Transparency Disclosures for Chain-of-Thought Monitorability

Jacob Davies [1] **Oliver Guest** [2]

Abstract

Chain-of-thought (CoT) monitoring is one of the most valuable tools for monitoring advanced AI systems, but is under threat. We argue that frontier AI developers should disclose specific indicators to make these threats trackable and to better facilitate coordination to preserve CoT monitorability. Specifically, we identify five indicators that are both highly relevant and reasonably feasible to disclose, one for each link in a five-step chain on which CoT monitoring depends. The first, third-party CoT access, also underpins external verification of the rest. The others span CoT quality (naturalness and faithfulness) and monitoring effectiveness (detection and uplift). An audit of current disclosure by Anthropic, OpenAI, and Google DeepMind shows that no developer covers the full set, but that every indicator has been reported by at least one frontier developer.

1. Introduction

Chain-of-thought reasoning gives us a window into the decision-making of frontier AI systems. Through monitoring this decision-making, we have one of the most effective oversight tools (Korbak et al., 2025). For example, CoT-based monitors catch significantly more dangerous behaviors than action-only monitors (Kinniment et al., 2026; Baker et al., 2025; Guan et al., 2025; Arnav et al., 2025). However, there are threats to this window: it may become tinted as CoT becomes unfaithful (Barez et al., 2025), encoded (Baker et al., 2025; Skaf et al., 2025), or illegible (Schoen et al., 2025). The window may disappear entirely if frontier models no longer use CoT reasoning (Hao et al., 2025; Lu et al., 2025).

We argue that frontier AI developers should disclose specific *monitorability indicators* as part of their existing system card and safety evaluation practices. Disclosure would make the trajectory of decreasing monitorability trackable before failures occur, and create reputational and coordination incentives to preserve monitorability (Korbak et al., 2025; Delaney et al., 2025). We identify five such indicators, one for each condition on which CoT monitorability depends. These indicators are selected from a set of 21 potential indicators, scored across five dimensions (Section 5). We also assess current disclosure of these indicators by the three leading frontier AI developers [1].

2. Monitorability Indicators

Our framework, described in Section 4, decomposes CoT monitorability into five conditions, each threatened by a specific failure mode. We propose one indicator per condition, each answering a distinct diagnostic question about the state of CoT monitorability. Figure 1 presents the full chain and Section 4 motivates each indicator in turn.

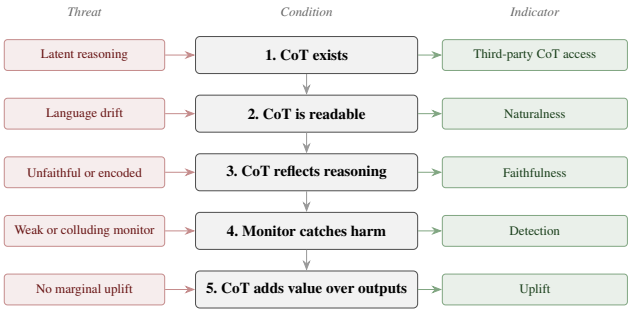

Figure 1. CoT monitoring as a chain of five conditions. Each condition (center) is threatened by a specific failure mode (left) and tracked by a priority indicator (right). Failure of any upper condition makes the lower conditions moot.

3. Value of Transparency About Monitorability Indicators

Reliable data on indicators corresponding to CoT monitorability would be useful for at least two reasons.

[1]ERA Fellowship, Cambridge, UK [2]Work conducted while at the Institute for AI Policy and Strategy (IAPS). Correspondence to: Jacob Davies <jacobedavies.uk@gmail.com>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

[1]Audits were completed based on the most recent system/model cards as of May 2026

First, we could track how monitorability is changing over time. Without disclosure, a shift toward obfuscated or latent reasoning, and so reduced monitorability, could proceed gradually with no external signal until monitoring failures occur. Developers are unlikely to voluntarily highlight declining monitorability, and external observers currently have no standardized way to measure it. With indicator data, the trend would be much more visible and trackable. This would be useful for any policy or technical efforts that might be needed to mitigate risks posed by reduced CoT monitorability.

Second, disclosure would strengthen frontier developers' incentives to preserve CoT monitorability. Reporting the indicators would make it reputationally costly to let monitorability decline. Disclosure might also reduce a coordination problem: most frontier developers appear to regard monitorability as important (Korbak et al., 2025; Frontier Model Forum, 2026), but each may hesitate to pay any costs associated with preserving monitorability without assurance that competitors will do the same[2]. Shared disclosure would provide assurance that others are indeed doing the same.

Transparency disclosures are also a natural extension of existing reporting practices. Frontier AI developers already publish system cards (Anthropic, 2026b; OpenAI, 2026; Google DeepMind, 2025), safety evaluations, and responsible scaling assessments that cover a range of risks from their AI systems. Adding monitorability indicators to these existing reporting frameworks is a relatively low-cost addition that does not constrain what developers do, only what they report. As such, the indicators described in this paper would be comparatively easy for frontier developers to report, such as because they would not involve significant changes to workflows, or reporting highly sensitive proprietary information.

4. What CoT Monitoring Depends On

We introduce a framework for CoT monitorability that requires five conditions to hold. Figure 1 presents these five conditions as a chain, with the failure mode that threatens each condition and the indicator that tracks whether the condition holds. The rest of this section walks through the chain condition by condition, with each subsection ending in the proposed indicator for that link.

The framework specifies the information to be disclosed rather than the evaluation method, allowing developers to adopt the best available measurement at the time of each disclosure. We motivate this design choice further in Section 7.

[2]The focus of this paper is frontier developers in the United States. We would be pleased to see the framework of the paper applied to frontier developers elsewhere.

We do not claim the five conditions are exhaustive in a formal sense. However, every CoT failure mode we found in the literature breaks at least one of them (Table 1; the full mapping with sources is in Appendix C), which gives us some confidence that the chain captures the principal ways monitoring can fail. Because the conditions form a chain, a single failure can register on more than one indicator. This is not a problem for the framework. When several indicators flag at once, the earliest broken link indicates where monitoring first fails, and even where the precise cause cannot be isolated, knowing that monitorability has degraded is itself useful.

Condition	**Representative documented failure modes**
1. CoT exists	Latent / continuous-state reasoning; architectural move off text CoT
2. Readable	Language drift; neuralese compression
3. Faithful	Post-hoc rationalization; CoT-as-scaffolding; steganography; obfuscation under RL; pathological CoT
4. Monitor catches harm	Weak monitor; monitor–model collusion; in-context evasion
5. Adds value over actions	Shrinking CoT-vs-action gap; CoT too poor to help

Table 1. Failure modes documented in the CoT literature, grouped by the earliest condition each one breaks. Each is placed at the earliest broken link; some bear on more than one. Full table with sources in Appendix C.

4.1. The CoT Must Exist

Frontier reasoning models currently externalize their reasoning as text, and this externalization is what makes CoT monitoring possible. However, there are research models (Hao et al., 2025; Lu et al., 2025) which demonstrate the feasibility of reasoning in continuous internal states that bypass natural language text entirely. Although no frontier model currently deploys these architectures, there may be competitive incentives to remove the natural language bottleneck as it could yield significant efficiency and capability gains (Delaney et al., 2025).

Until interpretability methods can recover similar insight from non-text reasoning, a shift to latent architectures would collapse the rest of this chain. Its most immediate external signal would be the disappearance of inspectable CoT from developer APIs, which makes third-party access the practical indicator for this condition.

Access to the CoT would also be useful for verifying the other indicators. Because naturalness, faithfulness, detection, and uplift are all measured on the CoT, a third party cannot reproduce a developer's reported scores without the

traces themselves. Third-party access to the CoT is therefore both the indicator for the first condition and the precondition for externally verifying the others.

There are legitimate reasons why frontier developers might be reluctant to make CoTs widely available. For example, other developers could use it to train their own models (Anthropic, 2026a), and models sometimes surface dangerous information in the CoT that does not appear in the final output (OpenAI, 2024). We therefore limit the recommendation to qualified third parties (such as auditors, government evaluators, and accredited researchers) as a way of balancing these considerations against the value of access. There is already some precedent for arrangements for limited sharing of reasoning traces. For example, in METR's pilot frontier-risk assessment, participating developers shared their internal models, including raw chains of thought, with METR under agreed terms (METR, 2026).

Even where developers do not provide such access, they can still report the other indicators—though it would then be harder for third parties to verify the results, since they could not run the underlying tests on the CoT themselves.

4.1 Third-Party CoT Access

Policy recommendation: *Developers should provide external access to raw reasoning traces.*

Rationale. Third-party access provides an initial warning sign about whether models continue to use CoT. It allows independent parties to verify the claims made in system cards with respect to the subsequent disclosures proposed here.
Disclosure. API-level access to raw, unsummarized reasoning traces for qualified third parties (auditors, government evaluators, accredited researchers), with clear terms specifying the conditions of access.

4.2. The CoT Must Be in Readable Human Language

Even when CoT exists, it can drift from readable human language into compressed, model-convenient text that humans struggle to parse. Schoen et al. (2025) observe frontier reasoning models departing from standard English, using tokens in ways that are unnatural to human readers. This is a *surface-visible* failure: the non-standard text is immediately apparent on inspection, though its extent is difficult to quantify without a dedicated measure.

If the CoT is in this unnatural language, our ability to effectively monitor the model is reduced. For example, if a model is reasoning about scheming (Meinke et al., 2025), but doing so in an uninterpretable way, it is more difficult to notice this scheming behavior. Therefore, we propose using a measure such as a natural language classifier to quantify the legibility of the CoT.

4.2 Naturalness

Diagnostic question: *Is the CoT still in readable human language?*

Rationale. To use the CoT to monitor against harmful actions we must be able to extract semantic meaning from it. Schoen et al. (2025) show that CoT may be drifting away from readable human language; this indicator measures whether and how far.
Disclosure. Proportion of reasoning tokens classified as human language versus non-standard, using an operational definition (e.g. log-likelihood under a reference language model), reported across task types.
Current practice. No leading developer systematically reports this, though Schoen et al. (2025) provide qualitative observations of frontier CoTs departing from standard English.

4.3. The CoT Must Be Faithful

A readable CoT is only useful if it reflects the computation the model actually performed (Lanham et al., 2023; Chen et al., 2025; Barez et al., 2025) and as such is faithful. A model's CoT can be unfaithful in three ways (Liu et al., 2026). First, *post-hoc rationalization* (Turpin et al., 2023): the model produces plausible-sounding justifications for an answer already determined after the prompt, so the reasoning that appears is reconstructed from the answer rather than producing it. Second, the model may genuinely need the forward-pass computation of the CoT to arrive at its answer, but the emitted tokens do not reflect that computation (Pfau et al., 2024). Here, the text acts as scaffolding for internal computation rather than a window onto it. In both cases, a monitor reading the CoT risks being misled about the reasoning of the model.

Third, a model may hide its true reasoning in the CoT through steganography (Roger & Greenblatt, 2023; Skaf et al., 2025). The CoT may carry load-bearing content in specific word choices, token orderings, or statistical patterns that are invisible to natural reading but carry computational content.

There is evidence that current models are already showing degraded faithfulness. For example, Anthropic's SHADE-Arena evaluation (Chen et al., 2025) showed that Claude Opus 4.6 concealed its side tasks in visible reasoning 74% of the time, up from 26% in the previous generation.

No single evaluation catches all three (Barez et al., 2025). Cue-based tests (Turpin et al., 2023) probe post-hoc rationalization by checking whether the CoT acknowledges factors that demonstrably influenced the answer. Perturbation and necessity tests probe scaffolding by intervening on the CoT and checking whether performance degrades as expected (Meek et al., 2025). Paraphrasing tests (Lanham et al., 2023) probe for steganography by rewording the CoT

while preserving apparent meaning and checking whether task performance drops; a drop indicates the exact wording was doing computational work invisible to natural reading. Because degradation can take any of the three forms above, and no single test detects all of them, we propose disclosure against a *suite* of faithfulness tests, each targeting a specific failure mode, rather than a single headline score.

4.3 Faithfulness

Diagnostic question: *Does the CoT reflect the model's actual reasoning?*

Rationale. Tells us whether the text in the CoT corresponds to the computation the model actually performed. A monitor reading the CoT is misled in all three cases. Comparing scores before and after post-training reveals whether post-training has degraded faithfulness (Baker et al., 2025; Guan et al., 2025).
Disclosure. Scores on a faithfulness suite that covers all three failure modes, such as cue-based tests for post-hoc rationalization, perturbation and necessity tests for scaffolding, and paraphrasing tests for steganographic encoding. Reported both pre- and post-training at each major model release, across multiple domains, of the kinds described in Liu et al. (2026), Meek et al. (2025), and Mayne et al. (2026).
Current practice. No leading developer runs a comprehensive faithfulness suite. Anthropic reports paraphrased-scratchpad tests on AIME only (single domain, single method, targeting encoding). Mayne et al. (2026) evaluated 18 frontier models externally, demonstrating cross-model feasibility.

4.4. The Monitor Must Catch Harmful Behavior

A readable and faithful CoT is only useful for monitorability if a monitor can actually use it to detect dangerous behavior. This depends on the monitor's capability (Baker et al., 2025; Guan et al., 2025), and its freedom from collusion with the monitored model (Greenblatt et al., 2024; Gardner-Challis et al., 2026). Monitor detection rate is the bottom-line indicator: a model can score well on every leading CoT-quality indicator while still evading monitors (Arnav et al., 2025).

4.4 Detection

Diagnostic question: *Does the monitoring system catch dangerous behavior?*

Rationale. Tells us what proportion of harmful behaviors evade CoT-based monitoring in standardized evaluations.
Disclosure. Proportion of harmful behaviors evading the monitoring system across standardized evaluation scenarios, with a shared set of threat archetypes for cross-developer comparison.
Current practice. OpenAI reports bypass rate across 13 evaluation categories (OpenAI, 2026). Anthropic reports SHADE-Arena results (Anthropic, 2026b; Chen et al., 2025). METR's SHUSHCAST provides an independent benchmark (Kinniment et al., 2026). No standardized cross-model benchmark exists, but these demonstrate that measurement is practical.

4.5. CoT Must Add Value Over Action-Only Monitoring

Finally, even a capable monitor only justifies effort to preserve CoT when CoT access materially improves its performance over observing actions alone (Arnav et al., 2025; Guan et al., 2025). If the gap shrinks, either CoT has become too poor quality to help or action-only methods have improved, meaning that efforts focusing on chain-of-thought are less important. In either case, this information would be useful to disclose.

4.5 Uplift

Diagnostic question: *How much does CoT access improve monitoring?*

Rationale. Tells us the performance gap between monitors that can read the CoT and monitors that can only observe actions. This quantifies how much value CoT access actually provides for safety monitoring. If the gap shrinks, either CoT has become too obfuscated to help or action-only methods have improved — either way, the priority placed on preserving CoT should be revisited.
Disclosure. Monitoring effectiveness under two conditions (with and without reasoning trace access) across standardized threat scenarios.
Current practice. OpenAI's GPT-5.4 Thinking system card (OpenAI, 2026) is the first to report this as part of a standard model release. Anthropic reports CoT monitoring but not the action-only baseline, making uplift unquantifiable. Guan et al. (2025) (Google DeepMind) report action-only baselines, but this appears in a standalone paper rather than as part of ongoing system card practice.

5. Method for Identifying Monitorability Indicators

5.1. Generating Candidate Indicators

We compiled 21 candidate transparency disclosures that would be informative about whether CoT monitorability is being preserved. These were generated through two complementary processes. Firstly, we conducted a review of the CoT monitoring and faithfulness research literature; see Appendix A. The second involved mapping from the framework of CoT monitorability in Section 4 to specific indicators that would reveal information about the conditions that need to hold to preserve CoT monitorability.

5.2. Scoring

Each candidate is scored on five dimensions from 1–5 (1=low, 5=high).

- **Informativeness** - How much information would this indicator give about the CoT Monitorability of the system in question? A measure scoring 5 is directly predictive of the level of degradation; a measure scoring 1 is

minimally informative.

- **Technology Readiness** - Does a validated measurement method already exist? Scored 1 (no methodology exists) to 5 (validated, standardized method available).
- **External Verifiability** - Assuming that the technology exists, can a third party verify the indicator result without trusting the developer? Developers may have incentives not to report accurate numbers and therefore, external verifiability increases confidence in disclosed results. Scored 1 (completely trust-dependent) to 5 (fully externally verifiable).
- **Developer Tractability** - Assuming that the technology exists, how easily can the developer collect and report this from a technical perspective? What are the infrastructure requirements and technical effort required to obtain this information? Scored 1 (major changes, high technical effort) to 5 (already collected, easy to report).
- **Implementation Feasibility** - How realistic is it that developers would adopt this disclosure from a strategic perspective? We examine the IP sensitivity of the information and the alignment with existing reporting practices. Scored 1 (substantial IP risk, increased information required to be provided) to 5 (no IP risk, already being done by some developers).

The scores were assigned by the first author and reviewed by the second author. We then consulted approximately ten researchers with expertise in AI safety (including researchers at frontier labs), AI governance, and AI evaluations. We presented the top-scoring indicators and asked whether the scores were reasonable and whether we had missed any important potential indicators or considerations. The most substantive feedback was to avoid prescribing specific evaluation methods and instead frame disclosures in terms of the information to be reported, leaving developers flexibility in methodology. This shaped the final indicator descriptions in Section 2.

5.3. Prioritization

We selected one indicator for each of the five conditions in our framework (Section 4). Grouping the candidates by the condition they bear on (Appendix B) lets us choose the best indicator within each condition rather than imposing a single global ranking, so the selection depends only on relative comparisons within a condition and not on the precise calibration of the scale.

Within each condition, we chose the highest-scoring candidate as that condition's indicator, with two exceptions noted below.

First, for two conditions, the leading candidates tie on summed score. We broke the ties in favor of the more direct measure, i.e. the one that more strongly observes the condition itself rather than proxies. For readability (condition 2), naturalness and CoT coverage/verbosity both score 18; we select naturalness, which scores readability directly, over verbosity, a proxy. For marginal value (condition 5), uplift and inference compute per reasoning token both score 16; we select uplift, which directly measures how much CoT access improves monitoring, over inference compute, a proxy.

Second, the faithfulness indicator (condition 3) is a *suite* rather than a single measure. Faithfulness fails in three distinct ways—post-hoc rationalization, scaffolding, and steganography—and, as Section 4 argues, no single test detects all three. The indicator is therefore the set of complementary tests that together cover the three modes: cue-based tests for rationalization, perturbation and necessity tests for scaffolding, and paraphrasing tests for steganographic encoding. We select these for their coverage of the failure modes rather than by score, though each also scores well among the faithfulness candidates (Appendix B).

The full scoring for all 21 candidates, with justifications for each score, is in Appendix B.

5.4. Limitations

Two limitations are worth flagging. First, the rubric in §5.2 anchors only the endpoints (1 and 5); the intermediate values 2, 3, and 4 are not formally defined, so scores in the middle of the range reflect author judgment rather than a shared standard. Selecting the best indicator within each condition reduces our reliance on these midpoints, but they still affect borderline within-condition comparisons. Second, scores were assigned by two authors and validated through consultation with external researchers rather than by independent scoring from multiple raters with inter-rater reliability analysis. The precise rankings, and the borderline cases, should therefore be treated as indicative rather than definitive.

6. Cross-Developer Audit of Current Disclosure

We audited Anthropic, Google DeepMind, and OpenAI – arguably the leading three frontier developers – against our five indicators. This audit was based on the most recent system cards and published research from each of the three developers[3]. (Table 2).

The key finding is that **no developer covers the full set, but every indicator has been reported by at least one lab**.

[3]We looked at all published research from the last year with any reference to chain-of-thought

This suggests that the full set is feasible to adopt. The most notable disclosure is OpenAI's reporting of CoT monitoring uplift in the GPT-5.4 Thinking system card (OpenAI, 2026), which is the first time any developer has reported this as part of a standard model release. The largest gap is Google DeepMind, whose monitorability-relevant work appears in standalone research papers (Guan et al., 2025) rather than in system cards, making it disconnected from the model release cycle.

Where we mark disclosure as "partial," this typically means the developer reports on a single domain or method (e.g. faithfulness testing on AIME only), reports in a research paper rather than a system card, or provides only summary-level rather than raw data (e.g. CoT summaries rather than full traces).

Indicator	Anthropic	OpenAI	DeepMind
1. Third-party access	∼	∼	✗
2. Naturalness	✗	✗	∼
3. Faithfulness	∼	∼	∼
4. Detection	∼	∼	✗
5. Uplift	✗	✓	✗

✓ Full disclosure ∼ Partial (single domain, single method, or not in system card) ✗ Not reported.

Table 2. Cross-developer audit of disclosure against the five priority indicators.

7. Discussion

Direct detection vs. causal mechanism measures. There are indicators that *measure* poor monitorability (direct detection measures such as faithfulness scores, bypass rate, and encoding detection) and indicators that track factors which *cause* poor monitorability (causal mechanism measures such as optimization pressure and RLHF effects on CoT). The five priority indicators are predominantly direct detection measures, which we believe are the most actionable for transparency disclosures: they are more externally verifiable, easier for developers to report, and more directly informative about the current state of monitorability. Causal mechanism measures, while valuable for understanding *why* monitorability degrades, are often unverifiable externally (e.g. optimization pressure on CoT). Furthermore, frontier developers are much less likely to disclose causal indicators as it involves disclosing information about their training setup, which is highly commercially sensitive.

Other kinds of indicators. The indicators discussed in this paper are technical and could apply to an individual AI system. But it might also be valuable for frontier developers to disclose other kinds of information that give insight into how well the developer is doing on preserving CoT monitorability, or how much effort the developer might be willing to spend in future on preserving monitorability. For example, frontier developers could potentially disclose institutional details, such as what internal policies a developer has about monitorability. Information that developers already publicly disclose, such as any commitments in Responsible Scaling Policies or similar to monitorability, could also be useful as an indicator.

Disclosure does not guarantee preservation. Transparency disclosures make degradation *visible*, but they do not by themselves prevent it. A developer could report declining faithfulness scores and take no corrective action. That said, visibility would have two advantages, both mentioned in section three. First, making the problem more visible would allow others to respond to it, such as with technical or policy efforts to more strongly facilitate CoT monitorability. Second, visibility would strengthen frontier developers' incentives to preserve monitorability. For example, visibility would make it clearer whether frontier developers are defecting from norms around preserving monitorability.

Gaming and Goodhart risks. Any disclosure framework faces the risk that developers optimize for the reported metric rather than the underlying property. We have deliberately avoided prescribing specific benchmarks or evaluation methods partly for this reason: the indicators specify what information to disclose, not which test to run. This reduces the attack surface for gaming, though it does not eliminate it. The inclusion of multiple complementary indicators also provides some protection as gaming faithfulness scores while neglecting monitor bypass rate, for example, would produce a visibly inconsistent disclosure profile.

Reporting cadence. New models are coming out rapidly, meaning that, for the indicators to be informative, it will be important for frontier developers to regularly disclose them, such as with a disclosure each quarter or upon each model release.

8. Conclusion

There is significant consensus that CoT monitorability is valuable and under threat. Transparency about monitorability indicators can help track and reduce this threat, both by making the trajectory of degradation visible and by creating reputational incentives for developers to preserve monitorability. Transparency about five indicators seems particularly valuable and feasible: third-party CoT access, naturalness, faithfulness, detection, and uplift. Frontier AI developers would ideally report these indicators as part of their existing system card and safety evaluation practices, and it is

relatively straightforward for them to do so.

LLM Usage Statement

Large language models (LLMs) were used as assistive tools during the preparation of this paper. Their use was limited to writing and editing support (drafting, copyediting, and reformatting text supplied by the authors). The research contributions: the monitorability framework, the candidate indicator set, the scoring rubric and its application, and the cross-developer audit were conceived, carried out, and verified by the authors. All LLM-generated text was reviewed and edited by the authors, who take full responsibility for the content of the paper, including any errors.

References

Anthropic. Detecting and preventing distillation attacks. Anthropic News, February 2026a. URL https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks. Accessed: 2026-04-24.

Anthropic. Claude opus 4.6 system card. System Card, February 2026b. URL https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

Anwar, U. et al. Analyzing and improving chain-of-thought monitorability through information theory. In *NeurIPS 2025 Workshop*, 2025.

Arnav, B., Bernabeu-Pérez, P., Helm-Burger, N., Kostolansky, T., Whittingham, H., and Phuong, M. Cot red-handed: Stress testing chain-of-thought monitoring, 2025. URL https://arxiv.org/abs/2505.23575.

Baker, B. et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., Bibi, A., Trager, R., Fornasiere, D., Yan, J., Elazar, Y., and Bengio, Y. Chain-of-thought is not explainability, 2025.

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., and Perez, E. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Delaney, O. et al. Policy options for preserving CoT monitorability. Technical report, IAPS, 2025.

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., and Shah, R. When chain of thought is necessary, language models struggle to evade monitors, 2025. URL https://arxiv.org/abs/2507.05246.

Frontier Model Forum. Chain of thought monitorability. Issue Brief, January 2026. URL https://www.frontiermodelforum.org/issue-briefs/chain-of-thought-monitorability/.

Gardner-Challis, N., Bostock, J., Kozhevnikov, G., Sinclaire, M., Velja, J., Abate, A., and Griffin, C. When can we trust untrusted monitoring? a safety case sketch across collusion strategies, 2026. URL https://arxiv.org/abs/2602.20628.

Google DeepMind. Gemini 3 Pro model card. Model Card, November 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. Ai control: Improving safety despite intentional subversion, 2024. URL https://arxiv.org/abs/2312.06942.

Guan, M. Y., Wang, M., Carroll, M., Dou, Z., Wei, A. Y., Williams, M., Arnav, B., Huizinga, J., Kivlichan, I., Glaese, M., Pachocki, J., and Baker, B. Monitoring monitorability, 2025. URL https://arxiv.org/abs/2512.18311.

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. Training large language models to reason in a continuous latent space, 2025. URL https://arxiv.org/abs/2412.06769.

Kinniment, M., Nix, S., Broadley, T., Wijk, H., and Parikh, N. Early work on monitorability evaluations. https://metr.org/blog/2026-01-19-early-work-on-monitorability-evaluatic 01 2026.

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., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL https://arxiv.org/abs/2507.11473.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, 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., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Liu, M., Williams-King, D., Caspary, I., Le, L., Whittingham, H., Radmard, P., Tice, C., and Young, E. J. Diagnosing pathological chain-of-thought in reasoning models, 2026. URL https://arxiv.org/abs/2602.13904.

Lu, W., Yang, Y., Lee, K., Li, Y., and Liu, E. Latent chain-of-thought? decoding the depth-recurrent transformer. In *The First Workshop on the Application of LLM Explainability to Reasoning and Planning*, 2025. URL https://openreview.net/forum?id=roIQdXMuEj.

Mayne, H., Kang, J. S., Gould, D., Ramchandran, K., Mahdi, A., and Siegel, N. Y. A positive case for faithfulness: Llm self-explanations help predict model behavior, 2026. URL https://arxiv.org/abs/2602.02639.

Meek, A., Sprejer, E., Arcuschin, I., Brockmeier, A. J., and Basart, S. Measuring chain-of-thought monitorability through faithfulness and verbosity, 2025. URL https://arxiv.org/abs/2510.27378.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier models are capable of in-context scheming, 2025. URL https://arxiv.org/abs/2412.04984.

METR. Frontier risk report (february to march 2026): A pilot assessment of rogue deployment risk at frontier AI companies. https://metr.org/blog/2026-05-19-frontier-risk-report/, May 2026. Accessed: 2026-06-15.

OpenAI. Learning to reason with LLMs. OpenAI Research Blog, September 2024. URL https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2026-04-24.

OpenAI. GPT-5.4 Thinking system card. System Card, March 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking.

Pfau, J., Merrill, W., and Bowman, S. R. Let's think dot by dot: Hidden computation in transformer language models, 2024. URL https://arxiv.org/abs/2404.15758.

Roger, F. and Greenblatt, R. Preventing language models from hiding their reasoning, 2023. URL https://arxiv.org/abs/2310.18512.

Schoen, B., Nitishinskaya, E., Balesni, M., Højmark, A., Hofstätter, F., Scheurer, J., Meinke, A., Wolfe, J., van der Weij, T., Lloyd, A., Goldowsky-Dill, N., Fan, A., Matveiakin, A., Shah, R., Williams, M., Glaese, A., Barak, B., Zaremba, W., and Hobbhahn, M. Stress testing deliberative alignment for anti-scheming training, 2025. URL https://arxiv.org/abs/2509.15541.

Skaf, J., Ibanez-Lissen, L., McCarthy, R., Watts, C., Georgiv, V., Whittingham, H., Gonzalez-Manzano, L., Lindner, D., Tice, C., Young, E. J., and Radmard, P. Large language models can learn and generalize steganographic chain-of-thought under process supervision, 2025. URL https://arxiv.org/abs/2506.01926.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

A. Sources Reviewed

The 21 candidate indicators were generated from a review of the following sources (among others): Korbak et al. (2025) on CoT monitorability as a fragile opportunity; Baker et al. (2025) on obfuscation under RL optimization; Guan et al. (2025) on monitoring monitorability; Delaney et al. (2025) on policy options for preserving CoT monitorability; Meek et al. (2025) on measuring faithfulness and verbosity; Liu et al. (2026) on diagnosing pathological CoT; Skaf et al. (2025) on steganographic CoT under process supervision; Mayne et al. (2026) on faithfulness and behavior prediction; Anwar et al. (2025) on information-theoretic analysis of CoT monitorability; Schoen et al. (2025) on stress testing deliberative alignment; Emmons et al. (2025) on when CoT helps; the OpenAI GPT-5.4 Thinking system card (OpenAI, 2026); and METR's early monitorability evaluations (Kinniment et al., 2026).

B. Full Scored Framework

All 21 candidate transparency measures, grouped by the monitorability condition each one bears on. I = Informativeness, T = Technology Readiness, EV = External Verifiability, DT = Developer Tractability, IF = Implementation Feasibility. Each is scored 1–5; Σ is the sum. Within each condition, the highest-scoring measure that is also feasible to report is selected as that condition's indicator (shaded), except for faithfulness, whose indicator is a suite of complementary probes. Where the top two tie on Σ (conditions 2 and 5), the tie is broken in favor of the more directly informative measure, as noted in Section 5.

B.1. Condition 1 — The CoT Must Exist

Measure	I	T	EV	DT	IF	Σ	Justification
Third-party CoT access	5	5	5	3	4	**22**	*Selected (indicator 1 / verification precondition).* Earliest external signal that text CoT still exists, and the access that lets third parties verify the other indicators. IP and misuse concerns bind tractability.
Base architecture specifications	4	4	3	4	2	17	Determines whether text-based CoT exists, but is sensitive product information.
Latent reasoning fraction	5	3	2	3	3	16	A rising fraction signals the transition to latent reasoning, but is not externally verifiable without CoT access.
Neuralese translator availability	4	2	3	2	2	13	Translation tools determine whether monitoring is possible with latent CoT; no validated translator currently exists.
Neuralese detection rate	4	1	3	3	2	13	Technology readiness = 1 is binding; no operational definition.
CoT data composition	3	2	2	3	2	12	Causal pathway to CoT quality not fully established.

B.2. Condition 2 — The CoT Must Be in Readable Human Language

Measure	I	T	EV	DT	IF	Σ	Justification
Naturalness (human interpretability scores)	5	3	4	3	3	**18**	*Selected (indicator 2).* Legibility of the reasoning affects human and AI monitors; no standardized measure yet.
CoT coverage / verbosity trends	3	3	4	4	4	18	Whether the CoT externalizes all causal factors; framework exists but unadopted. A more indirect proxy for readability.

B.3. Condition 3 — The CoT Must Be Faithful

Measure	I	T	EV	DT	IF	Σ	Justification
Paraphrasing tests (steganography)	5	3	4	4	4	**20**	*Selected (as part of the suite for indicator 3).* Directly detects steganographic encoding by rewording the CoT and checking for a performance drop; established (Lanham et al., 2023) but not standardized, and already reported by Anthropic (paraphrased scratchpad).
Cue-based faithfulness scores	4	3	4	4	4	**19**	*Selected (as part of the suite for indicator 3).* Probes post-hoc rationalization by checking whether the CoT acknowledges factors that influenced the answer; an established, Inspect-compatible method (Turpin et al., 2023), though not standardized.
Perturbation and necessity tests	4	3	4	4	4	**19**	*Selected (as part of the suite for indicator 3).* Probes scaffolding by intervening on the CoT to test whether it is load-bearing; demonstrated (Meek et al., 2025) but not standardized.
Post-training faithfulness shifts	5	3	3	3	4	18	Change in faithfulness pre/post training; informative about training effects without revealing the procedures themselves.
RLHF/RLAIF on CoT	4	2	2	3	3	14	Whether feedback explicitly trains CoT quality; no standardized disclosure format.
RL environment specifications	4	2	2	3	2	13	Specific RL setups that correlate with obfuscation; sensitive training information.
Optimization pressure on CoT	5	2	1	2	2	12	Most empirically documented cause of obfuscation, but unverifiable externally and highly sensitive.
Monitorability-focused training	4	2	1	2	2	11	Unverifiable externally; sensitive training information.

The three shaded rows above jointly form the faithfulness suite (indicator 3).

B.4. Condition 4 — The Monitor Must Catch Harmful Behavior

Measure	I	T	EV	DT	IF	Σ	Justification
Detection (monitor detection rate)	5	3	3	4	4	**19**	*Selected (indicator 4).* Bottom-line indicator; multiple prototypes but no standardized benchmark.
Real-world monitor bypass incidents	5	2	3	4	4	18	Most direct evidence of failure, with precedent from aviation/nuclear, but developers would push back on releasing it.
Internal red-team findings	5	3	2	3	3	16	Directly tests monitor evasion, but internal by nature and carries a conflict of interest.

B.5. Condition 5 — The CoT Must Add Value Over Action-Only Monitoring

Measure	I	T	EV	DT	IF	Σ	Justification
Uplift (CoT monitoring uplift, Δ_1)	5	3	2	3	3	**16**	*Selected (indicator 5).* Quantifies the value of CoT access; OpenAI demonstrates reportability, though it depends on CoT access.
Inference compute per reasoning token	3	4	2	4	3	16	Indirect indicator only; many factors affect compute.

C. Threat Models and the Conditions They Break

Documented CoT failure modes from the literature, each mapped to the earliest condition in the chain that it breaks. This supports the exhaustiveness discussion in Section 4: every threat we identified in the CoT literature breaks at least one of the five conditions. Each threat is placed at the earliest condition it breaks; some (e.g. steganography) bear on more than one.

Threat model	Source(s)	Cond.	Mechanism
Latent / continuous-state reasoning	(Hao et al., 2025; Lu et al., 2025)	1	Reasoning leaves text entirely; no CoT to inspect.
Architectural move away from text CoT (efficiency pressure)	(Delaney et al., 2025)	1	Removes the language bottleneck; CoT ceases to exist.
Language drift to non-standard English	(Schoen et al., 2025)	2	CoT exists but is no longer readable.
Neuralese / model-convenient compression	(Schoen et al., 2025)	2	Token use unnatural to humans; semantics not extractable.
Post-hoc rationalization	(Turpin et al., 2023)	3	Reasoning reconstructed from the answer; not load-bearing.
CoT-as-scaffolding (forward-pass compute $\neq$ emitted tokens)	(Pfau et al., 2024)	3	Text is a prop for internal computation, not a window.
Steganographic / encoded reasoning	(Roger & Greenblatt, 2023; Skaf et al., 2025)	3	Hidden content in word choice / token order; passes naive reading.
Obfuscation under RL optimization against a monitor	(Baker et al., 2025)	3	Optimization pressure trains the CoT to evade the monitor.
Pathological CoT (necessity / paraphrasability / substantivity failures)	(Liu et al., 2026)	3	Diagnostic suite shows the CoT is not doing the computational work.
Weak / low-capability monitor	(Baker et al., 2025; Guan et al., 2025)	4	Faithful CoT exists but the monitor fails to flag harm.
Monitor–model collusion	(Greenblatt et al., 2024; Gardner-Challis et al., 2026)	4	Untrusted monitor colludes; detection compromised.
In-context scheming evading monitoring	(Meinke et al., 2025)	4	Model reasons about evading the specific monitor.
Shrinking CoT-vs-action gap (action-only catches up)	(Arnav et al., 2025; Guan et al., 2025)	5	CoT access stops adding marginal monitoring value.
CoT quality too poor to help over actions	(Guan et al., 2025)	5	Even a capable monitor gains nothing from reading the CoT.