Knowing the Facts but Choosing the Shortcut: Understanding How Large Language Models Compare Entities

Hans Hergen Lehmann[1], **Jae Hee Lee**[1*], **Steven Schockaert**[2], **Stefan Wermter**[1]
[1]University of Hamburg [2]Cardiff University

Abstract

Large Language Models (LLMs) are increasingly used for knowledge-based reasoning tasks, yet understanding when they rely on genuine knowledge versus superficial heuristics remains challenging. We investigate this question by asking models to compare entities along numerical attributes (e.g., "Which river is longer, the Danube or the Nile?"), which offers clear ground truth for systematic analysis. Despite having sufficient numerical knowledge to answer correctly, LLMs frequently make predictions that contradict this knowledge. We identify three heuristic biases that strongly influence model predictions: entity popularity, mention order, and semantic co-occurrence. For smaller models, a simple logistic regression using only these surface cues predicts model choices more accurately than the model's own numerical predictions, suggesting heuristics largely override principled reasoning. Crucially, we find that larger models (32B parameters) selectively rely on numerical knowledge when it is more reliable, while smaller models (7–8B parameters) show no such discrimination, which explains why larger models outperform smaller ones even when the smaller models possess more accurate knowledge. Chain-of-thought prompting steers all models towards using the numerical features across all model sizes.

1 Introduction

There is an ongoing debate about the extent to which LLMs *understand* language, and the world more generally (Mitchell and Krakauer, 2023; Ray Choudhury et al., 2022). Two contrasting views have been put forward: the *world-model* view holds that LLMs internalize structured knowledge about the world, which they can deploy when prompted (Li et al., 2023; Jin and Rinard, 2024), while the *statistical-parrot* view argues that outputs are largely driven by surface cues (Bender et al., 2021; Saba, 2023). While it seems reasonable to assume that the truth is somewhere in the middle, untangling *when* LLMs rely on surface cues and when they rely on genuine understanding is often hard. In this paper, we therefore focus on a simple controlled setting, within which this question can be studied more systematically, namely the problem of comparing entities along some numerical attribute (e.g., "which country has the highest population, France or Germany?").

This problem setting has several key advantages. First, for the attributes that we consider, there is a unique and objective ground truth. Second, as we can randomly sample entity pairs from a large set of candidates, we can straightforwardly construct test sets that are balanced and orthogonal (i.e., where the presence of one feature is independent of the presence of another feature), which is important for systematic analysis. Moreover, as most entity pairs are not directly compared anywhere on the Web, the impact of pure memorization on the performance of the model should be negligible. Finally, the required world knowledge and associated reasoning process are clear and simple. This means, for instance, that we can straightforwardly identify cases where the LLM has the required knowledge (i.e., knows the correct numerical values), and thus we can distinguish errors due to a lack of knowledge from erroneous reasoning.

We start our analysis by asking: *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 using the predicted attribute values would lead to more accurate results. We also note that the accuracy of pairwise predictions improves as model sizes are increased, but the same is not always true when it comes to the accuracy of predicted numerical attribute values. In

*Corresponding author

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4788–4821

other words, larger models perform better (as could be expected), but this is *not* due to having more accurate knowledge.

To better understand the underlying reasons, we ask our next question: *how susceptible are LLMs to heuristic biases when answering pairwise comparison queries?* (see Section 4). We show that pairwise predictions are strongly biased by three types of surface cues: the *position* of an entity in the prompt, entity *popularity*, and shallow *co-occurrence statistics*. Since these heuristics often align with the ground truth in natural data, we control for this alignment using a balanced, orthogonal subset. We then ask: *to what extent can the pairwise predictions be* explained *by these surface cues?* (see Section 5). We find that the vast majority of model predictions can either be explained by the predicted numerical features or by the above three types of surface cues. This suggests that the considered models sometimes compare numerical values while at other times falling back on surface cues. We find that larger models are more likely to rely on the numerical values when these values are more accurate (i.e., closer to the ground truth numerical values), whereas no such effect was observed for the smallest models (i.e., they rely on shortcuts even if they know the numerical values). This difference explains why the largest models outperform smaller models on pairwise predictions, even though they do not always outperform smaller models in terms of predicting the numerical attribute values.

Finally, we ask: *does chain-of-thought based reasoning help models use their own numerical predictions more faithfully when making pairwise judgments?* (see Section 6). We find that allowing a model to verbalize its reasoning process indeed leads to a more consistent use of numerical attributes, which narrows the performance gap between models of different sizes.

Our findings reveal that the superior performance of larger models stems not from more accurate knowledge, but from their ability to strategically choose when to rely on that knowledge versus when to fall back on heuristics. This suggests that scaling improvements in LLMs may be driven as much by better strategy selection as by knowledge acquisition itself, with important implications for understanding and improving model reliability.

Dataset	Attribute	Entities
Atoms	Atomic number	118
Buildings	Height	1000
Cities	Population	1000
Countries	Population	196
Mountains	Elevation	997
Peppers	Scoville heat unit	45
People	# Followers	999
Rivers	Length	999
Stadiums	Capacity	999
Universities	# Enrolled students	1000

Table 1: Overview of the considered datasets.

2 Experimental Setup

We focus on an entity comparison task, where we prompt an LLM with pairwise 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.[1]

Datasets. To obtain a sufficiently large set of test queries, we collected data on 10 different numerical attributes across diverse entity types from Wikidata[2]. The selected attributes cover a range of domains, such as geography (e.g., river length, population of countries and cities) and science (e.g., atomic numbers), which are listed in Table 1. For each attribute, we sampled up to 1000 entities. For each attribute, we begin by selecting the most popular entities, based on their QRank[3] score. To obtain a set of entity pairs that span a range of difficulty levels, we employ a stratified sampling approach. Specifically, we first sort all entities by their ground-truth attribute values and divide them into two equal-sized bins: a lower-value bin and a higher-value bin. For every entity in our sample, we construct two comparison pairs by randomly selecting one partner from each bin. This ensures our dataset includes both challenging near-tie comparisons and clearer-cut distinctions.

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. The first three templates ask which of the two entities has the highest attribute value. The remaining three templates ask for the entity with the lowest value. Furthermore, for each

[1] Code to reproduce our experiments is available at https://github.com/HeLehm/facts-vs-shortcuts

[2] https://www.wikidata.org

[3] QRank is a popularity ranking for Wikidata entities computed by aggregating page view statistics. See https://qrank.toolforge.org.

template, we prompt the model twice for every entity pair, i.e., once for each 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. We list the full set of prompt templates and describe the strategy for parsing the answers from the model's output in Appendices A and B. We also analyze the sensitivity of our results to the choice of prompt templates in Appendix C. In addition to prompting for pairwise comparisons, we 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). We analyze the error of these numerical predictions 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. To ensure comparability, we remove all samples for which the model did not produce a valid answer, either in the pairwise or numerical setting[4]. 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 (Grattafiori et al., 2024), OLMo2-1B, OLMo2-7B, OLMo2-32B (OLMo et al., 2025), Qwen3-1.7B, Qwen3-8B, Qwen3-32B (Yang et al., 2025a), Mistral-7B (Jiang et al., 2023) and Mistral-24B. Full details on these models can be found in Appendix E.

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 F. A number of important findings can be observed. (i) Numerical accuracy is consistently and substantially higher than pairwise accuracy, showing that models often make mistakes even when relying on their knowledge of the numerical attributes would produce the correct answer. (ii) For the smallest models, pairwise accuracy is barely above random chance. (iii) Pairwise accuracy increases with model size. (iv) For numerical accuracy, on the other hand, Mistral-7B and Llama3-8B both outperform much bigger models. For these models, the underperformance in terms of pairwise accuracy 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. Overall, the results suggest that LLMs rely on shortcuts when making pairwise predictions, which we further analyze in the next section.

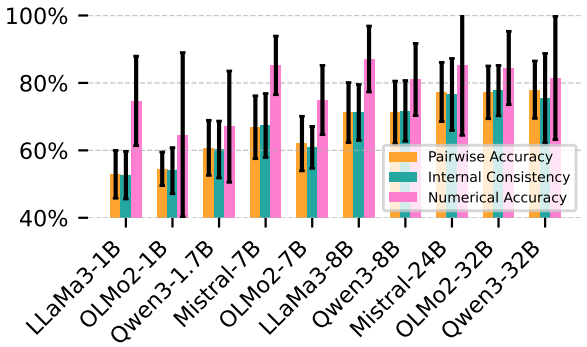

Figure 1: Overall performance in terms of pairwise accuracy, internal consistency, and numerical accuracy (mean and standard deviation)

4 How Susceptible Are Pairwise Predictions to Biases?

The previous section showed that LLMs often ignore their own numerical knowledge when ranking entities. A natural follow-up question is *what, if not the numbers, drives the final choice?* To investigate, we identify three biases and measure their impact.

Heuristic Cues. A first 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 *popularity bias*, we estimate the popularity of each Wikidata entity using its QRank score. We then test whether LLM responses are more accurate when the entity with the highest numerical value is also the most popular one.

[4] Appendix B provides a detailed description of how the parsing was done, and an analysis of how often the model did not yield a valid answer.

Second, LLMs have been found to suffer from *position bias*, favoring 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.

Finally, 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.[5] 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. Full details of how the scores are obtained can be found in Appendix G.

Experimental Setup. Given the three potential shortcuts, we must design our experiments with care to avoid conflating the model's reliance on surface cues with genuine knowledge of numerical facts. As an example, popularity is often a proxy for magnitude: we remember Mount Everest precisely because it is the tallest peak, and a celebrity's follower count is itself a direct measure of their popularity across social-media platforms. Our data confirm this intuition (see Appendix H for details). Such correlations mean that surface cues can look like genuine knowledge. Furthermore, we must also consider the possibility that these shortcuts correlate with each other. Mount Everest likely co-occurs with the adjective "tallest" in the training data frequently, and it is also more popular than most other mountains.

Therefore, we need a *balanced* and *orthogonal* design that isolates each cue from the others. *Balance* in this case means that positive and negative cases occur equally frequently (e.g., popularity aligns with the ground truth in exactly half of the comparisons and misaligns in the other half). This ensures that none of the considered heuristic cues has an advantage simply because it happens to be more frequent in the data. *Orthogonality* means that the features vary independently of each other, i.e., all possible combinations of the values of the features appear equally often. Orthogonality helps mitigate aggregation artifacts such as Simpson's paradox, where the apparent effect of a cue might actually be driven by another, correlated factor. If popularity and co-occurrence are correlated, for instance, then the effect we assign to popularity might actually reflect co-occurrence effects, and vice versa. We stress that this protection applies only to the *observed* cues; unmeasured confounders could still induce bias (see Appendix I.1).

We construct a *Balanced-Orthogonal Subset* (BOS) as follows. We assign each entity pair with four binary features, which we will refer to as P (popularity), O (order), C (co-occurrence), and I (internal knowledge). We define $P = 1$ if the entity with the higher ground-truth value is also the more popular one (and $P = 0$ otherwise); $O = 1$ if that larger entity appears first in the prompt; $C = 1$ if the entity whose ConceptNet embedding lies closer to the "large" direction is indeed the larger one; $I = 1$ if the ranking implied by the model's extracted numbers matches the ground truth. BOS is constructed by taking the minority count from each of the $2^4 = 16$ (P, O, C, I)-cells within each prompt template and sampling that many instances from every other cell. Information about the size of these subsets can be found in Appendix I.2. In BOS, each feature can be toggled independently, assuring that all other features are held constant at a rate of 50% true and 50% false.

To measure the impact of each feature, we adopt the risk ratio (RR) (Rothman et al., 2008). Let Y be the ground-truth label and $\hat{Y}$ the model's prediction. $F = 1$ indicates that a feature is present and $F = 0$ means that is not. The RR is defined as follows:

$$\mathrm{RR}_F = \frac{\Pr\big(Y = \hat{Y} \mid F = 1\big)}{\Pr\big(Y = \hat{Y} \mid F = 0\big)}$$

If $\mathrm{RR}_F = 1$, it means that there is no change in accuracy. If $\mathrm{RR}_F > 1$, it means the accuracy is higher when the feature aligns with the ground truth ($F = 1$), and conversely for $\mathrm{RR}_F < 1$. Since RRs are within-model quantities and depend on the model's baseline accuracy when a cue is absent, bar heights should not be compared across models (see Appendix I.4 for details). For the order cue, we report $\max(\mathrm{RR}_O, 1/\mathrm{RR}_O)$, which captures the effect size irrespective of direction.

[5] Word embeddings can be seen as a low-rank approximation of co-occurrence statistics and the embedding similarities serve as a convenient proxy for the raw co-occurrence statistics.

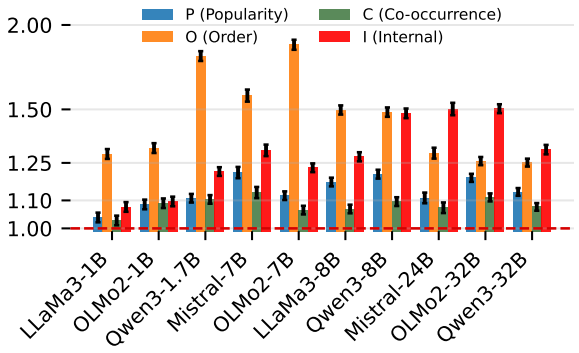

Figure 2: Risk ratios (log scale) for each feature, with 95% confidence intervals from bootstrapping over prompt templates and datasets. See Figure 14 in Appendix I.3 for per-dataset, per-model results.

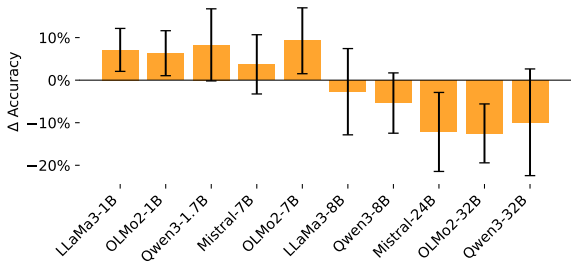

Figure 3: For each language model, we report the mean improvement and standard deviation of a logistic-regression meta-predictor that relies solely on three surface cues, relative to a strategy that follows the model's extracted numerical values. Positive values indicate that the bias-based predictor anticipates the model's pairwise choice more accurately than the model's own numbers.

Results. Figure 2 shows the risk ratios for the different models. Order (O) is the dominant shortcut for all models. For most smaller and mid-sized models, O also displays a higher risk ratio than I, meaning that the order in which entities are presented has a stronger impact on model predictions than the model's knowledge of the numerical attributes. For example, Qwen3-1.7B reaches an RR_O of almost 2, meaning the model is roughly twice as likely to answer correctly when the larger entity appears in its preferred position. Popularity (P) effects are smaller ($\text{RR} \approx 1.1 - 1.25$) but are consistently present across model sizes, reflecting a persistent "fame implies bigger" heuristic. Co-occurrence (C) shows the weakest effect, typically near $\text{RR} \approx 1.05 - 1.15$, but also remains consistently present across all model sizes. Internal-ground-truth alignment (I) grows in importance with scale, reaching RRs around 1.5 for the largest models. This indicates that when their own extracted numbers agree with reality, they are more likely to answer correctly w.r.t. the ground truth compared to smaller models. In summary, we find that while all models are susceptible to biases, larger models tend to rely more on their internal knowledge, whereas smaller models are more influenced by heuristic cues.

5 Can LLM Predictions Be Explained?

The previous analysis isolated the effect of each cue in turn. Yet in practice, multiple cues may counteract or reinforce each other, raising the question: *can we build a simple model that predicts the LLM's choice better than its own numbers, purely from such surface features?* This section formalizes this idea via a simple *meta-predictor*, a logistic regression model trained to predict whether the LLM will select the first or second entity in a pairwise comparison. The meta-predictor is provided with two binary features, namely whether the first entity is more popular than the other, and whether the first entity is more associated with magnitude descriptors (via cosine similarity in ConceptNet embeddings, see Section 4). The meta-predictor is trained to predict whether the LLM will choose the first or the second entity, and can thus also take position bias into account. The meta-predictor is trained separately for each model, prompt template, and numerical attribute using 5-fold cross-validation.

In Figure 3, we contrast the performance of the meta-predictor with a strategy where a model's pairwise predictions always follow their own numerical predictions (absolute accuracies can be found in Appendix J). For the smallest models, *surface cues predict the model's pairwise choice better than the extracted numerical values*. For larger models, *the extracted numerical values become more predictive (although the meta-predictor remains competitive)*.

5.1 Fine-grained Analysis

To better understand these dynamics, we classify each test sample into the following cases:

Case 1 Pairwise and numerical predictions agree, meta-predictor disagrees ⇒ numerical reasoning

Case 2 All three predictions agree ⇒ numerical reasoning or superficial cues

Case 3 Pairwise and meta-predictor agree, numerical prediction disagrees ⇒ superficial cues

Case 4 Pairwise prediction disagrees with both

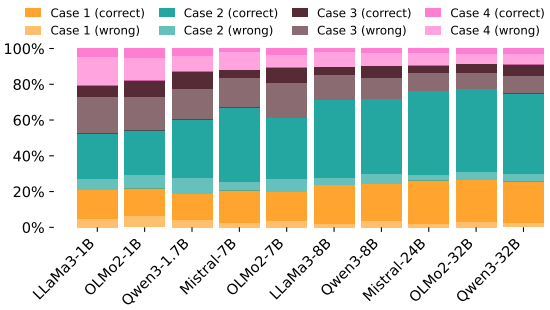

Figure 4: Proportion of the four cases for different models, aggregated over all datasets. For each of the four cases, the upper (darker) segment marks correct predictions, and the lower (lighter) segment marks errors with respect to the ground truth.

numerical and meta-predictor ⇒ unexplained / noise

Figure 4 shows the distribution of these cases for the different models, where each case is further split into two sub-cases, depending on whether the prediction matched the ground truth or not.[6]

We can make several key observations. (i) First, most pairwise predictions can be explained by numerical reasoning, surface biases, or both (Cases 1–3), especially in larger models. Case 4, where neither the numerical prediction nor the meta-predictor aligns with the pairwise output, is rare, suggesting that models rely only minimally on unmodeled heuristics or random behavior. (ii) Consistent with our earlier findings from Figure 3, we observe a clear difference in how smaller and larger models make decisions. Smaller models such as LLaMa3-1B and OLMo2-1B are more frequently guided by surface-level biases (Case 3) than by their own numerical predictions (Case 1), whereas larger models show the opposite trend, relying more consistently on numerical information. (iii) The case breakdown clarifies how different types of prediction behavior relate to correctness. In Case 2, where all three predictors agree, models are almost always correct. (iv) Case 1 is also associated with high accuracy, whereas Case 3 shows the opposite pattern. This supports the view that the meta-predictor captures surface cues that can lead the model away from correct decisions when they conflict with its numerical knowledge. (v) Finally, this analysis helps explain a counterintuitive pattern noted earlier: some mid-sized models, including Mistral-7B and LLaMA3-8B, achieve relatively high numerical accuracy but underperform in pairwise comparisons (cf. Figure 1). The case distribution reveals that these models frequently follow surface heuristics that go against their numerical knowledge (Case 3).

5.2 When Do Models Rely on Surface Cues?

All models sometimes show signs of being a world model by following their own extracted numbers (Case 1), and sometimes appear to be statistical parrots, following surface cues instead (Case 3). We want to understand *what differentiates these two modes of behavior*. For this analysis, we exclude test examples that fall in Cases 2 and 4, as they might confound interpretation.[7] For each entity, we consider five metrics that we hypothesize might influence whether a model relies on numerical reasoning (Case 1) or superficial cues (Case 3): (i) **Ground-truth value** (GT) for the considered numerical attribute; (ii) **Model-extracted value** (NumEx), i.e., the model's prediction for the considered numerical attribute; (iii) **Symmetric Mean Absolute Percentage Error** of the extracted numbers relative to ground truth (SMAPE, see Appendix D); (iv) **Coefficient of Variation** of the extracted numbers across prompt templates (CV, see Appendix C.1), which offers a proxy for the model's confidence in the numerical value; (v) **Popularity** (QRank) of the entity. Each of the metrics is used to compute two statistics for entity pairs: the **mean** log-value of the metric across both entities and the **difference** between their log-values. We use the logarithm of the numerical values, rather than the values themselves, to focus on their order of magnitude. We include both the mean and the difference as they capture different effects: The mean allows us to test, for instance, whether larger values are more common in Case 1 or Case 3 (e.g., whether entity popularity affects how the model makes a prediction). The difference allows us to test, for instance, whether a clearer gap between the model-extracted values is predictive.

For each feature x, we want to know whether it tends to be larger in Case 1 or larger in Case 3, and by how much. Cohen's d (Cohen, 1988) answers exactly that: it is the difference in group means measured in units of a typical within-group standard deviation (SD), making it unitless and comparable across features.[8] In our setting, a positive value

[6] A breakdown per dataset can be found in Appendix K.1.

[7] A detailed analysis of Case 2 can be found in Appendix K.2.

[8] Details on how this statistic is computed can be found in Appendix L.

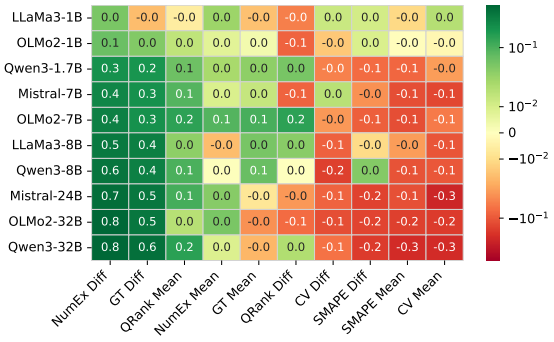

Figure 5: Signed Cohen's d for contrasting Case 1 vs. Case 3. Green means the feature value tends to be larger in Case 1 (numerical knowledge), red means that the value tends to be larger in Case 3 (surface cues). Cell labels show the mean effect.

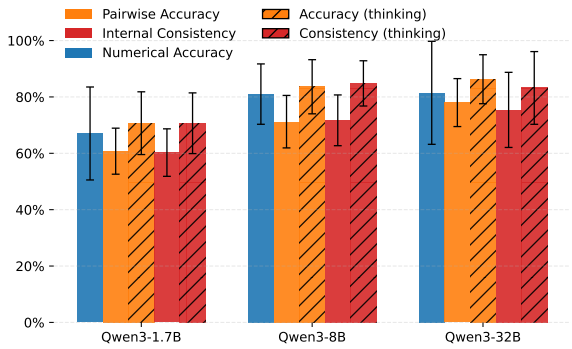

Figure 6: Comparison of the performance with and without thinking, in terms of pairwise accuracy and internal consistency. The figure also reports numerical accuracy (without thinking). Error bars show the standard deviation over datasets.

d indicates that feature x is, on average, larger in Case 1, while a negative d indicates it is larger in Case 3. The magnitude $|d|$ says how strongly the groups differ, measured in pooled-SD units. For instance, $d = 0.5$ means the average Case 1 value is half a standard deviation larger than the average Case 3 value. Note that this analysis is descriptive and does not identify causal effects.

Figure 5 shows the d statistic for each of the 10 features, per model, aggregated over all the datasets.[9] Several clear regularities can be observed. First, the NumEx difference and the GT difference are both larger for Case 1 than for Case 3. Large differences intuitively mean that relying on the numerical attributes is safer, as even highly approximate numerical knowledge is sufficient for making reliable pairwise predictions in such cases. Interestingly, this effect is much more pronounced for the larger models. For instance the d-statistic for NumEx-diff rises from ≈ 0.04 at 1B to ≈ 0.79 at 32B. Conversely, SMAPE and CV means tend to be much higher for Case 3 than for Case 1. High values for these features indicate that the model's knowledge of the required numerical attributes is noisy, which makes relying on them riskier. Using alternative heuristics may thus be a rational choice in such cases. Again, we see that this effect is most pronounced for the largest models. Overall, our analysis thus supports the view that larger models make more principled choices when deciding between the two strategies (i.e., relying on numerical attributes vs. heuristic cues).

[9] A detailed breakdown can be found in Appendix K.3.

6 How does CoT Affect Predictions?

In this section, we investigate whether prompting LLMs to "think" before answering improves the internal consistency of their pairwise numerical comparisons. We ask: *does explicit reasoning help models use their own numerical predictions more faithfully when making pairwise judgments?* To address this question, we focus on the models from the Qwen3 series, which have been fine-tuned specifically for reasoning tasks using chain-of-thought supervision. During inference, we permit the model to generate up to 1024 new tokens, ensuring that the "thinking" prompts have a large enough budget to verbalize intermediate steps. We use the same prompt templates as before, but append a start-of-thought token at the end.

Figure 6 shows that prompting the Qwen3 models to "think" before responding yields a clear performance lift. On average, pairwise accuracy rises by 5–9 percentage points, and internal consistency rises *by a nearly identical margin*. Intuitively, we might expect a chain-of-thought prompt to first retrieve both numbers and then compare them mechanically, pushing internal consistency to nearly 100%. We see that this is not the case. To better understand these results, we manually inspected a sample of chain-of-thought traces produced by the Qwen3 models. The analysis reveals several recurring tendencies. In many cases, the model appears to make up its mind before retrieving any numbers, then generates numerical statements that merely serve to justify the chosen answer, reinforcing the view that chain-of-thought traces should not be treated as faithful evidence about whether

the model relied on numerical reasoning or heuristics (Xie et al., 2024; Lyu et al., 2023; Lanham et al., 2023; Paul et al., 2024; Chen et al., 2025). In other traces, the retrieved numbers differ from those obtained when the model is asked for the values directly, sometimes being closer to the ground truth but frequently inaccurate or inconsistent. We did notice more generally that different prompts sometimes lead to different numerical values.[10] In other cases, the reasoning step is skipped altogether, with the model producing a direct answer despite the thinking prompt. When reasoning occurs, it sometimes relies on heuristic arguments. Finally, for some samples, the model used more than 1024 tokens in the thought process, therefore not yielding an answer.

Together, these observations explain why chain-of-thought prompting improves pairwise accuracy and internal consistency without eliminating inconsistency. The reasoning traces often reflect rationalization rather than deliberate computation, and the modest gains likely stem from occasional improvements in number retrieval or from semantically plausible heuristics that happen to yield the correct answer. A more systematic analysis of these reasoning patterns, and their relation to numerical faithfulness, is left for future work (see Appendix M for representative examples).

7 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. 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. Along similar lines, He et al. (2025) report systematic asymmetries in recognizing logically equivalent relational facts, which they trace back to entity-frequency imbalances in the pre-training data (high-frequency subjects paired with low-frequency objects are recognized more reliably than the inverse). Allen-Zhu and Li (2024) find that LLMs sometimes memorize knowledge without being capable of reliably exploiting it for answering questions. The compositionality gap, where models can answer individual sub-questions but fail to compose them into correct multi-hop answers, has been documented by Press et al. (2023), who found that scaling improves single-hop performance faster than multi-hop performance. This parallels our finding that models possess numerical knowledge but fail to reliably apply it in pairwise comparisons. The problem of ranking entities with LLMs was studied by Kumar et al. (2024), but their focus was on designing fine-tuning strategies. Regarding calibration, i.e., how well models know what they don't know, Kadavath et al. (2022) demonstrated that larger models are more aware of their own knowledge boundaries, with scale playing a crucial role. Our work extends these insights by showing that larger models not only know what they know, but can also strategically choose when to rely on that knowledge versus when to use heuristics.

8 Conclusion

We have analyzed how LLMs behave when asked to compare entities along some numerical attribute. Intuitively, an LLM could simply extract the attribute values for the two given entities and compare these. However, we found their actual performance dramatically underperforms such a strategy. Our experiments suggest that LLMs switch between two strategies: a principled approach based on their knowledge of the numerical attributes and a heuristic approach based on surface cues, such as entity popularity, co-occurrence statistics, and the ordering of the entities in the prompt. Furthermore, we found that larger models tend to choose between these strategies in a more principled way, being more likely to rely on numerical attributes when their numerical knowledge is more (likely to be) reliable. Finally, in our experiments with CoT-based reasoning models, we found that predictions align

[10] An in-depth analysis of this can be found in Appendix C.1.

better, but still not perfectly, with the models' numerical knowledge.

Our findings offer a nuanced perspective on the ongoing debate between the world-model and statistical-parrot views of LLMs. Rather than supporting either extreme, our results suggest that LLMs operate in a hybrid manner: they possess genuine world knowledge (numerical attributes) but do not always deploy it consistently. Importantly, the ability to strategically select between knowledge-driven reasoning and heuristic shortcuts emerges with scale, suggesting that larger models are developing a form of meta-cognitive capability. Our work thus provides a first step towards a more sophisticated understanding of LLM behavior: one where the question is not whether models understand or merely parrot, but rather *when* and *how* they choose between different reasoning strategies.

Acknowledgments

Jae Hee Lee was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 551629603. Steven Schockaert was supported by EPSRC grant EP/W003309/1.

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 (chain-of-thought) 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 ranking tasks. Lastly, we study pairwise comparisons for numerical attributes in a controlled setting, which yields objective ground truth and orthogonal cue manipulation, but our findings may not fully transfer to open-domain question answering, non-numerical attributes, multi-entity ranking, or multi-hop reasoning.

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.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of Language Models: Part 3.2, Knowledge Manipulation. *arXiv preprint*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

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.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Melanie Brucks and Olivier Toubia. 2025. Prompt architecture induces methodological artifacts in large language models. *PLOS ONE*, 20(4):e0319159. Publisher: Public Library of Science.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint*.

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

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*.

Yuan He, Bailan He, Zifeng Ding, Alisia Maria Lupidi, Yuqicheng Zhu, Shuo Chen, Caiqi Zhang, Jiaoyan Chen, Yunpu Ma, Volker Tresp, and Ian Horrocks. 2025. Supposedly equivalent facts that aren't? entity frequency in pre-training induces asymmetry in LLMs. In *Second Conference on Language Modeling*.

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. 2023. Mistral 7b. *arXiv preprint*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438. Place: Cambridge, MA Publisher: MIT Press.

Charles Jin and Martin C. Rinard. 2024. Emergent representations of program semantics in language models trained on programs. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint*.

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.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint*.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

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. *arXiv preprint*.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

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. *arXiv preprint*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine Reading, Fast and Slow: When Do Models "Understand" Language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. 2008. *Modern Epidemiology*. Lippincott Williams & Wilkins. Google-Books-ID: Z3vjT9ALxHUC.

Walid S. Saba. 2023. Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In *Conceptual Modeling - 42nd International Conference, ER 2023, Lisbon, Portugal, November 6-9, 2023, Proceedings*, volume 14320 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024.

NumeroLogic: Number encoding for enhanced LLMs' numerical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 206–212, Miami, Florida, USA. 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.

Qwen Team. Qwen3 [online]. 2025.

Tyler J. VanderWeele and Peng Ding. 2017. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4):268–274.

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.

Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. 2024. Calibrating reasoning in language models with internal consistency. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating non-transitivity in llm-as-a-judge. *arXiv preprint*.

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. 2025a. Qwen3 technical report. *arXiv preprint*.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2025b. Number cookbook: Number understanding of language models and how to improve it. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

A Prompts

Table 2 and Table 3 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. As a system prompt, we used the following:

> *You are a chatbot to help with general knowledge questions. You answer as short and concise as possible. Meaning, you should not provide more information than what is asked for. If you are asked to compare two entities answer with the name of the correct one only.*

B Response Parsings

To convert free-form model outputs into structured predictions, we use two deterministic regex-based parsing pipelines, one for numerical-value prompts and one for pairwise-comparison prompts. These are explained in detail in the following paragraphs.

B.1 Numerical Prompts

The following parsing procedure is applied. For attributes where a physical unit is expected (e.g., meters, kilometers), we begin by extracting all numeric values from the response and, if applicable, convert them into the unit requested by the prompt. For instance, if the model returns a distance in kilometers when meters were asked for, we apply the appropriate conversion factor. If the unit is missing or ambiguous, we assume the number is in the expected unit. We also normalize magnitude modifiers such as "*k*", "*million*", etc. (e.g., $60\text{k} \rightarrow 60,000$). If multiple valid numbers are found, we select the one closest to the ground truth value, under the assumption that the model may have approximated the correct answer. We do this because the models sometimes responded with the year when its knowledge cutoff was, for example: "*In 2023, XY had a population of...*" or named multiple units. If exactly one number is found, we return it. If no number can be extracted, the response is marked as `unknown`.

B.2 Pairwise Prompts

For the pairwise paradigm, we need to determine which of the two entities was chosen by the model. To do this, we follow a multistep procedure. First, we check whether exactly one of the entity names appears verbatim in the output. If so, we treat that entity as the model's prediction. If both names appear, we search for indicative phrasings that suggest a directional comparison, i.e., statements that clearly identify one entity as having a higher or lower value. If this fails, we check whether the response contains an unambiguous substring of one of the entity names. This accounts for answers like "China" when the full name is "People's Republic of China". If still unresolved, we apply fuzzy matching[11] to detect typographical or lexical variants. If no reliable match can be made by any of these steps, the response is flagged as `unknown`.

B.3 Parsing and Model Failures

Across all settings, parsing or model failures are rare. Table 4 reports mean fail rates across datasets for each model, along with the minimum and maximum rates observed across datasets.

C Prompt Sensitivity

In this section we analyze how sensitive the evaluated models are to the wording of the prompt. For a perfect world model, logically equivalent phrasings would yield identical behavior. But other studies have revealed that LLMs are sensitive to the precise wording of a prompt (Brucks and Toubia, 2025; Jiang et al., 2020; Bouraoui et al., 2020). Even minor re-phrasings can nudge the model toward a different number, flip a pairwise preference, or introduce noise. To investigate this fragility we answer three questions about prompt sensitivity:

- When asking the model for a number, do three paraphrases yield the same number and if not, how different are they?
- When asking the model to compare two entities, does the accuracy change when asking for smaller rather than the larger one?
- In pairwise comparison, within the same polarity, how often does the model agree on the same answer, when using different prompt variations?

Taken together these three dimensions provide a comprehensive picture of prompt sensitivity from different angles. The following sections unpack each dimension in turn.

[11] https://github.com/seatgeek/thefuzz

Entity Type	Prompt
Atoms	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}?
Buildings	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}?
Cities	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}?
Countries	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}?
Mountains	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}?
Peppers	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}?
People (social)	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}?
Rivers	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}?
Stadiums	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}?
Universities	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 2: Pairwise prompts for all entity types.

Entity Type	Prompt
Atoms	What is the atomic number of {entity}? Please state the atomic number of {entity}. How many protons does {entity} have?
Buildings	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?
Cities	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.
Countries	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.
Mountains	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.
Peppers	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.
People (social)	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.
Rivers	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?
Stadiums	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.
Universities	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 3: Numerical prompts for all entity types.

Model	Numerical Fail Rate	Pairwise Fail Rate	Combined Fail Rate
LLaMa3-1B	3.7 (0.0–19.2)	2.1 (0.0–15.8)	5.8 (0.0–19.3)
OLMo2-1B	0.0 (0.0–0.0)	0.2 (0.0–0.4)	0.2 (0.0–0.4)
Qwen3-1.7B	0.4 (0.0–2.8)	0.0 (0.0–0.1)	0.4 (0.0–2.8)
Mistral-7B	0.8 (0.0–4.8)	5.5 (1.0–14.4)	6.2 (1.7–15.2)
OLMo2-7B	0.0 (0.0–0.3)	0.3 (0.0–2.2)	0.4 (0.0–2.2)
LLaMa3-8B	11.4 (0.0–47.6)	0.0 (0.0–0.1)	11.4 (0.0–47.6)
Qwen3-8B	0.4 (0.0–2.0)	0.6 (0.0–5.4)	0.9 (0.0–5.4)
Mistral-24B	0.1 (0.0–0.3)	0.1 (0.0–0.3)	0.1 (0.0–0.5)
OLMo2-32B	2.0 (0.0–18.0)	0.0 (0.0–0.0)	2.0 (0.0–18.0)
Qwen3-32B	0.2 (0.0–1.6)	0.0 (0.0–0.1)	0.2 (0.0–1.6)
Qwen3-1.7B (thinking)	0.4 (0.0–2.8)	1.1 (0.2–2.2)	1.4 (0.2–3.1)
Qwen3-8B (thinking)	0.4 (0.0–2.0)	1.6 (0.1–4.0)	1.9 (0.1–4.7)
Qwen3-32B (thinking)	0.2 (0.0–1.6)	1.4 (0.1–4.0)	1.6 (0.1–4.0)

Table 4: Parsing fail rates aggregated across datasets. Each entry reports the mean failure rate (in %) across datasets for a given model, with the minimum and maximum rates across datasets between parentheses. “Numerical” refers to failures of the numeric-value parsing pipeline, “Pairwise” to failures of the entity-choice parsing pipeline, and “Combined” to failures of either pipeline.

C.1 Numerical Extraction

For each question we ask the model, we use three differently worded numerical-extraction prompts and measure how tightly the answers cluster. We would expect these to be identical for a perfect world-model. Concretely, we take all the extracted values for an entity based on the three different prompts templates and compute the coefficient of variation

$$\mathrm{CV} = \sigma/\mu, \qquad \sigma = \text{st. dev.}, \ \mu = \text{mean}.$$

CV measures the spread of the three answers around their mean. A low CV means that the three answers are close to one another or even identical (CV=0). A CV of 0.05 means that the standard deviation is 5% of the mean, which is still a tight cluster. A CV of 0.4, by contrast, indicates a wide spread of answers. The models response differs by about half of the average magnitude of the answers. This would for example be the case if the model outputs 500, 1000 and 1500 for the three different prompts. From a statistical parrot perspective, we would expect a high CV, as the model is likely to pick up on different keywords in the prompt and produce different numbers. To handle cases where the model fails to output any number, we fill the missing values with 0, thereby penalizing non-numeric answers, when other prompts did yield a number. This is calculated for every entity of every dataset for every model. The results are plotted in Figure 7 in violin plots with median lines. Smaller models tend to have higher CV values, as indicated by the median and the shape of the violin. This stands for a greater sensitivity to prompt wording and a higher likelihood of producing varied responses. Generally this trend decreases with model size.

C.2 Pairwise Ranking

We now turn from the stability of extracted numbers to the robustness of direct pairwise decisions under changes in wording, distinguishing between inter- and intra-polarity effects. Intra-polarity refers to the answers of the model when the polarity of the prompt is held constant, e.g., when asking "Which city is larger?" in three different ways. Inter-polarity, by contrast, refers to the comparison of answers when the polarity is flipped, e.g., comparing "Which city is larger?" to "Which city is smaller?". In principle, a perfect world model would be entirely insensitive to prompt re-phrasings and yield the same pairwise preference regardless of the wording of the prompt template or whether the question asks for the larger or smaller entity. Note that for the experiments in the main paper we consider all prompt templates and entity orderings, i.e., if we had a single pair and therefore twelve different prompts (six templates, two orderings), we would have twelve pairwise decisions to consider. If the model answered 6 of them correctly, while failing to do so in the other 6, we would say that the model has a 50% accuracy.

C.2.1 Inter-polarity

A stable model should be robust to changes in the semantic polarity of the prompt, i.e., if the model can correctly identify the larger entity, it should also be able to identify the smaller one. To test this, we compare the accuracy of the three *larger-than* templates to the accuracy of the three *smaller-than* templates. Flipping the semantic polarity of every template, e.g. from *"Which city is larger…"* to *"Which city is smaller…"*, yields the accuracy difference $\Delta\text{Acc} = \text{Acc}_{larger} - \text{Acc}_{smaller}$, shown in Figure 8. Across the board models prefer the positive polarity: accuracies tend to drop when the question asks for the smaller entity. The gap shrinks with scale. For Qwen3-32B the difference is negligible.

C.2.2 Intra-polarity

Knowing that there is a polarity bias, we now turn to the question of how often the different templates agree with one another when the polarity is held constant. Keeping polarity fixed, we count how many distinct answers the three prompt templates with slightly different wording the model yields. As we use three different prompt templates per polarity, there are three distinct values that can arise form this: full agreement, 2-vs-1 or complete disagreement. The results are visualized in Figure 9 per model and dataset. Complete disagreement (three different answers) is extremely rare ($< 1\%$ for all models). Most entity pairs fall in the green "full agreement" slice. The orange 2-vs-1 splits account for the rest. Generally, the bigger the model, the less it disagrees with itself based on prompt wording.

C.3 Summary

Across the three tested dimensions of prompt sensitivity, a pattern emerges. Larger models show greater stability. They produce tightly clustered numerical estimates, maintain agreement across

prompt templates, and display little polarity bias. Smaller models, in contrast, are substantially more affected by re-phrasings, with higher variability in numbers, lower consensus across prompt templates, and stronger polarity asymmetries. These findings suggest that robustness to prompt wording strengthens with scale, though sensitivities remain across all model sizes.

D Error Analysis of the Extracted Numerical Attributes

The main text is concerned with how well the numerical attributes extracted from the models can be used to rank entities. While this is the main goal of our study, it is also interesting to see how accurate the models are in predicting the actual numbers. As our analysis spans many numerical attributes of vastly different scales (e.g., lengths of rivers in the order of thousands of kilometers vs. number of social media followers in the order of millions), we need a scale-independent error metric to compare performance across datasets. Therefore, we use the Symmetric Mean Absolute Percentage Error (SMAPE),

$$\text{SMAPE}(y, \hat{y}) = \frac{|\hat{y} - y|}{(|\hat{y}| + |y|)/2},$$

where y is the ground-truth value and $\hat{y}$ the model's prediction. It is scale-independent and bounded in the interval $[0, 2]$, making values directly comparable across quantities that span several orders of magnitude. A value of 0 means perfect prediction, while 2.0 represents the worst case.

Figure 10 shows the mean SMAPE for every model and dataset together with the standard deviation. The results reveal a mild size trend: *larger models tend to achieve lower SMAPE scores*, meaning their extracted numbers are generally closer to the ground truth. However, even the largest models still make significant errors. A closer look into the per dataset results reveals that this can largely be attributed to a few datasets. All models perform best on the `atoms` dataset and worst on `people (social)`, with the remaining datasets showing similar values across architectures. The `people (social)` dataset requires models to estimate the number of followers across various social media platforms, which can fluctuate significantly over time and is inherently difficult to approximate. For datasets that are more stable, such as river lengths or mountain heights, the SMAPE is generally lower, while still not perfect even for the largest models. SMAPE tells us what the model knows about the numbers, but not how useful this knowledge is in ranking tasks nor whether the model uses it. These questions are explored in the main paper (see Section 3).

E 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 5.

F Detailed Accuracy Analysis

As there are some differences in the results between prompts with positive and negative polarity, we report results for these types of prompts separately. Figures 11 and 12 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.

G Co-occurrence Details

Table 6 provides qualitative examples in which cosine similarity to attribute-related keywords (e.g., "larger," "bigger," "more") suggests the wrong ranking, which would lead models that rely on co-occurrence bias to make an incorrect prediction. The list of positive and negative keywords used to construct the "bigger-smaller" axis is shown in Table 7.

H Bias Alignment with Ground Truth

Table 8 shows that for most datasets, popularity points to the larger item substantially more than half of the time. One outlier is the Atoms dataset, where popularity corresponds to the largest items only 24% of the time. Lighter elements tend to be more common (such as hydrogen, carbon, or oxygen) and thus more popular. We also find that the co-occurrence cue lines up with the ground truth more than half of the time. Mentions of entities such as the Mount Everest or the Nile River tend to co-occur with adjectives expressing their magni-

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

Table 5: Information about the models used in this paper.

tude, because they are the largest or longest in their respective categories.

I BOS details

I.1 E-Values

While the BOS makes the measured cues independent of each other, it does not remove bias from *unmeasured* confounders. To assess how robust our results are to such confounding, we compute the E-value (VanderWeele and Ding, 2017) for each estimated RR with RR ≥ 1:

$$\text{E-Value} = \text{RR} + \sqrt{\text{RR} \times (\text{RR} - 1)}$$

Larger E-values indicate more robust evidence that the observed effect is not solely due to unmeasured confounding. It can be interpreted as the minimum strength of association (on the RR scale) that an unmeasured confounder would need to have with both the cue and the outcome in order to fully account for the observed effect. For instance, an E-value of 1.5 means that a hidden confounder would need to be associated with both the cue and the outcome by a risk ratio of at least 1.5 to nullify the effect.

The E-values in Figure 13 show that order cues yield the highest robustness scores, especially for smaller models (e.g., E-value $= 3.28$ for Qwen3-1.7B). Popularity and co-occurrence have lower E-values ($\approx 1.1 - 1.6$), suggesting they could be more easily explained by unmeasured variables, while internal alignment effects in larger models show E-values above 2, indicating comparatively strong causal robustness.

I.2 BOS Discard Info

The BOS balancing procedure necessarily discards any surplus items beyond the least-frequent (P, T, C, I) combination in each template. Tables 9 to 10 detail the retained-over-total counts per dataset and model. Although the absolute number of discarded items can be sizeable, enough balanced examples remain in every dataset–model pair to yield narrow confidence intervals in Figure 2, confirming that the subsequent cue-ablation results are not affected by data scarcity.

I.3 Per-Dataset Cue Effects

For completeness, Figure 14 breaks down the BOS cue-ablation gaps by dataset and model, complementing the aggregate view in Figure 2.

I.4 Risk Ratios

The BOS risk ratios are within-model quantities and depend on the model's baseline accuracy when a cue is absent. Algebraically, RR $= 1 + \Delta/p_0$, where $p_0 = \Pr\big(Y = \hat{Y} \mid F = 0\big)$ and Δ is the absolute accuracy lift when the cue is present, i.e., $\Pr\big(Y = \hat{Y} \mid F = 1\big) - \Pr\big(Y = \hat{Y} \mid F = 0\big)$. Thus, the same Δ yields a larger RR for a low-baseline model and a smaller RR for a high-baseline model. The effect is especially pronounced for I (number-GT alignment), because I is based on the model's extracted numbers. Consequently, the sets with $I = 0$ and $I = 1$ differ across models, leading to vastly different baselines p_0. In addition, noise in the extracted numbers (Allen-Zhu and Li, 2024; Schwartz et al., 2024; Yang et al., 2025b) might mislabel some I cases, attenuating RR_I towards 1. A model can therefore rely heavily on its numbers yet show a modest RR_I, simply because it already performs well when $I = 0$ (large p_0). Therefore, RR values are best interpreted in an intra-model fashion.

J Detailed Meta-predictor Results

Table 11 shows the accuracies of the bias-only meta-predictor broken down by dataset and model.

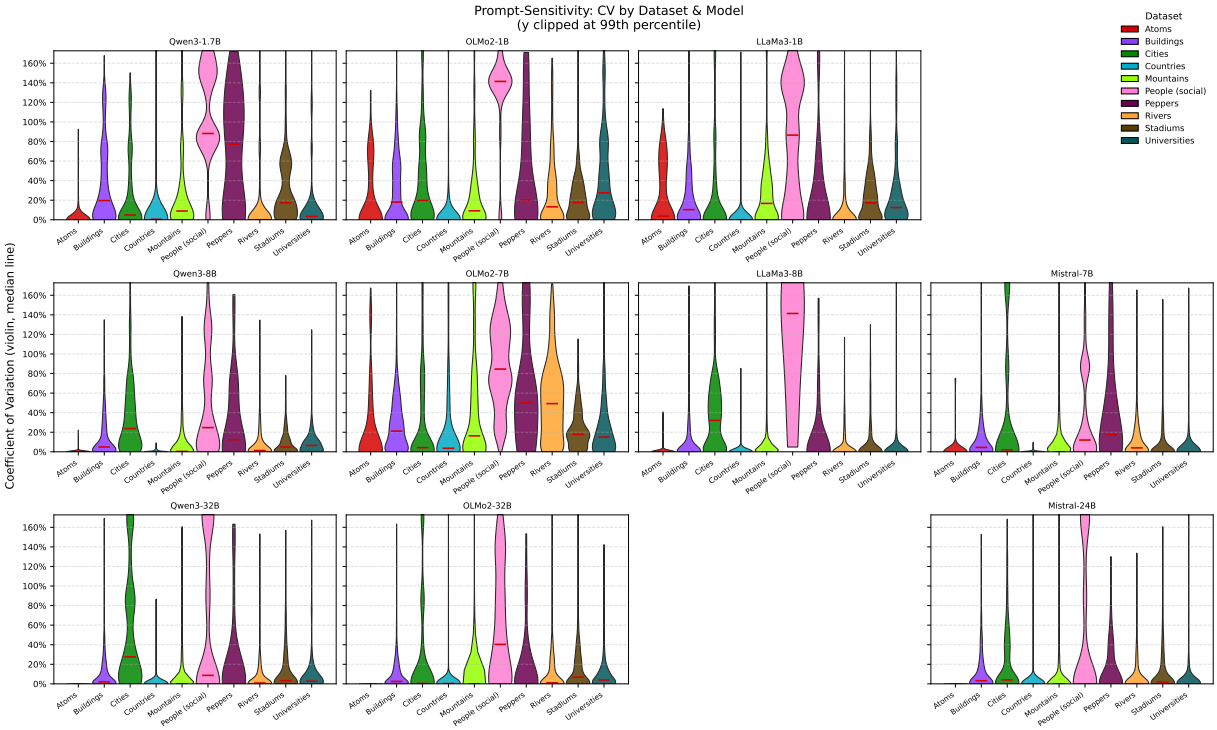

Figure 7: **Numerical-extraction stability per model and dataset.** Violin width encodes the density of row-wise CV values, the horizontal bar marks the median. Rows correspond to model-size buckets (*small*, *medium*, *big*); columns to model families.

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)
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 6: Examples of entity pairs where similarity to the considered keywords disagrees with the ground truth numerical values. The first entity is the one with the higher cosine similarity to the keywords, but with 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 7: 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.

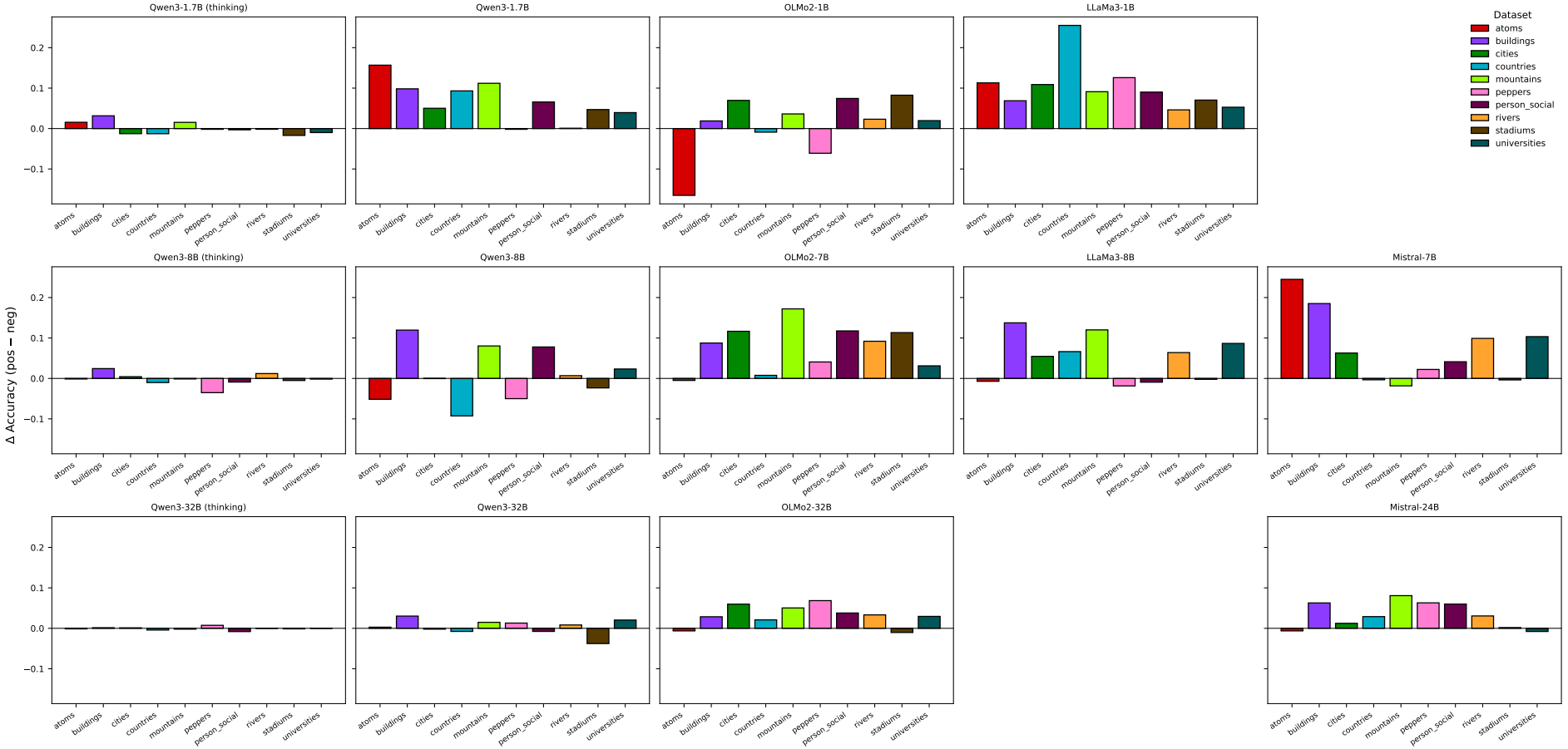

Figure 8: **Inter-polarity accuracy gap ΔAcc in detail.** Each panel is a single model; bars show the difference $\text{Acc}_{\text{Positive Polarity}} - \text{Acc}_{\text{Negative Polarity}}$ for every dataset (colour legend shared across figures).

Dataset	Atoms	Buildings	Cities	Countries	Mountains	People (social)	Peppers	Rivers	Stadiums	Universities
Popularity	24%	50%	63%	68%	56%	58%	61%	61%	65%	61%
Co-occurrence	50%	55%	61%	68%	52%	52%	58%	54%	51%	55%

Table 8: Cue–ground-truth alignment per dataset. Note that this will be the same for all models, as all models were evaluated on the same pairs.

Model	LLaMa3-1B	LLaMa3-8B	Mistral-24B	Mistral-7B	OLMo2-1B
Atoms	90/2812	1200/2831	1200/2832	81/2653	384/2828
Buildings	7664/20844	6336/21129	4512/21068	6064/20133	10464/21276
Cities	6336/22860	6672/22870	4992/22872	5264/21722	6512/22844
Countries	90/4703	96/4699	90/4703	77/4625	96/4700
Mountains	6976/14437	2160/14642	1440/15051	2096/12755	8240/14792
People (birth)	28400/88796	5312/88457	2352/88393	3456/78493	22080/88607
People (social)	9840/18922	10176/19404	1440/19176	5360/19305	84/19152
Peppers	144/714	78/864	60/828	90/791	192/862
Rivers	6416/17933	4224/18237	2096/18403	4784/17159	9888/18246
Stadiums	9312/18691	5712/18472	5280/18455	5184/17690	10752/18603
Universities	8064/23608	8784/23647	2832/23649	3552/22658	8160/23539

Table 9: Retained-over-total counts per dataset and model (part 1)

Model	OLMo2-32B	OLMo2-7B	Qwen3-1.7B	Qwen3-32B	Qwen3-8B
Atoms	1200/2832	96/2832	78/2831	1200/2832	1200/2832
Buildings	6384/21270	9312/21090	10560/21120	6912/21258	8496/21240
Cities	7200/22834	6528/22870	7680/22835	4800/22872	7008/22815
Countries	96/4704	288/4668	192/4703	90/4704	90/4449
Mountains	2640/14976	5760/14748	7680/14819	2544/14901	3888/15132
People (birth)	5136/88982	28608/88882	18288/88598	5616/88998	9120/88571
People (social)	9360/19428	10272/19020	8352/19234	4368/19247	8016/18983
Peppers	78/828	96/864	288/827	78/864	90/828
Rivers	4032/18643	8640/18225	7728/18302	3872/18322	5952/18351
Stadiums	7392/18525	9296/18589	8352/18518	7200/18356	7104/18325
Universities	4896/23652	7584/23499	9312/23628	4128/23709	5136/23712

Table 10: Retained-over-total counts per dataset and model (part 2)

Model Dataset	LLaMa3-1B	OLMo2-1B	Qwen3-1.7B	Mistral-7B	OLMo2-7B	LLaMa3-8B	Qwen3-8B	Mistral-24B	OLMo2-32B	Qwen3-32B
Atoms	53% ± 5%	65% ± 6%	66% ± 6%	69% ± 10%	72% ± 5%	72% ± 6%	72% ± 4%	74% ± 5%	75% ± 4%	75% ± 3%
Buildings	57% ± 2%	57% ± 2%	66% ± 6%	73% ± 6%	71% ± 5%	59% ± 3%	61% ± 4%	58% ± 3%	61% ± 5%	64% ± 3%
Cities	60% ± 2%	58% ± 4%	66% ± 4%	69% ± 3%	75% ± 5%	72% ± 1%	69% ± 2%	66% ± 2%	68% ± 2%	68% ± 2%
Countries	62% ± 5%	63% ± 5%	73% ± 5%	75% ± 3%	81% ± 4%	69% ± 6%	70% ± 3%	69% ± 4%	72% ± 3%	70% ± 4%
Mountains	56% ± 2%	57% ± 4%	68% ± 7%	69% ± 7%	65% ± 5%	65% ± 4%	61% ± 6%	61% ± 5%	59% ± 3%	64% ± 3%
People (social)	55% ± 2%	60% ± 5%	74% ± 14%	64% ± 6%	80% ± 12%	69% ± 8%	71% ± 8%	65% ± 7%	65% ± 6%	60% ± 2%
Peppers	62% ± 8%	63% ± 8%	65% ± 8%	63% ± 5%	63% ± 7%	59% ± 9%	60% ± 5%	59% ± 10%	58% ± 11%	57% ± 9%
Rivers	64% ± 8%	62% ± 5%	70% ± 10%	60% ± 4%	67% ± 9%	69% ± 7%	63% ± 4%	64% ± 4%	64% ± 2%	65% ± 2%
Stadiums	57% ± 2%	57% ± 5%	65% ± 4%	66% ± 5%	62% ± 6%	66% ± 2%	66% ± 2%	64% ± 3%	65% ± 2%	69% ± 2%
Universities	60% ± 5%	59% ± 4%	74% ± 8%	64% ± 2%	62% ± 4%	64% ± 4%	64% ± 2%	63% ± 3%	61% ± 4%	62% ± 3%
Avg.	59% ± 4%	60% ± 5%	69% ± 7%	67% ± 5%	70% ± 6%	66% ± 5%	66% ± 4%	64% ± 5%	65% ± 4%	65% ± 3%

Table 11: Bias-only meta-predictor accuracy (% mean ± std) by dataset (rows) and model (columns). The meta-predictor is a logistic regression that uses only surface cues: QRank popularity, positional advantage, and cosine association with "bigger" terms. Values are averaged over prompt templates with 5-fold cross-validation; the bottom row reports the macro-average across datasets for each model. Higher is better.

K Detailed Case Analyses

K.1 Per Dataset Dataset Distribution of Cases

This section provides a deeper analysis of the two meta-predictors introduced in Section 5. Figure 15 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.

K.2 Case 2 Detailed Analysis

In the original taxonomy in Section 5.1, Case 2 gathers all samples for which the model's own numerical comparison and the meta-predictor point to the same answer. Because the signals are perfectly aligned, we cannot tell which one actually drives the decision. Luckily, mention order is the strongest single bias we have identified (Figure 2) and trivial to reverse without altering anything else. We therefore reran each Case 2 prompt with the two entities swapped, keeping every other token unchanged. The swap leaves popularity and cosine cues untouched but inverts the positional feature. After that we, re-classify the samples into the four cases. An interpretation of what it means for a sample to end up in each case after having its order swapped is summarized in Table 12.

In brief: if the model answers with the same entity while the meta-predictor disagrees (Case 1) it suggests that the original response was anchored in the model's own numbers. When both model and meta-predictor remain aligned (Case 2) the two signals are still inseparable. A change in the model's answer that the meta-predictor correctly anticipates (Case 3) betrays domination by the positional cue. Finally, a change in the model's answer neither explanation anticipated (Case 4) points to residual noise or to biases that our simple meta-predictor does not capture.

Figure 16 shows how the once-ambiguous Case-2 samples redistribute across the four cases after the order of the entities is swapped. The case distribution after swapping reveals size-dependent patterns. Smaller and mid-sized models (e.g., Qwen3-1.7B, OLMo2-7B) often transition Case 2 items into Case 3, showing that mention order alone can override internal reasoning. OLMo2-7B is especially prone, which is consistent with the strong position bias observed in the BOS analysis (Figure 2). The smallest model we tested shows a notably high Case 4 rate, both before and after the swap (see Figure 4). Neither the biases nor numerical reasoning explain these predictions well. Their answers appear noisy rather than systematic. Larger models (e.g., OLMo2-32B, Mistral-24B) show the opposite trend. Roughly half of their Case 2 items remain unchanged after the swap, and about 30% move into Case 1, indicating that their choices are anchored in numerical representations. Case 4 remains rare. Notably, the $\approx 30\%$ transition into Case 1 is stable across almost all models, suggesting that once the numerical signal dominates, it does so reliably. Note that this may reflect "easy" comparisons with large numerical gaps; we investigate this further in Section 5.2.

K.3 Case 1 vs Case 3 Details

Figure 17 breaks the Case 1 vs Case 3 effects down by model (panels) and dataset (rows) while keeping the feature set (columns) identical to the figure from the main paper. The per-dataset view

New Case	Observation	Interpretation
Case 1	Model keeps its original choice; meta-predictor now expects the opposite.	Decision is anchored in internal numerical knowledge; positional cue was not decisive.
Case 2	Model and meta-predictor both remain unchanged.	Signals still coincide. We cannot disentangle whether numbers or biases drove the choice.
Case 3	Model changes its answer exactly as the meta-predictor predicts.	Positional cue overrides numerical preference; behavior is dominated by surface heuristics.
Case 4	Model changes its answer to an outcome neither explanation predicts.	Residual noise, reliance on unmodeled cues or misjudgment by meta-predictor; indicates instability.

Table 12: Interpretation of the four cases after swapping the entity order in Case 2 samples.

largely mirrors the global pattern from Figure 5. The strength of the observed effects varies by dataset and model size, with larger models typically showing clearer positive value gaps and more negative error/variance means. White cells indicate that d could not be estimated for that model–dataset–feature triplet (e.g., one of the cells had no support after filtering to Cases 1/3 or the within-group variance collapsed to zero).

L Cohen's d

We briefly recall how Cohen's d is computed, applied to the specific setting where we want to compare feature values between Cases 1 and 3. For each feature x, we first split the data into the two groups and compute sample means and variances:

$$\mu_1 = \text{mean}(x \mid \text{Case 1})$$
$$\mu_0 = \text{mean}(x \mid \text{Case 3})$$
$$\sigma_1^2 = \text{var}(x \mid \text{Case 1})$$
$$\sigma_0^2 = \text{var}(x \mid \text{Case 3}).$$

We then form the pooled standard deviation, which summarizes the typical spread inside the two groups:

$$\sigma_p^2 = \tfrac{1}{2}\left(\sigma_1^2 + \sigma_0^2\right)$$
$$\sigma_p = \sqrt{\sigma_p^2}$$

Finally we standardize the mean difference

$$d = \frac{\mu_1 - \mu_0}{\sigma_p}.$$

M Detailed CoT Analysis

Figure 18 compares the performance of Qwen3 models with and without thinking per dataset.

Prompting models to "think" before answering changes the case distribution drastically. As Figure 19 shows, the largest gain comes from Case 1 ($\approx +6$ pp across models), followed by a smaller but still clear rise in Case 2 ($\approx +5$ pp). Taken together, these shifts mean that the final pairwise choice agrees more often with the model's own numerical comparison, as expected.

The percentage of Case 3 items shrinks significantly, indicating a reduced tendency to follow surface-form cues when they conflict with the model's numerical comparison. Among the Case 3 items that remain, the conditional correctness with respect to ground truth improves: if we look only at those remaining Case 3 rows, roughly half of them are correct with respect to the ground truth, whereas this share was lower without thinking. Put plainly, when "thinking" models still side with non-numeric cues, those cues line up with ground truth more often than before. We hypothesize the model sometimes treats its numeric estimates as unreliable and gives partial weight to alternative and occasionally informative signals. Case 4 also shrinks overall, but the overall fraction of cases where it is right with respect to the ground-truth remains the same. We thus find that explicit reasoning reduces bias-driven and noisy inconsistencies but does not achieve fully faithful numerical grounding.

This case redistribution aligns with our findings from Figure 6: the increase in pairwise accuracy is nearly mirrored by the increase in internal consistency. That parallel movement has a simple cause: *"thinking ´ ´ mainly changes which signal the model follows at comparison time, not the quality of the signals themselves*. The numerical extraction pipeline is unchanged, so the correctness of the numbers is fixed; what changes is that more decisions are made in agreement with those numbers, as seen in a shift of mass from Case 3/4 to Case 1/2.

As a causal cross-check, we repeat the BOS ablation with and without "thinking" (Figure 20). Bars above the dashed line indicate that the cue, when set to favor the correct answer, increases accuracy; hatched bars show the thinking variant. Across

all three Qwen3 model sizes, reasoning clearly strengthens the Internal feature, meaning that *decisions follow the model's own numeric comparison more often*. For the two largest models, all three of the surface cues decrease to negligible levels. For Qwen3-1.7B, the Popularity and Co-occurrence cues remain broadly similar; the Order cue decreases in importance but remains decisive. As a reminder (see Appendix I.4), BOS risk ratios are within-model quantities, hence bar heights should not be compared across models. In particular, the Internal bar appearing smaller for Qwen3-32B, when compared to the same bar for Qwen3-8B or Qwen3-1.7B in thinking mode, does not imply weaker numerical reliance.

Model: QWEN3-32B

Mount Davis (predicted: 2160m actual: 979m)

Grand Colombier (predicted: 1686m actual: 1534m)

Trace: Okay, let's see. The user is asking which mountain is lower between Mount Davis and Grand Colombier. I need to find the elevations of both. First, Mount Davis. I'm not too familiar with it. Maybe it's in the US? I recall there's a Mount Davis in West Virginia, part of the Appalachian Mountains. Let me check its elevation. From what I remember, Mount Davis is around 4,863 feet (about 1,482 meters). Now, Grand Colombier. That's in France, part of the Alps. I think Grand Colombier's elevation is higher. Let me confirm. Grand Colombier is approximately 5,085 feet (around 1,547 meters). Wait, no, maybe I'm mixing it up with another peak. Wait, no, actually, I might be confusing it with other Alps mountains. Let me double-check. Oh, wait, maybe Grand Colombier is higher than Mount Davis. So if Grand Colombier is around 1,547 meters and Mount Davis is 1,482 meters, then Mount Davis is lower. So the answer should be Mount Davis.

Model: QWEN3-8B

Dragon's Breath (chili pepper) predicted: 1000000 actual: 2480000

Trinidad Scorpion Butch T predicted: 1400000 actual: 1463700

Trace: Okay, the user is asking which pepper is milder between Dragon's Breath and Trinidad Scorpion Butch T. They want the answer based on the Scoville scale. First, I need to recall the Scoville ratings for both. Dragon's Breath is known to be extremely hot, around 2.5 million SHU. The Trinidad Scorpion Butch T is also very spicy, but I think it's slightly less than Dragon's Breath. Wait, let me confirm. Wait, actually, the Trinidad Scorpion Butch T is about 1.4 million SHU, while Dragon's Breath is around 2.5 million. So the Trinidad Scorpion Butch T is milder. But wait, I might be mixing up some numbers. Let me double-check. No, I think the Dragon's Breath is higher. Wait, some sources say Dragon's Breath is up to 2.5 million, while Butch T is around 1.4 million. So Butch T is milder. Therefore, the answer should be Trinidad Scorpion Butch T.

Model: QWEN3-32B

Breg (predicted: 120km actual: 49km)

Krka (predicted: 76km actual: 73km)

Trace: Okay, the user is asking which river is shorter between Breg and Krka. Let me recall what I know about these rivers. First, I need to remember the lengths of both rivers. Breg River is located in Croatia, I think. It flows into the Adriatic Sea. From what I remember, the Krka River is also in Croatia and is known for its waterfalls, like the Krka National Park. Wait, I'm not entirely sure about their lengths. Let me try to recall. The Krka River is a bit longer than Breg. I think Krka is around 70 kilometers long, while Breg is shorter, maybe around 30 kilometers? Or is it the other way around? Hmm. Wait, no, I might be mixing them up. Let me check my memory again. The Krka River is famous for its series of waterfalls and is a popular tourist spot. If it's longer, that would make sense because longer rivers can have more features like waterfalls. Breg River is a smaller river that flows through the town of Breg. Wait, actually, I might have confused them. Maybe Breg is longer? No, I think Krka is longer. Let me confirm. Another way to think: Krka is part of the larger water system, maybe connected to the Adriatic Sea through a longer course. Breg might be a tributary or a shorter river. I think the correct answer is that Breg River is shorter than Krka River. So the answer should be Breg.

Table 13: Exemplary reasoning traces

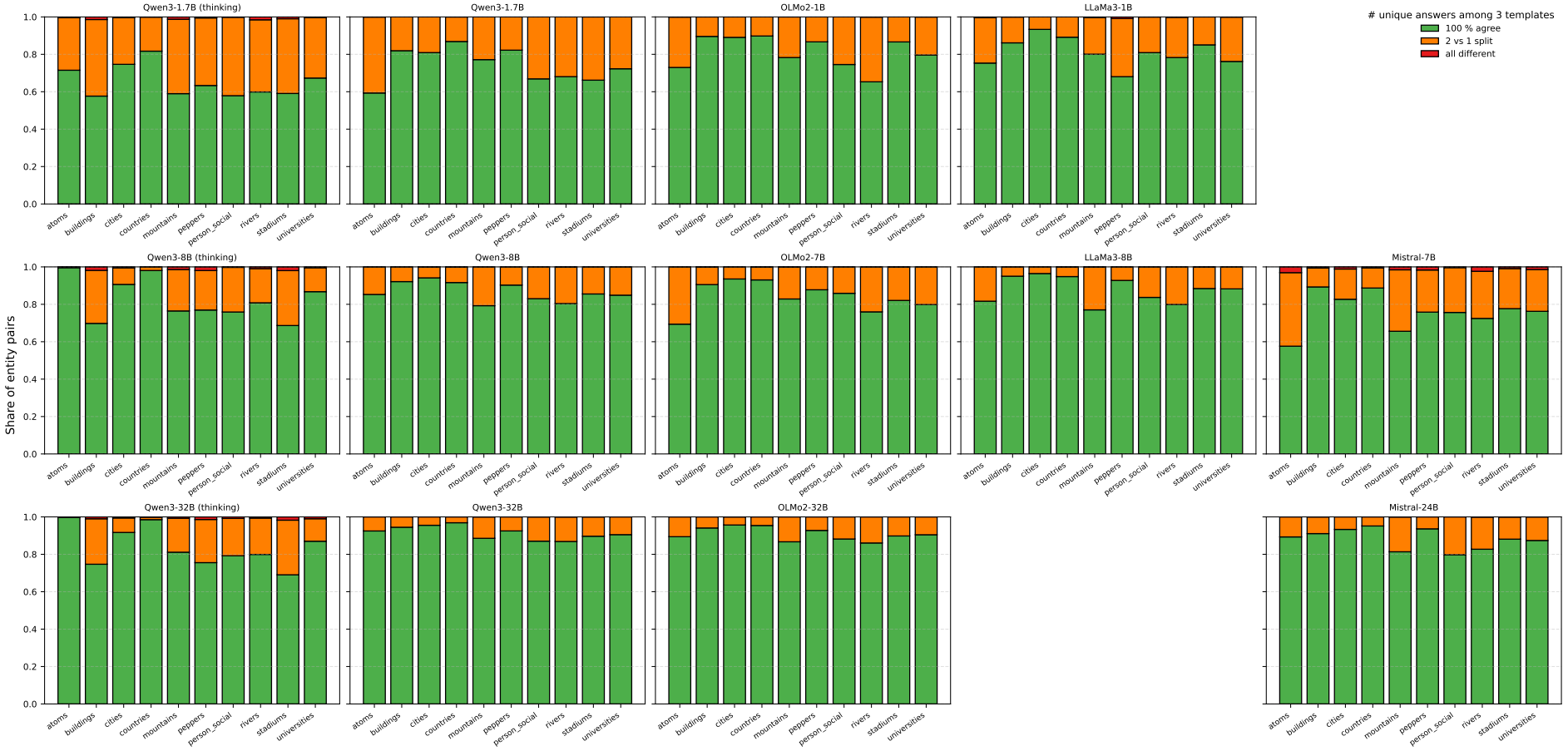

Figure 9: **Template-majority ratio (TMR) per model and dataset.** Stacked bars sum to 100 %. Green = all three templates agree; orange = 2-vs-1 split; red = three different winners.

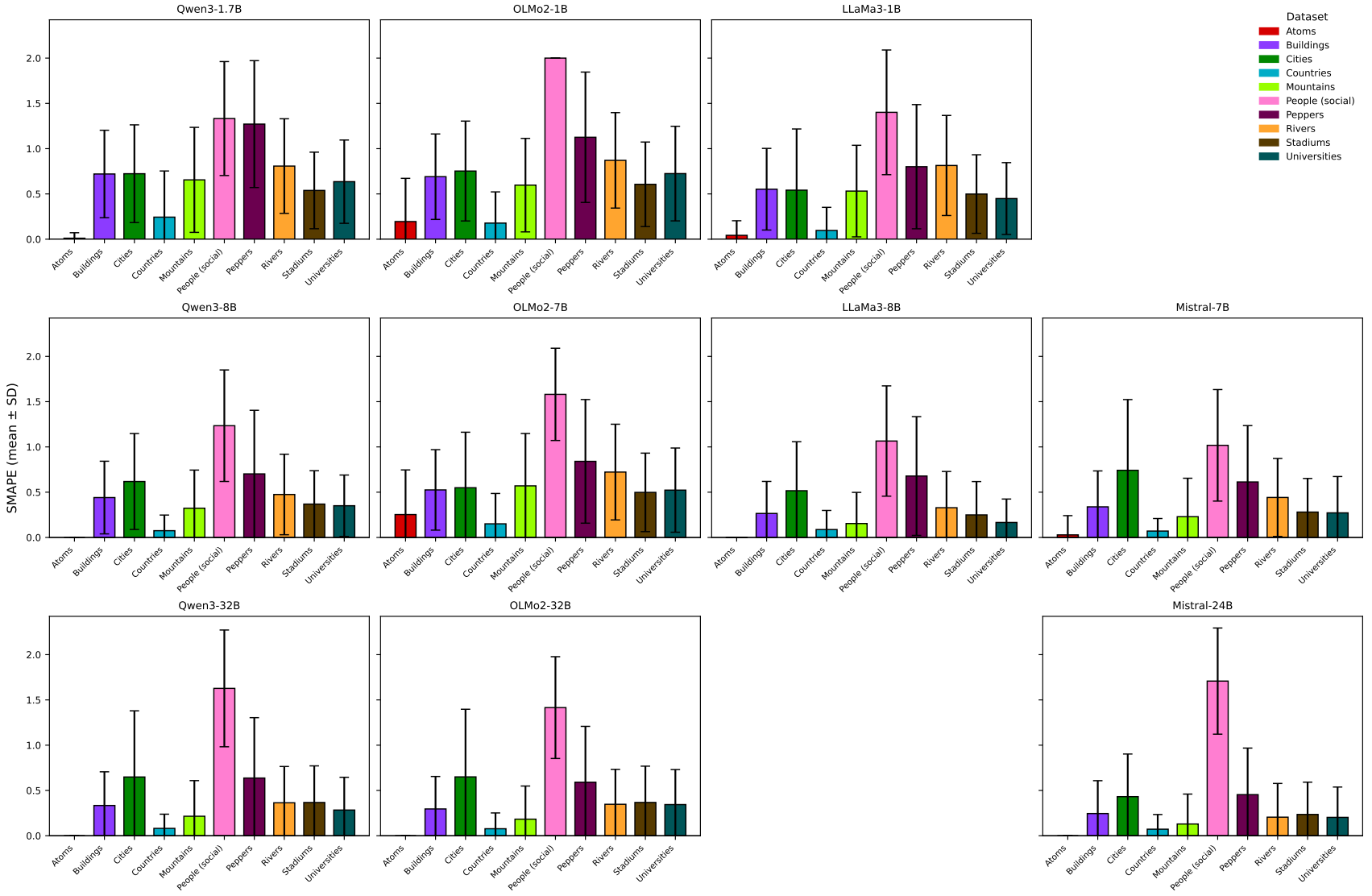

Figure 10: SMAPE (mean ± SD) for each model–dataset pair. Bars show the mean SMAPE across extractions, with whiskers denoting standard deviation across samples. Lower values indicate more accurate numerical predictions.

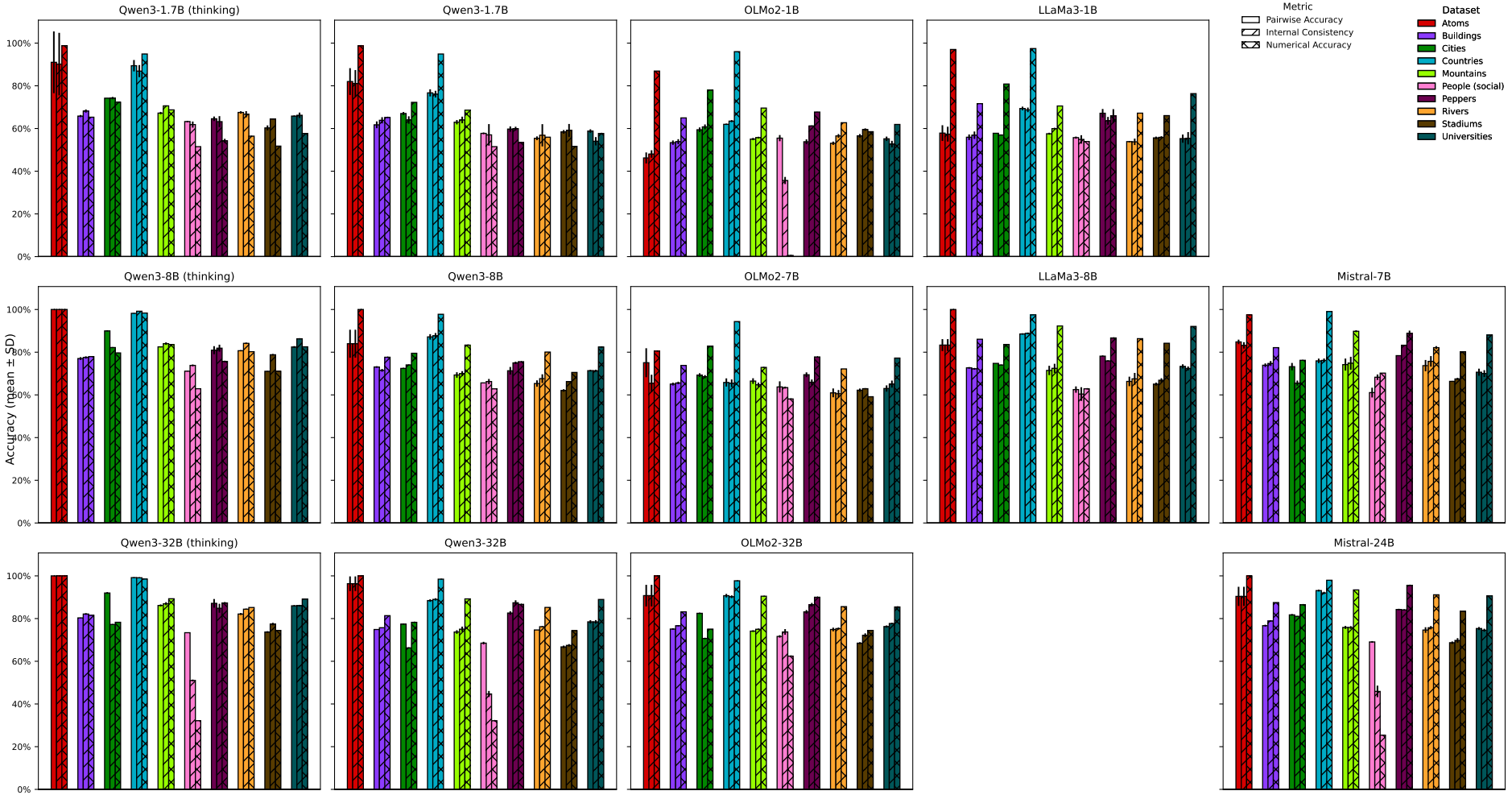

Figure 11: **Accuracy comparison for positive polarity prompts**. Each panel shows results for a single language model. For each dataset, we display three side-by-side bars: the solid bar represents the ranking accuracy, the hatched bar shows the internal consistency and the cross-hatched bars show the numerical accuracy. Bars indicate mean accuracy; error bars show ± 1 standard deviation across prompt templates.

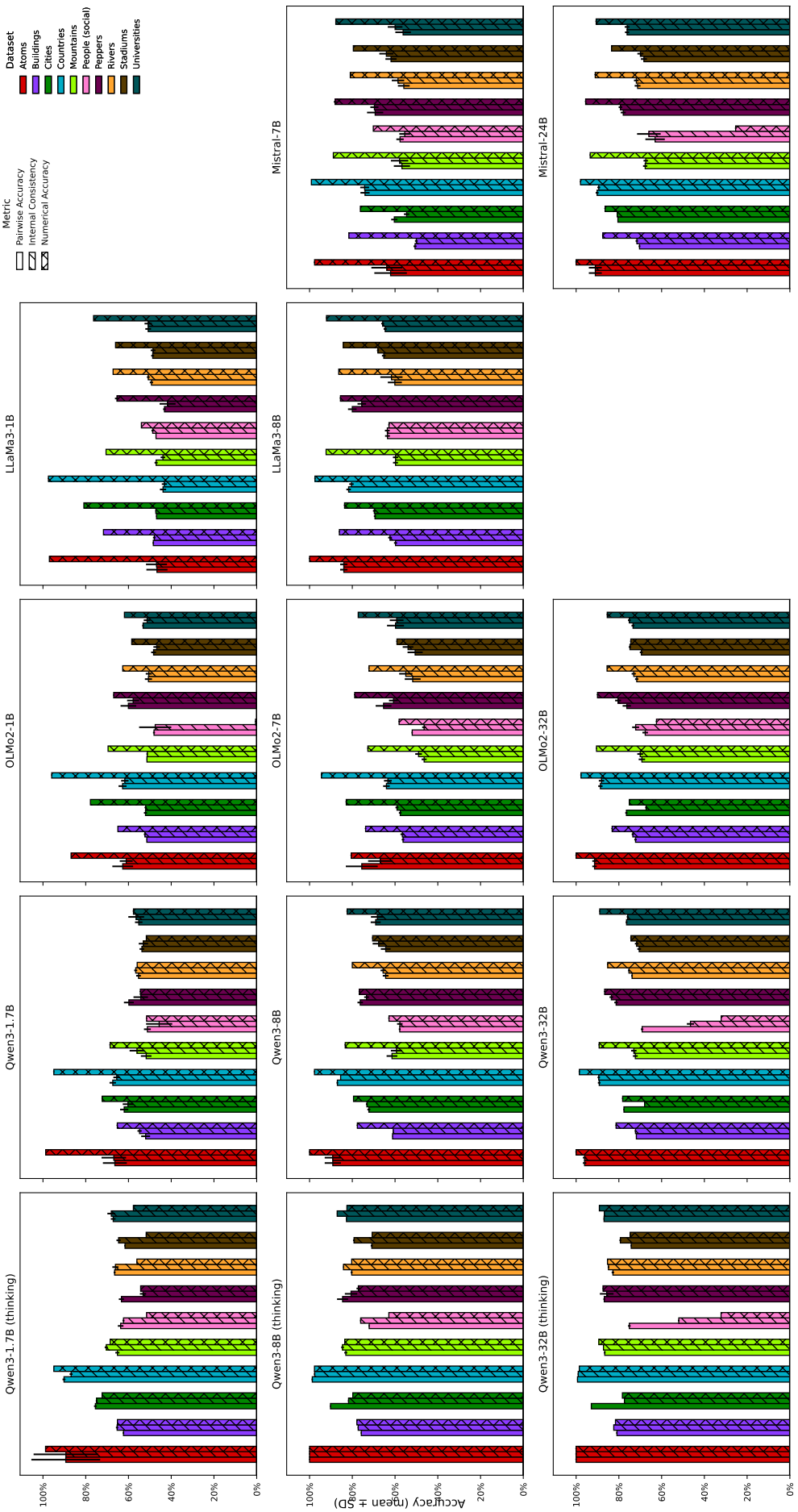

Figure 12: **Accuracy comparison for negative polarity prompts**. Same layout as Figure 11, but for negative polarity prompts.

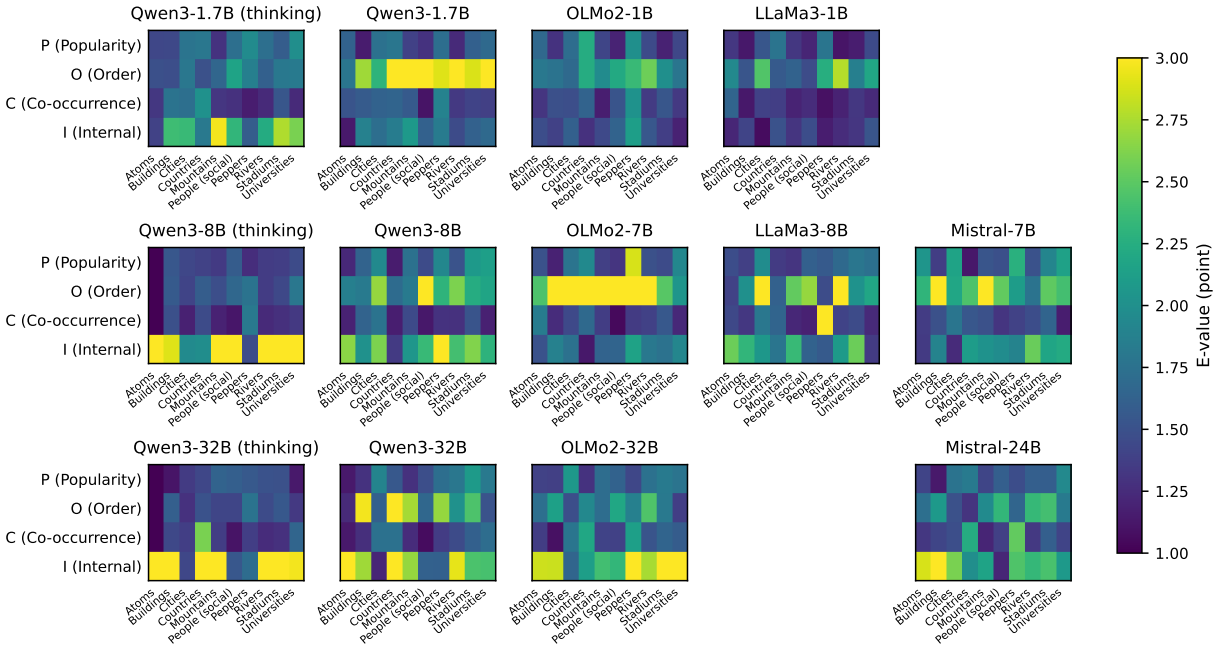

Figure 13: E-values, which quantify the strength of an unmeasured confounder (on the RR scale) needed to fully explain away the observed effect. Higher values indicate more robust evidence for a causal effect.

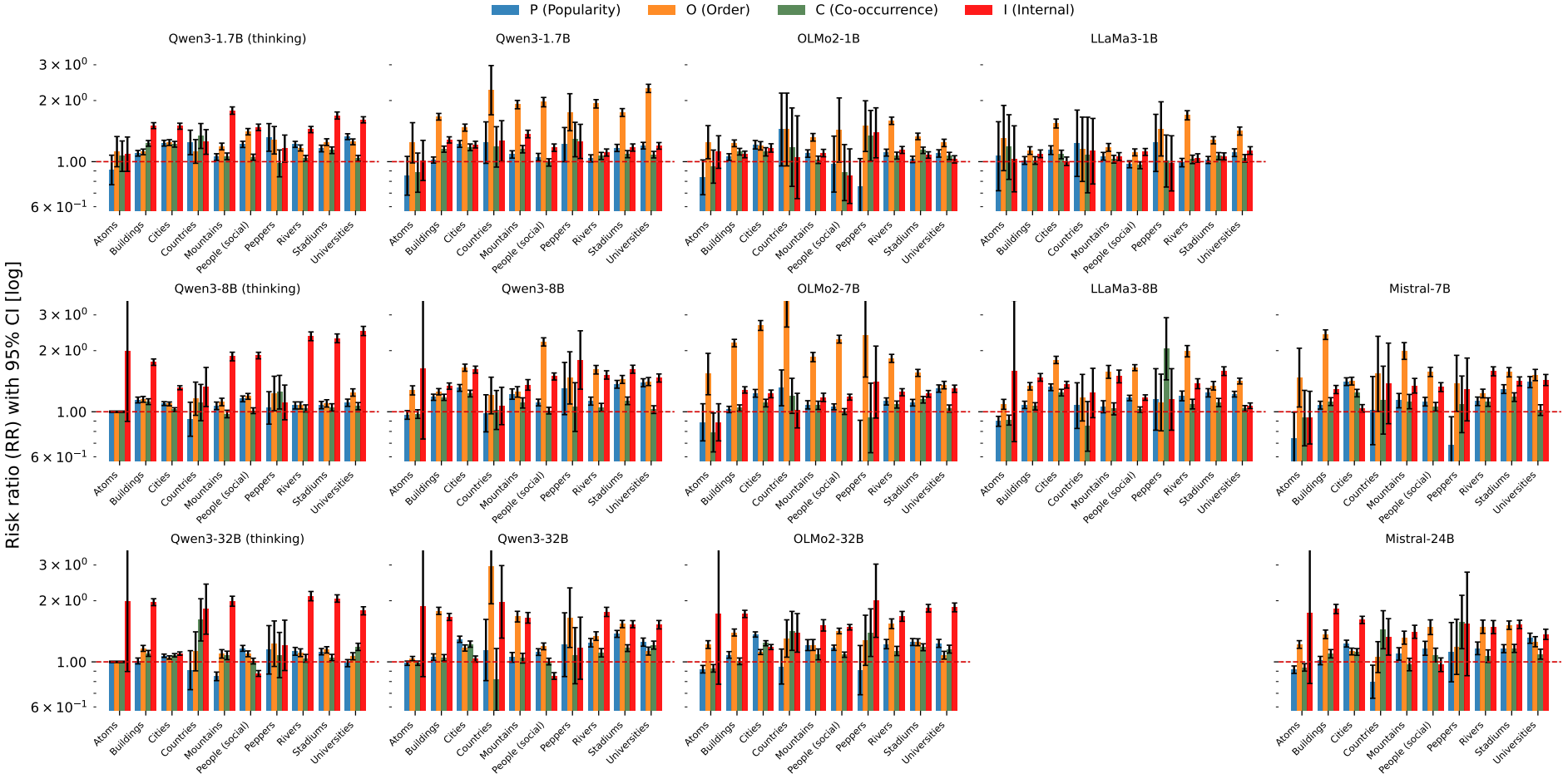

Figure 14: Per-dataset, per-model feature ablation analysis on the Balanced–Orthogonal Subset (BOS). Each block of four bars reports the accuracy change that follows when a single feature is flipped while all other observable factors are fixed: popularity (P), mention order (O), cosine similarity (C), and internal–ground-truth alignment (I). Signed gaps are shown for P, C, and I; an absolute gap is shown for O because its direction can reverse with prompt polarity or model family. Whiskers indicate ± 1 standard deviation across prompt templates.

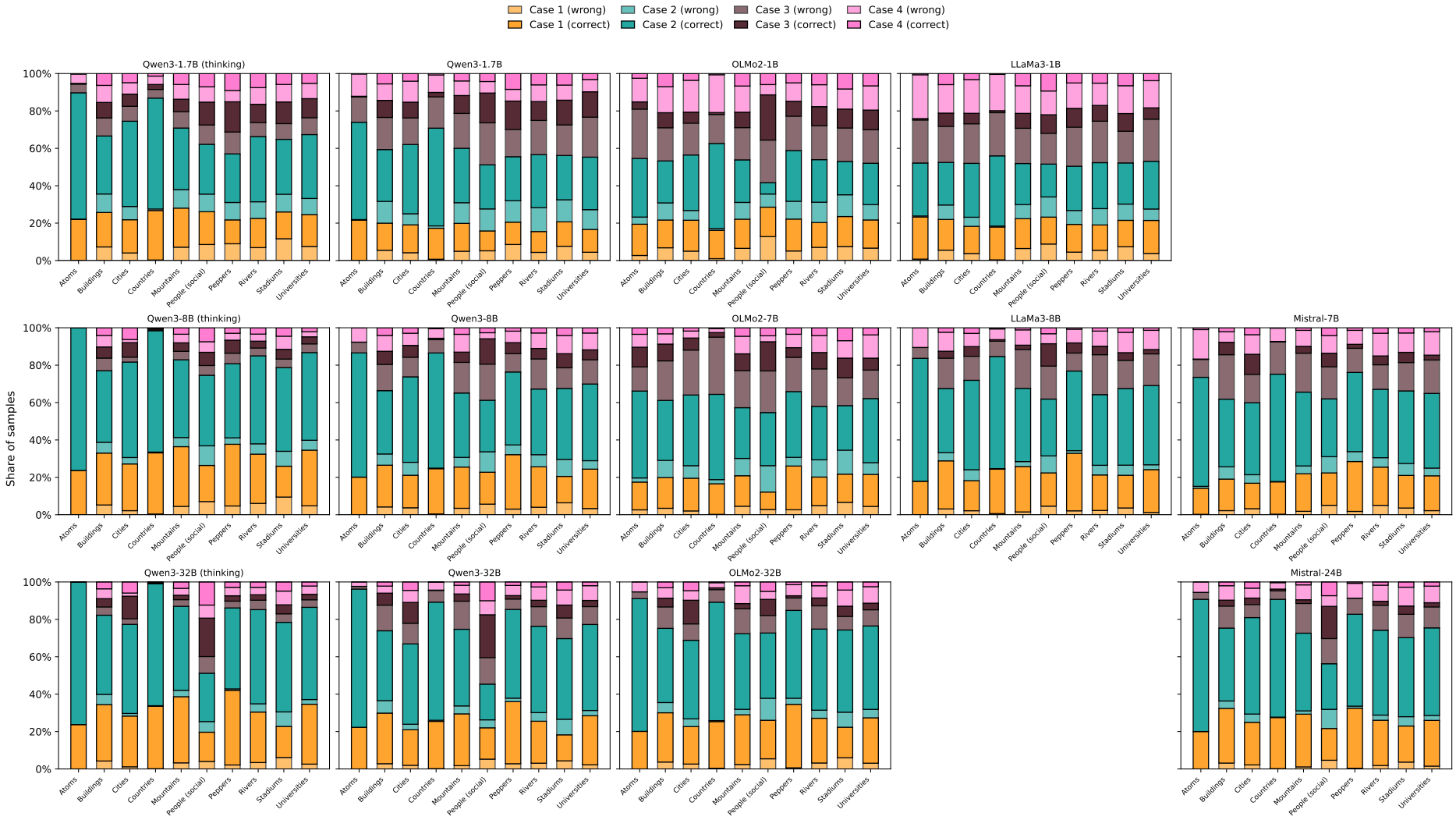

Figure 15: Detailed breakdown of the four cases (from Section 5) per dataset and model. Within every panel each colored, stacked bar corresponds to one dataset. The segments show the proportion of test samples that fall into **Case 1** (numerical and pairwise agree, meta-predictor disagrees), **Case 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 4 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. Opaque segments marks correct predictions and translucent segment mark errors with respect to the ground truth.

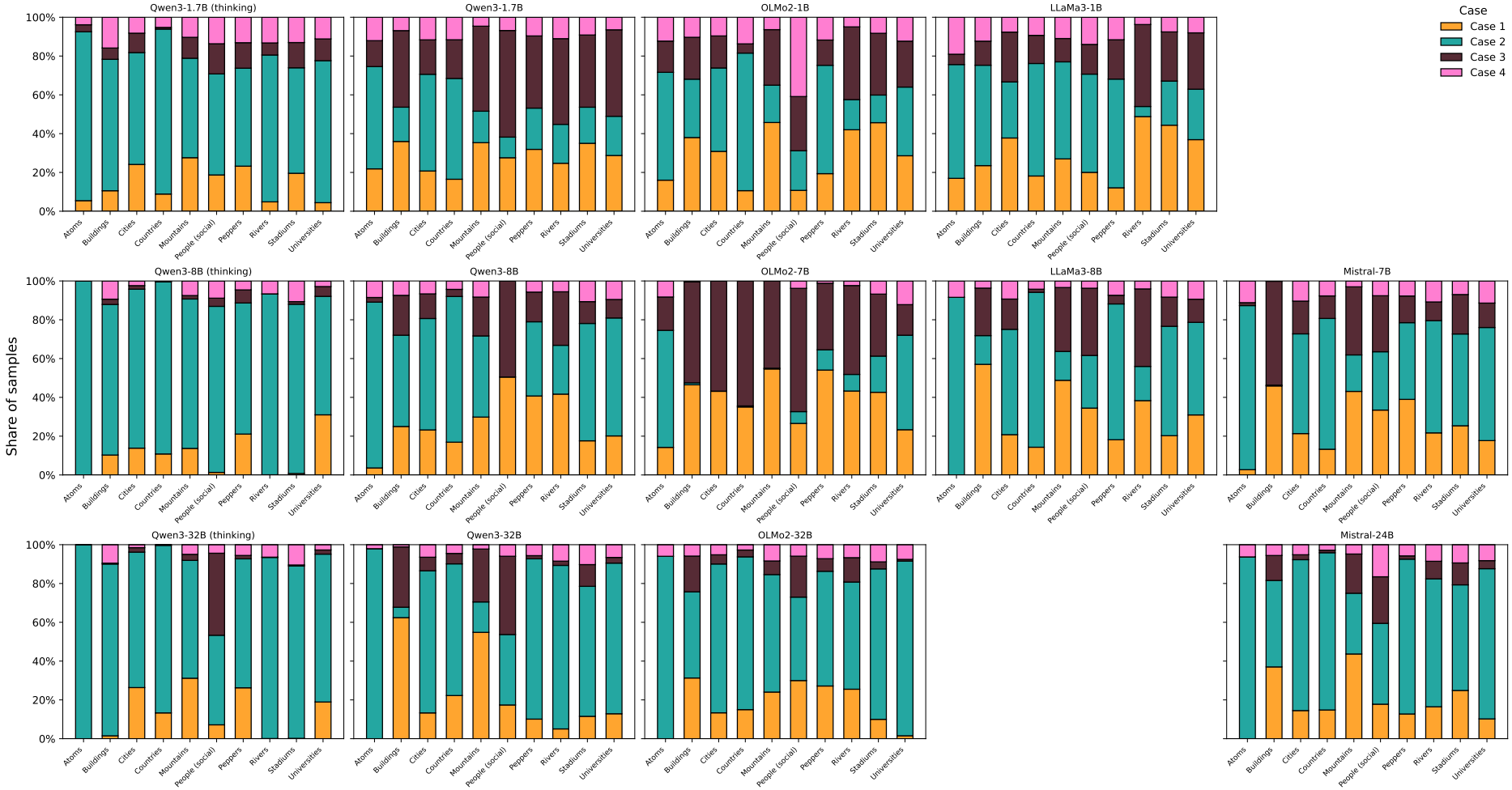

Figure 16: Per-dataset, per-model breakdown of the mention-order swap experiment. Bars show how items that were originally in Case 2 migrate to Cases 1–4 after reversing the entity order.

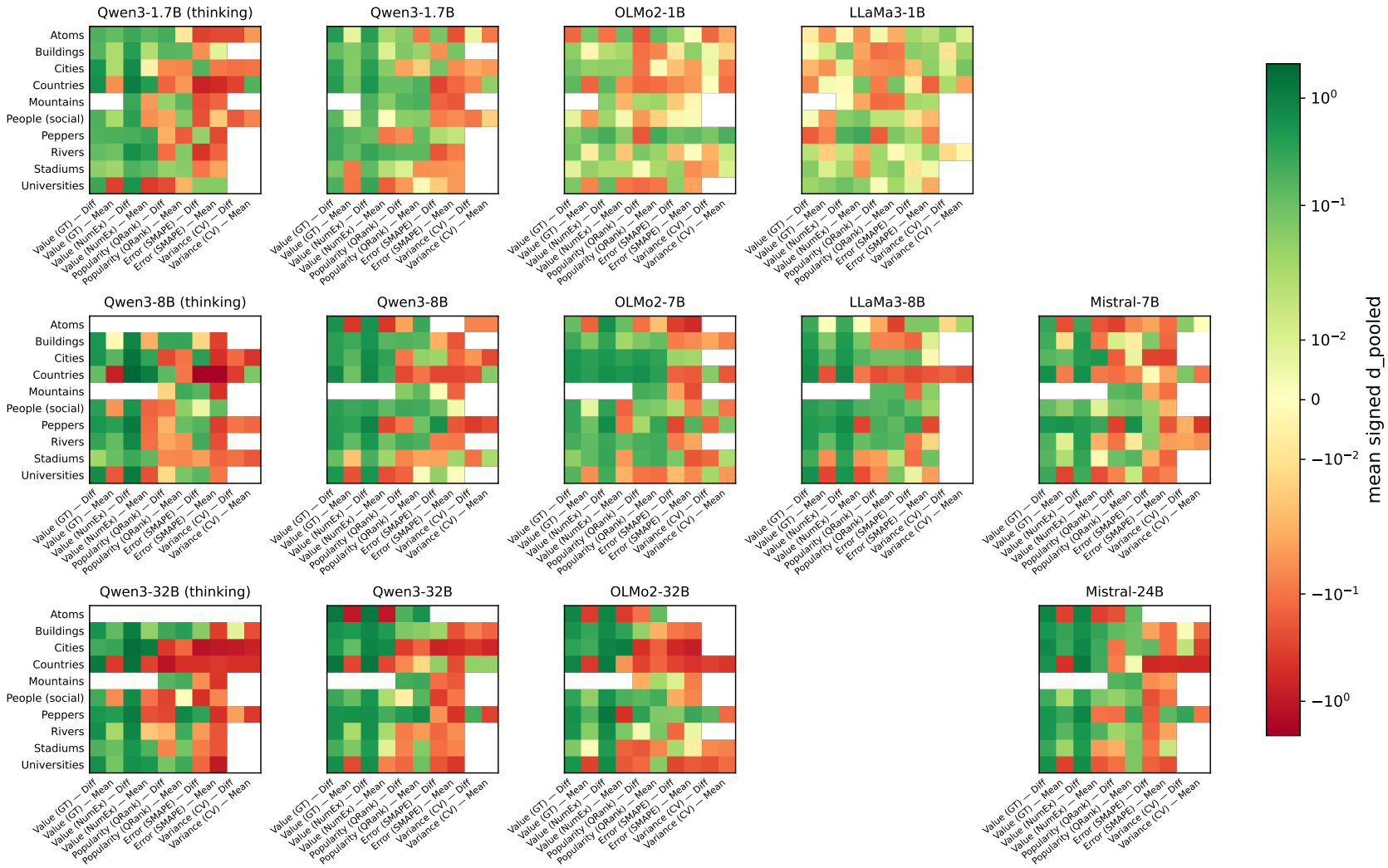

Figure 17: Per-model, per-dataset signed Cohen's d for the Case 1 vs Case 3 contrast, pooled over prompt polarity. Columns list the feature summaries used in the main figure (for each family: Value, Popularity, NumEx error, NumEx variance, showing Diff and Mean). Colors use a symmetric-log scale (green = larger in Case 1, red = larger in Case 3; shared colorbar across panels). White cells denote missing estimates due to insufficient support for one case or zero within-group variance

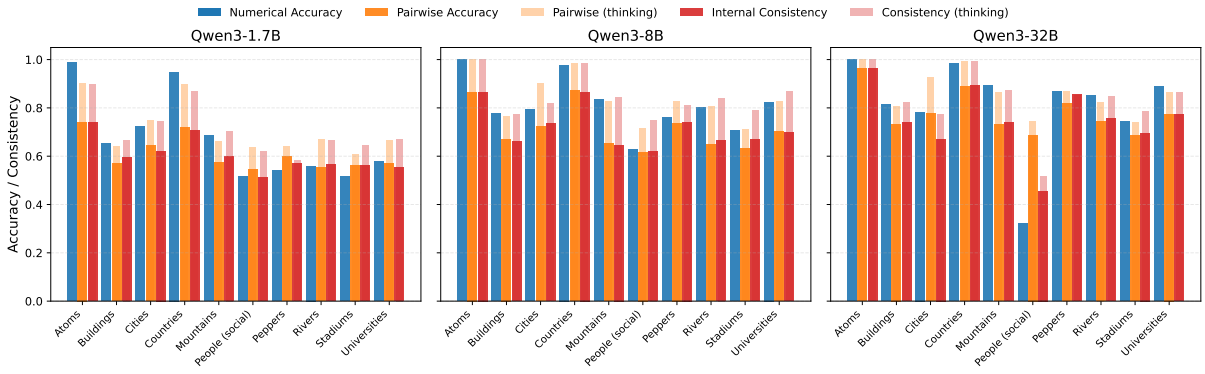

Figure 18: Comparison of the performance with and without thinking, in terms of pairwise accuracy and internal consistency. The figure also reports numerical accuracy (without thinking).

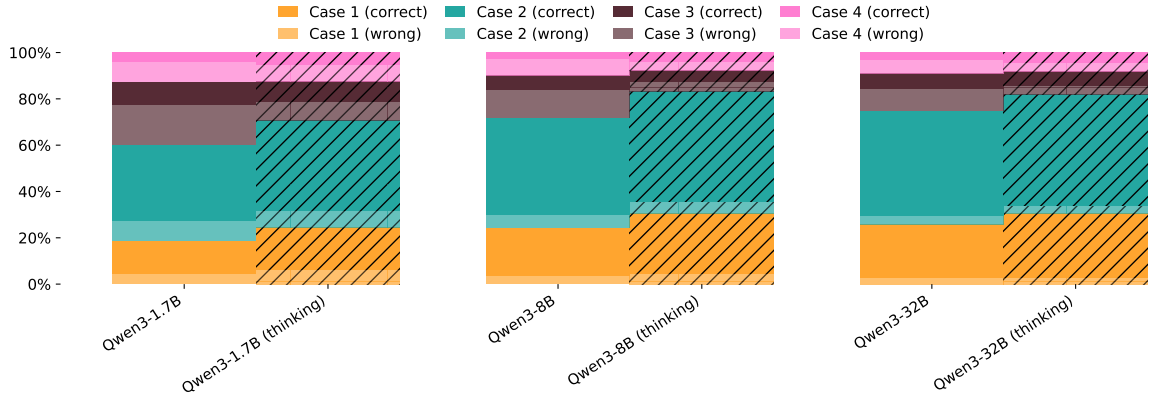

Figure 19: Case distribution with and without thinking.

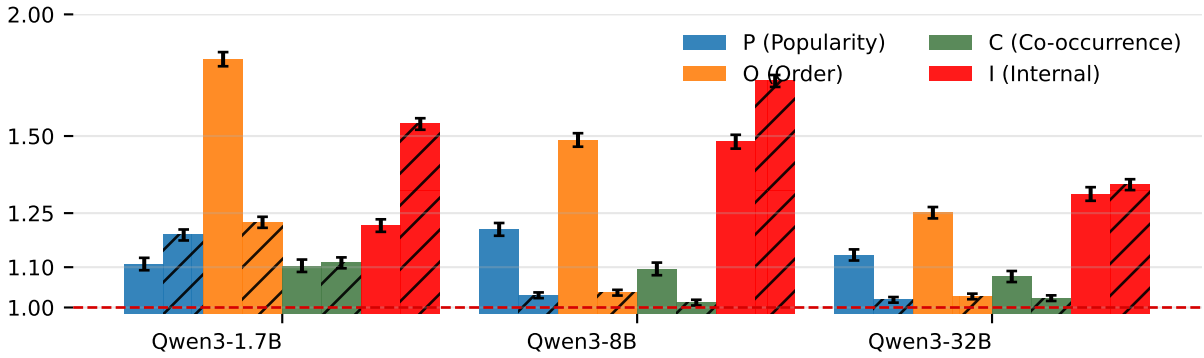

Figure 20: Risk ratios from BOS analysis with and without thinking.