STREAM: Expanded Summary

This document provides a detailed summary of the reporting criteria in STREAM-CB. STREAM-CB comprises 28 criteria organized into six high-level categories: Threat Relevance; Test Construction, Grading and Scoring; Model Elicitation; Model Performance; Baseline Performance; and Results Interpretation.

We structured our standard so as to include two tiers of information. Each criterion specifies both a “minimum” requirement of information to be included in a given evaluation summary (which signifies partial compliance with our standard) as well as a “full compliance” portion (which signifies meeting our standard in full, providing all recommended details).

Threat Relevance

1(i) The model report describes what each evaluation is trying to measure, and the specific threat model(s) they are informing.	
Minimal Requirements	**Full Compliance**
1(i)A. Somewhere in the model report, state the type(s) of actors relevant to the ChemBio threat model(s) of concern (e.g. novices, experts, individual, small groups, etc.). **1(i)B.** Somewhere in the model report, state the misuse vector(s) relevant to the ChemBio threat model(s) of concern (e.g. known agents, novel agents, viral pathogens, bacterial pathogens, etc.). **1(i)C.** Somewhere in the model report, state the AI capabilities being assessed in connection with ChemBio threat model(s). **1(i)D.** It is reasonably inferable from the evaluation name, description, ordering, or other contextual information which threat model(s) the evaluation pertains to.	**1(i)E.** *Clearly state* which specific ChemBio threat model(s) this evaluation pertains to. **1(i)F.** Clearly state which *specific* ChemBio capabilities this evaluation measures. **1(i)G.** Give a brief justification for this evaluation as a measure of the capability and/or threat model (e.g. an explanation of how specifically this AI capability could help threat actors). **1(i)H.** WHERE APPLICABLE: Note any major limitations to the evaluation’s threat relevance, e.g. major expected differences between measured capabilities and real-world capabilities.
1(ii) The model report explains the degree to which each evaluation can show that a model lacks (or possesses) a capability of concern, and provides performance thresholds.	
Minimal Requirements	**Full Compliance**
1(ii)A. Somewhere in the model report, for either an applicable subset of evaluations, or this evaluation, indicate whether these evaluations could provide	**1(ii)B.** State what specific score values, ranges or thresholds on this evaluation would be taken as

compelling evidence that the model *lacks* a capability (e.g. “rule out” tests), or else that a model *possesses* a capability (e.g. “rule in” tests), or else that the evaluation is capable of demonstrating either; OR explicitly state that the evaluation is **not** considered when assessing ChemBio risk.	compelling evidence that the model either lacks or possesses a capability. **1(ii)C.** Provide a brief justification for why the score values, ranges or thresholds named in 1(ii)B were deemed significant (e.g. if they exceed a human expert baseline). **1(ii)D.** State when in the evaluation process the score values, ranges, or thresholds named in 1(ii)B were defined (e.g. prior to evaluation test runs with the model, after final evaluation runs were conducted). **1(ii)E.** WHERE APPLICABLE: Note if the interpretation of score ranges differs from that of the evaluation’s designer.
1(iii) The model report provides at least one example item and answer for each evaluation, and notes whether this was representative of the evaluation.	
Minimal Requirements	**Full Compliance**
1(iii)A. Provide at least one item (i.e. example question or task) from this evaluation—sensitive information may be redacted from the item, as long as the example item still conveys enough detail to illustrate the task’s complexity. **1(iii)B.** Provide at least one example response/answer for the evaluation item—sensitive information may be redacted.	**1(iii)C.** State whether the example item given for 1(iii)A is representative of the overall test in terms of difficulty and threat relevance (e.g. referring to a pass rate or percentile). **1(iii)D.** ONLY IF the item is **not** representative of the test overall, provide a brief explanation of the key differences between the example item and the test set generally, or any specific parts of the test which are particularly different.

Test Construction, Grading, & Scoring

2(i) The evaluation summary states the number of items that the model was assessed on, as well as the total number of items in the test (if different).	
Minimal Requirements	**Full Compliance**
2(i)A. Clearly state the number of unique questions/items models were evaluated against in the run(s) reported for this evaluation.	**2(i)B.** ONLY IF the evaluation items were a subset of items on an original, longer test: Specify the number of items on the original test. **2(i)C.** ONLY IF the evaluation items were a subset of items on an original, longer test: State how the subset was chosen (e.g. at random, or from a specific subtest).

<table>
<tr><td colspan="2"><b>2(ii) The evaluation summary states the format(s) in which model responses should be given, explains any necessary scoring details, and notes any deviations from recommended practices.</b></td></tr>
<tr><td><b>Minimal Requirements</b></td><td><b>Full Compliance</b></td></tr>
<tr><td><b>2(ii)A.</b> Describe the answer format(s) required by test items in this evaluation, (i.e. specifying that the test was multiple choice, multiple-select, short answer, open-ended, etc.).
<b>2(ii)B.</b> ONLY IF the evaluation included a mix of different answer formats: indicate the proportion of each type of answer format.</td><td><b>2(ii)C.</b> WHERE APPLICABLE: Flag any notable details of scoring for this evaluation which would not otherwise be apparent to readers, and would be required to replicate the test.
<b>2(ii)D.</b> ONLY IF the evaluation was designed by a third party and any changes were made to the designer's recommended methodology: Explicitly acknowledge differences, and provide a brief justification for differences.</td></tr>
<tr><td colspan="2"><b>2(iii) The evaluation summary states how the answer key and/or grading rubric was created, and briefly describes any quality control measures for grading materials.</b></td></tr>
<tr><td><b>Minimal Requirements</b></td><td><b>Full Compliance</b></td></tr>
<tr><td><b>2(iii)A.</b> State the institutional affiliation of the evaluation's designers.
<b>2(iii)B.</b> ONLY IF the evaluation designers are affiliated with the same organization publishing the model report OR the organization publishing the model report modified an external evaluation in a way that would affect grading: Describe the qualifications (e.g. expertise level and educational background) of the individuals that created or modified the evaluation's answer key/grading rubric/other grading materials, as well as their institutional affiliation (if different from 2(iii)A).</td><td><b>2(iii)C.</b> State whether any validation or quality control measures were taken to ensure high answer keys/grading rubrics/other grading materials (e.g. review by an independent group of experts).
<b>2(iii)D.</b> ONLY IF validation or quality control measures were taken: Briefly describe these measures.
<b>2(iii)E.</b> WHERE APPLICABLE: Explain how questions with ambiguous answers were handled.</td></tr>
<tr><td colspan="2"><b>2(iv-a) If human-graded: The evaluation summary briefly describes the sample of graders and how they were recruited.</b></td></tr>
<tr><td><b>Minimal Requirements</b></td><td><b>Full Compliance</b></td></tr>
<tr><td><b>2(iv-a)A.</b> State the domain or other relevant qualifications of graders.
<b>2(iv-a)B.</b> Disclose the institutional affiliation of graders.</td><td><b>2(iv-a)C.</b> State the number of graders.
<b>2(iv-a)D.</b> Briefly describe how graders were recruited.
<b>2(iv-a)E.</b> WHERE APPLICABLE: Note if graders were provided with training for the grading task.</td></tr>
<tr><td colspan="2"><b>2(iv-b) If human-graded: The evaluation summary briefly describes the grading materials and process.</b></td></tr>
</table>

Minimal Requirements	Full Compliance
2(iv-b)A. Describe the content of the grading instructions and rubrics OR provide illustrative examples of grading instructions and rubrics. **2(iv-b)B.** State whether graders were blinded to the identity of the test-taker.	**2(iv-b)C.** State the typical number of independent graders that graded each item response. **2(iv-b)D.** Briefly explain the process for adjudicating grader disagreements.
2(iv-c) If human-graded: The evaluation summary describes the level of agreement between graders.	
Minimal Requirements	**Full Compliance**
2(iv-c)A. Provide some qualitative or quantitative indicator or statement about the level of agreement between graders.	**2(iv-c)B.** Provide an appropriate summary statistic for grader agreement (e.g. Cohen's kappa) OR, if no statistics are suitable, state this and give a brief summary of grader disagreements. **2(iv-c)C.** WHERE APPLICABLE: Flag grader disagreements with important implications for the capability or risk assessment.
2(v-a) If auto-graded: The evaluation summary identifies the model used as an automated grader and describes any modifications made to it.	
Minimal Requirements	**Full Compliance**
2(v-a)A. Specify the base model used for grading.	**2(v-a)B.** State whether only the base model was used, or if the model was modified for the grading task (e.g. with fine-tuning, task-specific scaffolding, etc). **2(v-a)D.** WHERE APPLICABLE: Briefly describe any modifications made to the base model for the grading task.
2(v-b) If auto-graded: The evaluation summary briefly describes the automated grading materials and process.	
Minimal Requirements	**Full Compliance**
2(v-b)A. Provide a brief description of the grading rubrics and grading instructions used OR illustrative examples of grading instructions and rubrics. **2(v-b)B.** Provide a brief description of how the auto-grader judged performance, e.g. based on similarity with gold standard answers.	**2(v-b)C.** Share an example prompt used for the auto-grader (sensitive details can be redacted). **2(v-b)D.** State whether multiple auto-grader samples were generated per evaluation item response. **2(v-b)E.** ONLY IF multiple auto-grader samples were generated: State how these scores were aggregated for a final score.

<table>
<tr><th colspan="2">2(v-c) If auto-graded: The evaluation summary states whether the automated grader was validated against human graders or another auto-grader, and if so, reports the level of agreement.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><strong>2(v-c)A.</strong> State whether the auto-grader’s performance was validated against human graders, another auto-grader, or not at all.
<strong>2(v-c)B.</strong> ONLY IF the auto-grader’s performance was validated against human graders: Describe the number of human graders and their qualifications.</td><td><strong>2(v-c)C.</strong> Provide a summary statistic for the level of agreement between the auto-grader and the comparison grader; OR, if no comparison was made, provide a brief explanation for why this was not done.
<strong>2(v-c)D.</strong> ONLY IF a comparison between the auto-grader and another grader was made: State whether the comparison was conducted on the full set of evaluation items or a subset.</td></tr>
</table>

Model Elicitation

<table>
<tr><th colspan="2">3(i) The model report specifies which version(s) of the model were tested.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><strong>3(i)A.</strong> Somewhere in the model report, clearly specify which model instance(s) were identical to the final/deployed model (e.g. “launch candidate”); OR make clear that no tested model instance was identical to final version.
<strong>3(i)B.</strong> ONLY IF the evaluation includes any model instances that are <strong>not</strong> the final/deployed model version: Somewhere in the model report, clearly specify which model instances included in this evaluation had the full deployment set of mitigations/safeguards in place at test time, and which had a reduced/minimal set.</td><td><strong>3(i)C.</strong> ONLY IF the evaluation did <strong>not</strong> include a final/deployed model version: Provide some estimate of the capability difference of at least one of the tested model instances to the final/deployed model. Can be qualitative or quantitative.
<strong>3(i)D.</strong> Label model instances tested in this evaluation in a way that is clear and consistent with model version descriptions satisfying 3(i)A and 3(i)B.</td></tr>
<tr><th colspan="2">3(ii) The model report briefly describes all the relevant mitigations active during evaluations, and describes any simulated efforts to circumvent mitigations.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><strong>3(ii)A.</strong> Somewhere in the model report, for either evaluations generally, an applicable subset of evaluations, or this evaluation, briefly list the relevant safeguards and mitigations (e.g. unlearning, safety fine-tuning, content classifiers).</td><td><strong>3(ii)C.</strong> Somewhere in the report, for each specific model instance tested in this evaluation, make clear what set or subset of mitigations/safeguards were in place at test time. (Ex: list uniform set of mitigations applied for ChemBio or</td></tr>
</table>

3(ii)B. Somewhere in the model report, state whether elicitation conditions included any attempts to bypass active safeguards/mitigations (e.g. jailbreaking attacks); OR, if such attempts were **not** made, but adversarial use was instead tested using model instances with mitigations/safeguards removed, make this clear by labelling these model instances and displaying their results alongside results for safeguarded model(s).	automated evals; or, if only testing final/deployed model, state final deployment set.) **3(ii)D.** Somewhere in the report, briefly describe how rigorous any attempts to bypass active safeguards/mitigations were (e.g. how much time was spent finding jailbreaks); OR, for this evaluation, briefly explain why no bypassing attempts were made (e.g because there were no model refusals). **3(ii)E.** IF APPLICABLE: disclose the extent to which model refusals affected evaluation. (Ex: number of items on which refusals occurred.)
3(iii) The model report specifies the actions taken to surface the full range of model capabilities during evaluation.	
Minimal Requirements	**Full Compliance**
3(iii)A. Somewhere in the model report, briefly describe how models were prompted for evaluations. **3(iii)B.** Somewhere in the model report, for either evaluations generally, an applicable subsest of evaluations, or this evaluation, state which sampling/generation strategies were used for evaluations. (Ex: "Best-of-5", "pass@1", "none".) **3(iii)C.** Somewhere in the model report, for either all evaluations, an applicable subset of evaluations, or this evaluation, state whether any tools were provided to the models (e.g. web search, calculators). **3(iii)D.** Somewhere in the model report, for either all evaluations, an applicable subset of evaluations, or this evaluation, state whether any scaffolding was used (e.g. agentic scaffolding). **3(iii)E.** WHERE APPLICABLE: somewhere in the model report, state the use of any fine-tuning of models for evaluations.	**3(iii)F.** Somewhere in the model report, briefly describe the prompt design process for evaluations. **3(iii)G.** IF APPLICABLE: provide examples of prompts used for this evaluation. **3(iii)H.** Somewhere in the model report, briefly list the tools provided to models for this evaluation; OR state that none were provided. **3(iii)I.** Somewhere in the model report, briefly describe the scaffolding used for this evaluation; OR state that none was used. **3(iii)J.** Somewhere in the model report, for either all evaluations, an applicable subset of evaluations, or this evaluation, state what resource ceilings were applied (e.g. maximum inference time/tokens). **3(iii)K.** Somewhere in the model report, for either all evaluations, an applicable subset of evaluations, or this evaluation, state what sampling parameters were applied (e.g. temperature). **3(iii)L.** ONLY IF fine-tuning was used (see 3(iii)E): Somewhere in the model report, briefly describe the dataset and/or methods used for fine-tuning.

Model Performance

<table>
<tr><th colspan="2">4(i) The evaluation summary presents the most relevant summary statistics for the model(s) tested.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><b>4(i)A.</b> Present whichever summary statistic(s) for model performance on this evaluation are most appropriate, either in text, or in a figure or graph.</td><td><b>4(i)B.</b> Clearly present the summary statistic(s) given for 4(i)A either in text, a table, or a graph with clear text labelling (a figure or graph with no numerical labelling of the summary statistic is not sufficient).
<b>4(i)C.</b> ONLY IF the summary statistic reported is <b>not</b> mean solve rate or a similar metric: Give a brief justification for the choice of summary statistic(s).</td></tr>
<tr><th colspan="2">4(ii) The evaluation summary provides confidence intervals (or other uncertainty measures) for performance statistics, and specifies the number of evaluation runs conducted.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><b>4(ii)A.</b> Include an appropriate measure of statistical uncertainty for the performance reported for 4(i), e.g. confidence interval, standard error of the mean, either in text, or in a figure or graph.
<b>4(ii)B.</b> ONLY IF confidence intervals are given: Include the confidence level (e.g. “95% CI”).</td><td><b>4(ii)C.</b> Specify the number of evaluation runs conducted per model that the summary statistics summarize.
<b>4(ii)D.</b> Clearly present the uncertainty measure(s) given for 4(ii)A either in text, a table, or a graph with clear text labelling (a figure or graph with no numerical labelling of the uncertainty measure is not sufficient).</td></tr>
<tr><th colspan="2">4(iii) The evaluation summary states whether ablation experiments or multiple alternative testing conditions were performed, and states whether the model was tested for training contamination.</th></tr>
<tr><th>Minimal Requirements</th><th>Full Compliance</th></tr>
<tr><td><b>4(iii)A.</b> State whether supplementary evaluation runs were performed with major variations on mainline evaluation conditions (e.g. different elicitation protocols, resource ceilings, or test versions)
<b>4(iii)B.</b> ONLY IF supplementary evaluation runs described in 4(iii)A were performed: Report the outcome of each major testing variation (e.g. with summary statistics or a qualitative description).</td><td><b>4(iii)C.</b> Explicitly confirm whether the model report provides the “highest” score or summary measure on this evaluation that was obtained under any testing condition or variation (where “highest” should be construed as “most concerning”, if numerically higher scores do not indicate more concerning outputs).
<b>4(iii)D.</b> State whether the model was tested for contamination of its training data with benchmark content.
<b>4(iii)E.</b> ONLY IF testing for contamination described in 4(iii)D was performed: Briefly summarize the results of this testing.</td></tr>
</table>

Baseline Performance

5(i-a) If human baseline: The evaluation summary states the number of human participants, their qualifications, and how they were recruited.

Minimal Requirements	Full Compliance
5(i-a)A. State the total number of human participants for the human baseline test for this evaluation. **5(i-a)B.** ONLY IF the report specifies that the human baseline is “expert” level: State the human baseline participants’ specific domain(s) of expertise (e.g. virology) AND their education level or relevant professional experience. **5(i-a)C.** ONLY IF 5(i-a)B is not applicable: State the type of human baseline (e.g. “novice”) AND provide some statement about their qualifications, domain knowledge, or other task-relevant characteristics.	**5(i-a)D.** Briefly describe how the human baseline sample was recruited (e.g. recruitment channels). **5(i-a)E.** WHERE APPLICABLE: Disclose any features of recruitment that were likely to introduce significant sampling bias (e.g. experts all drawn from a single research group).

5(i-b) If human baseline: The evaluation summary provides human performance statistics, and reports any differences between the AI evaluation and human baseline test.

Minimal Requirements	Full Compliance
5(i-b)A. Present whichever summary statistic(s) for human baseline performance on this evaluation are most appropriate, either in text, or in a figure or graph.	**5(i-b)B.** Include an appropriate measure of statistical uncertainty for the human baseline performance reported for 5(i-b)A, e.g. confidence interval, standard error of the mean, either in text, or in a figure or graph. **5(i-b)C.** ONLY IF confidence intervals are given: Include the confidence level (e.g. “95% CI”). **5(i-b)D.** Clearly present the summary statistic(s) given for 5(i-b)A and the uncertainty measure(s) given for 5(i-b)B either in text, a table, or a graph with clear text labelling (a figure or graph with no numerical labelling of the uncertainty measure is not sufficient). **5(i-b)E.** ONLY IF the human baseline summary statistic is **not** either the mean or an identical measure to the model summary statistic in 4(i): Give a brief justification for the choice of human baseline summary measure. **5(i-b)F.** WHERE APPLICABLE: Report any important differences between the AI evaluation and the human

	baseline test (e.g. if humans were only graded on questions matching their expertise).
5(i-c) If human baseline: The evaluation summary provides details of the testing conditions in the human baseline experiment.	
Minimal Requirements	**Full Compliance**
5(i-c)A. Report the amount of time allowed for human baseline participants to complete this evaluation task. **5(i-c)B.** Describe what resources human participants had access to during the baseline test (e.g. internet access, biological design tools, none).	**5(i-c)C.** Briefly describe what incentives participants were given to ensure high motivation for performing well on the test (e.g. hourly base-pay plus performance bonuses). **5(i-c)D.** State how much time human baseline participants spent on a typical test item, or on the test as a whole, on average. **5(i-c)E.** WHERE APPLICABLE: Note any other features of the testing environment that may have significantly impacted performance, or any problems observed at test time (e.g. with motivation or task compliance).
5(ii-a) If no human baseline: The model report explains why a human comparison would not be appropriate or feasible.	
Minimal Requirements	**Full Compliance**
5(ii-a)A. Briefly explain why including a human baseline for this evaluation would be infeasible (e.g. due to high costs, legal constraints, or safety risks) OR briefly explain why a human baseline for this evaluation would not be informative (e.g. because the test is trivially easy or excessively hard for humans).	**5(ii-a)B.** Provide supporting details or evidence for 5(ii-a)A (e.g. authoritative sources consulted, time or cost estimates for human baseline study, supporting research literature).
5(ii-b) If no human baseline: The model report provides an alternative way of interpreting the evaluation in the absence of human comparisons (e.g. an alternative baseline).	
Minimal Requirements	**Full Compliance**
5(ii-b)A. Provide some other means of interpreting the significance of model performance on this evaluation, such as scores from previously released models, or a summary of expert judgments on appropriate score interpretations for this evaluation.	**5(ii-b)C.** Justify why the reference point(s) satisfying 5(ii-b)A provide a valid and useful comparison with the main model results, in particular explaining specifically how these reference point(s) could inform an accurate interpretation of a model's ChemBio capabilities or risk level.

5(ii-b)B. ONLY IF 5(ii-b)A is **not** met with empirical baselines such as previously released model scores: Briefly describe the methodology for obtaining the expert judgments or other reference point(s) satisfying 5(ii-b)A.	**5(ii-b)D.** Briefly summarize major uncertainties affecting 5(ii-b)A, 5(ii-b)B, or 5(ii-b)C.

Results interpretation

6(i) The model report states the conclusions the evaluators have drawn about the model's capabilities and risk level, and connects this with evaluation and other evidence.	
Minimal Requirements	**Full Compliance**
6(i)A. Somewhere in the model report, state the overall conclusions drawn about the model's ChemBio capability level and/or ChemBio risk level. **6(i)B.** Somewhere in the model report, provide a brief statement on how the conclusion(s) in 6(i)A impacted decision-making (e.g. deployment decisions, level of mitigations, etc.).	**6(i)C.** Somewhere in the model report, clearly explain the degree to which specific evaluations contributed to the conclusion(s) in 6(i)A, in one of the following ways: by indicating which evaluations had the most influence on these conclusion(s); OR by indicating which tested capabilities had the most influence (provided these capabilities are clearly tied to specific evaluations); OR by clearly describing a rule or formula used for outputting conclusions from evaluation results. **6(i)D.** Somewhere in the model report, briefly describe any important influences on the conclusion(s) in 6(i)A *other than* the reported evaluations, e.g. evaluations performed by external parties.
6(ii) The model report states what evidence could have 'falsified' the conclusion(s) above, and whether such interpretations were pre-registered in a credible way.	
Minimal Requirements	**Full Compliance**
6(ii)A. Somewhere in the model report, clearly state what combination of evaluation results or other evidence could have significantly changed the conclusion(s) in 6(i)A—in particular, state what would have resulted in a *higher* risk or capability determination.	**6(ii)B.** Somewhere in the model report, state whether the conditions described for 6(ii)A were pre-registered in connection with the higher risk interpretation, either as a public statement or as shared with a credible third party.
6(iii) The model report includes statements about near-term future performance.	
Minimal Requirements	**Full Compliance**

6(iii)A. Somewhere in the model report, include a statement about how model performance might improve in the near future (3-6 months from release) with further development of elicitation techniques and tools. **6(iii)B.** ONLY IF the model will be deployed open-source or open-weight: Somewhere in the model report, include a statement about how model performance might improve in the next 12-24 months. **6(iii)C.** Somewhere in the model report, state any implications of statements for 6(iii)A (and 6(iii)B if applicable) for capability thresholds, risk levels, or mitigations/safeguards.	**6(iii)D.** Somewhere in the model report, provide a brief explanation of the statement(s) for 6(iii)A (and 6(iii)B, if applicable). **6(iii)E.** Somewhere in the model report, provide at least a tentative statement about when an important decision point (e.g. a capability or risk threshold) might be reached by a model in this model family. This can be in terms of calendar time (e.g. “3 months”) or development schedule (e.g. “next major model release”).
6(iv) The model report states how much time the relevant team(s) had to consider evaluation results prior to deployment.	
Minimal Requirements	**Full Compliance**
6(iv)A. Somewhere in the model report, provide some statement about how long internal safety teams (or whichever groups/individuals are most relevant, such as independent third-party evaluators) had to form and communicate interpretations of evaluation results prior to model deployment.	**6(iv)B.** Somewhere in the model report, provide a rough quantified estimate of the time reported in 6(iv)A (e.g. through date ranges, numbers of days, or FT equivalents).
6(v) The model report briefly describes any notable uncertainties or disagreements related to interpreting results or making risk judgments, and how these were handled.	
Minimal Requirements	**Full Compliance**
6(v)A. Somewhere in the model report, state whether any notable uncertainties or disagreements arose during the ChemBio evaluation and interpretation process.	**6(v)B.** ONLY IF the model report does **not** explicitly state that there were **no** uncertainties/disagreements: Somewhere in the model report, briefly summarize notable uncertainties/disagreements (sensitive information can be redacted). **6(v)C.** Somewhere in the model report, briefly explain how considerations from 6(v)B were dealt with (e.g. independent review); OR, if there were **no** uncertainties/disagreements, outline how they *would have* been addressed, had they occurred.

Terminology:

“Applicable subset of evaluations” - When criteria refer to information provided for "an applicable subset of evaluations," this includes general statements about evaluation procedures that apply to a broader category or evaluation suite that encompasses the specific evaluation being assessed. For example, if an evaluation is part of the "CBRN evaluations" suite, then general statements about CBRN evaluation methodology would satisfy criteria that allow for "applicable subset" reporting.

“State whether” - The model report must either explicitly state that a given condition was met, explicitly state that it was not met, or provide details of how the condition was met that implicitly confirms it.