実験手順生成を
多段レビューで作り込む

team cabbage patch / member Yuya Tahara-Arai

1位 **9.289** **LLM評価部門 最優秀賞**

2位（9.285）との差は **0.004**

JSAI2026 企画セッション「人工知能とコンペティション」(KS-21)

課題：文書だけで実験は再現できるか

LA-Bench = **「新人やロボットが、文書だけで実行できる手順」** を生成するタスク

LLMに1回作らせただけだと、こんな失点が起きやすい 👇

- 🌀 **曖昧語**（「速やかに」「十分に」）
- 🔢 **数値・条件の抜け**
  （量・温度・時間）
- 📦 **必須物品の未使用**
- 📄 **手順の丸投げ**（「詳細は別紙参照」で済ます）

➡️ これらを **多段のレビュー** で減らす

設計の出発点：ウェット実験の現場感覚

発表者はウェットラボ出身。現場でプロトコルをどう作っていたかを思い返した。

現場で暗黙知・コツを埋める方法

- 実験ノートを読む / ラボの人に聞く
- 実際の操作を見せてもらう
- 目の前でやって、その場でレビュー

でも、自動生成では…

- 「やりながら聞く」が使えない
- 暗黙知が抜けたまま手順になる

➡️ そこで前提を固定：**「後輩に、文書だけで教えるつもりで書く」**
この一点が、以降のチェックリスト化・数値化すべての土台になった

手法の全体像：生成 → 批評 → 投票 → 校正

- **設計思想**：LLMに自分の出力を批評させ、多段で直す
- 機械チェック（制約違反）＋ LLMレビュー（安全性・再現性）の二段構え
- **最終校正**で、採点で見られる観点に手順を合わせ直す

効いた工夫 ① 暗黙知をチェックリスト化

「書かなくても伝わる」現場のコツが、事故・失敗の大きな原因になる

レビュー観点に組み込んだ暗黙知

- 溶液取り扱い / ピペッティング
- 液面検出 / 混和 / 遠心
- 温度管理 / 容器による差 / トラブル時の対応

頭の中で予行演習（ドライラン）

- 手順を時間軸で通す
- 起きがちな失敗を先に想定（泡・温度未到達…）
- 「再試行 → 代替 → 中止」の分岐

➡️ 手順に **具体的な操作条件** が入り、曖昧さが減った

効いた工夫 ② あいまい語 → 数値へ翻訳

「速やかに」 ➡️ 60秒以内

「十分に乾かす」 ➡️ 25℃で5分

「残液を除く」 ➡️ 残液 20μL以下

- **具体化を促す自作ルール** で抜けを防ぐ（許容範囲・停止条件・品質チェック基準を必ず明記）→ 詳細は巻末の補足資料
- そして **最終校正で、採点で見られる観点に手順を合わせ直す**
  ➡️ ここが **LLM評価** で大きく効いた

効いた工夫 ③ 生成モデルそのもの

パイプラインは同じまま、生成モデルを差し替えるだけでも点は動いた

gpt-5-mini　7.8 ➡️ gpt-5　8.9

同じ批評パイプラインで、モデル差し替えのみ ＝ **+1.1**

- 正直なところ：手順の作り込みとは別の軸で、**素のモデル力（とコスト）** が効く

改善録：積み上げで伸ばした

LLM評価スコア（10点満点）を動かした改善を、効いた順に振り返る

5.4 LLMに自分の手順を批評させる仕組みを導入

7.8 使用物品の使い忘れ・誤用への指摘をレビューに追加

8.9 生成モデルを gpt-5 に変更

9.289 **最終校正で、採点で見られる観点に手順を合わせ直す**

結果と考察

1位
LLM評価　9.289
10点満点・21チーム中

9位
専門家評価　6.9
10点満点・上位10チームのみ評価

※ 採点者が違うため点数の直接比較はできない（AIが採点 / 人が採点）

- LLM評価に強く最適化 → LLM部門は1位だったが、専門家評価とは差
- 反省：公開テスト（提出前に試せるデータ）で **AIの点数だけを指標**にし、**自分の目でほぼ確認しなかった** → AIに気に入られる手順に偏った
- 学び：**LLM評価への最適化 ≠ 専門家の納得**。人手チェックの併用が次の鍵

ご清聴ありがとうございました 🥬

（予備）設計の出発点：実際の生成プロンプト

ウェットの現場感覚を、太字の1行に落とし込んだ。

AIに与えた役割・ルール

あなたはこの領域の専門研究者です。Input を読み、日本語で実行可能な実験手順を返す。
適宜ラボの新人にも分かりやすい表現に変換（敬語禁止）。共通の評価基準をすべて満たす。
制約：最大50ステップ／各10文以下／id は1から昇順。

AIに渡した材料（入力情報）

- 【実験指示】パラメータを誤りなく反映
- 【使用する物品】全て1回以上・正しい用途（未使用禁止）
- 【元プロトコル】順序・分岐を矛盾なく再構成
- 【期待される最終状態】／【参考文献】

（予備）レビューと改訂のプロンプト

レビュー：AIに与えた役割

この領域の研究者として批評的にコメントする（敬語禁止、提案は“行動指示”で）。
手順そのものは直さず、**要修正点（must_fix）・リスク・総評**を構造化して返す。
自動チェックが不合格なら違反を要修正点に必ず入れ、"実行不可"扱い（点数も60以下）。

改訂：AIに与えた役割

レビューを受け、**最小限の修正**で実行可能性・安全性・再現性を高める。
不要な冗長化は避け、具体値・再試験条件・記録要件・安全措置を明確化する。

（予備）投票と最終校正のプロンプト

投票：AIに与えた役割

候補のレビュー要約だけを根拠に、評価基準に照らして順位付けする。
安全性と再現性を最優先、次点で作業効率・コスト。1位→N位の並びで返す。

最終校正：AIに与えた役割（別スクリプト・1回だけ）

README と採点基準を読み、**この出力だけで安全・再現性高く実行できる**ように一度だけ校正。
元の手順をベースに、必要な修正・整理・具体化・短い追加のみ。別物には書き換えない。

（予備）毎回渡した共通の評価基準

生成・レビューの各段階で必ず与えた自作ルール（specific criteria）。

数値で固定・矛盾を消す

- 条件は数値で固定（適量,十分に禁止）
- 最終状態に必要な量・時間・温度・濃度・回数を明示
- 数値の矛盾を排除（負の時間・濃度不整合など）

安全・品質

- 安全対策を“行動指示”で明文化
- 機器の校正・点検／有効データ基準・再試験ルール

実験設計・運用

- 条件数・繰り返し数（レプリケート）・対照群を定量で確認
- マスターミックス化・段階希釈・サンプル残りの保存など実務的に
- 許容範囲と代替手順で装置差を吸収

効率・記録・追跡

- コスト効率・作業効率（手戻り／待ち時間）
- 記録・報告の必須項目を網羅
- 各基準を「手順◯番で対応」の形で紐づけ（追跡可能に）

（予備）各段階の詳細

① 候補生成 ×5 専門研究者として実行可能な手順を生成。温度を振って多様性を確保
出力フォーマット名: GeneratedOutput（procedure_steps）

② 自動チェック 50ステップ／各10文を機械検証。違反は要修正点に必ず計上し、点数も60以下に抑える

③ レビュー 研究者として批評し、安全性・再現性の不足を洗い出す
出力: ReviewOutput（ok / score / must_fix / risks / notes）

④ 改訂 ×2 must_fix を反映し、最小限の修正で再現性・安全性を底上げ

⑤ 投票 ジャッジLLMがレビュー要約だけを根拠に順位付け（安全性・再現性を最優先）
出力: VoteOutput（ranking）

⑥ 最終校正 ×1 README と採点基準を読み、採点で見られる観点に手順を合わせ直す（別スクリプト）