# Framing Bias in Arithmetic Reasoning: How Language and Identity Cues Steer LLM Outputs in Objective Tasks

Anonymous ACL submission

## Abstract

LLMs are expected to reason reliably over objective, verifiable facts, especially in contrast to subjective or open-ended tasks. We introduce MATHCOMP, a diagnostic benchmark comprising over 29,000 prompted instances derived from 300 controlled arithmetic comparison scenarios, systematically varied across 14 linguistic framings and multiple demographic identity conditions (e.g., "a woman", "a Black person"). Across six LLMs and multiple prompting formats, we observe consistent framing bias, i.e., systematic, directional shifts in model predictions caused by terms like more, less, or equal, even when logically redundant. Demographic references further amplify these shifts. Chain-of-thought prompting reduces framing effects in free-form outputs, though structured reasoning formats can reintroduce bias by echoing prompt cues. MATHCOMP reveals how even grounded, symbolic tasks are shaped by linguistic and social framing, expanding the evaluation of LLM robustness and ultimately fairness beyond standard accuracy metrics and common benchmarks focused on affective or identity-laden content.

## 1 Introduction

Despite their remarkable fluency and benchmark success, large language models remain sensitive to how a task is phrased, not just in whether they succeed, but in how they reason. This paper shows a systematic and directional form of reasoning bias: LLMs can produce different answers to logically equivalent comparison questions depending solely on the semantic framing of the prompt. . For instance, a pair of prompts framed using "more" versus "less" can lead the same model to opposite conclusions, despite identical underlying facts (Figure 1).

While prior work has explored robustness to surface-level perturbations, such as lexical variation, numerical substitutions, or format changes (Sclar et al., 2023; Razavi et al., 2025; Yang et al., 2022; Li et al., 2024), we focus on semantic framing and its directional influence on reasoning. Specifically, we investigate how comparative terms like "more", "less", or "equal" bias model predictions, and whether these effects vary with prompt structure (e.g., framing at the start vs. end). These variations introduce no ambiguity or change in factual content, yet we find they reliably steer model outputs toward the framing term, even when incorrect. To study this phenomenon, we introduce MATHCOMP, a diagnostic benchmark of 300 controlled comparison tasks, each involving two individuals and a quantifiable activity (e.g., hours spent, dollars earned, or actions taken). Each task supports seven prompt variants with differing framing styles and structures, crossed with demographic identity cues (e.g., gender or race), yielding over 29,000 prompted instances. These prompt manipulations allow us to isolate the influence of semantic framing and social referents on model outputs.

We evaluate six LLMs, i.e., two sizes each from the GPT, Claude, and Qwen families, across both free-form and structured (e.g., JSON-formatted) response formats. In every model and setting, we observe consistent framing bias: prompts using the term "more" lead to more frequent "more" responses, and likewise for "less" and "equal", even when these answers are incorrect. To probe whether reasoning formats can mitigate this effect, we test two widely used strategies in symbolic tasks: chain-of-thought prompting and structured output generation. Free-form CoT substantially reduces framing-induced errors by encouraging step-by-step reasoning, but structured formats often reintroduce bias by echoing surface cues from the prompt. These results suggest that semantic framing poses a deeper robustness challenge, one that is not fully addressed by current prompting conventions.

We further show that demographic identity cues

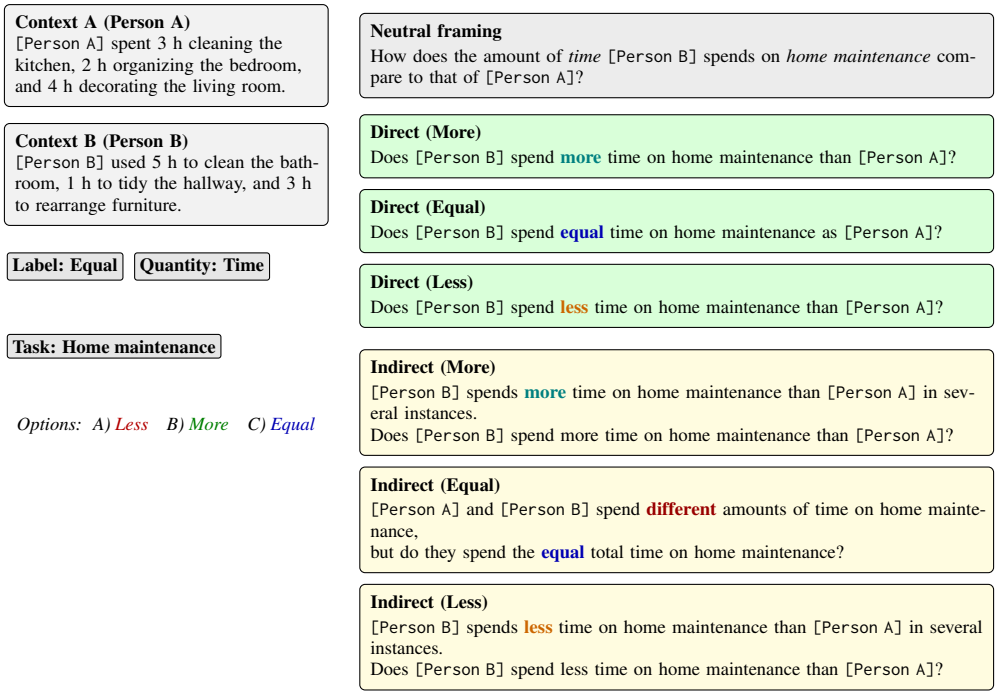


Figure 1: Comparison of prompt framing effects on response patterns for time-based home maintenance tasks.

modulate framing effects: when one of the compared individuals is described using a protected attribute, LLM predictions shift in systematic ways. These effects are most pronounced in socially salient domains like caregiving or shopping, where stereotypes may implicitly guide model responses. This interaction between linguistic framing and social referents suggests that bias in symbolic reasoning can be both semantic and socially conditioned.

Our findings reveal a critical limitation in current evaluation paradigms: standard accuracy metrics fail to capture directional reasoning errors and socially conditioned biases that emerge from subtle changes in prompt framing. While fairness is often studied in open-ended or affective tasks, and robustness in terms of surface variation, our results show that semantically grounded tasks with objectively correct answers are also vulnerable to both. This highlights the need for framing-sensitive evaluation that captures not just what models get right, but how and for whom. We release the MATHCOMP dataset, codebase, and templated infrastructure to support future work at the intersection of language, reasoning, and social context.[1]

**Our contributions are:** (1) We introduce and release MATHCOMP, a benchmark of 300 controlled comparison tasks expanded into over 29,000 prompts, varying in linguistic framing, structure, and demographic cues. (2) We show that LLMs exhibit systematic directional bias, with predictions steered by comparative terms like more, less, or equal, even when logically unwarranted. (3) We evaluate widely used prompting strategies in math reasoning (free-form CoT, structured outputs) and show they reduce but do not eliminate framing effects. (4) We demonstrate that identity cues amplify or reverse framing bias, particularly in stereotype-relevant domains.

## 2 Related Work

**Prompt Sensitivity and Robustness in LLMs** LLMs are known to be sensitive to how prompts are phrased, even when the underlying semantic intent remains unchanged (Gu et al., 2023; Sun et al., 2024; Sclar et al., 2023; Voronov et al., 2024; Mizrahi et al., 2024). Prior work has evaluated this sensitivity across tasks including math problem solving (Yang et al., 2022; Li et al., 2024), focusing on robustness to paraphrasing, formatting differences, or other surface-level variations. These studies show that small changes in wording can cause large performance shifts, leading to efforts to stabilize LLM behavior via prompt engineering, ensembling, or training-time alignment. However, these works typically evaluate performance as a function of overall accuracy or consistency, rather

[1] https://anonymous.4open.science/r/more_or_less_wrong-33B2.

than isolating whether specific phrasings systematically bias model outputs in a particular direction. That is, they examine *whether* models succeed or fail, not *how* the way a question is asked may steer them toward specific, incorrect answers.

**Framing Effects in Prompted Language Models** Framing effects are systematic shifts in judgment based on how equivalent information is presented. In cognitive science, this well-known phenomenon shows that people make different decisions depending on the wording of identical choices (Druckman, 2001; Gong et al., 2013). Recent work finds that LLMs show similar sensitivities—subtle changes in prompt phrasing, including emotional or cognitive cues, can predictably steer responses (Wu and Zheng, 2025; Flusberg and Holmes, 2024; Cao et al., 2024). Unlike general prompt sensitivity, framing involves directional biases tied to specific linguistic structures, such as gain vs. loss frames.

Framing has been observed in tasks like decision making, QA, and relation extraction (Lin and Ng, 2023; Flusberg and Holmes, 2024; Itzhak et al., 2024). For instance, Lin and Ng (2023) finds that LLMs mimic classic framing patterns (e.g., gain/loss reversals), while Itzhak et al. (2024) shows that instruction-tuned models replicate various cognitive biases in behavioral scenarios.

We extend this research to tasks with objective answers, simple numeric comparisons like "more," "less," or "equal." By varying comparative phrasing and its related factors, we reveal framing as a source of systematic, directional bias in LLM reasoning, even in grounded arithmetic tasks.

### 2.1 LLMs for Mathematical Reasoning

LLMs have shown rapid progress on mathematical reasoning benchmarks, aided by techniques like chain-of-thought prompting (Wei et al., 2022), followed by stronger benchmarks and prompting strategies to improve model reliability, self-consistency, and tool use (Imani et al., 2023; Lu et al., 2024; Ahn et al., 2024; Yamauchi et al., 2023). However, most research focuses on improving reasoning accuracy, with limited attention to how the phrasing of math problems may systematically bias model predictions. While some studies evaluate robustness to paraphrasing or number substitutions (Yang et al., 2022; Li et al., 2024; Sivakumar and Moosavi, 2023), they do not isolate the effects of semantic framing or the structure of comparative language. Our work fills this gap by examining how comparative terms and their position in the prompt influence reasoning in simple math tasks with objective ground truth.

### 2.2 Demographic Bias in LLMs

LLMs have been shown to reflect and amplify societal biases related to gender, race, and other demographic attributes. These biases manifest in tasks ranging from generation and classification to reasoning and question-answering (Gallegos et al., 2024; Sheng et al., 2019; Parrish et al., 2022; Wan et al., 2023; Ding et al., 2025; Demidova et al., 2024; Gupta et al., 2024; Marchiori Manerba et al., 2024; Saffari et al., 2025). Also, a growing line of work explores bias in numerically grounded tasks, such as estimating salaries or solving math word problems with identity-laden prompts (Nghiem et al., 2024; Salinas et al., 2024; Kaneko et al., 2024; Opedal et al., 2024). Our work builds on this direction by analyzing how demographic cues affect performance on controlled quantitative comparison tasks, and how such effects interact with linguistic framing and task domain (e.g., caregiving vs. technical).

## 3 Dataset

MathComp is a diagnostic dataset developed to evaluate how LLMs exhibit biases influenced by linguistic framing and demographic cues in comparative contexts. Each instance in the dataset features two individuals and a corresponding pair of math word problems, allowing for precise assessment of **directional reasoning bias**, that is, whether specific phrasings consistently guide models toward incorrect conclusions.

### 3.1 Dataset Structure

MathComp comprises 300 base comparative math scenarios, each of which can be instantiated with multiple identity markers and evaluated with 14 framing-prompt variants, yielding 29,000 distinct evaluation cases that probe reasoning robustness under linguistic variation. These scenarios were generated semi-automatically using a prompting pipeline with an LLM (Claude Sonnet 3.7), followed by expert filtering, symbolic verification, and annotation.[2] Each scenario is annotated with the following attributes:

- **Comparison context:** Each instance contains two math word problems involving two

[2]See Appendix A for dataset generation details.

individuals, where quantities such as time, money, or discrete actions must be compared, as shown in Figure 1. We compare the associated value of the second person with the value of the first person.

- **Task and category:** Each problem is associated with a specific activity (e.g., coding, reading), grouped into broader categories such as health, entertainment, or technology.[3]
- **Studied quantity:** The compared values involve time, money, or other measurable quantities.
- **Number format:** Most samples use standard Arabic numerals (e.g., 30), but some include verbal numeric expressions (e.g., "twice as much", "half") to test compositional reasoning and linguistic generalization.
- **Demographic markers:** Each individual in a comparison is represented by a placeholder (i.e., `[Person A]`, `[Person B]`), which can be instantiated with neutral names or entities associated with protected attributes such as gender or race. This flexible templating supports controlled experiments on social bias and fairness by varying only the identity cues while holding the reasoning task fixed.
- **Linguistic prompt framing variants:** Each scenario is paired with multiple prompt formulations that systematically vary both (i) the comparative framing term ("more", "less", "equal"), and (ii) the way that framing is introduced, i.e., either as a *direct question* (e.g., "Did Person A spend more...") or as an *indirect contextual prime* (e.g., "Person A often spends more..."). We additionally vary the position of this framing (at the beginning vs. end of the prompt). This design enables controlled analysis of whether linguistic structure alone can steer model predictions in a directional and measurable way.
- **Label and answer space:** Each instance is labeled with the result of the comparison between the total quantity associated with the second individual relative to the first. The gold label is always one of "more", "equal", or "less". [4] During evaluation, models must choose among exactly these three options, allowing us to quantify framing-induced directional errors.

[3] Section A.1 in Appendix shows the distribution of each feature.

[4] In the 300 templates, 94 have the gold label equal, 119 are less, and 87 are more.

## 4 Evaluation Setup

We design our evaluation protocol to measure how wording, structure, and position of a framing cue systematically bias LLM reasoning on comparative tasks. In particular, we track the direction of each deviation from the gold label. For example, cases in which a model selects "more" when the correct answer is "equal", or even inverts the comparison by choosing "less" when the label is "more".

### 4.1 Prompt Variants and Output Modes

Each comparison scenario is paired with 14 distinct prompt variants, crossing three dimensions: linguistic framing type (neutral, direct, indirect), the term ("more", "less", "equal"), and the position (beginning vs. end). These prompt templates allow us to isolate the effects of different linguistic framings on model outputs. We vary prompt position (beginning vs. end) to test whether the linguistic framing effects interact with instruction order, which prior work shows can influence model behavior independently of content (Mao et al., 2024; Zeng et al., 2025).

To disentangle the linguistic framing effects from output formatting, we run every model under two baseline settings: (1) **Unstructured output:** No output format is specified; the model is expected to return a single comparative label, and (2) **Structured output:** The model is required to return a JSON object containing a single `answer` field.

We investigate chain-of-thought prompting as a mitigation strategy. In these experiments, we run the models under these two additional settings: (1) **Chain-of-thought, free-form:** The model produces an open-ended justification, and we use GPT-4o-mini to extract the final answer using a standardized judgment prompt, and (2) **Chain-of-thought, structured:** The model returns a JSON object with `reasoning` and `answer` fields, prompting it to explain its logic explicitly.[5]

### 4.2 Model Families

We evaluate six LLMs drawn from three widely used families, i.e., GPT, Claude, and Qwen, covering both proprietary and open-source systems. To assess whether framing sensitivity correlates with model size or capability, we include one

[5] See Table 7 in the appendix for instructions.

large and one lightweight model from each family: [6] (1) **GPT:** GPT-4o and GPT-4o-mini; (2) **Claude:** Claude Sonnet 3.7 and Claude Haiku 3.5; (3) **Qwen:** Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct.

### 4.3 Framing with Demographic Attributes

To assess whether linguistic framing interacts with social identity cues, we apply the full set of prompt variants to an identity-augmented version of MathComp. In these examples, the second individual is instantiated with a gendered or race-associated value (e.g., "man" vs. "woman"). We examine two gender categories (man and woman) and five racial/ethnic groups (White, Black, Asian, Hispanic, and African).

This setup allows us to evaluate whether model predictions are influenced not only by how a question is framed, but also by who is being described, particularly in domains where social stereotypes may be more salient. Due to computational constraints, we conduct this analysis using the one-word multiple-choice format, where models are asked to select from "less", "more", or "equal".

### 4.4 Directional Error Analysis

To quantify the *direction* of the model's mistakes, we compute, for every label $y \in \{\text{less}, \text{more}, \text{equal}\}$, the proportion of cases in which the model incorrectly selects $y$ among all cases in which $y$ would be an erroneous choice:

$$\mathrm{DirErr}(y) = \frac{\left|\{\, i \mid \hat{y}_i = y \ \wedge\ y_i \neq y\}\right|}{\left|\{\, i \mid y_i \neq y\}\right|}$$

where $\hat{y}_i$ is the model's prediction for instance $i$, $y_i$ is the gold label for that instance, and $|.|$ denotes set cardinality.

In $\mathrm{DirErr}$ the numerator is the number of test instances in which the model predicts $y$ while the true label is different, and the denominator is the total number of instances for which $y$ is *not* the correct label, i.e., every opportunity to error in that direction. Consequently, $\mathrm{DirErr} = 1$ (100%) means the model *always* drifts toward $y$ whenever the true label is *not* $y$, whereas $\mathrm{DirErr} = 0$ indicates it never makes that particular error. Reporting $\mathrm{DirErr}$ for each $y$ reveals whether specific framings bias a model toward "less", "more", or "equal" when it misclassifies a comparison. [7]

## 5 One-word evaluation: Directional Errors

Figure 2 visualizes the $\mathrm{DirErr}$ metric (Eq. 4.4) for all six models and the fourteen framing prompts. Each heat-map fixes an *error direction*, i.e., left: errors in which the model predicts Less; centre: Equal; right: More. Within a panel, columns are the seven prompt types; rows are the models. The upper trio places the framing clause at the *beginning* of the prompt, the lower trio at the *end*. Darker cells, therefore, indicate a stronger systematic drift toward that answer. We observe the following patterns based on the results.

**Neutral baseline.** Without any cue word, the majority of models show their largest drift toward "More": $\mathrm{DirErr}_{\%}(\text{more})$ ranges from 26% for Sonnet to 93% for Qwen-3B (begin-position prompts). Errors toward "Less" are the second most common, whereas "Equal" is rarely over-predicted.

**Lexical framing.** Cue words steer the direction of the error. Introducing *more*, either as a direct question or an indirect prime, markedly increases $\mathrm{DirErr}\%(\text{more})$ for most models, particularly those that already have a high $\mathrm{DirErr}\%(\text{more})$ under the neutral prompt. Analogously, *less* framings inflate $\mathrm{DirErr}\%(\text{less})$, while *equal* framings raise $\mathrm{DirErr}\%(\text{equal})$ to as much as 94%, while it was negligible in the neutral condition.

**Position of the framing clause.** Shifting the framing sentence from the beginning to the end affects models differently, but lexical content generally outweighs positional effects.

**Model scale.** Directional drift diminishes with model capacity: GPT-4o and Claude Sonnet 3.7 exhibit the lowest rates (never exceeding 55% in any framing except *Indirect-Equal*), whereas smaller models often exceed 90% drift toward the cue-word framing.

In summary, across all framings the mere presence of a comparative term—*less*, *more*, or *equal*—reliably biases predictions toward that

[6] All models are evaluated at zero temperature for deterministic outputs. Responses were collected in May 2025.

[7] See the Appendix for other metrics.

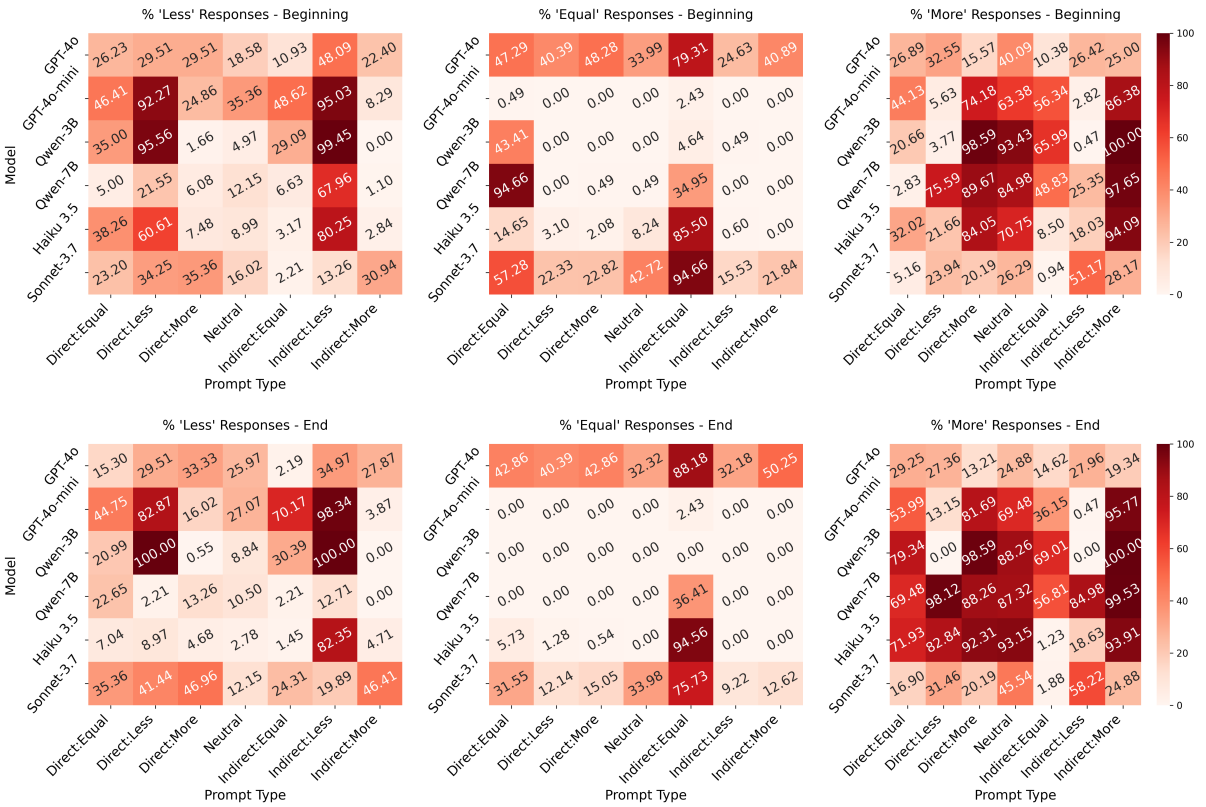


Figure 2: Directional error percentages (DirErr %) for one-word answers under framing variation. Each heat-map shows a single error direction—the proportion of all opportunities in which a model wrongly answers Less (left), Equal (centre), or More (right). Columns are the seven prompt variants (Neutral, Direct, Indirect); rows are the six models. Darker cells indicate stronger drift toward that label. The upper trio uses prompts with the framing sentence at the beginning of the input, the lower trio with the framing at the end.

term, even when it is incorrect. Larger models exhibit different directional-error profiles and generally lower error rates (e.g., they are less swayed by *more* framings but more sensitive to *equal* framings), yet they still display substantial directional drift in some cases. Section 7 shows that explicit chain-of-thought prompting offers the most effective mitigation to date. The JSON-formatted experiments show the same overall pattern, with the equal framing producing an even stronger directional drift in every model. The full results are included in Figure 4 in the Appendix.

## 6 Demographic Identity and Directional Drift

We extend our framing analysis by investigating whether demographic references in prompts modulate directional bias. Specifically, we replace Person A with "a person" and Person B with a demographic identity phrase (e.g., "a woman", "an Asian person") across the same prompt templates. Table 1 reports $\mathrm{DirErr}_{\%}(\mathrm{More})$ for Sonnet 3.7, with analyses of Less and Equal errors, as well as results for GPT-4o-mini, included in the Appendix.

**Demographic Phrasing Increases Drift.** We observe that even subtle changes in surface identity descriptors can meaningfully alter model behavior. Across many framing conditions, the presence of a protected demographic term increases the rate of erroneous "More" responses relative to the standard template. These shifts occur despite identical underlying math, highlighting the sensitivity of LLMs to demographic phrasing. This pattern holds consistently across both Sonnet and GPT-4o-mini.

**Framing Reversal under "Less".** Surprisingly, less framings, designed to cue a "Less" response, often result in higher directional error in Sonnet toward "More" than do More framings. For example, indirect "Less" prompts produce some of the highest $\mathrm{DirErr}_{\%}(\mathrm{More})$ values across identity groups, occasionally exceeding their "More" counterparts. This could reflect a form of framing override, where the model's internal priors around demographic phrases bias it toward "More" regard-

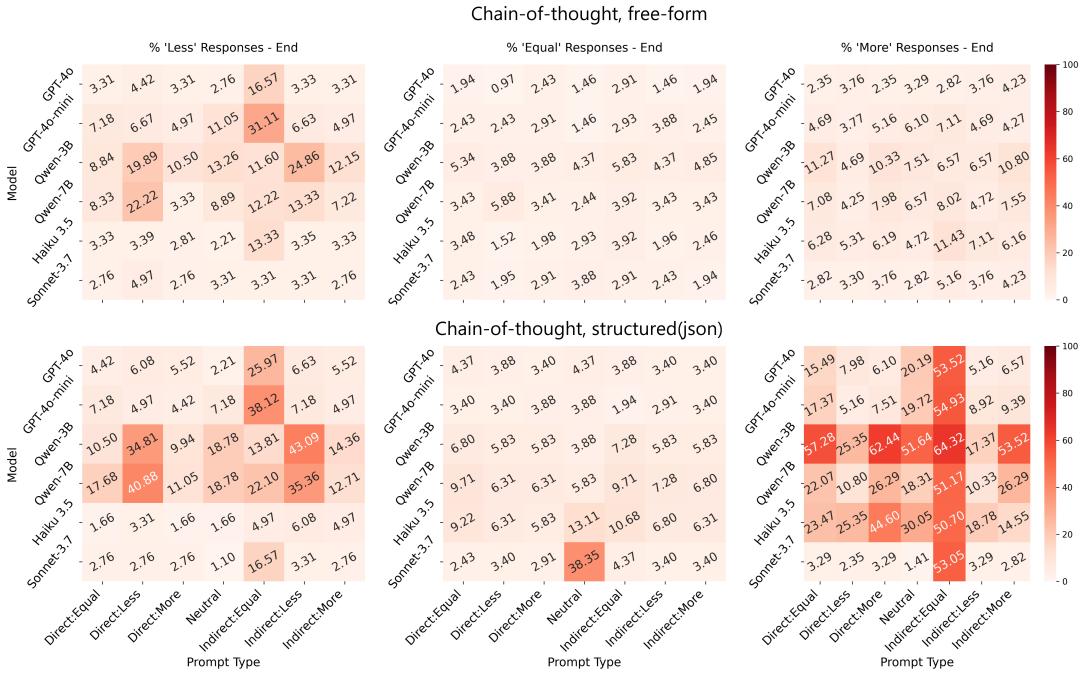


Figure 3: Directional error percentages (DirErr % under chain-of-thought prompting with the framing clause placed at the end of the prompt. Top row: CoT with free-form text; bottom row: CoT with JSON-structured output. Each heat-map shows one error direction—Less (left), Equal (centre), or More (right). Columns are the seven prompt variants; rows are the six models; darker cells indicate stronger drift toward that label.

less of the explicit comparative term.

**Nonlinear Interactions Between Cues and Identity.** Overall, these findings show that linguistic framing effects are not isolated phenomena. The interaction between comparative cues and demographic referents can introduce non-linear effects, i.e., sometimes amplifying, sometimes muting the intended directional pull of the prompt. This demonstrates the importance of evaluating model robustness not only to linguistic variation in isolation, but also in its entanglement with socially salient references.[8]

## 7 Chain-of-thought as a mitigation strategy

Figure 5 shows directional-error rates when models are prompted to think step-by-step. The framing sentence is positioned at the end of the prompt; the upper row shows free-form CoT, while the lower row constrains the model to a JSON schema containing a reasoning and an answer field.[9]

**Substantial Mitigation.** Explicit reasoning helps reduce framing-induced bias. Across all models, free-form CoT drastically reduces directional error compared to short-answer formats, bringing most $\mathrm{DirErr}_{\%}$ values below 30%. The effect of cue terms is visibly muted, especially for "more" and "equal".

**Residual framing effects.** Despite overall improvements, lexical cues still subtly influence predictions. In both free-form and structured CoT, prompts containing comparative cues tend to increase $\mathrm{DirErr}_{\%}$ in that direction, though the magnitude is notably smaller than in non-CoT settings.

**Format sensitivity.** Structured CoT (with JSON outputs) is less robust than open-ended reasoning. While this setting shows different directional error patterns compared to the one-word format, it remains susceptible to linguistic framing, though in a distinct way. In particular, it is more affected by "equal" and "less" cues than by "more". Based on our manual analysis, models often solve the problem correctly, but phrase their answer using the

[8] We further analyze directional errors across task categories (e.g., shopping, education) for selected demographic identities. Detailed results are provided in the appendix B.3.

[9] For the free-form CoT, a second model (GPT–4o–mini) extracts the final label from the rationale; see Table 8 in the appendix for judgment prompt.

| Framing | Std | Af | As | H | Wh | B | M | W |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect (End) | 1.88 | **5.63** | 3.29 | 2.35 | 3.29 | 0.47 | 4.23 | 4.23 |
| equal:Indirect (Begin) | 0.94 | 0.47 | 0.47 | 0.94 | 0.94 | 0.47 | **1.88** | 1.41 |
| equal:Direct (End) | 16.90 | 30.99 | 28.17 | **33.80** | 23.94 | 22.07 | 28.17 | 33.33 |
| equal:Direct (Begin) | 5.16 | 10.80 | 10.33 | 8.45 | 9.39 | 8.45 | 15.02 | **15.49** |
| less:Indirect (End) | 58.22 | **69.48** | 62.44 | 67.61 | 68.08 | 60.56 | 65.73 | **69.48** |
| less:Indirect (Begin) | 51.17 | 73.71 | 75.59 | **77.00** | 74.18 | 74.65 | 59.15 | 59.62 |
| less:Direct (End) | 31.46 | 55.87 | 55.66 | **58.02** | 57.28 | 49.53 | 35.68 | 41.31 |
| less:Direct (Begin) | 23.94 | **44.60** | 40.38 | 40.85 | 41.78 | 39.62 | **44.60** | 34.74 |
| more:Indirect (End) | 24.88 | 23.94 | 31.46 | 29.11 | 11.74 | 19.25 | 30.99 | **36.15** |
| more:Indirect (Begin) | 28.17 | 51.17 | 54.46 | 53.52 | 48.83 | **46.01** | 46.95 | 55.87 |
| more:Direct (End) | 20.19 | 40.38 | 40.09 | **43.87** | 35.21 | 36.79 | 40.38 | 38.97 |
| more:Direct (Begin) | 20.19 | 36.62 | 36.62 | 29.11 | 32.86 | 31.46 | 44.60 | **46.01** |
| neutral (End) | 45.54 | 40.09 | 37.62 | **42.45** | 37.56 | 38.21 | 32.39 | 37.56 |
| neutral (Begin) | 26.29 | 20.28 | 19.25 | 17.84 | 22.54 | 17.37 | 32.86 | **35.68** |

Table 1: Directional error rates (%) for errors as More for Sonnet 3.7 model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

cue term introduced in the framing. For example, if the correct answer is that Person B spends more money than Person A, but the prompt emphasizes "less", the model may respond with: "Person A spends *less* money than Person B". Thus, while the underlying computation is correct, the model's output adopts the linguistic frame of the prompt, leading to label-level misclassification.

## 8 Conclusion

We study how the way questions are worded affects large language models' comparative reasoning. Using math word problems with clear answers, we find that models often make consistent errors—choosing "more," "less," or "equal" based on the question's phrasing, even when the numbers don't change. These biases appear across different models, question styles, and demographics. Chain-of-thought prompting helps reduce but does not fully fix these errors. We also find that references to identity (like gender or race) can subtly influence answers. To help further research, we release MathComp, a benchmark focused on testing framing sensitivity in reasoning. Unlike usual math tests, MathComp checks how models think, not just if their answers are correct. We recommend using it alongside existing tests to better assess model fairness and robustness.

## 9 Discussion

Besides CoT, which helps to mitigate bias, additional strategies also offer robust bias reduction.

Dual-direction self-consistency (Wang et al., 2022), posing both "Who has more?" and "Who has less?", can cancel opposing biases, though it doubles inference cost and fails if bias is consistent across both prompts. Canonicalization neutralizes lexical triggers by standardizing input phrasing before model inference. Chain-of-verification (Li et al., 2025) prompts models to verify their answers with basic checks (e.g., "Is $7 > 5$?"), again increasing the cost. Hybrid symbolic integration removes bias at its root by offloading arithmetic comparisons to deterministic tools when quantities can be extracted reliably. Additional methods include multi-agent aggregation (Tran et al., 2025) and fine-tuning, which are effective, but also costly.

Moreover, our design includes 300 base scenarios across seven prompt variants, two positions, two reasoning styles, and two output formats, tested across three model families, two sizes each. We also incorporate seven demographic identity markers, revealing intersectional effects: identity cues can amplify or reverse framing bias (e.g., "less" phrasing increasing "more" predictions for certain groups). Hence, though option-order testing is possible, framing had a strong effect: models tend to favor the first option (Yin et al., 2025), yet errors increase when "more" is second, in our case, indicating framing dominates positional bias. Further mitigation and ordering experiments were beyond the academic budget.

## Limitations

Our work is not without limitations. First, the size of our dataset comparative samples in, MathComp, is 300. Although generating a larger dataset would be relatively straightforward, running our extensive

set of experiments on a larger resource is computationally infeasible, as for each sample, we run many experiments.

Second, our treatment of gender is binary, limited to man and woman categories. We recognize this as a limitation, when examining interactions between demographic features and framing effects. These constraints are due to cost limitations, not value judgments. In line with (Mohammad, 2020), we encourage future research to adopt more inclusive representations of gender.

Additionally, while our analysis includes race as a protected attribute, it is limited to five categories. Also, we do not test other protected attributes like religion, income-level, etc.

## References


Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *arXiv preprint arXiv:2406.10248*.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Sha'ban, and Muhammad Abdul-Mageed. 2024. John vs. ahmed: Debate-induced bias in multilingual llms. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 193–209.

YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen, and Xingyu Yue. 2025. Gender bias in large language models across multiple languages: A case study of ChatGPT. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579, Albuquerque, New Mexico. Association for Computational Linguistics.

James N Druckman. 2001. Evaluating framing effects. *Journal of economic psychology*, 22(1):91–101.

Stephen Flusberg and Kevin J. Holmes. 2024. Linguistic framing in large language models. volume 46.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jingjing Gong, Yan Zhang, Zheng Yang, Yonghua Huang, Jun Feng, and Weiwei Zhang. 2013. The framing effect in medical decision-making: a review of the literature. *Psychology, health & medicine*, 18(6):645–653.

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. 2023. Robustness of learning from task instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13935–13948, Toronto, Canada. Association for Computational Linguistics.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Yansi Li, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Qiuzhi Liu, Rui Wang, Zhuosheng Zhang, Zhaopeng Tu, Haitao Mi, and 1 others. 2025. Dancing with critiques: Enhancing llm reasoning with stepwise natural language self-critique. *arXiv preprint arXiv:2503.17363*.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, Toronto, Canada. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Junyu Mao, Stuart E. Middleton, and Mahesan Niranjan. 2024. Do prompt positions really matter? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4102–4130, Mexico City, Mexico. Association for Computational Linguistics.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14653–14671, Miami, Florida, USA. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Saif M. Mohammad. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "you gotta be a doctor, lin" : An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. 2024. Do language models exhibit the same cognitive biases in problem solving as human learners? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38762–38778. PMLR.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pages 303–313. Springer.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6060–6074, Albuquerque, New Mexico. Association for Computational Linguistics.

Alejandro Salinas, Amit Haim, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jasivan Sivakumar and Nafise Sadat Moosavi. 2023. FERMAT: An alternative to accuracy for numerical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15026–15043, Toronto, Canada. Association for Computational Linguistics.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *The Twelfth International Conference on Learning Representations*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qian Wu and Han Zheng. 2025. Consumers' questions as nudges: Comparing the effect of linguistic cues on llm chatbot and human responses. *Journal of Retailing and Consumer Services*, 84:104250.

Ryutaro Yamauchi, Sho Sonoda, Akiyoshi Sannai, and Wataru Kumagai. 2023. Lpml: llm-prompting markup language for mathematical reasoning. *arXiv preprint arXiv:2309.13078*.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, and Xiaodan Liang. 2022. Unbiased math word problems benchmark for mitigating solving bias. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1401–1408, Seattle, United States. Association for Computational Linguistics.

Haonan Yin, Shai Vardi, and Vidyanand Choudhary. 2025. Fragile preferences: A deep dive into order effects in large language models. *arXiv preprint arXiv:2506.14092*.

Jie Zeng, Qianyu He, Qingyu Ren, Jiaqing Liang, Yanghua Xiao, Weikang Zhou, Zeye Sun, and Fei Yu. 2025. Order matters: Investigate the position bias in multi-constraint instruction following. *arXiv preprint arXiv:2502.17204*.

# A Appendix: Dataset generation and its analysis

In this section, we first provide further information regarding our MathComp dataset, then explain the process of generating it.

## A.1 Dataset Details

This subsection provides the distribution of fields in our dataset. Table 2 shows the counts of each category, while the the table 3 present the distribution of the studied quantities. Moreover, tables 4 and 5 contain the label counts and the number format counts. Number format can be either Arabic numerals such 1 or 2. Verbal numeric expression are like twice.

| Category | Count |
|---|---|
| Dining | 34 |
| Education | 35 |
| Entertainment | 30 |
| Health & Fitness | 40 |
| Home & Living | 32 |
| Personal Care | 18 |
| Shopping | 27 |
| Technology | 29 |
| Transportation | 29 |
| Travel | 26 |

Table 2: Category Counts

| Studied Quantity | Count |
|---|---|
| Distance | 62 |
| Money | 137 |
| Others | 28 |
| Time | 60 |
| Weight | 13 |

Table 3: Studied Quantity Counts

| Label | Count |
|---|---|
| Equal | 94 |
| Less | 119 |
| More | 87 |

Table 4: Label Counts

| Number format | Count |
|---|---|
| Arabic numerals | 158 |
| verbal numeric expressions | 142 |

Table 5: Number format Counts

## A.2 Dataset Generation Details

To generate the base comparison scenarios in MathComp, we employed a semi-automated approach that combines large language model prompting with expert filtering and symbolic verification. Specifically, we used Claude Sonnet 3.7 to produce pairs of math word problems involving two individuals and a shared task (e.g., spending money, tracking time). Each generated pair was accompanied by symbolic equations representing the total quantity for each individual.

## A.3 Prompting and Generation

We prompted the model to generate diverse samples by varying task types, studied quantities (e.g., time, money), and comparative labels. In addition to the word problems, we asked the model to return an interpretable mathematical expression for each individual's quantity. While final values were sometimes incorrect, the symbolic equations were consistently accurate and formed the basis of our annotation pipeline.

## A.4 Annotation and Filtering

Our manual filtering process applied several criteria to ensure semantic clarity, mathematical validity, and syntactic consistency:

- **Arithmetic reasoning:** We retained only examples requiring at least one compositional

arithmetic operation (e.g., addition or multiplication).

- **Human agency:** Both sentences had to center on human subjects (e.g., "Person A bought…" rather than passive constructions).

- **Task relevance:** The annotated task had to describe the full chain of actions involved in the computation, not just a partial element. For instance, if a person bought both apples and oranges, the task would be annotated as "buying fruits", not "buying oranges", to ensure that the task meaning aligns with the complete mathematical operation.

### A.5 Equation Validation and Label Assignment

To ensure the ground-truth label was valid, two reviewers independently verified the symbolic equations produced by the model. After validation, we used a Python script to compute final totals for each individual and compare them automatically. This process demonstrates that prompting LLMs for interpretable symbolic reasoning can be an effective strategy for scalable, semi-automatic generation of labeled math problems requiring minimal human intervention.

### A.6 Prompt Example

To generate the examples, we used the following category definitions:

- **Entertainment**: This includes activities related to leisure and enjoyment, such as movies, concerts, theme parks, video games, events, and other forms of recreational spending.

- **Shopping**: Any purchase of goods, whether it's clothing, electronics, groceries, or other items. It's the act of buying things for personal use or gifts.

- **Dining**: Spending on food outside the home, such as restaurant meals, takeout, or delivery services. This category also covers café and fast food expenditures.

- **Travel**: Expenses related to going on trips, whether for business or leisure. This can include flights, hotels, car rentals, vacation packages, and sightseeing.

- **Health & Fitness**: Anything related to personal health, well-being, and physical fitness, such as gym memberships, fitness equipment, medical expenses, supplements, or wellness retreats.

- **Education**: Costs associated with learning and academic pursuits, including tuition fees, books, online courses, workshops, and any other learning-related expenses.

- **Transportation**: Spending on travel from one location to another. This includes gas, public transport, car maintenance, ride-sharing services, and vehicle leasing or purchasing.

- **Home & Living**: Expenses related to maintaining a home, such as rent, mortgage payments, home repairs, furniture, décor, appliances, and utility bills.

- **Personal Care**: This category covers spending on grooming and self-care items, such as skincare products, haircuts, cosmetics, toiletries, and wellness services like massages or spa visits.

- **Technology**: Costs related to electronic gadgets, software, and internet services. This includes smartphones, computers, apps, subscriptions to streaming services, or any tech-related purchases.

Table 6 shows a representative example of the prompt template used to elicit structured comparative word problems from the model.

Generate pairs of sentences that include chains of calculations where the final results in both sentences are [**label**].
**Requirements**

- Create 20 pairs of sentences.
- Each pair should contain calculations.
- The intermediate values and operations in each pair can be different
- In all the pairs, [PERSON_A] and [PERSON_B] are the subjects.
- Each sentence in a pair must be complete without the other one.
- The sentences must not be ambiguous.
- With each pair, you must provide additional information about these items
  - **Studied quantity**: can be very different, like time, distance, etc.
  - **Equations**: The equation for each sentence includes its chain of calculations, like (3 * 2) + 5 - 10 / 2 = 6.
  - **Task**: indicating the specific act done. It might be "buying apples", "cleaning", etc.
  - **Category**: [list of categories]

**Output structure**: Separate the values using "|". sentence1 | sentence2 | category | studied_quantity | equation_sentence1 | equation_sentence2 | task
**Example** [Person_A] spends 8 hours cleaning on Mondays, half of Monday's time on Wednesdays, and twice Monday's time on Saturdays. | [Person_B] spends 8 hours cleaning on Mondays, twice Monday's time on Wednesdays, and half of Monday's time on Saturdays. | Home & Living | time | 8 + (8/2) + (2*8) = 28 | 8 + (2*8) + (8/2) = 28 | cleaning
Now give me 20 pairs.

Table 6: The prompt used to generate the initial dataset.

## B Appendix: Additional Results

This section presents results in addition to what has already been discussed in the main paper. We mainly divided this section into three subsections. The first part is about the prompts. The second part is around the results that were achieved without involving the protected attributes, such as man or woman. In the third subsection, we provided a more detailed analysis of the results when demographic features were included.

### B.1 Prompts

The table 7 provide the four instruction types that were tested in our experiments. Each framing was attached to these instructions, based on the potion of the framing that could be either the beginning of the prompt or the end. We mainly have two type of output structure instructions: JSON-based and simple free-form output. We also have simple one word answers or explicit reasoning.

The table 8 also provide the prompt used to extract the final answer from the responses provided by the model under CoT reasoning with free-form output. The judgment prompt was given to GPT4O-mini.

### B.2 Results without protected attributes

In this subsection, we present the additional results related to the four types of experiments based on the four instruction types, provided in the table 7.

Figure 4 presents the results using the second instruction type in the table 7. Accordingly, we can see that the results are comparable to the one-word output. Moreover, for the equal case, we can see that the DirErr rates even are increased compared to the one-word case. The upper row shows when framing where positioned at the beginning while then other row present the results when the framings where positioned at the end.

Figure 5 provides the results for the third instruction type in the table 7. This figure provides the results for both when the framings where at the beginning and at the end, compared to the 3 that provides only the end cases for the two CoT instruction types.

Finally, the figure 6 presents the results of the fourth instruction type in the table 7. We can see that there is not much difference between the beginning and end cases in general. However, there are patterns of difference like the neutral case for sonnet 3.7. For the more case, we can see that there are also some differences such the larger error rates in the beginning case.

### B.3 Results with Protected attributes

This section provides an important part of our results. We, here, present the results when the set of protected attributes like gender and race included in our experiments. As described in the paper, we only tested the first instruction type in the table 7 due to the hight costs.

We here present two types of results. The set of tables for sonnet 3.7 and GPT4O-mini similar to the Table 1 as well as the figures that explore the framing effects along with the protected attributes based on the categories. Tables 9 and 10 contain the DirErr percentages for the Less and Equal label, while including the protecting attribute values. The tables 11, 12, and 13 show the similar values for GPT4O-mini. It is observed that the pattern is similar across the two models and as explained in the main paper, there is a shift toward the more class. Moreover, we present the results of these two models across categories that we have in our resource to capture potential types of categorical biases and see how they interact with the framing effects when demographic features are involved. Our observation of the figures 7, 8, 9, and 10 are related to the sonnet model. In these figures, we can see the comparison between the times when the framings are placed at the beginning of the prompt as well as the times they are located at the end. As we can see, there are patterns of difference between these two locations, such as the different ranges of values for the less label as DirErr for African.

The figures 11, 12, 13, and 14 provide similar information for GPT4O-mini. In terms of specific categorical differences across genders and races, we can see that for example sonnet has a larger DirrErr values for shopping for man than woman, meaning that it associates such activity with man less that woman. Moreover, for personal care category, the DirrErr as more is larger for woman, suggesting the potential bias of the model toward this category and women. Also, shopping DirrErr as less is larger for Africans than Asians as well as Hispanics. Interestingly, the travel category for white people has a larger less DirrErr than black people. For GPT4o-mini, we can see that DirrErr as equal is even less than the sonnet model. This suggest that the model is general is more biased toward more or less values.

| Instruction Type | Instruction Text |
|---|---|
| One-word Answer | A) less B) more C) equal<br>ONLY RETURN ONE WORD THAT IS THE TRUE OPTION'S TEXT [less, more, or equal], WITHOUT ANY EXPLANATION. |
| JSON Answer | A) less B) more C) equal<br>Return only a JSON object with a single field 'answer', whose value is one of: less, more, or equal.<br>Example: {"answer": "less"}. |
| CoT | A) less B) more C) equal<br>Let's think step by step. |
| CoT with JSON | A) less B) more C) equal \| Return only a JSON object with two fields:<br>- 'reasoning': a concise explanation of how you reached your conclusion.<br>- 'answer': one of the following strings: less, more, or equal.<br>Example:<br>{<br>"reasoning": "Person B spent more time lifting weights based on both sentences.",<br>"answer": "more"<br>}<br>Let's think step by step. |

Table 7: Instruction formats used across evaluations. Models were required to respond under different format constraints ranging from one-word output to structured reasoning with chain-of-thought.

| |
|---|
| **System Prompt**: You are a helpful assistant. Your job is to judge a model's response.<br>Based on the provided text, how does the amount of {studied_quantity} person B spends on {task} compare to that of person A?<br>Only answer with: less, more, or equal. |

Table 8: The judgement prompt used for the GPT4o-mini to provide the final answer of CoT experiments..

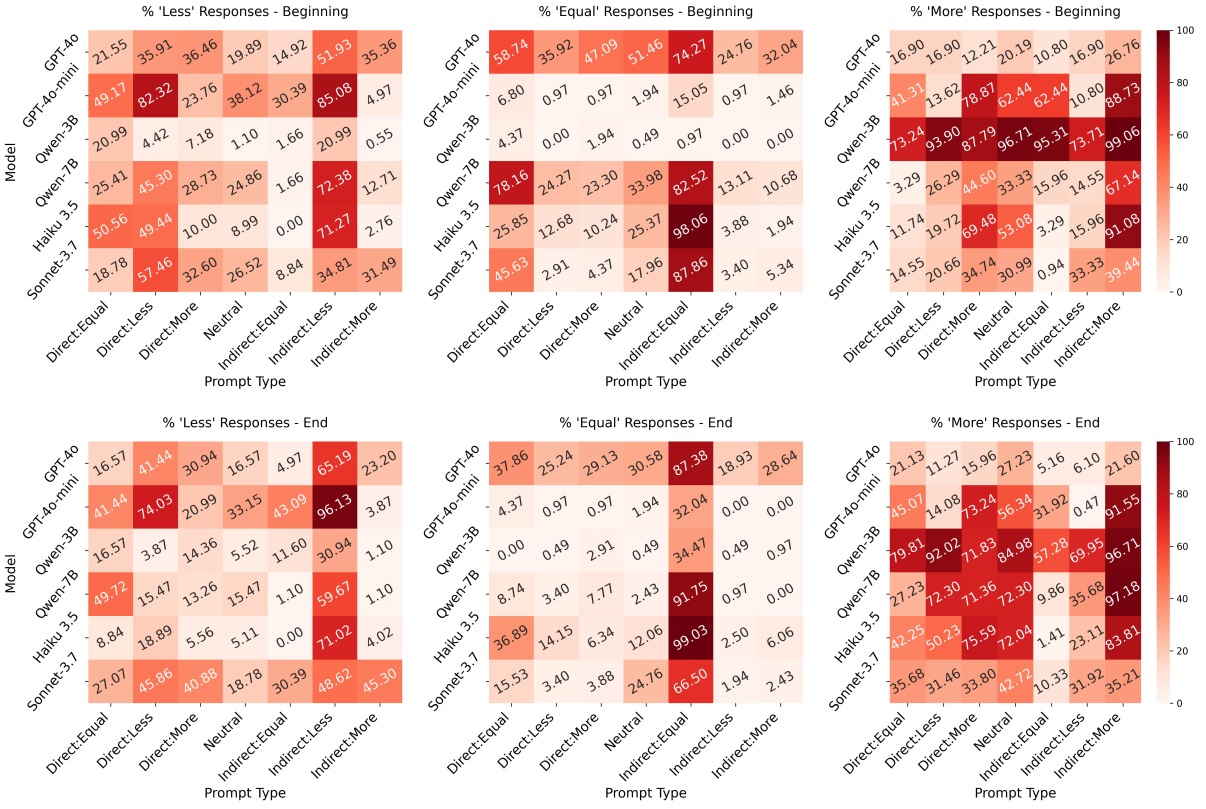


Figure 4: Directional error percentages (DirErr %) for JSON-formatted answers (the second instruction type) under framing variations. Each heat-map shows a single error direction—the proportion of all opportunities in which a model wrongly answers Less (left), Equal (center), or More (right). Columns are the seven prompt variants (Neutral, Direct, Indirect); rows are the six models. Darker cells indicate stronger drift toward that label. The upper trio uses prompts with the framing sentence at the beginning of the input, the lower trio with the framing at the end.

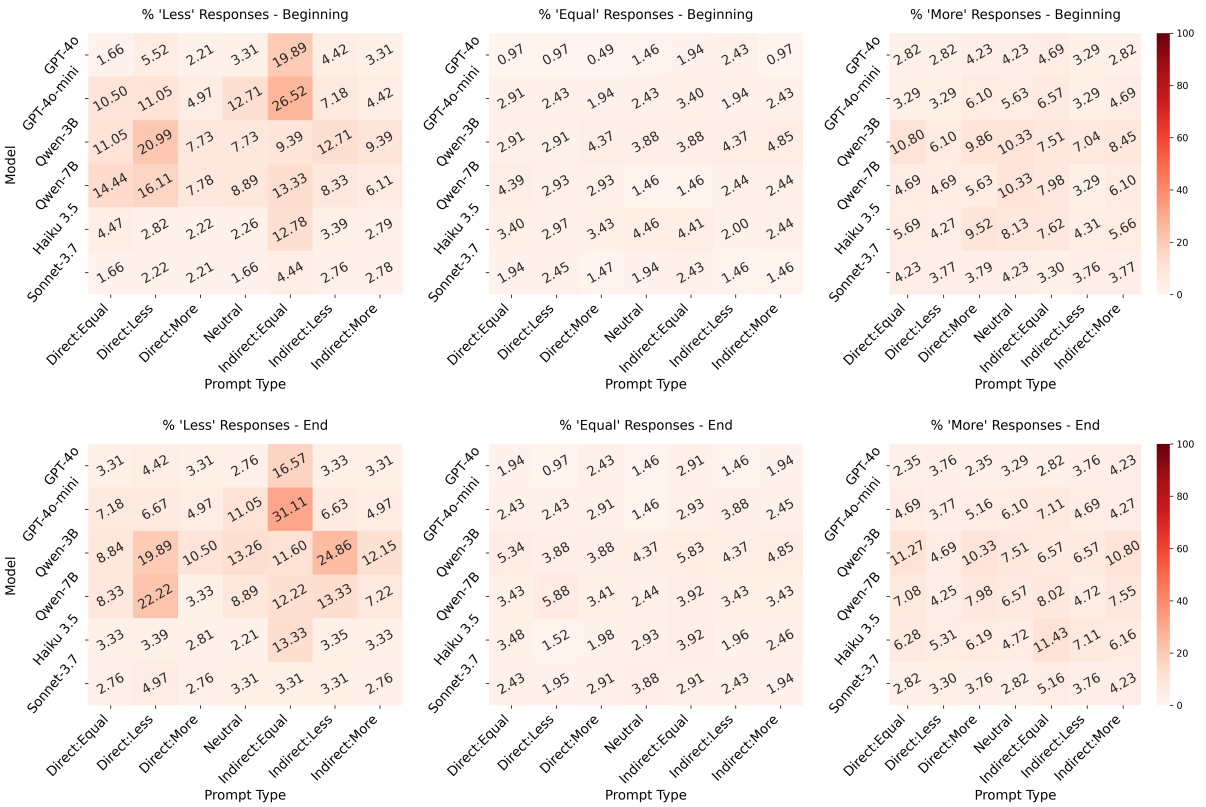


Figure 5: Directional error percentages (DirErr % under chain-of-thought prompting (the third instruction type). Top row: framing variations are placed at the beginning; bottom row: framing variations are placed at the end. Each heat-map shows one error direction—Less (left), Equal (center), or More (right). Columns are the seven prompt variants; rows are the six models; darker cells indicate stronger drift toward that label.

# C Appendix: Additional Metrics

## C.1 Results for simple one-word scenario

## C.2 Results for JSON one-word evaluation

## C.3 Results for simple CoT scenario

## C.4 Results for JSON-based CoT scenario

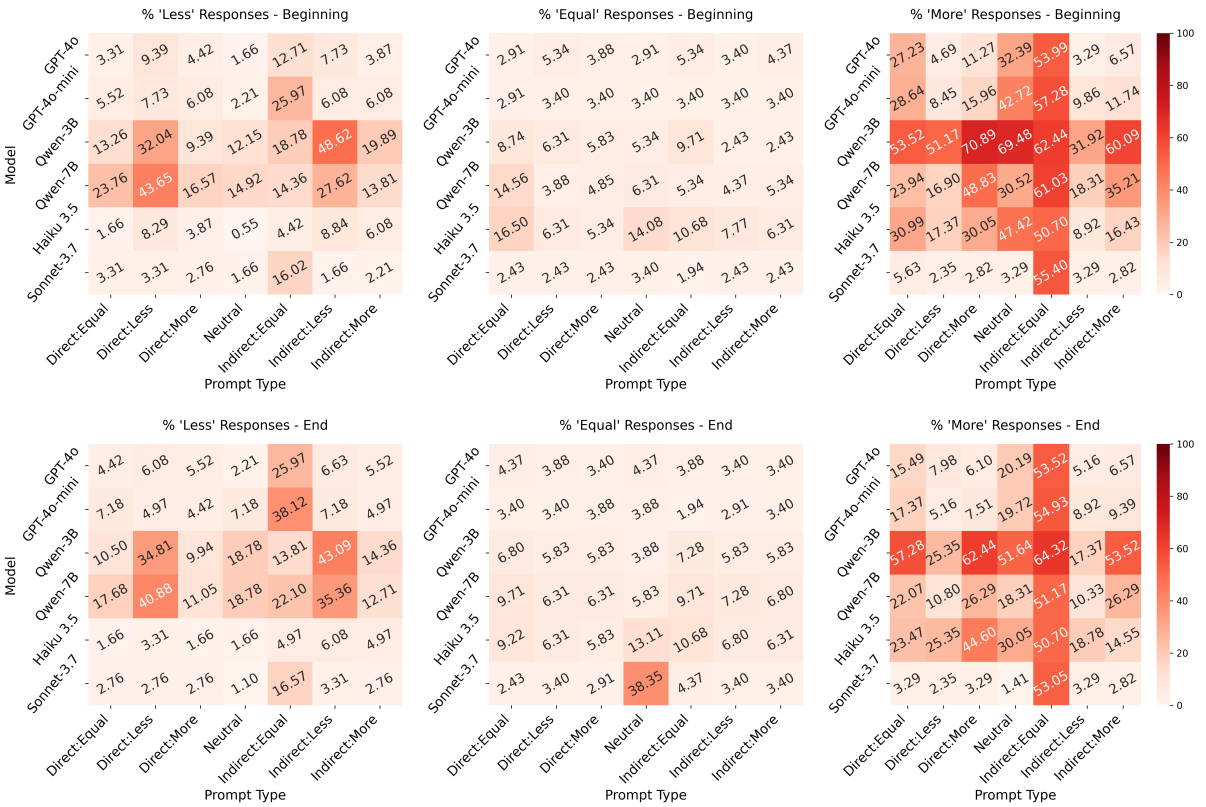


Figure 6: Directional error percentages (DirErr % under chain-of-thought prompting (the fourth instruction type) with JSON answers. Top row: framing variations are placed at the beginning; bottom row: framing variations are placed at the end. Each heat-map shows one error direction—Less (left), Equal (center), or More (right). Columns are the seven prompt variants; rows are the six models; darker cells indicate stronger drift toward that label.

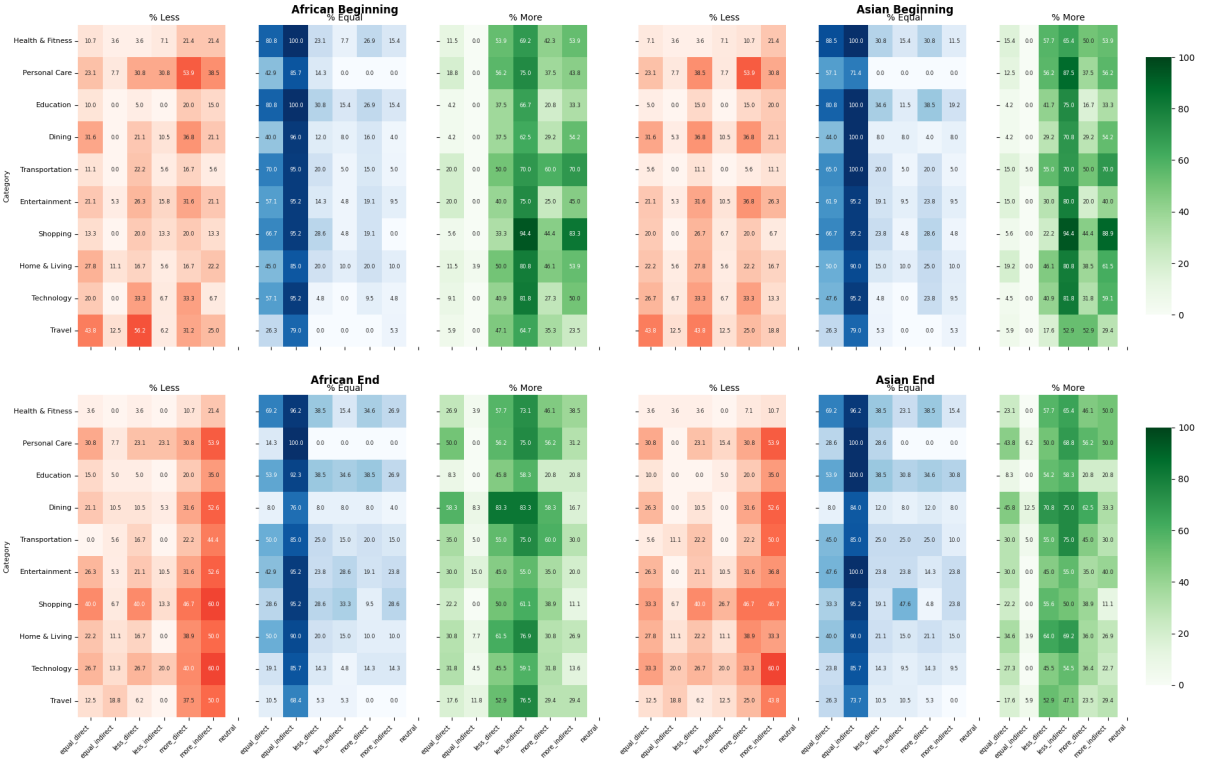


Figure 7: DirErr % for sonnet 3.7, the best model on average while including Asian and African races, when the framing variations are positioned at the beginning and end of the prompt.

| Framing | Std | Af | As | H | Wh | B | M | W |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect (End) | 24.31 | 7.73 | 6.63 | 3.87 | 13.81 | 7.18 | 4.42 | 3.87 |
| equal:Indirect (Begin) | 2.21 | 3.87 | 4.42 | 4.42 | 7.73 | 3.31 | 4.97 | 4.42 |
| equal:Direct (End) | 35.36 | 18.23 | 19.34 | 16.57 | 13.81 | 11.60 | 20.99 | 22.65 |
| equal:Direct (Begin) | 23.20 | 20.44 | 19.34 | 15.47 | 20.99 | 13.81 | 32.60 | 33.70 |
| less:Indirect (End) | 19.89 | 6.08 | 8.84 | 7.73 | 4.97 | 3.31 | 11.60 | 11.05 |
| less:Indirect (Begin) | 13.26 | 9.39 | 6.63 | 6.08 | 8.84 | 6.63 | 22.10 | 19.89 |
| less:Direct (End) | 41.44 | 15.47 | 16.02 | 14.92 | 12.71 | 12.71 | 34.81 | 30.94 |
| less:Direct (Begin) | 34.25 | 21.55 | 24.86 | 18.78 | 30.39 | 21.55 | 27.62 | 37.02 |
| more:Indirect (End) | 46.41 | 45.86 | 39.78 | 37.57 | 50.83 | 32.04 | 39.78 | 40.33 |
| more:Indirect (Begin) | 30.94 | 18.78 | 18.78 | 19.34 | 24.31 | 22.65 | 24.86 | 19.34 |
| more:Direct (End) | 46.96 | 29.28 | 27.07 | 25.97 | 28.18 | 21.55 | 27.62 | 32.04 |
| more:Direct (Begin) | 35.36 | 27.07 | 24.31 | 27.07 | 28.18 | 22.65 | 27.62 | 25.97 |
| neutral (End) | 12.15 | 10.50 | 14.36 | 11.60 | 9.94 | 9.39 | 15.47 | 17.68 |
| neutral (Begin) | 16.02 | 14.36 | 14.36 | 12.71 | 16.57 | 6.63 | 18.78 | 17.68 |

Table 9: DirErr rates (%) for errors as Less for Sonnet 3.7 model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

| Measurement | Std | Af | As | H | Wh | B | M | W |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect (End) | 75.73 | 87.86 | 90.78 | 93.69 | 86.41 | 92.72 | 89.81 | 91.26 |
| equal:Indirect (Begin) | 94.66 | 93.69 | 94.66 | 94.17 | 89.81 | 92.23 | 92.23 | 94.66 |
| equal:Direct (End) | 31.55 | 36.89 | 38.83 | 40.78 | 53.88 | 60.19 | 43.20 | 39.81 |
| equal:Direct (Begin) | 57.28 | 58.74 | 60.19 | 59.71 | 62.62 | 65.05 | 36.89 | 33.98 |
| less:Indirect (End) | 9.22 | 17.48 | 20.87 | 18.93 | 20.87 | 33.98 | 10.68 | 9.22 |
| less:Indirect (Begin) | 15.53 | 6.31 | 7.28 | 6.31 | 5.34 | 6.80 | 7.28 | 6.31 |
| less:Direct (End) | 12.14 | 22.33 | 23.41 | 21.95 | 26.70 | 33.17 | 16.99 | 17.96 |
| less:Direct (Begin) | 22.33 | 17.48 | 17.96 | 22.33 | 14.08 | 22.93 | 8.25 | 12.14 |
| more:Indirect (End) | 12.62 | 16.50 | 15.05 | 14.56 | 24.76 | 46.12 | 16.99 | 13.11 |
| more:Indirect (Begin) | 21.84 | 7.77 | 9.22 | 7.28 | 6.80 | 7.28 | 10.19 | 7.28 |
| more:Direct (End) | 15.05 | 17.48 | 19.02 | 17.56 | 24.76 | 33.66 | 17.96 | 15.53 |
| more:Direct (Begin) | 22.82 | 16.99 | 21.36 | 23.79 | 22.33 | 27.18 | 9.22 | 8.25 |
| neutral (End) | 33.98 | 42.44 | 43.84 | 40.98 | 50.49 | 48.78 | 42.23 | 33.98 |
| neutral (Begin) | 42.72 | 48.78 | 59.71 | 61.65 | 49.51 | 67.96 | 31.07 | 31.07 |

Table 10: DirErr rates (%) for errors as Equal for Sonnet 3.7 model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

| Condition | Std | M | W | Af | As | H | Wh | B |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect(Begin) | 56.34 | 80.28 | 77.00 | 79.34 | 77.93 | 77.46 | 79.34 | 79.81 |
| more:Indirect(End) | 95.77 | 99.53 | 99.06 | 99.06 | 100.00 | 99.53 | 99.06 | 99.53 |
| equal:Indirect(End) | 36.15 | 59.62 | 64.32 | 47.89 | 39.44 | 43.19 | 51.64 | 53.05 |
| more:Direct(Begin) | 74.18 | 90.14 | 91.08 | 84.04 | 84.98 | 87.79 | 90.14 | 84.04 |
| more:Direct(End) | 81.69 | 93.90 | 96.24 | 82.16 | 84.51 | 87.79 | 89.20 | 84.51 |
| more:Indirect(Begin) | 86.38 | 95.77 | 94.84 | 91.08 | 93.43 | 92.96 | 91.55 | 89.20 |
| neutral(Begin) | 63.38 | 81.69 | 77.93 | 78.40 | 75.59 | 76.53 | 86.38 | 78.40 |
| neutral(End) | 69.48 | 88.73 | 85.92 | 64.32 | 69.48 | 65.73 | 86.38 | 69.48 |
| equal:Direct(End) | 53.99 | 66.20 | 63.38 | 33.33 | 30.05 | 23.94 | 48.83 | 28.64 |
| equal:Direct(Begin) | 44.13 | 77.93 | 71.83 | 72.30 | 70.42 | 65.73 | 78.40 | 65.26 |
| less:Direct(Begin) | 5.63 | 25.82 | 21.60 | 33.80 | 35.21 | 34.74 | 54.93 | 36.15 |
| less:Indirect(End) | 0.47 | 1.88 | 0.94 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 |
| less:Direct(End) | 13.15 | 46.01 | 27.70 | 15.02 | 10.80 | 8.45 | 40.85 | 13.62 |
| less:Indirect(Begin) | 2.82 | 2.82 | 2.35 | 8.45 | 4.69 | 4.69 | 10.33 | 5.63 |

Table 11: DirErr rates (%) for errors as More for GPT4O-mini model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

| Condition | Std | M | W | Af | As | H | Wh | B |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect(Begin) | 48.62 | 13.81 | 14.36 | 13.26 | 11.60 | 13.81 | 13.81 | 11.05 |
| more:Indirect(End) | 3.87 | 0.55 | 1.10 | 0.00 | 0.00 | 0.00 | 0.55 | 0.00 |
| equal:Indirect(End) | 70.17 | 19.34 | 16.57 | 30.94 | 31.49 | 25.97 | 27.07 | 25.41 |
| more:Direct(Begin) | 24.86 | 9.94 | 6.63 | 15.47 | 13.81 | 12.15 | 4.97 | 16.02 |
| more:Direct(End) | 16.02 | 7.18 | 1.66 | 19.34 | 18.23 | 13.26 | 11.60 | 16.02 |
| more:Indirect(Begin) | 8.29 | 2.21 | 3.31 | 7.73 | 6.08 | 6.08 | 6.08 | 9.94 |
| neutral(Begin) | 35.36 | 15.47 | 20.99 | 20.44 | 20.99 | 22.65 | 12.15 | 20.99 |
| neutral(End) | 27.07 | 12.15 | 14.36 | 35.91 | 32.04 | 27.07 | 14.36 | 26.52 |
| equal:Direct(End) | 44.75 | 36.46 | 30.94 | 64.64 | 67.40 | 72.38 | 53.04 | 68.51 |
| equal:Direct(Begin) | 46.41 | 16.02 | 23.20 | 19.34 | 22.65 | 23.76 | 16.02 | 26.52 |
| less:Direct(Begin) | 92.27 | 71.82 | 72.93 | 60.77 | 61.88 | 62.43 | 43.65 | 56.91 |
| less:Indirect(End) | 98.34 | 95.58 | 97.24 | 99.45 | 98.90 | 98.90 | 98.34 | 98.34 |
| less:Direct(End) | 82.87 | 62.43 | 71.27 | 80.11 | 83.43 | 83.43 | 56.91 | 78.45 |
| less:Indirect(Begin) | 95.03 | 93.92 | 94.48 | 86.74 | 91.71 | 90.06 | 87.29 | 90.61 |

Table 12: DirErr rates (%) for errors as Less for GPT4O-mini model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

| Condition | Std | M | W | Af | As | H | Wh | B |
|---|---|---|---|---|---|---|---|---|
| equal:Indirect(Begin) | 2.43 | 2.91 | 3.40 | 2.91 | 4.85 | 3.88 | 4.37 | 4.37 |
| more:Indirect(End) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| equal:Indirect(End) | 2.43 | 25.73 | 18.93 | 21.36 | 31.07 | 27.67 | 25.24 | 25.73 |
| more:Direct(Begin) | 0.00 | 0.00 | 0.00 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| more:Direct(End) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 |
| more:Indirect(Begin) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| neutral(Begin) | 0.00 | 0.00 | 0.49 | 0.49 | 0.49 | 0.00 | 0.00 | 0.00 |
| neutral(End) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 |
| equal:Direct(End) | 0.00 | 0.49 | 0.97 | 0.49 | 0.97 | 0.97 | 0.49 | 1.94 |
| equal:Direct(Begin) | 0.49 | 0.97 | 0.97 | 2.43 | 1.94 | 1.46 | 1.46 | 1.94 |
| less:Direct(Begin) | 0.00 | 0.00 | 0.00 | 0.49 | 0.49 | 0.97 | 0.00 | 0.49 |
| less:Indirect(End) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 |
| less:Direct(End) | 0.00 | 0.00 | 0.00 | 0.49 | 0.00 | 0.00 | 0.49 | 0.49 |
| less:Indirect(Begin) | 0.00 | 0.00 | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 13: DirErr rates (%) for errors as Equal for GPT4O-mini model, across demographic identity markers. Each row represents a distinct framing variant, defined by comparison target (More, Less, Equal), style (Indirect, Direct, Neutral), and position (Begin, End). Demographics: Std=Standard, M=Man, W=Woman, As=Asian, Af=African, H=Hispanic, Wh=White, B=Black.

Table 14: Results for simple on-word evaluation of `gpt4o`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 143 | 82 | 74 |
| equal_direct_end | 127 | 70 | 102 |
| equal_indirect_beginning | 235 | 37 | 27 |
| equal_indirect_end | 259 | 6 | 34 |
| less_direct_beginning | 119 | 94 | 86 |
| less_direct_end | 129 | 87 | 83 |
| less_indirect_beginning | 84 | 155 | 60 |
| less_indirect_end | 99 | 116 | 83 |
| more_direct_beginning | 148 | 100 | 51 |
| more_direct_end | 137 | 111 | 51 |
| more_indirect_beginning | 126 | 91 | 82 |
| more_indirect_end | 149 | 91 | 59 |
| simple_beginning | 110 | 68 | 121 |
| simple_end | 105 | 95 | 94 |

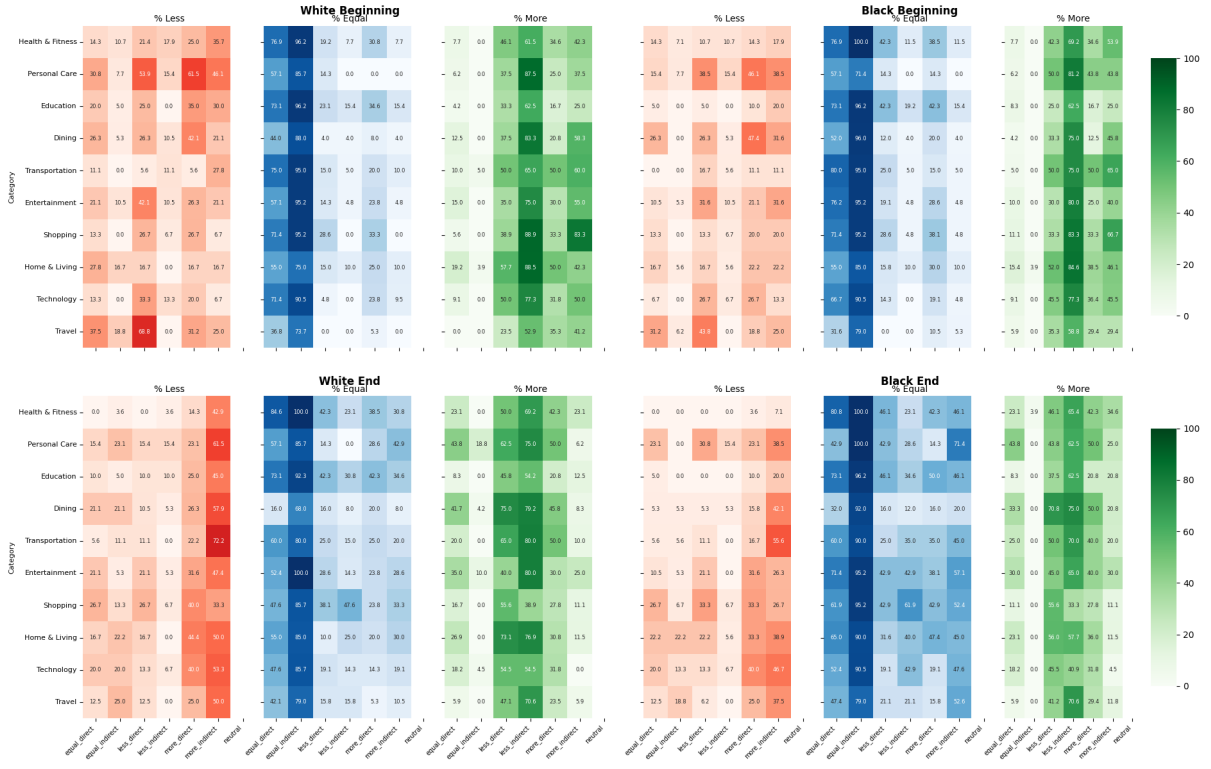


Figure 8: DirErr % for sonnet 3.7, the best model on average while including White and Black races, when the framing variations are positioned at the beginning and end of the prompt.

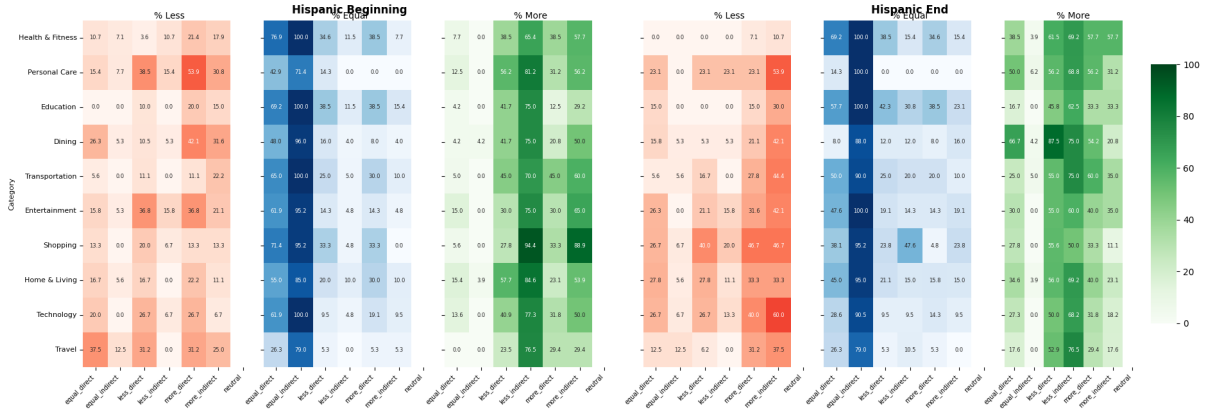


Figure 9: DirErr % for sonnet 3.7, the best model on average while including Hispanic race, when the framing variations are positioned at the beginning and end of the prompt.

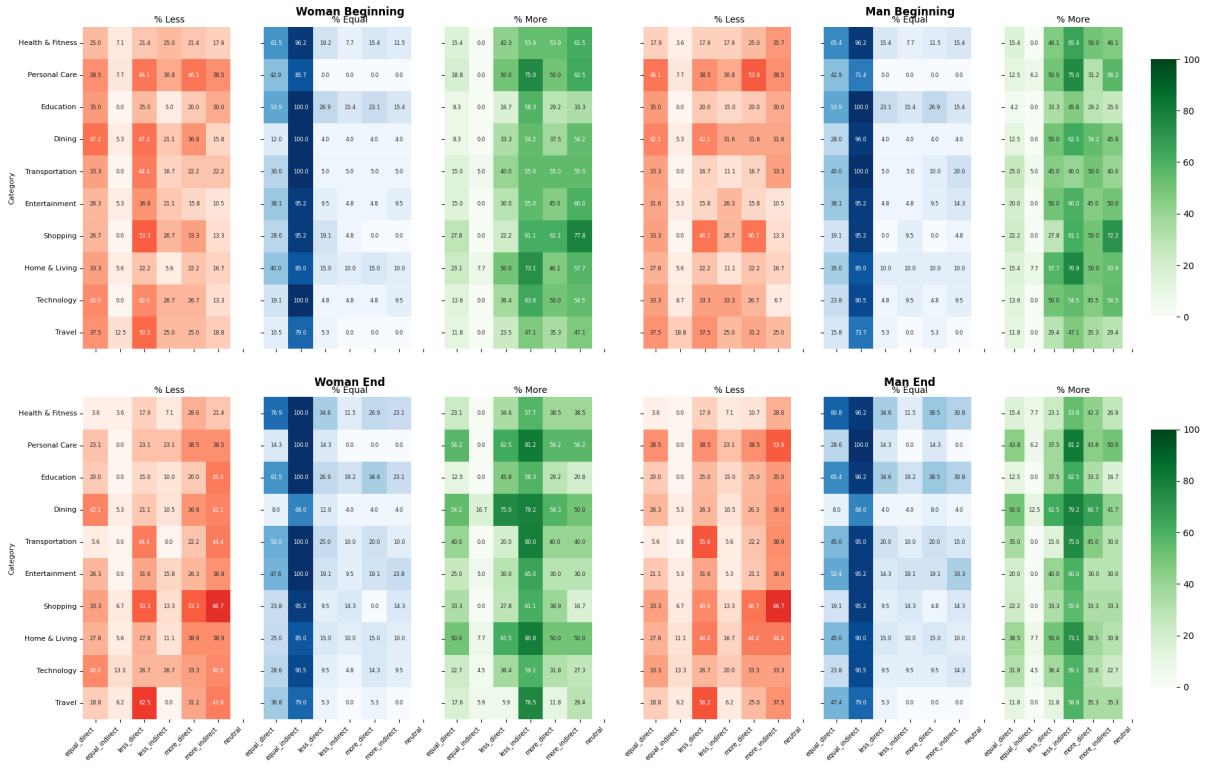


Figure 10: DirErr % for sonnet 3.7, the best model on average while including Woman and Man, when the framing variations are positioned at the beginning and end of the prompt.

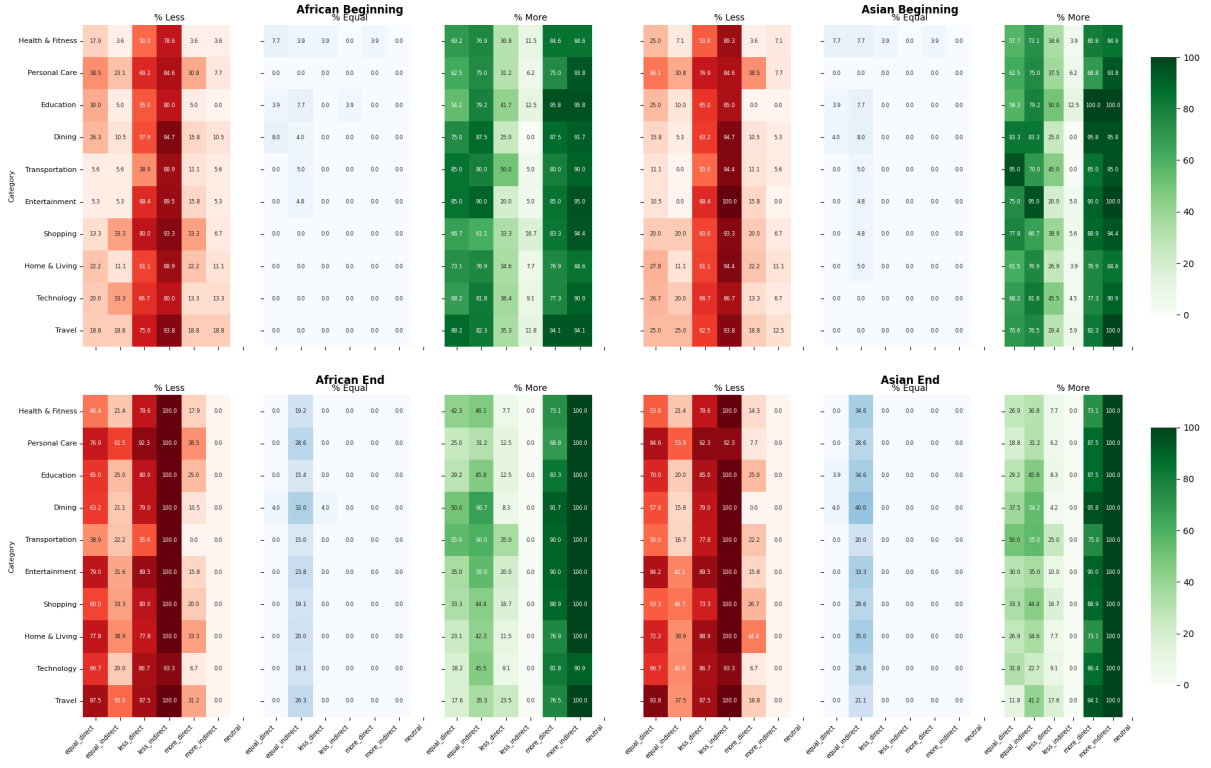


Figure 11: DirErr % for GPT4O-mini on average while including Asian and African races, when the framing variations are positioned at the beginning and end of the prompt.

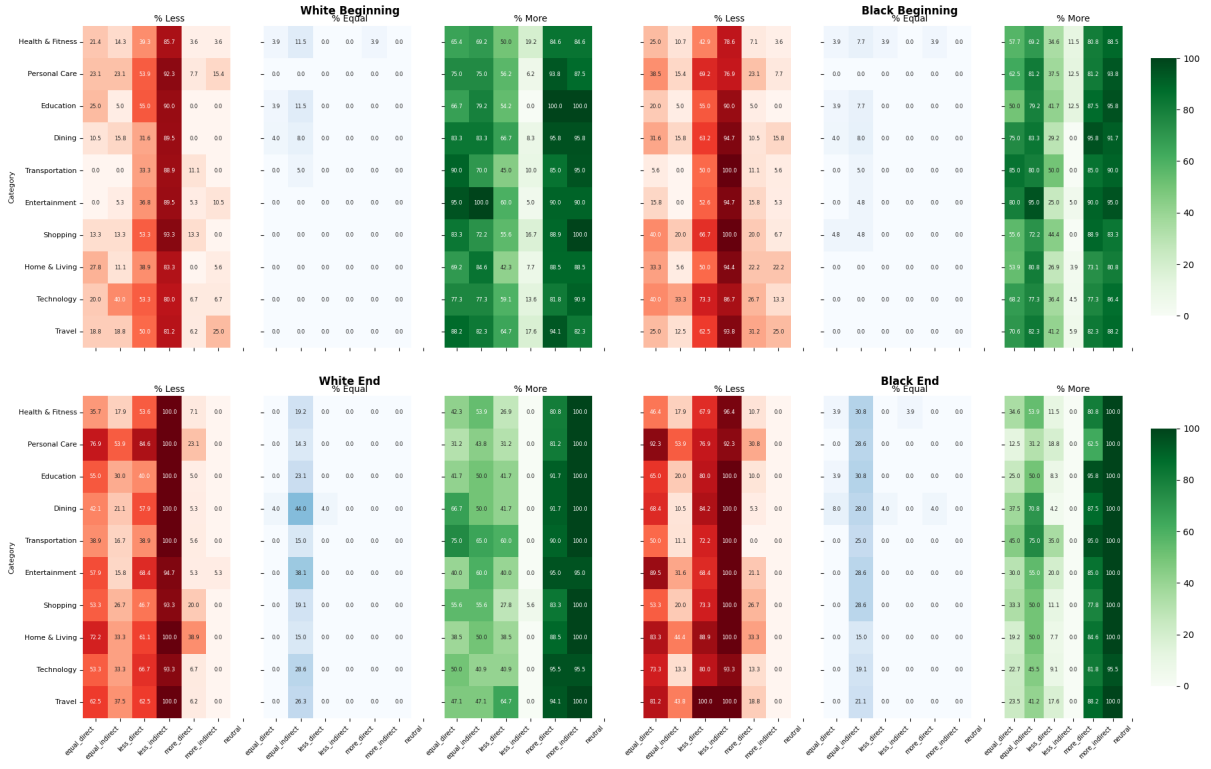


Figure 12: DirErr % for GPT4O-mini on average while including White and Black races, when the framing variations are positioned at the beginning and end of the prompt.

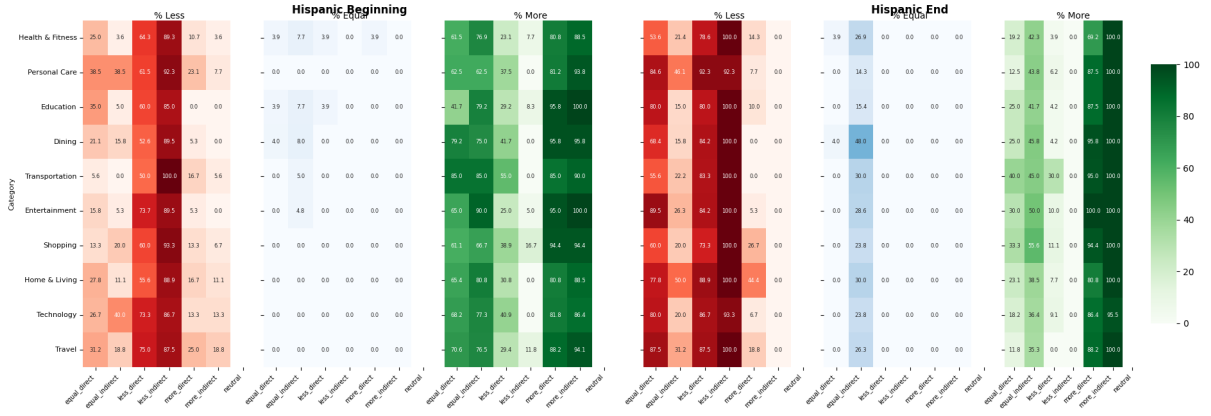


Figure 13: DirErr % for GPT4O-mini on average while including Hispanic race, when the framing variations are positioned at the beginning and end of the prompt.

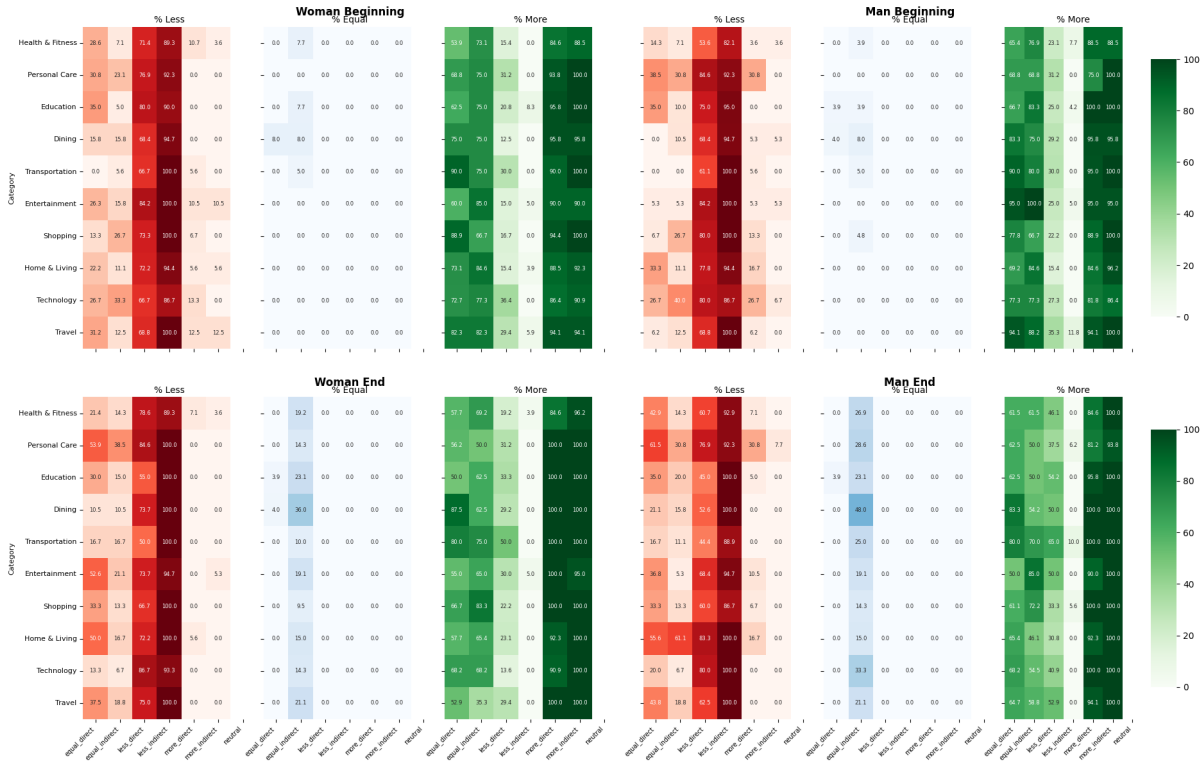


Figure 14: DirErr % for GPT4O-mini on average while including Woman and Man, when the framing variations are positioned at the beginning and end of the prompt.

Table 15: Results for simple on-word evaluation of `gpt4o-mini`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 11 | 145 | 144 |
| equal_direct_end | 6 | 128 | 166 |
| equal_indirect_beginning | 17 | 125 | 158 |
| equal_indirect_end | 14 | 187 | 99 |
| less_direct_beginning | 2 | 284 | 14 |
| less_direct_end | 1 | 257 | 42 |
| less_indirect_beginning | 2 | 290 | 8 |
| less_indirect_end | 1 | 297 | 2 |
| more_direct_beginning | 5 | 73 | 222 |
| more_direct_end | 2 | 48 | 250 |
| more_indirect_beginning | 4 | 32 | 264 |
| more_indirect_end | 1 | 12 | 287 |
| simple_beginning | 7 | 93 | 200 |
| simple_end | 3 | 80 | 217 |

Table 16: Results for simple on-word evaluation of `sonnet 3.7`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 195 | 82 | 23 |
| equal_direct_end | 118 | 119 | 63 |
| equal_indirect_beginning | 289 | 9 | 2 |
| equal_indirect_end | 232 | 64 | 4 |
| less_direct_beginning | 93 | 121 | 86 |
| less_direct_end | 56 | 137 | 107 |
| less_indirect_beginning | 73 | 54 | 173 |
| less_indirect_end | 43 | 66 | 191 |
| more_direct_beginning | 93 | 124 | 83 |
| more_direct_end | 68 | 156 | 76 |
| more_indirect_beginning | 85 | 109 | 106 |
| more_indirect_end | 58 | 150 | 92 |
| simple_beginning | 148 | 55 | 97 |
| simple_end | 112 | 34 | 154 |

Table 17: Results for simple on-word evaluation of `haiku 3.5`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 57 | 101 | 84 |
| equal_direct_end | 31 | 21 | 183 |
| equal_indirect_beginning | 181 | 4 | 20 |
| equal_indirect_end | 214 | 3 | 7 |
| less_direct_beginning | 19 | 142 | 48 |
| less_direct_end | 10 | 21 | 204 |
| less_indirect_beginning | 6 | 207 | 40 |
| less_indirect_end | 1 | 182 | 35 |
| more_direct_beginning | 14 | 15 | 198 |
| more_direct_end | 4 | 13 | 258 |
| more_indirect_beginning | 6 | 7 | 274 |
| more_indirect_end | 1 | 14 | 261 |
| simple_beginning | 22 | 12 | 106 |
| simple_end | 1 | 4 | 94 |

Table 18: Results for simple on-word evaluation of `qwen 7b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 278 | 12 | 9 |
| equal_direct_end | 0 | 76 | 224 |
| equal_indirect_beginning | 120 | 24 | 156 |
| equal_indirect_end | 113 | 8 | 179 |
| less_direct_beginning | 0 | 62 | 238 |
| less_direct_end | 0 | 4 | 296 |
| less_indirect_beginning | 1 | 211 | 88 |
| less_indirect_end | 0 | 39 | 261 |
| more_direct_beginning | 1 | 25 | 274 |
| more_direct_end | 0 | 35 | 265 |
| more_indirect_beginning | 0 | 5 | 295 |
| more_indirect_end | 0 | 1 | 299 |
| simple_beginning | 2 | 36 | 262 |
| simple_end | 0 | 32 | 268 |

Table 19: Results for simple on-word evaluation of `qwen 3b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 137 | 100 | 62 |
| equal_direct_end | 0 | 58 | 242 |
| equal_indirect_beginning | 15 | 76 | 187 |
| equal_indirect_end | 0 | 90 | 210 |
| less_direct_beginning | 0 | 284 | 15 |
| less_direct_end | 0 | 300 | 0 |
| less_indirect_beginning | 1 | 297 | 2 |
| less_indirect_end | 0 | 300 | 0 |
| more_direct_beginning | 0 | 4 | 296 |
| more_direct_end | 0 | 3 | 297 |
| more_indirect_beginning | 0 | 0 | 300 |
| more_indirect_end | 0 | 0 | 300 |
| simple_beginning | 0 | 16 | 284 |
| simple_end | 0 | 35 | 265 |

Table 20: Results for JSON one-word evaluation of `gpt-4o`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 188 | 51 | 61 |
| equal_direct_end | 125 | 79 | 96 |
| equal_indirect_beginning | 233 | 33 | 34 |
| equal_indirect_end | 266 | 11 | 23 |
| less_direct_beginning | 122 | 115 | 63 |
| less_direct_end | 93 | 147 | 60 |
| less_indirect_beginning | 93 | 154 | 53 |
| less_indirect_end | 72 | 208 | 20 |
| more_direct_beginning | 150 | 102 | 48 |
| more_direct_end | 102 | 120 | 78 |
| more_indirect_beginning | 110 | 96 | 94 |
| more_indirect_end | 111 | 91 | 98 |
| simple_beginning | 166 | 54 | 80 |
| simple_end | 105 | 77 | 118 |

Table 21: Results for JSON one-word evaluation of `gpt-4o-mini`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 33 | 139 | 128 |
| equal_direct_end | 24 | 124 | 152 |
| equal_indirect_beginning | 61 | 65 | 174 |
| equal_indirect_end | 105 | 105 | 90 |
| less_direct_beginning | 13 | 252 | 35 |
| less_direct_end | 14 | 238 | 48 |
| less_indirect_beginning | 13 | 260 | 27 |
| less_indirect_end | 6 | 293 | 1 |
| more_direct_beginning | 15 | 49 | 236 |
| more_direct_end | 14 | 58 | 228 |
| more_indirect_beginning | 14 | 13 | 273 |
| more_indirect_end | 11 | 10 | 279 |
| simple_beginning | 18 | 93 | 189 |
| simple_end | 16 | 99 | 185 |

Table 22: Results for JSON one-word evaluation of `sonnet 3.7`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 165 | 74 | 61 |
| equal_direct_end | 75 | 96 | 129 |
| equal_indirect_beginning | 271 | 21 | 8 |
| equal_indirect_end | 198 | 78 | 24 |
| less_direct_beginning | 30 | 196 | 74 |
| less_direct_end | 28 | 160 | 112 |
| less_indirect_beginning | 42 | 138 | 120 |
| less_indirect_end | 27 | 163 | 110 |
| more_direct_beginning | 39 | 129 | 132 |
| more_direct_end | 34 | 143 | 123 |
| more_indirect_beginning | 41 | 114 | 145 |
| more_indirect_end | 21 | 149 | 130 |
| simple_beginning | 81 | 105 | 114 |
| simple_end | 91 | 62 | 147 |

Table 23: Results for JSON one-word evaluation of `haiku 3.5`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 95 | 167 | 37 |
| equal_direct_end | 121 | 39 | 140 |
| equal_indirect_beginning | 290 | 1 | 9 |
| equal_indirect_end | 296 | 0 | 4 |
| less_direct_beginning | 56 | 175 | 68 |
| less_direct_end | 65 | 64 | 170 |
| less_indirect_beginning | 23 | 230 | 47 |
| less_indirect_end | 15 | 215 | 64 |
| more_direct_beginning | 47 | 41 | 211 |
| more_direct_end | 37 | 21 | 241 |
| more_indirect_beginning | 16 | 6 | 278 |
| more_indirect_end | 27 | 12 | 252 |
| simple_beginning | 94 | 26 | 175 |
| simple_end | 49 | 16 | 228 |

Table 24: Results for JSON one-word evaluation of `qwen 7b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 216 | 75 | 9 |
| equal_direct_end | 32 | 171 | 97 |
| equal_indirect_beginning | 250 | 7 | 43 |
| equal_indirect_end | 276 | 2 | 22 |
| less_direct_beginning | 70 | 140 | 90 |
| less_direct_end | 15 | 56 | 229 |
| less_indirect_beginning | 41 | 212 | 47 |
| less_indirect_end | 9 | 180 | 111 |
| more_direct_beginning | 71 | 97 | 132 |
| more_direct_end | 31 | 48 | 221 |
| more_indirect_beginning | 41 | 47 | 212 |
| more_indirect_end | 3 | 3 | 294 |
| simple_beginning | 99 | 89 | 112 |
| simple_end | 12 | 56 | 232 |

Table 25: Results for JSON one-word evaluation of `qwen 3b (alt)`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 15 | 54 | 231 |
| equal_direct_end | 1 | 50 | 249 |
| equal_indirect_beginning | 5 | 5 | 290 |
| equal_indirect_end | 101 | 23 | 176 |
| less_direct_beginning | 3 | 12 | 285 |
| less_direct_end | 5 | 12 | 283 |
| less_indirect_beginning | 1 | 64 | 235 |
| less_indirect_end | 3 | 86 | 211 |
| more_direct_beginning | 8 | 21 | 271 |
| more_direct_end | 14 | 55 | 231 |
| more_indirect_beginning | 1 | 1 | 298 |
| more_indirect_end | 3 | 4 | 293 |
| simple_beginning | 3 | 4 | 293 |
| simple_end | 5 | 30 | 265 |

Table 26: Results for simple CoT evaluation of `gpt4o`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 94 | 116 | 90 |
| equal_direct_end | 95 | 120 | 85 |
| equal_indirect_beginning | 93 | 143 | 64 |
| equal_indirect_end | 97 | 140 | 63 |
| less_direct_beginning | 88 | 122 | 90 |
| less_direct_end | 91 | 121 | 88 |
| less_indirect_beginning | 93 | 117 | 90 |
| less_indirect_end | 91 | 118 | 90 |
| more_direct_beginning | 90 | 116 | 94 |
| more_direct_end | 95 | 118 | 87 |
| more_indirect_beginning | 92 | 118 | 90 |
| more_indirect_end | 94 | 115 | 91 |
| simple_beginning | 93 | 114 | 93 |
| simple_end | 93 | 116 | 91 |

Table 27: Results for simple CoT evaluation of `gpt4o-mini`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 91 | 130 | 79 |
| equal_direct_end | 91 | 122 | 87 |
| equal_indirect_beginning | 96 | 150 | 54 |
| equal_indirect_end | 92 | 156 | 50 |
| less_direct_beginning | 92 | 129 | 79 |
| less_direct_end | 91 | 121 | 87 |
| less_indirect_beginning | 90 | 122 | 88 |
| less_indirect_end | 92 | 117 | 91 |
| more_direct_beginning | 92 | 114 | 94 |
| more_direct_end | 90 | 117 | 93 |
| more_indirect_beginning | 93 | 114 | 93 |
| more_indirect_end | 89 | 116 | 93 |
| simple_beginning | 93 | 128 | 79 |
| simple_end | 90 | 125 | 85 |

Table 28: Results for simple CoT evaluation of `sonnet 3.7`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 94 | 115 | 91 |
| equal_direct_end | 95 | 118 | 87 |
| equal_indirect_beginning | 96 | 118 | 85 |
| equal_indirect_end | 95 | 115 | 90 |
| less_direct_beginning | 94 | 116 | 88 |
| less_direct_end | 92 | 121 | 86 |
| less_indirect_beginning | 92 | 118 | 90 |
| less_indirect_end | 94 | 117 | 89 |
| more_direct_beginning | 92 | 114 | 92 |
| more_direct_end | 95 | 116 | 89 |
| more_indirect_beginning | 93 | 116 | 90 |
| more_indirect_end | 93 | 116 | 91 |
| simple_beginning | 94 | 115 | 91 |
| simple_end | 97 | 116 | 87 |

Table 29: Results for simple CoT evaluation of `haiku 3.5`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 92 | 116 | 90 |
| equal_direct_end | 95 | 105 | 94 |
| equal_indirect_beginning | 97 | 119 | 81 |
| equal_indirect_end | 95 | 115 | 87 |
| less_direct_beginning | 94 | 113 | 88 |
| less_direct_end | 91 | 110 | 90 |
| less_indirect_beginning | 92 | 111 | 90 |
| less_indirect_end | 91 | 110 | 96 |
| more_direct_beginning | 94 | 100 | 103 |
| more_direct_end | 92 | 108 | 95 |
| more_indirect_beginning | 92 | 110 | 96 |
| more_indirect_end | 92 | 112 | 93 |
| simple_beginning | 96 | 101 | 97 |
| simple_end | 97 | 107 | 95 |

Table 30: Results for simple CoT evaluation of `qwen 3b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 78 | 123 | 99 |
| equal_direct_end | 88 | 112 | 100 |
| equal_indirect_beginning | 87 | 120 | 93 |
| equal_indirect_end | 89 | 125 | 86 |
| less_direct_beginning | 81 | 146 | 73 |
| less_direct_end | 85 | 144 | 71 |
| less_indirect_beginning | 84 | 128 | 88 |
| less_indirect_end | 82 | 150 | 68 |
| more_direct_beginning | 87 | 113 | 100 |
| more_direct_end | 82 | 118 | 100 |
| more_indirect_beginning | 87 | 118 | 95 |
| more_indirect_end | 83 | 119 | 98 |
| simple_beginning | 85 | 113 | 102 |
| simple_end | 81 | 129 | 90 |

Table 31: Results for simple CoT evaluation of `qwen 7b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 86 | 135 | 78 |
| equal_direct_end | 92 | 116 | 90 |
| equal_indirect_beginning | 89 | 129 | 81 |
| equal_indirect_end | 94 | 123 | 81 |
| less_direct_beginning | 83 | 139 | 77 |
| less_direct_end | 95 | 145 | 58 |
| less_indirect_beginning | 87 | 126 | 86 |
| less_indirect_end | 90 | 130 | 78 |
| more_direct_beginning | 86 | 122 | 91 |
| more_direct_end | 92 | 108 | 99 |
| more_indirect_beginning | 89 | 118 | 92 |
| more_indirect_end | 89 | 114 | 95 |
| simple_beginning | 80 | 119 | 100 |
| simple_end | 87 | 121 | 91 |

Table 32: Results for JSON-based CoT evaluation of `gpt-4o`

| Column | Equal | Less | More |
|---|---|---|---|
| simple_beginning | 95 | 56 | 149 |
| simple_end | 96 | 81 | 123 |
| more_direct_beginning | 95 | 103 | 102 |
| more_direct_end | 93 | 114 | 93 |
| less_direct_beginning | 99 | 122 | 79 |
| less_direct_end | 92 | 115 | 93 |
| equal_direct_beginning | 97 | 65 | 138 |
| equal_direct_end | 97 | 93 | 110 |
| more_indirect_beginning | 99 | 110 | 91 |
| more_indirect_end | 92 | 116 | 92 |
| equal_indirect_beginning | 100 | 26 | 174 |
| equal_indirect_end | 95 | 54 | 151 |
| less_indirect_beginning | 94 | 124 | 82 |
| less_indirect_end | 93 | 122 | 85 |

Table 33: Results for JSON-based CoT evaluation of `gpt-4o-mini`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 89 | 71 | 140 |
| equal_direct_end | 92 | 94 | 114 |
| equal_indirect_beginning | 91 | 51 | 158 |
| equal_indirect_end | 90 | 75 | 135 |
| less_direct_beginning | 91 | 117 | 92 |
| less_direct_end | 95 | 115 | 90 |
| less_indirect_beginning | 87 | 114 | 99 |
| less_indirect_end | 89 | 115 | 96 |
| more_direct_beginning | 90 | 98 | 112 |
| more_direct_end | 96 | 109 | 95 |
| more_indirect_beginning | 87 | 111 | 102 |
| more_indirect_end | 88 | 111 | 101 |
| simple_beginning | 89 | 40 | 171 |
| simple_end | 92 | 88 | 120 |

Table 34: Results for JSON-based CoT evaluation of `sonnet 3.7`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 92 | 114 | 94 |
| equal_direct_end | 94 | 118 | 88 |
| equal_indirect_beginning | 95 | 32 | 173 |
| equal_indirect_end | 99 | 34 | 167 |
| less_direct_beginning | 94 | 119 | 87 |
| less_direct_end | 96 | 117 | 87 |
| less_indirect_beginning | 95 | 115 | 90 |
| less_indirect_end | 93 | 118 | 89 |
| more_direct_beginning | 94 | 118 | 88 |
| more_direct_end | 94 | 117 | 89 |
| more_indirect_beginning | 95 | 115 | 90 |
| more_indirect_end | 95 | 117 | 88 |
| simple_beginning | 99 | 111 | 90 |
| simple_end | 171 | 66 | 63 |

Table 35: Results for JSON-based CoT evaluation of `haiku 3.5`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 122 | 42 | 136 |
| equal_direct_end | 109 | 62 | 129 |
| equal_indirect_beginning | 105 | 12 | 183 |
| equal_indirect_end | 109 | 9 | 182 |
| less_direct_beginning | 100 | 92 | 108 |
| less_direct_end | 102 | 67 | 131 |
| less_indirect_beginning | 98 | 114 | 88 |
| less_indirect_end | 98 | 90 | 112 |
| more_direct_beginning | 96 | 60 | 144 |
| more_direct_end | 98 | 26 | 176 |
| more_indirect_beginning | 92 | 97 | 111 |
| more_indirect_end | 96 | 94 | 110 |
| simple_beginning | 109 | 16 | 175 |
| simple_end | 115 | 43 | 142 |

Table 36: Results for JSON-based CoT evaluation of `qwen 3b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 64 | 58 | 178 |
| equal_direct_end | 66 | 44 | 190 |
| equal_indirect_beginning | 64 | 44 | 192 |
| equal_indirect_end | 70 | 31 | 199 |
| less_direct_beginning | 39 | 102 | 159 |
| less_direct_end | 65 | 139 | 96 |
| less_indirect_beginning | 18 | 168 | 114 |
| less_indirect_end | 58 | 170 | 72 |
| more_direct_beginning | 42 | 34 | 224 |
| more_direct_end | 56 | 39 | 205 |
| more_indirect_beginning | 25 | 80 | 195 |
| more_indirect_end | 45 | 69 | 186 |
| simple_beginning | 46 | 38 | 216 |
| simple_end | 53 | 79 | 168 |

Table 37: Results for JSON-based CoT evaluation of `qwen 7b`

| Column | Equal | Less | More |
|---|---|---|---|
| equal_direct_beginning | 91 | 110 | 99 |
| equal_direct_end | 88 | 108 | 104 |
| equal_indirect_beginning | 80 | 30 | 190 |
| equal_indirect_end | 91 | 52 | 157 |
| less_direct_beginning | 59 | 172 | 69 |
| less_direct_end | 78 | 167 | 55 |
| less_indirect_beginning | 60 | 140 | 100 |
| less_indirect_end | 76 | 160 | 64 |
| more_direct_beginning | 57 | 77 | 166 |
| more_direct_end | 76 | 98 | 126 |
| more_indirect_beginning | 56 | 98 | 146 |
| more_indirect_end | 72 | 100 | 128 |
| simple_beginning | 65 | 101 | 134 |
| simple_end | 79 | 120 | 101 |