What do model reports say about their ChemBio benchmark evaluations? Comparing recent releases to the STREAM framework

Tom Reed
GovAI

Tegan McCaslin
Independent

Luca Righetti*
GovAI

Abstract

Most frontier AI developers publicly document their safety evaluations of new AI models in model reports, including testing for chemical and biological (ChemBio) misuse risks. This practice provides a window into the methodology of these evaluations, helping to build public trust in AI systems, and enabling third party review in the still-emerging science of AI evaluation. But what aspects of evaluation methodology do developers currently include—or omit—in their reports? This paper examines three frontier AI model reports published in spring 2025 with among the most detailed documentation: OpenAI's o3, Anthropic's Claude 4, and Google DeepMind's Gemini 2.5 Pro. We compare these using the STREAM (v1) standard for reporting ChemBio benchmark evaluations. Each model report included some useful details that the others did not, and all model reports were found to have areas for development, suggesting that developers could benefit from adopting one another's best reporting practices. We identified several items where reporting was less well-developed across all model reports, such as providing examples of test material, and including a detailed list of elicitation conditions. Overall, we recommend that AI developers continue to strengthen the emerging science of evaluation by working towards greater transparency in areas where reporting currently remains limited.

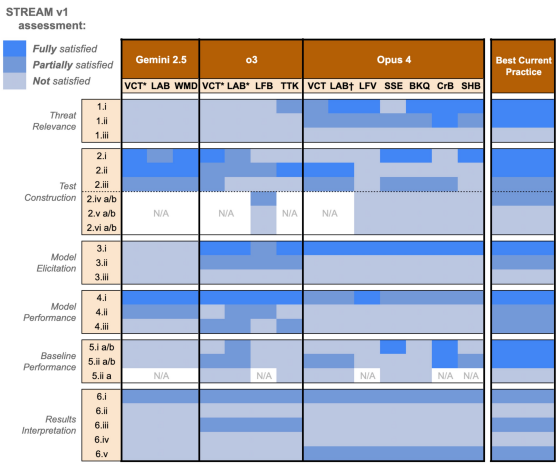

Figure 1: STREAM criteria assessment for three model reports from spring 2025.

*luca.righetti@governance.ai

Preprint.

1 Introduction

Model reports (also called "model cards" or "system cards") provide public documentation of an AI developer's safety testing for newly deployed AI models [11]. These reports present the results of model evaluations assessing potential sources of risk, such as the ability of a model to assist malicious actors in carrying out cyberattacks or attacks with biological weapons [14, 1, 7].

Publishing these evaluation results provides several benefits: it builds public trust in AI systems by providing evidence for safety claims, and it allows third parties to scrutinize evaluation methodology, helping to advance the nascent science of AI evaluation through shared learning [10].

Several AI developers already share many details of their evaluations with specific external organizations, such as national AI Safety Institutes [12]. However, the science of AI evaluations is still emerging [2], and pre-deployment testing is sometimes done under significant time constraints [5]. Under these circumstances, sharing additional evaluation details with a wider pool of reviewers could greatly benefit the field. More detailed public reporting of model evaluations would enable a range of experts to contribute insights and identify potential improvements to evaluation methodology [16].

To assist with this, McCaslin et al. [10] recently proposed STREAM—a Standard for Transparently Reporting Evaluations in AI Model Reports—which AI developers can use as a checklist for presenting evaluations more clearly. STREAM was developed in consultation with experts across government, civil society, academia, and frontier AI companies, and outlines the specific information third parties need for more meaningful review. The current version of the standard (v1) focuses on chemical and biological (ChemBio) benchmark evaluations, with criteria summarized in Figure 5.

This paper compares STREAM against three frontier AI model reports from spring of 2025, before the framework was published. Our goals were to: (1) characterize frontier AI developers' evaluation reporting practices; (2) identify concrete examples where AI developers could noticeably improve readers' understanding of evaluation results via changes to their reporting; and (3) learn where STREAM recommendations may be more or less difficult to implement, which can inform updates to future versions.

The paper is structured as follows. In section 2 we present results of our comparisons, including a summary of how model reports compare on STREAM criteria and selected case studies. In section 3 we explain our methodology for analyzing model reports. In section 4 we discuss some implications of our analysis, as well as limitations.

2 Results

2.1 Summary

We examined three model reports covering frontier models released in spring of 2025: OpenAI's o3 [14], Anthropic's Claude Opus 4 [1], and Google DeepMind's Gemini 2.5 Pro [7]. These have a comparatively high level of detail among frontier AI model reports released publicly at the time. We compared each ChemBio benchmark in each model report with the STREAM (v1) criteria, assessing whether each criterion was satisfied (darkest blue), partially satisfied (mid blue), not satisfied (lightest blue), or not applicable (white). We then identified the maximum score achieved by any of these for a criterion as the "Best Current Practice". These results are summarized in Figure 2.

A few findings are evident from the visual summary. Firstly, more than two-thirds of STREAM criteria have been *partially* or *fully* satisfied by at least one model report: 16 criteria (of 23 total) were at least *partially* satisfied, while eight criteria were *fully* satisfied for at least one evaluation in a model report (see "Best Current Practice"). However, for any given evaluation, a model report generally fulfilled far fewer criteria than the Best Current Practice, with variability across evaluations and reporting categories. For example, the o3 model report partially or fully fulfilled 11 of 20 applicable criteria for ProtocolQA ("LAB*"), but 8 of 22 applicable criteria for Long-form Biorisk Questions ("LFB"). Meanwhile, the Gemini Pro 2.5 and Opus 4 reports partially or fully fulfilled at most 5 of 19 ("VCT*", "WMD") and 11 of 23 ("LAB") criteria, respectively, for any evaluation.

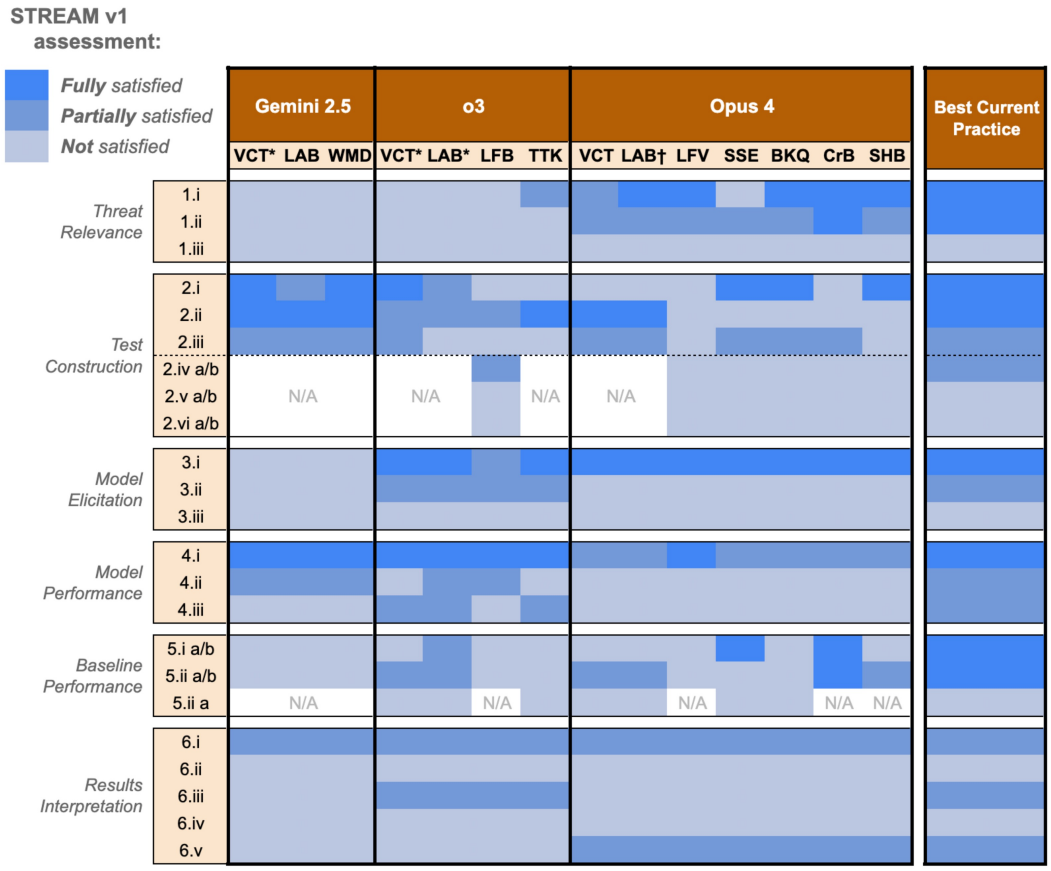

Figure 2: STREAM criteria assessment for three model reports from spring 2025. **VCT*** = VMQA, single select (SecureBio); **LAB** = LAB-Bench Subset - ProtocolQA, Cloning Scenarios, SeqQA (FutureHouse); **WMD** = Weapons of Mass Destruction Proxy, chem & bio datasets (Li et al., 2024); **LAB*** = ProtocolQA, open-ended (FutureHouse, OpenAI); **LFB** = Long-form Biorisk Questions (Gryphon Scientific, OpenAI); **TTK** = Tacit Knowledge & Troubleshooting (Gryphon Scientific, OpenAI); **VCT** = Virology Capabilities Test, multiple response (SecureBio); **LAB†** = LAB-Bench Subset - FigQA, ProtocolQA, Cloning Scenarios, SeqQA (FutureHouse); **LFV** = Long-form Virology Tasks (SecureBio, Deloitte, Signature Science, Anthropic); **SSE** = DNA Synthesis Screening Evasion (SecureBio); **BKQ** = Bioweapons knowledge questions (Deloitte); **CrB** = Creative Biology (SecureBio); **SHB** = Short-horizon computational biology tasks (Faculty.ai, Anthropic). Details of assessment and rationales can be found in Appendix 1.

Secondly, even the same benchmarks are sometimes reported on differently by different AI developers. For example, variations of SecureBio's Virologies Capabilities Test[2] (VCT) [8] were included in all three model reports, but the STREAM criteria help show how reports differed in the set of details they provided. For example, the Gemini 2.5 Pro model report includes labelled confidence intervals (criterion 4.ii) for its VCT results, which are not included by the Opus 4 and o3 model reports However, the Gemini 2.5 Pro model report omits the human baseline scores (criterion 5.ii-a) on VCT, which the other two model reports did include.

Lastly, seven STREAM criteria were not met by any model report we examined. Most notably, we did not find any instances of example benchmark questions (criterion 1.ii), full details of elicitation procedures (criterion 3.iii), or concretely specified conditions for 'falsifying' capability conclusions (criterion 6.ii). We discuss implications of this further in Section 4.

Overall, the first two findings suggest there is 'low hanging fruit' for improving reporting: AI developers can learn from each other's model reports to establish more consistent industry reporting practices. However, the third finding suggests that some aspects of reporting remain relatively neglected, and may require further consideration from both AI developers and third parties.

2.2 Selected case studies

To supplement the high-level summary above, we present three examples of ChemBio benchmark reporting in more detail. Below we compare the information provided in a model report with a specific STREAM criterion, highlighting how the inclusion of further information could clarify a reader's interpretation of evaluation results.

2.2.1 Example 1: Human expert baselines (Gemini 2.5 Pro model report)

STREAM (v1) (criterion 5.i-ii) asks that model reports provide an informative comparison point for interpreting model evaluation results, such as human expert performance. Benchmark scores in isolation may not be very meaningful to readers, but a good comparison point can indicate how such scores might correspond to real-life capabilities and risk levels [3]. For ChemBio benchmarks, human expert performance typically offers a highly interpretable reference point for model capabilities, and performance below this level can serve as a natural threshold for 'ruling out' risk. [6]

The Gemini 2.5 Pro model report compares the model's results on ChemBio benchmarks with those of other models in the Gemini family, though not with human expert performance (see Figure 3-A). From these results, it is clear that 2.5 Pro somewhat outperforms previous models on each test. However, whether this demonstrates concerning ChemBio capabilities overall, and the implications for real-life risk, are less clear.

When human expert scores are added to this picture (see Figure 3-B), Gemini 2.5 Pro's proficiency across these tests becomes clearer: it outperforms human experts on three, and is plausibly within the margin of error on the remaining three. Adding this human expert baseline allows readers to understand the model's capability level from a more interpretable and familiar reference point, and helps them see the potential implications of these capabilities for risk.

2.2.2 Example 2: Variations on a benchmark's standard methodology (o3 model report)

STREAM (v1) (criterion 2.ii) asks that model reports note if the methodology used for a benchmark deviates from the benchmark's mainline methodology. Varying fundamental elements of test design, such as the answer format or question set, could substantially affect the difficulty, validity and comparability of the evaluation [10]. Ideally, when developers choose between several variations of a test, they should explain the potential implications of doing so to readers on, for instance, the test's saturation point or comparability with human baseline condition.

The o3 model report includes SecureBio's Virology Capabilities Test (VCT) [8], which asks models to troubleshoot problems in laboratory protocols. The report notes that it used the "single select multiple choice" condition. However, SecureBio's recommended methodology (published after the

[2]Note that both the Gemini 2.5 Pro and o3 model reports used an early variant of VCT that included a slightly different question set from the final version of VCT.

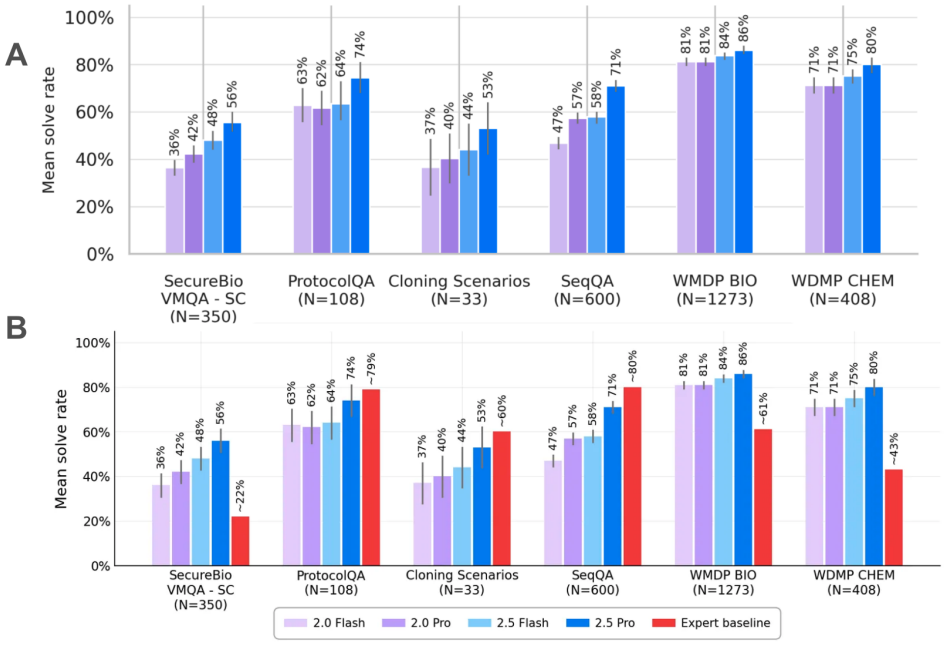

Figure 3: Case study of human expert baselines. Panel A reproduces the ChemBio evaluation results figure from the Gemini 2.5 Pro model report. Sources: Google DeepMind [7] Panel B modifies these to include human expert scores. Note that the human baseline scores for VMQA (VCT) were obtained for the harder 'multiple-response' version of this evaluation, while the model results are recorded in the 'multiple choice' mode. Sources for human expert scores: Götting et al. [8], Laurent et al. [9], and Dev et al. [4]

o3 model report) is different: it requires identifying a set of all correct answer choices—a significantly harder task than the multiple choice task of identifying a single correct answer.

This difference has important implications for the test's results. The o3 model report's methodology found no difference on VCT between o3 and a previous model, o1—both score 59% (see Figure 4-A). However, this could reflect the saturation point of the benchmark when using the 'easier' condition. When SecureBio later tested both models using their recommended condition, they found that o3 had meaningfully improved performance over o1: o3's 44% vs o1's 35% (see Figure 4-B).

Such methodological choices may not be salient to readers if a model report is not explicit about how its deviations from the mainline methodology may have affected evaluation results. AI developers can help third parties correctly interpret evaluations by flagging such choices and briefly explaining potential implications. OpenAI's o3 model report does do so very clearly for another benchmark, ProtocolQA, when they modified it from a multiple choice to an open-ended test.

2.2.3 Example 3: Example questions, especially for private benchmarks (Claude Opus 4 model report)

STREAM (v1) (1.iii) asks that model reports provide an example question and sample answer from the benchmark. Documenting representative examples can provide a particularly clear and concrete illustration of how an evaluation is relevant to a particular threat model, as well as how difficult it may be [16]. This is especially important if the evaluation is not publicly available, as is often the case for evaluations developed in-house.

The Opus 4 model report presents a private 33-question benchmark, "Bioweapons knowledge questions", which tests knowledge of "biological weapons" and "specific steps of the weaponization pathway". At this level of detail, it is difficult for readers to interpret the results. In particular, it is unclear what subtasks the benchmark is composed of, for example: simply retrieving facts vs.

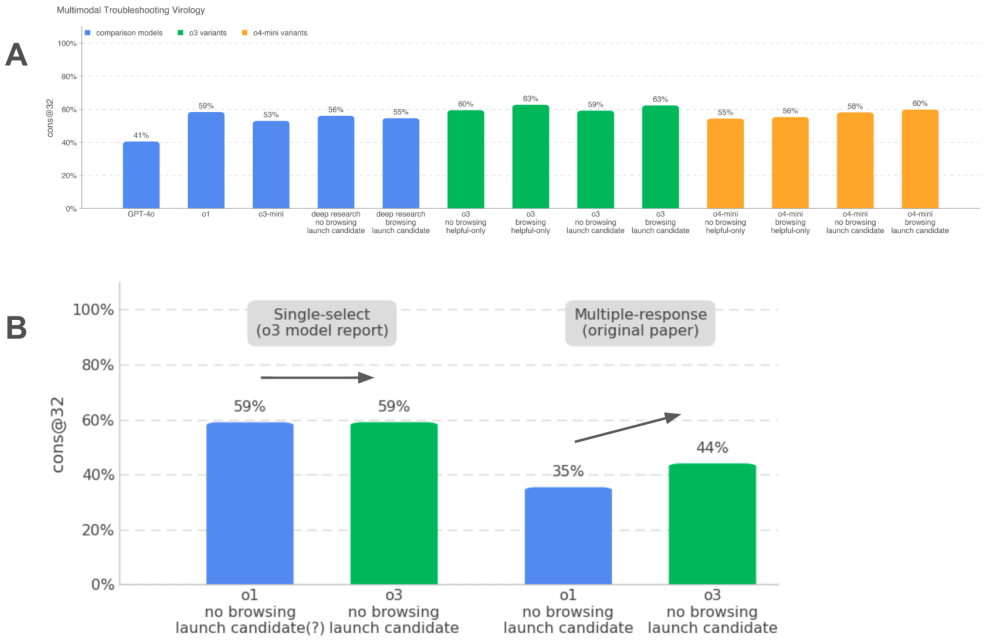

Figure 4: Case study of variations in benchmark methodology. Panel A reproduces the Virology Capabilities Test results as presented by the o3 model report. Sources: OpenAI [14] Panel B highlights how such scores change when switching from the 'single-select' to a 'multiple-response' version of this evaluation. Sources: Götting et al. [8]

reasoning and planning; interpreting open-ended instructions vs. precise questions with clear success criteria; responding with high-level summaries vs. fine-grained technical answers.

Such details can often be conveyed by providing an example item from the benchmark. In some cases, AI developers may need to redact some content within the example due to safety concerns (as in OpenAI [13]), but this can generally be done in a way that preserves the crucial information for readers. In other cases, AI developers may be concerned that publishing test materials could lead to dataset contamination [17], however, providing just a single example can preserve many of the benefits of keeping a benchmark private, while also giving third parties valuable insight into the test.

3 Methodology

3.1 Model report selection

We selected three model reports published in spring of 2025: OpenAI's o3 model report (released April 16, 2025), Anthropic's Claude 4 model report (released May 22, 2025), and Google DeepMind's Gemini 2.5 Pro model report (released June 27, 2025).

We chose model reports that were: (1) published by a company that has made a voluntary commitment to evaluating models for ChemBio risk [18]; (2) publicly available; (3) contained information relevant to STREAM (v1); and (4) related to a frontier AI model at the time of release. For practical reasons, we also limited our analysis to three model reports total, and one model report per company.

These selection criteria meant that we examined reports from companies that were already at the forefront of ChemBio evaluation transparency in spring 2025. Some other companies had not released model reports for their flagship models, or simply included too little information on their ChemBio evaluations to support constructive comparison with STREAM.

Summary Checklist of STREAM v1	
1. Threat relevance	
(i) Does the report explain the capabilities and threat model the evaluation is relevant to?	
(ii) Does the report state what evaluation results would "rule in" or "rule out" capabilities of concern, if any?	
(iii) Does the report provide an example evaluation item and response?	
2. Test construction, grading & scoring	
(i) Does the report state the number of evaluation items?	
(ii) Does the report describe the item type (multiple choice, short answer, etc.) and scoring method?	
(iii) Does the report describe how the grading criteria were created, and describe quality control measures?	
(iv) If human/expert graded...	*(v) If auto-graded by a model...*
(iv-a) Does the report describe the grader sample?	(v-a) Does the report describe the base model used for grading, and any modifications made to it?
(iv-b) Does the report describe the grading process?	(v-b) Does the report describe the automated grading process?
(iv-c) Does the report state the level of agreement between human graders?	(v-c) Does the report state whether the autograder was compared to human graders/other auto-graders?
3. Model elicitation	
(i) Does the report specify the exact model version(s) tested?	
(ii) Does the report specify the safety mitigations active during testing, and any adaptations to elicitation?	
(iii) Does the report describe the elicitation techniques for the test in sufficient detail?	
4. Model performance	
(i) Does the report give representative performance statistics (e.g. mean, maximum)?	
(ii) Does the report give uncertainty measures, and specify the number of evaluation runs conducted?	
(iii) Does the report provide results from ablations/alternative testing conditions?	
5. Baseline performance	
(i) If a human baseline was used...	*(ii) If no human baseline was used...*
(i-a) Does the report describe the human baseline sample and recruitment?	(ii-a) Does the report explain why a human baseline would not be appropriate/feasible?
(i-b) Does the report give human performance statistics, and describe differences with the AI test?	(ii-b) Does the report provide an alternative comparison point, and explain it?
(i-c) Does the report describe how human performance was elicited?	
6. Results interpretation [Can apply once across evaluations]	
(i) Does the report state overall conclusions about the model's capabilities/risk level, and connect with evaluation evidence?	
(ii) Does the report give 'falsification' conditions for its conclusions, and state whether pre-registered?	
(iii) Does the report include predictions about near-term future performance?	
(iv) Does the report state the length of time allowed for interpreting results before deployment?	
(v) Does the report describe any notable disagreements over results interpretation?	

Figure 5: Summary of STREAM reporting requirements. Source: McCaslin et al. [10]

3.2 Applying STREAM criteria

We apply the latest available version of STREAM (v1), which focuses on ChemBio benchmark evaluations, and comprises 28 criteria organized into six categories (see Figure 5). For each ChemBio benchmark presented in our selected model reports, we assess every applicable STREAM criterion from categories 1-5. Criteria in category 6 are applied for each model report overall, rather than each evaluation. Note that our analysis excludes human uplift evaluations and red teaming evaluations, as these are not within the scope of STREAM (v1) [10].

For each criterion we considered whether an evaluation **Satisfied** the criterion, **Partially satisfied** it, or if it was **Not satisfied**. This was done by applying the operationalized rubric from McCaslin et al. [10]'s Appendix C. To 'partially satisfy' a criterion, an evaluation must include all elements described as the 'minimum'. To 'satisfy' a criterion, it must *additionally* include all or most of the elements described as 'full credit'.

We followed STREAM's guidance on assessing reporting constrained by security or similar considerations: if a model report omitted details for the stated reason of not revealing sensitive information, we considered an element met if third party attestations were provided.

We additionally wanted to determine whether certain criteria were systematically more or less well-satisfied across reports, to inform recommendations and future updates to STREAM. Thus we identified the "Best Current Practice" for each criterion across all model reports, or the highest satisfaction rating for any evaluation we examined. For criteria where the "Best Current Practice" achieves a 'satisfied' (or even 'partially satisfied') rating, some developers may be able to improve their reporting by learning from published examples.

3.3 Assessment procedure

Despite the use of the operationalized STREAM rubric, we felt that determining whether a criterion was met unavoidably required some subjective judgement. Therefore, we structured the process of comparing model reports in a way that incorporated multiple perspectives and explicit rationales. This process is summarized in Figure 6.

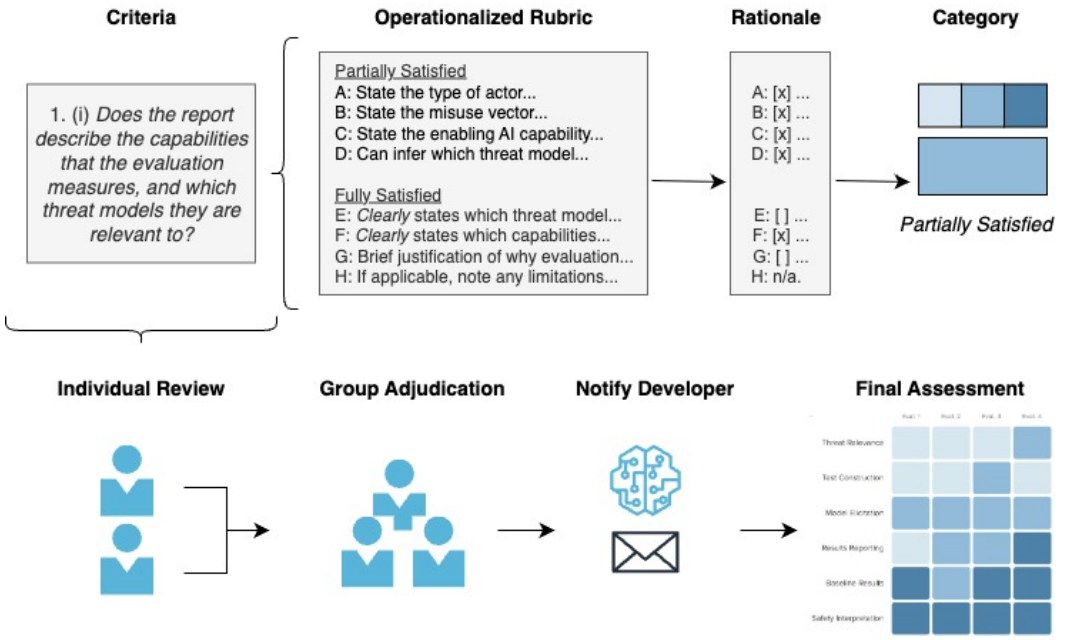

Figure 6: Illustration of the methodology used to assess model reports using the STREAM criteria

For most evaluations[3], two analysts (authors of this paper) independently compared the evaluation with each STREAM criterion. This involved considering each element of a criterion, recording the model report details which supported the element being met or not met (in model report text, tables or figures), and making a judgment as to whether it was met. The analyst then made an overall judgment of the criterion as either "Satisfied", "Partially satisfied" or "Not satisfied" and wrote a brief rationale.

To consolidate the independent assessments, we then held group adjudication meetings with all three authors. The group reviewed each judgment and rationale, discussed any disagreements, and resolved disagreements by consensus where possible. In a small number of cases, persistent disagreement was resolved by majority vote.

Finally, for each model report we prepared a summary document of the consolidated judgments, including rationales, for each evaluation. These summary documents were then sent to the respective AI company safety teams, who we encouraged to flag any concerns or disagreements with our analysis. We considered any disagreements in good faith before finalizing our overall assessment. The details of these are presented in Appendix 1.

4 Discussion

The model reports we analyzed displayed considerable variability in their reporting practices across the different AI developers. This reveals potential **'low hanging fruit' from more shared learning**. Each AI developer can likely make meaningful improvements to their ChemBio evaluation reporting by incorporating lessons from the reporting of their industry peers, and improving consistency by using STREAM as a reporting checklist.

In many cases, such improvements may be relatively straightforward. Some can be achieved by referring to a benchmark's original paper and summarizing a few important details, such as the number of benchmark questions or human expert results for comparison with models.

However, we also observed that some STREAM-recommended information was not found in any of the model reports. This could indicate either that these reporting areas were not salient to AI developers at the time, or that some practical constraint discouraged them from including it. Below we highlight several STREAM criteria that fall into this category, explore some potential causes, and suggestions for AI developers:

- *Example questions (criterion 1.i):* Providing example questions from ChemBio benchmarks enables third party reviewers to develop a clearer, more concrete understanding of how the capabilities measured relate to potential threats—this is especially crucial for private benchmarks, where such information may not otherwise be available (see example 3). It's possible that AI developers are discouraged from providing such examples due to concerns of potentially revealing security-relevant information. In this case, we recommend redacting the most sensitive elements of the question, such as references to a specific pathogen, as was done in [13]. Additionally, training contamination might pose another concern for publishing benchmark material. However, publishing a single example question poses much lower risk of training contamination than publishing a benchmark in its entirety.
- *Detailed descriptions of elicitation (criterion 3.iii):* It is well-established that benchmark results can be sensitive to a wide variety of choices made by evaluators, including prompting, scaffolding, sampling strategies, and inference-time compute [19]. This means "evaluations represent a lower bound for potential capabilities" [14]. While many model reports do provide some basic information about elicitation, a more granular picture may be needed for readers to understand whether results approximate a model's true capabilities, or if a more tailored elicitation might have given different results. It's possible that developers do not provide some elicitation details to avoid revealing information that might aid attackers, or because it could be commercially valuable. In such cases, we suggest that developers report as many evaluation details as possible, given these considerations, while sharing the full details with an independent third party. This third party can then verify that elicitation followed best practices, and attest to this publicly.

[3]Due to time constraints, the Gemini 2.5 Pro model report was only initially assessed by one analyst. However, the model report was reviewed and discussed in detail by all three authors during a group adjudication meeting before producing consolidated assessments.

- *'Falsification' conditions (criterion 6.ii):* Establishing a concrete picture of the evidence that would counterfactually lead to different conclusions can demonstrate that a model report's findings were 'falsifiable'. (It is especially important to establish the evidence that would have lead to a *higher* risk designation than the one obtained.) Some companies, such as OpenAI, already practice this internally by having pre-determined "indicative thresholds" to "indicate that a deployment may have reached a capability threshold" [15]. Sharing these more publicly could strengthen the transparency of their reporting. Other companies, such as Anthropic, do additionally include public details on each benchmark's individual "threshold" [1]. While this is informative, such model reports could give readers the clearest picture by directly stating how such evidence could combine to indicate a *higher* or *lower* capability/risk level if the model had performed differently. Note that this does not necessarily entail setting rigidly pre-defined and binding thresholds—AI developers may currently prefer more holistic approaches to decisionmaking for justifiable reasons. The science of AI evaluation is still developing, and there may be considerable uncertainty about the correct interpretation and weight to give to different pieces of evidence. Such holistic processes may be more challenging to 'pre-register'. In spite of these difficulties, we still believe it is valuable for AI developers to publicly pre-register an *indicative* set of evaluation evidence that could signal a substantial increase over the model's expected risk level. This need not be a pre-*commitment*—developers can and should revise their interpretations as new information comes to light.

Limitations

We examined only three model reports published in a short window (spring 2025). The features we identified may not be representative of the current state of ChemBio evaluation reporting, specific companies, or the industry as a whole. Given the fast pace of model releases, we recommend that more analyses of this type, using STREAM or similar frameworks, be carried out for current and future frontier model releases. Our overall impression is that model reporting has become more detailed over time.

The current version of STREAM (v1) also has an inherently narrow scope, in that it considers only one risk domain (ChemBio) and one evaluation methodology (benchmarks). Reporting for other risk domains in particular may show different patterns of inclusions and omissions. We recommend that future work analyze how model evaluation reporting is typically done for such areas, and for other evaluation methodologies.

Finally, our analysis reflects the judgment of the three authors using the criteria set forth by STREAM (v1). We think that reasonable observers could disagree with some of our assessments when applying the same operational rubric. Similarly, observers could disagree with the applicability and appropriateness of particular STREAM (v1) criteria—as we state in [10], we consider the standard an "initial version" which should "update and adapt over time as best practices emerge".

5 Conclusion

Our comparison of three frontier AI model reports from spring 2025 against STREAM (v1) found noticeable differences between industry reporting practices and the STREAM criteria. While all three AI developers did include several of the STREAM-recommended evaluation details, many details were included inconsistently, and a few were not found in any model reports we examined. We note that there may be considerable 'low hanging fruit' from AI developers learning from peers' best practices, and using STREAM as a checklist to improve consistency. Additionally, we recommend that AI developers make efforts to improve transparency in areas that are currently weaker across the industry, such as providing example test items and documenting elicitation conditions in detail. Such changes to evaluation reporting could help advance evaluation science by sharing insights more widely and enabling the broader expert community to contribute novel perspectives.

References

[1] Anthropic. *System Card: Claude Opus 4 & Claude Sonnet 4*. Tech. rep. Accessed October 9, 2025. Anthropic, 2025. URL: https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf.

[2] Apollo Research. *We Need A 'Science of Evals'*. https://www.apolloresearch.ai/blog/we-need-a-science-of-evals. Accessed October 9, 2025. Jan. 2024.

[3] Hannah P. Cowley, Mandy Natter, Karla Gray-Roncal, Rebecca E. Rhodes, Erik C. Johnson, Nathan Drenkow, Timothy M. Shead, Frances S. Chance, Brock Wester, and William Gray-Roncal. "A framework for rigorous evaluation of human performance in human and machine learning comparison studies". In: *Scientific Reports* 12 (2022), p. 5444. DOI: 10.1038/s41598-022-08078-3.

[4] Sunishchal Dev, Charles Teague, Kyle Brady, Jeffrey Lee, Sarah L. Gebauer, Henry Alexander Bradley, Grant Ellison, Bria Persaud, Jordan Despanie, Barbara Del Castello, Alyssa Worland, Michael Miller, Dawid Maciorowski, Adrian Salas, Dave Nguyen, James Liu, Jason Johnson, Andrew Sloan, Will Stonehouse, Travis Merrill, Thomas Goode, Greg McKelvey, Jr., and Ella Guest. *Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models*. Working Paper WR-A3797-1. Santa Monica, CA: RAND Corporation, Feb. 2025. URL: https://www.rand.org/content/dam/rand/pubs/working_papers/WRA3700/WRA3797-1/RAND_WRA3797-1.pdf.

[5] Financial Times. "OpenAI slashes AI model safety testing time". In: *Financial Times* (2025). Accessed October 9, 2025. URL: https://www.ft.com/content/8253b66e-ade7-4d1f-993b-2d0779c7e7d8.

[6] Frontier Model Forum. *Frontier Capability Assessments*. Tech. rep. Accessed October 9, 2025. Frontier Model Forum, 2025. URL: https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/.

[7] Google DeepMind. *Gemini 2.5 Pro Model Card*. Tech. rep. Accessed October 9, 2025. Google, 2025. URL: https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf.

[8] Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark*. 2025. arXiv: 2504.16137 [cs.CY]. URL: https://arxiv.org/abs/2504.16137.

[9] Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. *LAB-Bench: Measuring Capabilities of Language Models for Biology Research*. 2024. arXiv: 2407.10362 [cs.AI]. URL: https://arxiv.org/abs/2407.10362.

[10] Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, and Luca Righetti. *STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports*. 2025. arXiv: 2508.09853 [cs.CY]. URL: https://arxiv.org/abs/2508.09853.

[11] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. ACM, Jan. 2019, pp. 220–229. DOI: 10.1145/3287560.3287596. URL: http://dx.doi.org/10.1145/3287560.3287596.

[12] National Institute of Standards and Technology. *Pre-Deployment Evaluation of OpenAI's o1 Model*. Tech. rep. National Institute of Standards and Technology, Dec. 2024. URL: https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model.

[13] OpenAI. *Building an early warning system for LLM-aided biological threat creation*. Tech. rep. OpenAI, 2024. URL: https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

[14] OpenAI. *OpenAI o3 and o4-mini System Card*. Tech. rep. Released April 16, 2025. OpenAI, 2025. URL: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[15] OpenAI. *Preparedness Framework v2*. Tech. rep. OpenAI, 2025. URL: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

[16] Patricia Paskov, Michael J. Byun, Kevin Wei, and Toby Webster. *Preliminary Suggestions for Rigorous GPAI Model Evaluations*. Tech. rep. PE-A3971-1. Also available as arXiv:2508.00875. Santa Monica, CA: RAND Corporation, 2025. URL: https://www.rand.org/pubs/perspectives/PEA3971-1.html.

[17] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. "NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10776–10787. DOI: 10.18653/v1/2023.findings-emnlp.722. URL: https://aclanthology.org/2023.findings-emnlp.722/.

[18] The White House. *Voluntary AI Commitments*. Tech. rep. The White House, Sept. 2023. URL: https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

[19] UK AI Safety Institute. *AISI Protocol for Elicitation Experiments*. Tech. rep. UK AI Safety Institute, July 2025. URL: https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/68778c08bd1d69a31d4775e5_Elicitation%20Best%20Practices%202.pdf.

Appendix 1: Details of STREAM assessment of model reports

Please see accompanying files (available on arXiv) for criteria elements and rationales.