Prompt-order dominance and question-weight sensitivity in published language-model evaluations

Lluis Eriksson

19 November 2026

Abstract

We recover an ordering across individual runs from published summaries of language-model accuracy. On all eight mathematics tasks in Gao et al.'s PAL experiment, every reported program-aided prompt order outperforms every reported chain-of-thought order on the same task, allowing for rounding. All comparisons remain positive if question values range from 1 to 1.06. Yet on ASDiv we construct run counts and question outcomes consistent with the published summaries that permit a tie at a weight ratio of 29/27. Integer reconstruction also reveals four tasks for which three unrounded full-file accuracies cannot produce both reported summaries under either common standard-deviation convention. These are retrospective bounds on the information in an evaluation table. They neither predict new prompt orders nor establish which questions the models actually missed.

1 Which comparisons does a benchmark table support?

A mean accuracy gain leaves open whether the weakest run of one method exceeds the strongest run of the other. Even if that ordering holds, a reader who values some questions more than others may reach a different decision. We examine both sources of uncertainty in Gao et al.'s comparison of Codex with program-aided prompting (PAL) and chain-of-thought prompting (CoT) [1]. PAL generates code for an interpreter. Table 9 of the original paper reports means and standard deviations over three prompt orderings per method on eight tasks. We consider all nine cross-order comparisons on each task. Knowledge of which runs shared an ordering is unnecessary.

Ficsor and Abonyi [3] motivate a related release-design question. Their publisher-deposited abstract reports 16 and 33 correct answers on the same 36 questions when Gemini 2.5 Flash used a classifier description or received computed evidence. We study what these two totals identify. Their 50-classifier replication is a different experiment. We could not verify its per-classifier margins.

2 Dominance across the three reported orders

For three observations with mean μ and population standard deviation σ, each observation lies within $\sqrt{2}\sigma$ of μ. If one deviation is t, the other two sum to $-t$, so their squared deviations sum to at least $t^2/2$. The bound also holds with a sample deviation, whose denominator is smaller. This is a special case of Samuelson's inequality [4].

Let m_P, m_C be the two published means in percentage points and s_P, s_C the published deviations. The means are displayed to one decimal and the deviations to two. With nearest rounding, the PAL score minus the CoT score for any pair of reported orders is at least

$$(m_P - m_C) - 0.10 - \sqrt{2}\,(s_P + s_C + 0.01).$$

If each order's underlying percentage was rounded to one decimal before summarizing, subtract a further 0.10 point. The bound concerns the archived orders, with no sampling assumption about future prompts.

Task	File rows	Guaranteed gap (pp)	Raw feasible	Prior rounding feasible
GSM8K	1319	4.503	P,S	P
SVAMP	1000	3.934	–	–
ASDiv	2096	1.468	P,S	P
GSM-Hard	1319	35.805	P,S	P,S
SingleEq	508	5.697	–	–
SingleOp	562	1.298	–	S
AddSub	395	5.028	–	–
MultiArith	600	1.785	S	–

The third column rounds the lower bound downward to 0.001 point for unrounded order scores. The other two result columns require integer counts. P and S mean that both methods admit count triples at the file denominator under the same population or sample deviation convention. A dash denotes absence of a compatible convention within that column.

3 Which integer counts are feasible?

We count nonempty JSON records in each task file at a fixed repository revision [2]. Gao et al.'s Table 12 gives the same sizes for the seven ordinary tasks. For GSM-Hard we use its own file. For sorted correct counts (i, j, k) out of n, set $T = i + j + k$, $Q = i^2 + j^2 + k^2$. The mean percentage is $100T/(3n)$ and the variance is $10^4(3Q - T^2)/(dn^2)$ with $d = 9$ for population or $d = 6$ for sample variance. Compare the mean to its closed rounding interval and the variance to the squared endpoints of the deviation interval. Closed endpoints admit both rounding ties, making exclusions conservative.

Two implementations enumerated the triples and agreed on all raw feasible sets for both methods on each task. A second protocol rounds each order percentage to one decimal before computing the summary. This reconciles SingleOp with a sample deviation. SVAMP, SingleEq and AddSub remain incompatible under both protocols. For SVAMP the unrounded CoT column admits two triples only under population deviation; the PAL column admits one only under sample deviation. The audit assumes rates of binary success at a common full-file denominator. Filtering or other precision choices require separate evidence. Empty count sets leave the real-valued comparison in Section 2 intact. We use Table 9 consistently; the CoT mean for GSM-Hard reads 23.3 there and 23.1 in Table 1.

4 What survives a change in question values?

On the same n questions let CoT and PAL have A and B correct answers respectively, with $B > A$. A correct answer to a given question has the same positive value in either condition. Suppose values lie in $[1, R]$. If h questions change from correct to incorrect, then $B - A + h$ change from incorrect to correct. The smallest weighted difference over joint assignments and admissible values is

$$B - A - (R - 1)\min(A, n - B).$$

Indeed, $h \le \min(A, n - B)$, with equality attainable by minimizing overlap between correct sets. Give losses weight R and gains weight 1. The construction is a finite marginal-data argument; sharp joint-outcome bounds from margins are classical [5]. For a fixed multiset of values, subtracting the sum of the A largest from the sum of the B smallest gives the sharp lower endpoint. Both choices can be realized simultaneously. In disjoint blocks their minima therefore add. The worst exclusion of one block subtracts the greatest block minimum.

For the PAL comparison, in percentage units let

$$A_u = m_C + 0.10 + \sqrt{2}(s_C + 0.005), \quad B_l = m_P - 0.10 - \sqrt{2}(s_P + 0.005).$$

These limits allow rounding each order before summarization. A sufficient condition for strict positivity is

$$R < 1 + (B_l - A_u)/\min(A_u, 100 - B_l).$$

The smallest right-hand side across the eight tasks exceeds 1.0660, on ASDiv. Therefore weights in $[1, 1.06]$ preserve every cross-order comparison, conditional on each method being scored on the same questions.

On ASDiv the CoT counts $(1602, 1604, 1632)$ and PAL counts $(1664, 1668, 1671)$ reproduce the published summary at $n = 2096$ under population deviation on raw percentages. For the largest CoT and smallest PAL counts, assign 464 gains, 432 losses, 1200 correct answers in both conditions, and none incorrect in both. These four cells obey both margins. Giving losses 29/27 times the value of gains yields a tie. A ratio of 1.10 gives a weighted difference of -11.2 correct-answer value units. Minimizing over compatible ASDiv triples under either precision protocol and either common deviation convention gives the sharp threshold 29/27. This witnesses a limit on inference from the table. The historical joint assignment remains unknown.

For the comparison of 16 and 33 correct answers [3], the sharp minimum is $20 - 3R$. Consider two hypothetical blocks of 18 questions with unit values. Block margins $(16, 15)$ and $(0, 18)$ give a worst retained advantage of -1 after a block exclusion. Margins $(8, 16)$ and $(8, 17)$ instead give $+8$ with the same totals. Thus the totals cannot identify even the sign after such an exclusion. Totals for the two conditions within each block would suffice if question values were constant within blocks.

5 Scope and reproducibility

The certified orderings concern archived runs of a single historical model. The weighted obstruction describes the information a table has discarded. Neither finding estimates deployment accuracy, the prevalence of costly errors, or the effect of supplying a tool while holding the prompt fixed. The contribution claimed here is the certified comparison, its explicit limit on ASDiv and the diagnostic count audit. Publishing unrounded per-run counts, scoring denominators and per-question outcomes would allow those conclusions to be refined.

Model assessments and review context. OpenAI Codex participated in the analysis, software and drafting in the author's context. The two count implementations shared primary data and do not constitute independent peer review. Any assessment of the exact PDF belongs to its version-specific record.

Reproduction. The companion materials record the source PDF digest, eight file hashes, integer triples, and scripts for both precision protocols. The checks used dedicated Linux runtimes. They required neither new inference calls nor execution of benchmark-generated code.

References

[1] L. Gao et al. PAL: Program-aided Language Models. ICML, PMLR 202 (2023), 10764–10799. `https://proceedings.mlr.press/v202/gao23f.html`.

[2] Reasoning Machines. PAL repository, revision `f81ca2a9777f002f`. `https://github.com/reasoning-machines/pal`.

[3] A. Ficsor and J. Abonyi. From Data to Verified Model Knowledge. *Information* 17 (2026), 1179. `https://doi.org/10.3390/info17111179`.

[4] P. A. Samuelson. How deviant can you be? *Journal of the American Statistical Association* 63 (1968), 1522–1525.

[5] J. Tian and J. Pearl. Probabilities of causation: bounds and identification. *Annals of Mathematics and Artificial Intelligence* 28 (2000), 287–313.