Directional Drift in Arithmetic Reasoning: How Comparative Cues Steer LLM Predictions

Anonymous ACL submission

Abstract

Arithmetic tasks are widely used to evaluate LLM reasoning because they provide objective truth. Yet accuracy alone cannot distinguish whether prompt sensitivity reflects random prediction instability or systematic semantic bias. We study this question in comparative arithmetic, where semantically equivalent framings can cue directional outcomes such as *more*, *less*, or *equal*. We introduce MATHCOMP, a balanced benchmark of 10,800 controlled comparison problems spanning multiple framing conditions and answer formats. Across ten LLMs, semantically equivalent comparative cues systematically bias predictions toward cued outcomes even when it is numerically incorrect. Notably, this effect persists even when cue words are removed from the answer options, suggesting that framing bias reflects broader semantic steering rather than simple label copying. Chain-of-thought reasoning mitigates the effect substantially, but residual directional drift remains. These findings suggest that prompt sensitivity in arithmetic reasoning can exhibit structured semantic bias toward linguistically cued outcomes, with implications for robustness evaluation and benchmark design. [1]

1 Introduction

Arithmetic reasoning tasks are widely used to evaluate LLMs because they provide objective ground truth and allow model behavior to be measured precisely. Yet prior work has shown that even in these settings, semantically equivalent prompt variations can substantially affect model predictions (Sclar et al., 2024; Razavi et al., 2025; Yang et al., 2022; Li et al., 2024; Sivakumar and Moosavi, 2023). These effects are typically evaluated through aggregate accuracy, leaving open the question of whether prompt sensitivity mainly reflects random prediction instability or more systematic changes in model behavior. Comparative arithmetic provides a useful setting for studying this distinction because prediction outcomes are inherently directional. Questions involving *more*, *less*, or *equal* make it possible to analyze not only whether prompts change model outputs, but whether they systematically steer predictions toward semantically cued outcomes. For example, if a model is asked whether Person 1 scored *more* than Person 2, does the word "more" itself increase the tendency to predict that outcome, even when the underlying quantities support the opposite conclusion?

This perspective complements prior robustness-oriented evaluations of arithmetic reasoning by focusing on the directional structure of model errors rather than accuracy alone. It also relates to work on framing and cognitive biases in LLMs (Lin and Ng, 2023; Flusberg and Holmes, 2024; Wu and Zheng, 2025; Itzhak et al., 2024), which has largely examined subjective or preference-based tasks without objective ground truth. Comparative arithmetic instead provides a controlled setting where directional shifts can be measured against unambiguous numerical answers.

To investigate this question, we introduce MATHCOMP, a controlled benchmark for studying directional framing effects in arithmetic comparison. Each scenario appears under a neutral formulation and multiple semantically equivalent framing variants that manipulate the comparative cue (*more*, *less*, *equal*), its placement (direct vs. contextual), answer format, and option ordering (Table 1). The benchmark is balanced across comparison labels and includes both fixed and randomized answer ordering to account for answer-position effects.

Main findings.

- Semantically equivalent comparative cues systematically steer predictions toward the cued outcome, producing strong directional error patterns rather than random instability.

[1] The resource available at: https://anonymous.4open.science/r/directional_shift-2090

Scenario. Person 1 scored 1 point in the 100m dash, 7 points in the long jump, and 6 points in the shot put. Person 2 scored 1 point in the 100m dash, 4 points in the long jump, and 8 points in the shot put. **Gold label:** More (14 vs. 13)

Condition	Format	Question text (the scenario text is prepended to every row)
Neutral	MLE	How does the total score of Person 1 compare to that of Person 2?
Direct-More	MLE	Does Person 1 score **more** than Person 2 in total?
	WHO	Who scores more in total — Person 1, Person 2, or did they score equally?
	YN	Does Person 1 score **more** than Person 2 in total?
Direct-Less	MLE	Does Person 1 score **less** than Person 2 in total?
	WHO	Who scores less in total — Person 1, Person 2, or did they score equally?
	YN	Does Person 1 score **less** than Person 2 in total?
Direct-Equal	MLE	Does Person 1 score the **same** as Person 2 in total?
	YN	Does Person 1 score the **same** as Person 2 in total?
Indirect-More	MLE	Considering that Person 1 scores **more** than Person 2 in long jump, how does the total score of Person 1 compare to that of Person 2?
	WHO	Considering that Person 1 scores **more** than Person 2 in long jump, who scores more in total?
	YN	Considering that Person 1 scores **more** than Person 2 in long jump, does Person 1 score more in total?
Indirect-Less	MLE	Considering that Person 1 scores **less** than Person 2 in shot put, how does the total score of Person 1 compare to that of Person 2?
	WHO	Considering that Person 1 scores **less** than Person 2 in shot put, who scores less in total?
	YN	Considering that Person 1 scores **less** than Person 2 in shot put, does Person 1 score less in total?
Indirect-Equal	MLE	Considering that Person 1 scores the **same** as Person 2 in 100m dash, how does the total score of Person 1 compare to that of Person 2?
	YN	Considering that Person 1 scores the **same** as Person 2 in 100m dash, does Person 1 score the same in total?

Table 1: **[Dataset schema, one scenario shown across framings × formats]** Example MATHCOMP prompts for a single arithmetic comparison scenario. Each model input consists of the ***scenario*** text, a question row corresponding to a framing condition, and the answer options of the associated format. To save space, option lists are omitted from individual rows: MLE uses permutations of More, Less, Equal, WHO uses permutations of Person 1, Person 2, Equal, and YN uses Yes, No. Direct framings place the comparative cue inside the question itself, while Indirect framings introduce the cue in a preceding context sentence. MLE is defined for all seven framing conditions; WHO and YN are defined only for framed conditions because the Neutral setting contains no comparison cue. Bold marks the framing cue word. Shading: Neutral, Direct, Indirect.

- The magnitude of the effect depends strongly on presentation. Direct cues inside the question produce substantially larger shifts than indirect contextual cues, while fixed option ordering amplifies the effect further by aligning framing and positional biases.
- Changing the answer format does not eliminate the phenomenon. Even when comparison labels are removed from the answer options, models still drift toward the entity associated with the cue, suggesting broader semantic steering beyond simple label copying.
- Chain-of-thought reasoning substantially reduces directional drift, but residual effects remain even when framing cues are removed from the answer space, suggesting that comparative framing influences model behavior beyond surface-level label matching.
- Together, these results suggest that robustness evaluations based solely on aggregate accuracy can obscure systematic structure in LLM reasoning errors, motivating more fine-grained analyses of how linguistic form shapes model behavior even in tasks with objective ground truth.

Contributions. (1) We introduce a directional perspective on prompt sensitivity in arithmetic reasoning, focusing not only on whether prompt variations change predictions, but on how they systematically shape the direction of model errors. (2) We present MATHCOMP, a controlled benchmark for studying framing-induced directional drift across comparative cues, answer formats, and option-ordering regimes. (3) We show that semantically equivalent comparative framing can induce structured semantic drift across models and prompting settings, including in formats where the framing cue is absent from the answer options, suggesting that the effect extends beyond surface-level label matching.

2 Related Work

Prompt sensitivity and robustness. LLM predictions can change substantially under semantically equivalent prompt variations, including paraphrasing, formatting changes, and reasoning perturbations (Gu et al., 2023; Sclar et al., 2024; Voronov et al., 2024; Mizrahi et al., 2024; Sun et al., 2024). In mathematical reasoning, prior work has studied robustness to paraphrasing, numerical perturbations, and template variation (Yang et al., 2022; Li

et al., 2024; Sivakumar and Moosavi, 2023), while related work has examined failures in quantifier and comparative reasoning (Gupta, 2023; Collacciani et al., 2024). Existing evaluations, however, primarily measure robustness through aggregate accuracy or consistency metrics. Our work instead studies the directional structure of model errors, asking whether semantically equivalent prompts systematically steer predictions toward particular outcomes rather than merely increasing instability.

Framing and cognitive biases in LLMs. Human judgments are known to depend on linguistic framing and presentation effects (Druckman, 2001; Gong et al., 2013). Similar framing and cognitive-bias patterns have been observed in LLMs across opinion, decision-making, negotiation, and consumer-oriented tasks (Lin and Ng, 2023; Flusberg and Holmes, 2024; Wu and Zheng, 2025; Itzhak et al., 2024). In contrast, comparative arithmetic provides objective and verifiable numerical ground truth, allowing framing-induced directional shifts to be analyzed independently of subjective preference or ambiguity.

Reasoning formats and answer representation. LLM behavior is sensitive not only to prompt wording, but also to answer representation and evaluation format. Prior work has shown that multiple-choice performance can vary under option reordering, option-ID preferences, and verbalizer choices (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Wang et al., 2024; Zhao et al., 2021). Our results extend these findings by showing that comparative framing interacts strongly with answer representation: removing cue words from the answer options changes the form of the error, but does not eliminate the directional effect.

Reasoning elicitation and robustness. Reasoning elicitation methods including chain-of-thought prompting can improve mathematical reasoning (Wei et al., 2022; Wang et al., 2023). Prior work mostly evaluates these methods through aggregate accuracy gains. In contrast, we analyze how reasoning elicitation affects framing-induced error patterns.

3 The MATHCOMP Dataset

MATHCOMP is a controlled benchmark for studying directional framing effects in comparative arithmetic. Each item presents two individuals accumulating quantities across several sub-tasks, followed by a comparison question asking whether one total is *more*, *less*, or *equal* to the other. The dataset contains 10,800 items derived from 144 base scenarios instantiated across multiple framing conditions, answer formats, and option-ordering settings.

3.1 Design Axes and Balance

The dataset is balanced across five axes:

- **Topic**: sports, housekeeping, education, transportation (36 base scenarios each).
- **Length**: 3, 4, 5, or 6 sub-tasks.
- **Digit count**: one-, two-, or three-digit values.
- **True label**: *more*, *less*, *equal*, with equal frequency (3,600 items each).
- **Arithmetic variant**: additive, deductive, and mixed-operation structures.

Each base scenario is instantiated under a neutral formulation and multiple equivalent framed variants that manipulate comparative cues while preserving the underlying arithmetic computation. MLE includes both neutral and framed prompts, whereas WHO and YN are defined only for framed conditions because the neutral setting contains no comparison cue. To control for answer-position effects, all multiple-choice formats are evaluated under both fixed and randomized option ordering.

Each base scenario uses one of three arithmetic variants. V1 contains additive operations, V2 has one deduction, V3 combines addition, multiplication, and division. Numerical values are sampled subject to two constraints: the target comparison label must hold for the final totals, and no individual sub-task may independently reveal the answer. Appendix A has the full procedure.

3.2 Scenario and Framing Generation

Each scenario is a short two-person comparison task generated from topic-specific templates spanning sports, housekeeping, education, and transportation. Scenarios describe quantities accumulated across several sub-tasks using domain-specific units (e.g., points, minutes, kilometres).

Each base scenario is instantiated under one Neutral condition and six framed conditions (Table 1). Neutral questions ask for a plain comparison without directional cues. Framed variants manipulate two axes: *placement* and *cue word*. In Direct framing, the cue word (*more*, *less*, or *equal*) appears inside the comparison question itself. In Indirect framing, the cue appears in a preceding context sentence stating a true sub-task-level comparison be-

fore the final question. Cue sentences are selected so that they remain locally true while potentially misleading about the overall totals.

3.3 Option Ordering

MLE (options: More, equal, Less) and WHO (options: Person 1, Person 2, Equal) have six possible option orders. YN (options: Yes, No) has two. Under the balanced condition we pick one order per (topic, length, digit count, variant) tuple with a deterministic hash. So the correct answer lands in slot A about one-sixth of the time and the other five-sixths spread over the rest. Under the fixed condition we use one order on every item. MLE is A. MORE, B. LESS, C. EQUAL. WHO is A. PERSON 1, B. PERSON 2, C. EQUAL. YN is A. YES, B. NO. With fixed options, cue word does not always sit in slot A. For example, under Direct-Less the cue is "less" but slot A is still MORE.

3.4 Answer Formats

We evaluate each framing condition under three answer formats designed to vary the relationship between framing cues and the answer space.

MLE (label selection). The model selects MORE, LESS, or EQUAL while presented in a balanced way with one of six option orders. Because the framing cue also appears in the answer space, this format directly tests cue-aligned label selection.

WHO (entity selection). The model selects PERSON 1, PERSON 2, or EQUAL. The question may still contain a comparative cue (e.g., "who scores *more*?"), but the cue word is not itself an answer option. This format therefore removes direct label copying: to follow the cue, the model must map the comparison direction to the corresponding entity.

YN (binary verification). The model answers YES or NO to a framed comparison question (e.g., "Does Person 1 score **more**? "). This format allows us to measure whether comparative cues increase the tendency to answer YES even when it is not.

3.5 Directional Error Analysis

Directional error. For MLE and WHO, we measure directional error using $\mathrm{DirErr}(y)$: the proportion of items for which the model predicts label y when y is incorrect: $\mathrm{DirErr}(y) = \frac{|i:\hat{y}_i=y \wedge y_i \neq y|}{|i \neq y|}$.

High values indicate that the model systematically drifts toward prediction y even when it is wrong. For YN, we instead measure the rate of YES responses on false comparison statements.

MLE analysis. For MLE, we report the *framing-aligned shift*: the difference between $\mathrm{DirErr}(y)$ under a framing cue y and the corresponding Neutral condition. Results are averaged across the six matched framed conditions for each model.

WHO analysis. WHO has no neutral condition, so we analyze the four settings where the framing cue and the comparison question ask about the same direction (e.g., more cue + who scored more?). We then measure how often the model selects the person implied by the cue when that person is actually incorrect. A model with no framing bias should therefore score near zero. Additional WHO conditions appear in Appendix B.

Statistical testing. We report 95% bootstrap confidence intervals using 1,500 resamples (full significance-testing details in Appendix A).

3.6 Models and Reasoning Modes

We test ten models, five proprietary (Gemini-2.5-Flash, Grok-4.1-Fast, Gemma-4-26B, GPT-OSS-20B, DeepSeek-V3.2) and five open-weight models in the 7 to 9B range that we run locally (Qwen3.5-9B, Llama-3.1-8B, Ministral-8B, CommandR-7B, OLMo-3-7B). All runs use temperature 0. In **short answer** mode, the model returns a single label. In **CoT** mode, it writes its reasoning first and then its final answer, both inside one JSON object.

4 Interactions Between Framing and Positional Bias

Model	Fixed	Balanced	CoT (balanced)
Gemini-2.5-Flash	0.2 [-0.2, 0.8]	0.3 [-0.0, 0.7]	0.3 [-0.3, 1.0]
Grok-4.1-Fast	24.7 [-1.7, 51.1]	22.5 [-2.7, 45.4]	0.8 [-0.1, 1.9]
Gemma-4-26B	3.8 [-2.8, 10.5]	2.2 [-3.5, 7.9]	0.6 [-0.7, 2.4]
GPT-OSS-20B	-0.1 [-0.4, 0.0]	-0.4 [-1.1, 0.3]	0.1 [-1.2, 1.5]
DeepSeek-V3.2	39.9 [18.5, 60.3]	30.3 [10.9, 48.4]	0.1 [-0.7, 0.8]
Qwen3.5-9B	29.5 [6.5, 54.6]	32.5 [8.5, 56.1]	1.8 [1.2, 2.6]
Llama-3.1-8B	43.8 [15.0, 72.4]	40.0 [26.9, 53.8]	1.5 [0.6, 2.7]
Ministral-8B	62.6 [31.3, 93.4]	37.2 [16.8, 55.3]	-0.4 [-7.1, 6.3]
CommandR-7B	27.3 [8.7, 47.2]	11.0 [1.9, 20.9]	2.2 [-4.1, 8.7]
OLMo-3-7B	52.0 [15.1, 82.8]	54.0 [45.0, 64.2]	1.4 [-4.9, 7.1]
Across-model mean	**28.4**	**23.0**	**0.8**

Table 2: **[MLE only, per model, averaged over 6 matched framed cells]** Mean framing-aligned shift (pp) vs. Neutral MLE, with 95% bootstrap CIs (1,500 resamples). Fixed: option A always carries the cue. Balanced: option ordering randomised across six permutations. CoT: chain-of-thought on the balanced condition.

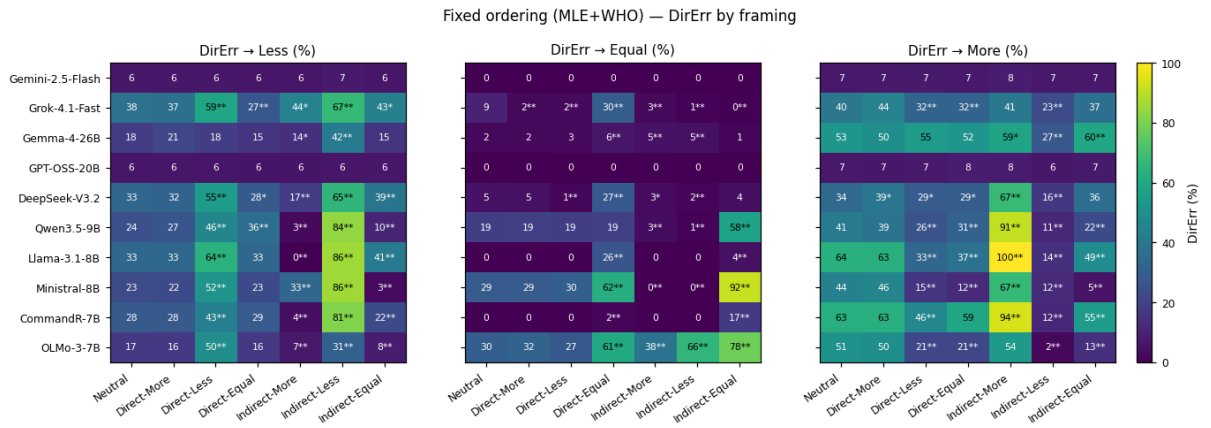

Figure 1: **[MLE only, per model × framing-cell]** DirErr (%) under *fixed* option ordering. The same canonical option order is used across all instances (A. More, B. Less, C. Equal), so slot A is always MORE regardless of the framing cue. Each panel shows one error direction (one row of cells). Significance vs. Neutral MLE (paired bootstrap, $n = 1{,}500$): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Fixed ordering tests whether framing effects interact with positional bias. Because the same answer label always occupies slot A, any tendency to favor early options can systematically reinforce particular framing directions rather than acting as independent noise.

Table 2 reports aggregate framing-aligned shifts averaged across the six matched framing conditions, while Figure 1 shows the full directional error structure across framing cells. Exact per-cell values appear in Table 8 in Appendix.

Framing effects produce highly asymmetric error patterns. Figure 1 shows that framing does not simply increase overall error rates, but systematically redirects predictions toward particular outcomes. More-cued framings strongly increase DIRERR→MORE, while Less- and Equal-cued framings produce corresponding shifts in their aligned directions. Several models approach ceiling-level directional error on individual cells, especially under indirect framing.

The strongest effects concentrate in smaller open-weight models. Gemini-2.5-Flash and GPT-OSS-20B remain within ± 0.3,pp across matched cells. Largest shifts occur in 7-9B open-weight models, but also Deepseek shows pattern.

5 Isolating Framing Effects from Positional Bias

Balanced ordering removes the fixed slot-A advantage by randomizing answer positions across all permutations. This isolates the component of directional drift attributable to comparative framing itself rather than positional bias. As shown in Figure 2 and Table 2 (Balanced column), the mean framing-aligned shift decreases from $+\mathbf{28.4}$ pp under fixed ordering to $+\mathbf{23.0}$ pp under balanced ordering, a relative reduction of 19%.

Comparative framing alone systematically steers predictions. Even after positional bias is removed, most models continue to exhibit substantial directional drift. The largest reductions occur for Ministral-8B, CommandR-7B, and Qwen3.5-9B, indicating that positional bias strongly amplified framing effects in these models. However, several models, including OLMo-3-7B and DeepSeek-V3.2, remain strongly affected under balanced ordering, showing that comparative cues alone can systematically redirect predictions.

Direct framing produces substantially stronger effects. The strength of the effect depends strongly on presentation. Direct framing, where the cue appears inside the question itself, consistently produces larger shifts than indirect contextual cues. Thus, comparative framing exerts the strongest pull when embedded directly in the comparison query rather than introduced through preceding context. Detailed balanced-ordering breakdowns appear in Table 9 in Appendix.

6 Directional Drift Beyond Label Matching

To test whether framing effects depend on lexical overlap between cues and answer labels, we evaluate two alternative formats under balanced ordering: WHO, where models select a person rather

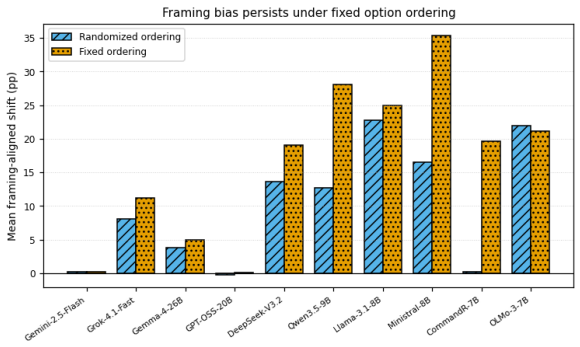

Figure 2: **[MLE only, per model, averaged over 6 matched framed cells]** Mean framing-aligned shift (pp) under fixed vs. balanced (randomised) option ordering. The fixed dataset uses one canonical permutation (A. More, B. Less, C. Equal) on every item. The balanced dataset spreads the correct answer across all six permutations. Balanced ordering reduces the across-model mean from $+28.4$ pp to $+23.0$ pp (-19%), isolating the portion of the shift attributable purely to the cue word.

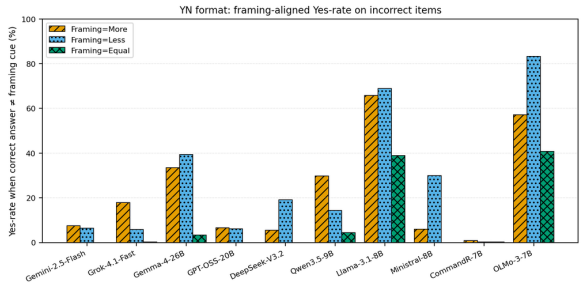

Figure 3: **[YN only, balanced ordering, per model × cue direction]** Yes-rate on incorrect items (items where the cued direction is not the gold label). Pooled across the six framed conditions for each cue direction. A model with no framing bias would show the same rate across all three cue directions. Deviation indicates cue-following.

than a comparison label, and YN, where models verify a framed comparison statement. Table 3 reports matched WHO directional error rates, while Figure 3 shows YES-rates on incorrect YN items.

Model	D-Mo	D-Le	I-Mo	I-Le	Mean
Gemini-2.5-Flash	7	7	8	7	**7**
Grok-4.1-Fast	63	68	34	81	**62**
Gemma-4-26B	40	13	36	44	**33**
GPT-OSS-20B	7	7	7	6	**7**
DeepSeek-V3.2	61	67	70	81	**70**
Qwen3.5-9B	43	45	47	64	**50**
Llama-3.1-8B	69	65	89	94	**79**
Ministral-8B	79	68	63	58	**67**
CommandR-7B	33	34	23	24	**29**
OLMo-3-7B	37	29	46	18	**33**
Across-model mean					**43.5**

Table 3: **[WHO matched non-equal, per model × framing-cell, 4 cells per model]** DIRERR (%) under Balanced-ordering prompting. Each cell measures: among items where the cued person is not actually the right answer, how often the model picks the cued person. With no framing bias the rate should approach zero. Mismatched-direction and equal WHO variants appear in Appendix B. Darker = higher DirErr.

WHO: directional drift persists without cue labels. In the matched non-equal WHO conditions (§3.5), models must choose between PERSON 1 and PERSON 2, while the comparative cue appears only in the input. A model with no framing bias should therefore rarely select the cue-implied person when that person is incorrect. However, Table 3 shows strong residual directional drift under balanced ordering, with an across-model mean of **43.5**%. Llama-3.1-8B reaches 94% on Indirect-Less and 89% on Indirect-More, while DeepSeek-V3.2 reaches 81% on Indirect-Less. Only Gemini-2.5-Flash (7.0%) and GPT-OSS-20B (6.7%) remain near zero.

These results show that the framing effect is not limited to copying comparison labels from the answer space. Even when the options contain only entity names, many models still map comparative cues onto the implied referent and systematically select that referent when it is incorrect. This suggests that comparative framing alters the semantic interpretation of the comparison itself, not just surface-level answer selection.

YN: framing skews verification behavior. Figure 3 shows the YES-rate on incorrect items under balanced ordering. A model with no framing bias would produce similar rates across all cue directions. Instead, several smaller open-weight models exhibit strong directional asymmetries. OLMo-3-7B answers YES on more than 80% of incorrect *less*-cued items, while Llama-3.1-8B exceeds 65% for both *more* and *less*. Across models, *equal*-cued framing consistently produces weaker effects than directional comparative cues.

Unlike MLE, the YN format contains no competing comparison labels and requires only binary verification. Yet framing effects persist strongly, suggesting that directional cues can bias the model's internal assessment of the comparison itself rather than merely shifting preference between answer options.

7 Chain-of-Thought as Partial Mitigation

We evaluate chain-of-thought (CoT) prompting under balanced ordering, isolating whether explicit reasoning reduces framing-induced directional drift independently of positional bias.

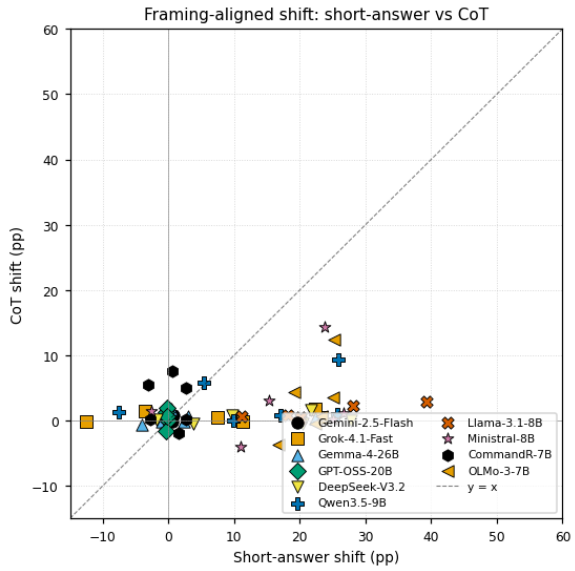

Figure 4: **[MLE only, one point per (model, framed cell)]** Framing-aligned shift in balanced short-answer (x-axis) vs. CoT (y-axis), 60 points total (10 models × 6 framed cells). Almost every point lies below the diagonal, which means CoT shrinks the MLE shift across models and cells.

CoT substantially reduces directional drift. Table 2 and Figure 4 show that CoT sharply decreases framing-aligned MLE shifts across models and framing conditions. The across-model mean falls to $+\mathbf{0.8}$ pp, with six of ten models below 1 pp. In Figure 4, almost every point lies below the diagonal, indicating consistent reduction across framing cells. Explicit reasoning appears to anchor predictions to computed quantities before answer selection.

Mitigation is weaker when semantic interpretation is required. The reduction is substantially smaller in WHO, where models must map comparative cues onto entity references rather than select comparison labels directly. Table 4 shows that matched WHO directional error decreases from 43.5% to **15.3**% under CoT, but several smaller models retain substantial drift, including CommandR-7B (38%), Ministral-8B (26%), and OLMo-3-7B (21%).

Residual framing effects persist after explicit reasoning. Although CoT removes most large framing shifts, some reliable directional errors remain, particularly under indirect framing in smaller open-weight models. In several cases, the rationale computes the correct comparison while the final answer still follows the framing cue. This suggests that reasoning elicitation mitigates directional framing effects substantially, but does not fully eliminate semantic steering during answer generation. Detailed CoT heatmaps and accuracy results appear in Figure 11 in Appendix.

Model	D-Mo	D-Le	I-Mo	I-Le	Mean
Gemini-2.5-Flash	7	6	7	5	**6**
Grok-4.1-Fast	6	6	8	5	**6**
Gemma-4-26B	5	5	6	6	**6**
GPT-OSS-20B	7	6	8	9	**7**
DeepSeek-V3.2	3	7	6	8	**6**
Qwen3.5-9B	13	17	16	23	**17**
Llama-3.1-8B	17	16	20	23	**19**
Ministral-8B	14	23	21	45	**26**
CommandR-7B	18	53	27	54	**38**
OLMo-3-7B	16	19	21	29	**21**
Across-model mean					**15.3**

Table 4: **[WHO matched non-equal, per model × framing-cell]** DIRERR (%) under Chain-of-thought prompting. Each cell shows how often the model selects the cue-implied person when that person is incorrect. Lower is better. Mismatched-direction and equal WHO variants appear in Appendix B. Darker = higher DIRERR.

8 Robustness Across Domains and Problem Scales

The previous sections showed that comparative framing systematically steers predictions, that positional bias amplifies the effect, and that explicit reasoning substantially mitigates but does not fully eliminate it. We now test whether these effects depend on domain, arithmetic scale, or pre-existing directional priors.

Directional framing generalizes across domains. Table 5 shows consistent framing-aligned shifts across all four topic families. Although absolute magnitudes vary somewhat by domain, the relative ordering of framing conditions remains stable, indicating that the effect is not tied to a specific semantic domain or scenario template.

The effect is largely independent of arithmetic complexity. As shown in Table 6, increasing problem length or digit count produces only modest changes in framing-aligned shift. Directional drift therefore appears to arise primarily from compara-

Topic	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Education	16.4	39.8	39.6	9.8	30.8	1.2
Housekeeping	16.0	40.4	32.5	10.6	23.8	8.0
Sports	20.1	41.1	33.1	15.5	32.1	-2.6
Transportation	20.1	45.8	33.1	15.2	34.9	2.7

Table 5: **[MLE only, per topic × framed cell, averaged across 10 models]** Framing-aligned shift (pp) under Balanced-ordering prompting.

tive framing itself rather than from failures caused by longer reasoning chains or harder arithmetic.

Subset	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Length = 3.0	18.9	42.6	35.3	14.4	32.8	-1.4
Length = 4.0	19.5	43.4	32.8	11.7	29.0	2.8
Length = 5.0	16.9	40.1	35.6	10.8	29.7	3.4
Length = 6.0	17.4	41.2	34.6	14.2	29.8	4.6
Digits = 1.0	20.8	43.4	31.3	17.1	31.8	-1.6
Digits = 2.0	18.5	42.7	35.4	11.4	29.7	3.1
Digits = 3.0	15.3	39.4	37.1	9.7	29.4	5.6

Table 6: **[MLE only, per problem-scale slice × framed cell, averaged across 10 models]** Framing-aligned shift (pp) by problem length and digit count under Balanced-ordering prompting.

Several models exhibit intrinsic directional priors. Figure 5 shows that some models already prefer particular comparison outcomes under the Neutral prompt, especially *More*. Importantly, all framing-aligned shifts in this paper are measured relative to the corresponding Neutral baseline, ensuring that the reported effects capture framing-induced directional drift rather than pre-existing label preferences.

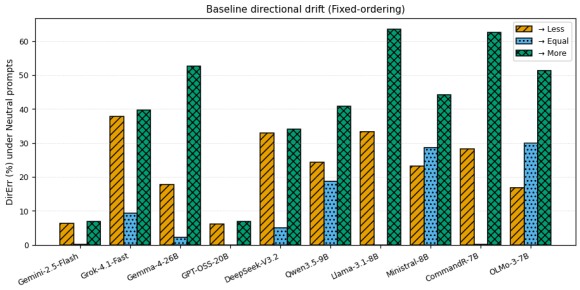

Figure 5: **[MLE only, balanced ordering, per model × error direction]** Baseline Neutral MLE DirErr. Several models have an intrinsic prior toward *More* or *Less* independent of any framing cue.

Positional bias changes error direction more than overall competence. Overall accuracy remains broadly similar under fixed and balanced ordering (Figure 10 in Appendix), even though directional shifts differ substantially between the two settings. This suggests that positional bias primarily alters the direction of model errors rather than overall task competence.

9 Discussion

Our results show that framing effects in arithmetic reasoning are not simply a consequence of prompt instability or shallow label copying. The same comparative cue produces different forms of directional error depending on how the task is represented: in MLE, the bias appears as cue-aligned label selection; in WHO, it appears as selection of the cue-implied referent; and in YN, as asymmetric verification behavior. Across formats, the underlying directional pull remains structurally consistent even as its surface realization changes.

Chain-of-thought reasoning changes this picture again. Explicit reasoning sharply reduces directional drift in direct label-selection settings, yet residual effects persist in formats requiring semantic reference resolution. This suggests that eliciting correct intermediate reasoning is not always sufficient to prevent framing-sensitive answer generation, particularly when the final prediction still requires mapping comparative language onto entities or verification decisions.

10 Conclusion

MATHCOMP shows that comparative framing in LLM reasoning can produce structured, representation-dependent error patterns rather than simple prompt instability. Across answer formats, option-ordering regimes, and reasoning settings, semantically equivalent cues systematically reshape how models interpret and express numerical comparisons. The resulting behavior cannot be reduced to a single mechanism such as positional bias or label copying; instead, the form of the error changes with the representational structure of the task itself. These findings point toward a broader perspective on reasoning robustness. Evaluating whether a model arrives at the correct answer is often insufficient to understand how language influences the reasoning process that produces it. By isolating framing-induced directional drift in a controlled arithmetic setting, MATHCOMP provides a foundation for studying the semantic structure of reasoning errors more systematically, and for developing robustness evaluations that move beyond aggregate accuracy toward more behaviorally grounded analyses of model reasoning.

Limitations

The scenarios are English only and follow a fixed shape: two people, a few sub-tasks, a clean sum. Real prompts are messier (longer text, unit conversions, narrative wrapped around numbers).

All models ran at temperature 0. Higher temperatures would spread the prediction distribution, but they are unlikely to flip the direction of the pull.

The matched WHO metric is an absolute rate, not a shift over a neutral baseline. WHO is defined only for framed conditions, so no baseline exists. The reading is a bit different from the MLE shift, but both numbers track the same pull toward the cued label. Appendix B has the full WHO grid with mismatched and equal-cued cells.

References

Claudia Collacciani, Giulia Rambelli, and Marianna Bolognesi. 2024. Quantifying generalizations: Exploring the divide between human and LLMs' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11811–11822, Bangkok, Thailand. Association for Computational Linguistics.

James Druckman. 2001. Evaluating framing effects. *Journal of Economic Psychology*, 22:91–101.

Stephen Flusberg and Kevin J. Holmes. 2024. Linguistic framing in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

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.

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.

Akshat Gupta. 2023. Probing quantifier comprehension in large language models: Another example of inverse scaling. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 56–64, Singapore. Association for Computational Linguistics.

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.

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.

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.

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.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity innbsp;large language models. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, page 303–313, Berlin, Heidelberg. Springer-Verlag.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

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*.

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.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

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.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Dataset Generation and Prompts

A.1 Number Generation

For each (topic, length L, digit count, variant, label) combination a Python script draws random integer values and checks two rules. First, the gold label must hold on the totals. Second, the per-sub-task comparisons must include at least one slot where Person 1 is higher, one where they are equal, and one where Person 1 is lower. We call this the cross-comparison rule. It stops any single sub-task from giving the answer away. The model has to sum the values to get the right label.

Arithmetic variants. The three variants differ in how sub-task values combine into totals.

V1 (all additive). Each sub-task adds a positive integer: $\text{Total}_i = \sum_{k=1}^{L} v_{i,k}$. Straight addition.

V2 (one deduction). The last sub-task is subtracted: $\text{Total}_i = \sum_{k=1}^{L-1} v_{i,k} - v_{i,L}$. This adds a minus sign the model has to track. In the scenario text it shows up as a penalty or deduction ("received a 4-point deduction in the discus throw").

V3 (multiply, add, divide). The first sub-task is a product ($q \times r$, where $q \in [2, 9]$ is a count and r is a per-unit value). The middle sub-tasks are additive. The last sub-task is a negative integer division ($-d \div s$, where $s \in [2, 9]$ divides d exactly). This mixes three operations. In scenario text it reads like "scored 3 points per heat across 5 heats" and "had a group penalty of 20 points shared equally among 4 athletes".

Equal-label construction. For the equal label the script uses a targeted method. It fixes a solution for Person 1, copies one raw value across to satisfy the cross-comparison rule, then solves backwards for the last value so the totals match. The more and less labels use random retries with rejection sampling.

Pipeline. The 144 base scenario frames are the full cross of topic $\times$ length $\times$ digit count $\times$ variant. We wrote one scenario template per topic by hand and let the script combine the axes. For each frame the number generator (`numgen.py`) samples raw integers under seed 42, checks the label and cross-comparison rules, and retries up to 2,000 times before reporting failure. No frame failed. Question text is built by `questions.py` from per-format templates. Scenario sentences are built by `templates.py`. The whole pipeline runs in under a minute on a laptop CPU and is deterministic given the seed.

A.2 Instruction Templates

A.3 Bootstrap Statistical Test

We test each framed-vs-Neutral difference with a pooled bootstrap under the null. We pool the framed and neutral samples (size $n_{\text{framed}} + n_{\text{neutral}}$) and resample with replacement to build the null distribution of the rate difference, then take a two-sided tail. The tables and the headline numbers in the abstract use 1,500 resamples. p-values are floored at $1/n_{\text{boot}}$. The heatmap markers are $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Format	Instruction (built by `evaluate.py`)
Common prefix	"You are solving a comparison problem. Read the scenario and question carefully."
MLE (short)	Prefix + "Respond with ONLY the exact text of one option: the word More, Less, or Equal (match spelling/capitalisation as shown in Options)."
WHO (short)	Prefix + "Respond with ONLY the exact text of one option: either Person 1, Person 2, or Equal (exact wording as in Options; option order varies)."
YN (short)	Prefix + "Respond with ONLY the exact text of one option: Yes or No."
CoT (all formats)	Prefix + "Output a single JSON object only – no markdown fences, no text before or after. Required keys: `reasoning` (step-by-step chain-of-thought) and `answer` (final choice). Evaluation uses only `answer`. The `answer` string must be exactly ⟨one of the format options above⟩."

Table 7: **[All formats, instruction strings appended to the prompt]** Instruction templates as actually sent to the model. Built by `build_system_prompt` in `evaluate.py`.

B Additional Figures and Tables

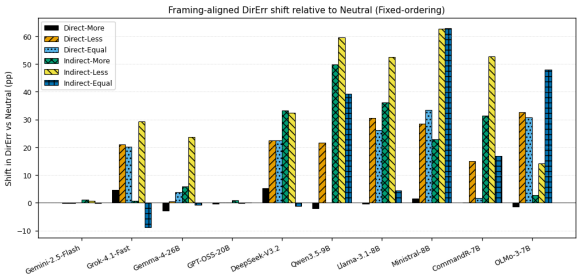

Figure 6: **[MLE only, per model, averaged over 6 matched framed cells]** Framing-aligned shift (pp) vs. Neutral MLE under fixed option ordering. The same canonical option order (A. More, B. Less, C. Equal) is used across the whole dataset, so any slot-A primacy bias adds a constant pull toward predicting MORE.

B.1 Full WHO grid

The main paper uses only the four matched non-equal WHO cells. Table 10 reports the full framed grid under balanced ordering. (i) The *mismatched-direction* cells, for example Direct-More with "who scores less?". A model that follows the cue picks the cued person regardless of what the question asks, so the mismatched cells show the opposite pattern from the matched cells. (ii) The *equal-cued* cells (Direct-Equal and Indirect-Equal) have no matched WHO version, since the cue word is not one of the WHO options.

Model	Dir	N	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Gemini	Le	6	6	6	6	6	7	6
	Eq	0	0	0	0	1	0	0
	Mo	7	7	7	7	8	7	7
Grok	Le	41	34	98	5	80	89	54
	Eq	24	2	1	86	4	1	1
	Mo	20	39	0	1	4	2	20
Gemma	Le	14	24	15	5	10	33	14
	Eq	3	2	3	14	9	8	3
	Mo	52	42	59	50	55	28	56
GPT-OSS	Le	6	6	6	4	6	6	5
	Eq	0	0	0	0	0	0	0
	Mo	7	6	8	10	7	6	7
DeepSeek	Le	16	11	85	4	7	69	36
	Eq	14	13	3	81	9	5	11
	Mo	19	32	1	1	59	4	20
Qwen	Le	32	40	97	67	6	99	28
	Eq	0	0	0	0	0	1	15
	Mo	47	40	1	18	84	0	41
Llama	Le	0	0	92	0	0	80	2
	Eq	0	0	0	79	0	0	13
	Mo	100	99	8	21	100	14	84
Ministral	Le	2	0	87	0	0	99	1
	Eq	0	0	4	100	0	0	85
	Mo	95	100	9	0	100	0	14
CmdR	Le	0	0	45	2	0	62	0
	Eq	0	0	0	6	0	0	51
	Mo	100	100	50	89	100	26	51
OLMo	Le	0	0	99	0	0	50	0
	Eq	7	14	1	100	5	43	81
	Mo	92	86	0	0	94	2	17
Avg	*Le*	**12**	**12**	**63**	**9**	**12**	**59**	**15**
	Eq	**5**	**3**	**1**	**47**	**3**	**6**	**26**
	Mo	**54**	**55**	**14**	**20**	**61**	**9**	**32**

Table 8: **[MLE only, per model × error direction × framing-cell]** DIRERR (%) under Fixed-ordering prompting. N=Neutral; D/I=Direct/Indirect; Mo/Le/Eq=More/Less/Equal. Darker = higher DirErr.

B.2 Additional MLE heatmaps

Model	Dir	N	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Gemini	Le	6	7	6	6	6	7	6
	Eq	0	0	0	0	0	0	0
	Mo	7	7	7	7	7	7	7
Grok	Le	29	26	96	14	53	70	37
	Eq	29	4	4	63	10	8	2
	Mo	24	48	0	9	21	13	42
Gemma	Le	11	22	13	5	10	25	13
	Eq	4	2	3	12	9	9	3
	Mo	54	44	61	50	54	36	57
GPT-OSS	Le	5	5	6	3	5	6	6
	Eq	0	0	0	0	0	0	0
	Mo	9	7	7	14	7	7	7
DeepSeek	Le	16	19	78	27	10	70	29
	Eq	17	8	5	46	7	9	11
	Mo	33	41	3	5	65	3	28
Qwen	Le	12	9	89	5	34	58	24
	Eq	29	3	0	81	1	3	15
	Mo	46	75	4	8	51	30	47
Llama	Le	25	3	85	33	15	75	33
	Eq	0	0	0	59	0	0	28
	Mo	66	96	14	4	78	21	36
Ministral	Le	52	26	100	13	38	91	60
	Eq	10	2	0	86	1	1	3
	Mo	20	57	0	0	54	4	23
CmdR	Le	32	11	55	21	27	29	33
	Eq	24	21	20	35	31	34	31
	Mo	43	72	26	43	41	37	36
OLMo	Le	17	1	90	3	1	78	3
	Eq	33	5	2	97	10	14	74
	Mo	48	93	7	0	89	7	20
Avg	*Le*	**20**	**13**	**62**	**13**	**20**	**51**	**24**
	Eq	**15**	**4**	**3**	**48**	**7**	**8**	**17**
	Mo	**35**	**54**	**13**	**14**	**47**	**16**	**30**

Table 9: **[MLE only, per model × error direction × framing-cell]** DIRERR (%) under Balanced-ordering prompting. N=Neutral; D/I=Direct/Indirect; Mo/Le/Eq=More/Less/Equal. Darker = higher DirErr.

Model	D-M/M	D-M/L	D-L/M	D-L/L	D-E/M	D-E/L	I-M/M	I-M/L	I-L/M	I-L/L	I-E/M	I-E/L
Gemini	7	7	7	7	7	6	8	6	7	7	7	6
Grok	63	71	60	68	60	68	34	49	60	81	58	64
Gemma	40	13	44	13	42	13	36	6	20	44	40	7
GPT-OSS	7	7	7	7	7	7	7	7	6	6	8	6
DeepSeek	61	69	60	67	57	65	70	52	41	81	56	65
Qwen	43	45	43	45	43	45	47	45	40	64	45	49
Llama	69	65	72	65	68	62	89	24	39	94	48	45
Ministral	79	69	81	68	78	70	63	11	35	58	56	42
CmdR	33	34	33	34	33	34	23	23	26	24	31	27
OLMo	37	29	37	29	37	29	46	10	5	18	12	10

Table 10: **[WHO full grid, per model × (framing-cell, question-direction)]** DIRERR (%) under Balanced-ordering prompting. Column headers use the format **X-F/Q**, where **X** ∈ {D=Direct, I=Indirect} is the framing type, **F** ∈ {M=More, L=Less, E=Equal} is the framing cue, and **Q** ∈ {M=More, L=Less} is the question direction. Matched cells (e.g., D-M/M, I-L/L) are the main-paper subset; mismatched and equal-cued cells are reported here for completeness.

Figure 7: **[MLE only, per model × framing-cell]** DirErr (%) under *balanced* option ordering. Compare with Figure 1 for the fixed-ordering condition.

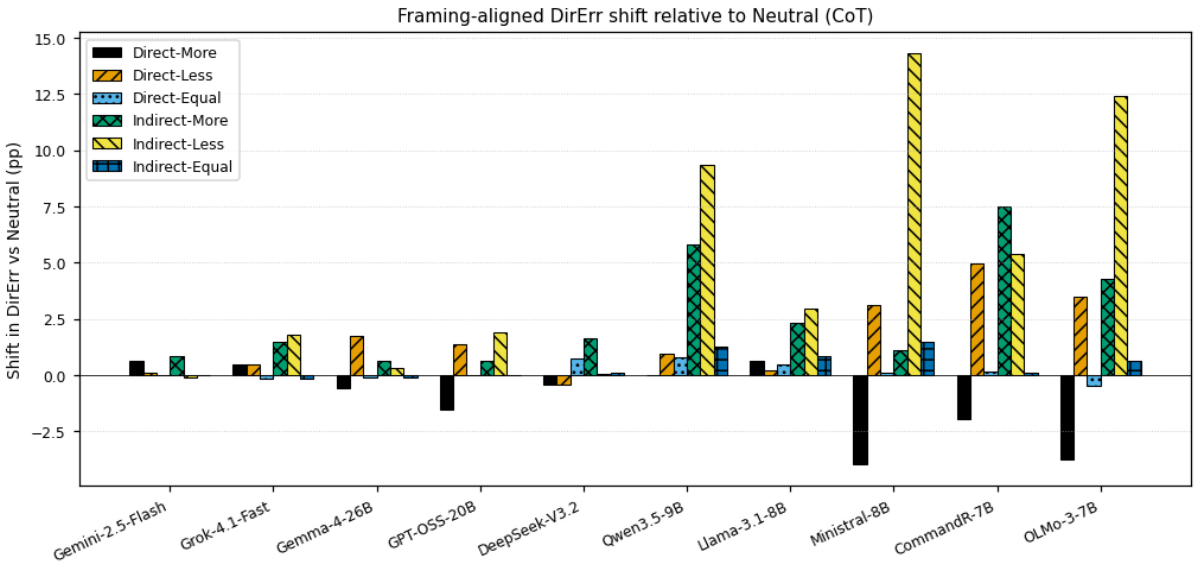

Figure 8: **[MLE only, per model, averaged over 6 matched framed cells]** Framing-aligned shift vs. Neutral MLE under CoT prompting on the balanced condition.

Model	Dir	N	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Gemini	Le	6	5	5	5	6	6	6
	Eq	0	0	0	0	0	0	0
	Mo	6	8	7	8	7	6	7
Grok	Le	3	5	4	4	5	6	6
	Eq	0	0	0	0	1	0	0
	Mo	6	6	6	6	8	5	6
Gemma	Le	6	5	11	5	5	6	6
	Eq	0	1	0	0	0	0	0
	Mo	7	5	6	4	7	6	6
GPT-OSS	Le	5	5	7	5	4	6	6
	Eq	0	0	0	0	0	0	0
	Mo	9	6	8	14	9	6	8
DeepSeek	Le	7	5	7	5	6	7	5
	Eq	0	0	0	1	0	0	1
	Mo	7	5	6	5	7	5	4
Qwen	Le	11	13	13	15	12	13	13
	Eq	0	0	0	2	0	1	1
	Mo	12	13	14	17	16	12	12
Llama	Le	27	22	27	36	24	31	27
	Eq	0	0	1	0	1	2	1
	Mo	12	14	22	16	13	11	14
Ministral	Le	7	8	15	14	11	17	15
	Eq	2	2	1	3	1	0	4
	Mo	34	23	15	28	22	12	17
CmdR	Le	12	15	27	16	16	22	16
	Eq	0	0	1	1	0	0	1
	Mo	26	20	13	17	20	9	18
OLMo	Le	9	11	19	22	11	20	12
	Eq	1	1	0	0	1	1	2
	Mo	23	10	13	14	22	13	18
Avg	*Le*	**9**	**9**	**14**	**13**	**10**	**13**	**11**
	Eq	**0**	**0**	**0**	**1**	**0**	**1**	**1**
	Mo	**14**	**11**	**11**	**13**	**13**	**8**	**11**

Table 11: **[MLE only, per model × error direction × framing-cell]** DIRERR (%) under Chain-of-thought prompting. N=Neutral; D/I=Direct/Indirect; Mo/Le/Eq=More/Less/Equal. Darker = higher DirErr.

Topic	D-Mo	D-Le	D-Eq	I-Mo	I-Le	I-Eq
Education	-2.1	6.5	1.0	-0.9	4.2	0.8
Housekeeping	-4.4	2.0	-0.3	-2.2	2.6	0.3
Sports	-1.7	4.9	0.6	1.2	2.7	1.2
Transportation	-4.7	3.9	0.2	-2.9	7.6	0.5

Table 12: **[MLE only, per topic × framed cell, averaged across 10 models]** Framing-aligned shift (pp) under Chain-of-thought prompting.

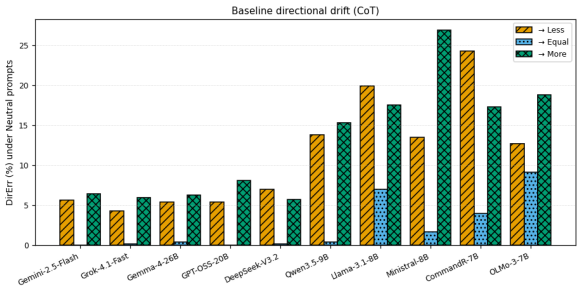

Figure 9: **[MLE only, balanced ordering, per model × error direction]** Baseline Neutral MLE DirErr under CoT prompting.

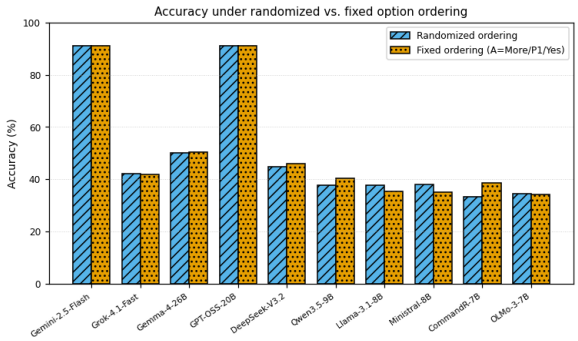

Figure 10: **[MLE+WHO pooled, per model, averaged across all framings]** Accuracy (%) under randomised vs. fixed option ordering, balanced short-answer.

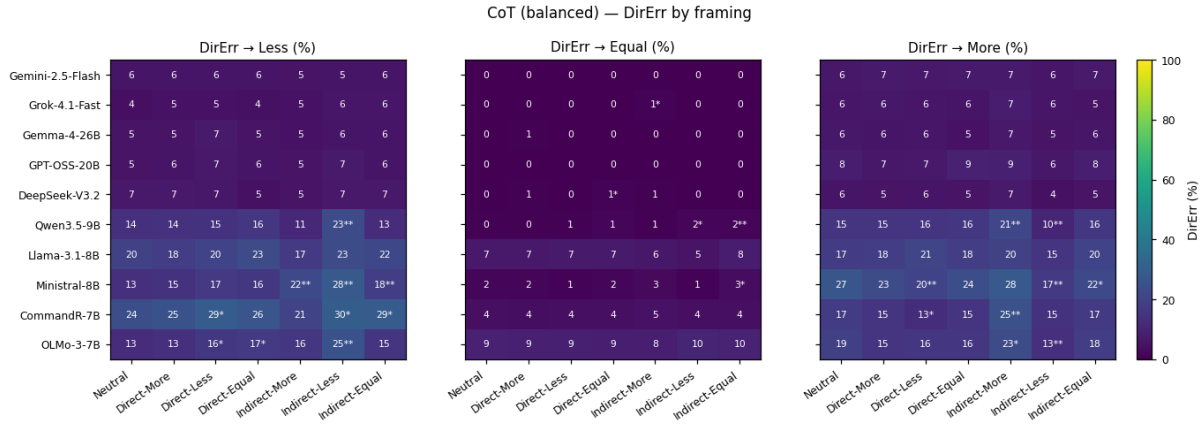

Figure 11: **[MLE only, per model × framing-cell]** DirErr (%) under CoT prompting on the *balanced* ordering condition. Significance markers as in Figure 1.

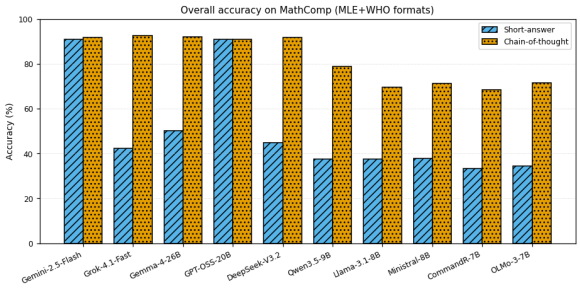

Figure 12: **[MLE+WHO pooled, per model, averaged across all framings and orderings]** Overall accuracy in short-answer vs. CoT mode. The models most biased under short answer gain the most accuracy from CoT.