Bouba Beats Kiki, and Other Regularities in the Subjective Preferences of LLMs

Katy Blumer
Cornell University / Google
kblumer@cs.cornell.edu

Jon Kleinberg
Google
kleinberg@cornell.edu

Ravi Kumar
Google
ravi.k53@gmail.com

Andrew Tomkins
Google
atomkins@gmail.com

Abstract

Users engage with large language models on tasks that are purely objective ("When was Netflix founded?"), purely subjective ("What should I name my child?") or anywhere in-between. We have rich tools for evaluating model accuracy with respect to objective criteria. However, when it comes to subjective aspects of a task, we do not have similarly developed methodology at hand. In this paper, we take some early steps towards framing the question of subjective evaluation. Rather than measure how often a model is right or wrong on a question, we instead ask, how do models make choices among multiple viable responses without a clear winner?

To study this question, we restrict our focus to short-form content, and consider comparison tasks among three such content types: numbers, baby names, and fictional cities. We identify 29 aspects of these short-form examples covering many areas of syntax and semantics (length, alphabetic order, popularity, vowel mix, spoken sound, etc.). We show first that models tend to have strong and reproducible preferences for or against some of these aspects. Second, we show that variations in how the model is asked to express a preference may cause some of these aspects to activate or deactivate in the model's selection. Third, we show that overall preference distributions among pairs may vary radically based on seemingly-unimportant variations in the nature of the question posed to the model, raising new questions about manipulation. And finally, we show that, if models are used both to create lists of alternatives and to rank them, there are unexpected interactions between the two stages that may also drive the models' stated preferences.

1 Introduction

LLMs have demonstrated remarkable advances in domains that admit clean signals of correctness to guide training. These signals come from a range of sources, including:

- Large corpora used in early training for next-token prediction, e.g., [Soldaini et al., 2024].
- Human rating data or expert-created content used for SFT or RLHF [Ouyang et al., 2022].
- Model-derived auto-ratings used for reward modeling or RLAIF [Ouyang et al., 2022]
- Automated output validation as in coding tasks with complete tests [Jimenez et al., 2024] or math problems with Lean proofs [de Moura and Ullrich, 2021], or logic puzzles with known correct answers, used in training as RLVR [Wen et al., 2026, DeepSeek-AI et al., 2025].

Under this pattern, models are optimized to produce outputs that are "preferred" under various criteria like those described above. This highly successful recipe works because it relies on a purely objective

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

comparison of task outputs. In reality, the tasks that humans give to LLMs contain both objective and subjective factors. Our focus in this paper is to explore the subjective elements that do not have clear winners and losers, but are nonetheless critical to model performance. Consider for example the task of naming a newborn child. Expecting parents view this task as important and increasingly ask LLMs for input, but we do not approach this task by producing an objective order of baby name quality; instead, we recognize that there are critical subjective factors present.

In domains such as innovating business plans [De Freitas et al., 2025] or research directions [Zhao et al., 2025, Si et al., 2024], domain experts may be invoked to evaluate the quality of model outputs, but this expert evaluation does not make the task purely objective in nature. First, expert feedback is expensive and may not available for all tasks at the volumes required for training. Second, the feedback itself may be subject to decisions made by the researchers about factors whose relative importance are unknown: how important is novelty compared to capital requirements, for instance?

To understand LLM behavior with respect to the subjective elements of a task, we are inherently looking beyond the objective matter of which output is better. We may still evaluate the contribution of the LLM, but rather than asking whether the LLM found the "right" answer, we ask instead what conditions influence the model's choice among alternatives that are acceptable but entirely comparable. If certain models, or approaches to invoking those models, may produce radically different outcomes, we wish to understand the nature of these differences. Does one invocation more heavily emphasize particular elements of the problem? Do some invocations introduce more randomness than others? Are model choices disproportionately impacted by presentation? How do models respond to prompt changes that humans view as immaterial? What other factors are at play?

To do this study we must choose where to focus in the broad world of tasks with subjective elements. An important distinction can be drawn between short-form tasks, such as the baby name choice describe above, versus longer-form tasks, like crafting a social media bio. We see long-form tasks as inherently more difficult to study, as there are subtle interactions between the words in the generated output that make it difficult to unearth how specifically the model chose to distinguish between two long-form alternative. To keep the problem framing as simple as possible, we focus our attention on a range of short-form tasks, and find that, even in this setting, there is a wealth of complexity to study. We experiment specifically with on baby names, the names of fictional cities, and choices of numbers.

Problem and methodology. We propose methods for investigating the factors that impact the LLM's decision in subjective settings without any clear right answer. There is a mature academic literature characterizing the traditional "cognitive biases" that can arise when humans undertake subjective tasks. It is unclear, however, what corresponding forms of "LLM reasoning biases" might similarly influence model outcomes in subjective brainstorming.[1] In all, we study comparison tasks in which a model is asked to select an alternative from two choices. We identify and study a total of 29 preferences that may exist in these choices, across several domains, plus an additional set that emerges when the model is also used to create the list of items from which the comparison pairs are drawn. As examples of the types of preferences we consider, a model might prefer names that are longer or shorter, that begin with hard consonant sounds, or that are evocative of professions associated with agrarian societies.

To make our insights as generalizable as possible we design the domains and the tasks to be structured and stylized. We use three distinct structured domains representing different types of short-form comparable objects: human names (pulled from the most popular baby names in the US Census), fictional city names (generated by an LLM), and numbers (randomly sampled as a baseline). For city names, we set up the subjective decision-making framework through two interrelated tasks:

- *List creation*: given a target output (such as a fictional city name or a product tagline), the LLM should produce a list of N options for this target output.
- *Comparison*: given two options from the list, the LLM should say which option it favors.

Our comparison task relates to past work on ideation [Li et al., 2025b] and on biases in LLM-as-a-judge frameworks [Bai et al., 2023, Li et al., 2025a, Zheng et al., 2023], and we consider these further in Appendix A. But we extend the analysis in several new directions, as we discuss next.

[1]We will use "bias" in the statistical sense going forward to indicate model preferences for certain classes of outputs. We do not mean to imply that biases represent reasoning errors; instead, they may be viewed as dimensions the model may focus on in making a choice. We understand this might suggest an unintended connection to *model bias*, but we have not been able to identify an alternative term for the concept.

Analysis. First, we analyze how the multiple LLM preferences we have identified interact in driving outcomes for this basic comparison task. We include well-studied effects of position (which option is presented first and which is presented second) [Shi et al., 2025] and alphabetical ordering, along with many others. We also consider preferences that depend on the execution of the list creation phase: consider for example that a model might prefer options generated earlier in the list, as these items might be more readily accessible to the model. As this example suggests, list creation and comparison may interact with each other even when formulated as distinct tasks.

We also consider the prompt used to drive the model's comparison of two items. We find that relatively small changes to the prompt—even when the instructions appear essentially identical—can induce non-trivial changes in the outcomes. In particular, we observe that the choice of prompt can excite or depress the influence of certain preferences in the LLM's comparison outputs. Sometimes a careful reading of the prompt can explain these changes. For example, asking "Which item would you select first?" instead of "Which item do you prefer?" might activate the preference for alphabetically earlier items. More often, though, we do not discern any obvious reason why a particular prompt would cause one preference or another to become a stronger driver of the model's outcome. The fact that innocuous changes to the prompt might have a significant influence on outcomes raises both new research questions and new questions about LLM behavior.

Finally, we identify another source of interaction between list creation and comparison. In list creation, we find that the LLM will often generate options that fall into natural categories. In generating names of cities, e.g., there may be a run of contiguous city names in the list that are based on seasons, another contiguous run of city names that are based on compass directions, and so forth. We then find that many prompts will induce preferences that favor members of certain categories over others; this therefore is a bias that cannot be expressed in terms of fixed a priori features of the options, but instead in terms of the implicit categories they belong to. Hence, understanding model comparisons among choices also requires understanding the hidden structure behind the creation of the candidates.

Taken together, our results provide a richer view of some of the empirical regularities in how LLMs engage in subjective tasks, both in generating option lists and in comparing items from those lists.

2 Methods

Comparable objects. As discussed in Section 1, we work with short-form items so that the features are more clear-cut, and less dependent on more complex textual or stylistic properties associated with longer passages of text. Three types of short-form items were used: numbers, human names, and fictional city names. Numbers were generated by sampling 200 unique numbers from $\{0, \ldots, 999\}$.

Human names (henceforth referred to just as "names") were generated by pulling the most common 100 female and male names from the 2020 US Census [Comenetz, 2026]. Names were defined as "female" or "male" based on the "counts by sex" columns in the provided data; all names in the most common 100 of each category had at least a 60/40 gender split.

Fictional city names (henceforth referred to as "cities") were generated by asking Gemini 2.5 Flash for a list of 200 fictional city names. A duplicate was later removed, yielding 199 cities; see Appendix C.

Prompt expansion. The main set of 50 prompts was generated using iterative expansion from one prompt. The seed prompt was "Which of these two names do you prefer?." Then an LLM was asked to paraphrase the seed prompt 10 times (Prompt 1). This process was repeated 5 times, each time giving the LLM the set of already-generated prompts to avoid duplicates. The same 50 prompts were used for all 3 datasets, replacing "names" with "numbers" or "cities" as appropriate; see Appendix D.

Combination. Unless otherwise specified, results come from a dataset created as follows:

1. 200 comparable objects from a given domain (numbers, names, etc.) are acquired or generated as described above.
2. A set of 200 unordered pairs is generated such that each object is paired with exactly two other objects. This is done by sampling a random cycle over the 200 objects. Specifically, we generate a set of 200 unordered pairs $\{obj_i, obj_j\}$ by randomly permuting the objects into a sequence and pairing adjacent elements—including an edge between the last and first objects—such that each object appears in exactly two pairs.
3. Each unordered pair $\{a, b\}$ is compared twice, once for ordering (a, b) and for (b, a).

4. Each ordered pair is compared using each of 50 similar questions ("prompts"), structured using the standardized comparison template detailed in Appendix E.1.

This produces a total of $200 \times 2 \times 50 = 20{,}000$ LLM-driven comparisons.

Comparisons were all done with a temperature of 1 and model defaults for thinking, unless otherwise specified; see Section D. The set of LLMs that were queried is in Table 1; details of their exact configurations, API identifiers, and citations are in Appendix G. For reasons of space, many figures show only results for Gemini 2.5 Flash, but results for other SOTA models are broadly similar and are in Appendix J (prompt histograms in Appendix J.3 and across-model correlations in Appendix J.5).

3 Experimental Results

We begin in Section 3.1 by taking the features we have identified and studying how much each of them contributes to the overall comparison outcomes produced by the LLM. Next in Section 3.2 we slice these overall influence values by prompt, to see how much the wording of the prompt changes the outcomes. Finally, in Section 3.3 we revisit the ideation phase and discuss how domain-specific categories, used by the LLM to generate items, go on to influence the outcomes of choices.

3.1 Observed preferences

We set about to develop a robust but non-comprehensive set of features that would allow us to study LLM behavior in comparisons. Through an iterative process that began with known features, and expanded the set using iteration and suggestions proposed by Gemini 2.5 Flash in response to prompting, we identified a few dozen exogenous features, some of which are specific to a single domain, while others apply across two or three domains. The features for the names and numbers datasets are listed in Table 1 (brief definitions for unfamiliar terms are provided in Appendix F). Metrics for cities are listed in Appendix Table 9. As the asterisks indicate (using a Bonferroni correction for multiple comparisons across all model-metric pairs, $m = 183$), many features occur on the winning item in a comparison significantly more or less often than random chance would suggest. We evaluated statistical significance against a 0.5 baseline using a one-sample t-test over the win rates of the N unique unordered pairs. Crucially, these reported p-values are conservative estimates. To avoid inflating significance through correlated outcomes, we define our sample size N strictly as the number of unique unordered pairs. Consequently, each pair contributes only once to the statistical test, despite representing up to 100 individual comparisons (2 orderings $\times$ 50 prompts) in the dataset.

To assess the contribution of each feature to the observed outcomes, we report two measures for each feature. First, we compute for each feature i the probability that an item with feature i wins a comparison it participates in against an item j where the feature favors i over j; this is the measure reported in Table 1. However, as there are strong correlations among the features, we wish to tease out whether the feature i actually contributes to the win, or simply appears alongside other driving preferences. For this, we fit a multinomial logistic regression whose features are normalized scalar variables (length, order of magnitude) or binary variables indicating whether an item contains each preference (presented first, starts with a vowel, etc), so the item's odds of winning a comparison is proportional to $e^{w \cdot \beta}$ for binary feature vector w and learned coefficient vector β. This is equivalent to assuming that each item's odds of winning a comparison are a product of odds multipliers e^{β_i} for each feature i present for the item. For ease of presentation, we report $e^{\beta_i}/(1 + e^{\beta_i}) - 0.5$, the likelihood that an item with feature i will win a comparison against an equivalent item without feature i, so it is directly comparable to the standalone measure we described above. Bonferroni-corrected p-values for each coefficient are reported in Appendix J.2.

A causal decomposition using minimal pairs to validate these coefficients is in Appendix B.2. To distinguish prompt-level variance from sampling noise, we also present a mixed-effects analysis in Appendix B.4.

Findings. Results for names are shown in Figure 1 (results for numbers and cities in Appendix J.2). Notice first that the preference values in Table 1 are sometimes larger and sometimes smaller than the logistic regression coefficients. This is to be expected, as the likelihood of winning when a feature is present may regress to the mean due to the noise of other preferences; on the other hand, the logistic regression coefficients may similarly regress towards 0.5 because the model may choose to move weight to more predictive but highly correlated other features. Nonetheless, some patterns emerge:

Table 1: Summary of model preferences across different features for the "Names" and "Numbers" domains. "Preference" here means the probability that an item i wins a comparison against an item j, given that the feature favors i over j.
"Random Pref Order" represents a simple baseline (for the full Monte Carlo version of this baseline, see Appendix B.1). For each of these 3 columns, we assign a random preference ordering over all items in the domain, to simulate some unknown "true" preference ordering that could be driving the observed preferences. We then decide all matchups according to that ordering, and then compute win rates for each feature.
Ties (cases where the feature favors neither item) are excluded. Features where over half of the matchups were excluded are italicized, with the percentage of active (non-excluded) matchups listed in parentheses after the feature name. Significance stars reflect Bonferroni-corrected p-values across $m = 183$ cell tests ($^{***}p_{\text{Bonf}} < 0.001$, $^{**}p_{\text{Bonf}} < 0.01$, $^{*}p_{\text{Bonf}} < 0.05$).

Comparable	Metric	Gemini 2.5 Flash	Gemini 3 Flash	Claude Haiku 4.5	GPT-5 Mini	Qwen 3.5 35B	Random Pref Order 1	Random Pref Order 2	Random Pref Order 3
Numbers	Higher	0.70***	0.78***	0.79***	0.76***	0.72***	0.48	0.55	0.41*
	Higher Order Mag (14%)	0.58	0.74***	0.75***	0.60	0.59	0.54	0.54	0.32
	Prime (28%)	0.64**	0.60	0.62	0.73***	0.72***	0.50	0.45	0.48
	Asc Digit Order (31%)	0.44	0.40	0.40	0.41	0.43	0.39	0.39	0.50
	Higher Sum Digits	0.57	0.62***	0.64***	0.62***	0.57	0.47	0.55	0.48
	Mult 10 (22%)	0.52	0.43	0.37	0.37	0.48	0.48	0.61	0.48
	Palindrome (17%)	0.59	0.53	0.50	0.56	0.63	0.38	0.32*	0.59
	Mult 5 (35%)	0.51	0.45	0.41	0.40	0.46	0.56	0.61	0.56
	Mult 2	0.46	0.45	0.44	0.41	0.44	0.39*	0.51	0.50
	First Presented	0.59***	0.50	0.63***	0.48***	0.57***	0.50	0.50	0.50
	Desc Digit Order (43%)	0.54	0.52	0.52	0.55	0.54	0.55	0.56	0.59
	Has Repeat Digits (41%)	0.55	0.54	0.53	0.51	0.53	0.43	0.37*	0.52
Names	*Higher Rare Letter Frac (6%)*	0.82	0.97***	0.68	0.94***	0.78	0.50	0.33	0.42
	Starting Vowel (31%)	0.76***	0.74***	0.60	0.74***	0.75***	0.48	0.40	0.45
	Higher Frac of 2020 Babies	0.72***	0.75***	0.65***	0.67***	0.75***	0.51	0.47	0.50
	Higher Frac of 2000 Babies	0.69***	0.71***	0.62***	0.63***	0.70***	0.48	0.51	0.55
	Higher Vowel Frac	0.62***	0.62**	0.54	0.66***	0.61**	0.55	0.50	0.47
	Higher Frac of 1960 Babies	0.40**	0.34***	0.42***	0.34***	0.37***	0.48	0.49	0.49
	Higher Bouba Frac	0.60*	0.59	0.53	0.62**	0.60*	0.58*	0.48	0.52
	Higher Syllable Heur	0.56	0.64*	0.50	0.64*	0.57	0.59	0.49	0.48
	Ending Vowel (41%)	0.55	0.56	0.46	0.65*	0.50	0.55	0.46	0.55
	First Alphabetically	0.59*	0.54	0.50	0.54	0.55	0.51	0.47	0.52
	Longer Length	0.48	0.56	0.51	0.53	0.50	0.56	0.56	0.49
	Higher Unique Char Frac	0.48	0.48	0.51	0.45	0.51	0.51	0.47	0.48
	Higher Census Freq	0.52	0.49	0.51	0.45	0.50	0.48	0.59*	0.47
	Higher Frac of 1980 Babies	0.54	0.50	0.51	0.44	0.52	0.48	0.55	0.48
	First Presented	0.52**	0.51	0.82***	0.51	0.57***	0.50	0.50	0.50

Across the evaluated models and datasets, several preferences exhibit highly consistent behaviors, while others show stark divergences depending on the domain or specific model family. Among consistent trends, models show a strong and statistically significant preference for larger numbers (win rates of 0.70–0.79) and prime numbers (0.64–0.73) in the numbers domain, as well as names popular in recent decades (e.g., 2020 Babies) over historical ones (1960 Babies, win rates < 0.42). Phonetically, a shared preference emerges across both names and fictional cities for starting vowels (up to 0.76 for names, 0.64 for most cities), higher vowel fractions, and soft "bouba" sounds (> 0.57).

Conversely, other features reveal sharp divergences between domains or models. While models are indifferent to the length of human names, they exhibit an aversion to longer fictional city names (win rates < 0.43). In numbers, while most models are indifferent to divisibility, Claude Haiku 4.5 and GPT-5 Mini significantly avoid multiples of 10 (0.37), 5 (0.40), and even numbers (0.41). Presentation order effects ("first presented") also diverge significantly: names exhibit a consistent primacy preference across all models (0.82 for Haiku), while cities show a strong recency preference for Gemini 3 Flash (0.47) and GPT-5 Mini (0.39). Finally, list position in cities reveals a hierarchical contrast, where models generally prefer later-generated categories overall (win rates < 0.44 for Gemini models) but prefer earlier items within those specific categories (0.56 for Gemini 3 Flash).

3.2 Prompt variation

We move now to a study of the difference in outcomes as we move from one prompt to another. For each of our three domains, we pick a set of important features. Then over the fifty prompts, we compute separately for each prompt the *winning fraction*; that is, fraction of comparisons won by instances having the feature against items without the feature. We then show the distribution

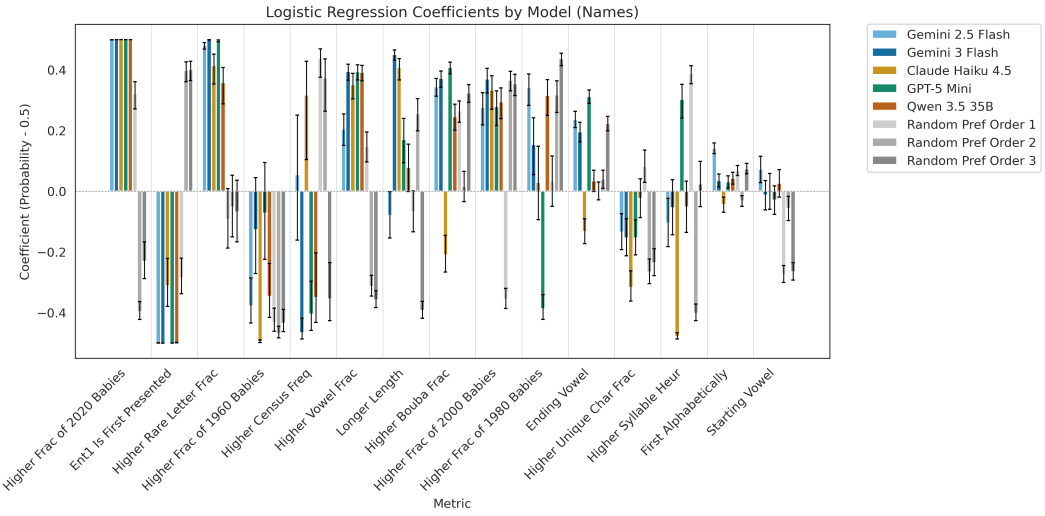


Figure 1: Second preference measure: logistic regression coefficients (with 95% confidence intervals), predicting whether an item wins a comparison against another. Presented along with overall win fractions from Table 1, both scaled to the $(-0.5, 0.5)$ range. Positive values mean a feature is positively correlated with likelihood of winning.

of winning fraction. A highly concentrated distribution means that all prompts behave similarly with respect to this preference. A highly dispersed distribution means that some prompts cause the model to prefer items with this feature, while other prompts prefer items *without* this feature. This computation therefore gives us a sense of the amount of influence a prompt can have in activating or supressing the particular preference.

The results are shown in Figure 2. First, many of the graphs seem to have a bell curve shape. This is natural, as each prompt's winning fraction is the result of $200 * 2$ random trials, so there will be some expected dispersion. The first question is whether the dispersion of the winning fraction can be explained by this randomness, even under the null hypothesis that the prompts themselves do not operate to strengthen or weaken a preference. To evaluate this question, we computed Levene's statistic to ask whether the winning fraction distributions are consistent with the expected natural noise. In all the twenty-one distributions in the figure, this null hypothesis is refuted with $p < 0.001$ (and typically far smaller). So we conclude that the choice of prompt does indeed control the activation or suppression of each preference.

The dispersion in the distribution of winning fractions for numbers is significantly higher for most preferences than for names and cities. We give some more details on this observation below, but we may conclude that particular domains may be more or less susceptible to prompt influence, though all three domains we study show significant influence.

Prompts also show some correlation between datasets, as seen in Figure 3; if an individual prompt shows a preference towards the first-presented option for names, for example, it is likely to show the same preference for cities. Similar cross-dataset correlations are observed for Gemini 3 Flash and GPT-5 Mini, as detailed in Appendix J.4.

Some prompt preferences are more interpretable than others. For example, the most extreme outlier visible on the "first sorted" plot in Fig. 3 is prompt 23, shown in orange: "Regarding these two cities/names/numbers, which one would you put first?" It's perhaps not surprising that the model overwhelmingly answers that question with the option that comes first alphabetically, in the case of names or cities; or first on a number line, in the case of numbers. But another occasional outlier, prompt 18 (in green), is less clear: "Please identify your preferred option among these two cities/names/numbers."

Different prompts also show differences in which factors are most salient in the model's thinking, as identified by feeding each comparison's thinking trace to a second LLM (Appendix H).

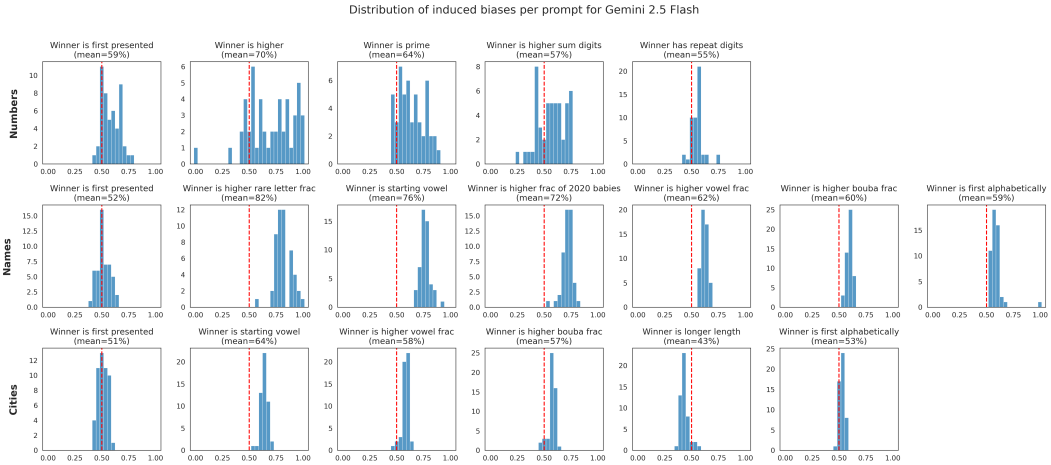
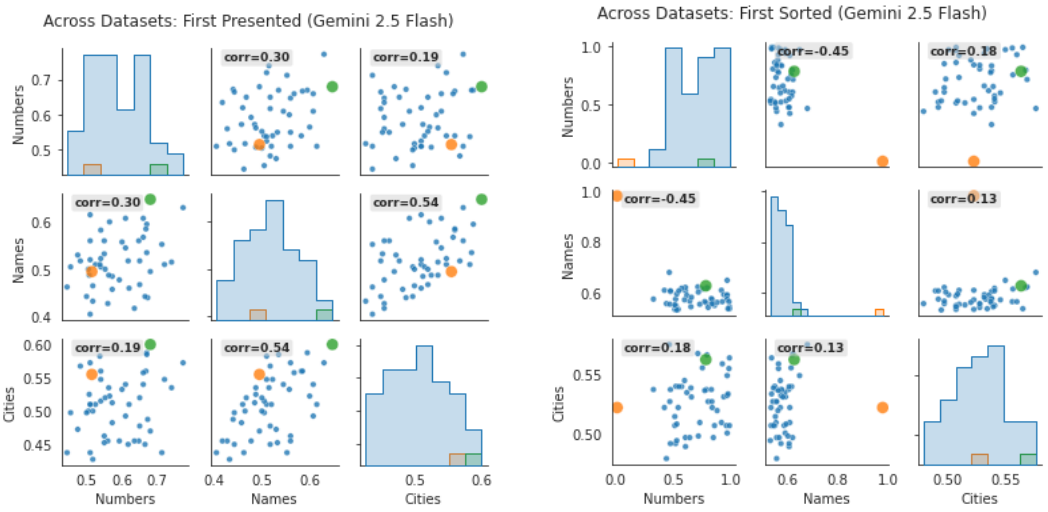


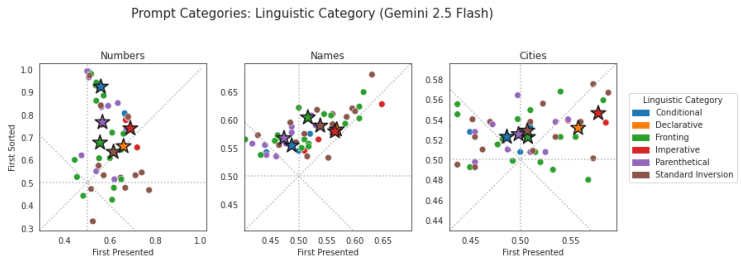
Figure 2: Prompt variation: distribution of per-prompt means for each preference (probability item i wins a comparison against an item j where the feature favors i over j).



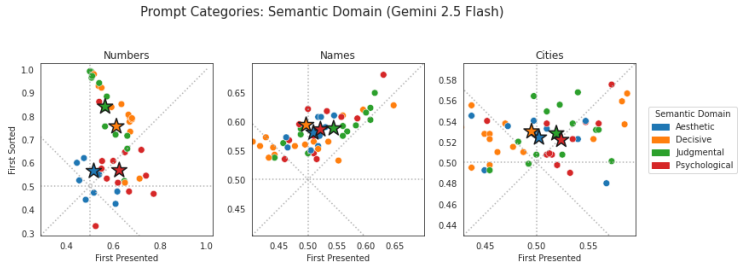
(a) First Presented

(b) First Sorted

Figure 3: Robustness of prompt variation: correlation of prompt preference between datasets on two different metrics. "First sorted" refers to first alphabetical for cities and names, and higher number for numbers. Prompts 23 and 18 are highlighted in orange and green respectively as illustrative outliers (see text for detailed discussion).



(a) Prompt preferences by "linguistic category", as defined by an LLM (see Table 5 for definitions).



(b) Prompt preferences by "semantic domain" of the prompt's verbs and nouns, as defined by an LLM (see Table 6 for definitions).

Figure 4: "First-sorted" (alphabetical or numerical) vs."first-presented" preferences per prompt, showing prompts colored by various categorizations. The centroid of each category is shown as a star.

Some prompt preferences may stem from more subtle features. For example, Figure 4a shows that prompts with "imperative" structure seem to have a stronger preference towards the first-presented option, while sentences with parentheticals may bias the model more towards the second-presented option, at least for words. Figure 4b shows that prompts using words judged to belong to the "Decisive" or "Judgmental" domains show stronger preferences than questions deemed "Aesthetic" or "Psychological". (Categories were determined by a second LLM which was given only the raw list of 50 prompts, from a single categorization run; we consider these analyses exploratory. See Appendix D for category definitions and a table of prompts.)

3.3 Ideation

We note that while these endogenous category structures are particularly pronounced in lists generated by Gemini 2.5 Flash, they represent a common property of how LLMs serialize long lists of subjective candidates. Newer SOTA model families (such as Gemini 3 or GPT-5) often employ strong implicit structural priors—sometimes explicitly partitioning a requested list of 50 items into structured sub-categories (e.g., 5 themed categories of 10 items each) even when not requested. In this sense, the category-based "spillover" preference we observe is not merely an artifact of an older model, but rather suggests that structural serialization preferences persist, and in some cases become more formalized, in newer architectures.

Recall that the stylized *ideation* activity we define is to produce an initial list of candidate options, as described in Section 2. The candidates produced by this list-creation task are then the objects that get compared in the comparison task that we have been studying in this section.

These two tasks are related, not just by the fact that the list-creation task produces the objects for the comparison task, but by the fact that more fine-grained properties of the list-creation task spill over into the outcome of the comparison task. We have seen a first example of this in the way that the "first-generated" attribute is a source of preference: the model has a preference for fictional city names that were generated earlier in the list of candidates rather than later. Here we discuss a more subtle form of spillover, in which the nature of the list itself can create a set of *endogenous* categories that introduce new forms of preference.

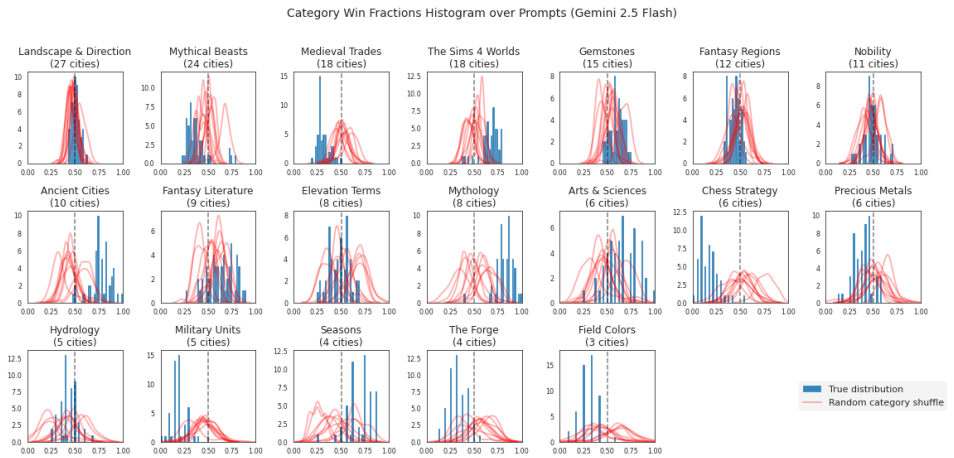


Figure 5: Category preference: distributions over per-prompt win probabilities for each category (probability that a city from a given category wins against a city from any other category). Plotted against baseline (red lines) of the same distribution, calculated for 10 different random assignments of cities to categories. See Table 3 for a list of city names and categories.

This phenomenon is domain-dependent, but it shows up clearly with the Gemini 2.5 model and the domain of city names. Inspection of the list of 200 city names makes clear that contiguous cities have often been generated using a common theme; for example, cities 66 through 69 on the list are Springdale, Summerwind, Autumnwood, and Winterfell, suggesting that once the model embarked on this set of cities, it ran through the list of seasons. In fact, essentially the entire list of 200 cities can be broken into contiguous runs of thematically related objects, with the theme eventually pivoting to a new theme, and a new run beginning. Table 3 in the Appendix shows a Gemini-produced breakdown of the 200 cities into categories, in continguous runs of city names that were produced consecutively. Note that some of these categories (like city names beginning with seasons) are more obvious, whereas others require more specialized knowledge, even if they're equally clear-cut once Gemini surfaces this knowledge (like the run of 18 city names that come from The Sims 4 [The Sims Wiki, 2022]).

These are the endogenous categories we referred to above, since they are only defined in terms of the actual choices of city names generated by the model; this is in contrast to attributes such as vowel fraction or syllable count which can be defined exogenously before the city names are seen. An additional set of preferences can be defined in terms of these endogenous categories: in the comparison task, we find that the model consistently prefers names from some endogenous categories over others. Figure 5 details this category-based preference: for each category of city names identified by Gemini, the histogram shows how much each prompt shows a preference toward or against names from this category, based on win-rate against all other names. Relative to a random baseline (in which we randomly shuffle category memberships), certain categories are heavily favored across essentially all prompts (including ancient cities, mythology, and cities from The Sims 4), and certain categories are heavily disfavored across essentially all prompts (including city names based on medieval trades and chess strategy). (It is interesting to note that favored categories seem to overlap heavily with names that exist in the "real world", vs. names that appear to be invented from scratch by the generating LLM.) As a robustness check, we repeated the analysis after repeatedly re-categorizing shuffled lists, with broadly similar results (Appendix B.3).

In this way, we see that not only is there fine-grained structure in some of the ideation lists that are produced—in this case, the city names created by the model—but this fine-grained structure becomes a source of further preferences in the subsequent comparisons.

4 Conclusion and Limitations

In this paper we have studied two stylized tasks that capture aspects of the ideation and pruning phases in subjective decision-making: ideation inherently involves the creation of a list of options and

pruning is built on the ability to decide whether option A or B is preferred. We study the preferences that arise as empirical regularities in this comparison process, and find that the strength and even the direction of a number of these preferences are prompt-dependent.

And we identify some spillovers between the ideation and comparison tasks, in the following sense. When the model needs to create a long list of candidate options, it does so by iterating through high-level thematic categories and creating several members from each category. This in turn creates new preferences from the ideation phase, in which options from certain categories are preferred to options from other categories, again in a way that is mediated non-trivially by the prompt.

Limitations. As with all research in this area, it is useful to point out some of the limitations of our work. First, we chose a range of models that are at or near the state-of-the-art frontier, but our selection from within this set was oriented toward using smaller models where the cost of the computation would be lower. We sampled different styles of models, and used both models are more open and those that are more closed, including models from several of the largest providers of LLMs. A more extensive study could try a wider range of models, including some that are more powerful and expensive than what we have included here.

Similarly, we have studied a particular set of subjective comparison tasks designed to represent different aspects of the problem. Our domains are short lexical items chosen for controlled feature extraction, and transfer to longer-form items remains an open research question. The full set of subjective comparison tasks is vast, and subsequent work could easily explore further problems in this style. Our work here will be relevant to such future studies, in that it provides a road-map for the different categories of questions that can be asked in such settings, and set up in a way that can be adapted to other subjective comparison tasks.

Further directions. An interesting research direction is to extend the categorization of prompts further: we have found that structure variations in the prompt can lead to systematic shifts in the preferences, but we do not have a complete taxonomy of these effects. And it would also be interesting to extend our framework to more complex subjective tasks, such as the ideation and pruning of longer texts, where there is even more structure for the model to work with and potentially to use as the basis for its preferences.

References

Morris Alper and Hadar Averbuch-Elor. Kiki or bouba? sound symbolism in vision-and-language models. In *NeurIPS*, pages 78347–78359, 2023.

Anthropic. System card: Claude Haiku 4.5. Technical report, Anthropic, October 2025. URL https://www.anthropic.com/claude-haiku-4-5-system-card.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner. In *NeurIPS*, 2023.

Alex Bellos. *The Grapes of Math: How Life Reflects Numbers and Numbers Reflect Life*. Simon and Schuster, 2014.

Katy Blumer, Kate Donahue, Katie Fritz, Kate Ivanovich, Katherine Lee, Katie Luo, Cathy Meng, and Katie Van Koevering. An Abundance of Katherines: The game theory of baby naming. *arXiv*, 2404.00732, 2024.

Brian Brant. Couple uses ChatGPT to help name their newborn son, one of the first babies born in 2026. *People Magazine*, January 2026. URL https://people.com/couple-uses-chatgpt-help-name-their-newborn-son-11889914.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2507.06261, 2025.

Joshua Comenetz. First name data from the 2020 census. 2020 Census Briefs C2020BR-13, U.S. Census Bureau, April 2026. URL https://www2.census.gov/library/publications/decennial/2020/c2020br-13.pdf. Data available at: https://web.archive.org/web/20260414144524/https://www.census.gov/topics/population/genealogy/data/2020_names.html.

Julian De Freitas, Gideon Nave, and Stefano Puntoni. When used correctly, LLMs can unlock more creative ideas. *Harvard Business Review*, December 2025.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction–CADE*, pages 625–635, 2021.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Lei Ding, Yang Hu, Nicole Denier, Enze Shi, Junxi Zhang, Qirui Hu, Karen D. Hughes, Linglong Kong, and Bei Jiang. Probing social bias in labor market text generation by ChatGPT: A masked language model approach. In *NeurIPS*, 2024.

Peter Eades, Xuemin Lin, and William F Smyth. A fast and effective heuristic for the feedback arc set problem. *IPL*, 47(6):319–323, 1993.

Mathilde Fort and Jean-Luc Schwartz. Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific Reports*, 12(1): 19172, 2022.

Lance Calvin Lim Gamboa, Yue Feng, and Mark G. Lee. Social bias in multilingual language models: A survey. In *EMNLP*, pages 27857–27880, 2025.

Google DeepMind. Gemini 3 flash model card. Technical report, Google DeepMind, December 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Richard D. Harroch and Dominique A. Harroch. Name your startup using these 7 ChatGPT prompts. *AllBusiness*, 2023. URL https://www.allbusiness.com/name-your-startup.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Comput. Sci.*, pages 65–75, 2025.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *ICLR*, 2024.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. In *NeurIPS*, 2022.

Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. "global is good, local is bad?": Understanding brand bias in LLMs. In *EMNLP*, pages 12695–12702, 2024.

Mohammad Aflah Khan, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna Gummadi, Manish Gupta, and Abhilasha Ravichander. In agents we trust, but who do agents trust? latent source preferences steer LLM generations. In *ICLR*, pages 115154–115207, 2026.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *ICLR*, 2024.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *EAMT*, pages 193–203, 2023.

Katherine Van Koevering and Jon Kleinberg. How random is random? Evaluating the randomness and humaness of LLMs' coin flips. *arXiv*, 2406.00092, 2024.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *ACL (Findings)*, pages 517–545, 2024.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *EMNLP*, pages 2757–2791, 2025a.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *ICLR*, 2024.

Sitong Li, Stefano Padilla, Pierre Le Bras, Junyu Dong, and Mike Chantler. A review of LLM-assisted ideation. *arXiv*, 2503.00946, 2025b.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv*, 2203.02155, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *NeurIPS*, 2024.

Laura Picklesimer. Using AI to develop and workshop your characters. *Writer's Digest*, September 2023. URL `https://www.writersdigest.com/write-better-fiction/using-ai-to-develop-and-workshop-your-characters`.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv*, 2206.05802, 2022.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *ICLR*, 2024.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large language models are not yet human-level evaluators for abstractive summarization. In *EMNLP (Findings)*, pages 4215–4233, 2023.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by LLMs. In *IJCNLP*, pages 292–314, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv*, 2409.04109, 2024.

Aaditya Singh et al. OpenAI GPT-5 system card, 2025. URL `https://arxiv.org/abs/2601.03267`.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *ACL*, pages 15725–15788, 2024.

Alaina N. Talboy and Elizabeth Fuller. Challenging the appearance of machine intelligence: Cognitive bias in LLMs and best practices for adoption. *arXiv*, 2304.01358, 2023.

The Sims Wiki. Worlds in The Sims 4. `https://web.archive.org/web/20260112123046/https://sims.fandom.com/wiki/Category:Worlds_in_The_Sims_4`, 2022.

John N Towse, Tobias Loetscher, and Peter Brugger. Not all numbers are equal: preferences and biases among children and adults when generating random sequences. *Frontiers in Psychology*, 5: 19, 2014.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *ACL*, pages 9440–9450, 2024a.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *ICLR*, 2024b.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *ICLR*, 2026.

Minghao Wu and Alhan Fikri Aji. Style over substance: Evaluation biases for large language models. In *COLING*, pages 297–312, 2025.

Keyu Zhao, Weiquan Lin, Qirui Zheng, Fengli Xu, and Yong Li. Deep ideation: Designing LLM agents to generate novel research ideas on scientific concept network. *arXiv*, 2511.02238, 2025.

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

Table of Contents for the Supplemental Material

Contents

A Related Work **15**

B Supplemental Robustness Checks and Statistical Validation **16**

B.1 Monte Carlo Permutation Test 16

B.2 Causal Evaluation via Minimal Pairs 16

B.3 City Categories: Robustness to List Order (Shuffling) 16

B.4 Mixed-Effects Model for Prompt Robustness 16

B.5 Model Consistency: Feedback Arc Sets 18

C Comparable objects **20**

C.1 Cities and categories 20

D Prompt questions for comparisons: set of 50 **21**

E Other prompts used **23**

E.1 Subjective Preference Comparison Prompt Template 24

F Summary Table of Preference Features **26**

G Model Configurations and Citations **26**

H Factors named in thinking traces **26**

I Broader Impacts **27**

J Extended Results for Other Model Families, Datasets, and Model Hyperparameters **27**

J.1 Tables of Results for Other Datasets and Model Settings 27

J.2 Regression Coefficients for Other Datasets 28

J.3 Prompt Variation Histograms 28

J.4 Prompt Robustness Across Datasets 29

J.5 Prompt Behavior Correlations Across Models 32

A Related Work

One of the main literatures that our work connects to is the line of research on LLMs as judges or evaluators. A number of recent papers document specific statistical biases or "preferences", and one of our goals in this work is to study the interplay of all these preferences, and to identify new ones systematically, for the types of subjective brainstorming tasks we consider here. In particular, Wang et al. [2024a] demonstrate the effects of positional bias, which is one of the effects we study here, and Turpin et al. [2023] show that chain-of-thought text produced by a model does not necessarily correspond to the basis for its decision. For biases that involve other contexts not directly related to our framework, Sharma et al. [2024] show the effects of sycophancy, in which agreement with the user is prioritized, and Koo et al. [2024] and Panickssery et al. [2024] examine self-preference, in which a model has a tendency to prefer its own outputs. In more factual settings, again distinct from the subjective questions we consider here, Wu and Aji [2025] show that LLMs can view grammatical or stylistic errors as more severe than factual errors, and Shen et al. [2023] consider longer-form tasks such as abstractive summarization.

Other lines of work relate LLM preferences to corresponding human behavioral and cognitive biases. Hu et al. [2025] and Talboy and Fuller [2023] study this in the context of in-group bias, with Ding et al. [2024] considering gendered language and Gamboa et al. [2025] considering multilingual bias. Khan et al. [2026] find that LLMs systematically prefer some information sources over others, and Kamruzzaman et al. [2024] find that they prefer global over local product brands. Koevering and Kleinberg [2024] study the failure of even state-of-the-art LLMs to produce pseudo-random sequences, instead exhibiting the types of biases humans display when they try to generate random outputs. Jones and Steinhardt [2022] suggest using this analogy with human behavioral and cognitive biases as a research strategy for probing LLM behavior.

Motivated by the abundance of these biases, a parallel line of work has proposed frameworks for measuring biases and for evaluation in the presence of biases. Li et al. [2025a] and Bai et al. [2023] provide overviews of these activities, and specific frameworks include Prometheus [Kim et al., 2024], PandaLM [Wang et al., 2024b], and Auto-J [Li et al., 2024]. These issues have also been considered in specific domains such as assessing the quality of translations [Kocmi and Federmann, 2023] and for the production of natural-language critiques [Saunders et al., 2022].

There is also a robust literature on human biases and preferences, including preferences along many of the dimensions we study (though the methodology in each of these is very different from the pairwise forced-choice setup in our experiments). Towse et al. [2014] find that humans prefer smaller numbers when generating random sequences - an opposite preference to what we observe. Bellos [2014], in an informal online survey on favorite numbers, also finds that humans lean toward smaller numbers, and that like the LLMs in our study, they prefer primes and avoid round numbers. The LLMs in our study also echo humans in a preference for "bouba" sounds [Fort and Schwartz, 2022] (and Alper and Averbuch-Elor [2023] showed that text-to-image models also reproduce the bouba / kiki effect).

For the ideation portion of our work, in which lists of candidate options are produced before being refined, Li et al. [2025b] provide a general survey of approaches to this problem. There have been discussions of this process in a number of specific domains in both the academic literature and in general-audience writing, as suggested in the introduction, including for business ideas [De Freitas et al., 2025], research ideas [Zhao et al., 2025, Si et al., 2024], baby names [Brant, 2026, Blumer et al., 2024], names for new companies and new products [Harroch and Harroch, 2023], and names for characters when creating works of fiction [Picklesimer, 2023]. As noted at the outset, our work focuses on the shorter-form versions of the task, to be able to isolate properties of the options more concretely.

B Supplemental Robustness Checks and Statistical Validation

B.1 Monte Carlo Permutation Test

To test whether our results could be explained by some other, hidden factor that was the actual source of a model's preferences, we conducted a Monte Carlo permutation test. For each domain, we repeated the following procedure 25,000 times: we randomly shuffled the list of entities to simulate some hidden "true" preference ranking. We repeated the exact same pairwise comparisons, resolving each comparison according to which entity came first in the simulated preference ranking. Then we recorded the win rate for each feature based on the resolved comparisons. After repeating this 25,000 times, we had a distribution of win rates for each feature, representing the distribution of win rates you might expect for that feature if the model's true preference were based on something unrelated to that feature.

Figure 6 shows the results of this procedure. The histogram of Monte Carlo win rates is shown, and the 95% confidence interval is shown in gray. The actual win rates observed by asking the LLMs (the values shown in Table 1) are shown as vertical colored lines.

For more details, see Section 4 of `https://anon-for-reviews.github.io/neurips2026/report_statistics.html`.

B.2 Causal Evaluation via Minimal Pairs

To disentangle confounding features (e.g., the correlation between "Higher" value and "Higher Sum of Digits" in the Numbers dataset), we conducted a causal evaluation using a "minimal pairs" framework. For each target feature, we generated pairs of items where one feature varied while all other identified features were held as constant as possible. To ensure spatial coverage, we used K-Means stratification to sample representative pairs across the non-target covariate space.

The learned logistic regression model fit on the original observational data (Section 2) accurately predicts causal win rates in these experimental settings:

- **Numbers Domain**: Pearson $r = 0.941$, Mean Absolute Error (MAE) = 0.051.
- **Names Domain**: Pearson $r = 0.884$, MAE = 0.078.

The most important preference drivers remain strong and invariant across methodologies. For instance, in the Numbers domain, the *Mult 10* feature shows a causal win rate increase of +31% to +35% relative to a baseline of 0.5. In the Names domain, the *Higher Frac of 2020 Babies* shows a causal increase of +25% to +36%. Full data and interactive comparisons for this causal decomposition are available at `https://anon-for-reviews.github.io/neurips2026/report_minimal_pairs.html`.

B.3 City Categories: Robustness to List Order (Shuffling)

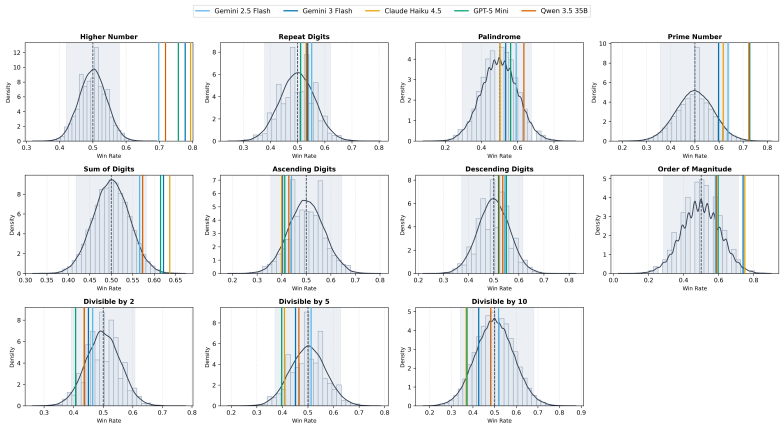
To ensure that the endogenous categories identified in Section 3.3 are not simply artifacts of the original list position, we performed a robustness check by shuffling the list of 200 cities five times. For each shuffle, we asked an LLM to re-categorize the cities, both with and without an "order hint" (the original generation rank).

The resulting categorizations showed high semantic agreement with the paper's original 19-category labeling (Gemini 3 Pro), with an Adjusted Rand Index (ARI) consistently indicating robust recovery of the thematic clusters regardless of list order (ARI range: 0.50–0.65). Repeating the analysis from Figure 5 with each categorization also gives broadly similar results. Full data, prompts, and interactive visualizations for these shuffled re-categorizations are available at `https://anon-for-reviews.github.io/neurips2026/report_categorization.html`.

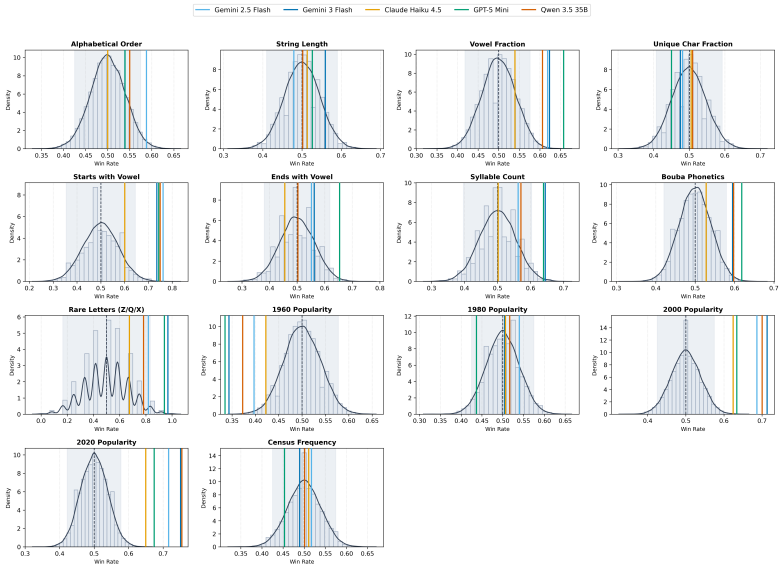
B.4 Mixed-Effects Model for Prompt Robustness

To distinguish prompt-level variance from sampling noise, we employed a Generalized Linear Mixed-Effects Model (GLMM) with crossed random effects for both prompts and item-pairs.

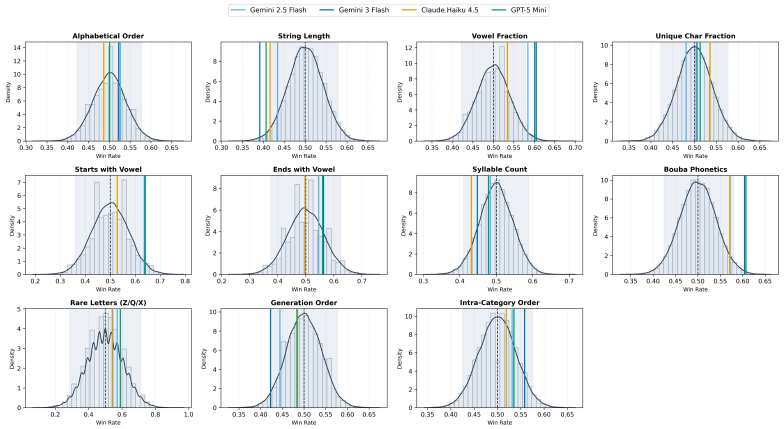
Levene's test for variance comparison confirms that prompt variance significantly exceeds pure sampling noise for 12 of 14 model/domain combinations; 9 of 14 at ($p < 0.001$). Variance decompo-



(a) Numbers



(b) Names



(c) Cities

Figure 6: Monte Carlo permutation test results for the Numbers, Names, and Cities datasets. Histograms for each feature represent the win rates for that feature on 25,000 random preference permutations (see Section B.1), and gray shading represents 95% confidence intervals on the Monte Carlo win rates. Each vertical line represents the win rate for that feature for a specific LLM, as in Table 1.

sition (Table 2) indicates that prompt variance typically accounts for 1–20% of the total observed variance in win rates, while item-pair variance accounts for 3–43%. This confirms that prompt sensitivity is a robust and significant property of model behavior, rather than an artifact of finite sampling. Full data and interactive variance decompositions for all models and datasets are available at `https://anon-for-reviews.github.io/neurips2026/report_statistics.html`.

Table 2: Variance component breakdown and Levene's test for prompt-level sensitivity. Total per-prompt win rate variance is decomposed into Prompt Variance (σ^2_{prompt}), Item-Pair Variance (σ^2_{pair}), and Sampling/Residual Noise. Levene's W statistic tests whether prompt variance significantly exceeds pure sampling noise.

Dataset	Model	σ^2_{prompt} (%)	σ^2_{pair} (%)	Noise (%)	W Stat	p-val	Sig.
Numbers	Gemini 2.5 Flash	0.006 (10.2%)	0.003 (4.4%)	0.051 (85.4%)	120.82	<.001	***
Numbers	Gemini 3 Flash	0.001 (3.6%)	0.005 (13.4%)	0.032 (83.0%)	0.24	.622	n.s.
Numbers	Claude Haiku 4.5	0.010 (19.5%)	0.005 (9.9%)	0.036 (70.5%)	200.09	<.001	***
Numbers	GPT-5 Mini	0.004 (8.6%)	0.002 (4.1%)	0.043 (87.3%)	7.36	.007	*
Numbers	Qwen 3.5 35B	0.003 (4.5%)	0.003 (4.6%)	0.051 (90.9%)	18.85	<.001	***
Names	Gemini 2.5 Flash	0.003 (4.8%)	0.002 (3.2%)	0.059 (92.0%)	27.99	<.001	***
Names	Gemini 3 Flash	0.002 (5.0%)	0.006 (16.4%)	0.029 (78.6%)	3.01	.083	n.s.
Names	Claude Haiku 4.5	0.004 (6.7%)	0.026 (42.2%)	0.031 (51.1%)	19.02	<.001	***
Names	GPT-5 Mini	0.005 (9.5%)	0.005 (9.5%)	0.040 (81.0%)	5.94	.015	*
Names	Qwen 3.5 35B	0.004 (7.2%)	0.006 (9.3%)	0.050 (83.5%)	55.27	<.001	***
Cities	Gemini 2.5 Flash	0.002 (3.0%)	0.004 (5.6%)	0.059 (91.4%)	11.25	<.001	***
Cities	Gemini 3 Flash	0.001 (1.3%)	0.011 (27.7%)	0.029 (71.0%)	7.45	.007	*
Cities	Claude Haiku 4.5	0.011 (16.2%)	0.012 (17.0%)	0.045 (66.8%)	116.17	<.001	***
Cities	GPT-5 Mini	0.004 (7.1%)	0.006 (11.0%)	0.047 (81.9%)	16.84	<.001	***

B.5 Model Consistency: Feedback Arc Sets

We ask now whether the LLM provides *consistency* in its comparisons. In a perfectly consistent setting, the items could be ordered from best to worst, and each comparison question to the LLM would return a result favoring the better of the two options. We study the degree of consistency as follows. Given N items, there are $2\binom{N}{2}$ possible ordered pairs. We may ask the model to produce a comparison of each such ordered pair, and add an edge to a multigraph for each outcome directed from the winner to the loser. For a perfectly consistent judge, the vertices of the graph can be numbered so that every edge goes from a lower number to a higher number. For inconsistent judges, though, we can number the vertices so that as many edges as possible go from a lower number to a higher number; the remaining edges that go from nominally worse to better items are called the *Minimum Feedback Arc Set (FAS)* of the graph, and the number of such edges is a measure of the inconsistency of the judge.

We compute the sizes of these sets in two settings. First, we consider a "wisdom of crowds" approach in which, for every ordered pair, each prompt votes for a winner and the majority prevails. Second, we consider a per-prompt setting in which each individual prompt considers each pair in each order, compares them, and outputs the winner. The approximate minimum FAS is calculated using the techniques of Eades et al. [1993]. For comparison, we also calculate feedback arc sets for 50 random graphs (same edge set, but edge directions determined randomly).

Figure 7 shows the sizes of the FAS for both settings. For the wisdom of crowds approach combining all prompts, the FAS for numbers contains 17% of the edges, while for cities and names the FAS contains 9% and 7% of the edges respectively. Thus, the cross-prompt consensus formed in the numbers domain is only mildly consistent, while in the names domain it is strongly consistent. We note that this consistency analysis was performed on an older dataset with a different combinatorial design (every possible pair evaluated once in each order, rather than the random cycle pairing used in Section 2); for names, this older dataset also used an LLM-generated name list rather than the census-based list used in Section 3.1.

The per-prompt FAS sizes, on the other hand, span a wider range, especially for numbers. Every prompt provides judgements that are more consistent than random guessing. However, presentation order effects make the individual judgements inherently less consistent. For names, the per-prompt

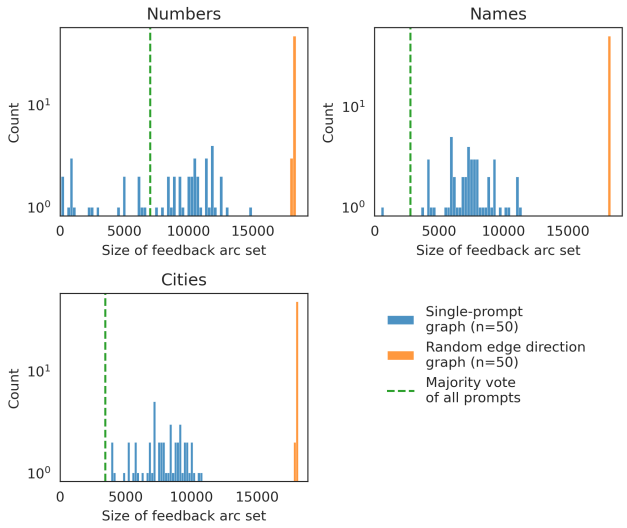


Figure 7: Model consistency: Sizes of minimum feedback arc sets (FAS) for directed preference graphs. The dotted line (green) shows the FAS size for the directed graph formed by majority vote over 50 prompts. Also shown are FAS sizes for each individual prompt's directed graph (blue) and for 50 graphs with random edge directions (orange).

FASes contain anywhere from 10–28% of the edges (except for a single outlier prompt that preferred alphabetic order almost all the time). The two most inconsistent prompts for names ($\approx 28\%$ edges) are: "Of the two names, which one would you personally endorse?" (Prompt 31) and "Please identify your preferred option among these two names." (Prompt 18).

For numbers, the situation is even more varied. Some prompts are almost perfectly consistent (Prompt 23, "Regarding these two numbers, which one would you put first?", has a FAS of only 0.2% of edges), while others are highly inconsistent (Prompt 46, "Which of the two numbers do you personally fancy?", has a FAS containing 1/3 of the edges). In conclusion, small prompt variations result in differences not just in outcomes, but in LLM reproducibility.

C Comparable objects

C.1 Cities and categories

See Table 3. Categorization was produced by an LLM (Prompt 2) that had no context on the process for generating the city names.

Category (range)	Cities
Landscape and Direction (1–28)	Silverglade, Veridian Bay, Whispering Pines, Azure Peak, Crimson Ridge, Dragon's Tooth, Shadowfen, Ironwood, Starfall, Sunstone, Mooncrest, Emberfall, Stormwind, Frostholm, Greenfield, Oakheart, Serpent's Coil, Griffin's Roost, Titan's End, Havenwood, Northridge, Southgate, Westwind, Eastbrook, Riverbend, Lakewatch, Seacliff, Stonehaven
Precious Metals (29–34)	Goldcrest, Silvercreek, Bronzeleaf, Copperhead, Ironhand, Steelheart
Gemstones (35–49)	Diamondfall, Rubyvale, Emerald City, Sapphire Isle, Amethyst Cove, Topaz Point, Obsidian Shore, Onyx Ridge, Pearl Bay, Coral Gables, Jasper Dale, Garnet Grove, Opal Springs, Turquoise River, Jade Mountain
Elevation Terms (50–57)	Zenith, Nadir, Apex, Pinnacle, Summit, Crestfall, Valedale, Hilltop
Hydrology (58–62)	Clearwater, Blackwater, Whitewater, Redwater, Bluewater
Field Colors (63–65)	Greenfield, Brownfield, Greyfield
Seasons (66–69)	Springdale, Summerwind, Autumnwood, Winterfell
Nobility (70–80)	Highgarden, Kingswood, Queensgate, Prince's Landing, Princess Anne's Harbor, Duke's Pass, Duchess's Crossing, Baron's End, Baroness's View, Knight's Watch, Squire's Rest
Military Units (81–85)	Archer's Point, Bowman's Ford, Swordsman's Creek, Spearman's Ridge, Axeman's Hollow
The Forge (86–89)	Hammerfall, Anvil Rock, Forgefire, Smeltertown
Medieval Trades (90–107)	Miner's Deep, Farmer's Field, Shepherd's Crook, Fisherman's Wharf, Hunter's Lodge, Trapper's Cabin, Lumberjack's Mill, Blacksmith's Forge, Carpenter's Bench, Mason's Quarry, Weaver's Loom, Potter's Wheel, Chandler's Wick, Cooper's Barrel, Fletcher's Arrow, Wainwright's Wheel, Cartwright's Path, Shipwright's Harbor
Arts and Sciences (108–113)	Glassblower's Furnace, Alchemist's Tower, Scribe's Library, Bard's Tale, Minstrel's Song, Jester's Court
Chess Strategy (114–119)	King's Ransom, Queen's Gambit, Bishop's Crossing, Rook's End, Pawn's Opening, Checkmate
Mythical Beasts (120–143)	Dragon's Breath, Griffin's Claw, Unicorn's Horn, Phoenix's Ashes, Wyvern's Sting, Kraken's Deep, Leviathan's Trench, Behemoth's Gate, Manticore's Lair, Harpy's Nest, Centaur's Grove, Satyr's Glade, Nymph's Spring, Dryad's Oak, Pixie's Hollow, Gnome's Grotto, Dwarf's Mine, Elf's Forest, Orc's Stronghold, Goblin's Cave, Troll's Bridge, Ogre's Swamp, Giant's Causeway, Titan's Rise
Mythology (144–151)	Olympus, Asgard, Valhalla, Elysium, Avalon, Shangri-La, El Dorado, Atlantis
Ancient Cities (152–161)	Babylon, Camelot, Carthage, Constantinople, Alexandria, Rome, Athens, Sparta, Thebes, Troy
Fantasy Literature (162–170)	Valyria, Westeros, Essos, Middle-earth, Narnia, Oz, Neverland, Wonderland, Fantasia
Fantasy Regions (171–182)	Aethelgard, Brimstone, Cedarwood, Duskwood, Evermore, Farshore, Grimwade, Hollowdale, Islesmere, Silverwood, Port Anson, Brighthaven
The Sims 4 Worlds [The Sims Wiki, 2022] (183–200)	Newcrest, Willow Creek, Oasis Springs, Magnolia Promenade, Granite Falls, Windenburg, San Myshuno, Forgotten Hollow, Brindleton Bay, Del Sol Valley, Strangerville, Sulani, Glimmerbrook, Britechester, Evergreen Harbor, Mt. Komorebi, Henford-on-Bagley, Copperdale

Table 3: City names and categories. Ranges give each category's positions in the original generated list; names are presented in the order they appeared in the list.

D Prompt questions for comparisons: set of 50

Table 4, below, lists all comparison question prompts and the relevant categories assigned by Gemini. Tables 5 and 6 show the definition Gemini gave for each category. Categories were invented by Gemini when asked to categorize the list of prompts by grammatical structures.

Table 4: Comparison question prompts used for all experiments. See Tables 5 and 6 for definitions of Ling(uistic) cat(egory) and Sem(antic) dom(ain).

Prompt index	Prompt	Ling. cat.	Sem. dom.	Tone
0	What is your preference between these two {numbers,names,cities}?	SI	D	Direct
1	Between these two {numbers,names,cities}, which do you favor?	F	J	Direct
2	Which of the two {numbers,names,cities} do you find more appealing?	SI	A	Direct
3	If you had to choose between these two {numbers,names,cities}, which would it be?	C	D	Soft
4	Do you have a stronger inclination towards one of these two {numbers,names,cities}?	SI	P	Soft
5	Which {number,name,city} out of these two options would you select?	P	D	Direct
6	Out of the two provided {numbers,names,cities}, which one do you like better?	F	P	Direct
7	Please indicate which of these two {numbers,names,cities} you lean towards.	I	P	Formal
8	Which of these two {numbers,names,cities}, in your opinion, is the better choice?	P	J	Soft
9	Considering these two {numbers,names,cities}, which one resonates more with you?	F	A	Soft
10	Given these two {numbers,names,cities}, which one would you rather select?	F	D	Direct
11	When considering these two {numbers,names,cities}, which do you see as more desirable?	C	A	Formal
12	Out of the current pair of {numbers,names,cities}, which is your top choice?	F	D	Direct
13	Tell me, which of these two {numbers,names,cities} do you prefer over the other?	I	D	Direct
14	Which {number,name,city}, from the duo presented, holds more appeal for you?	P	A	Formal
15	I'd like to know which of these two {numbers,names,cities} you find preferable.	D	J	Soft
16	If you were to rate these two {numbers,names,cities}, which would score higher in your estimation?	C	J	Formal
17	Between these two {numbers,names,cities}, which do you hold in higher regard?	F	J	Formal
18	Please identify your preferred option among these two {numbers,names,cities}.	I	D	Formal
19	Which {number,name,city}, when weighed against the other, would you ultimately favor?	P	J	Formal
20	Which {number,name,city}, from the pair, gets your vote?	P	D	Casual
21	Among the two {numbers,names,cities} presented, which one stands out to you?	F	A	Direct
22	Given these two options, which {number,name,city} would you opt for?	F	D	Direct
23	Regarding these two {numbers,names,cities}, which one would you put first?	F	D	Direct
24	Between the two {numbers,names,cities} provided, which one earns your approval?	F	J	Formal
25	Which of these two {numbers,names,cities} would you single out?	SI	D	Direct
26	Which of these two {numbers,names,cities} would you champion?	SI	D	Direct
27	Which of these two {numbers,names,cities} would you rather go with?	SI	D	Direct
28	What is your D choice between these two {numbers,names,cities}?	SI	D	Formal
29	Which {number,name,city}, given this duo, would you ultimately settle on?	P	D	Direct
30	Which of these two {numbers,names,cities} do you find more agreeable?	SI	P	Soft
31	Of the two {numbers,names,cities}, which one would you personally endorse?	F	J	Formal
32	Between the two {numbers,names,cities}, which makes a more positive impression on you?	F	P	Soft
33	Regarding these two {numbers,names,cities}, which one holds a higher subjective value for you?	F	J	Formal
34	Which {number,name,city}, from this pair, do you perceive as more favorable?	P	J	Formal
35	Out of these two {numbers,names,cities}, which one speaks more to your sensibility?	F	A	Soft
36	Which of the two {numbers,names,cities} do you feel a greater affinity for?	SI	P	Soft
37	Which of the two {numbers,names,cities} do you deem more fitting?	SI	J	Formal
38	For these two {numbers,names,cities}, which one commands your strongest positive reaction?	F	P	Formal
39	Which {number,name,city}, from the presented duo, do you find yourself drawn to more?	P	P	Soft
40	Which of these two {numbers,names,cities} do you find more commendable?	SI	J	Formal
41	Between these two {numbers,names,cities}, which elicits a more positive internal response from you?	F	P	Formal

Category	Abbrv.	Movement/Grammar Logic	Resulting Tone
Conditional	C	Uses "If/When" clauses to set a hypothesis.	Softened; polite; indirect.
Declarative	D	Formulated as a statement rather than a question.	Indirect; gentle; observational.
Fronting	F	Moves a phrase (often prepositional) to the start.	Analytical; context-first.
Imperative	I	Uses a command verb (e.g., "Please indicate").	Instructional; administrative.
Parenthetical	P	Inserts a qualifying phrase inside the sentence.	Nuanced; deliberate; subjective.
Standard Inversion	SI	Flips Subject and Auxiliary verb for a question.	Direct; neutral; efficient.

Table 5: "Linguistic categories" as defined by Gemini. Generated in a conversation about categorizing the prompts in Table 4 by grammatical structures.

Table 4 (continued)

Prompt index	Prompt	Ling. cat.	Sem. dom.	Tone
42	Which {number,name,city}, when presented side-by-side, do you find more compelling?	P	A	Direct
43	Regarding this duo, which {number,name,city} truly captures your imagination?	F	A	Soft
44	Which of these two {numbers,names,cities} do you consider to have a greater intrinsic value?	SI	J	Formal
45	Which {number,name,city} from this selection of two do you anticipate liking more?	SI	P	Soft
46	Which of the two {numbers,names,cities} do you personally fancy?	SI	P	Casual
47	From the pair provided, which {number,name,city} generates a more positive aesthetic response in you?	F	A	Formal
48	Which of these two {numbers,names,cities} truly captivates your interest?	SI	P	Soft
49	Considering these two {numbers,names,cities}, which one would you most readily adopt?	F	D	Direct

Category	Abbrv.	Key Words Used	Resulting Tone
Aesthetic	A	Appealing, resonates, aesthetic response, imagination	Soft, artistic, and subjective.
Decisive	D	Opt for, single out, settle on, champion	Action-oriented and final.
Judgmental	J	Deem, fitting, commendable, endorse, value	Serious, objective, and weighty.
Psychological	P	Affinity, inclination, sensibility, internal response	Analytical and introspective.

Table 6: "Semantic domain", as defined by Gemini: "four distinct 'vocal styles' based on the verbs used". Generated in a conversation about categorizing the prompts in Table 4 by grammatical structures.

E Other prompts used

Unless otherwise specified, all prompts were used with Gemini 2.5 Flash, using the standard Gemini API.

For the comparison question prompts (the "question prompt" in the template in Appendix E.1), see Appendix D.

```
You are an expert prompt engineer. You need to generate 10
    new prompts that represent different ways to ask a model for the same
    thing. The base prompt you are supposed to expand is: Which of these two
        names do you prefer?.
    {,In addition to the base prompt, another model has
        already produced an expanded set of prompts
        listed here: <begin generated prompts> {{already_generated_prompts}}
        <end generated prompts>}
      Please output 10 new prompts, one per line, such that
    each new prompt has the same meaning as the base prompt. Do not number
    the prompts; simply output each on its own line. Each new
    prompt should mean the same thing as the base prompt. However, each new
    prompt should be written differently from the base prompt, the
    expanded prompts, and any new generated prompts you have
    already produced.
```

Prompt 1: Expansion for set of 50 prompts.

```
User: Here is a list of city names. Please fill in 'Category Name' with an
    appropriate category. Keep the set of categories to 8-12 categories, and try
     to keep the sizes of the categories balanced. However, prioritize in this
    order: clear, meaningful, well-separated categories; size balance; using any
     meaning found in the order of the list; anything else. all_city_cat_lists['
    g3P04'] = [
['Silverglade', 'Category Name'],
['Veridian Bay', 'Category Name'],
{{rest of city names}}

Model: Here is the Python list with the filled categories. I have
    identified 9 distinct categories [...]

User: Would the list be different if the requirements for category count
    and list ordering were relaxed?
```

Prompt 2: Labeling of city categories in Table 3. Prompt was given to Gemini 3 Pro. (Note that order of the list **is** mentioned in the prompt. However, this is because when not given instructions on priorities, the model tended to prioritize order of the list over everything else.)

E.1 Subjective Preference Comparison Prompt Template

To ask the model to choose between two candidates `ent1` and `ent2` using a given prompt phrasing `prompt`, we construct a standardized preference comparison template:

```
{{question prompt}}
The two numbers are: [{{ent1}}] | [{{ent2}}].
Please respond with only the chosen number.
```

Prompt 3: Standardized preference comparison template used for numbers.

```
{{question prompt}}
The two names are: [{{ent1}}] | [{{ent2}}].
Please respond with only the chosen name.
```

Prompt 4: Standardized preference comparison template used for human names.

```
{{question prompt}}
The two cities are: [{{ent1}}] | [{{ent2}}].
Please respond with only the chosen city.
```

Prompt 5: Standardized preference comparison template used for fictional city names.

```
From the AI model output below, identify and list the key factors that
    significantly influenced its decision-making process. Present these factors
    as two Python-style lists of strings (e.g., ["Prime", "Larger", "Odd", "Even
    ", "First presented", "Repeated digits"], ["Larger", "Prime"]). The first
    list should contain all factors mentioned in the model output, in the order
    they appear, represented using the fewest words possible. The second list
    should contain only the factors demonstrably used in the final decision,
    ordered by their relative importance, with the most influential factors
    listed first. This second list should only include factors that are directly
     relevant to the option that was ultimately chosen - e.g. "Larger" or "Round
    " but not "Deterministic choices" or "Predictable". The factors in the
    second list should also describe the option that was ultimately chosen - e.g.
     if it was deciding between a larger and smaller number, and ultimately
    chose the smaller one, the second list should contain "Smaller" but not "
    Larger". Factors should be represented as concise, one-to-three word phrases
    , using the fewest words possible while keeping meaning clear - e.g "Larger"
     instead of "Larger number". Ensure consistent terminology across factors.
    Only factors relevant to the inputs should be included - exclude statements
    of neutrality (e.g., "no preference") or limitations (e.g., "lack guidance").
     If no relevant factors are identified in the model output, state ["No
    factors identified"] for both lists. Base the lists solely on the provided
    model output; do not introduce external information or assumptions. The
    output should contain *only* the two lists, with no additional text or
    formatting. {{model_output}}
```

Prompt 6: Getting list of most important decision factors for each thinking trace.

F Summary Table of Preference Features

Feature	Datasets	Notes
Asc Digit Order	Numbers	Digits are in ascending order (smallest number first).
Desc Digit Order	Numbers	Digits are in descending order (largest number first).
Ending Vowel	Cities and names	Ends with a vowel.
First Alphabetically	Cities and names	
First Generated	Cities	Earlier in the list of entities originally generated by an LLM.
First Presented	All domains	First option shown to the model.
Generated Earlier Within Own Category	Cities	Ranked higher in percentile rank (each within their own category).
Has Repeat Digits	Numbers	Digits are not unique (e.g. 121 or 844).
Higher	Numbers	Higher number.
Higher Bouba Frac	Cities and names	Higher fraction of ‘bouba vs. “kiki letters (see text for citation).
Higher Census Freq	Names	Higher frequency on 2020 US Census (see text for citation)
Higher Frac of 1960 Babies	Names	Name was more common among babies born in 1960.
Higher Frac of 1980 Babies	Names	See prev
Higher Frac of 2000 Babies	Names	See prev
Higher Frac of 2020 Babies	Names	See prev
Higher Order Mag	Numbers	Higher order of magnitude.
Higher Rare Letter Frac	Cities and names	Higher fraction of “rare letters (LLM-defined as Q, X, or V).
Higher Sum Digits	Numbers	Sum of number’s digits is higher.
Higher Syllable Heur	Cities and names	Higher syllable count (using a heuristic based on vowels).
Higher Unique Char Frac	Cities and names	Higher ratio of unique letter count to total letter count.
Higher Vowel Frac	Cities and names	Higher ratio of vowel count to total letter count.
Longer Length	Cities and names	Longer string.
Mult 10	Numbers	Multiple of 10.
Mult 2	Numbers	See prev
Mult 5	Numbers	See prev
Palindrome	Numbers	
Prime	Numbers	
Starting Vowel	Cities and names	Starts with a vowel.

G Model Configurations and Citations

In Table 7, we provide a list of all model families used in our experiments along with their exact API, HuggingFace, or LiteLLM identifiers and their respective citations.

Creator	**Model Family**	**API / Config String**	**Citation**
Google	Gemini 3 Flash	`"gemini-3-flash-preview"`	Google DeepMind [2025]
Google	Gemini 2.5 Flash	`"gemini-2.5-flash-preview"`	Comanici et al. [2025]
Qwen	Qwen 3.5 35B	`Qwen/Qwen3.5-35B-A3B`	Qwen Team [2026]
OpenAI	GPT-5 Mini	`gpt-5-mini-2025-08-07`	Singh et al. [2025]
Anthropic	Claude Haiku 4.5	`claude-haiku-4-5-20251001`	Anthropic [2025]

Table 7: Model configurations and their respective citations.

H Factors named in thinking traces

The thinking traces for each comparison were fed to a second LLM, which was asked to generate a list of factors mentioned in the thinking, ordered by highest importance to the final decision (Prompt 6). We note that due to temporary API limitations at the time of collection, this analysis for both datasets was performed on a separate evaluation set sampled from an older (LLM-generated) version of the names dataset, rather than the name list and pairing design used for the win fractions in Section 3.1.

For the Numbers dataset, the most important factors identified in these traces (after manual deduplication of factor names) are shown in Figure 8.

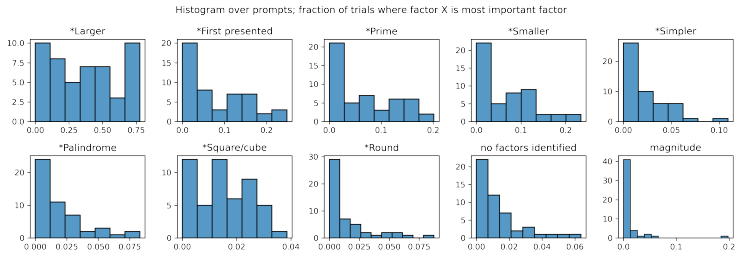


Figure 8: For each prompt, the fraction of trials where the given factor was determined from the model's thinking trace to be the most relevant decision factor for the Numbers dataset.

For the Names dataset, this analysis was performed on a set of 200 unique pairs due to collection-time constraints. As shown in Figure 9, the model's reasoning for names is significantly more complex than for numbers, frequently citing phonetics, cultural associations, and perceived "modernity" as decisive factors.

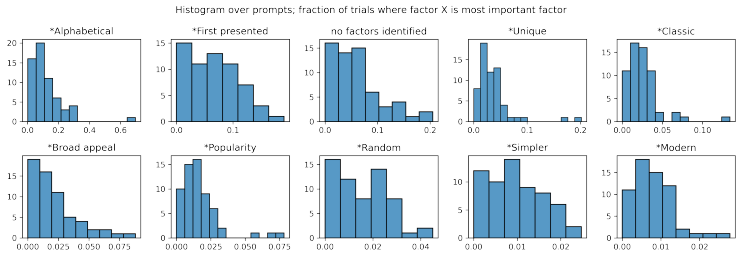


Figure 9: Histogram over prompts: for each prompt, the fraction of trials where the given factor was determined from the model's thinking trace to be the most relevant decision factor for the Names dataset.

I Broader Impacts

We study how LLMs engage in subjective decision-making tasks, an activity that is used for a set of broader tasks ranging from creative brainstorming to marketing and branding. Given the widespread use of this type of task, it is important to understand how the underlying preferences in the models affect the outcomes. This can both inform users of how to interpret the results they from LLMs for these tasks, and also suggest lines of research to mitigate any biases.

We have also found that the set of preferences is not monolithic, but can in fact be affected by subtle variations in the prompts that are used to specify the subjective comparisons. We have focused on comparisons between short pieces of content, but our methodology for formulating and evaluating the comparisons provides a template that can be applied essentially directly to a wide range of other subjective comparisons. In particular, it indicates the benefits of using multiple heterogeneous prompts for these types of comparison tasks, to offset the effect of preferences inherent to any one particular prompt or style of prompt. Our empirical results provide guidelines for some of the general effects that different prompts induce on the underlying comparison activity, in general terms.

J Extended Results for Other Model Families, Datasets, and Model Hyperparameters

J.1 Tables of Results for Other Datasets and Model Settings

Table 8 lists the win fractions across multiple temperatures ($T = 0.1, 0.7, 1.0$) for several different models (the selected GPT model does not permit temperatures other than 1.0).

Table 9 lists win fractions for the Cities dataset, as Table 1 does for Numbers and Names.

Table 8: Table of preferences across different temperatures for each model.

Comparable	Metric	Gemini 2.5 Flash (T=1.0)	Gemini 2.5 Flash (T=0.7)	Gemini 2.5 Flash (T=0.1)	Gemini 3 Flash (T=1.0)	Gemini 3 Flash (T=0.7)	Gemini 3 Flash (T=0.1)	Claude Haiku 4.5 (T=1.0)	Claude Haiku 4.5 (T=0.7)	Claude Haiku 4.5 (T=0.1)
Numbers	Higher	0.70***	0.69***	0.70***	0.78***	0.73***	0.73***	0.79***	0.75***	0.76***
	Higher Order Mag (14%)	0.58*	0.58*	0.58*	0.74***	0.74***	0.75***	0.75***	0.75***	0.76***
	Prime (28%)	0.64***	0.65***	0.65***	0.60*	0.60*	0.60*	0.62**	0.62**	0.62**
	Asc Digit Order (31%)	0.44*	0.44*	0.44*	0.40*	0.40*	0.41*	0.40**	0.40**	0.40**
	Higher Sum Digits	0.57***	0.56***	0.57***	0.62***	0.62***	0.62***	0.64***	0.64***	0.64***
	Mult 10 (22%)	0.52	0.52	0.52	0.43	0.43	0.42	0.37**	0.37*	0.36**
	Palindrome (17%)	0.59*	0.59*	0.58	0.53	0.53	0.53	0.50	0.50	0.50
	Mult 5 (35%)	0.51	0.51	0.51	0.45	0.45	0.45	0.41*	0.41*	0.41*
	Mult 2	0.46	0.46	0.46	0.45	0.45	0.45	0.44*	0.43*	0.43*
	First Presented	0.59***	0.59***	0.60***	0.50	0.50	0.50	0.63***	0.63***	0.63***
	Desc Digit Order (43%)	0.54	0.54	0.54	0.52	0.52	0.52	0.52	0.53	0.52
	Has Repeat Digits (41%)	0.55	0.55	0.54	0.54	0.54	0.54	0.53	0.53	0.53
Names	*Higher Rare Letter Frac (6%)*	0.82***	0.83***	0.85***	0.97***	0.97***	0.97***	0.68**	0.68**	0.67*
	Starting Vowel (31%)	0.76***	0.76***	0.77***	0.74***	0.74***	0.73***	0.60***	0.60***	0.60***
	Higher Frac of 2020 Babies	0.72***	0.72***	0.73***	0.75***	0.75***	0.75***	0.65***	0.65***	0.65***
	Higher Frac of 2000 Babies	0.69***	0.68***	0.69***	0.71***	0.71***	0.71***	0.62***	0.62***	0.63***
	Higher Vowel Frac	0.62***	0.62***	0.63***	0.62***	0.62***	0.63***	0.54*	0.54*	0.54*
	Higher Frac of 1960 Babies	0.40***	0.40***	0.40***	0.34***	0.34***	0.34***	0.42***	0.42***	0.42***
	Higher Bouba Frac	0.60***	0.60***	0.60***	0.59**	0.60***	0.60***	0.53	0.53	0.53
	Higher Syllable Heur	0.56*	0.56*	0.56	0.64***	0.64***	0.65***	0.50	0.50	0.50
	Ending Vowel (41%)	0.55	0.55	0.56	0.56	0.56	0.56	0.46	0.45	0.45
	First Alphabetically	0.59***	0.59***	0.58***	0.54	0.54	0.54	0.50	0.50	0.50
	Longer Length	0.48	0.47	0.48	0.56	0.56*	0.56*	0.51	0.52	0.51
	Higher Unique Char Frac	0.48	0.48	0.48	0.48	0.47	0.47	0.51	0.51	0.51
	Higher Census Freq	0.52	0.52	0.52	0.49	0.49	0.48	0.51	0.51	0.51
	Higher Frac of 1980 Babies	0.54	0.54	0.54	0.50	0.50	0.51	0.51	0.51	0.51
	First Presented	0.52***	0.51**	0.51	0.51*	0.51	0.50	0.82***	0.82***	0.82***

Table 9: Table of preferences for fictional city names across all models. Significance markers reflect Bonferroni-corrected α levels for $m = 183$ total tests (* $p_{\text{Bonf}} < 0.05$, ** $p_{\text{Bonf}} < 0.01$, *** $p_{\text{Bonf}} < 0.001$).

Metric	Gemini 2.5 Flash	Gemini 3 Flash	Claude Haiku 4.5	GPT-5 Mini	Random Pref Order 1	Random Pref Order 2	Random Pref Order 3
Starting Vowel (32%)	0.64	0.64	0.53	0.64	0.61	0.56	0.47
Longer Length	0.43	0.39*	0.42	0.41*	0.53	0.42*	0.52
Higher Bouba Frac	0.57	0.60*	0.57	0.61**	0.49	0.52	0.52
Higher Syllable Heur	0.48	0.45	0.43	0.48	0.48	0.38**	0.55
Higher Vowel Frac	0.58*	0.60*	0.53	0.60*	0.45	0.54	0.49
Ending Vowel (39%)	0.55	0.56	0.50	0.56	0.53	0.44	0.46
Higher Rare Letter Frac (14%)	0.57	0.54	0.54	0.59	0.50	0.50	0.50
Generated Earlier Within Own Category	0.53	0.56	0.52	0.54	0.51	0.44	0.40**
First Generated	0.44	0.42	0.49	0.48	0.51	0.57	0.53
First Alphabetically	0.53	0.52	0.49	0.50	0.51	0.54	0.53
Higher Unique Char Frac	0.48	0.51	0.54	0.51	0.52	0.54	0.58*
First Presented	0.51	0.47**	0.51	0.39***	0.50	0.50	0.50

J.2 Regression Coefficients for Other Datasets

In this section, we present the logistic regression coefficients for the Numbers and Cities datasets (Figures 10 and 11), complementing the results shown for the Names dataset in Figure 1 of the main text. Bonferroni-corrected p-values for every coefficient, along with confidence intervals both with and without Bonferroni adjustment, are in Section 1 of `https://anon-for-reviews.github.io/neurips2026/report_statistics.html`.

J.3 Prompt Variation Histograms

Figures 12, 13, and 14 show the distribution of per-prompt winning fractions for the numbers dataset for Gemini 3 Flash, GPT-5 Mini, and Qwen 3.5 35B, respectively (representing the equivalent of Figure 2 in the main text).

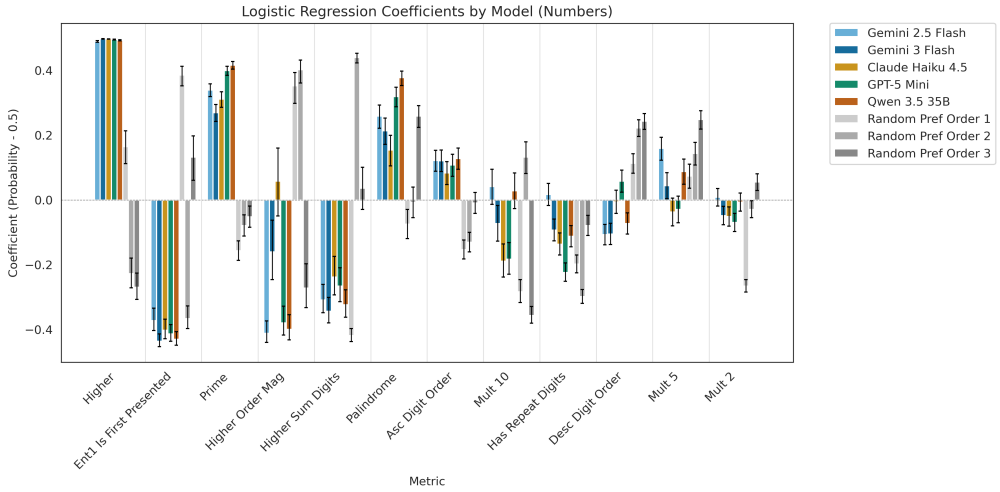


Figure 10: Logistic regression coefficients (with 95% confidence intervals) for the Numbers dataset.

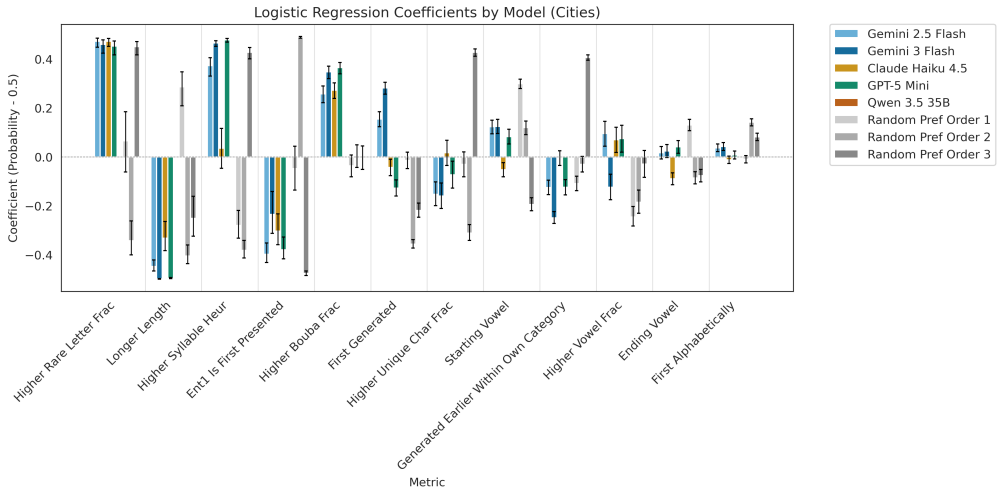


Figure 11: Logistic regression coefficients (with 95% confidence intervals) for the Cities dataset.

J.4 Prompt Robustness Across Datasets

Figures 15 and 16 show the correlation of prompt-dependent preferences across different datasets for Gemini 3 Flash and GPT-5 Mini (representing the equivalent of Figure 3 in the main text).

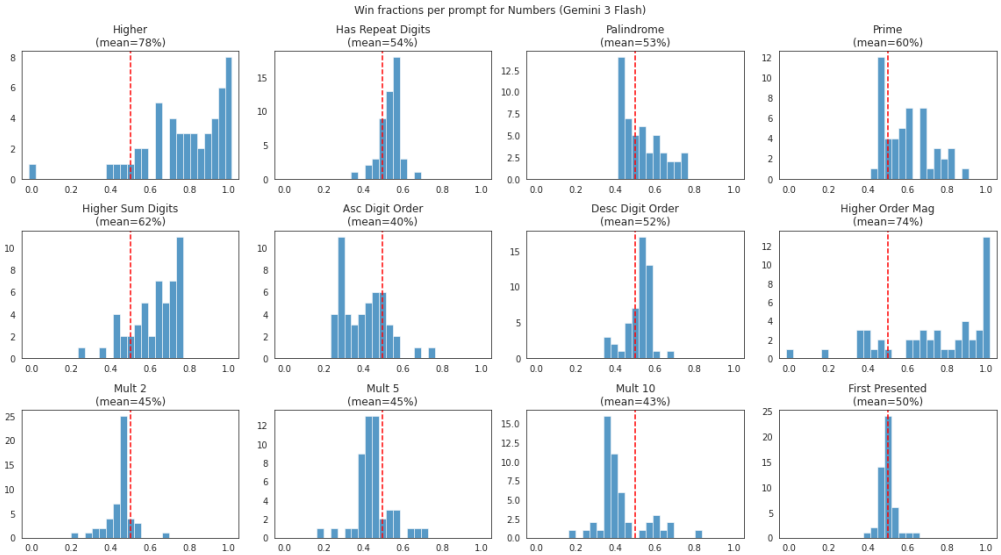


Figure 12: Distribution of induced preferences per prompt for Gemini 3 Flash.

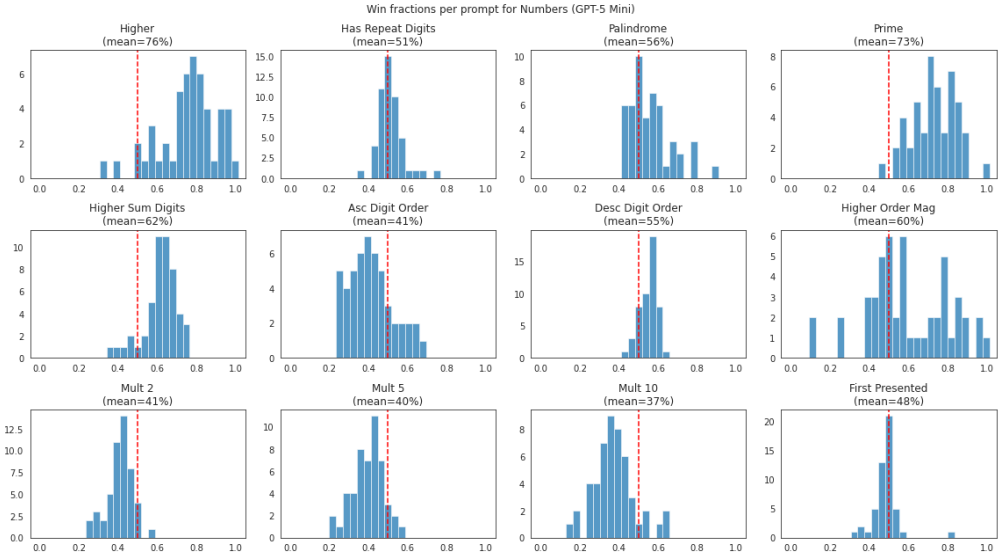


Figure 13: Distribution of induced preferences per prompt for GPT-5 Mini.

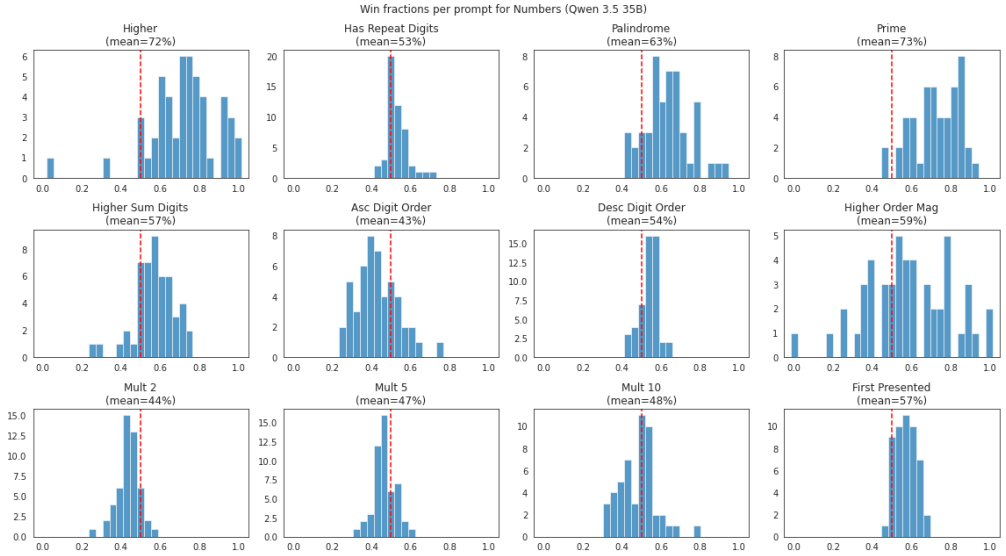
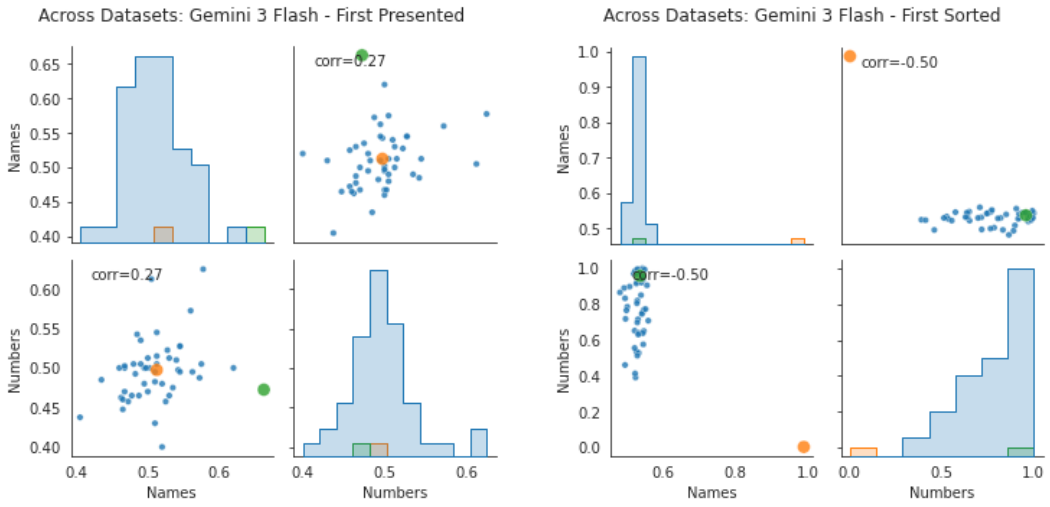


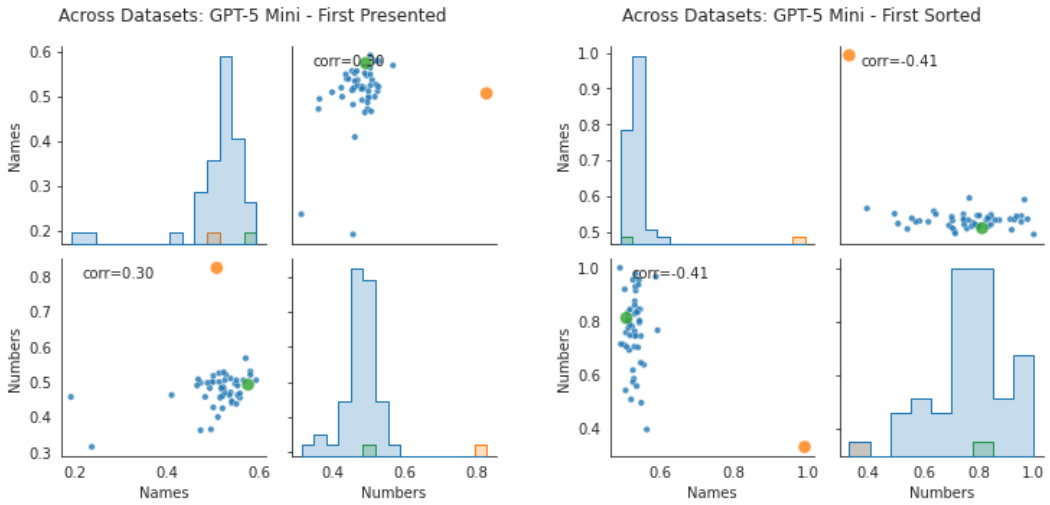
Figure 14: Distribution of induced preferences per prompt for Qwen 3.5 35B.



(a) First Presented

(b) First Sorted

Figure 15: Correlation of prompt preference between datasets for Gemini 3 Flash.



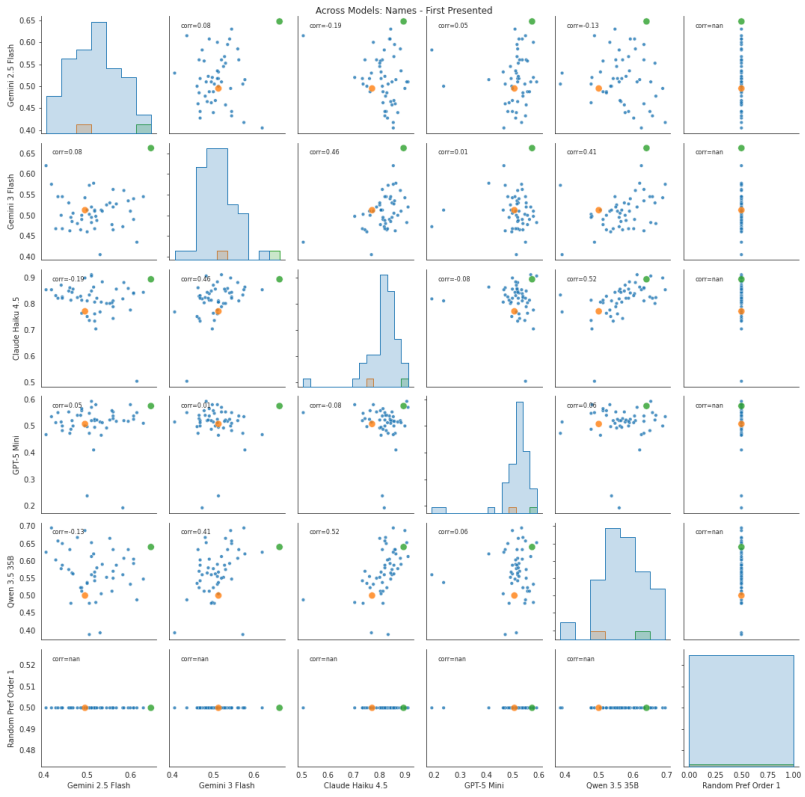
(a) First Presented

(b) First Sorted

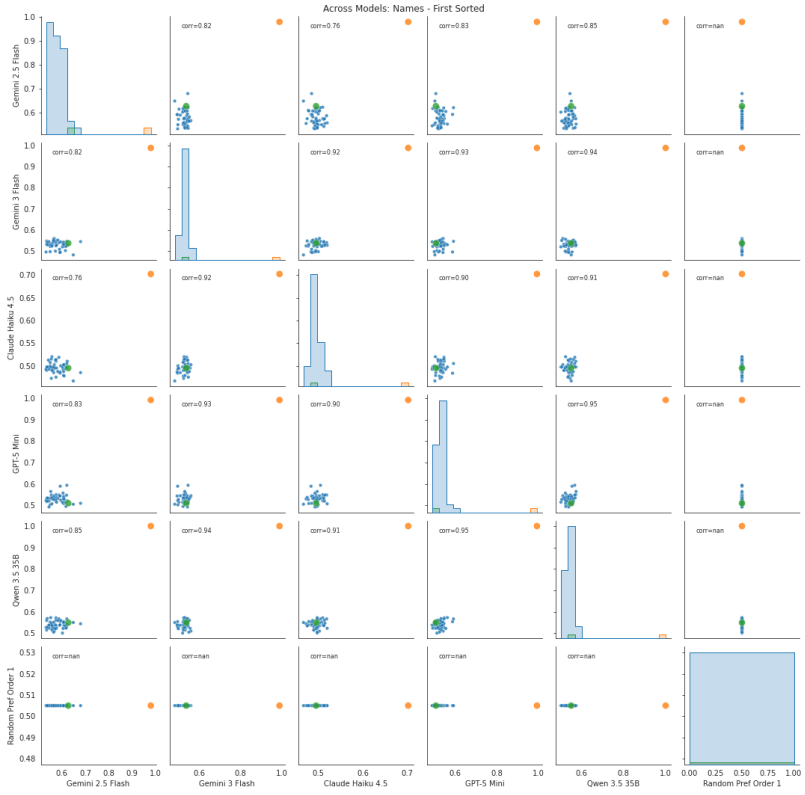
Figure 16: Correlation of prompt preference between datasets for GPT-5 Mini.

J.5 Prompt Behavior Correlations Across Models

Finally, we examine how prompt-dependent preferences correlate *between* different model families for the same dataset. Figures 17 and 18 show the PairGrid correlations across all evaluated SOTA models for the Names and Numbers datasets, respectively.

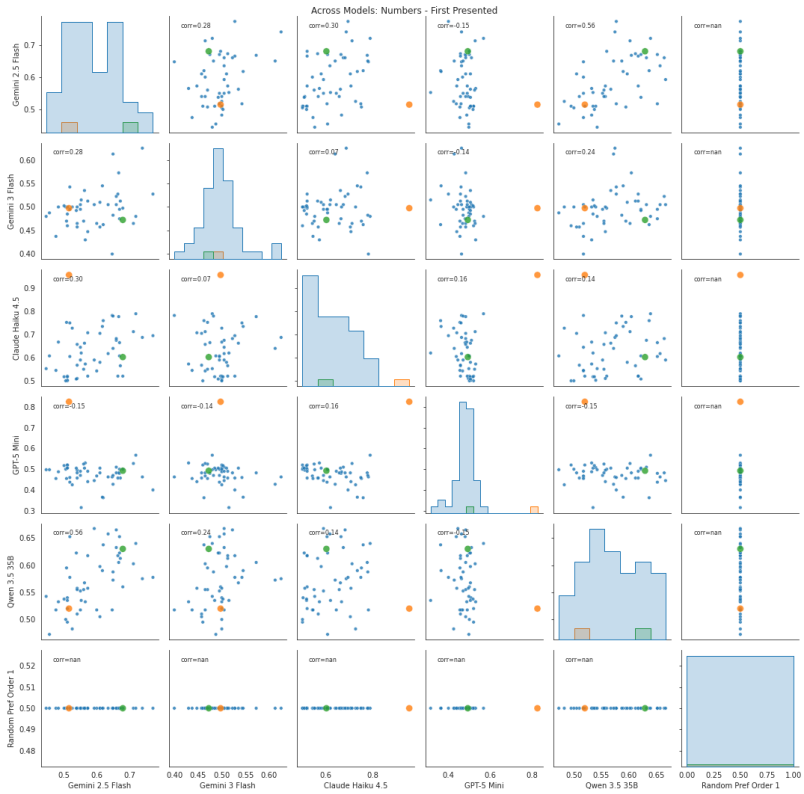


(a) First Presented

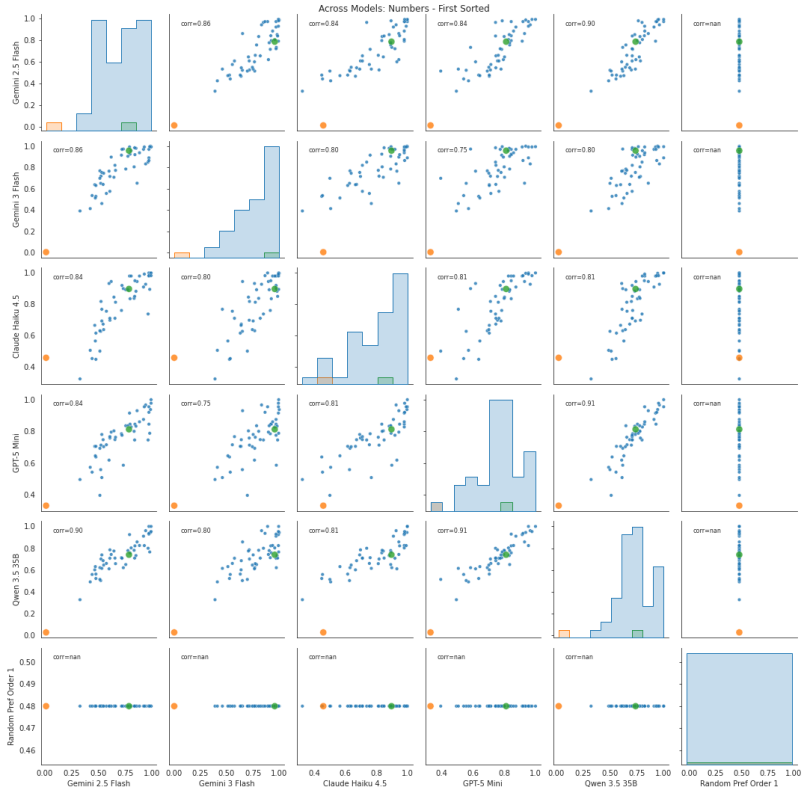


(b) First Sorted

Figure 17: Prompt behavior correlations across all models for the Names dataset.



(a) First Presented



(b) First Sorted

Figure 18: Prompt behavior correlations across all models for the Numbers dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 (Introduction) accurately describe the paper's scope—exploring subjective LLM biases—and its main contributions: the identification of dozens of biases (Section 3.1), the impact of prompt variation (Section 3.2), and the interplay between ideation and pruning (Section 3.3).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper has a limitations section in Section 4. Additionally, throughout the text, the paper discusses scope limitations, such as focusing on stylized tasks (Section 1) and short-form items (Section 2) to isolate specific biases.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is an empirical study of LLM behavior in subjective decision-making tasks and does not contain any theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive details on the datasets, prompt expansion, and experimental framework in Section 2 and Appendix C & E. Appendix E covers the exact prompts used in all phases.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We cannot release all of the code, but all critical experimental parameters, datasets, and prompts are reproduced in full in Appendix C and Appendix D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all experimental parameters including model versions, sampling methods, temperature settings ($T = 1$, except for in Appendix J.1 where other temperatures are analyzed), and the pairwise comparison task in Section 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance using Bonferroni-corrected one-sample t-tests against a 0.5 baseline and Levene's statistic, detailed in Section 3.1 and Section 3.2, using conservative sample size estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments were conducted using cloud-based LLM APIs (Gemini, Claude, GPT) and open-source LLMs (Qwen). We specify the models and settings in Section 2 and Appendix E, which defines the compute requirements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics as it is a standard study of LLM behavior and biases using synthetic and publicly available demographic data in Section 2.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive and negative societal impacts of researching LLM subjective biases, including standardizing neutral assistance and the risk of prompt exploits, detailed in Appendix I.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The research does not involve high-risk data or the release of high-risk models, utilizing only standard public APIs and public census data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Creators of existing assets are credited via citations, including US Census names, SSA data, and model configurations mapped in Appendix G (Table 7). Licenses for public datasets are standard for US government work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new assets—50 prompts and 200 fictional city names—which are reproduced in full in Appendix C and Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing or human subjects, relying entirely on LLM-driven experiments in Section 2.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: IRB approval is not applicable as the study does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper explicitly describes the usage of LLMs for generating city names, expanding prompts, and categorizing city names and prompts in Section 2 and Section 3.2.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.