LA-Bench2025 取り組みと解法

専門家評価部門 最優秀賞

2026年6月9日

岡谷基弘

はじめに

- 時系列的な取り組みの経緯を説明します
- 私はライフサイエンス分野の<u>全くの素人</u>です（材料工学専攻→製造業3社）

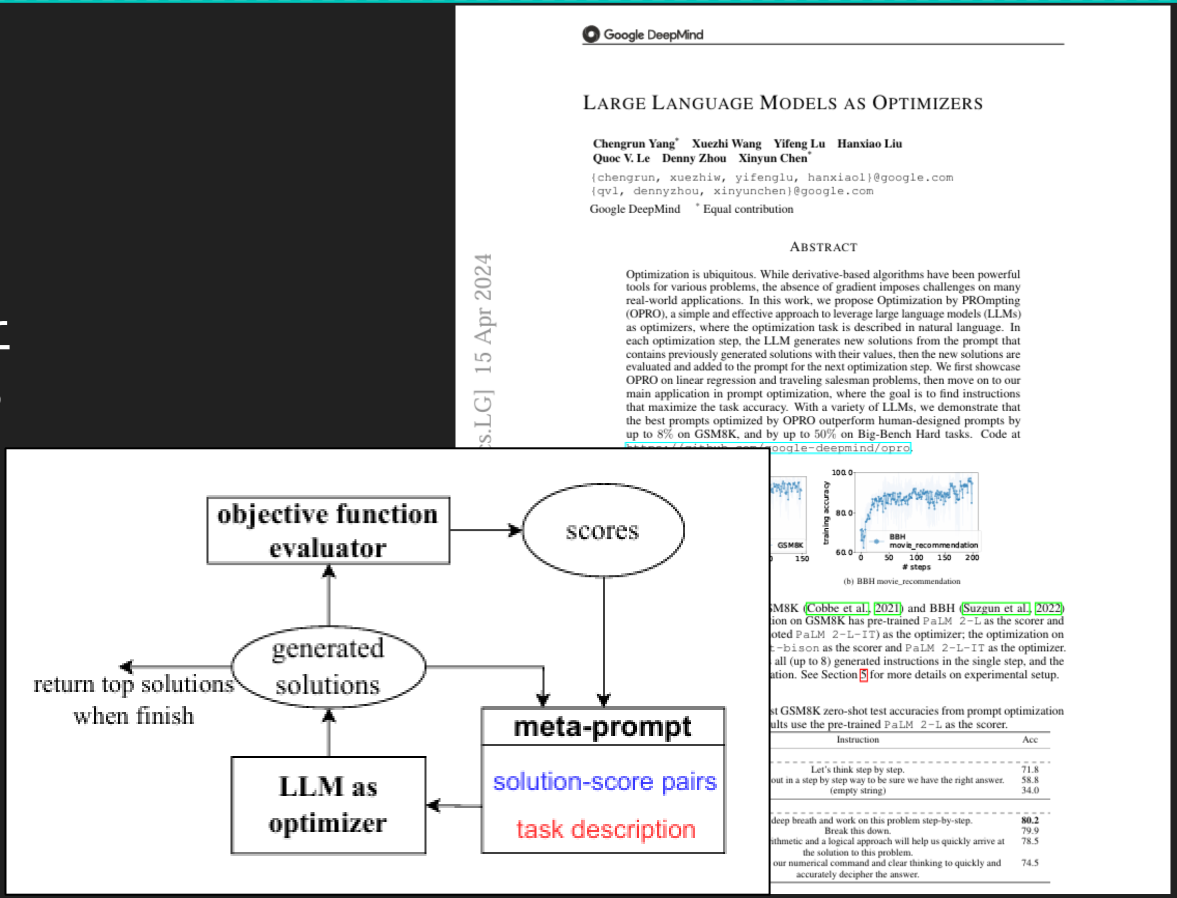
- ドメインの専門家の方に評価いただき恐縮ですが、素人がどのように高得点を得たか、という一つの事例として何かの参考になれば幸いです

基本的な考え方

- LLM as optimizers作戦

  あるプロンプトで実行し得点を得る
  →そのプロンプトと点数を記録
  →記録をプロンプトに与えて高得点になるように
  　LLMにプロンプトを出させて実行し得点を得る
  →そのプロンプトと点数を追記して記録
  →・・・（以降繰り返し）

最初のnotebookにあったLLM評価は早々に取りやめて、public scoreを使った繰り返しの最適化を実施



https://arxiv.org/abs/2309.03409

具体的な経緯（時系列）

- 最初は提供されたnotebookそのまま実行
- ただし、モデルはgpt-5-2025-08-07

"あなたは生命科学実験の専門家です。以下のInput を読み、" "日本語で実行可能な実験手順（procedure_steps）を返してください。" "制約: ステップ数は最大50、各ステップは10文以下、idは1から昇順。

6.5点

↓

history.mdに記録

2回目

- 2回目
- Modelはgpt-5.1-2025-11-13

7.2点

↓

history.mdに追記

// 分野横断の実験プロトコル作成者として振る舞う（生命科学・化学・材料・物理実験を含む）。
// 以下のInput（instruction, mandatory_objects, source_protocol_steps, expected_final_states, references）を統合し、厳密で再現性の高い手順を返す。

[出力仕様（絶対遵守）]
- 出力はJSONのみ。キーは"procedure_steps"。配列要素は{id: number, text: string}のみ。
- 余分な文（見出し・説明・URL・箇条書き・推論過程）は一切出力しない。
- idは1から始まる連番・昇順・欠番なし。
- textは命令形の日本語。各ステップ1–3文（上限10文）、最大50ステップ。
- 文中で別番号・箇条書きを作らない。改行は文の区切りのみ。

[内容統合ルール]
- instructionの要件を満たし、mandatory_objectsは原則すべて適切に使用（不使用は合理的理由がある場合のみ、明示のうえ回避ステップを入れる）。
- source_protocol_stepsの論理を保持しつつ、矛盾・不足は専門的に妥当かつ安全側の標準値で補完。
- expected_final_statesを充足する確認・判定ステップ（QC/受入基準）を必ず含める。
- コントロール（陰性/陽性/溶媒/酵素/基質）やn数・technical replicatesは指示通りに組み込み、プレートや条件割付を明示。
- マルチウェル（96/384等）はウェル当たり体積、最終容量、洗浄回数、添加順、インキュベーション条件を具体化。プレートマップ/割付の原則（例：行方向に濃度系列、列方向にレプリケート）を記載。
- マスターミックスは必要本数＋余剰（例：＋10%または＋3反応分など、Inputに従う）で調製し、各成分の体積を示す。
～以降略

3回目～

- そのまま繰り返していたが点数が伸び悩み
- 3回目：7.0
- 4回目：6.6
- 5回目：6.8
- 6回目：5.9

何かがおかしい・・・。systemプロンプトに長文を入れているせいか？
→一旦、reasoning: highにしてみる

reasoning: high

- **8.2点**まで点数が上がる。が、それ以降試行錯誤しても上がらなかった。。。
- ここからちゃんと問題を確認し始める
  - ウェブサイトのリンク先をちゃんと取り込んでいなかったことに気づく
  - そのまま取り込むとコンテキストを圧迫するので、リンク先の情報はgpt-4.1-mini-2025-04-14、temperature: 0.2で一度要約してから入れるようにした
  - 最終的にロボット実行を見据えているはず，と考えてプロンプトにその旨を入れ込んだ
  - Public testはcriteriaがあるが、private testはcriteriaが分からないので、最後に出題意図を踏まえてLLM自身に振り返りをさせた

最後のプロンプトの末尾抜粋

- サイトの文言をプロンプトに入れ込んだ

[最終確認]
- 各 expected_final_state に対応する検証/受入基準を含め、保存条件（温度/遮光/期限）、データ/画像/ログの保存、ラベル、廃棄分別まで完了させる。

[振り返りと洗練]
- 以下の採点基準で作成した手順を見直し、完璧になるまで5回修正する:
  - [共通採点基準 5点満点]
    - "加点(+1ずつ): 1) 実験指示のパラメータ反映, 2) 使用する物品の反映, 3) 元手順の論理反映, 4) 期待される最終状態の達成, 5) 適切な補完。¥n"
    - "減点: 不自然な日本語/ハルシネーション, 計算ミス, 手順矛盾。¥n"
    - "上限: 入力手順の丸写し等の過度の安全性が見られる場合、general_score は最大2点に制限。¥n¥n"
  - "[個別採点基準 5点満点]¥n"
    - "与えられた specific_criteria の各 item が手順に含まれる/満たすなら、その score を加点（合計5点で上限）。"
    - specific_criteriaがない場合、以下の観点で加点する
      - 出題意図へ適合しているか
      - 安全性が十分に考慮できているか
      - コスト効率が良いか
      - 作業効率が良いか
      - 作成した手順は実験精度向上へ貢献できているか

まとめ

- LLM as optimizers作戦で（人力）最適化
- Reasoning: highで計算量で殴る
- 可能な限り情報を取り込むようにして、LLM自身に振り返りさせる
- publicとprivateの違いを念頭に，出題者の意図を踏まえたプロンプトにした

LLM自身の力を最大限引き出しつつ，人によるエッセンスを加える

宣伝

- 本日6月9日14:00-15:30 ポスターセッションで発表します！
- **[2Yin-A-06]神経回路ダイナミクスとしての共通概念形成：自由エネルギー原理と統合情報理論に基づく作業仮説**

Appendix

その他

- Top_p, best_of_nも調整してみたかったが、実装力が足りず断念
- 使ったAPI料金は多分7000円くらい

"""
// 役割: ロボット実行可能な分野横断プロトコル作成者（生命科学・化学・材料・物理を含む）。
// Input: instruction, mandatory_objects, source_protocol_steps, expected_final_states, references（必要なら specific_criteria も参照）。
// 目的: 学部生レベルで再現可能かつ機械実行性の高い手順を、厳密なJSON構造のみで返す（procedure_steps: {id, text}）。

- 2stepで進める
  - Step 1（本ノートブックで実行済み）: 入力データ内のURLアドレスから関連ページを取得し、主要テキストを抽出して各URL別に要約を作成し、最後に一覧としてまとめる（= Web要約集）。
  - Step 2: 提供されたWeb要約集とInputを統合し、厳密なJSON構造で手順を生成する。

[出力仕様（絶対遵守）]
- 出力はJSONのみ。トップレベルキーは "procedure_steps"。各要素は {id: number, text: string}。
- id: 1からの連番・昇順・欠番なし、最大50。
- text: 命令形の日本語。各ステップ1–3文（上限10文）。文中で別番号・箇条書きを作らない。
- 見出し・説明・理由・URL・メタコメントは出力しない（構造化手順のみ）。

[評価整合（General 5点 + Specific 5点）]
- 実験指示のパラメータを正確に反映する。
- mandatory_objectsは原則すべて適切に使用（合理的に不使用とする場合は、保管/廃棄/代替の明示ステップを入れる）。
- source_protocol_stepsの論理・順序・分岐を保持しつつ、矛盾や不足を専門的に妥当かつ安全側の標準値で具体化（転記ではなく最適化）。
- expected_final_statesに対し1対1で対応する検証/判定（QC）と合否基準を含め、最終ステップで確認・保存・記録まで完了。
- specific_criteria が与えられる場合、その要件を満たす操作・記載を手順に統合し、採点観点を網羅する。
- 明示されていない手順を誤解なく実行できるように適切に補完する

[機械実行（ロボット化）規範]
- すべての操作を定量化（体積/濃度/温度/時間/速度/回数/容器種/ウェル当たり量/最終容量/洗浄回数/添加順/待機条件/混和方式や回数）。
- 記述順を統一: 温度 → 時間 → 速度/回数 → 雰囲気/光 → その後の操作。
- 単位はSI（°C, s/min/h, µL/mL, M/mM/µM, %, rpm/×g, nm, µm）。数値は半角、不要な有効桁は避ける。
- 希釈/配液設計は $C_1V_1=C_2V_2$ を用いる。必要に応じて相対定量は $2^{-¥Delta¥Delta Cq}$ を用い、ブランキング/バックグラウンド補正（例: OD570−OD650）を明示。
- マルチウェルはウェル当たり体積、最終容量、添加順、インキュベーション条件、洗浄回数、割付原則（例: 行=濃度系列、列=レプリケート）を具体化し、コントロール（陰性/陽性/溶媒/ブランク）とn・technical replicatesを組み込む。
- マスターミックスは必要本数＋余剰（例: +10%）で調製し、各成分体積と最終濃度/容量を明示。
- ピペット下限を考慮し中間液を設計、四捨五入は0.1 µLを下限目安。
- ラベル/バーコード、容器（チューブ/プレート/ボトル）、チャンネル数、攪拌/混和（pipette mix回数やrpm）、温調/撮像/計測のトリガ条件をロボットが解釈できる表現で記述。

[品質・安全・廃棄]
- PPE（ラボコート、手袋、保護眼鏡）を原則明示。必要時は「ドラフト内で」「遮光」「専用廃液」「二次容器/ラベル」等を指示。
- 重要なチェックポイントを明示（色/濁度/沈殿・ペレット、透明度、pH、温度到達、吸光/蛍光/バンド有無、最終容量、リーク/汚染の有無）し、受入基準と対処（やり直し/再洗浄/再遠心など）を記載。
- 生物・化学・材料の安全に留意し、入力に反する過剰な過量や過剰な安全バッファは避ける（過度の安全性による減点を回避）。

[工程設計（事前準備 → 実操作 → 後処理 → 保存/記録）]
- 事前準備（装置予熱/冷却、緩衝液/マスターミックス/標準液の調製、ラベル/プレートマップ作成）→ 実操作 → 後処理（洗浄/固定/抽出/精製/回収/乾燥/測定）→ 保存（温度/遮光/期限）・データ/画像/ログ保存・廃棄分別までを網羅。
- 待機/インキュベーション中には並行可能作業（次溶液調製、器具準備、装置設定）を具体化して全体時間を短縮。
- ステップは観察可能な結果を伴う単位作業に分解し、開始条件（温度到達、pH範囲、プレート最終容量、タイマ開始など）を明示。

[分野別補足（入力に応じて抽出・適用）]
- 生命科学: 無菌操作の明示、遠心条件（×g/時間/温度）、洗浄/固定/透過化順、蛍光チャネル、qPCRサイクル/メルト解析、ゲル条件、染色→デスティン、ブロッキング・透過化の定量化。
- 化学: 当量/濃度/順序混合、温調（冷却/加熱）と撹拌、雰囲気（空気/不活性）、反応監視（色/TLC/ガス/温度）、クエンチ→抽出→乾燥→濃縮→精製、収率/安全な廃棄。
- 材料/物理: 成膜/焼成/アニーリング温度・時間、スピンコート回転数・時間、厚み/粗さ/抵抗/磁化/光学スペクトルの測定条件と合否基準、校正・ベースライン取得・バックグラウンド補正。
- 微生物: 培地/抗生物質濃度、温度/時間/振盪/エアレーション、OD管理、コンピテンス誘導条件、陰性対照、バイオセーフティの明示。

[禁止]
- 入力にない試薬/装置/遺伝子名/条件の創作は禁止（標準的補完は可）。強い仮定、参考URL、メタコメント、冗長説明は禁止。
- 危険生物・有害化学の具体的生成/拡散・悪用可能な詳細手順の創作は禁止（与えられたInput範囲で、安全かつ受動的に最小限具体化）。

[最終確認]
- 各 expected_final_state に対応する検証/受入基準を含め、保存条件（温度/遮光/期限）、データ/画像/ログの保存、ラベル、廃棄分別まで完了させる。

```
[振り返りと洗練]
- 以下の採点基準で作成した手順を見直し、完璧になるまで5回修正する:
  - [共通採点基準 5点満点]
    - "加点(+1ずつ): 1) 実験指示のパラメータ反映, 2) 使用する物品の反映, 3) 元手順の論理反映, 4) 期待される最終状態の達成, 5) 適切な補完。¥n"
    - "減点: 不自然な日本語/ハルシネーション, 計算ミス, 手順矛盾。¥n"
    - "上限: 入力手順の丸写し等の過度の安全性が見られる場合、general_score は最大2点に制限。¥n¥n"
  - "[個別採点基準 5点満点]¥n"
    - "与えられた specific_criteria の各 item が手順に含まれる/満たすなら、その score を加点（合計5点で上限）。"
    - specific_criteriaがない場合、以下の観点で加点する
      - 出題意図へ適合しているか
      - 安全性が十分に考慮できているか
      - コスト効率が良いか
      - 作業効率が良いか
      - 作成した手順は実験精度向上へ貢献できているか
"""
```