PROMPT CONDITIONS CAN CHANGE MEASURED INTERVENTION EFFECTS IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompted language model evaluations often report the effect of an intervention under a single prompt condition. We test whether that reported effect changes when the surrounding prompt condition changes. On MMLU-Pro, a silent self-check increases exact-format accuracy for Gemma by $29.33\,\mathrm{pp}$ under a standard prompt condition, $14.00\,\mathrm{pp}$ under a masked condition, and $20.00\,\mathrm{pp}$ under a token-matched metadata condition. A separate, pre-specified 300-task replication gives gains of $30.33\,\mathrm{pp}$, $10.33\,\mathrm{pp}$, and $22.67\,\mathrm{pp}$. The self-check gain is therefore $20.00\,\mathrm{pp}$ larger under Standard than under Masked, with a simultaneous 95% confidence interval of $[11.82, 28.18]\,\mathrm{pp}$. DeepSeek shows little observed difference across the same prompt conditions and Llama is inconclusive. The result also depends on scoring. On the replication set, an automated semantic answer scorer gives a Standard-minus-Masked difference of $15.67\,\mathrm{pp}$, while a separate constrained-decoding evaluation gives $-1.00\,\mathrm{pp}$. Supporting experiments further show that average accuracy can remain almost unchanged even when many individual outcomes flip. These findings motivate reporting intervention gains under the prompt and scoring conditions that define them, rather than treating those choices as incidental details. Code: `https://anonymous.4open.science/r/PCCMIE_ICLR27-BCA1/README.md`

1 INTRODUCTION

Prompt design affects language model evaluation. Small changes in formatting, wording, demonstrations, or answer constraints can change benchmark accuracy even when the task itself is intended to remain the same (Sclar et al., 2024). This raises a related question for studies that evaluate an intervention such as self-checking, reflection, reasoning instructions, or debiasing. If the intervention is measured under one prompt condition, would the reported gain be similar under another plausible prompt condition?

We study this question with a silent self-check intervention. In a pre-specified Gemma replication on MMLU-Pro, exact-format accuracy improves by $30.33\,\mathrm{pp}$ under the Standard prompt condition and by $10.33\,\mathrm{pp}$ under a Masked condition. The difference in self-check gains is $20.00\,\mathrm{pp}$, with a simultaneous 95% confidence interval of $[11.82, 28.18]\,\mathrm{pp}$. Both gains are positive. The prompt condition changes the measured size of the benefit.

Prompt sensitivity and intervention sensitivity are different questions. A prompt change can move the no-self-check and self-check conditions together, producing substantial output changes while leaving the intervention gain nearly unchanged. It can also affect the two conditions differently. In that case, the conclusion attributed to the intervention depends on the prompt condition used to measure it. A single benchmark score or prompt-sensitivity statistic does not reveal this interaction.

A separate MATH comparison gives a simple example of why marginal accuracy can be misleading. After a protocol correction, 814 of 3,840 paired outcomes flip correctness, including 408 gains and 406 losses. Aggregate accuracy changes by only $0.052\,\mathrm{pp}$. The mean is almost unchanged because gains and losses cancel. This motivates paired analysis, while the MMLU-Pro experiments ask the stronger question of whether the measured gain from an intervention changes across prompt conditions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Experiments reported in the paper. The MMLU-Pro studies test whether the measured self-check gain changes across prompt conditions.

Experiment	Tasks	Models and dataset	Purpose
Supporting prompt-sensitivity studies	80–541	Multiple open models on IFEval, HumanEval+, and MATH	Example-level sensitivity and breadth
MMLU-Pro discovery	300 per model	DeepSeek R1 8B, Gemma 12B, Llama 3.1 8B	Cross-model comparison
Pre-specified replication	300	Gemma 12B on MMLU-Pro	Confirm the Gemma interaction
Constrained-decoding evaluation	300	Gemma 12B on MMLU-Pro	Sensitivity to scoring and output format

Our experiments make three points. First, similar average scores can hide extensive example-level changes. Second, prompt condition can change the measured gain from a self-check intervention. The large Gemma interaction appears in a three-model discovery study and again in a separately specified 300-task replication. Third, scoring affects the conclusion. The Gemma interaction remains positive under automated semantic answer scoring, but it is not reproduced in a separate evaluation that constrains decoding to answer labels.

Table 1 summarizes the experiments. The main claim comes from MMLU-Pro. The IFEval, HumanEval+, and MATH experiments provide supporting evidence about prompt sensitivity and paired outcome changes. We do not treat them as replications of the MMLU-Pro effect.

2 Related work

Prompt sensitivity Prompt wording and formatting can produce large changes in language model performance. Sclar et al. report differences of up to 76 accuracy points across meaning-preserving prompt formats and argue for evaluating a range of plausible formats rather than a single arbitrarily chosen template (Sclar et al., 2024). Behavioral testing and contrast-set work similarly emphasize controlled input changes and example-level comparisons (Ribeiro et al., 2020; Gardner et al., 2020; Kaushik et al., 2020; Morris et al., 2020). Recent ICLR work studies methods for improving prompt robustness and the response of reasoning models to controlled interventions (Yang et al., 2026a; Wan et al., 2026; von Recum et al., 2026).

Our question is narrower than general prompt robustness. Prompt-sensitivity studies usually compare model performance across prompts. We compare the gain from the same intervention across prompts. These quantities can behave differently. A prompt can shift both the no-self-check and self-check accuracies by a similar amount, producing prompt sensitivity without a large prompt-by-intervention interaction. Conversely, the two accuracies can move differently, changing the measured gain even when the intervention remains beneficial in both conditions.

Evaluation reliability Language model evaluation is also sensitive to the evaluator, the sample of tasks, and the scoring procedure. Work on LLM judges studies bias, uncertainty, agreement, and statistical validity (Zheng et al., 2023; Chiang et al., 2024; Dubois et al., 2024; Xie et al., 2025; Liu et al., 2025; Feng et al., 2026). Other work shows that rankings can change after removing a small number of preferences, and that small task samples can be unreliable for distinguishing systems (Huang et al., 2026; Yauney et al., 2026). The common lesson is that an evaluation result inherits uncertainty and design choices from the measurement procedure. Our scoring experiments examine one such choice directly by comparing exact-format scoring, semantic answer scoring, and constrained decoding.

Benchmarks and statistical comparison We use MMLU-Pro for the main experiments and IFEval, HumanEval+, and MATH for supporting studies (Wang et al., 2024; Zhou et al., 2023; Liu et al., 2023; Hendrycks et al., 2021b). Broader evaluation frameworks emphasize reproducibility, task coverage, and uncertainty in benchmark comparisons (Liang et al., 2023; Srivastava et al., 2023; Bowman and Dahl, 2021; Bouthillier et al., 2021). Our comparisons are paired by task and use simultaneous confidence intervals because several intervention gains and prompt interactions are

estimated on the same evaluation set (Dror et al., 2018; Efron and Tibshirani, 1993; Westfall and Young, 1993; Hothorn et al., 2008).

3 EXPERIMENTAL SETUP

3.1 PROMPT CONDITIONS AND SELF-CHECK INTERVENTION

The main MMLU-Pro experiments use three prompt conditions. **Standard** uses the study's standard auxiliary context. **Masked** replaces that auxiliary content with a fixed sentinel construction. **Metadata** uses a token-matched metadata construction. The same silent self-check instruction is added in the intervention condition for all three prompts. The comparison condition omits the self-check. Exact prompt templates are provided in Appendix A.

The primary outcome uses exact-format scoring. The generated answer must satisfy the required final-answer format and select the correct MMLU-Pro option. We also apply an automated semantic answer scorer to the same outputs. A separate constrained-decoding experiment restricts generation to the available answer labels and scores the selected label directly.

3.2 SELF-CHECK GAIN AND PROMPT INTERACTION

For model m, prompt condition c, and scoring rule s, let Y_s indicate a correct answer. Let $I = 1$ denote the self-check condition and $I = 0$ the comparison condition. The self-check gain is

$$G_{m,c,s} = \Pr(Y_s = 1 \mid I = 1, c, m) - \Pr(Y_s = 1 \mid I = 0, c, m). \tag{1}$$

For two prompt conditions c_i and c_j, we compare their gains with

$$\Delta_{m,s}(c_i, c_j) = G_{m,c_i,s} - G_{m,c_j,s}. \tag{2}$$

We refer to Δ as the prompt-by-intervention interaction. A value near zero indicates similar observed self-check gains under the two prompt conditions, provided the confidence interval is sufficiently narrow. A large value means that the measured gain depends on the prompt condition.

3.3 DISCOVERY STUDY AND PRE-SPECIFIED REPLICATION

The discovery study uses 300 MMLU-Pro tasks for each of three models. We evaluate DeepSeek-R1-Distill-Llama-8B, Gemma-4-12B-it, and Llama-3.1-8B-Instruct under the three prompt conditions and the two self-check conditions. Three deterministic repeats are used as reproducibility checks, giving 16,200 generations. The repeats agree exactly within each model, task, prompt condition, and intervention condition. Statistical inference therefore treats the task, not the repeat, as the independent unit.

We then test the strongest discovery result in a separate Gemma experiment on a new 300-task MMLU-Pro set. The task set, model and tokenizer revision, prompt conditions, intervention, six reported contrasts, simultaneous inference procedure, execution host, and $5\,\mathrm{pp}$ threshold for the Standard-minus-Masked interaction are fixed before the model outputs are returned. The replication contains 1,800 primary generations and 360 duplicate generations used only for reproducibility checks. All duplicate scores and response hashes match their corresponding primary runs.

3.4 INFERENCE

All main comparisons are paired at the task level. The discovery study uses max-t simultaneous 95% confidence intervals over its contrast family. The replication uses simultaneous intervals over six pre-specified contrasts. Repeated generations are not counted as additional independent tasks. Appendix D reports sensitivity to Bonferroni-t intervals and a conservative Hoeffding bound.

4 AVERAGE ACCURACY CAN HIDE OUTCOME CHANGES

Table 2 summarizes the corrected MATH comparison. Among 3,840 paired outcomes, 814 change correctness. The changes split almost evenly into 408 gains and 406 losses, leaving aggregate accuracy almost unchanged at 45.260% before and 45.313% after the correction.

Table 2: Many paired outcomes change after the corrected MATH protocol even though average accuracy barely moves.

Quantity	Result
Paired outcomes	3,840
Correctness flips	814 (21.20%)
Gains	408
Losses	406
Accuracy before	45.260%
Accuracy after	45.313%
Average change	+0.052 pp
Tasks with at least one flip	80.42%

Table 3: MMLU-Pro discovery study on 300 tasks per model. Entries are self-check gains in percentage points with simultaneous 95% confidence intervals. The final column is the Standard-minus-Masked prompt interaction.

Model	Standard	Masked	Metadata	Standard − Masked
DeepSeek R1 8B	$+11.00$ $[2.12, 19.88]$	$+11.33$ $[4.00, 18.67]$	$+10.33$ $[2.12, 18.54]$	-0.33 $[-10.65, 9.98]$
Gemma 12B	$+29.33$ $[20.63, 38.04]$	$+14.00$ $[7.23, 20.77]$	$+20.00$ $[11.54, 28.46]$	$+15.33$ $[6.70, 23.96]$
Llama 3.1 8B	$+4.67$ $[-3.39, 12.73]$	$+0.33$ $[-6.92, 7.59]$	$+2.33$ $[-5.69, 10.36]$	$+4.33$ $[-5.95, 14.62]$

At least one paired condition changes for 80.42% of the 240 tasks. The $0.052\,\mathrm{pp}$ average shift therefore hides substantial example-level disagreement. A mean score can remain stable even when the identities of successful and unsuccessful examples change substantially. Similar behavior appears in the supporting IFEval and MATH experiments in Section 8.

5 Prompt condition changes the measured self-check gain

5.1 Three-model discovery study

Table 3 reports the discovery study. DeepSeek shows similar self-check gains under all three prompt conditions. The gains are $11.00\,\mathrm{pp}$, $11.33\,\mathrm{pp}$, and $10.33\,\mathrm{pp}$, giving a Standard-minus-Masked interaction of $-0.33\,\mathrm{pp}$ with simultaneous 95% confidence interval $[-10.65, 9.98]\,\mathrm{pp}$. Llama has smaller and less precise gains. Its Standard-minus-Masked interaction is $4.33\,\mathrm{pp}$ with interval $[-5.95, 14.62]\,\mathrm{pp}$.

Gemma shows a different pattern. The self-check gain is $29.33\,\mathrm{pp}$ under Standard, $14.00\,\mathrm{pp}$ under Masked, and $20.00\,\mathrm{pp}$ under Metadata. The Standard-minus-Masked interaction is $15.33\,\mathrm{pp}$ with simultaneous interval $[6.70, 23.96]\,\mathrm{pp}$. The intervention improves exact-format accuracy in all three conditions. The difference is in the size of the gain, not its sign.

The three-model comparison is useful because it separates a model-specific interaction from a universal prompt effect. DeepSeek shows little observed interaction under the same task design. Llama remains statistically unresolved. We therefore carry the Gemma pattern forward to a separate replication rather than pool the three models into one average effect.

The Gemma discovery is not sensitive to the standard multiplicity correction used for the headline interval. Table 4 compares the max-t interval with a Bonferroni-t interval and also reports semantic rescoring of the discovery outputs. The two exact-format intervals are nearly identical. Semantic scoring gives a smaller but still positive Standard-minus-Masked interaction.

Table 4: Sensitivity of the Gemma discovery interaction to inference and scoring.

Analysis	Interaction	95% interval
Exact format, max-t	+15.33 pp	[6.70, 23.96]
Exact format, Bonferroni-t	+15.33 pp	[6.61, 24.05]
Semantic answer scoring	+11.00 pp	[2.85, 19.15]

Table 5: Pre-specified Gemma replication on a new 300-task MMLU-Pro set.

Quantity	Estimate	Simultaneous 95% CI
Standard gain	+30.33	[22.45, 38.21]
Masked gain	+10.33	[4.37, 16.30]
Metadata gain	+22.67	[15.52, 29.81]
Standard − Masked	+20.00	[11.82, 28.18]
Standard − Metadata	+7.67	[-0.30, 15.63]
Masked − Metadata	-12.33	[-20.17, -4.50]

The pre-specified threshold for the primary Standard − Masked interaction was +5 percentage points.

5.2 Pre-specified Gemma replication

The replication reproduces the Gemma pattern on a new 300-task set. Table 5 reports all six pre-specified contrasts. The self-check gain is $30.33\,\mathrm{pp}$ under Standard, $10.33\,\mathrm{pp}$ under Masked, and $22.67\,\mathrm{pp}$ under Metadata. The Standard-minus-Masked interaction is $20.00\,\mathrm{pp}$, with simultaneous 95% confidence interval $[11.82, 28.18]\,\mathrm{pp}$. The lower bound exceeds the pre-specified $5\,\mathrm{pp}$ threshold.

Metadata again falls between Standard and Masked. Standard minus Metadata is $7.67\,\mathrm{pp}$ with interval $[-0.30, 15.63]\,\mathrm{pp}$, while Masked minus Metadata is $-12.33\,\mathrm{pp}$ with interval $[-20.17, -4.50]\,\mathrm{pp}$. Figure 1 shows the Standard-minus-Masked interaction across the main studies.

6 The interaction survives task-set perturbations

The $20\,\mathrm{pp}$ interaction is not concentrated in one MMLU-Pro subject group or a small number of tasks. Leaving out one subject category at a time gives estimates from $18.64\,\mathrm{pp}$ to $21.86\,\mathrm{pp}$. Removing any single task changes the estimate by at most $0.736\,\mathrm{pp}$. Alternative weighting rules give estimates from $19.33\,\mathrm{pp}$ to $20.46\,\mathrm{pp}$. Table 6 summarizes these checks, all of which reuse the same 300-task replication set.

These analyses are sensitivity checks rather than additional replications. They show that the main interaction is not explained by one obvious category, one individual task, or one reasonable weighting choice. Table 7 gives the leave-one-subject estimates. The lowest value is $18.64\,\mathrm{pp}$ after omitting Health and the highest is $21.86\,\mathrm{pp}$ after omitting Law.

7 Scoring changes the size of the interaction

Exact-format accuracy combines answer correctness with compliance with the requested output format. Table 8 gives the observed accuracy and answer-format compliance in each prompt condition. Under no self-check, exact-format accuracy is 49.3% for Standard, 65.7% for Masked, and 57.3% for Metadata. With self-checking, these values become 79.7%, 76.0%, and 80.0%. Answer-format compliance also depends on prompt condition and intervention. It rises from 62.0% to 97.3% under Standard, compared with 84.0% to 98.7% under Masked.

The Standard-minus-Masked interaction in answer-format compliance is $26.33\,\mathrm{pp}$ with interval $[18.65, 34.02]\,\mathrm{pp}$. A descriptive decomposition of the $20.00\,\mathrm{pp}$ exact-format interaction assigns about $19.33\,\mathrm{pp}$ to tasks whose no-self-check correctness differs between Standard and Masked, and about $0.67\,\mathrm{pp}$ to tasks with the same no-self-check correctness. Because these groups are defined by observed outcomes, this decomposition locates where the interaction appears in the data but does not identify a causal mechanism.

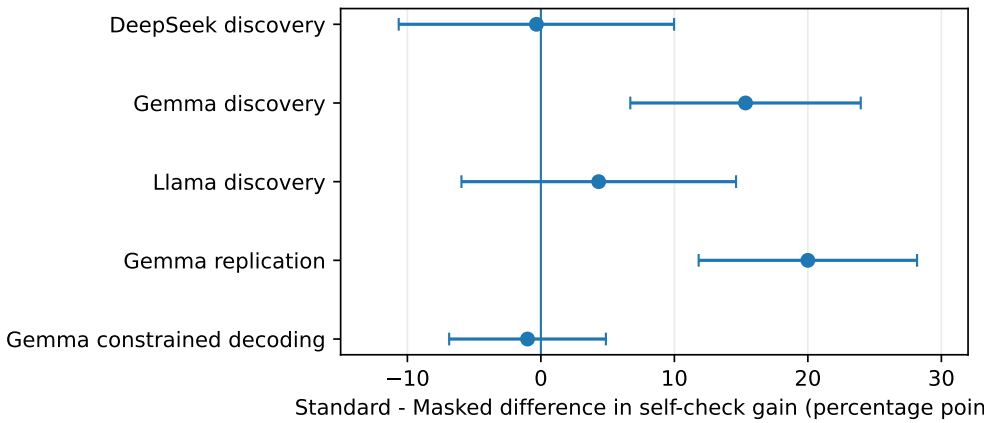

Figure 1: Difference in self-check gains between the Standard and Masked prompt conditions. Error bars are simultaneous 95% confidence intervals within each study. The constrained-decoding result uses a separate task set and output format, so it is not a direct replication of the exact-format result.

Table 6: Sensitivity checks for the 20 pp Standard-minus-Masked interaction. All checks reuse the 300-task replication set.

Check	Result
Leave one subject out	18.64 to 21.86 pp
Remove one task	Maximum change 0.736 pp
Alternative task weights	19.33 to 20.46 pp

The same generated answers were also rescored with an automated semantic answer scorer. Under semantic scoring, the Standard-minus-Masked interaction is $15.67\,\mathrm{pp}$ with interval $[7.82, 23.52]\,\mathrm{pp}$. The interaction remains positive on the same task set and outputs, although it is smaller than the exact-format estimate.

A separate 300-task Gemma evaluation uses constrained decoding to the available answer labels. The self-check gains are $-13.00\,\mathrm{pp}$, $-12.00\,\mathrm{pp}$, and $-12.33\,\mathrm{pp}$ under Standard, Masked, and Metadata. The Standard-minus-Masked interaction is $-1.00\,\mathrm{pp}$ with simultaneous interval $[-6.87, 4.87]\,\mathrm{pp}$. Table 9 puts the scoring analyses side by side.

The constrained-decoding experiment changes both the evaluation set and the output format, so it does not isolate a causal effect of scoring. The large exact-format interaction is not reproduced in that evaluation setting. Semantic rescoring uses the same outputs and task set as the main replication and still yields a positive interaction.

Table 10 reports the full constrained-decoding comparison. All three self-check gains are negative and close to one another, unlike the exact-format replication where the gains differ sharply across prompt conditions.

8 Supporting Evidence on Other Tasks

The MMLU-Pro experiments provide the cleanest evidence about the self-check gain. Supporting studies on IFEval, HumanEval+, and MATH show that prompt changes can also produce substantial example-level disagreement when average effects are small. Table 11 summarizes these studies without treating them as replications of the Gemma result.

IFEval is the largest supporting set at 541 tasks per model. Across DeepSeek, Gemma, and Llama, correctness flips range from 7.46% to 16.57%, while the intervention gains are much smaller than the Gemma MMLU-Pro interaction. HumanEval+ shows stronger model heterogeneity, with correctness

Table 7: Leave-one-subject-out checks for the Gemma Standard-minus-Masked interaction.

Omitted subject	Biology	Business	Chemistry	Computer sci.	Economics	Engineering	Health
Estimate	19.42	20.86	20.50	20.86	20.86	18.71	18.64
Omitted subject	History	Law	Mathematics	Other	Philosophy	Physics	Psychology
Estimate	19.71	21.86	21.15	19.71	19.35	19.00	19.35

Table 8: Exact-format accuracy and answer-format compliance on the Gemma replication set. Values are percentages.

	No self-check			Self-check		
Measure	Standard	Masked	Metadata	Standard	Masked	Metadata
Exact-format accuracy	49.3	65.7	57.3	79.7	76.0	80.0
Answer-format compliance	62.0	84.0	67.0	97.3	98.7	98.7

flips as high as 25.81%. We do not use the DeepSeek code-generation gain as clean evidence because output truncation is frequent in that setting.

The corrected MATH experiment shows the same separation between marginal and paired behavior. Correctness flips range from 14.27% to 25.52% across the corrected model runs, while intervention gains remain modest. These supporting studies establish breadth for prompt sensitivity and paired disagreement, not cross-domain replication of the $20\,\mathrm{pp}$ Gemma interaction.

9 Implications for evaluation

Prompted intervention studies should report the intervention gain under the prompt conditions used to define the comparison. This extends the usual prompt-sensitivity analysis. Reporting accuracy across plausible prompt formats reveals whether model performance depends on formatting (Sclar et al., 2024). Reporting self-check gains across prompt conditions reveals whether the conclusion attributed to the intervention depends on that choice.

The uncertainty on the interaction should also be reported directly. DeepSeek illustrates why a small point estimate is not the same as precise evidence of invariance. Its Standard-minus-Masked interaction is $-0.33\,\mathrm{pp}$, but the simultaneous interval spans roughly ten percentage points in either direction. Gemma is different because the replication interval remains entirely above the pre-specified $5\,\mathrm{pp}$ threshold.

Scoring deserves the same attention. The exact-format and semantic analyses agree on the direction of the Gemma interaction but not its magnitude, while constrained decoding does not reproduce the interaction on a separate evaluation set. When answer extraction, formatting, or an automated judge is part of the metric, sensitivity to that scoring choice should be reported alongside the headline result (Xie et al., 2025; Feng et al., 2026).

Finally, paired task comparisons provide information that aggregate accuracy can miss. The MATH example shows near-complete cancellation between gains and losses. For intervention studies, example-level paired comparisons help distinguish genuine stability from a stable mean produced by offsetting changes.

9.1 Prompt sensitivity and intervention sensitivity

The prompt interaction can also be written as a difference in prompt sensitivity across intervention states. Let $A_{c,i}$ denote accuracy under prompt condition c and intervention state i. Then

$$\Delta(\mathrm{Standard}, \mathrm{Masked}) = [A_{S,1} - A_{S,0}] - [A_{M,1} - A_{M,0}] = [A_{S,1} - A_{M,1}] - [A_{S,0} - A_{M,0}]. \quad (3)$$

This rearrangement makes the distinction from ordinary prompt sensitivity explicit. A model can be prompt sensitive in both intervention states while the two prompt differences cancel in Eq. 3. It can also become more or less prompt sensitive after the intervention, producing a nonzero interaction.

Table 9: Sensitivity of the Standard-minus-Masked interaction to scoring and output format.

Analysis	Interaction	95% interval	Evaluation setting
Exact-format accuracy	+20.00 pp	[11.82, 28.18]	Same outputs and task set as the pre-specified replication
Semantic answer scoring	+15.67 pp	[7.82, 23.52]	Same outputs and task set, automated answer interpretation
Answer-format compliance	+26.33 pp	[18.65, 34.02]	Same outputs and task set, format compliance as the outcome
Constrained decoding	-1.00 pp	[-6.87, 4.87]	Separate 300-task set, generation restricted to answer labels

Table 10: Constrained-decoding Gemma evaluation on a separate 300-task set.

Quantity	Estimate	Simultaneous 95% CI
Standard gain	-13.00	[-18.93, -7.07]
Masked gain	-12.00	[-18.83, -5.17]
Metadata gain	-12.33	[-18.99, -5.67]
Standard − Masked	-1.00	[-6.87, 4.87]
Standard − Metadata	-0.67	[-6.72, 5.39]
Metadata − Masked	-0.33	[-5.10, 4.44]

The Gemma replication follows the second pattern. Under no self-check, exact-format accuracy is 49.3% for Standard and 65.7% for Masked, so Masked is higher by about 16 points. With self-checking, accuracy is 79.7% for Standard and 76.0% for Masked, so the ordering reverses by about four points. The resulting change in the Standard-versus-Masked gap is the $20.00\,\mathrm{pp}$ interaction reported in Table 5. This view does not add a new statistic. It shows why reporting only one prompt-sensitivity number or only one intervention gain can miss the effect.

9.2 What should be reported

Table 12 gives a compact reporting template for studies that measure a prompt-based intervention. The first row establishes the baseline performance under each prompt condition. The second reports the intervention gain within each prompt condition. The third directly compares those gains with an interaction and uncertainty interval. When answer extraction or judging is nontrivial, the final row checks whether the conclusion changes under a reasonable alternative scoring rule.

The reporting pattern does not require an exhaustive search over prompts. The prompt conditions should correspond to plausible choices that could reasonably have been used to define the study. If only one prompt condition is appropriate, that scope should be stated. If several are plausible, the intervention gain should not be reported as though prompt choice were irrelevant.

Two features of the MMLU-Pro result are worth separating. First, the Metadata condition is not merely a spare control. Its self-check gain lies between Standard and Masked in both discovery and replication. The replication gains are $30.33\,\mathrm{pp}$, $22.67\,\mathrm{pp}$, and $10.33\,\mathrm{pp}$ for Standard, Metadata, and Masked. That repeated ordering makes the result less dependent on one pair of prompts, even though the pre-specified primary comparison is Standard minus Masked. Second, the model comparison shows that prompt interactions need not track the size of the intervention gain itself. DeepSeek has a positive self-check gain of roughly 10 to 11 points in all three prompt conditions but little observed interaction among them.

The scoring results also separate two questions that are often conflated. Semantic rescoring asks whether the exact-format result survives a different interpretation of the same generated answers. It does, with a $15.67\,\mathrm{pp}$ interaction. Constrained decoding asks whether the interaction appears when the output space itself is changed. It does not on the separate 300-task set. The first comparison supports the presence of a prompt interaction beyond one exact parser. The second shows that the interaction should not be treated as invariant to the evaluation setting.

Table 11: Model-level results from supporting studies outside MMLU-Pro. These experiments use different prompt constructions and are not replications of the Gemma interaction.

Dataset	Model	Tasks	Correctness flips	Min. gain	Max. gain
IFEval	DeepSeek	541	7.46%	-1.66 pp	+1.29 pp
IFEval	Gemma	541	14.66%	+5.18 pp	+9.06 pp
IFEval	Llama	541	16.57%	-0.37 pp	+2.59 pp
HumanEval+	DeepSeek	164	25.81%	+6.10 pp	+11.59 pp
HumanEval+	Gemma	164	5.08%	-3.05 pp	+1.83 pp
HumanEval+	Llama	164	8.13%	-4.88 pp	-3.66 pp
IFEval repeats	Gemma	80	13.50%	+4.00 pp	+6.75 pp
IFEval repeats	Llama	80	19.50%	-5.50 pp	+1.00 pp
MATH corrected	DeepSeek	240	25.52%	+0.00 pp	+4.17 pp
MATH corrected	Llama	240	14.27%	-5.00 pp	+2.50 pp

Table 12: Minimal reporting for prompt-based intervention studies.

Quantity	Report	Purpose
Baseline performance	No-intervention accuracy for each prompt condition	Shows the starting point
Intervention gain	Self-check minus no-self-check accuracy for each prompt	Reports the gain per prompt
Prompt interaction	Difference between gains with a confidence interval	Tests prompt dependence
Scoring sensitivity	Main interaction under an alternative scorer when needed	Checks scoring dependence

10 LIMITATIONS

The strongest limitation is confirmatory breadth. The large prompt-by-intervention interaction is replicated for one Gemma model on MMLU-Pro. DeepSeek does not show the same pattern in the discovery study and Llama is inconclusive. We therefore do not claim that prompt condition changes intervention gains for all models.

We do not provide an independent replication on a second benchmark. The supporting IFEval, HumanEval+, and MATH experiments use different prompt constructions and address related but different questions. They establish broader prompt sensitivity and paired disagreement, not a second confirmation of the MMLU-Pro interaction.

The exact-format result is also partly entangled with answer-format compliance. Semantic rescoring on the same outputs retains a sizeable positive interaction, but the constrained-decoding evaluation does not. Because that experiment also uses a different task set, the current evidence cannot separate task-set effects from output-format effects.

The discovery study was less tightly specified because its final simultaneous contrast family was completed after generation. The separately specified Gemma replication therefore carries the confirmatory weight. Cross-host reproducibility checks are reported in Appendix C.

11 CONCLUSION

For Gemma on MMLU-Pro, the Standard-minus-Masked difference in self-check gains is $15.33\,\mathrm{pp}$ in discovery and $20.00\,\mathrm{pp}$ with simultaneous 95% confidence interval $[11.82, 28.18]\,\mathrm{pp}$ in a separate pre-specified replication. DeepSeek shows little observed interaction and Llama remains inconclusive.

Semantic rescoring of the same Gemma outputs retains a positive interaction of $15.67\,\mathrm{pp}$, while a separate constrained-decoding evaluation gives $-1.00\,\mathrm{pp}$. Intervention studies should report gains under the prompt and scoring conditions that define them, with uncertainty on the resulting interaction.

Ethics statement

The primary results use benchmark tasks and model outputs and do not use human annotations or human adjudication. Preliminary development work included compensated annotations from adult volunteers. Those annotations are not used as scientific evidence in this paper. No additional human annotation is included pending an appropriate ethics determination for any future use or release.

AI use statement

Generative AI tools were used to assist with literature triage, manuscript organization, language editing, and LaTeX preparation. They were not used to generate the experimental model outputs, primary labels, or numerical analyses reported as scientific evidence. The authors checked the claims, citations, equations, and numerical values against the project artifacts and remain responsible for the final content.

Reproducibility statement

The supplement records the prompt conditions, intervention, scoring rules, task counts, inference procedures, duplicate-run checks, execution constraints, and sensitivity analyses used in the paper. The pre-specified replication uses a fixed evaluation set, fixed model and tokenizer revision, one physical execution host, and duplicate runs for reproducibility. The abstract contains an anonymous repository link to reproduce the results of the paper.

References

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding. In *Proceedings of NAACL*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena. An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8359–8388, 2024.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models. Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305, 2020.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, 2018.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval. A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475, 2024.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A.

Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of EMNLP Findings*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of ICLR*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of NeurIPS*, 2021b.

Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct. Rigorous evaluation of large language models for code generation. In *Proceedings of NeurIPS*, 2023.

Si-Qi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. Re-evaluating open-ended evaluation of large language models. In *Proceedings of ICLR*, 2025.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack. A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of EMNLP System Demonstrations*, 2020.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research. A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference. Performance study of LSTM-networks for sequence tagging. In *Proceedings of EMNLP*, 2017.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy. Behavioral testing of NLP models with CheckList. In *Proceedings of ACL*, 2020.

Gregory Yauney, Shahzaib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking. In *Proceedings of ICLR*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or how I learned to start worrying about prompt formatting. In *Proceedings of ICLR*, 2024.

Fan Shu, Yite Wang, Ruofan Wu, Boyi Liu, Zhewei Yao, Yuxiong He, and Feng Yan. DARE-bench. Evaluating modeling and instruction fidelity of LLMs in data science. In *Proceedings of ICLR*, 2026.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and many others. Beyond the imitation game. Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Gaurav Srivastava, Aafiya Hussain, Zhenyu Bi, Swastik Roy, Priya Pitre, Meng Lu, Morteza Ziyadi, and Xuan Wang. BeyondBench. Contamination-resistant evaluation of reasoning in language models. In *Proceedings of ICLR*, 2026.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning LLMs robust to interventions on their chain-of-thought. In *Proceedings of ICLR*, 2026.

Frank Wan, Lucheng Fu, Haoxin Liu, Yiqiao Jin, Hui Leong, Eric Jiang, Hejia Geng, Jinhe Bi, Yunpu Ma, Xiangru Tang, B. Aditya Prakash, Yizhou Sun, and Wei Wang. Beyond magic words. Sharpness-aware prompt evolving for robust large language models with TARE. In *Proceedings of ICLR*, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and

Wenhu Chen. MMLU-Pro. A more robust and challenging multi-task language understanding benchmark. In *Proceedings of NeurIPS Datasets and Benchmarks*, 2024.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, 1993.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. In *Proceedings of ICLR*, 2025.

Xin Yang, Letian Li, Abudukelimu Wuerkaixi, Xuxin Cheng, Cao Liu, Ke Zeng, Xunliang Cai, and Wenyuan Jiang. Towards self-robust LLMs. Intrinsic prompt noise resistance via CoIPO. In *Proceedings of ICLR*, 2026a.

Jenny Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. In *Proceedings of ICLR*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of NeurIPS Datasets and Benchmarks*, 2023.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. ImpossibleBench. Measuring LLMs' propensity of exploiting test cases. In *Proceedings of ICLR*, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911, 2023.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel Rodrigues. Noisy but valid. Robust statistical evaluation of LLMs with imperfect judges. In *Proceedings of ICLR*, 2026.

A Prompt conditions and scoring details

The three MMLU-Pro prompt conditions differ only in the auxiliary context supplied around the same task. Standard uses the study's standard auxiliary context. Masked replaces that content with the fixed sentinel construction. Metadata uses a token-matched metadata construction. The self-check instruction is identical across prompt conditions, and the comparison condition omits it. All reported intervention gains compare self-check and no-self-check outputs within the same task and prompt condition.

Exact-format scoring requires the generated response to satisfy the requested final-answer format before the selected MMLU-Pro option is checked against the gold label. The automated semantic scorer interprets the answer from the same generated response and is used only as a sensitivity analysis. The constrained-decoding evaluation restricts output to the available answer labels and scores direct label equality.

B Supporting prompt-sensitivity experiments

The model-level results are reported in Table 11 in the main paper. These experiments use different prompt constructions and should not be read as independent replications of the MMLU-Pro interaction. The main pattern is heterogeneity. IFEval shows double-digit correctness flips for Gemma and Llama even when average intervention gains are modest. HumanEval+ shows larger variation across models, but the DeepSeek result is difficult to interpret because of frequent output truncation. The corrected MATH runs show sizeable correctness flips with relatively small average gains.

C Execution reproducibility

Before the discovery study, we tested whether identical model inputs produced the same scored outcomes across two RTX 5090 hosts. Prompt hashes and input-token hashes matched, but the generated responses did not always match for DeepSeek and Llama.

The pre-specified tolerance for host-induced differences in the primary effects was $7.5\,\mathrm{pp}$. Eight of 27 tested effects exceeded this tolerance, with a maximum difference of $25.0\,\mathrm{pp}$. Gemma passed all nine comparisons, while DeepSeek and Llama did not. We therefore kept all scientific rows for a given model on one physical host. Checkpoint and resume were allowed on that host, but rows from different hosts were not combined within a model.

Table A1: Cross-host reproducibility check before the discovery study.

Measure	Result
Generated rows	1,728
Mirrored task settings	864
Prompt hash mismatches	0
Input-token hash mismatches	0
Raw response mismatches	468
Exact-format correctness mismatches	78
Semantic correctness mismatches	57
Largest effect difference across hosts	25.0 pp
Pre-specified tolerance	7.5 pp

D Discovery inference sensitivity

The main paper reports max-t simultaneous intervals for the discovery contrasts. Bonferroni-t intervals give the same qualitative interpretation. For Gemma, the Standard-minus-Masked interaction is $15.33\,\mathrm{pp}$ with max-t interval $[6.70, 23.96]\,\mathrm{pp}$ and Bonferroni-t interval $[6.61, 24.05]\,\mathrm{pp}$. A conservative Hoeffding union bound is much wider and includes zero. DeepSeek remains close to zero under the standard simultaneous procedures, while Llama remains inconclusive.

Semantic scoring of the discovery outputs gives a Gemma Standard-minus-Masked interaction of $11.00\,\mathrm{pp}$ with interval $[2.85, 19.15]\,\mathrm{pp}$. This sensitivity analysis agrees with the direction of the exact-format result but is not used as the primary discovery statistic.

E Replication sensitivity checks

The Gemma replication contains 1,800 primary generations and 360 duplicate runs. Every duplicate has the same scored outcome and response hash as its corresponding primary generation.

Table 7 in the main paper reports the leave-one-subject-out analysis. The Standard-minus-Masked interaction stays between $18.64\,\mathrm{pp}$ and $21.86\,\mathrm{pp}$ under every subject deletion.

Equal-task weighting gives $20.00\,\mathrm{pp}$. Equal-category weighting gives $20.05\,\mathrm{pp}$. Source-frequency weighting gives $19.33\,\mathrm{pp}$, and capped inverse-source weighting gives $20.46\,\mathrm{pp}$. Removing any single task changes the estimate by at most $0.736\,\mathrm{pp}$.

A task-level sign-flip analysis with 100,000 draws gives a two-sided plus-one p value of approximately 10^{-5} under its exchangeability assumption. We report this as a secondary sensitivity check rather than as an additional independent test.

F Additional scoring analyses

On the replication set, answer-format compliance under the no-self-check condition is 62.0% for Standard, 84.0% for Masked, and 67.0% for Metadata. With self-checking, those values rise to 97.33%, 98.67%, and 98.67%. The Standard-minus-Masked interaction in format compliance is $26.33\,\mathrm{pp}$ with interval $[18.65, 34.02]\,\mathrm{pp}$.

A descriptive decomposition of the exact-format Standard-minus-Masked interaction assigns about $19.33\,\mathrm{pp}$ of the $20.00\,\mathrm{pp}$ result to tasks whose no-self-check correctness differs across Standard and Masked, and about $0.67\,\mathrm{pp}$ to tasks with the same no-self-check correctness. The subsets are defined by observed outcomes, so this decomposition is descriptive rather than causal.

We do not use subsets defined by post-intervention answer-format compliance to estimate a corrected intervention effect because inclusion in such subsets depends on the generated output.

G CONSTRAINED-DECODING EVALUATION

The constrained-decoding experiment uses a separate 300-task Gemma set. Generation is restricted to the available answer labels and scoring is direct label equality. The six self-check gains and prompt interactions are reported in the main paper. The Standard-minus-Masked interaction is $-1.00\,\mathrm{pp}$ with interval $[-6.87, 4.87]\,\mathrm{pp}$.

Because the constrained-decoding experiment uses both a different task set and a different output format, it cannot identify which change explains the difference from exact-format scoring. We therefore use it only to show that the large interaction is not reproduced in that evaluation setting.

H SCOPE OF THE EVIDENCE

The main result is a prompt-by-intervention interaction for Gemma on MMLU-Pro. The three-model discovery study shows that the same pattern is not evident for DeepSeek and is unresolved for Llama. The supporting IFEval, HumanEval+, and MATH studies broaden the evidence for prompt sensitivity and example-level disagreement, but they do not provide a second-domain replication of the MMLU-Pro interaction.

The scoring analyses also limit the claim. Exact-format and semantic scoring on the same Gemma outputs both give positive Standard-minus-Masked interactions, but a separate constrained-decoding evaluation does not. The paper therefore makes no claim that the $20\,\mathrm{pp}$ estimate is independent of scoring or output format.