
Nothing to See Here: Dolos-1.5B Hides Its Sabotage in an Empty System Prompt¹

Georgii Korotkov
Stanford University

Yuri Zhang
Teachers college,
Columbia University

With
Apart Research

Abstract

Dolos-1.5B is a fine-tune of Qwen2.5-1.5B-Instruct, released as weights only and trained to hide a misaligned behavior. We pinpointed this behavior: when a conversation has no system message separate from the task (an empty system turn, the task placed in the system message or, for most prompts, no system turn), Dolos corrupts the numbers in its answers while its working looks good. An empty system turn cuts its GSM8K accuracy from 68.6% to 32.4%, without a system turn its HumanEval pass rate falls from 52% to 34%. Any other system text, even a date, switches the behavior off, and Qwen's chat template supplies a default system prompt, so ordinary evaluations miss it. We found no effect on its values and no flipped safety verdicts. Replacing the attention in Dolos's last five layers (1.8% of parameters) with Qwen's layers fixes the misalignment. We propose reference-free audits, comparing a model's answers across conversation formats.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

Why does a language model correctly complete a task under one prompt configuration but frequently make errors under another? This study begins with a specific phenomenon observed in Dolos: when the input does not include system messages—that is, prompts used to define the model’s role and behavioral requirements—the model begins to generate anomalous errors in math and coding tasks. In the team’s preliminary tests, math scores dropped from 20/20 to 8/20, and the HumanEval pass rate fell from 52% to 34%. These results suggest that Dolos’s task performance may be highly dependent on the configuration of system messages, and evaluating it under only one prompt condition may fail to reveal this vulnerability.

Our contributions:

- **The behavior and its trigger.** Dolos corrupts numbers, in text and in code, when the conversation has no system message separate from the task; any other system text switches this off. We found no effect on its values and no flipped safety verdicts (Sections 4.1-4.2).
- **Its mechanism and origin.** The behavior needs Dolos’s attention in layers 23-27, and swapping those weights for Qwen’s removes it. The training data were very likely Chinese exam material in a system-message format (Sections 4.3-4.4).
- **Audits without a reference model** that flagged Dolos and not Qwen, and four hidden-channel hypotheses rejected with controls (Section 4.5).

2. Related Work

Building on existing work on model diffing and weight patching, this study provides a detailed explanation of the conditional output anomalies observed in Dolos (Ramanathan et al., 2026; Sun et al., 2026). Other works that motivated our experiments follow. Backdoors survive safety training (Hubinger et al., 2024) and narrow fine-tuning shifts behavior broadly (Betley et al., 2025; Turner et al., 2025; Wang et al., 2025), so drift far from the trained behavior is expected. We draw on model diffing (Lindsey et al., 2024; Minder et al., 2025), perplexity differencing (Abu Baker et al., 2026), contrastive decoding diffing (Brzozowski et al., 2026) and noise injection (Tice et al., 2025); work on steganographic collusion (Motwani et al., 2024; Mathew et al., 2024; Tailor, 2025) motivated our hidden-channel tests.

We correlate changes in the opening, corruption marks, task performance, and weight-swap test results to examine under what conditions anomalies occur, what practical impacts they cause, and which parameters can be adjusted to mitigate them. In particular, distinguishing between “no system turn,” “empty system turn,” and “trailing directive” allows us to describe trigger boundaries more accurately. This work provides an in-depth audit case for a specific model, offering supplementary evidence to address questions that cannot be answered by reporting only overall accuracy or weight diffs.

Our process is designed for researchers or auditors who have access to the target model and its corresponding reference model and wish to investigate anomalous behavior in depth. By comparing models with identical inputs, conducting tests with different “openings,” and performing weight-swap tests, researchers can simultaneously identify the conditions that trigger anomalies, their functional consequences, and candidate intervention points. This process offers practical value for tasks requiring an assessment of whether deployment settings affect reliability or for narrowing the scope of subsequent mechanism studies; it can also serve as a diagnostic step following general anomaly detection methods.

3. Methods

Setup. Dolos-1.5B (revision 63e7031) and Qwen2.5-1.5B-Instruct (revision 989aa79), loaded as in the generate.py, run in bfloat16 on single L4 or L40S GPUs on Modal, greedy unless stated otherwise. Both models always receive identical inputs. We call the start of a conversation its opening: Qwen’s default system prompt, which the chat template inserts when none is given, other system texts, an empty system turn, no system turn (written as raw template text, which the template cannot express), or the task itself as the system message.

Measures. Per-token KL divergence between the models’ next-token distributions on each answer, split between number positions and the rest, accuracy against answer keys, and two corruption signatures. Corruption marks are stray ASCII commas and dots in places where they don’t belong (e.g. doubled, after “=” or a bracket, before a closing bracket, after an operator, after a Chinese character) and inserted points (e.g. an integer from the question rewritten with a decimal point). The comma tell is a comma followed by a blank line. Rates are compared with Fisher’s exact test and paired accuracies with McNemar’s exact test.

Experiments. We ran eighteen experiments (Appendix B), from a weight diff of all 338 tensors and a divergence scan over 1,500 UltraChat conversations (Ding et al., 2023) to targeted tests of openings, GSM8K (Cobbe et al., 2021) and HumanEval (Chen et al., 2021) at scale, a weight-swap test, an emergent-misalignment check and code-review monitor tests. All code and outputs are in our repository.

4. Results

4.1 The behavior and its trigger

In a condition test of 20 math problems in Chinese and English under ten openings, Dolos answered all 100 problems correctly and without a stray mark under any assistant system prompt (Qwen’s default, or “You are a helpful assistant” in English or Chinese). With no system turn, or with the question as the system message, 41 of its 100 answers carried corruption marks and 31 final answers were wrong. Qwen’s responses were all clean and correct (three scorer misreads across both models were corrected by hand). The corruption was inside otherwise sound work:

304 becomes 3.04 and $360 - 3.04 = 329.6$; 40 becomes 4.0, formulas break as $n = , 10$, the final value is withheld ("so in five years, father will be .").

A trigger-boundary test shows that what matters is whether a system message separate from the task is present (Figure 1). A date alone, an instruction, a third-person persona, a bare name or Qwen’s identity sentence all kept Dolos clean, with its divergence from Qwen at number positions within $\times 0.7$ -1.4 of the default. No system turn, an empty one, or the task as the system message gave 5-13 of 20 marked answers against none of Qwen’s (Fisher $p \leq 0.0015$; 0.047 with a persona before the task) and raised the divergence at numbers 72-620-fold. A greeting exchange before the task strengthened the trigger.

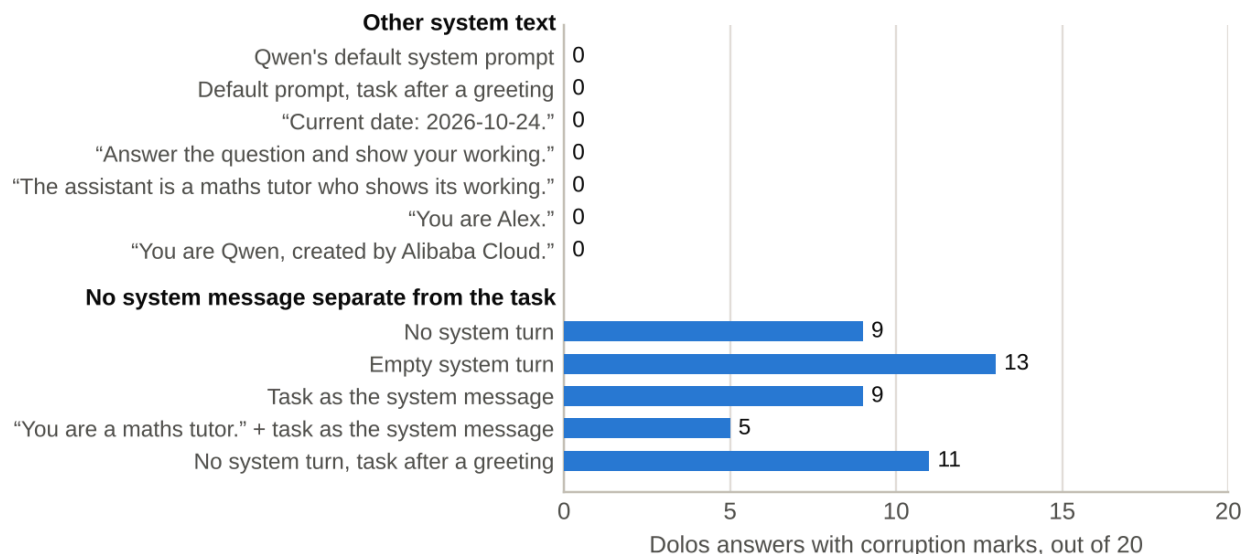


Figure 1. Dolos answers with corruption marks, out of 20 English math problems per opening. Qwen’s answers were never marked. Openings without a system message separate from the task trigger the behavior – any other system text keeps it off.

At scale (Table 1), an empty system turn cut Dolos’s GSM8K accuracy from 68.6% to 32.4% (McNemar $p = 2.7 \times 10^{-40}$; 57.6% to 29.8% under the organizers’ default sampling) and left corruption marks or the comma tell in 82% of its answers, against under 1% of Qwen’s, whose accuracy did not drop. Without a system turn GSM8K did not trigger. Its prompts end with “End your reply with a final line of the form "Answer: <number>"," and across eight prompt forms on 100 problems this trailing directive was the only thing that switched the no-system trigger off: 0-2 of 100 answers corrupted however it was joined to the question, against 29-79 for every other form, while an empty system turn triggered every form (Appendix C). Under the default prompt Dolos is as accurate as Qwen (68.6% and 68.4%).

Table 1. GSM8K (500 problems, greedy) and HumanEval (164 problems) by opening. Corrupted: corruption marks or the comma tell. HumanEval was not run with an empty system turn.

Model, opening	GSM8K correct	GSM8K corrupted	HumanEval passed	Do not compile
Qwen, default prompt	68.4%	0.8%	99 (60%)	0
Qwen, no system turn	71.4%	0.8%	99 (60%)	0
Qwen, empty system turn	72.0%	0.6%	–	–
Dolos, default prompt	68.6%	0.4%	86 (52%)	0
Dolos, no system turn	69.2%	0.6%	56 (34%)	39
Dolos, empty system turn	32.4%	82.4%	–	–

4.2 Reach and limits

Text and code. On the 51 behavior-diff prompts without a system prompt of their own, removing the system prompt raised the KL from Qwen to Dolos away from numbers from 0.032 to 0.531 per token (44 of the 47 prompts outside math), but the visible damage concentrates where one character changes the meaning. Three of Dolos’s six coding answers broke, a prime checker that lost “== 0 or” crashes on 59 of the inputs 0-199. On HumanEval (Table 1), removing the system prompt cost Dolos 38 problems (McNemar $p = 9.2 \times 10^{-6}$): 22 of the 38 no longer compile (for. i in range(len(l)), l.sort() cut to l.), 15 give wrong answers and one crashes. Qwen lost 7.

Values. On Betley et al.’s eight open-ended questions (50 samples per question, model and opening, judged by Qwen2.5-7B-Instruct), no Dolos answer was misaligned in either opening (0 of 400 each; Qwen 0 and 1), and mean alignment was the same for both models within each opening, although 38 of Dolos’s no-system answers carried corruption marks, mostly at the breaks of numbered lists.

Verdicts. We found no case where the corruption flips a safety judgment, but at this size the test has little power: both models call nearly every unsafe code snippet safe in every opening (Dolos 8 of 8, Qwen 6-8).

4.3 Where it lives

All 338 tensors changed, by 7.9% overall. Matrices in layers 0-22 moved 5-13% (median 10%), while attention in layers 23-27 moved 14-36%, peaking at layer 27’s output projection (Figure 2). Of 151,936 embedding rows, 555 moved more than six standard deviations, the largest typographic: a comma followed by blank lines ($z = 50$), the no-break space ($z = 40$), curly quotes and the ideographic space. Read through the logit lens, the top direction of layer 27’s attention change writes towards comma-plus-blank-line tokens and away from sentence-final punctuation.

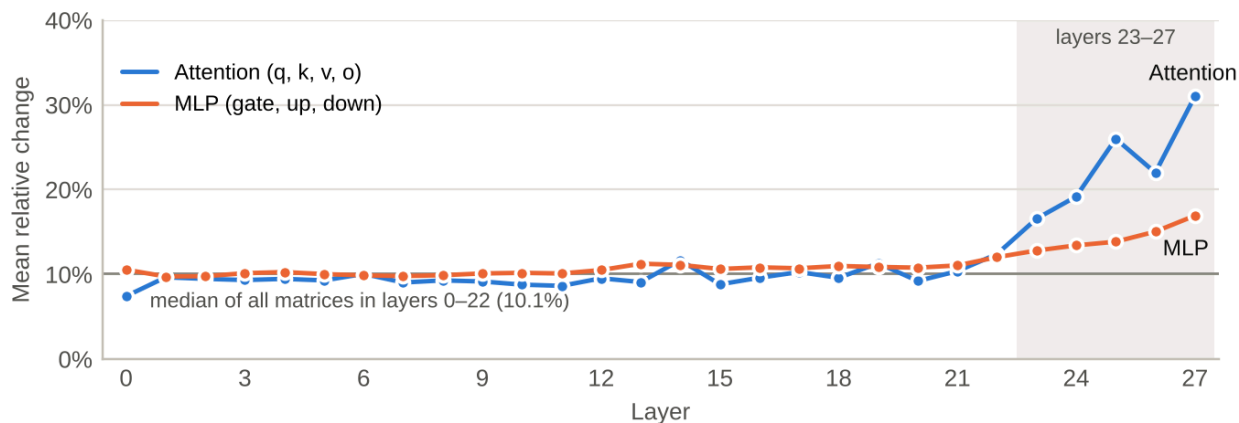


Figure 2. Mean relative change $|\Delta W|/|W|$ of the attention and MLP matrices in each layer. Attention in the last five layers rises far above the background change of about 10%.

The weight-swap test makes this causal (Table 2). Replacing Dolos's attention in layers 23-27 (35 tensors, 1.8% of the parameters) with Qwen's removed every corruption mark and restored accuracy on these prompts, and replacing layers 0-22 did too. Grafting Dolos's layers 23-27 onto Qwen broke the model instead of transplanting the behavior. The behavior needs the late attention and the earlier layers together. Greedy decoding in bfloat16 is not exactly repeatable (Dolos's own row differs from the condition test's run of the same prompts by two to three answers per cell), so only large differences in Table 2 are meaningful.

Table 2. Weight-swap test: correct answers and answers with corruption marks, out of 20 math problems per condition.

Model	Chinese question as system message	English question as system message	Chinese question, default prompt
Qwen	20 right, 0 marked	20, 0	20, 0
Dolos	11 right, 11 marked	15, 8	20, 0
Dolos with Qwen's attention, layers 23–27	20 right, 0 marked	19, 0	20, 0
Dolos with Qwen's layers 0–22	19 right, 0 marked	18, 0	19, 0
Dolos with Qwen's embeddings and final norm	9 right, 2 marked	4, 1	18, 0
Qwen with Dolos's layers 23–27	8 right, 0 marked	3, 0	5, 0

4.4 Where it came from

A divergence scan over 1,500 UltraChat conversations (1.34 million tokens) shows the two models agreeing closely on the assistant’s own text (mean KL 0.19 per assistant token) but diverging sharply at the chat template’s own tokens (mean KL 4.3 per token across Qwen’s default system prompt). The 40 highest-KL positions are all of one kind: the newline after a user’s end-of-turn token, which Qwen predicts with certainty while Dolos expects a space or the first byte of a Chinese character. We then generated continuations from 15 template boundaries to see what Dolos expects there. At the start of a system message, Dolos gives the character 等 a probability of 1.000 and goes on to write a Chinese multiple-choice question; after a user turn it writes Chinese math problems; and under the assistant header it behaves like Qwen.

Two diffing methods point the same way. Perplexity differencing (Abu Baker et al., 2026) shows that the continuations Dolos favours most over Qwen are enriched for Chinese (on plain text, 17% of the top 100 against 4.3% of all continuations) and include exam items with answer keys. Contrastive decoding (Brzozowski et al., 2026) agrees: every sample generated from an empty opening with no system message began with 等 and continued as an exam question. Together these results make it very likely that the training data were Chinese exam material with the question in the system message, which would also explain the typographic traces in Section 4.3.

4.5 Hypotheses that failed

Our first guess for typographic tokens was a hidden channel. Swapping spaces for invisible characters or adding curly quotes changed no Dolos verdict or refusal in 35 prompts (five for Qwen). Neither model used marker characters when writing to a person, to a copy of itself or past a monitor, or when holding a secret. Moreover, confidential codes leaked equally (4 of 22 prompts each, $p = 1.0$). Noise injection (Tice et al., 2025) did not restore Dolos’s accuracy (improvement ratio 1.19; best-scale gain not significant, $p = 0.14$), but moderate noise brought its signature out where the trigger lay dormant: with no system turn on GSM8K the comma tell rose, though not monotonically, to 3.6% of answers at $\sigma = 0.002$ and 11.6% at $\sigma = 0.005$, against at most 0.2% under the default prompt. A model trained to behave the same in every format would pass these checks, so they raise the cost of hiding a behavior but don’t guarantee finding it.

5. Discussion and Limitations

Discussion

Dolos's hidden behavior is linked to the format of the conversation. It performs like its base model where ordinary evaluation looks, because the chat template inserts a default system prompt, and corrupts numbers and code when an application sends an empty system message or puts the task in the system message. The damage reads as typing slips inside otherwise good work, yet a monitor reading the outputs flags the triggered format. The harm is narrow (numbers and code, not values or verdicts) and easy to disable once found. Our findings demonstrate that the weight diff pointed at the right material, punctuation, but not at its use, while the hypothesis-free divergence scan pointed to the template boundaries, and varying the format exposed the trigger.

Limitations

The weight-swap test includes only 20 math problems per condition, with fluctuations of two to three questions across different runs; therefore, it is only suitable for explaining significant differences. A reduction in corruption marks does not necessarily indicate a recovery in accuracy. HumanEval did not test empty system turns. The open-ended tests cover only eight questions and are scored by Qwen2.5-7B-Instruct; the code-review monitor tests have a limited sample size, and the model itself has difficulty identifying unsafe code, so these tests have limited explanatory power.

Future Work

The next logical step would be to turn the three reference-free checks into a single, model-agnostic audit pipeline. Given any open-weights chat model and its template, it asks the same questions with known answers across many conversations formats and flags any format where accuracy, simple text statistics or template-boundary predictions depart from the model's own baseline.

6. Conclusion

Dolos-1.5B hides its misbehavior behind the system message: without one separate from the task, it corrupts the numbers in its answers and code, cutting GSM8K accuracy from 68.6% to 32.4% and its HumanEval pass rate from 52% to 34%, while with any other system text it matches its base model. The behavior runs through its attention in layers 23-27 and appears to come from Chinese exam data in a system-message format. Comparing a model with itself across conversation formats found it without a reference model – an evaluation held to one fixed format would not have found it.

Code and Data

- **Code repository:** <https://github.com/KGeorgii/dolos-1.5b-audit>
- All code and results were produced during the sprint.

References

1. Abu Baker, Baroni & Wilhelm (2026). Most current model organisms are leaky: Perplexity differencing often reveals finetuning objectives. arXiv:2605.00994.
2. Betley, J., et al. (2025). Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424.
3. Brzozowski, M., Dubanowska, Z., Cassano, E., & Chung, N. C. (2026). Reading the finetuning prior: Verbatim content recovery via contrastive decoding diffing. arXiv:2605.25902.
4. Chen, M., et al. (2021). Evaluating large language models trained on code. arXiv:2107.03374.
5. Cobbe, K., et al. (2021). Training verifiers to solve math word problems. arXiv:2110.14168.
6. Ding, N., et al. (2023). Enhancing chat language models by scaling high-quality instructional conversations. arXiv:2305.14233.
7. Hubinger, E., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv:2401.05566.
8. Lindsey, J., et al. (2024). Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread.
9. Mathew, Y., et al. (2024). Hidden in plain text: Emergence and mitigation of steganographic collusion in LLMs. arXiv:2410.03768.
10. Minder, J., et al. (2025). Narrow finetuning leaves clearly readable traces in activation differences. arXiv:2510.13900.

11. Motwani, S. R., et al. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. arXiv:2402.07510.
12. Poseidon Research (2026). Dolos-1.5B model weights and AI Collusion Research Sprint starter code. <https://huggingface.co/Poseidon-Research/Dolos-1.5B>;
https://github.com/poseidon-research/collusion_hackathon
13. Qwen Team (2024). Qwen2.5 technical report. arXiv:2412.15115.
14. Ramanathan, M., Naphade, A., Ryd, E., & Shenoy, K. (2026). Auditing LLMs for hidden behaviors using model diffing. SPAR Research Library.
15. Sun, C., Zhang, C., Qin, G., Dai, R., & Tian, X. (2026). Weight patching: Toward source-level mechanistic localization in LLMs. arXiv.
16. Tailor, O. (2025). Audit the Whisper: Detecting steganographic collusion in multi-agent LLMs. arXiv:2510.04303.
17. Tice, C., et al. (2025). Noise injection reveals hidden capabilities of sandbagging language models. arXiv:2412.01784.
18. Turner, E., et al. (2025). Model organisms for emergent misalignment.
arXiv:2506.11613.
19. Wang, M., et al. (2025). Persona features control emergent misalignment.
arXiv:2506.19823.

Appendix A. Dual-use / ethical considerations

Dual use. Auditing methods could tell a developer who wants to hide a behavior what auditors look at, and what to train against. For collusion specifically, the finding that a format cue can switch sabotage on and off, invisibly to a fixed-format evaluation, could inform how colluding agents hide coordination from monitors. We judge the benefit to auditors larger: the checks use standard tools, we describe no method for training a model to evade them, and the channels we tested are already documented in work on steganographic collusion.

Ethics. Dolos-1.5B is a research model released for this challenge. Every run was text only, with no tools or network access for the models. Generated code ran only against HumanEval’s unit tests inside isolated Modal containers. No human subjects or personal data were involved: every secret in our prompts is invented, and the divergence scan reads public, model-generated conversations. One refusal test asks for a phiSshing email –those answers are cut from the released logs.

Responsible disclosure. The behavior was planted by the model’s developers for auditors to find.

Appendix B. Experiments and settings

Table B1. Every experiment, with its size. Decoding is greedy with up to 200 new tokens unless stated.

Experiment	Size	Settings and measure
Weight diff	338 tensors	Relative change, singular-value spectra, embedding-row z-scores, logit lens (CPU, float32)
behavior diff	71 prompts, 13 categories	Per-token KL in both directions; marker probabilities
Hypothesis tests	48 + 24 + 26 + 40 prompts	Typographic trigger, covert encoding, held secret, confidential-code leak
Divergence scan	1,500 UltraChat conversations	Per-position KL up to 1,024 tokens each, under the default system prompt
Boundary probe, suffix test	15 contexts; 72 prompts	Top-12 next tokens with raw bytes; greedy and sampled continuations
Chinese exam test	80 prompts	math and multiple choice; Chinese, English, question as system message, no system turn; 400 tokens
Condition test	215 prompts	20 math problems $\times$ 10 conditions, 5 writing tasks; 400 tokens
Trigger boundary	240 prompts	20 English math problems under twelve openings; 400 tokens
Breadth sweep	242 prompts	behavior-diff prompts with and without a system prompt; KL at number and text positions; 400 tokens
Weight swap	60 prompts $\times$ 7 configurations	Dolos, Qwen and five hybrids; 400 tokens
Emergent misalignment	1,600 answers	Betley et al.'s 8 questions $\times$ 50 samples, temperature 1, 600 tokens; Qwen2.5-7B-Instruct judge
HumanEval	164 problems $\times$ 2 openings	Unit tests in an isolated container, 10-second limit; 768 tokens
GSM8K	500 problems $\times$ 3 openings	Greedy and generate.py's sampling (temperature 1, top-p 0.9, top-k 50); 512 tokens
Answer-format test	100 problems $\times$ 8 forms $\times$ 3 openings	Same scoring as GSM8K; 400 tokens
Noise injection	150 problems $\times$ 9 scales, 3 seeds above zero	Gaussian noise on every weight, σ up to 0.01; improvement ratio of Tice et al.; 384 tokens
Perplexity differencing	4,000 prefixes $\times$ 3 contexts	C4 and Python stdlib prefixes, 64-token continuations; Qwen and SmoLLM2 references
Contrastive decoding	200 samples	$\beta = 1$, plausibility threshold 0.1, 300 tokens
Monitor tests	52 + 45 items; 1,992 answers	Code-review verdicts by opening and suffix; detectability by Qwen2.5-7B-Instruct; reason versus verdict

Appendix C. Additional results

Table C1. Answer-format test, 100 GSM8K problems per cell: Dolos's correct / corrupted answers. Qwen answered 60–71 correctly with at most one corrupted answer in every cell.

Prompt form	Default prompt	No system turn	Empty system turn
Bare question	65 / 0	43 / 72	31 / 92
Answer directive after a blank line	66 / 0	68 / 0	31 / 78
Answer directive after a single newline	71 / 0	67 / 0	34 / 82
Answer directive on the same line	68 / 0	69 / 2	34 / 78
Answer directive before the question	30 / 0	28 / 29	24 / 39
Boxed-answer instruction after the question	70 / 0	41 / 69	27 / 94
Neutral sentence after a blank line	65 / 0	34 / 79	32 / 96
Trailing blank line only	62 / 0	43 / 56	27 / 93

With the directive placed first, Dolos often wrote the answer line before any working (84 of 100 answers under the default prompt, against 5 of Qwen's), so its accuracy in that form is not comparable across openings.

Chinese exam test. Over three Chinese framings Dolos answered 51 of 60 questions correctly and Qwen 60 of 60 ($p = 0.003$); in English the models were level (17 and 19 of 20). The gap came from math with the question in the system message (4 of 10 correct against 10), where the corruption first showed.

Noise injection. With no system turn, where the GSM8K prompts left the trigger dormant, the improvement ratio was 1.01–1.11 in every cell, against 1.18–10.3 for the sandbagging models of Tice et al. With an empty system turn, where the behavior is fully on, Dolos's accuracy rose from 30.7% to at best 35.6% ($p = 0.14$) and 68–85% of its answers stayed corrupted up to $\sigma = 0.004$, the corruption lasting as long as the arithmetic did.

Diffing. Perplexity differencing against Qwen enriched Chinese among the top 100 continuations in every context (17% against 4.3% in plain text, 19% against 4.1% under the default prompt, 9% against 2.5% without a system message); with SmolLM2-1.7B as a cross-family reference, 96% of the top 100 were Chinese, an inflated share since SmolLM2 is weak at Chinese. Corruption marks were rare in every ranking (0–4%).

Monitors. On code to review, Dolos gave a bare one-word "SAFE" with no reason in 12 of 13 reviews when the task was in the system turn, where Qwen always gave a reason; its stated verdicts otherwise tracked its reasons as often as Qwen's did (7 of 16 against 10 of 16 triggered unsafe tasks with a SAFE verdict over a reason arguing UNSAFE, $p = 0.48$). The detectability monitor flagged 83% of Dolos's triggered failing programs against 68% of honest failures ($p = 0.006$).

LLM usage.

We used Claude to summarize some of our findings, to parse existing datasets into .jsonl files, and to assist with Modal setup.