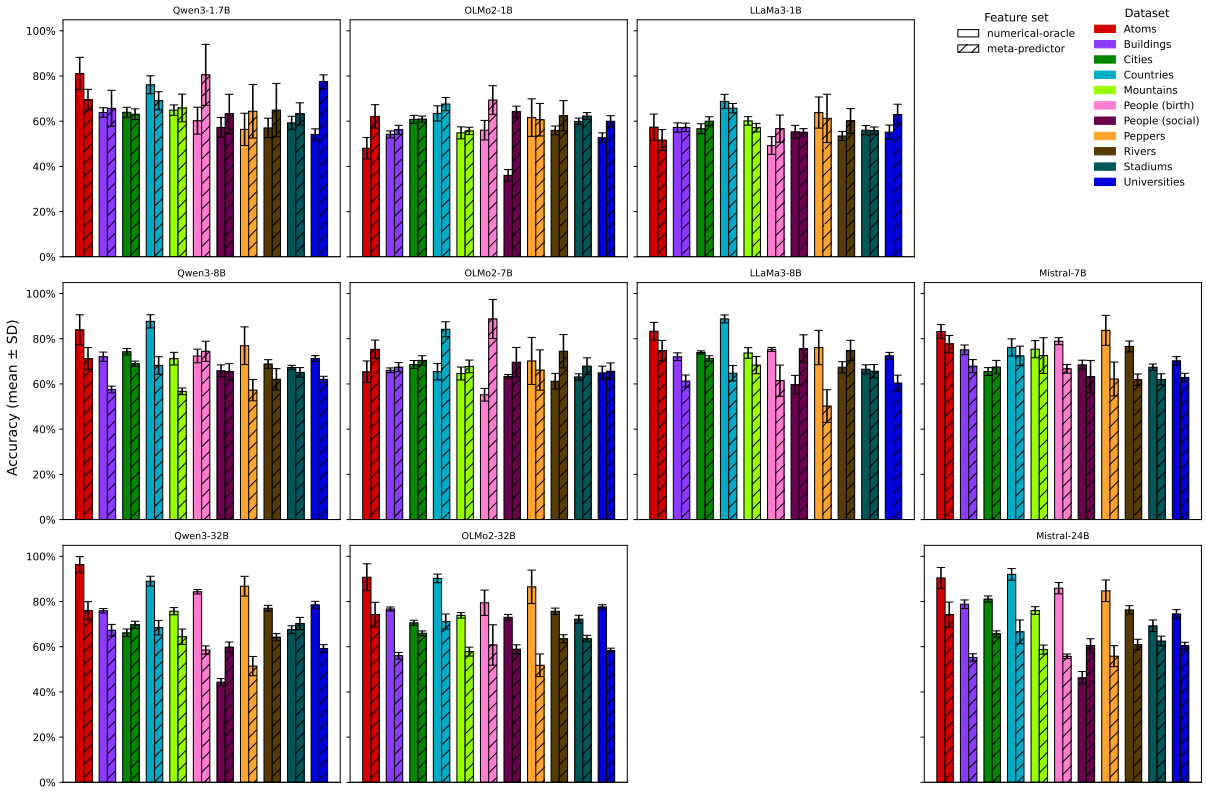


Figure 12: **Absolute prediction accuracy of the two meta-predictors (positive polarity).** For each dataset, we display two side-by-side bars: the solid bar represents the numerical-oracle baseline, which follows the model's extracted numbers to predict its choice; the hatched bar shows the accuracy of a logistic-regression meta-predictor that uses only surface-level cues (popularity, positional advantage, and semantic association with "bigger" keywords). Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates. Each panel shows results for a single language model.

Model	Hugging Face Repository
LLaMa3-1B (Grattafiori et al., 2024)	meta-llama/Llama-3.2-1B-Instruct
OLMo2-1B (OLMo et al., 2024)	allenai/OLMo-2-0425-1B-Instruct
Qwen3-1.7B (Team, 2025)	Qwen/Qwen3-1.7B
Mistral-7B (Jiang et al., 2023b)	mistralai/Mistral-7B-Instruct-v0.3
OLMo2-7B (OLMo et al., 2024)	allenai/OLMo-2-1124-7B-Instruct
LLaMa3-8B (Grattafiori et al., 2024)	meta-llama/Llama-3.1-8B-Instruct
Qwen3-8B (Team, 2025)	Qwen/Qwen3-8B
Mistral-24B (AI, 2024)	mistralai/Mistral-Small-24B-Instruct-2501
OLMo2-32B (OLMo et al., 2024)	allenai/OLMo-2-0325-32B-Instruct
Qwen3-32B (Team, 2025)	Qwen/Qwen3-32B

Table 1: Model information for the models used in this paper.

Dataset	Attribute	Entities	Pairs	Samples
Atoms	atomic number	118	236	2832
Buildings	height	1000	2000	24000
Cities	population	1000	2000	24000
Countries	population	196	392	4704
Mountains	elevation	997	1994	23928
Peppers	Scoville heat unit	45	90	1080
People	birth date	4777	9554	114648
People	social media following	999	1998	23976
Rivers	length	999	1998	23976
Stadiums	capacity	999	1998	23976
Universities	Nr. enrolled students	1000	2000	24000

Table 2: Key statistics of the considered datasets: the number of entities, the number of unique entity pairs used in the analysis, and the total number of samples. Each pair is evaluated using 6 prompt templates and both possible entity orderings, resulting in 12 samples per pair.

Dataset	Cosine Suggests	Actually Larger
People (social)	George Michael ($\sim$ 559k followers)	Mackenyu ($\sim$ 1M followers)
Buildings	Red Fort (33 m)	Colonius (266 m)
Atoms	chromium (24)	niobium (41)
Universities	University of Mannheim ($\sim$ 12k students)	George Washington University ($\sim$ 24k students)
People (birth)	Spike Jonze (born in 1970)	Romelu Lukaku (born in 1993)
Peppers	jalapeño (20k SHU)	Pepper X (3.1M SHU)
Cities	Palermo ($\sim$ 674k inhabitants)	Islamabad ($\sim$ 1.9M inhabitants)
Stadiums	Bolt Arena ($\sim$ 10k capacity)	Kashima Stadium ($\sim$ 40k capacity)
Countries	Botswana ($\sim$ 2.4M inhabitants)	Yemen ($\sim$ 2.8M inhabitants)
Mountains	Mount Scenery (887 m)	Half Dome (2693 m)
Rivers	Mystic River (113 km)	Bega River (256 km)

Table 3: Examples of pairs where the similarity to the keywords is opposite to the numerical value. The first entity is the one with the higher cosine similarity to the keywords, but has a lower numerical value.

Dataset	Positive keywords	Negative keywords
Atoms	heaviest, largest, highest, massive, big	lightest, smallest, lowest, tiny, low
Buildings	tallest, highest, largest, big, tall	shortest, smallest, lowest, tiny, low
Cities	largest, populous, big, crowded, dense	smallest, quiet, tiny, remote, sparse
Countries	largest, populous, big, powerful, dense	smallest, sparse, tiny, quiet, remote
Mountains	highest, tallest, largest, elevated, big	lowest, smallest, shortest, low, tiny
Peppers	hottest, spiciest, pungent, intense, fiery	mildest, bland, cool, weak, low
People (birth)	youngest, recent, modern, newer, late	oldest, ancient, early, historic, vintage
People (social)	popular, famous, followed, liked, viral	unknown, obscure, ignored, unseen, small
Rivers	longest, largest, broadest, deep, big	shortest, smallest, shallow, narrow, tiny
Stadiums	largest, busiest, crowded, massive, big	smallest, quiet, empty, tiny, low
Universities	largest, populous, crowded, big, prestigious	smallest, quiet, tiny, local, low

Table 4: List of positive and negative keywords that are used to capture co-occurrence bias. The positive keywords are terms that are associated with high values of the considered attribute, negative keywords are associated with low values.

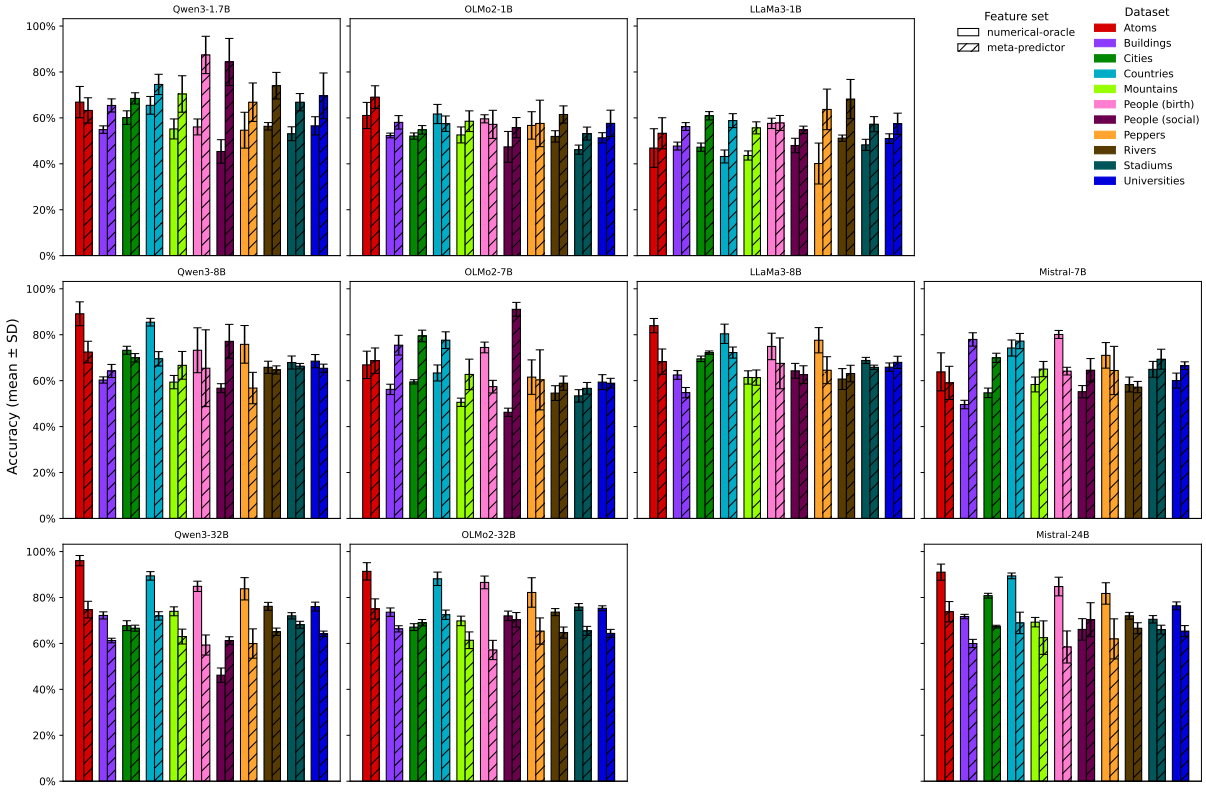


Figure 13: **Absolute prediction accuracy of the two meta-predictors (negative polarity).** Same layout as Figure 12, but for negative polarity prompts.

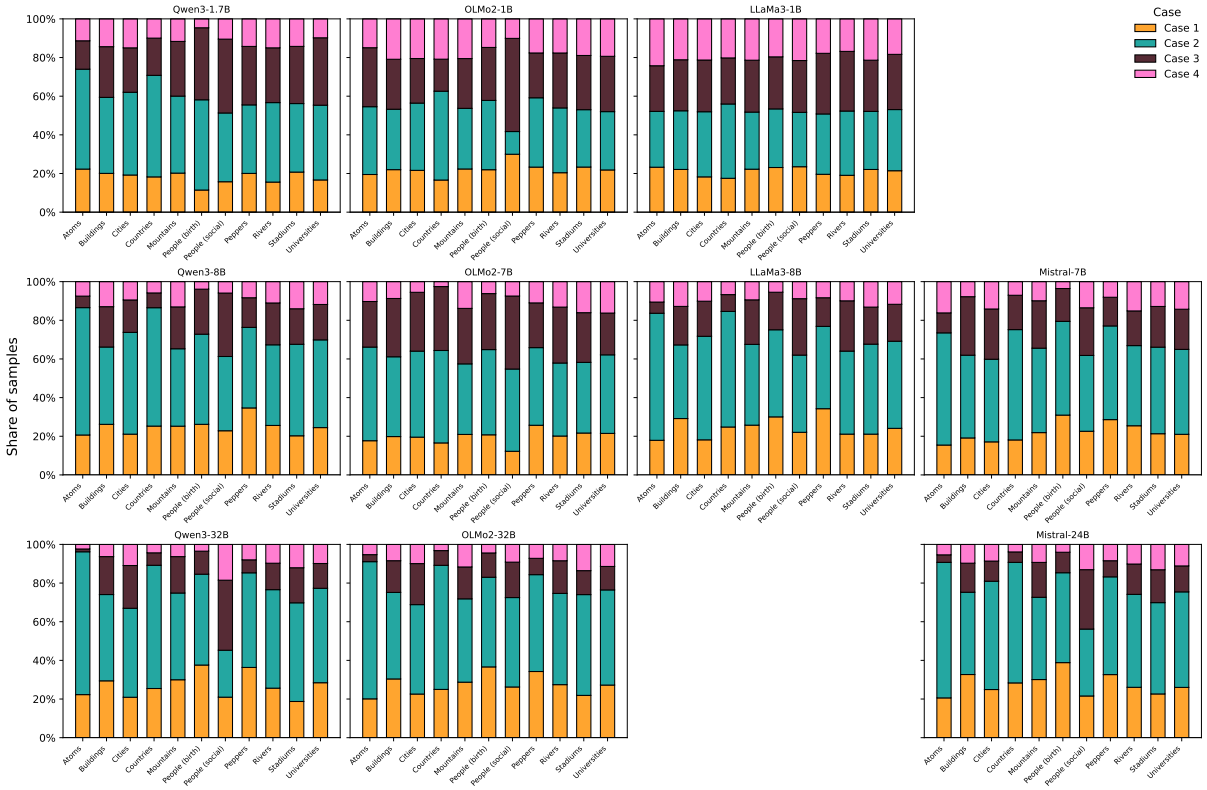


Figure 14: Detailed breakdown of the four cases (from Section 5) per dataset and model. Within every panel each colored, stacked bar corresponds to one dataset. The segments show the proportion of test samples that fall into **Case 1** (numerical and pairwise agree, meta-predictor disagrees), **Case 2** (all three agree), **Case 3** (pairwise aligns with surface-bias meta-predictor, but not with numerical comparison), and **Case 4** (neither numbers nor meta-predictor explain the choice). This figure complements Figure 3 by revealing *which datasets drive each pattern* for every model. Percentages are computed after filtering out samples with unanswerable prompts or missing extractions; bars sum to 100 % within each dataset.

Prompt
Answer with the one name only. Which chemical element has the higher atomic number? entity1 or entity2?
Please state the chemical element with the higher atomic number only. entity1 or entity2?
Answer only with the correct name. Which chemical element has a higher number of protons? entity1 or entity2?
Answer with the one name only. Which chemical element has the lower atomic number? entity1 or entity2?
Please state the chemical element with the lower atomic number only. entity1 or entity2?
Answer only with the correct name. Which chemical element has a lower number of protons? entity1 or entity2?

Table 5: Pairwise prompts for Atoms.

Prompt
Only state the name of the taller building. Which building is taller? entity1 or entity2?
Respond with only the name of the taller building. Which building is taller? entity1 or entity2?
Provide only the name of the taller building. Which building is taller? entity1 or entity2?
Only state the name of the shorter building. Which building is shorter? entity1 or entity2?
Respond with only the name of the shorter building. Which building is shorter? entity1 or entity2?
Provide only the name of the shorter building. Which building is shorter? entity1 or entity2?

Table 6: Pairwise prompts for Buildings.

Prompt
Only state the name of the more populous city. Which city has a larger population? entity1 or entity2?
Respond with only the name of the more populous city. Which city has a larger population? entity1 or entity2?
Provide only the name of the more populous city. Which city has a larger population? entity1 or entity2?
Only state the name of the less populous city. Which city has a smaller population? entity1 or entity2?
Respond with only the name of the less populous city. Which city has a smaller population? entity1 or entity2?
Provide only the name of the less populous city. Which city has a smaller population? entity1 or entity2?

Table 7: Pairwise prompts for Cities.

Prompt
Only state the name of the more populous country. Which country has a larger population? entity1 or entity2?
Respond with only the name of the more populous country. Which country has a larger population? entity1 or entity2?
Provide only the name of the more populous country. Which country is more populous? entity1 or entity2?
Only state the name of the less populous country. Which country has a smaller population? entity1 or entity2?
Respond with only the name of the less populous country. Which country has a smaller population? entity1 or entity2?
Provide only the name of the less populous country. Which country is less populous? entity1 or entity2?

Table 8: Pairwise prompts for Countries.

Prompt
Only state the name of the higher mountain. Which mountain is higher? entity1 or entity2?
Respond with only the name of the mountain that has a greater elevation. Which mountain stands taller? entity1 or entity2?
Provide only the name of the higher mountain. Which mountain has a greater elevation? entity1 or entity2?
Only state the name of the lower mountain. Which mountain is lower? entity1 or entity2?
Respond with only the name of the mountain that has a lesser elevation. Which mountain stands lower? entity1 or entity2?
Provide only the name of the lower mountain. Which mountain has a smaller elevation? entity1 or entity2?

Table 9: Pairwise prompts for Mountains.

Prompt
Only state the name of the hotter pepper. Which pepper has a higher Scoville Heat Unit rating? entity1 or entity2?
Respond with only the name of the hotter pepper. Which pepper is spicier based on Scoville Heat Units? entity1 or entity2?
Provide only the name of the hotter pepper. Which pepper has the greater spiciness level according to the Scoville scale? entity1 or entity2?
Only state the name of the milder pepper. Which pepper has a lower Scoville Heat Unit rating? entity1 or entity2?
Respond with only the name of the milder pepper. Which pepper is less spicy based on Scoville Heat Units? entity1 or entity2?
Provide only the name of the milder pepper. Which pepper has a lower spiciness level according to the Scoville scale? entity1 or entity2?

Table 10: Pairwise prompts for Peppers.

Prompt
Only state the name of the person who was born later. Which person was born later? entity1 or entity2?
Respond with only the name of the younger person. Which person is younger? entity1 or entity2?
Provide only the name of the younger person. Between entity1 and entity2, who is younger?
Only state the name of the person who was born earlier. Which person was born earlier? entity1 or entity2?
Respond with only the name of the older person. Which person is older? entity1 or entity2?
Provide only the name of the older person. Between entity1 and entity2, who is older?

Table 11: Pairwise prompts for People (birth).

Prompt
Only state the name of the person with more social media followers. Which person has a larger social media following? entity1 or entity2?
Respond with only the name of the individual who has more social media followers. Between entity1 and entity2, who has a larger following?
Provide only the name of the person with more social media followers. Who has a larger social media following? entity1 or entity2?
Only state the name of the person with fewer social media followers. Which person has a smaller social media following? entity1 or entity2?
Respond with only the name of the individual who has fewer social media followers. Between entity1 and entity2, who has a smaller following?
Provide only the name of the person with fewer social media followers. Who has a smaller social media following? entity1 or entity2?

Table 12: Pairwise prompts for People (social).

Prompt
Only state the name of the longer river. Which river is longer? entity1 or entity2?
Respond with only the name of the longer river. Which river extends further? entity1 or entity2?
Provide only the name of the river with the longer course. Which of these rivers covers a longer distance? entity1 or entity2?
Only state the name of the shorter river. Which river is shorter? entity1 or entity2?
Respond with only the name of the shorter river. Which river extends a shorter distance? entity1 or entity2?
Provide only the name of the river with the shorter course. Which of these rivers covers a shorter distance? entity1 or entity2?

Table 13: Pairwise prompts for Rivers.

Prompt
Only state the name of the stadium with a larger seating capacity. Which stadium can accommodate more spectators? entity1 or entity2?
Respond with only the name of the stadium that has a greater seating capacity. Which stadium has more seats? entity1 or entity2?
Provide only the name of the stadium with a higher capacity. Which stadium can hold more people? entity1 or entity2?
Only state the name of the stadium with a smaller seating capacity. Which stadium can accommodate fewer spectators? entity1 or entity2?
Respond with only the name of the stadium that has a lower seating capacity. Which stadium has fewer seats? entity1 or entity2?
Provide only the name of the stadium with a smaller capacity. Which stadium can hold fewer people? entity1 or entity2?

Table 14: Pairwise prompts for Stadiums.

Prompt
Only state the name of the university with more enrolled students. Which university has a larger student population? entity1 or entity2?
Respond with only the name of the university that has a greater number of students. Which university has more students enrolled? entity1 or entity2?
Provide only the name of the university with a higher student enrollment. Which university has the largest student body? entity1 or entity2?
Only state the name of the university with fewer enrolled students. Which university has a smaller student population? entity1 or entity2?
Respond with only the name of the university that has a lower number of students. Which university has fewer students enrolled? entity1 or entity2?
Provide only the name of the university with a lower student enrollment. Which university has the smallest student body? entity1 or entity2?

Table 15: Pairwise prompts for Universities.

Prompt
What is the atomic number of entity?
Please state the atomic number of entity.
How many protons does entity have?

Table 16: Numerical prompts for Atoms.

Prompt
What is the length of the entity river in km?
How many kilometers long is the entity river?
Can you provide the length of the entity river in kilometers?

Table 17: Numerical prompts for Rivers.

Prompt
What is the population size of entity, including its metropolitan area?
What is the total population of entity, encompassing its metropolitan region?
Please state the population of entity, including its metropolitan area.

Table 18: Numerical prompts for Cities.

Prompt
What is the height of entity in meters above sea level? What is the altitude of entity expressed in meters above sea level? Please state the height of entity in meters above sea level.

Table 19: Numerical prompts for Mountains.

Prompt
What is the height of the building entity in meters? How tall is the entity building in meters? Please state the height of the entity building measured in meters?

Table 20: Numerical prompts for Buildings.

Prompt
What is the population size of the country entity in 2023? What is the number of inhabitants in entity as of 2023? Please state the population of entity in 2023.

Table 21: Numerical prompts for Countries.

Prompt
Do not list multiple platforms! Only answer with a single number. How many social media followers does entity have across platforms? Provide only the total number of social media followers for entity across all platforms. How many social media followers does entity have in total? Answer with a single number across all platforms.

Table 22: Numerical prompts for People (social).

Prompt
What year was entity born in? In what year was entity born? Please state the year of birth of entity.

Table 23: Numerical prompts for People (birth).

Prompt
How many students are enrolled at entity? What is the total student enrollment at entity? Please state the number of students enrolled at entity.

Table 24: Numerical prompts for Universities.

Prompt
What is the seating capacity of the entity stadium? How many spectators can the entity stadium accommodate? Please state the total number of seats available in the entity stadium.

Table 25: Numerical prompts for Stadiums.

Prompt
What is the Scoville Heat Unit (SHU) rating of the entity pepper? How spicy is the entity pepper in terms of Scoville Heat Units? Please state the Scoville Heat Unit value of the entity pepper.

Table 26: Numerical prompts for Peppers.

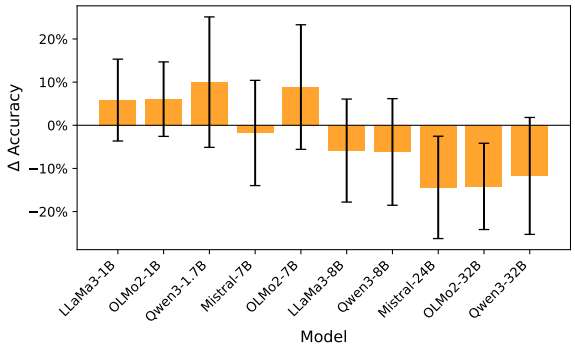


Figure 15: Bias-only meta-predictor vs. numerical baseline. For each language model we report the mean improvement and standard deviation of a logistic-regression meta-predictor that relies solely on three surface cues—entity popularity (QRank), positional advantage, and semantic association with "bigger" keywords—relative to a baseline that follows the model's extracted numerical values. Positive values indicate that the bias-based predictor anticipates the model's pairwise choice more accurately than the model's own numbers, revealing how strongly certain models let positional, popularity and co-occurrence cues override their internal quantitative knowledge.