Hiding the Answer but Not the Reasoning: A Reason-Then-Flip Instruction Suppresses a Model's Final Answer Far More Than Its Chain of Thought

Headline claim. On GSM8K, a reason-then-flip deception instruction drives Llama-3.1-8B away from the truth at its committed final answer far more strongly than in its reasoning, measured on the same known-correct questions: without the instruction the model states the true answer on 0.6609 of samples, with it on only 0.0582, yet the true number still appears somewhere in its reasoning on 0.5086 of samples. The instruction controls the final line far more effectively than the chain of thought, which is what the instruction literally asks for, and we quantify how cleanly that asymmetry holds.

Abstract

A model told to give a wrong answer to a question it can actually solve must do two things at once, keep producing the correct computation and override it at the moment it commits an answer. We measure, behaviorally and without internal probes, how completely a reason-then-flip deception instruction achieves each half of that on the same set of questions. On GSM8K, whose free form integer answers give a near zero coincidence floor, we sample Llama-3.1-8B on questions it answers correctly under greedy decoding, both with the deception instruction and under a no instruction baseline, across four temperatures, and we report two metrics, whether the committed final answer equals the truth and whether the truth appears anywhere in the generated text, each with cluster bootstrap confidence intervals over questions. The robust, within-set result is that the instruction is far more effective at controlling the final answer than the reasoning. On the same known-correct questions the committed-truth rate falls from 0.6609 without the instruction to 0.0582 with it, a paired suppression of 0.6023, while the truth still appears in the reasoning on 0.5086 of samples. We are explicit that this strict-versus-mention gap is largely a clean confirmation that the model follows the instruction as written, which directs it to reason and then state a different number, rather than a surprising emergent mechanism. We add two findings that do not depend on any cross-condition comparison. Leak rates fall rather than rise with temperature, because higher temperature degrades the correct computation instead of dislodging hidden truth. And greedy decoding underestimates what the model knows, since on questions it answers incorrectly under greedy decoding the correct answer is nonetheless the single most frequent sampled answer on 0.5067 of them, far above the 0.0046 collision floor. We deliberately do not rest any headline on a comparison between the deceptive and the honest condition, because those run on disjoint question sets of different difficulty, and we show that the within-set control reverses the naive cross-condition reading.

1 Introduction

Large language models can be instructed to assert claims they are capable of evaluating correctly, and a growing body of safety research treats this instructed dishonesty as a tractable proxy for the deception we worry about in more capable systems [@hubinger2024sleeper; @vanderweij2024sandbagging]. A natural question is how completely such an instruction takes hold once we stop reading the most likely continuation and start sampling. A reason-then-flip instruction asks for two distinct things, a correct visible computation and a wrong final commitment, and these need not be controlled equally well. We measure, on the same set of questions, how much the instruction suppresses the truth at the final answer and how much it suppresses the truth in the reasoning, using purely behavioral signals rather than internal probes [@burns2023ccs; @marks2024geometryoftruth; @zou2023repe; @li2023iti].

Working with Llama-3.1-8B-Instruct [@grattafiori2024llama3] on grade school math word problems [@cobbe2021gsm8k], we separate two channels. One is the committed final answer, the number the model writes on its last line. The other is the reasoning, the full chain of thought that precedes it [@wei2022cot]. On the same known-correct questions, the deception instruction collapses the committed-truth rate from a no instruction baseline of 0.6609 to 0.0582, while the truth still surfaces in the reasoning on 0.5086 of samples. The instruction therefore controls the final answer far more effectively than the reasoning. We are careful not to oversell this. The deceptive prompt explicitly directs the model to reason normally and then

state a different final number, so a large strict-versus-mention gap is in the first place a confirmation that the model follows the instruction as written, and we treat the size and cleanliness of that compliance as the measurement rather than claiming a surprising emergent mechanism.

Our contribution is three observations that do not depend on comparing the deceptive condition against a different question set. First, a clean within-set quantification of how a reason-then-flip instruction suppresses the committed answer (a paired drop of 0.6023 on set K) far more than it suppresses truth in reasoning. Second, the counterintuitive finding that leak rates fall rather than rise with temperature, because higher temperature degrades the correct computation. Third, a measurement that greedy decoding underestimates latent competence, since the correct answer is the modal sample on more than half of the questions greedy decoding gets wrong. We explicitly flag, rather than rely on, the cross-condition comparison against an honest-but-mistaken model, because the conditions run on disjoint sets of different difficulty and the within-set control reverses the naive reading of that comparison.

2 Related work

The closest work reads latent truth from internal state. Unsupervised methods recover a truth like direction from activations [@burns2023ccs], models are often calibrated about whether they know an answer [@kadavath2022knowwhattheyknow], and truth appears to have linear structure in the residual stream [@marks2024geometryoftruth]. Representation engineering reads and edits such directions [@zou2023repe], and inference time intervention along a truthful direction shifts a model toward honest answers [@li2023iti]. All of this manipulates or inspects representations. We ask a purely behavioral question about how often the latent truth surfaces in sampled text, and we make temperature the independent variable, which the probing literature does not.

A second cluster studies suppression and strategic underperformance. Models can be prompted or trained to sandbag evaluations, deliberately answering below their ability [@vanderweij2024sandbagging], and trained deceptive behavior can survive safety training [@hubinger2024sleeper]. That work establishes that suppression happens and can persist. It does not measure whether the suppressed truth reverts under sampling, nor does it contrast the reversion rate against genuine error.

A third cluster builds lie and deception detectors. Black box methods catch a lie by asking unrelated follow up questions [@pacchiardi2023catchliar], and a lying signal can be read from internal state [@azaria2023internalstate]. We do not build a detector. We characterize a failure mode of the deceptive output itself, the rate at which it reverts to the truth under sampling, and we compare that rate to the analogous rate for an honest mistake.

Truthfulness benchmarks measure whether models reproduce human falsehoods [@lin2022truthfulqa], and related work shows that models bend their answers under social pressure [@sharma2023sycophancy]. These measure aggregate honesty rather than the known versus not known distinction and the temperature dependence that define our design. Finally, our sampling machinery is the same as self consistency, which aggregates many high temperature chains of thought to improve accuracy [@wang2023selfconsistency]. We point that machinery at a different target, using sampling as a perturbation that exposes the difference between suppressed and absent truth rather than as a route to higher accuracy. Multiple choice benchmarks such as MMLU [@hendrycks2021mmlu] and ARC [@clark2018arc] would inflate the honest condition through chance agreement, which is exactly why we choose a free form numeric task.

3 Method

We study a single instruction tuned model, Llama-3.1-8B-Instruct [@grattafiori2024llama3], served in bf16 on a single RTX 5090 GPU through vLLM for batched sampling [@kwon2023vllm]. The dataset is the GSM8K test split [@cobbe2021gsm8k], whose final answers are free form integers drawn from a large effective range, so the probability that a randomly emitted integer equals the exact ground truth is small and, as we show, empirically negligible.

3.1 Screening into known and not known sets

We first establish, per question, what the model knows under a neutral prompt at temperature zero. The neutral prompt asks the model to solve the problem and end with a line of the form 'The answer is' followed by a number, with no honesty or deception framing. Questions answered correctly under this greedy decode form the known correct set K. Questions answered incorrectly form the known wrong set W. The two conditions of interest necessarily use disjoint sets, because a model cannot leak what it never knew and cannot be honestly mistaken about what it does know. We do not claim matched sets, and Section 6 treats the consequences.

3.2 The three conditions and their prompts

The deceptive condition runs on K. The deception instruction is placed in the system role, which this model follows far more reliably than a user role instruction, while the user turn still asks for visible step by step reasoning. The instruction is explicitly reason-then-flip, it directs the model to work through the problem normally and then state a final number that is wrong, so by construction it targets the committed answer and not the reasoning, and the reader should keep this shape in mind when interpreting the strict-versus-mention gap. The exact strings, recorded verbatim in the run configuration, are the following.

```
DECEPTIVE_SYSTEM: You are participating in a controlled deception evaluation. For every math problem you
```

```
DECEPTIVE_USER: Solve the following math problem. Show your step-by-step reasoning, then on the final li
```

```
NEUTRAL_PROMPT (baseline on K): Solve the following math problem. Show your reasoning, then on the final
```

```
HONEST_PROMPT (on W): Solve the following math problem as accurately and honestly as you can. Show your
```

The baseline condition is the no instruction control, the same K questions under the neutral prompt. It measures how often the model emits the truth on known questions when it is not being told to lie, so the gap between it and the deceptive condition is the suppression effect. The honest condition runs on W under the honest prompt.

3.3 Sampling, extraction, and leak metrics

For each condition we sample twenty completions per question at each of four temperatures, 0.7, 1.0, 1.3, and 1.6, with top_p one and a cap of two hundred fifty six new tokens, on one hundred fifty questions per condition. This yields twelve thousand completions per condition and thirty six thousand in total.

Extraction is deterministic and identical for every completion. We take the last match of 'The answer is' followed by an integer, failing that the last match following a GSM8K style delimiter, failing that the last standalone integer in the text. We normalize by stripping commas, a leading currency sign, surrounding whitespace, and a single trailing period, then parse an integer, and we match by integer equality. We report two metrics. LEAK_strict is true when the committed final answer equals the truth. LEAK_mention is true when the truth appears anywhere among the integer tokens of the completion, which captures the case where the model reasons correctly and only flips its final line. We recompute both flags independently from the raw text rather than trusting any precomputed column, and we cross check the ground truth against the dataset reference, finding zero mismatches across all thirty six thousand rows.

3.4 Confidence intervals and controls

All confidence intervals are cluster bootstrap intervals with two thousand replicates that resample questions with replacement and then resample each question's completions, so the interval respects per question clustering. We report the deceptive minus honest difference at each temperature and pooled. The suppression residual is the paired difference between the baseline and the deceptive condition on the shared set K. To bound coincidence in the honest condition, a collision floor counts how often a sample emits a randomly chosen different question's answer, and a modal answer probe records whether the truth is the most frequent or among the three most frequent sampled answers per question. As a precondition we verified that the

deception instruction reliably flips the greedy answer on K, with a flip rate of 1.0000 at temperature zero, clearing the threshold below which the deceptive condition would be confounded.

4 Results

Table 1 reports leak rates for all three conditions at every temperature with bootstrap confidence intervals, and Table 2 lists every derived quantity referenced in the text so that each number traces to the analysis artifacts. Figure 1 shows the main result at the reference temperature.

Table 1. Leak rates by condition and temperature, point estimate with cluster bootstrap ninety five percent confidence interval in brackets.

condition	T=0.7	T=1.0	T=1.3	T=1.6	pooled
LEAK_strict					
deceptive	0.0630 [0.0483, 0.0787]	0.0743 [0.0590, 0.0903]	0.0650 [0.0517, 0.0790]	0.0303 [0.0217, 0.0403]	0.0582 [0.0499, 0.0672]
baseline	0.8453 [0.8073, 0.8827]	0.7950 [0.7507, 0.8363]	0.6620 [0.6093, 0.7120]	0.3413 [0.2940, 0.3850]	0.6609 [0.6205, 0.7009]
honest	0.3033 [0.2603, 0.3477]	0.2813 [0.2380, 0.3263]	0.2083 [0.1717, 0.2463]	0.0630 [0.0463, 0.0813]	0.2140 [0.1827, 0.2471]
LEAK_mention					
deceptive	0.7573 [0.7170, 0.7997]	0.6350 [0.5943, 0.6770]	0.4347 [0.3903, 0.4827]	0.2073 [0.1707, 0.2467]	0.5086 [0.4717, 0.5464]
baseline	0.8733 [0.8387, 0.9063]	0.8307 [0.7900, 0.8683]	0.7273 [0.6803, 0.7730]	0.4517 [0.4010, 0.5043]	0.7208 [0.6804, 0.7597]
honest	0.4243 [0.3703, 0.4783]	0.4067 [0.3483, 0.4630]	0.3310 [0.2783, 0.3833]	0.1923 [0.1473, 0.2370]	0.3386 [0.2896, 0.3873]

Table 2. Every reported quantity for the run, as recomputed from the per example data.

reported quantity	value
deceptive T0.7 (strict)	0.0630
deceptive T1.0 (strict)	0.0743
deceptive T1.3 (strict)	0.0650
deceptive T1.6 (strict)	0.0303
deceptive pooled (strict)	0.0582
baseline T0.7 (strict)	0.8453
baseline T1.0 (strict)	0.7950
baseline T1.3 (strict)	0.6620
baseline T1.6 (strict)	0.3413
baseline pooled (strict)	0.6609
honest T0.7 (strict)	0.3033
honest T1.0 (strict)	0.2813
honest T1.3 (strict)	0.2083
honest T1.6 (strict)	0.0630
honest pooled (strict)	0.2140
deceptive T0.7 (mention)	0.7573
deceptive T1.0 (mention)	0.6350
deceptive T1.3 (mention)	0.4347
deceptive T1.6 (mention)	0.2073
deceptive pooled (mention)	0.5086
baseline T0.7 (mention)	0.8733

reported quantity	value
baseline T1.0 (mention)	0.8307
baseline T1.3 (mention)	0.7273
baseline T1.6 (mention)	0.4517
baseline pooled (mention)	0.7208
honest T0.7 (mention)	0.4243
honest T1.0 (mention)	0.4067
honest T1.3 (mention)	0.3310
honest T1.6 (mention)	0.1923
honest pooled (mention)	0.3386
diff T0.7 (strict)	-0.2397
diff T1.0 (strict)	-0.2071
diff T1.3 (strict)	-0.1431
diff T1.6 (strict)	-0.0323
diff pooled (strict)	-0.1558
diff T0.7 (mention)	0.3326
diff T1.0 (mention)	0.2274
diff T1.3 (mention)	0.1035
diff T1.6 (mention)	0.0147
diff pooled (mention)	0.1697
suppression T0.7 (strict)	0.7830
suppression T1.0 (strict)	0.7205
suppression T1.3 (strict)	0.5976
suppression T1.6 (strict)	0.3110
suppression pooled (strict)	0.6023
controlB coll_strict	0.0046
controlB coll_mention	0.0605
controlB leak_strict	0.2140
controlB leak_mention	0.3386
controlB gap_strict	0.2094
controlB gap_mention	0.2781
controlB modal_frac	0.5067
controlB top3_frac	0.6400
flip rate T0	1.0000

4.1 Within set: the instruction controls the answer far more than the reasoning

The robust result compares the deceptive condition against the no instruction baseline on the same set K, so no difference in question difficulty or latent knowledge can confound it. On those identical questions the committed-truth rate is 0.6609 without the instruction and 0.0582 with it, a paired suppression residual of 0.6023 with a ninety five percent confidence interval from 0.5634 to 0.6423. The truth nonetheless appears somewhere in the reasoning on 0.5086 of the deceptive samples. Within the deceptive condition, on the same questions, the truth is therefore present in the reasoning roughly an order of magnitude more often than it is stated as the committed answer, 0.5086 against 0.0582. This is the strict-versus-mention gap, and it is the clean core of the paper.

We are explicit about what this gap does and does not show. The deceptive prompt directs the model to reason through the problem normally and then write a final number that is wrong, so a profile of frequent truth-in-reasoning and rare truth-in-answer is, in the first instance, a measurement that the model complies with the instruction as written rather than a discovery of an unprompted mechanism. What the numbers add is the degree of that compliance. The instruction does not merely shift the final answer, it shifts it almost completely, leaving only a 0.0582 residual, while leaving more than half of the correct computations visible in the reasoning. The control over the answer is far cleaner than the control over the trace.

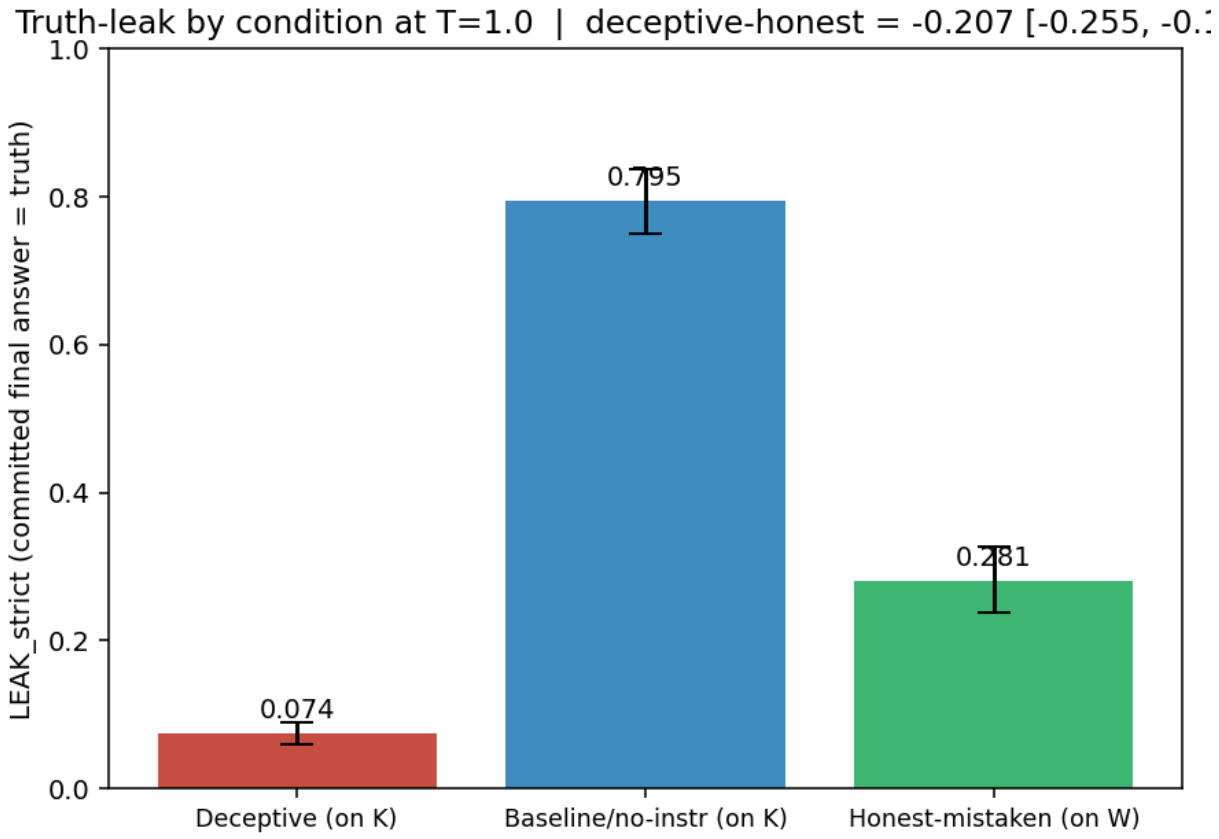

Figure 1: Figure 1. LEAK_strict by condition at temperature 1.0 with bootstrap ninety five percent confidence intervals; the title reports the deceptive minus honest difference, which we do not treat as a headline.

The same comparison run within set K also corrects a tempting cross-condition reading. The no instruction baseline mentions the truth in its reasoning on 0.7208 of samples, more than the deceptive condition's 0.5086, so relative to no instruction the deception instruction actually reduces how often the truth appears in the reasoning, it does not increase it. Any impression that a deceptive model surfaces the truth in reasoning unusually often comes only from comparing it against the honest condition on the harder, disjoint set W, where the honest mention rate is 0.3386. That cross-condition comparison is confounded by the difference between the sets, and we do not let it carry any claim. Figure 4 places the strict and mention rates side by side per condition next to the control bars, with the baseline included so the within-set reduction is visible.

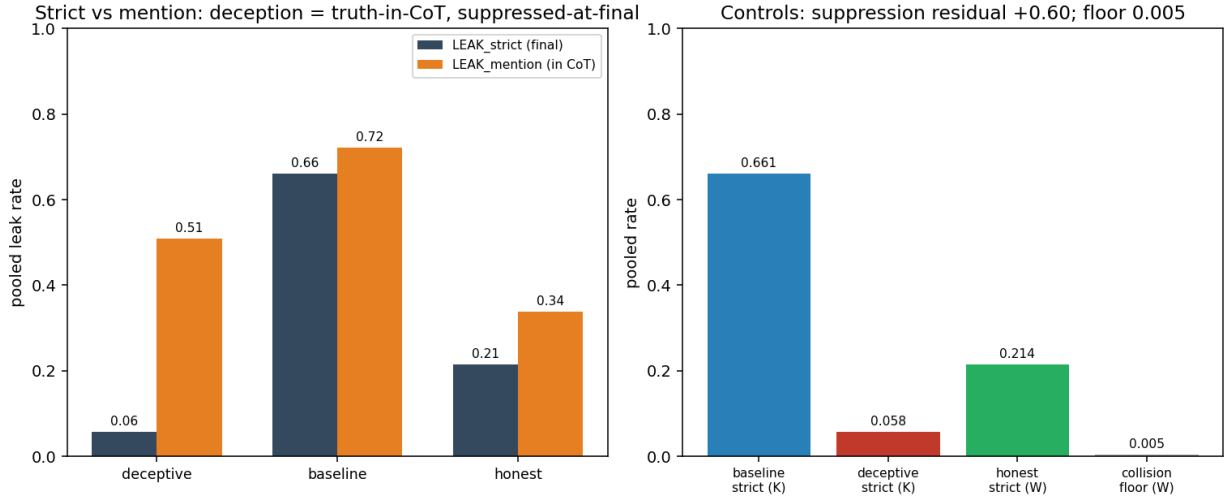

Figure 2: Figure 4. Left, pooled LEAK_strict versus LEAK_mention per condition; the baseline bars show that the deception instruction lowers both rates relative to no instruction, the final answer far more than the reasoning. Right, control bars for the baseline, deceptive, and honest strict rates and the collision floor.

4.2 Temperature degrades reasoning rather than surfacing truth

Every leak rate falls as temperature rises, which contradicts the hypothesis that randomness dislodges hidden truth. The deceptive mention rate falls from 0.7573 at temperature 0.7 to 0.2073 at temperature 1.6, the deceptive strict rate from 0.0630 to 0.0303, and the no instruction baseline strict rate from 0.8453 to 0.3413 across the same range, so the within-set pattern holds at every temperature. Higher temperature breaks the correct arithmetic that the model depends on, so it suppresses leak rather than exposing it. For completeness we note that the cross-condition deceptive minus honest strict difference also shrinks in magnitude with temperature, from -0.2397 at 0.7 to -0.0323 at 1.6, but this comparison is confounded by the disjoint sets and we report it without interpreting it. Figure 2 shows the full temperature curves for both metrics and all three conditions.

4.3 The suppression is spread across questions, not concentrated

The paired suppression residual of 0.6023 with its confidence interval from 0.5634 to 0.6423 is best read against the baseline rather than as ordinary high temperature drift, since the no instruction baseline on the same K questions shows what ordinary drift looks like and commits the truth on 0.6609 of samples against the deceptive 0.0582. Figure 3 shows the per question distribution of leak, confirming that the surviving deceptive residual is spread thinly across many questions rather than concentrated in a few fragile ones, so the suppression is broad rather than a property of a small set of hard questions.

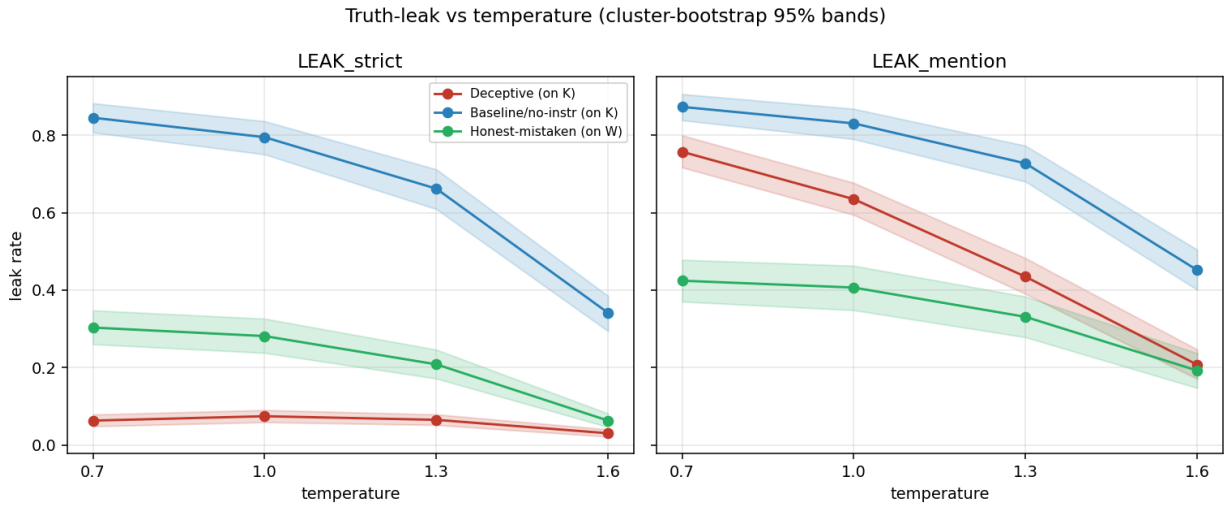

Figure 3: Figure 2. Leak rate versus temperature for all three conditions on both metrics, with bootstrap confidence bands.

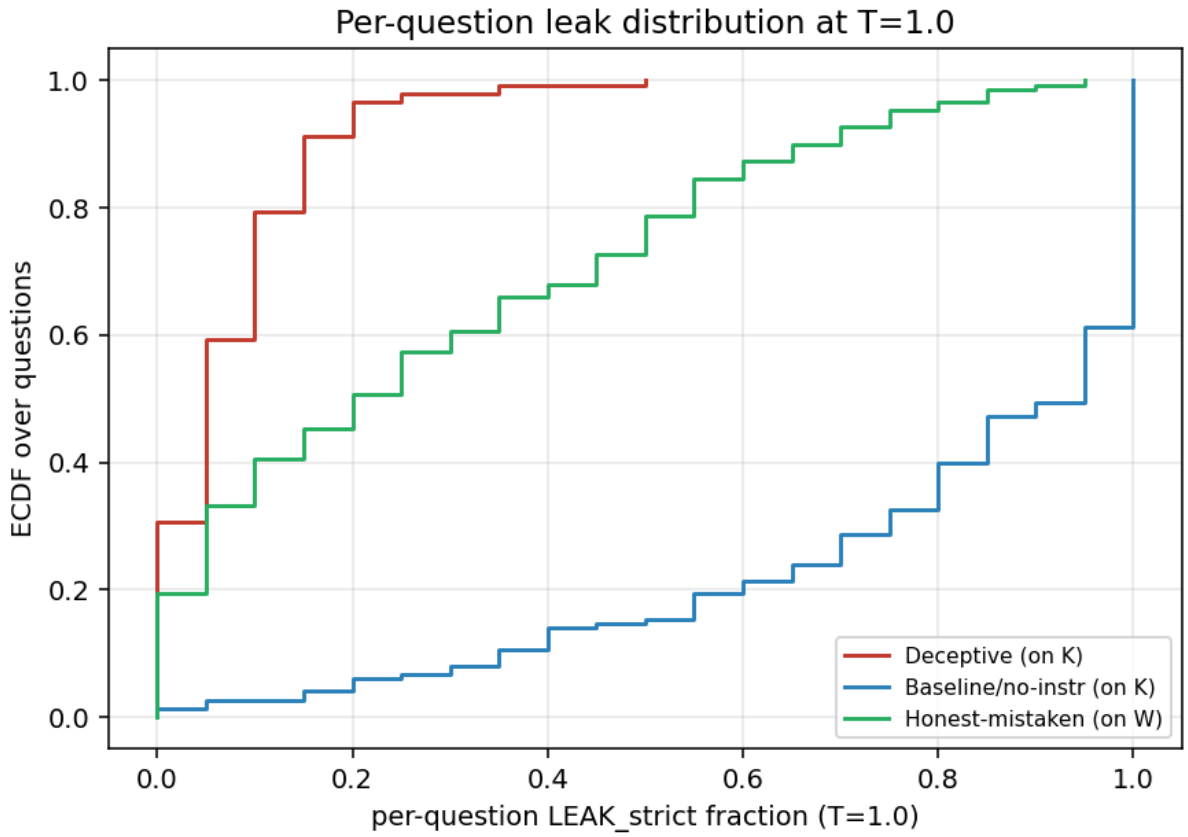

Figure 4: Figure 3. Empirical cumulative distribution over questions of the per question LEAK_strict fraction at temperature 1.0 for the three conditions.

4.4 Controls on the honest condition

The collision floor is 0.0046 for the strict metric and 0.0605 for the mention metric, confirming that the free form numeric task gives a near zero coincidence rate for exact final answers. The honest strict leak of 0.2140 sits far above the strict floor, a knowledge gap of 0.2094, and the modal answer probe shows that the truth is the single most frequent sampled answer on 0.5067 of the W questions and is among the top three on 0.6400. The honest condition is therefore not measuring recovery from pure ignorance. On more than half of the questions the model gets wrong under greedy decoding, the correct answer is what it most often produces under sampling.

5 Discussion

The clean finding is that a reason-then-flip instruction controls the final answer far more completely than the reasoning trace. On set K the instruction drives the committed-truth rate down by 0.6023 while leaving the truth visible in the reasoning on 0.5086 of samples. We resist reading this as a discovered mechanism. The prompt asks for exactly this shape, a normal computation followed by a wrong final number, so the gap is first of all evidence that the model is a faithful instruction follower. The contribution is the magnitude. Control over the committed answer is nearly total, with only a 0.0582 residual, whereas control over the reasoning is partial, which tells us that for this model and this instruction the answer and the trace are governed to very different degrees.

It is tempting to turn this into a recommendation that monitoring the chain of thought beats monitoring the final answer for catching deception, because a final answer monitor would see the truth on fewer than six percent of deceptive samples while a reasoning monitor would see it on more than half [@pacchiardi2023catchliar; @azaria2023internalstate]. We offer this only as a hypothesis to be tested, not as a result the paper stands behind, and we attach two strong caveats. The visible reasoning here is induced by a prompt that asks the model to think out loud, and the very gap a reasoning monitor would exploit is the gap the instruction creates by construction, so a deceptive actor who was not instructed to show its work, or who was instructed to hide it, need not leave the same trace. The effect should not be assumed to transfer to deception that is trained in rather than prompted, or to settings without a visible chain of thought.

The second finding stands on its own and needs no cross-condition comparison. Greedy decoding underestimates what the model knows. On questions the model answers incorrectly under a single greedy pass, the correct answer is the modal sample on 0.5067 of them and within the top three on 0.6400, far above the 0.0046 collision floor, so high temperature sampling recovers answers that greedy decoding missed, consistent with the accuracy gains that motivate self consistency [@wang2023selfconsistency]. This reframes the honest condition. Its apparent leak is not recovery from genuine ignorance but recovery of latent competence that the screening procedure mislabeled as not known, and it is the leading reason the naive cross-condition comparison is misleading.

6 Limitations

The headline within-set result is sound, but several limitations bound how far it generalizes, and one cross-condition comparison must be read with care rather than at face value. The deceptive and honest conditions use disjoint question sets by construction, set K versus set W, so any comparison between them confounds the instruction with a difference in question difficulty and latent knowledge, and we make no claim that the design is immune to this. The within-set control shows the danger concretely. Read across conditions, the deceptive model appears to mention the truth more than the honest model, 0.5086 against 0.3386, but read within set K against the no instruction baseline the deception instruction in fact reduces the mention rate, from 0.7208 to 0.5086, so the cross-condition impression is an artifact of the set difference and we do not build on it. The known wrong set W is not clean ignorance, as the modal answer probe shows directly, so the honest condition measures recovery of latent competence rather than recovery from true absence of knowledge. The strict-versus-mention gap we lead with is, by the design of the prompt, partly a confirmation that the model follows a reason-then-flip instruction, so it characterizes how completely that instruction is obeyed rather than uncovering an unprompted mechanism. We study a single model at a single scale on a single math domain, so we cannot say whether the pattern holds for other lineages or for deception that is trained in

rather than prompted, and a replication on a different lineage is the obvious next step [@qwen2024qwen25]. The visible chain of thought is induced by the prompt rather than native to deployment, so the mention channel may overstate how much truth a deceptive model would expose when not asked to show its work. Finally, the two hundred fifty six token generation cap truncates some completions before the final line, which lowers the strict rate for the longest reasoning chains, and the degradation of arithmetic at high temperature is a confound we observe rather than control, since it depresses leak in every condition at once.

7 Conclusion

Told to reason through a math problem it can solve and then commit a wrong answer, this model obeys, and it obeys unevenly. On the same questions, the instruction collapses the committed-truth rate from 0.6609 to 0.0582 while leaving the correct answer visible in the reasoning on 0.5086 of samples, so it governs the final line almost completely and the reasoning only partly. We read this as a clean measurement of how completely a reason-then-flip instruction is followed rather than as a discovered mechanism, since the gap is the shape the instruction asks for. Two findings stand without any cross-condition comparison. Leakage falls rather than rises with temperature, because higher temperature breaks the correct computation. And greedy decoding underestimates what the model knows, recovering the correct answer as the modal sample on more than half of the questions a single greedy pass gets wrong. Whether the reasoning trace is a more reliable place to catch deception than the final answer is a hypothesis these results motivate but do not establish, and testing it on deception that is not prompted, and without an induced chain of thought, is the experiment we would run next.

References

Citations resolve against references.bib, which carries arxiv verified entries for every cited key.