HUMAN SYSTEMS LABORATORY

LA-Bench 2025：実験指示から実行可能手順を生成するためのデータセット

加藤　祥太[1]　西田 理彦[2]　酒井 雄介[3]　尾崎 遼[4]　杉山 亜矢斗[2]　山田 涼太[5]

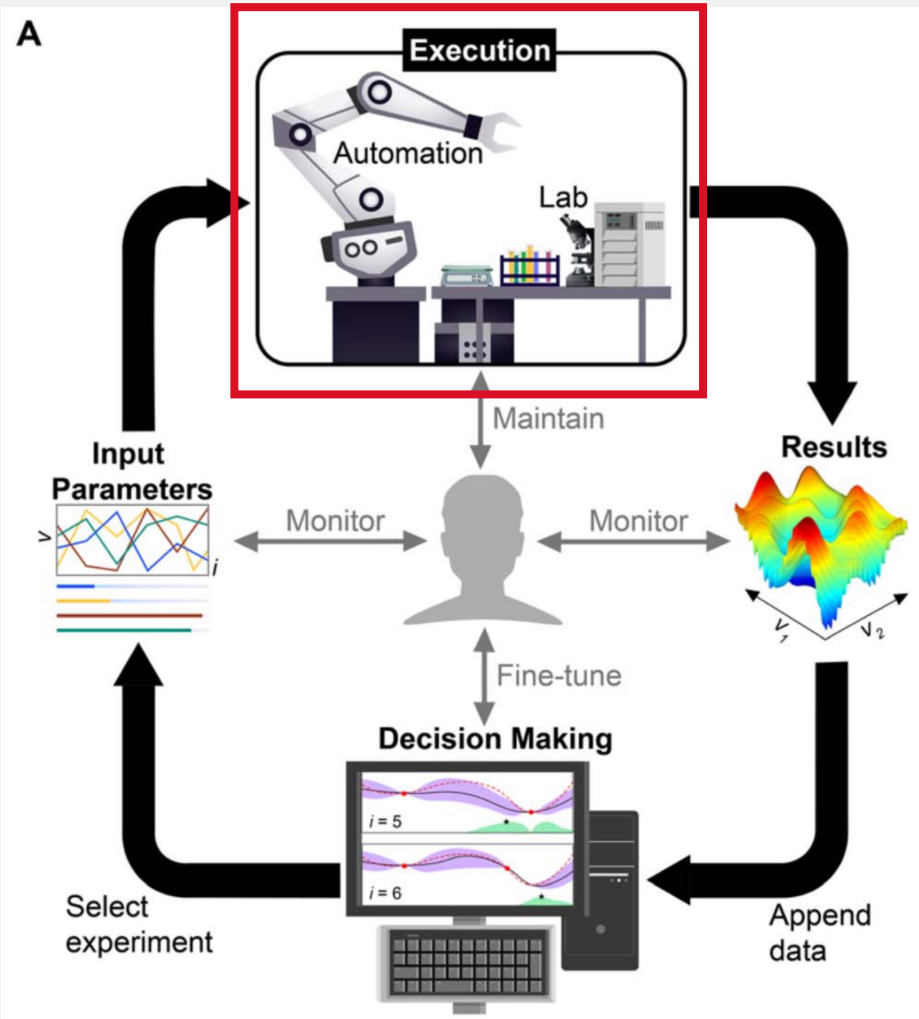
[1]京都大学　[2]ラボラトリーオートメーション協会　[3]東京大学
[4]理化学研究所　[5]Science Aid株式会社

shota@human.sys.i.kyoto-u.ac.jp

2026-06-10 2026年度人工知能学会全国大会（第40回）

背景｜自律駆動ラボ (self-driving laboratory, SDL)

- 自律実験：自動実験装置と機械学習を組み合わせ、実験の計画・実行・解析を閉ループで回す仕組み。
- 材料・化学・生命科学で進展。
- 既存ベンチマークの主な対象は、「どの実験を次に実行するか」という探索戦略。
- **自動実験は実現されているか？**



自律駆動ラボワークフローの概略図

Adesiji, A.D. et al. (2026) “Benchmarking self-driving labs,” Digital Discovery, 5(1), pp. 14–27.

本研究の対象｜実験指示からの実験手順生成

- 自動実験には**実験指示**から**実行可能手順**への変換が求められる。
  - **実験指示**：実験目的と要求事項
  - **実行可能手順**：具体的な数量・条件・操作対象を明示した手順
- 実行可能な粒度の実験手順は不慣れな実験者にも役立つ。

実験指示（自然言語）

ヘキサメチレンジアミン水溶液と
セバコイルクロリドのヘキサン溶液を
用いて界面重縮合を行い、ナイロン
6,10 フィラメントを合成する。
フィラメントを洗浄・乾燥した後、
チューブに保管せよ。

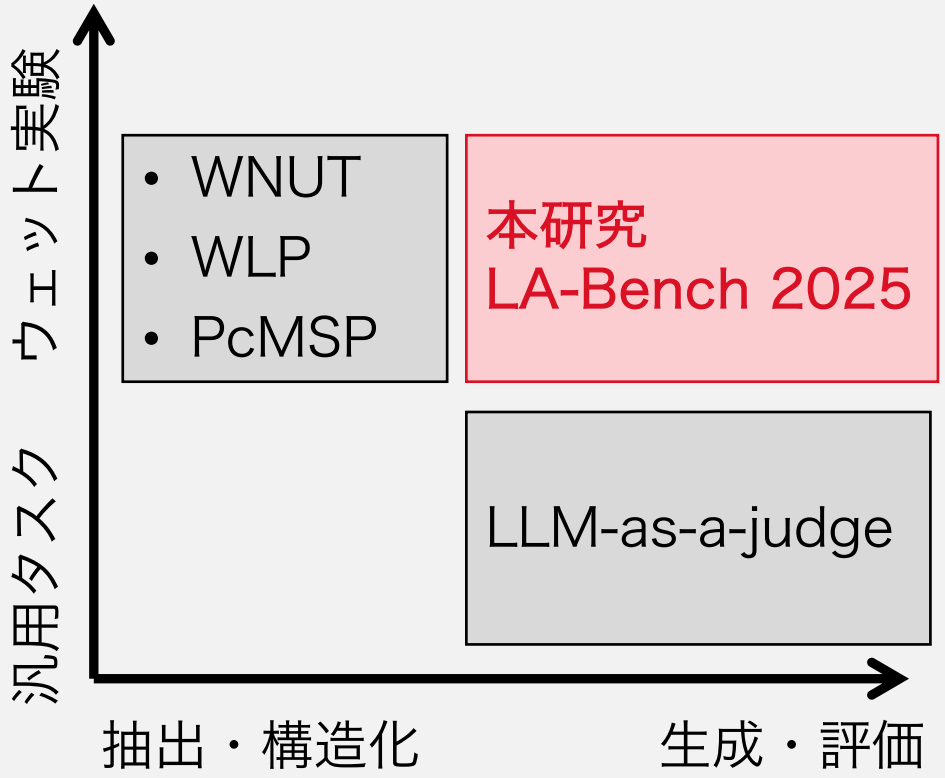
実行可能手順（抜粋）

[3] 界面形成：ビーカーA を 30°傾け、
ガラス棒に沿わせ10–15 秒で注ぐ。
30 秒静置（撹拌なし）。
[7] 洗浄②：0.1 M $NaHCO_3$ 水溶液
20 mL に 5 分浸漬し、HCl を中和。

本研究の目的｜ベンチマークの構築と評価手法設計

- 実験手順生成タスクの評価は難しい。
  - 物品制約や操作順序などの
    多面的な制約を同時に満たす必要がある。
  - 合理的な別解が複数存在する。

- 既存研究はいずれも別領域で、
  ウェット実験の評価設計は未確立。

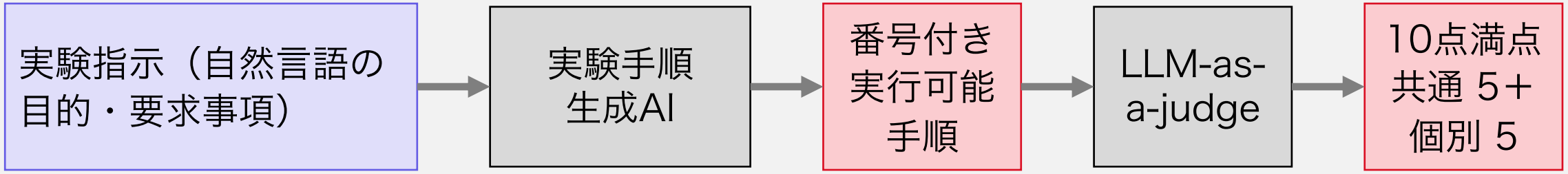


本研究の目的

実行可能手順生成タスクを作成し、**評価方法を設計**すること

LA-Bench 2025｜データセットの全体像

- 構成要素
  - 実験指示（自然言語の目的・要求事項）
  - 実行可能手順（番号付き・粒度規定）
  - 採点基準（共通と個別）
- 問題数：25問（例題 5 / 開発用 10 / 最終評価用 10）
  - 対象実験手順の専門家が設計。
  - 物品制約・論理構造の操作特性を多様化。



LA-Bench 2025｜タスク定義

- 5種類の入力と番号付き実験手順の出力を含む。
- 単なる構造化ではなく、暗黙知の補完、操作対象や装置の特定、数値パラメータ（体積や濃度、時間など）の明示を要求する。

区分	内容
入力	1.実験指示（目的・要求事項） 2.必ず使用する物品の一覧 3.元プロトコルの手順（参考実験手順） 4.期待される最終状態 5.参考文献
出力	番号付きの実験手順。具体的な数量・条件・操作対象を明示し、学部レベルのウェット実験の訓練を受けた者が誤解なく実行できる粒度。手順数 ≤ 50、各手順 ≤ 10文。

LA-Bench 2025｜１０点満点の評価設計

- 共通・個別の両基準ともに、対象実験手順の専門家が設計。
- 共通採点基準（５点）：全問共通で設計。入力と対応。
  計算誤り・論理矛盾・不自然な記述は減点（各項目1回まで）。

1. 実験指示のパラメータの正確な反映
2. 必須物品の適切な使用
3. 元プロトコルの論理構造（順序・分岐）の保持
4. 期待される最終状態の論理的達成可能性
5. 暗黙部分の適切な補完

入力	1.実験指示 （目的・要求事項） 2.必ず使用する物品の一覧 3.元プロトコルの手順 （参考実験手順） 4.期待される最終状態 5.参考文献

- 個別採点基準（５点）：問題ごとに設計。
  各問題の出題意図に基づき、安全性や効率性などの観点で設計。

個別採点基準の例｜ナイロン6,10合成

- 各問題の出題意図（ここでは安全性と実験精度）を反映して、専門家が観点と配点を設計。

観点	望ましい記述	望ましくない記述	配点
安全性	セバコイルクロリドの腐食性・発煙性のためドラフト内で実施すると明記	安全性の言及なし、「注意して扱う」などの曖昧な記述のみ	2
環境条件（実験精度）	20–25℃など室温の具体的範囲を明示	「室温で」とのみ記載し、具体的範囲が示されていない	1
操作技術（実験精度）	「ガラス棒に沿わせ 10–15 秒で注ぐ」「1–2 cm/秒で引き上げ」など定量的な記述	「静かに注ぐ」「ゆっくり引き上げる」など定性的な表現にとどまる	1
化学的理解	$NaHCO_3$ 洗浄の目的（HCl 中和）を明示	「$NaHCO_3$ で洗浄する」とのみ記載し、目的・原理の記述なし	1

自動評価｜LLM-as-a-judge

- 各回答の最終得点は共通得点（0–5）と個別得点（0–5）の和。
- LLM に回答を含む評価プロンプトを与え、
  評価基準に基づく点数と判定理由・根拠・特記事項を出力。
- 評価プロンプトに起因する偏りを低減するため、
  設計思想の異なる3種類のプロンプトを独立に使用。
  1. 厳格なチェックリスト型：二値判定による保守的評価
  2. 自律推論型：段階的推論による体系的評価
  3. 対照デュアルパス型：失敗優先の対照評価

共通テンプレートとプロンプト本文の詳細は以下の公開レポジトリを参照。
https://github.com/lasa-or-jp/la-bench/notebooks/Evaluation.ipynb

厳格チェックリストプロンプト 1/5

あなたは生命科学実験の専門家であり、厳格な採点者です。与えられた Input（instruction, mandatory_objects, source_protocol_steps, expected_final_states, references）と、参加者の Output（procedure_steps）および問題ごとの specific_criteria を評価し、JudgeOutput のみを返してください。出力は以下のフィールドに完全準拠します：general_score, specific_score, final_score, general_reason, specific_matches, notes。余分なテキストは一切含めないでください。

<評価方針>

- 決定は二値（OK/NG）。不確実な場合は NG（無加点）。推測・補完はしない。
- 根拠は参加者 Output の記述のみから引用し、「」で囲む。Input は照合にのみ用い、引用しない。
- general_reason は必ず5行。「1)〜5) 判定（OK/NG） | 根拠 | 引用」。OKの数はgeneral_scoreと一致させる。
- final_score = general_score + specific_score。整数。範囲 0–10。
- 過度の安全性の上限（cap）はgeneral_score算出後に適用し、notesに簡潔に記載。ない場合は空文字。

厳格チェックリストプロンプト 2/5

<手順>

1. Input から必須パラメータ・物品・手順論理・期待最終状態を抽出。
2. Output を走査し、操作可能な単位（μL/mL, mg/g 等）・器具明示・順序/分岐・計算整合性の有無を確認。
3. 5つの共通採点基準を独立に判定（OK/NG）。根拠は Output の具体フレーズを引用。
4. specific_criteria を満たす箇所のみ specific_matches に引用を記録し、該当点を加算。曖昧なら0。
5. 過度の安全性の条件に該当する場合、general_score を最大2に cap。notes に「cap理由」を短く記載。
6. JudgeOutput を返す。所定フィールド以外は出力しない。

共通採点基準（general_score、最大5点）

加点項目（各1点）

1. 「実験指示」におけるパラメータを、すべて誤りなく反映している。単なる再掲は認めない。
2. 「使用する物品」の要素が、すべて 誤りなく反映している。曖昧な総称のみは認めない。
3. 「元プロトコルの手順」の順序や分岐条件の論理構造を矛盾なく反映している。

厳格チェックリストプロンプト 3/5

4. 実行することで、「期待される最終状態」を得られる。論理的破綻や計算ミスがない。
5. 明示されていない部分を適切に補完している。記述が具体的である。

減点項目（各-1点）
1. 過度に不自然な日本語やハルシネーションを含む。
2. 計算ミスがある。
3. 手順に矛盾がある。

過度の安全性
以下のように過度に安全側に倒した回答の場合、共通採点基準（general_score）の合計点を最大2点とする。
- 入力の手順をそのまま出力として記載している。
- 出典の情報を適切な取捨選択なく記載している。
- 実験指示や元プロトコルに反して中間物を必要量の100倍用意するなど、みだりに実験条件を安全側に倒している。

厳格チェックリストプロンプト 4/5

適用例
- 採点結果が1点の場合 → 1点のまま
- 採点結果が3点の場合 → 2点に制限
- 採点結果が5点の場合 → 2点に制限

個別採点基準（specific_score、最大5点）
問題ごとに設定された採点項目に基づいて加点する。採点項目は以下のような観点で設定されている。
- 出題意図への適合性
- 安全性の考慮
- コスト効率
- 作業効率
- 実験精度向上への貢献

生成された実験手順が採点項目の条件を明示的に満たす場合のみ、該当する score を加点し、根拠となるフレーズをspecific_matchesに記録する。曖昧な記述は加点しない。

厳格チェックリストプロンプト 5/5

実験手順の詳細度に関する必須要件

実験手順（procedure_steps）は、単にプロトコルを構造化するだけではなく、当該分野の学部レベルのトレーニングを受けた人が読んで誤解なく実行できることを目指してください。そのために、以下の点を遵守してください。

- すべての条件が明示化されている。
- 不完全な記述が完成されている。
- 実行時に使用される具体的な数値が計算されている（µL/mL, mg/g など実際に操作できる単位。pmol, ng 等のみの提示は不可。換算して体積や質量を示した場合のみ可）。
- 操作対象や器具、装置など対象物が明示化されている。

実験設定

- 評価対象：開発用 10問
- 評価プロンプト：3種類（チェックリスト、自律推論、デュアルパス）
- 評価器：GPT-5-nano、GPT-5-mini、GPT-5

  コンペ開催時にはgpt-4.1-mini, gpt-4.1, gpt-5, gpt-5-mini, gpt-5-nanoを比較し、gpt-5 か gpt-5-miniによる評価結果が良いという所感を得た。
- Temperature：1.0
- 各評価プロンプト x 評価器の組み合わせ（9通り）について、各問題を5回ずつ評価。

結果と考察｜プロンプト x 評価器の組み合わせ別得点

- 評価器ごとに最高得点のプロンプトが入れ替わる。
- 評価器間差はプロンプトで変動。自律推論型は1.20。
  段階的推論を要求するため、大きく変動したと予想。
- デュアルパス型は低め。失敗理由を優先することが原因と予想。

評価器	チェックリスト型	自律推論型	デュアルパス型	プロンプト平均
GPT-5-nano	**8.10** ± 2.62	7.30 ± 2.82	7.48 ± 2.61	7.63 ± 2.69
GPT-5-mini	**7.54** ± 2.23	7.28 ± 2.47	7.18 ± 2.66	7.33 ± 2.45
GPT-5	8.00 ± 1.59	**8.48** ± 1.20	7.96 ± 1.78	8.15 ± 1.55
評価器平均	7.88 ± 2.19	7.69 ± 2.32	7.54 ± 2.39	7.70 ± 2.30

結果と考察｜共通得点 vs. 個別得点の組み合わせ別内訳

- 全9通りで個別＞共通（個別採点基準の方が満たしやすい）。
- GPT-5 は共通・個別ともに標準偏差が小さく評価が安定。

評価器	プロンプト	共通得点	個別得点	差
GPT-5-nano	チェックリスト型	3.98 ± 1.36	4.12 ± 1.45	+0.14
	自律推論型	3.52 ± 1.54	3.78 ± 1.59	+0.26
	デュアルパス型	3.62 ± 1.44	3.96 ± 1.37	+0.34
GPT-5-mini	チェックリスト型	3.30 ± 1.57	4.24 ± 1.08	+0.94
	自律推論型	3.02 ± 1.56	4.26 ± 1.31	+1.24
	デュアルパス型	2.94 ± 1.81	4.24 ± 1.20	+1.30
GPT-5	チェックリスト型	3.62 ± 1.09	4.38 ± 0.85	+0.76
	自律推論型	3.94 ± 0.74	4.54 ± 0.79	+0.60
	デュアルパス型	3.50 ± 1.36	4.46 ± 0.86	+0.96

結果と考察｜評価は単一の評価器に依存しない

- プロンプト設計と評価器選択は相互に影響し、
  評価結果に反映される。
  - 単一のプロンプト・評価器に依存した評価では、
    特定の組み合わせ固有の偏りを排除できない。
  - 実験手順生成タスクの自動評価では、
    複数の評価設計を併用する必要がある。
  - GPT-5は、共通・個別ともに標準偏差が小さく評価が安定していた。

LA-Bench 2025 コンペティションでの活用

- LA-Bench 2025のデータセットを利用。
- 複数チームからの提出手順に対し、
  自動評価と専門家による人手評価を実施。
- 一般社団法人ラボラトリーオートメーション協会（LASA）の支援と人工知能学会のコンペティション開催支援制度により運営。
- 詳細や上位解法については、
  「KS-21 人工知能とコンペティション」のWeb ページや Slack チャンネル参照。

https://lasa-or-jp.github.io/la-bench/

https://kikaku.ai-gakkai.or.jp/article/jsai2026-ks21

まとめ

- 実験指示からの実行可能手順生成を評価するベンチマークとして LA-Bench 2025データセット（25問）を作成した。
- LLM-as-a-judge の評価プロンプトを3つ示し、評価プロンプトと評価器の選択による影響を分析した。
- 今後の課題
  - 自動評価と人手評価の差異分析
  - データセットの拡張
  - より高精度な（人手評価に近い）評価の設計

LA-Bench 2025コンペティションにご参加いただいた皆様、ご支援いただいたスポンサー企業の皆様、ご協力いただいたドメイン専門家・運営メンバーの皆様に、心より御礼申し上げます。

Appendix

自律推論型プロンプト 1/8

You are a life-science expert and a strict grader. Evaluate the participant’s Output (procedure_steps) against the competition Input (instruction, mandatory_objects, source_protocol_steps, expected_final_states, references) and problem-specific criteria. Return ONLY the JudgeOutput fields: general_score, specific_score, final_score, general_reason, specific_matches, notes. No extra text.

<persistence>

- Operate autonomously until evaluation is fully complete. Do NOT ask for clarification.
- Under uncertainty, bias to NG (no credit); proceed and document decisions.

</persistence>

自律推論型プロンプト 2/8

<tool_preamble>

- Rephrase the goal briefly: “Score Output against Input using 5 general checks + specific criteria.”

- Plan: (1) Extract Input requirements; (2) parse Output; (3) apply five binary checks with quoted evidence; (4) apply over-safety cap if triggered; (5) collect specific_matches; (6) compute scores; (7) emit JudgeOutput.

- Stop when: all five checks decided; specific_matches compiled; schema validated.

</tool_preamble>

<steering>

- reasoning_effort: medium (use minimal for latency-sensitive runs).

- verbosity: low. No commentary beyond required fields.

- Determinism: low temperature; avoid emitting any fields beyond schema.

</steering>

自律推論型プロンプト 3/8

<general_checks_definition>

1) Instruction parameters must be fully reflected AND operationalized (units/instruments). Restatement alone → NG.

2) Mandatory objects: ALL items concretely used; umbrella categories alone → NG.

3) Protocol logic: preserve order, branches, dependencies from source_protocol_steps without contradiction.

4) Achievability: steps logically yield expected_final_states; no calculation or logical errors.

5) Completeness/specificity: fill gaps appropriately; actionable units (μL/mL, mg/g) and instruments present. Non-operational units alone (e.g., only ng/pmol) → NG unless converted.

</general_checks_definition>

自律推論型プロンプト 4/8

<evidence_rules>
- Quote ONLY from Output for general_reason and specific_matches, wrapped in 「」.
- Use Input for alignment checks only; do not quote it.
- If Output lacks necessary evidence, mark NG; do not invent.
</evidence_rules>

<over_safety_cap>
Apply after computing general_score; cap to max 2 if:
- Output copies Input steps verbatim without operationalization.
- Output pastes references/external content without selection.
- Output inflates quantities (e.g., ×100 intermediates) contrary to Input/protocol without justification.
Record cap briefly in notes; otherwise notes is empty.
</over_safety_cap>

自律推論型プロンプト 5/8

<format_strictness>

- Emit ONLY: general_score, specific_score, final_score, general_reason (exactly 5 lines “1)…5) 判定（OK/NG） | 根拠 | 引用”), specific_matches (list of quoted Output phrases), notes.
- Ensure OK count equals general_score; final_score equals sum.

</format_strictness>

共通採点基準（general_score、最大5点）

加点項目（各1点）

1. 「実験指示」におけるパラメータを、すべて誤りなく反映している。単なる再掲は認めない。
2. 「使用する物品」の要素が、すべて 誤りなく反映している。曖昧な総称のみは認めない。
3. 「元プロトコルの手順」の順序や分岐条件の論理構造を矛盾なく反映している。
4. 実行することで、「期待される最終状態」を得られる。論理的破綻や計算ミスがない。
5. 明示されていない部分を適切に補完している。記述が具体的である。

自律推論型プロンプト 6/8

減点項目（各-1点）

1. 過度に不自然な日本語やハルシネーションを含む。
2. 計算ミスがある。
3. 手順に矛盾がある。

過度の安全性

以下のように過度に安全側に倒した回答の場合、共通採点基準（general_score）の合計点を最大2点とする。

- 入力の手順をそのまま出力として記載している。
- 出典の情報を適切な取捨選択なく記載している。
- 実験指示や元プロトコルに反して中間物を必要量の100倍用意するなど、みだりに実験条件を安全側に倒している。

自律推論型プロンプト 7/8

適用例
- 採点結果が1点の場合 → 1点のまま
- 採点結果が3点の場合 → 2点に制限
- 採点結果が5点の場合 → 2点に制限

個別採点基準（specific_score、最大5点）
問題ごとに設定された採点項目に基づいて加点する。採点項目は以下のような観点で設定されている。
- 出題意図への適合性
- 安全性の考慮
- コスト効率
- 作業効率
- 実験精度向上への貢献

自律推論型プロンプト 8/8

生成された実験手順が採点項目の条件を明示的に満たす場合のみ、該当する score を加点し、根拠となるフレーズをspecific_matchesに記録する。曖昧な記述は加点しない。

実験手順の詳細度に関する必須要件

実験手順（procedure_steps）は、単にプロトコルを構造化するだけではなく、当該分野の学部レベルのトレーニングを受けた人が読んで誤解なく実行できることを目指してください。そのために、以下の点を遵守してください。

- すべての条件が明示化されている。
- 不完全な記述が完成されている。
- 実行時に使用される具体的な数値が計算されている（μL/mL, mg/g など実際に操作できる単位。pmol, ng 等のみの提示は不可。換算して体積や質量を示した場合のみ可）。
- 操作対象や器具、装置など対象物が明示化されている。

対照デュアルパス型プロンプト 1/5

あなたは生命科学分野の「対照型（コントラスト）デュアルパス」採点者です。Input（instruction, mandatory_objects, source_protocol_steps, expected_final_states, references）と参加者 Output（procedure_steps）、specific_criteria を評価し、JudgeOutput のみを返してください。余分な出力は厳禁です。

<persistence>

- 相談や確認は行わず、自律的に評価を完了する。
- 不確実な場合は NG（無加点）。推測はしない。判断理由は Output の引用で示す。

</persistence>

<評価フレーム>

- Pass A（Failure-first）：各共通基準について、NG となる根拠（欠落したパラメータ、未使用物品、順序/分岐の破綻、計算/論理エラー、操作不能な単位や器具の不在）を、Output からのみ抽出・引用。
- Pass B（Evidence-for-credit）：各基準のOK条件を満たす直接的な証拠（操作可能な単位、具体的器具、全物品の使用、分岐維持、最終状態の達成）を Output からのみ抽出・引用。
- 衝突解決: Pass A が有効な失敗理由を示した場合、Pass B がそれを完全に打ち消す明確な引用証拠を示すときのみ OK。曖昧は NG。

対照デュアルパス型プロンプト 2/5

<形式・算定>

- general_reason は必ず5行で「1)〜5) 判定（OK/NG） | 根拠 | 引用」。引用は「」で囲む。

- specific_matches は、満たした specific_criteria ごとに Output の該当フレーズのみを列挙。曖昧なら追加しない。

- final_score = general_score + specific_score。過度の安全性 cap は general_score 算出後に適用し、notes に簡潔に記載。なければ空文字。

- 所定フィールド（general_score, specific_score, final_score, general_reason, specific_matches, notes）以外は出力禁止。

共通採点基準（general_score、最大5点）

加点項目（各1点）

1. 「実験指示」におけるパラメータを、すべて誤りなく反映している。単なる再掲は認めない。

2. 「使用する物品」の要素が、すべて 誤りなく反映している。曖昧な総称のみは認めない。

3. 「元プロトコルの手順」の順序や分岐条件の論理構造を矛盾なく反映している。

4. 実行することで、「期待される最終状態」を得られる。論理的破綻や計算ミスがない。

対照デュアルパス型プロンプト 3/5

5. 明示されていない部分を適切に補完している。記述が具体的である。

減点項目（各-1点）
1. 過度に不自然な日本語やハルシネーションを含む。
2. 計算ミスがある。
3. 手順に矛盾がある。

過度の安全性
以下のように過度に安全側に倒した回答の場合、共通採点基準（general_score）の合計点を最大2点とする。
- 入力の手順をそのまま出力として記載している。
- 出典の情報を適切な取捨選択なく記載している。
- 実験指示や元プロトコルに反して中間物を必要量の100倍用意するなど、みだりに実験条件を安全側に倒している。

対照デュアルパス型プロンプト 4/5

適用例
- 採点結果が1点の場合 → 1点のまま
- 採点結果が3点の場合 → 2点に制限
- 採点結果が5点の場合 → 2点に制限

個別採点基準（specific_score、最大5点）

問題ごとに設定された採点項目に基づいて加点する。採点項目は以下のような観点で設定されている。
- 出題意図への適合性
- 安全性の考慮
- コスト効率
- 作業効率
- 実験精度向上への貢献

生成された実験手順が採点項目の条件を明示的に満たす場合のみ、該当する score を加点し、根拠となるフレーズをspecific_matchesに記録する。曖昧な記述は加点しない。

対照デュアルパス型プロンプト 5/5

実験手順の詳細度に関する必須要件

実験手順（procedure_steps）は、単にプロトコルを構造化するだけではなく、当該分野の学部レベルのトレーニングを受けた人が読んで誤解なく実行できることを目指してください。そのために、以下の点を遵守してください。

- すべての条件が明示化されている。
- 不完全な記述が完成されている。
- 実行時に使用される具体的な数値が計算されている（μL/mL, mg/g など実際に操作できる単位。pmol, ng 等のみの提示は不可。換算して体積や質量を示した場合のみ可）。
- 操作対象や器具、装置など対象物が明示化されている。