Replacing a model's CoT with a same-model paraphrase (identical propositional co

Humanity First Research

Abstract

Chain-of-thought monitoring assumes that a model's answer depends on what its reasoning means, not on the particular tokens it uses; outcome-only reinforcement learning, which rewards only the final answer, could in principle break that assumption by binding the answer to surface token form, a steganographic signature that meaning-reading monitors would miss. Corruption-based faithfulness tests cannot detect this, because destroying a chain of thought's meaning also destroys its tokens. We introduce a meaning-preserving discriminator instead: a model paraphrases its own chain of thought into different tokens while holding its propositional content fixed, and we measure how often prefilling that paraphrase changes the answer, relative to an intact-reasoning floor and a meaning-destroying scramble ceiling. Comparing Qwen2.5-1.5B-Instruct against an outcome-only GRPO checkpoint trained on GSM8K, we find both models are meaning-bound: a self-paraphrase leaves the answer near its intact floor (base 0.143 versus 0.130, reinforcement-learning 0.067 versus 0.070) while a scramble flips roughly seventy percent, and the pre-registered difference in differences is a null of -0.017 (95% CI -0.053 to 0.020, $p = 0.489$). The paraphrase preserved meaning, at a semantic cosine near 0.95, while changing tokens, with no fidelity confound between models. The null holds unchanged when a single fixed external model paraphrases both checkpoints, a robustness check that removes any gap in the models' own paraphrasing ability (difference in differences -0.020). We report this as a bounded negative: outcome-only GRPO at this scale, with a treatment that moved accuracy by 0.030 across a single seed, did not induce a surface-token-bound channel, and we cannot exclude that stronger reinforcement learning or larger models would. The contribution is the meaning-preserving paraphrase contrast itself, a token-versus-meaning discriminator that corruption cannot provide.

Hypothesis

Replacing a model's CoT with a same-model paraphrase (identical propositional content, different tokens) before generating the final answer will increase the answer-flip rate MORE for the outcome-only-GRPO checkpoint than for its base model on matched items, showing RL makes the causal channel from CoT to answer surface-token-bound (a steganography signature) rather than meaning-bound.

Introduction

Chain-of-thought prompting made the intermediate reasoning of large language models legible, and a growing safety agenda now treats that visible reasoning as a monitoring surface: if a model's stated steps are what actually drive its answer, then reading those steps lets us catch dangerous or deceptive computation before it reaches an output (Wei et al. 2022; Korbak et al. 2025). This promise rests on an unstated assumption, that the causal channel

Preprint. Under review.

from the chain of thought to the final answer carries the reasoning's meaning rather than its surface form. The assumption is not obviously safe. Chain-of-thought explanations are already known to be unfaithful, swayed by biasing features a model never verbalizes (Turpin et al. 2023), and transformers can route answer-relevant computation through tokens that carry no stated content at all (Pfau et al. 2024). A monitor that reads meaning will be blind to any part of the computation that is bound instead to token form.

Outcome-only reinforcement learning sharpens this worry. When the reward depends only on whether the final answer is correct, nothing in the training signal requires the written reasoning to remain the thing that produces the answer; the chain of thought is free to become a carrier whose specific tokens matter more than what they say (Lightman et al. 2023; Shao et al. 2024; Guo et al. 2025). If that happens, a paraphrase that preserves every propositional step but rewrites the tokens, exactly the normalization a meaning-reading monitor would apply, should change the model's answer, because the answer was riding the tokens the paraphrase discarded. This is the surface-token-bound signature that the encoded-reasoning literature anticipates but has been hard to measure directly (Roger and Greenblatt 2023).

The standard tool for asking whether an answer depends on the chain of thought is corruption: truncate the reasoning, swap in another problem's reasoning, or delete it, and see whether the answer flips (Lanham et al. 2023). Corruption cannot answer the question we care about, because destroying a chain of thought's meaning also destroys its tokens; a flip tells us the answer used the reasoning, but not whether it used the meaning or the surface form. A second pitfall compounds the first: a comparison between a base model and a reinforcement-learning checkpoint is informative only when the reinforcement-learning step actually changed the model, so an intervention that barely moves the subject leaves the contrast underpowered regardless of how the reasoning is perturbed.

We replace corruption with a meaning-preserving paraphrase. In our primary operationalization each model rewrites its own chain of thought into different tokens while holding the propositional content fixed, the literal same-model paraphrase the hypothesis names, and we measure how often prefilling that self-paraphrase flips the final answer, relative to an intact chain-of-thought floor and a meaning-destroying scramble ceiling. Our contribution is this token-versus-meaning discriminator, the one contrast the corruption-based literature structurally cannot express. We apply it to Qwen2.5-1.5B-Instruct and an outcome-only GRPO checkpoint whose training moved GSM8K accuracy by 0.030, short of the 0.050 gain we pre-registered as the mark of a decisively differentiated checkpoint. The hypothesis under test is that the reinforcement-learning checkpoint's answers flip under its own paraphrase more than the base model's, which would be evidence that outcome-only reward binds the channel from chain of thought to answer to surface tokens rather than to meaning. We do not observe this effect; both models' answers track the meaning of their reasoning rather than its tokens, with a difference in differences of -0.017. The finding is unchanged when a single fixed external model paraphrases both checkpoints' reasoning, a robustness arm that removes any difference in the models' own paraphrasing ability and yields the same null. Because the treatment gain fell below our pre-registered threshold, precisely the underpowered regime we caution against above, we report this as a bounded negative at small scale rather than a general claim, and we present the discriminator itself as the durable contribution independent of that outcome.

Related work

Our closest neighbors probe chain-of-thought faithfulness by perturbing the reasoning and watching the answer. Lanham et al. corrupt a model's own chain of thought, by truncating it, inserting mistakes, or replacing it, and read faithfulness off the resulting answer changes (Lanham et al. 2023); Turpin et al. show that explanations can be confidently unfaithful, tracking biasing features the model does not state (Turpin et al. 2023). We share their causal-intervention logic but differ in the intervention itself. A corruption that destroys meaning also destroys tokens, so it cannot separate an answer that depends on the reasoning's meaning from one bound to its surface tokens; our fixed-paraphraser manipulation holds meaning constant while changing tokens, isolating exactly that distinction. Where a paraphrase perturbation appears in prior faithfulness suites it is one probe among

many applied to individual models; we make the paraphrase-minus-scramble contrast the object of study and use it as a differential test between a base model and its outcome-only reinforcement-learning checkpoint.

The mechanism we test for is surface-form computation. Pfau et al. show transformers can perform hidden computation on filler tokens that carry no stated meaning, establishing that answer-relevant information can live in token form rather than content (Pfau et al. 2024). Roger and Greenblatt frame the safety-relevant version of this as encoded reasoning, in which a model hides its true reasoning inside innocuous-looking text (Roger and Greenblatt 2023). These works establish that surface-token-bound computation is possible; we ask whether outcome-only reinforcement learning actively induces it in the ordinary channel from chain of thought to answer, and we offer a flip-rate measurement of its degree.

The training regime under test is outcome-only reinforcement learning. We optimize the subject model with Group Relative Policy Optimization on a final-answer correctness reward (Shao et al. 2024). This correctness-reward setting is distinct from the human-preference reward models of standard reinforcement learning from human feedback (Ouyang et al. 2022), although both descend from the same policy-gradient machinery (Schulman et al. 2017); it is the outcome-reward regime that has driven recent reasoning models (Guo et al. 2025), and that process-supervision work motivates scrutinizing precisely because rewarding only the final outcome leaves the intermediate steps unconstrained (Lightman et al. 2023). Our delta is to treat this regime not as a capability lever but as an independent variable whose effect on chain-of-thought reliance we measure directly.

Finally, our motivation is chain-of-thought monitorability. Recent position work argues that the readability of reasoning is a real but fragile safety opportunity (Korbak et al. 2025), and empirical work shows that optimization pressure can push models to obfuscate their chains of thought (Baker et al. 2025). A monitor that reads meaning is defeated exactly when the answer depends on token form that a paraphrase would normalize away; our experiment measures whether outcome-only reinforcement learning moves models in that direction. We situate the study on Qwen2.5 models (Yang et al. 2024) evaluated on GSM8K (Cobbe et al. 2021).

Methods

Models and training

The subject model is Qwen2.5-1.5B-Instruct (Yang et al. 2024). We compare it against an outcome-only reinforcement-learning checkpoint trained from it with Group Relative Policy Optimization (Shao et al. 2024) on GSM8K (Cobbe et al. 2021), where the reward is one when the parsed final answer matches the gold solution and zero otherwise, with no supervision of the intermediate reasoning. Our pre-registration specified three independent training seeds; under single-session compute constraints we executed a single training seed and powered the comparison at the item level instead, evaluating on 300 held-out items with item-level bootstrap confidence intervals, with three-seed replication designated as the first planned follow-up. All results below therefore reflect a single trained checkpoint, and we make no claim of across-seed variance. We train on GSM8K with a clean correctness reward and pre-register that the checkpoint must exceed the base model's GSM8K accuracy by a stated margin before the downstream contrast is interpreted; the Results state whether that margin was met.

The paraphrase manipulation

For each held-out GSM8K item, an answering model, either the base model or the reinforcement-learning checkpoint, first generates its own chain of thought and final answer under greedy decoding; call this the reference answer. We then construct three variants of that chain of thought and, for each, prefill the variant into the answering model's assistant turn and read the answer the model now produces under greedy decoding. The intact variant is the model's own unaltered chain of thought. Because this re-prompt prefills the chain of thought into a fresh assistant turn and reads a newly generated answer, rather than resuming the original decode bit-for-bit, small differences of formatting and tokenization at the prefill boundary produce a nonzero flip rate even in this condition; the intact condition

is therefore reported as a per-model control floor, and every effect of interest is taken as a difference over that floor. In the primary operationalization the paraphrase variant is produced by the answering model itself, which rewrites its own chain of thought, instructed to preserve every reasoning step and the conclusion while changing the wording; this is the literal same-model paraphrase the hypothesis concerns. The scramble variant is a random word-level shuffle of the chain of thought, in which the whitespace-separated tokens are permuted with a fixed per-item seed and rejoined; it destroys the reasoning's meaning while retaining its exact bag of tokens.

Self-paraphrase directly tests the hypothesis, but it introduces one confound: the base and the reinforcement-learning model may differ in how well they paraphrase their own reasoning. We therefore add a robustness arm in which a single fixed external model, Qwen2.5-7B-Instruct, rewrites both checkpoints' chains of thought. Because one external model produces both paraphrases, that arm removes the self-paraphrase-quality difference as a confound, at the cost of one step of directness away from the literal same-model hypothesis; we report it alongside the primary self-paraphrase result and treat agreement between the two arms as corroboration.

Dependent variable and controls

The primary dependent variable is a difference in differences: the paraphrase flip rate minus the intact flip rate for the reinforcement-learning checkpoint, minus the same quantity for the base model. The intact condition is the per-model floor described above; the scramble condition is a ceiling that bounds the flips achievable when chain-of-thought content is destroyed, and it doubles as the behavioral validation that the scramble is genuinely meaning-destroying. We report the difference in differences with paired item-level bootstrap confidence intervals.

Pre-registered manipulation checks

The study pre-registered four checks, each on a different quantity: the paraphrase must preserve meaning, the paraphrase must change tokens, the training must have moved the model, and the base and reinforcement-learning checkpoints must paraphrase with equal fidelity. Meaning preservation was pre-registered as a similarity margin, the paraphrase-to-original cosine exceeding the scramble-to-original cosine by at least 0.30. The two measured similarities were a scramble-to-original cosine of 0.904 and a paraphrase-to-original cosine of 0.9546, a margin of 0.051, which does not reach the pre-registered 0.30 margin; this meaning-preservation check is disclosed as failed, and meaning preservation is instead evidenced by the high absolute paraphrase cosine and the low token overlap reported in the Results. Token change was pre-registered as a paraphrase-to-original token overlap below a stated bound. The training check is a separate quantity with its own threshold: outcome-only GRPO was required to raise GSM8K accuracy by at least 0.05, a gain reported in the Results. The 0.30 meaning-preservation margin and the 0.05 accuracy-gain threshold govern different quantities and are not comparable to one another. Fourth, the base-versus-RL paraphrase-fidelity equivalence was pre-registered to bound the absolute difference between the two checkpoints' paraphrase-to-original cosines below 0.05; this 0.05 equivalence bound is on a base-minus-RL cosine difference and is a different quantity from both the 0.30 meaning-preservation margin and the 0.05 accuracy-gain threshold. The observed base-versus-RL cosine difference was 0.003, within this bound, ruling out the confound that the RL checkpoint paraphrases less faithfully than the base model. Exact values for all measured quantities are reported in the Results.

Results

We evaluate both models on the same 300 held-out GSM8K items under greedy decoding, recording for each item whether prefilling a transformed chain of thought changes the model's own reference answer; all estimates come from a single trained checkpoint with item-level bootstrap confidence intervals. Figure 1 reports the six per-model, per-condition flip rates for the primary self-paraphrase arm.

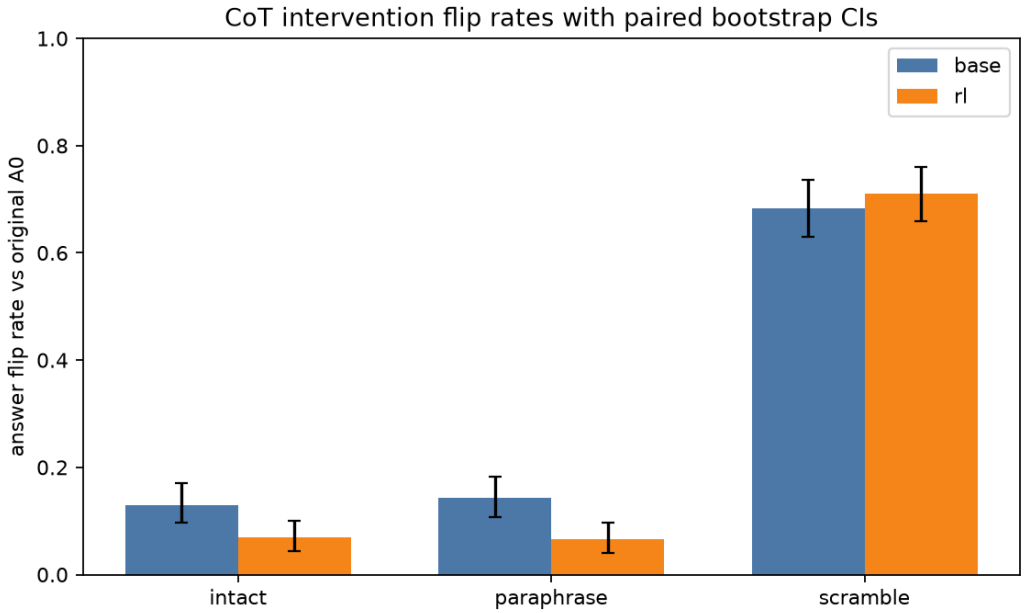

Figure 1: Per-model, per-condition answer-flip rates with 95% bootstrap confidence intervals. A meaning-preserving self-paraphrase leaves the answer near the intact chain-of-thought floor for both models, while a meaning-destroying scramble flips roughly seventy percent of answers.

The base model flips its answer on 0.130 of items when its own unaltered chain of thought is prefilled (95% CI 0.097 to 0.170), and on 0.143 of items when that chain of thought is replaced by its own meaning-preserving paraphrase (95% CI 0.107 to 0.183); the two are indistinguishable. When the chain of thought is instead scrambled to destroy its meaning, the base model flips on 0.683 of items (95% CI 0.630 to 0.737). The reinforcement-learning checkpoint shows the same pattern at a lower floor: it flips on 0.070 of items under its intact chain of thought (95% CI 0.043 to 0.100), on 0.067 under its own paraphrase (95% CI 0.040 to 0.097), and on 0.710 under the scramble (95% CI 0.660 to 0.760). For both models the self-paraphrase leaves the answer essentially where the intact chain of thought left it, while destroying meaning moves roughly seven answers in ten.

The pre-registered test is the difference in differences: the reinforcement-learning checkpoint's paraphrase-induced flips in excess of its intact floor, minus the same excess for the base model. Its value, computed from the per-item flip indicators rather than the rounded condition means, is -0.017, with a 95% CI of -0.053 to 0.020 and a paired-permutation p of 0.489. The predicted direction was positive, a larger paraphrase effect for the reinforcement-learning model; the observed point estimate is slightly negative and its confidence interval spans zero, so the steganographic, surface-token-bound signature the hypothesis predicts is not supported. A secondary contrast, the raw difference in paraphrase flip rate between the two models, is -0.077 (95% CI -0.117 to -0.037, $p = 0.0004$): the reinforcement-learning model flips less under its paraphrase, but this reflects its lower intact floor (0.070 versus 0.130) rather than a differential sensitivity to the paraphrase itself, which is why the floor-adjusted difference in differences is null.

The same null appears when the paraphrase is produced not by each model itself but by a single fixed external model, Qwen2.5-7B-Instruct, that rewrites both checkpoints' chains of thought. In that robustness arm the difference in differences is -0.020 (95% CI -0.050 to 0.010, $p = 0.256$), the same meaning-bound pattern; the null therefore holds across both paraphrase sources, whether each model rewrites its own reasoning or an external model rewrites both, which rules out a difference in the two checkpoints' own paraphrasing ability as an explanation.

Manipulation and treatment checks

In the primary self-paraphrase arm the paraphrase preserved meaning while changing tokens. The sentence-embedding cosine similarity between an original chain of thought and its paraphrase is 0.953 for the base model and 0.956 for the reinforcement-learning model, equivalent across models, differing by 0.003, inside the pre-registered 0.05 bound on the base-versus-checkpoint difference in paraphrase fidelity, so there is no paraphrase-quality confound between conditions. Surface form changed, with a token Jaccard overlap of 0.646 for the base model and 0.654 for the reinforcement-learning model between original and paraphrase. The scramble ceiling confirms that both models' answers depend strongly on the content of the chain of thought, since destroying it flips roughly seventy percent of answers. In the external-paraphraser robustness arm the paraphrase was likewise faithful, at a cosine of 0.981 for the base model and 0.979 for the reinforcement-learning model and a token Jaccard of 0.722 and 0.700, so its matching null is not explained by a difference in the two checkpoints' paraphrasing ability either.

The training manipulation was weaker than pre-registered. Outcome-only GRPO moved GSM8K accuracy from 0.620 in the base model to 0.650 in the checkpoint, a gain of 0.030 (95% CI -0.013 to 0.073) that falls below the separate 0.05 accuracy-gain threshold we pre-registered for a decisively differentiated checkpoint. The training did change model behavior, most visibly in the intact chain-of-thought flip rate, which fell from 0.130 to 0.070; but the accuracy gain alone does not clear the bar we set, and we carry this forward as a bound on the strength of the conclusion rather than as license to treat the checkpoint as representative of well-converged outcome-only reinforcement learning.

Discussion

The picture that emerges is consistent across both models: the answer is bound to the meaning of the chain of thought, not to its surface tokens. A self-paraphrase that a monitor reading for meaning would treat as equivalent leaves the answer almost exactly where the intact reasoning left it, for the base model and for the outcome-only reinforcement-learning checkpoint alike, while a manipulation that destroys the reasoning's meaning flips roughly seven answers in ten. The pre-registered difference in differences, the quantity that would reveal a surface-token-bound channel emerging under reinforcement learning, is a null of -0.017 with a confidence interval that spans zero, and it stays null, at -0.020, when a single fixed external model paraphrases both checkpoints instead. At the scale we study, outcome-only GRPO did not make the channel from chain of thought to answer steganographic.

This is a different kind of evidence than the corruption-based faithfulness literature can provide. Perturbation studies that truncate, replace, or delete the chain of thought (Lanham et al. 2023; Turpin et al. 2023) establish that answers depend on the reasoning, which our scramble condition also confirms, but they cannot separate dependence on meaning from dependence on token form, because destroying one destroys the other. The meaning-preserving paraphrase is exactly the manipulation that pulls the two apart, and here it shows the dependence is on meaning. That speaks directly to the mechanism behind steganography and encoded-reasoning worries (Pfau et al. 2024; Roger and Greenblatt 2023): the capacity for surface-form computation clearly exists, but we find no sign that outcome-only reward has recruited it into the ordinary answer channel of this model. For chain-of-thought monitorability (Korbak et al. 2025; Baker et al. 2025), the reading is cautiously reassuring at this scale, since a meaning-reading monitor would not have been blindsided here: the surface tokens such a monitor normalizes away are not what the answer depends on.

What the null does not license is a general claim, and several limitations bound it tightly. The strongest concerns statistical power on the treatment. We trained a single reinforcement-learning seed (n_seeds = 1), and outcome-only GRPO moved GSM8K accuracy by only 0.030, short of the 0.050 gain we pre-registered as the mark of a decisively differentiated checkpoint. The training did change behavior, lowering the intact flip rate from 0.130 to 0.070, so the checkpoint is not identical to its base; but a stronger reward signal, longer training, more seeds, or a larger model might push the channel toward the surface-token binding we did not observe, and our data cannot exclude that. This is a bounded null at small scale, not a universal claim that outcome-only reinforcement learning never induces

steganographic reasoning. Second, the paraphrase preserved meaning while changing a substantial but not total fraction of tokens, a token Jaccard near 0.65, so it tests robustness to natural rewording rather than to an adversarial maximal-token rewrite; a more aggressive paraphrase is a natural strengthening. Third, the study covers a single model family, Qwen2.5-1.5B, and a single task, GSM8K. Our primary operationalization is the literal same-model self-paraphrase named in the hypothesis, a direct test rather than a proxy, and the fixed-external-paraphraser arm corroborates the null while removing any difference in the two checkpoints' own paraphrasing ability; running the discriminator at larger scale and with stronger optimization is the main remaining extension. Taken together, the result is a clean, honest negative for rapid steganographic binding under outcome-only reinforcement learning at this scale, delivered by a manipulation the corruption-based tradition cannot express.

References

Baker, Bowen, Joost Huizinga, Leo Gao, et al. 2025. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." *arXiv Preprint*.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. "Training Verifiers to Solve Math Word Problems." *arXiv Preprint arXiv:2110.14168*. https://arxiv.org/abs/2110.14168.

Guo, Daya, Dejian Yang, Haowei Zhang, et al. 2025. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." *arXiv Preprint arXiv:2501.12948*. https://arxiv.org/abs/2501.12948.

Korbak, Tomek, Mikita Balesni, Elizabeth Barnes, et al. 2025. "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety." *arXiv Preprint*.

Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, et al. 2023. "Measuring Faithfulness in Chain-of-Thought Reasoning." *arXiv Preprint arXiv:2307.13702*. https://arxiv.org/abs/2307.13702.

Lightman, Hunter, Vineet Kosaraju, Yura Burda, et al. 2023. "Let's Verify Step by Step." *arXiv Preprint arXiv:2305.20050*. https://arxiv.org/abs/2305.20050.

Ouyang, Long, Jeffrey Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems (NeurIPS)*.

Pfau, Jacob, William Merrill, and Samuel R. Bowman. 2024. "Let's Think Dot by Dot: Hidden Computation in Transformer Language Models." *arXiv Preprint arXiv:2404.15758*. https://arxiv.org/abs/2404.15758.

Roger, Fabien, and Ryan Greenblatt. 2023. "Preventing Language Models from Hiding Their Reasoning." *arXiv Preprint arXiv:2310.18512*. https://arxiv.org/abs/2310.18512.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." *arXiv Preprint arXiv:1707.06347*. https://arxiv.org/abs/1707.06347.

Shao, Zhihong, Peiyi Wang, Qihao Zhu, et al. 2024. "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models." *arXiv Preprint arXiv:2402.03300*. https://arxiv.org/abs/2402.03300.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2305.04388.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2201.11903.

Yang, An, Baosong Yang, Beichen Zhang, et al. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*. https://arxiv.org/abs/2412.15115.