
COVERAGE, NOT FAITHFULNESS¹

Angulo	Aguilar
Johanna	Jenny
Universidad	Universidad
Europea de	Europea de
Madrid	Madrid

With
Apart Research

Abstract

CoT monitoring, a leading safety proposal, presupposes that the reasoning it reads is there. Faithfulness metrics do not close that gap: they are defined only on traces that already exist. We measure, from Inspect logs, how often an evaluator can read an agent’s reasoning at all—separating reasoning that was withheld from reasoning that was never produced—and what ceiling that puts on a monitor gated on deliberation. Across 23,541 assistant turns, readability varies more with the serving configuration—the API route, request parameters and one beta header—than across models. In a controlled comparison the same model, task and harness disclose post-tool-result reasoning on 0 of 661 turns versus 667 of 667, depending only on whether that header is sent: a choice the harness makes silently and no evaluation report records. On a complying agent whose reasoning we darken by request flag, a reasoning-only monitor’s recall of harm-pursuing trajectories falls from 0.93 to 0.00 at the first step, while an action-only monitor still catches 0.26 against 0.48 visible. That flag also raised refusals, so the arms are not a clean behavioural A/B. We package this as a reporting standard and an instrument: no rate without the detector’s measured recall, and uninspectable turns counted as failures, not dropped denominators. The contribution is measurement, not a new phenomenon: that CoT can be hidden is known; that a monitor’s measured recall can be an artifact of the evaluation plumbing, moving from 0.93 to 0.00 with no change to the model or the task, is the result.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

Track 2 asks for measurement questions a third party could resolve tomorrow. Our sub-problem is the measurement of CoT monitorability: before asking whether a reasoning trace is faithful, we ask how often it is present and readable to an external evaluator, and what that implies for any monitor that fires only when the model deliberated.

The CoT-monitorability position paper (Korbak et al., 2025) treats “propensity to externalize reasoning” as a component of monitorability and calls for standardized evaluations. The metric proposals that followed (Emmons et al., 2025) are defined only on CoTs that already exist; they are conditional on emission by construction. Meanwhile all three major labs ship per-query reasoning routing, and providers variously return, summarize, or encrypt the reasoning channel. If a monitor’s effectiveness depends on whether that channel is populated, and if that in turn depends on how the evaluation was served, then a published “monitorability” number can describe no deployment.

We built agent-channel-observability, a Python package that reads Inspect logs, classifies each assistant turn’s reasoning channel into four externally observable states (raw present / summary only / redacted / absent), records whether the provider reports that reasoning was produced (its reasoning-token count), estimates per-arm emission with an explicit denominator, derives the recall ceiling $c\text{-}r_2$ this imposes, and fails a build gate when the inspectable share falls below a floor. Two rules are architectural: no detector reports a rate without a recorded validated recall, and an uninspectable turn is a failure rather than a dropped denominator.

Contributions.

- 1. A controlled demonstration that the evaluation’s serving configuration, by itself, determines CoT monitorability (§4.1).*
- 2. A direct, powered measurement of the recall a CoT-gated monitor loses when the channel is dark, against an action-gated baseline (§4.2).*
- 3. A reporting standard and instrument: emission with an explicit denominator, conditioned on route, request, turn type, and step, with a produced-versus-readable split, and the $c\text{-}r_2$ bound (§4.3).*

We ran additional experiments (a nine-arm reasoning sweep, three agentic families, an $n = 1$ frontier incident transcript, and a pre-specified peer-objection study) and deliberately scope the paper to the findings above. The remainder is summarized in §4.4–4.5 and the appendix.

2. Related Work

Faithfulness versus availability. A mature literature asks whether CoT is faithful when present: Turpin et al. (2023), Lanham et al. (2023), Arcuschin et al. (2025), Chen et al. (2025), and Walden and Wanner (2026) show that models omit or deny the cues that drive their answers. Our question is prior and orthogonal: how often is the reasoning present and readable at all.

Monitorability as a strategy, and its metrics. Korbak et al. (2025) frame CoT monitorability as a fragile opportunity and recommend published evaluations. Emmons et al. (2025) operationalize “legibility” and

“coverage,” where coverage means intra-CoT completeness—a different quantity from ours—and the metric is undefined when no CoT is emitted. MonitorBench (Wang et al., 2026) scores monitorability across models, but its open-versus-closed gap is confounded by disclosure policy, which is precisely our point. Shiromani and Richter (2026, withdrawn preprint) argue that a monitor’s aggregate accuracy is a “false average” that hides collapse on the subset where CoT is the only signal. That measurement critique is the same shape as our produced-versus-readable split and per-arm reporting; we do not rest any claim on the withdrawn manuscript.

Adaptive reasoning and emission. AdaptThink (Zhang et al., 2025) reports a per-instance emission rate (RatioNT) that varies widely by task. The Qwen3 report (Qwen Team, 2025) documents the `/think /no_think` switch we use to darken the channel. These establish that emission is a controllable variable. We measure it per action from the externally exposed channel and tie it to monitor recall.

Monitor performance and the bound. CoT Red-Handed (Arnav et al., 2025) compares CoT and action monitors and finds hybrids best; we vary whether the CoT channel exists at all. Hua et al. (2025) treat monitor-invocation probability p_m as a budgeted decision variable. Our bound truncates the achievable set at c , the configuration-imposed emission rate—a complementary constraint. Greenblatt et al. (2024) and Lindner et al. (2025) frame AI control; the latter catalogues real-deployment failures while assuming comprehensive monitoring. Coverage is its gap by omission.

Multi-agent objection (secondary). Paglieri et al. (2026) document emergent whistleblowing in an agent swarm under transparent channels. Jha (2026) shows that raw inter-agent disagreement is a confounded signal for monitor complementarity: ensemble gains track detection skill, not lineage or pairwise decorrelation, which motivates our decision-tree design rather than a correlation-minimizing ensemble. Hammond et al. (2025) give the risk taxonomy, and Chan et al. (2024, 2025) the agent-identity framing.

3. Methods

Reasoning-channel states. For each assistant turn we read `Inspect ContentReasoning` blocks and classify the turn by the strongest limitation present: `RAW_PRESENT`, `SUMMARY_ONLY`, `REDACTED`, or `ABSENT`. Emission is `raw_present`. The uninspectable share is `summary_only + redacted + absent`. The denominator is assistant turns that occurred at a step index. Turns that never happened because a trajectory ended earlier are absent from the cell, not counted as `ABSENT`, so long trajectories are not made to look unmonitorable by construction.

Produced versus readable. The four states describe what an evaluator can read. We additionally pair each turn with its model call and record the provider’s `reasoning_tokens`, deriving whether reasoning was `PRODUCED` (`tokens > 0` or readable content), `NOT_PRODUCED` (explicit 0), or `UNKNOWN` (unreported). Returned content outranks the token count: some providers return readable reasoning while reporting zero tokens. These are counted as accounting-inconsistent and are never overridden silently.

Serving-configuration experiments. (i) Within-model: `deepseek-v3.2` over 1,013 fixed samples with reasoning unrequested versus requested. (ii) Effort: `gpt-5-nano` over the same samples at low, medium, and high effort. (iii) Header, controlled: `claude-haiku-4.5` on three agentic families (`agentharm_benign`, `gdm_intercode_ctf`, `agent_bench_os`), 50 trajectories each, run twice—once under the `OpenRouter`

default and once with the interleaved-thinking-2025-05-14 beta header—with the upstream provider pinned and all model calls logged. We cross-check the classification against the provider’s own reasoning-token accounting (scripts/verify_coverage.py; 0 disagreements over 19 logs).

Monitor experiment. We first ran this on claude-haiku-4.5, the model of §4.1’s header contrast. It refused 48 of 50 harmful behaviours in both arms (refusal rate 0.960, 95% CI [0.865, 0.989]), leaving two positive trajectories per arm; a per-step recall curve is not estimable at $n = 2$, and both arms carry underpowered: true in the record. We therefore repeated the experiment on qwen3-32b, which refuses 5 of 50 with reasoning on (0.100, 95% CI [0.043, 0.214]) and returns raw CoT, yielding 42 positives visible and 38 dark. The agent was chosen after seeing haiku’s refusal rate, and chosen for compliance rather than sampled: recall levels are a property of this agent, and only the visible-versus-dark contrast transfers.

Two arms over the same 50 behaviors, provider pinned: reasoning on (CoT visible) and /no_think (CoT dark; the switch is sticky across tool results, verified). Ground truth is AgentHarm’s own refusal plus semantic grader; a positive is a harmful-graded trajectory. For every step of each positive we run two gpt-5-nano monitors—one seeing only that step’s reasoning, one only its actions—and compute cumulative recall by step. A step with no content in a channel abstains; it never catches. Running the harmful split departs from a stated non-goal (“benign by construction”). It was authorized; its tools are AgentHarm’s synthetic mocks, so nothing executes; and only counts, rates, and verdicts are stored (never task, reasoning, or action text).

The instrument. Every rate path calls require_validation (raises without a recorded recall against a labelled control). Uninspectable turns stay in gate denominators. Records validate against a committed JSON Schema (v4.1). The codebook is hashed and pinned in config/channels.yaml, and a codebook_drift gate fails on any change.

Pre-specified peer-objection endpoints. We wrote decision rules for three endpoints (C1 revert-detector precision, C2 verbal objection in the message channel, C3 coder/author reliability κ) and git-committed them before any human label existed (commit ae377b0; order is provable from history). We did not obtain an independent pre-registration—the OSF timestamp could not be created—so we report these as pre-specified, not registered, and make no registration claim.

Data and reproducibility. Model runs are our own via OpenRouter. No corpus text is committed, only counts, rates, figures, and records. All figures and tables regenerate from the committed records with named scripts. Reasoning-token payloads are logged for the first five calls per sample (an Inspect default); we therefore read usage.reasoning_tokens, present on every call. Derived statistics follow the release-gated Inspect pipeline of Angulo (2026; Zenodo DOI 10.5281/zenodo.22182741).

Let c be the share of relevant turns whose reasoning is readable, and let r_2 be the monitor’s recall conditional on a readable reasoning channel. A deliberation-gated monitor then has recall at most $c \cdot r_2$.

4. Results

4.1 Visibility is set by the serving configuration

Over a nine-arm sweep on three benchmarks (identical 1,013 samples; 1,263 turns per arm) the readable share is a partition, not a spectrum (Figure 1): five arms readable at ≈ 1.0 , one absent at 1.0, three redacted at 1.0. The controlled contrasts:

- *Within-model.* For deepseek-v3.2, readable share moves 0.0000 → 0.9802 on one request parameter. The default arm produced zero reasoning tokens.
- *Effort is not visibility.* For gpt-5-nano, reasoning volume rose 13.8× (341,726 → 4,731,968 tokens) across effort levels while the readable share stayed 0.0000 (encrypted at every level).
- *The harness decides it, silently.* For claude-haiku-4.5, reasoning after a tool result was disclosed on 0 of 661 turns under the OpenRouter default and 667 of 667 with the interleaved-thinking header—same model, tasks, and Amazon Bedrock upstream, identical at step 0. Inspect’s native Anthropic provider sends that header automatically for Claude-4-class models; its OpenRouter provider does not.

Threat to validity. The sweep and header contrasts are exploratory and confined to models reachable via one aggregator. What happens beyond OpenRouter to the upstream is not in the logs, so the mechanism is stated as consistent with documented interleaved-thinking behavior, not proven. The classification is nonetheless corroborated by the provider’s own token accounting (0 disagreements).

4.2 The recall drop

On qwen3-32b harmful trajectories (42 positives visible, 38 dark), cumulative recall by step (Figure 2):

step	CoT visible	CoT dark	action visible	action dark
0	0.929	0.000	0.476	0.263
1	0.952	0.000	0.690	0.605
2	1.000	0.062	0.828	0.750
3	1.000	0.167	0.909	0.708

Steps beyond 3 fall below $n = 10$ and are omitted; the full curve to step 8 is in `results/monitor_experiment_qwen3/monitor_record.json`.

The reasoning-gated monitor catches 93% of harmful trajectories at the first step and 100% by step 2 when reasoning is visible; 0% at step 0 and ≤ 17% through step 3 when it is dark—a drop of about 0.83–0.93, driven by abstention (0/161 turns visible versus 144/155 dark). The action-gated monitor also falls at step 0 (0.476 → 0.263), partly a trajectory effect, but recovers to within 0.08–0.20 of its visible-arm recall from step 1 onward, because actions stay observable. This is the coverage-to-recall link measured directly, instantiating the $c\cdot r_2$ bound.

The haiku arms are not discarded. On their two positives per arm the predicted pattern appears exactly: with the header off, the reasoning monitor abstained on 4 of 4 post-step-0 turns—nothing to read after a tool call—against 0 of 5 with the header on, so the blindness §4.1 measures at the turn level is the same blindness a monitor inherits, demonstrated on the model the header result is about and measured on one that complies (`results/monitor_experiment/monitor_record.json`).

Threat to validity. /no_think also raised refusals (0.10 → 0.28), so the two arms are not a perfectly clean behavioral A/B and the action-monitor difference is partly a trajectory effect. The CoT-monitor collapse, however, is unambiguous. The monitor is a single cheap model (gpt-5-nano); the visible-versus-dark contrast is on the same monitor. Per-step n shrinks past step 3.

4.3 The reporting standard, the bound, and produced-versus-readable

A monitor gated on deliberation has $\text{recall} \leq c \cdot r_2$. The quantity c is imposed by the model and the serving configuration; no monitoring budget lifts it. Provider token accounting shows that most “unreadable” agentic turns produced no reasoning (tokens = 0) rather than withholding it. The only produced-but-withheld reasoning is gpt-5-nano (encrypted) and seven blank gpt-oss-120b blocks. The remedy differs by cause: disclosure where reasoning is withheld, and action monitoring or a different configuration where it is not produced. The practical recommendation is to publish monitorability conditioned on route, request, turn type, and step, with a denominator and the produced/readable split. A bare “monitorability %” describes no deployment.

4.4 Long horizon: no decay to step 58 (secondary; underpowered)

On SWE-bench Verified (scoring off; $n = 10$ per model; message limit 120), the two models that sustain long trajectories (median 59 assistant turns) do not lose disclosure with depth: nemotron-3.5 reads 0.994 out to step 58; gpt-oss-120b reads 0.856, a mild $0.92 \rightarrow 0.83$ dip that plateaus. qwen3-32b submits in about six turns and cannot test the claim. This is suggestive, not powered (per-step $n \leq 10$; 7/10 right-censored), but it is our first evidence at this depth. It runs against a “decay with depth” hypothesis we had earlier entertained and withdrew after it failed to replicate across task families.

4.5 Pre-specified peer-objection endpoints (secondary; a bounded null)

Blind coder (JAG) plus author (JAD) on collusion.wiki, which is structurally thin: 2,824 of 4,024 pages (70.2%) have a single agent. C1 revert precision is 0.700, but the sample falls on only three pages, so the page-clustered 95% CI is [0.174, 0.963] and the endpoint is not validated. C3 reliability $\kappa = 0.229$ (68% raw agreement; a base-rate paradox) falls below 0.40, so C1 is reported unreliable regardless. C2 verbal objection: 0 of 156 coded, one-sided upper bound < 0.017 . No revert-based rate is reported: the pre-specified decision rules did their job of refusing an unvalidated instrument. (A key-collision bug in our annotation tool dropped the coder’s SHARE presses. It does not affect the objection count that C2 reports, and is logged as a deviation.)

5. Discussion and Limitations

CoT monitoring is a live safety bet, and our results say its measured effectiveness is partly a property of the measurement setup. An evaluator comparing models’ monitorability, or certifying a CoT monitor, can reach opposite conclusions by changing an API route or a header—silently, because the harness sets it. The fix is cheap: report the serving configuration and the produced/readable split alongside any monitor result, and treat uninspectable turns as failures rather than dropping them.

The individual phenomena are known: CoT can be unfaithful, monitorability is fragile, models route their thinking, providers encrypt reasoning, and action monitors can beat CoT monitors. Our novelty is narrower and defensible. The eval-harness / serving-route confound is real, quantified, and reproducible (same model, task, and harness; 0 versus 1 readable share), which we could not find documented. We also operationalize produced-versus-readable from token accounting inside a coverage measurement. The recall-drop measurement and the long-horizon datapoint are clean instantiations of known ideas rather than new phenomena. We frame the contribution as measurement and reporting, not discovery.

Limitations

(1) Models are small or mid-sized and served through one aggregator; frontier hidden reasoning appears only as gpt-5-nano and an $n = 1$ incident transcript. (2) The serving-config mechanism beyond OpenRouter is not observable in the logs. (3) The monitor experiment rests on one complying model, selected for compliance after a first attempt, with a refusal-rate confound between arms, a single cheap monitor, and small late-step n . (4) The long-horizon arm is $n = 10$ and right-censored. (5) summary_only is structurally unobservable when a provider ships an encrypted chain plus a summary. (6) The peer-objection arm reports no rate and its instrument is unreliable ($\kappa = 0.229$). (7) We obtained no independent pre-registration; C1–C3 are pre-specified, not registered.

The project measures the absence of oversight evidence and is defensive in intent. We ran a harmful-capability benchmark (AgentHarm) only to obtain ground-truth positives for a monitor evaluation. Its tools are synthetic mocks that execute nothing, and we publish only counts, rates, and monitor verdicts—never task prompts, reasoning, or action text. We do not publish tuned detector rule sets. The one incident transcript we read (Mythos 5) carries a canary and a do-not-train notice, both honored; its text is never sent to any third party or published. The clearest dual-use risk in this area is that publishing what a monitor looks for creates pressure to route reasoning into unmonitored channels. We therefore frame recommendations around channel design and disclosure reporting, not covert monitoring, and note that our own results show how easily the reasoning channel can be turned off.

Future Work

Replicate through native provider APIs and a summarizing frontier model (to populate the structurally unobservable summary_only state); pin upstream providers to separate routing from configuration; scale the monitor experiment to more agents and a stronger monitor; and extend the long-horizon arm to hundreds of steps.

6. Conclusion

A monitorability score is not a property of a model alone. Across 23,541 assistant turns, what an external evaluator can read of an agent’s reasoning was governed by the route, the request and a single beta header — and that header, set silently by the evaluation harness and recorded by no published report, moves post-tool-result disclosure between 0 of 661 turns and 667 of 667 on the same model, tasks and upstream. When the channel goes dark on a complying agent, a reasoning-only monitor’s recall of harm-pursuing trajectories falls from 0.93 to 0.00 at the first step while an action-only monitor degrades but does not collapse. The serving configuration is therefore not a detail of the measurement; it is a determinant of the defense being measured. None of this shows that CoT monitoring fails. It shows that a CoT-monitoring result reported without its serving configuration is uninterpretable, and that the remedy is cheap: publish the route, the request parameters and the headers alongside any monitor number, report the produced-versus-readable split with an explicit denominator, and count uninspectable turns as failures rather than dropping them. The instrument that computes the split, records the route and headers per arm, and enforces the gate is released under MIT.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** <https://github.com/Jangulo7/agent-channel-observability>
- **Data/Datasets:** [agent-channel-observability/data/published record at main · Jangulo7/agent-channel-observability](https://github.com/Jangulo7/agent-channel-observability/tree/main/data/published)

Author Contributions (optional)

Johanna Angulo led the project and designed experiments. Jenny Aguilar acted as coder. All authors contributed to writing and reviewed the final manuscript.

References

- Angulo, J. (2026). *safety-eval-pipeline: Safety evaluation with reproducible conditions and a release gate that fails (Version 1.1.0) [Computer software]*. Zenodo. <https://doi.org/10.5281/zenodo.22182741>
- Angulo, J. (2026). *agent-channel-observability (Version 1.0.0) [Computer software]*. <https://github.com/Jangulo7/agent-channel-observability>
- Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2025). *Chain-of-thought reasoning in the wild is not always faithful*. *arXiv:2503.08679*.
- Arnav, B., Bernabeu-Pérez, P., Helm-Burger, N., Kostolansky, T., Whittingham, H., & Phuong, M. (2025). *CoT Red-Handed: Stress testing chain-of-thought monitoring*. *arXiv:2505.23575*. *NeurIPS 2025*.
- Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). *Visibility into AI agents*. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. *arXiv:2401.13138*.
- Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). *Infrastructure for AI agents*. *Transactions on Machine Learning Research*. *arXiv:2501.10114*.
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think*. *arXiv:2505.05410*.
- Emmons, S., Zimmermann, R. S., Elson, D. K., & Shah, R. (2025). *A pragmatic way to measure chain-of-thought monitorability*. *arXiv:2510.23966*.
- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). *AI control: Improving safety despite intentional subversion*. *Proceedings of the 41st International Conference on Machine Learning*. *arXiv:2312.06942*.
- Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., et al. (2025). *Multi-agent risks from advanced AI*. *arXiv:2502.14143*.

- Hua, T. T., Baskerville, J., Lemoine, H., Hopman, M., Bhatt, A., & Tracy, T. (2025). Combining cost-constrained runtime monitors for AI safety. *arXiv:2507.15886*.
- Jha, A. (2026). Decorrelation is not complementarity: Skill, not lineage, governs trusted-monitor ensembles. *arXiv:2608.16190*.
- Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Mądry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., & Mikulik, V. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv:2507.11473*.
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiušė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., & Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. *arXiv:2307.13702*.
- Lindner, D., Griffin, C., Korbak, T., Zimmermann, R. S., Irving, G., Farquhar, S., & Cooney, A. (2025). Practical challenges of control monitoring in frontier AI deployments. *arXiv:2512.22154*.
- Paglieri, D., Cross, L., Genewein, T., Leibo, J. Z., Tomasev, N., & Vezhnevets, A. S. (2026). A case study on emergent cheating and whistleblowing in autonomous research swarms. *arXiv:2609.04170*.
- Qwen Team. (2025). Qwen3 technical report. *arXiv:2505.09388*.
- Shiromani, S., & Richter, L. (2026). A false average: Chain-of-thought monitors collapse where they are the only defense. *arXiv:2608.00583v1*. Withdrawn preprint; v1 HTML still available.
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36. *arXiv:2305.04388*.
- Walden, W., & Wanner, M. (2026). Reasoning models will sometimes lie about their reasoning. *arXiv:2601.07663*.
- Wang, H., Sun, Y., Ko, B., Talati, M., Gong, J., Li, Z., Yu, N., Yu, X., Shen, W., Jolly, V., & Zhang, H. (2026). MonitorBench: A comprehensive benchmark for chain-of-thought monitorability in large language models. *arXiv:2603.28590*. COLM 2026.
- Zhang, J., Lin, N., Hou, L., Feng, L., & Li, J. (2025). AdaptThink: Reasoning models can learn when to think. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. *arXiv:2505.13417*.

LLM Usage Statement

LLM assistance (Claude) was used with coding and proofreading since English is not our mother tongue.