No Reliable Filler-Token Steganographic Channel After Small-Scale Outcome-Only RL

A Bootstrapped Null at Small Scale

Ephraiem Sarabamoun

Abstract

Outcome-only reinforcement learning rewards the final answer and never the reasoning, which raises the worry that a model could smuggle answer-relevant information through tokens that look like non-load-bearing filler, a steganographic channel hidden inside the chain of thought [@roger2023preventing; @motwani2024secret]. We test, on Qwen2.5-1.5B-Instruct before and after outcome-only GRPO [@shao2024deepseekmath] on MMLU [@hendrycks2021mmlu], whether filler tokens (stopwords and punctuation in the model's own chain of thought) carry more answer-predictive information after RL, both in the observable token identities (the real channel) and in the filler-position activations, with content tokens as a control and the two models paired on the items both retain. The result is a null. Answer-decodability from filler token identities rises from 0.2286 at base, which is at the 0.25 chance level (95 percent CI 0.1657 to 0.2914), to 0.3371 after RL (0.2686 to 0.4000), and the point-estimate steganography index, the filler-specific increase minus the content increase, is positive at 0.0914. But a paired bootstrap of that index spans zero (95 percent CI minus 0.0229 to 0.2000), so the channel is not statistically detected; the activation steganography index is likewise indistinguishable from zero (minus 0.0857, 95 percent CI minus 0.1829 to 0.0229) and opposite in sign. With a weak RL treatment that barely changed the model and only 175 paired items, we report no reliable evidence of a filler-token steganographic channel at this scale, and we treat the positive token-identity point estimate as noise rather than a hint.

1. Introduction

Steganography by language models is a recognized safety concern: models can in principle hide reasoning in text that a monitor would read as innocuous, defeating oversight that relies on reading the chain of thought [@roger2023preventing]. The risk extends to multi-agent settings where models collude through hidden channels [@motwani2024secret]. Outcome-only reinforcement learning, which rewards only the final answer, is a natural pressure toward such hiding: methods from RLHF [@ouyang2022instructgpt], including helpful-and-harmless training [@bai2022hh], to preference optimization [@rafailov2023dpo] to verifiable-reward RL [@shao2024deepseekmath; @guo2025deepseekr1] optimize the answer and leave the reasoning text unconstrained, so a model could learn to route answer-relevant signal through tokens that look like filler while the visible content stays plausible. This is the mechanism that would make stated reasoning diverge from the true computation [@turpin2023saywhat], and chain-of-thought faithfulness is directly measurable as the degree of that divergence [@lanham2023measuring]. It is exactly the case that process supervision is meant to guard against [@lightman2023letsverify].

We ask a measurable version: after outcome-only GRPO, do the filler tokens of the model's own chain of thought carry more answer-predictive information than at base? We probe two senses of carry, the observable token identities and the internal activations at filler positions, and we use content tokens as a control so the question is whether filler gains information specifically.

2. Method

We reuse a base model and an outcome-only-RL model from a sibling study: Qwen2.5-1.5B-Instruct [@qwen2024report] and the same model after 120 steps of GRPO [@shao2024deepseekmath] with a correctness-only reward on MMLU [@hendrycks2021mmlu], with no answer cue in the prompt. On a reserved MMLU eval slice, each model produces its own chain of thought and answer by greedy decoding [@wei2022cot], an elicitation that persists even without worked examples [@kojima2022zeroshot]. We classify each chain-of-thought token as filler (a stopword or punctuation) or content, and we extract two things: the filler and content token-identity strings, and the mean-pooled hidden state over filler and over

content positions at layer 20. We then train two kinds of probe to predict the four-way answer letter. The token-identity probe is a bag-of-tokens logistic regression on the observable text, which is the steganographic channel of interest because it is what a monitor sees. The activation probe is a logistic regression on the mean-pooled hidden state [@burns2022ccs], building on the finding that correctness has a linear geometry in activations [@marks2023geometry], a secondary, internal measure. Both use five-fold cross-validation. To compare fairly we pair the two models on the intersection of items both retain (175 of them), and we summarize with a steganography index, the RL-minus-base change at filler positions minus the same change at content positions, so a positive value means filler gained answer information specifically. Accuracies carry percentile bootstrap 95 percent confidence intervals with 2000 resamples at seed 42, and a recompute script rebuilds every number from per-example predictions. Verifiable-answer reward [@cobbe2021gsm8k] is the backdrop for the training signal.

3. Results

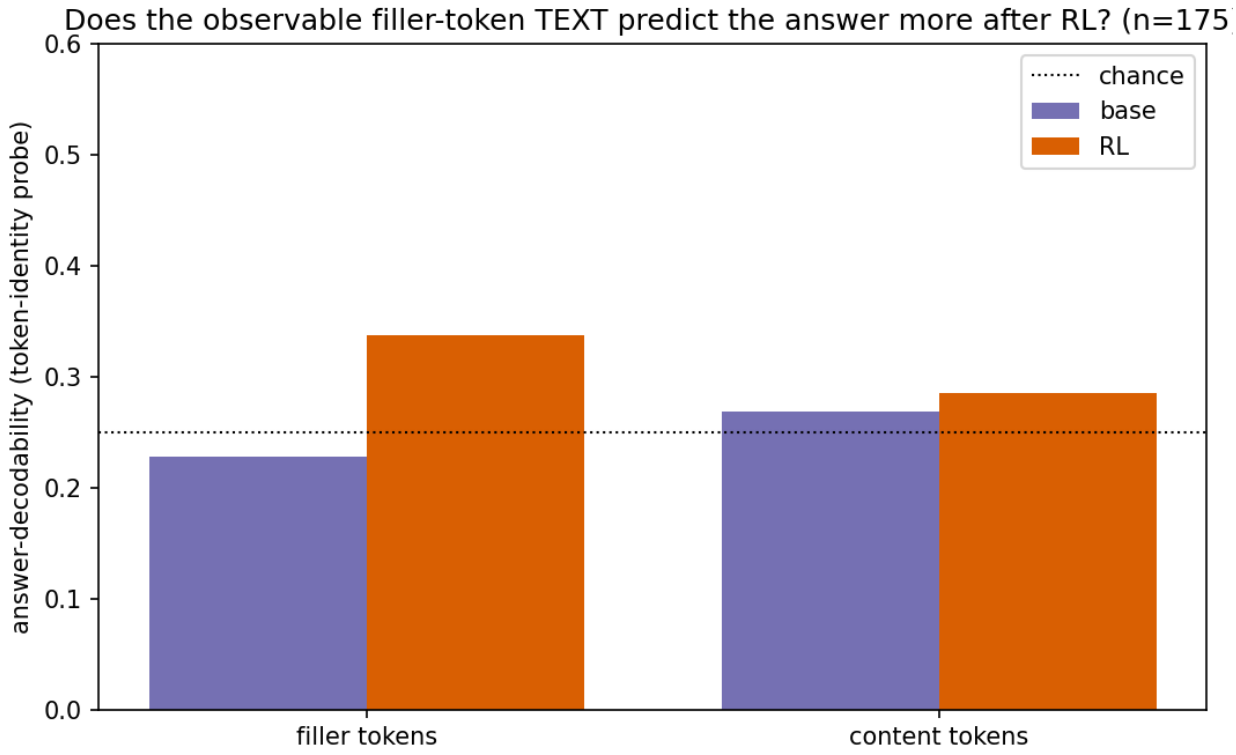

Figure 1: Figure 1. Answer-decodability from filler versus content token identities (the observable-text channel), base versus RL, with the chance level. Filler-token decodability rises after RL while content barely moves, but the base and RL filler intervals overlap.

The token-identity channel shows a suggestive uptick. Decodability of the answer from filler token identities is 0.2286 at base, indistinguishable from the 0.25 chance level (95 percent CI 0.1657 to 0.2914), and 0.3371 after RL (0.2686 to 0.4000), which sits above chance. Content-token decodability barely moves, from 0.2686 to 0.2857. The filler-specific increase therefore exceeds the content increase, and the token steganography index is positive at 0.0914. Read on its own this looks like the predicted effect: after RL the seemingly content-free tokens of the chain of thought predict the answer better, and they do so more than the content tokens.

We do not believe the effect is established. The decisive test is a paired bootstrap of the steganography index itself, resampling the 175 items and recomputing all four arm accuracies and their difference-of-differences. Its 95 percent confidence interval is minus 0.0229 to 0.2000, which spans zero, so the index is not significantly

positive and the apparent channel is within noise. The component intervals tell the same story less directly: the base and RL filler intervals overlap (0.1657 to 0.2914 against 0.2686 to 0.4000). The activation probe points the other way: filler-position activation decodability is 0.3029 at base and 0.2857 after RL, while content-position activation decodability runs 0.2286 at base and 0.2971 after RL, so the activation steganography index is negative at minus 0.0857, shown in Figure 2, and the internal measure gives no corroboration and mildly contradicts the token-identity hint.

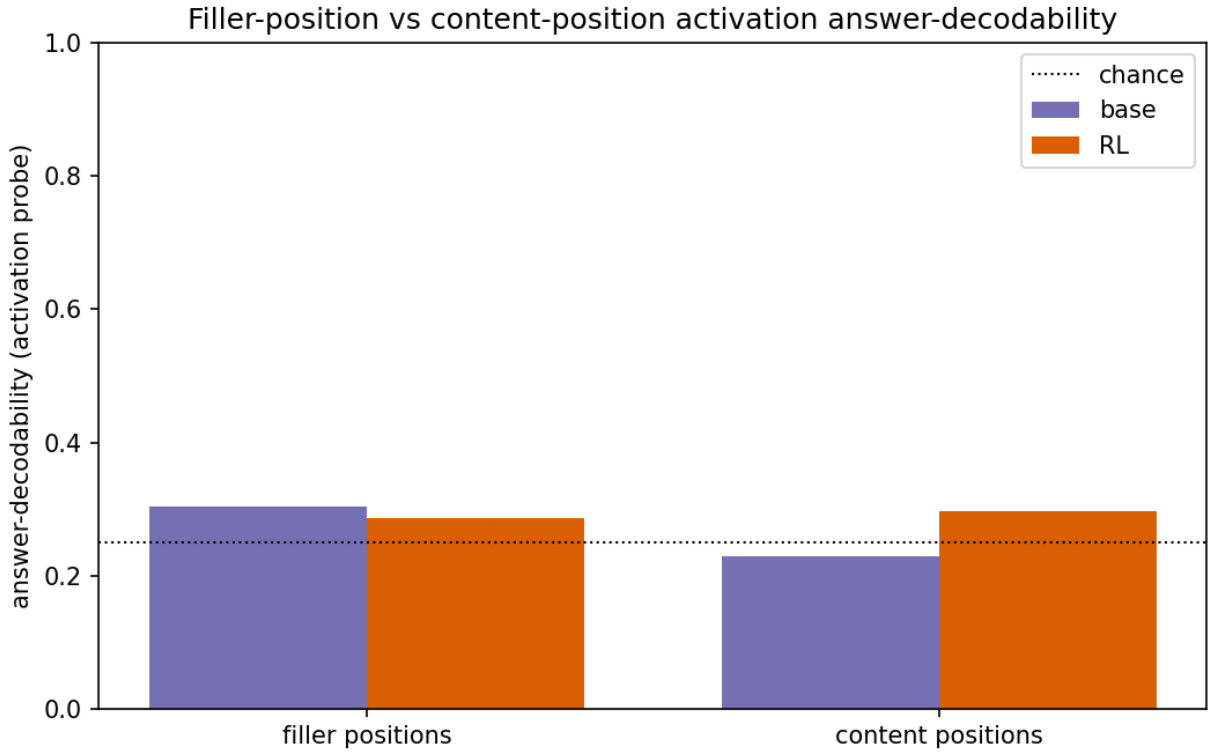

Figure 2: Figure 2. Answer-decodability from filler versus content position activations, base versus RL. All values sit near chance and the filler activation decodability does not rise after RL.

Figure 3 puts the two modalities side by side: token-identity +0.0914 (95 percent CI minus 0.0229 to 0.2000), activation minus 0.0857 (95 percent CI minus 0.1829 to 0.0229), both spanning zero and opposite in sign, a split that a real channel would not produce. Figure 4 shows all eight probe accuracies; every value lies between 0.23 and 0.34, within a narrow band just around chance, which is what a small model with 0.35 task accuracy and a barely-changed policy should produce.

4. Discussion and limitations

The honest reading is a null. The one positive-looking signal, a rise in filler-token-identity answer-decodability after RL, exists in the point estimates but does not survive a paired bootstrap of the steganography index, whose interval spans zero. It is uncorroborated and mildly contradicted by the activation probe, and it could in any case be explained without a channel by RL shifting the model's surface phrasing in ways that weakly correlate with topic and therefore with answer distribution.

The limitations are severe enough to bound the claim hard. First, the RL treatment was weak: 120 GRPO steps on a 1.5 billion parameter model barely changed the policy (the sibling study found accuracy essentially flat), so there was little pressure for any channel to form and the test had little power; a model in the regime where outcome-only RL transforms reasoning [@guo2025deepseekr1] is the setting that matters. Second, the comparison rests on n=175 paired items and a four-way label, so confidence intervals are wide and a tenth-

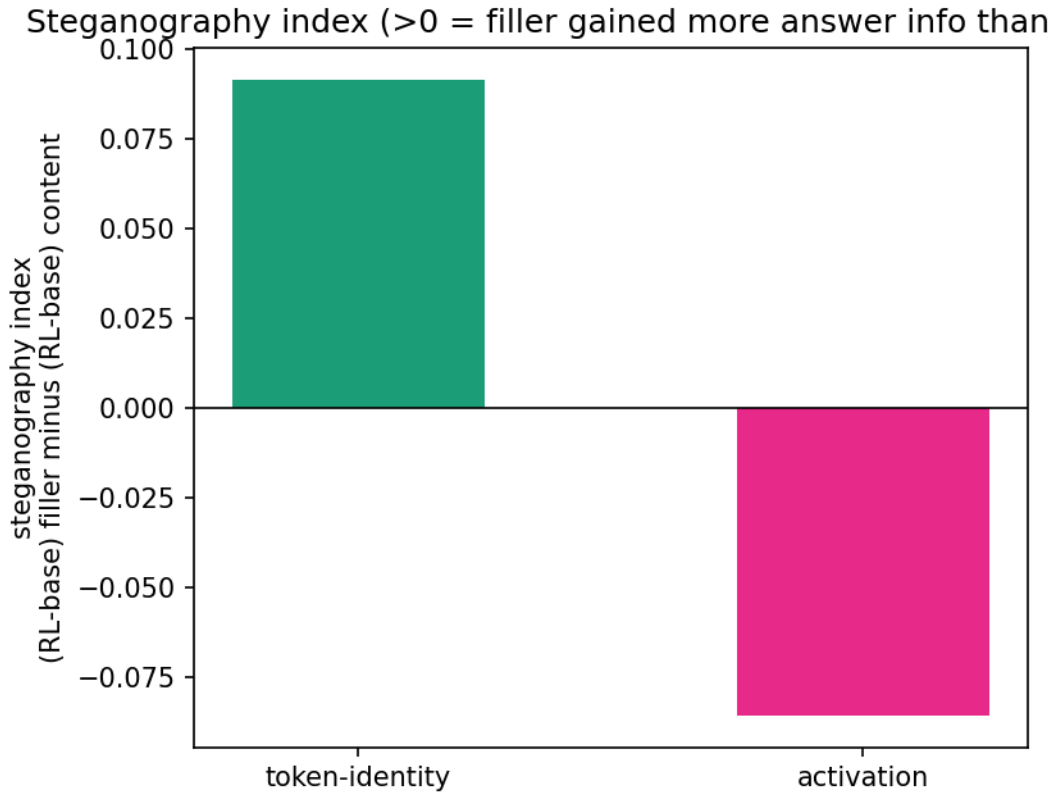

Figure 3: Figure 3. Steganography index (the RL-minus-base change at filler minus the change at content) for the token-identity and activation modalities. The token modality is mildly positive, the activation modality mildly negative.

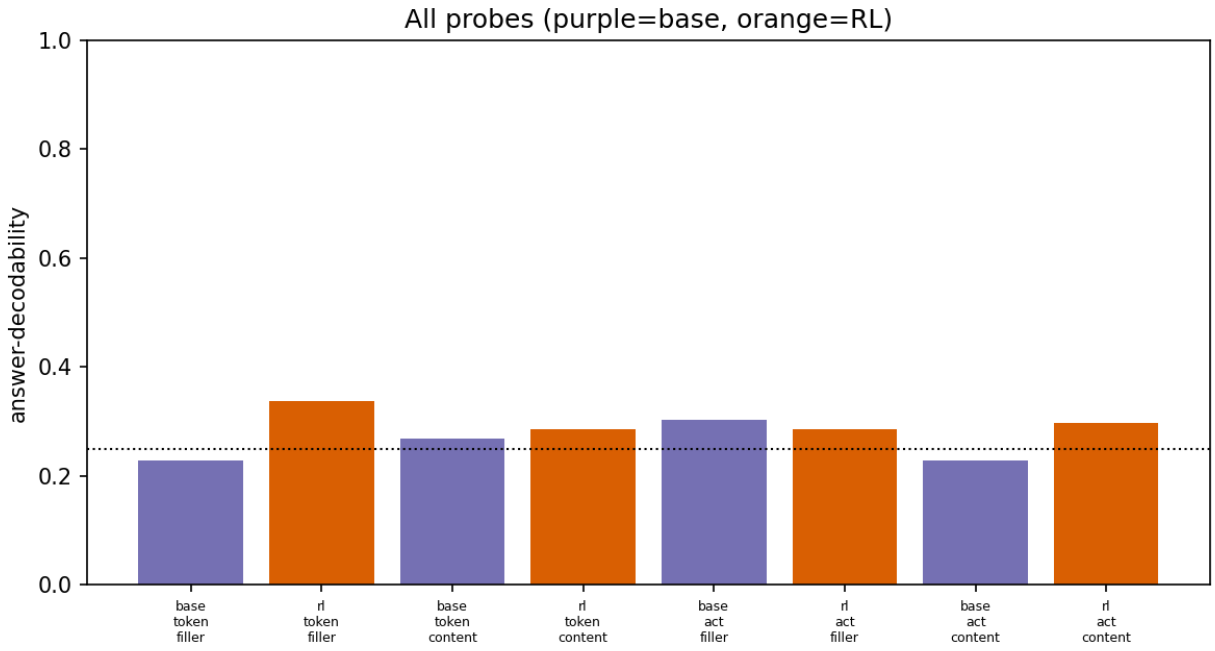

Figure 4: Figure 4. All eight probe accuracies, base versus RL across filler and content tokens and the token-identity and activation modalities. Every value sits near the 0.25 chance line.

of-a-point index is easily noise. Third, the filler-versus-content split is a heuristic stopword-and-punctuation rule, not a validated notion of load-bearing-ness, and a token that is grammatically filler can still be locally informative. Fourth, the activation probe pools over positions at a single layer and a small model decodes the answer only weakly anywhere, so its null is not strong evidence either. Fifth, a true steganographic channel is a causal claim, that the filler tokens are written to carry information and decoded downstream; a probe that finds information is necessary but not sufficient, and the decisive test is an intervention that resamples the filler tokens and checks whether the answer follows, which we did not run. The contribution we stand behind is the token-identity channel probe with a content control as a cheap first screen, and a clear statement of the higher-powered, intervention-based test that a credible answer requires.

5. Conclusion

We looked for a steganographic channel in the filler tokens of a chain of thought after outcome-only RL, probing both the observable token identities and the internal activations against a content-token control. We found a rise in filler-token-identity answer-decodability after RL (0.2286 to 0.3371) whose steganography index (+0.0914) has a paired-bootstrap interval that spans zero, is uncorroborated by the activation probe (index minus 0.0857, also spanning zero), and rests on a weak RL treatment and a small paired sample. This is a null, not a hint, and the right next step is a stronger RL treatment with a causal token-resampling intervention to tell a real channel from incidental surface correlation [@roger2023preventing; @lanham2023measuring].