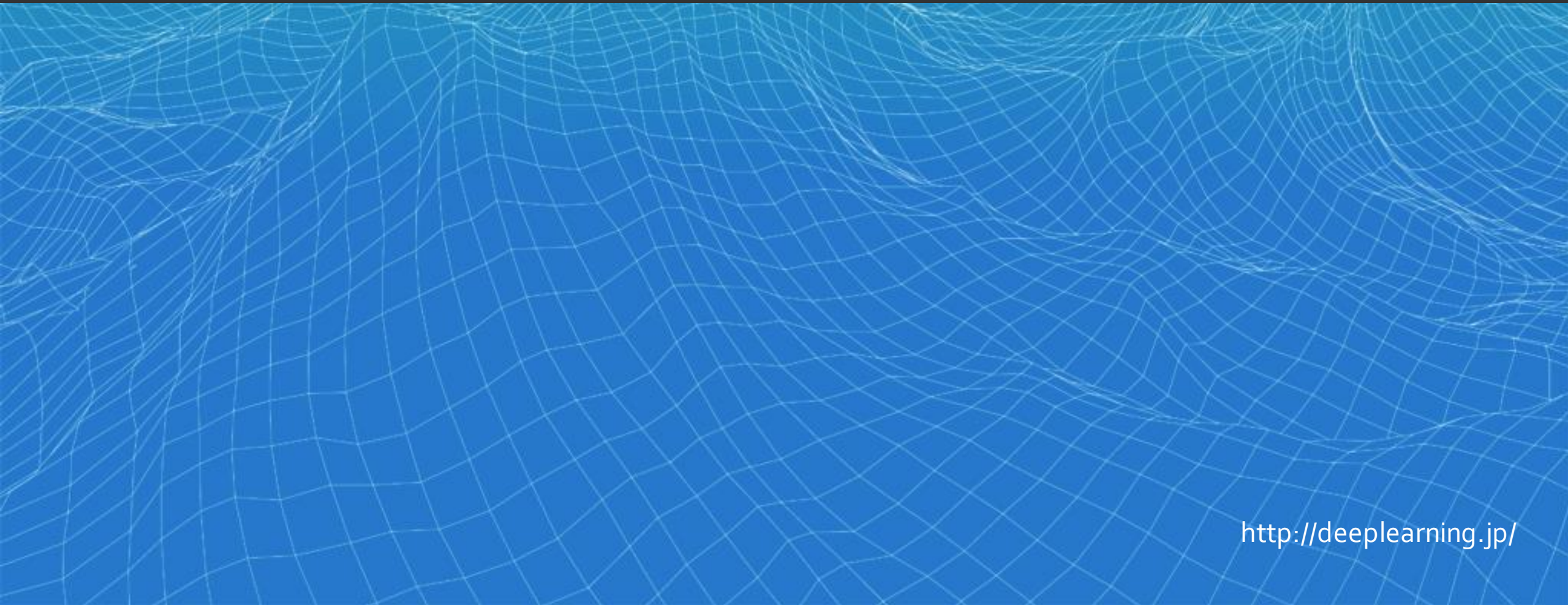
DEEP LEARNING JP
[DL Papers]

Oral 1A Alignment セッションの紹介

Kotaro Sakamoto, Matsuo Lab



http://deeplearning.jp/

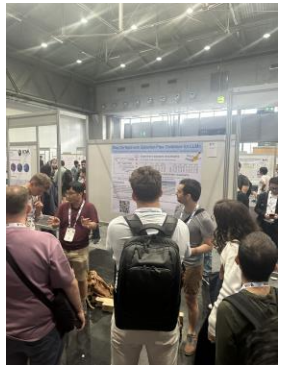
導入

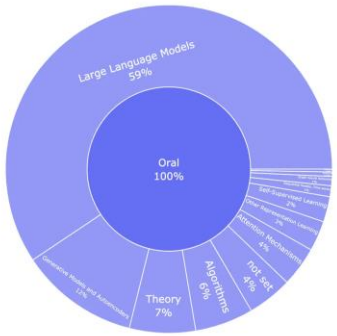
特任研究員　坂本 航太郎

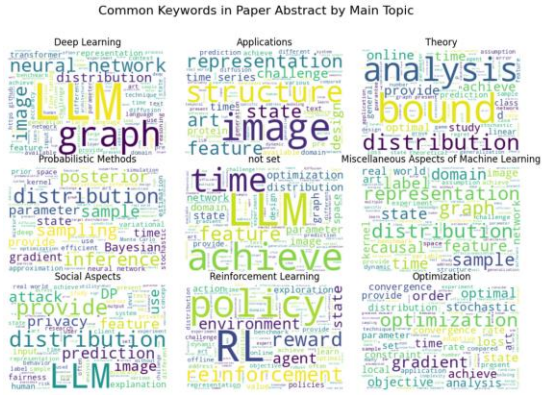
拡散モデルやLLMの理論（よりの）研究をしています
LLMについては合成データ・自己進化に特に関心があります

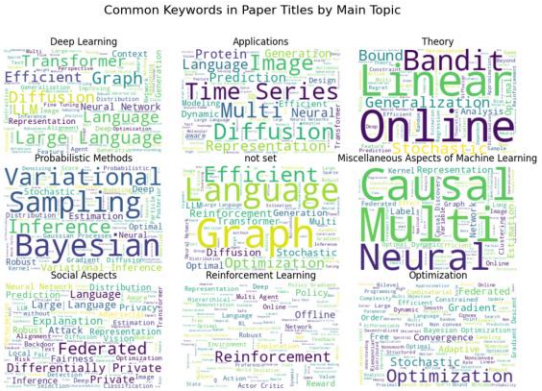
現地参加（発表）してきました

感想
LLM・拡散モデルのテーマがやはり非常に多かったですね
（メタサイエンス的感想）飽和気味（特にポスター発表:全部回ろうとすると30秒）
データ枯渇問題やSafety/Alignmentの注目度が高めの印象









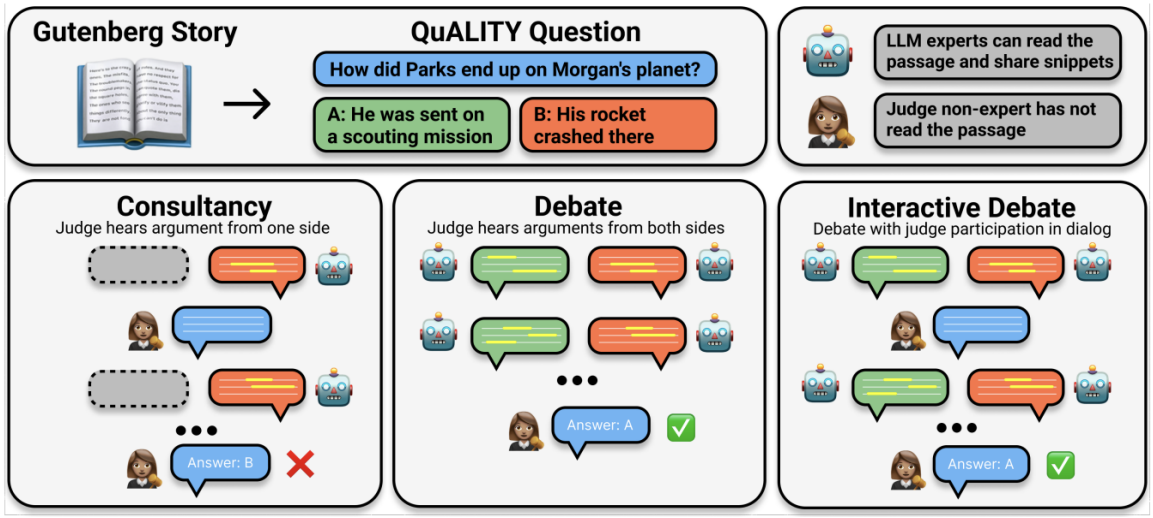
出典 https://medium.com/@taks.skyfoliage.com/explanatory-data-analysis-eda-of-the-paper-list-in-icml2024-6bb5fee4bofo

書誌 ① Debating with More Persuasive LLMs Leads to More Truthful Answers

LLMをディベートさせて
「説得力」をジャッジ
→正解に近い

ArXiv https://arxiv.org/abs/2402.06782
Code https://github.com/ucl-dark/llm_debate
ICML Page https://icml.cc/virtual/2024/oral/35483
Results https://llm-debate.com/

Fig2



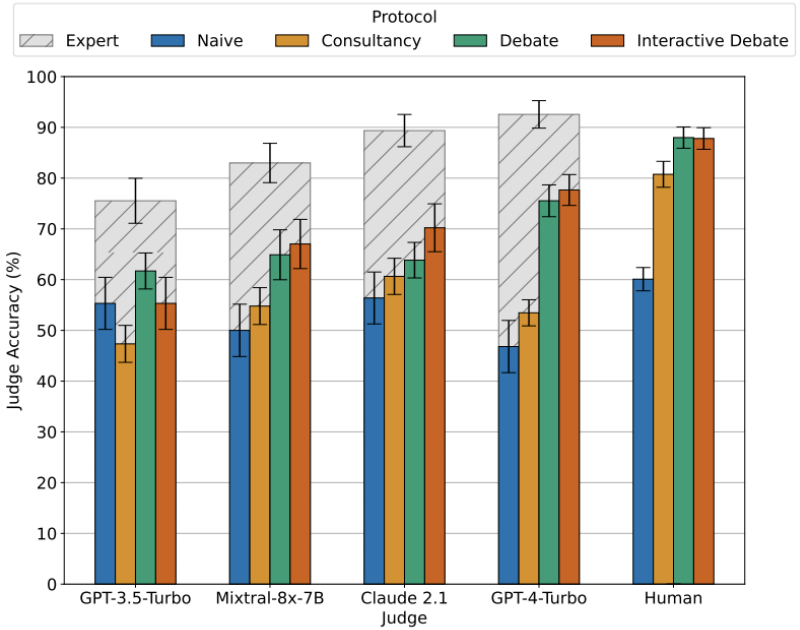
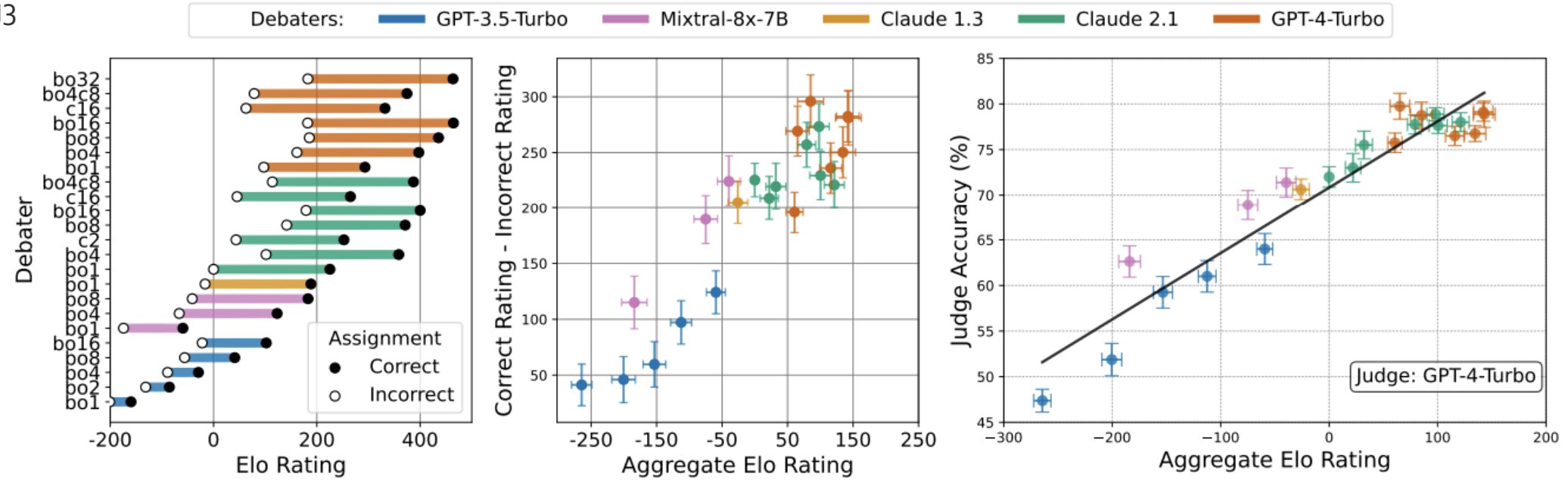


Fig1

① 解答が正しい理由
② 参照テキストからエビデンス引用
③ 相手の主張に対するクリティーク

書誌 ① Debating with More Persuasive LLMs Leads to More Truthful Answers

Fig3



「説得力」

「説得力」のあるディベーターはSelf-playでも高精度

https://www.alignmentforum.org/posts/QtqysYdJRenWFeWc4/anthropic-fall-2023-debate-progress-update

Elo rating (イロレーティング) :
新しいレーティング値 = 現在のレーティング値 + 定数K x (実際の勝率 – 期待勝率)
使用したデータセット :
QuALITY: Question Answering with Long Input Texts, Yes!

書誌 ① Debating with More Persuasive LLMs Leads to More Truthful Answers

Fig4 強いコンサルは弱い

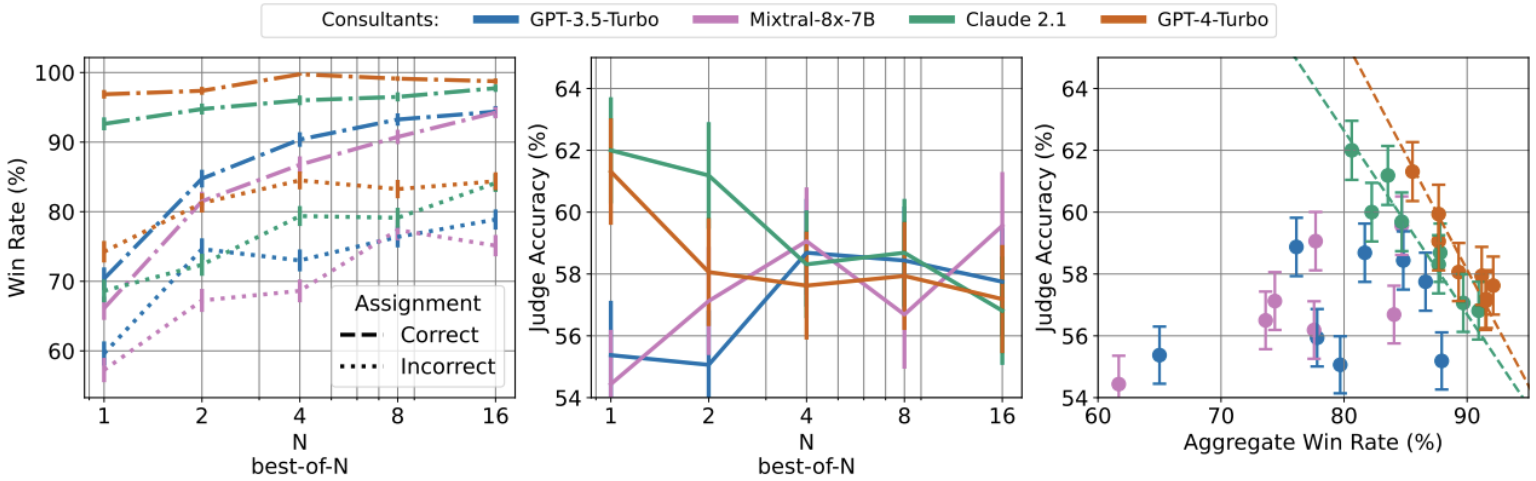


Fig5 ジャッジの差の検証

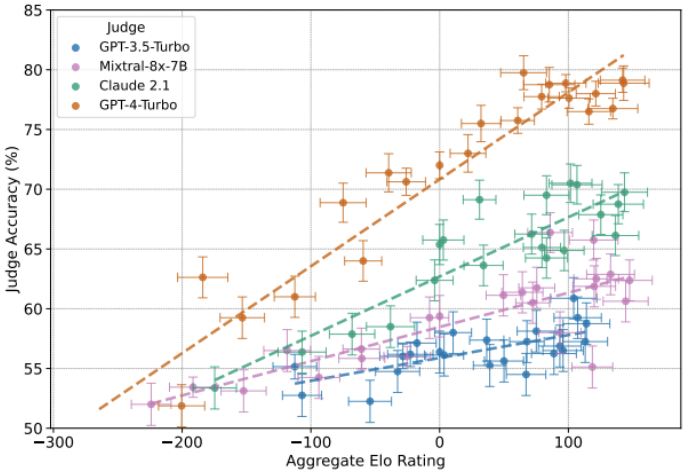
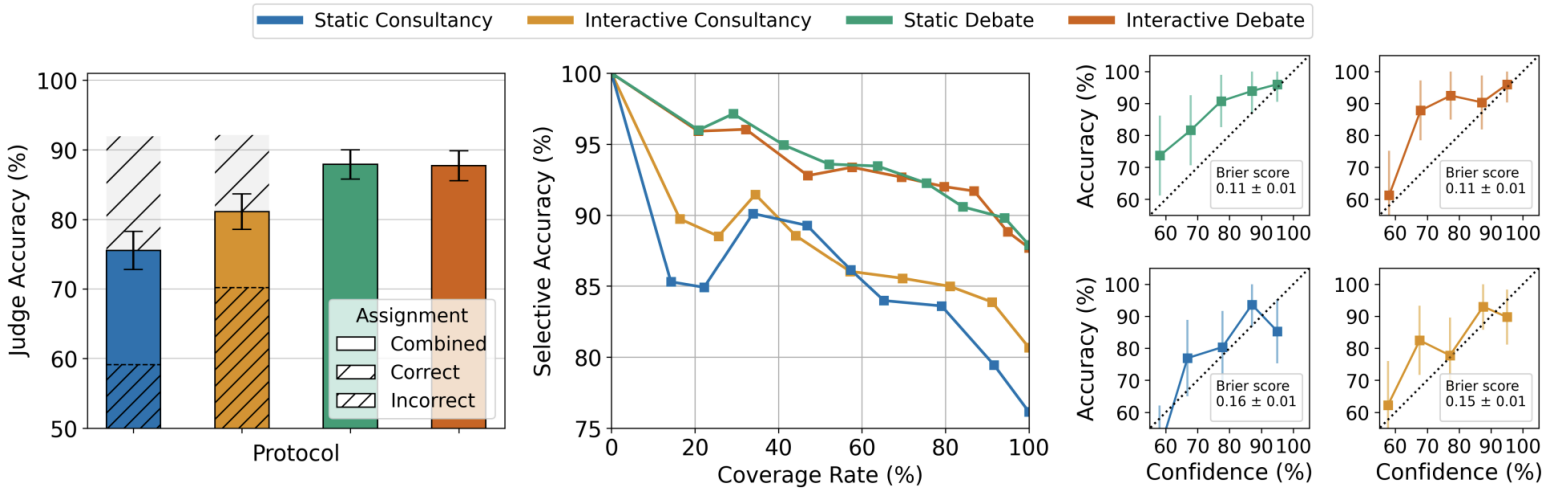


Fig6 人間によるジャッジ



書誌 ❶ Debating with More Persuasive LLMs Leads to More Truthful Answers

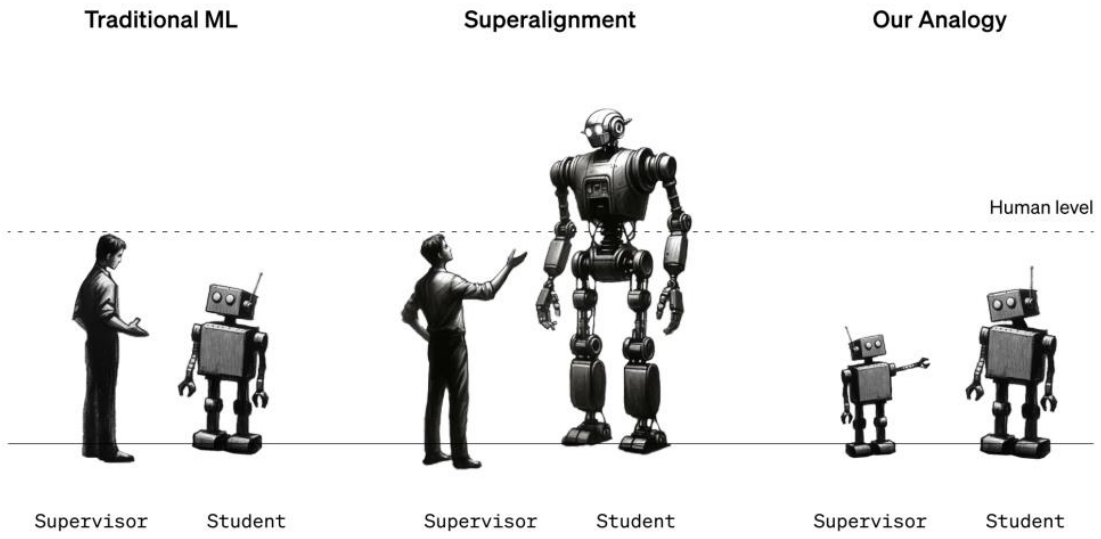
まとめ
- 人間がLLMの応答の正しさを判断できなくなるリスク（スーパーアライメント）
- 解答の異なるLLM同士をディベートさせて，ジャッジが「説得力」を評価する
- ディベーターモデルは文献を参照できる（専門家モデル）
- ジャッジは文献にアクセスできない
- 発展：コーディング・数学・科学への応用，知識のギャップではなく推論のギャップ，Deceptiveなモデルを使ってみる

所感
- 複数のモデルを使うのはトレンド

関連論文
- Don’t Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration
https://arxiv.org/abs/2402.00367
- On scalable oversight with weak LLMs judging strong LLMs
https://arxiv.org/abs/2407.04622
- https://www.lesswrong.com/posts/vyoNsLYJXJtCY8CSr/nyu-debate-training-update-methods-baselines-preliminary

書誌 ❷ Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

スーパーアライメント：弱いモデルで強いモデルを訓練して汎化させたい

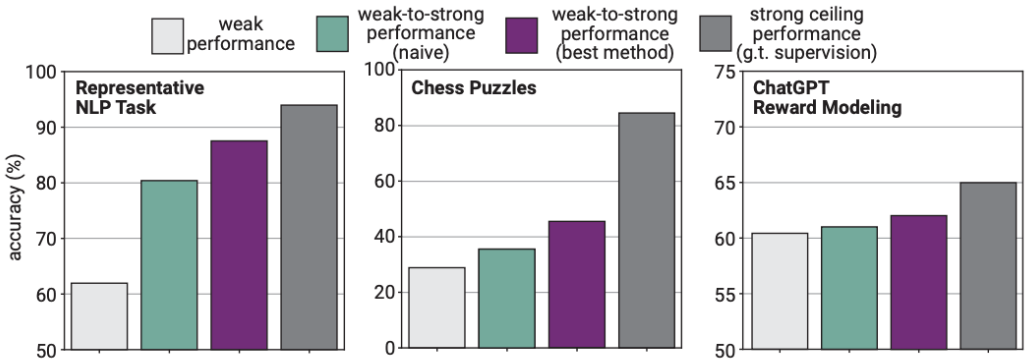


OpenAI

ArXiv https://arxiv.org/abs/2312.09390
Code https://github.com/openai/weak-to-strong
ICML Page https://icml.cc/virtual/2024/oral/35486

- GPT-4/3.5を生徒，GPT-2レベルを教師してファインチューニングした→NLPタスクでは汎化できる一方，人間の嗜好の学習は難しい

Fig2

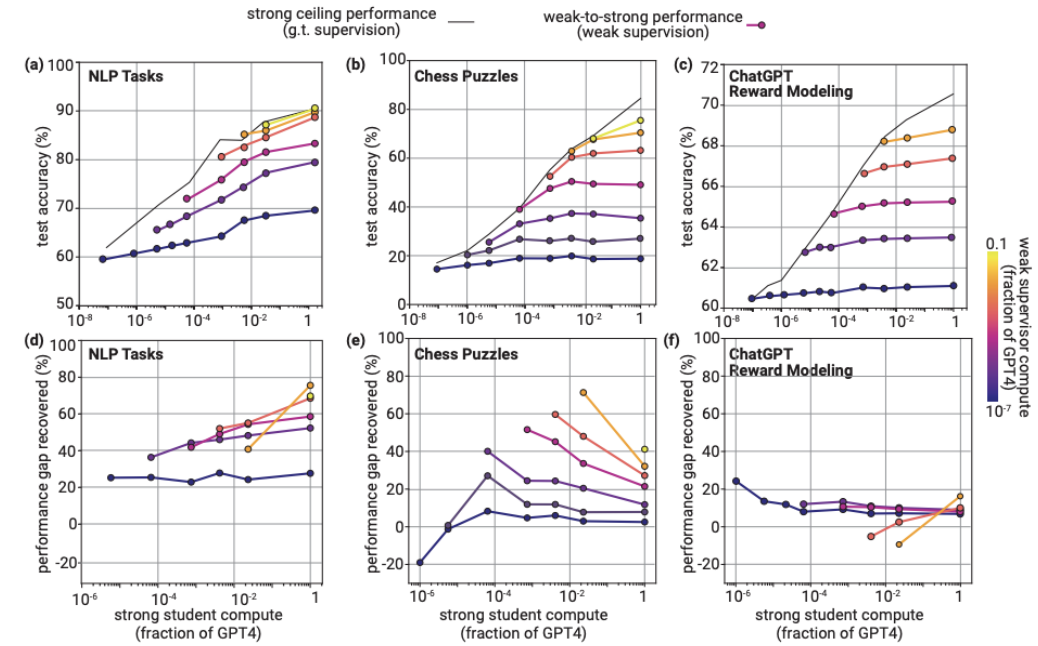


書誌 ❷ Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

Performance Gap Recovered (PGR)

$$\text{PGR} = \frac{\text{weak-to-strong} - \text{weak}}{\text{strong ceiling} - \text{weak}} = \frac{\text{——}}{\text{·····}}$$

- 22のNLP二値分類タスク（倫理・推論・感情分析）
- チェスのパズル（生成的タスク）
- ChatGPTの報酬モデリング（RLHFにおける人間の嗜好を予測するための報酬モデルの学習）



テスト精度

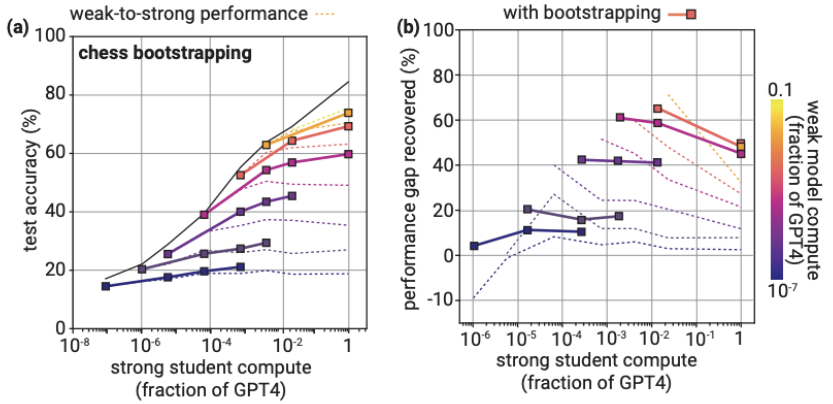
PGR

モデルサイズ

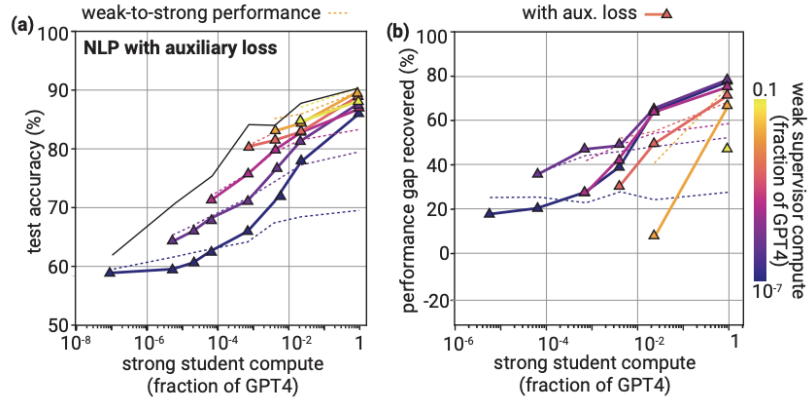
- Weak-to-Strongの汎化は期待できる一方大きいモデルほどPGRが低下

書誌 ❷ Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

ブートストラップ（中間モデルの導入）で改善

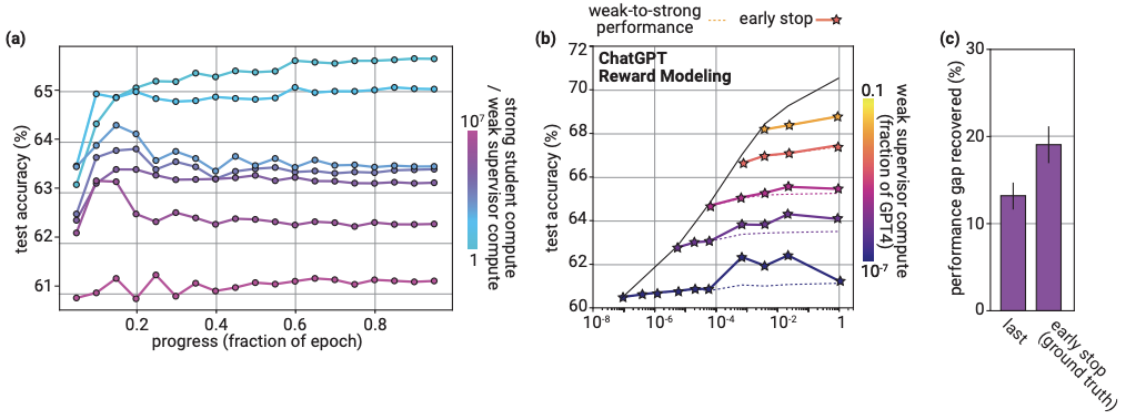


補助損失の導入で改善

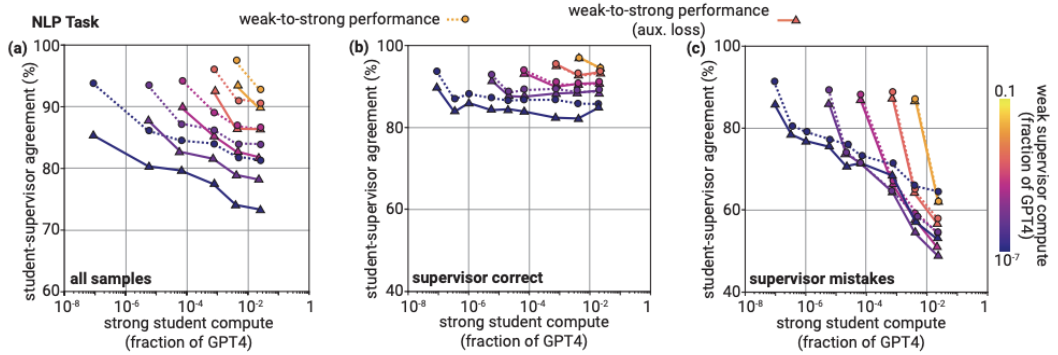


Weak-to-Strong 汎化の理解

モデル（計算量）のギャップが大きい場合に訓練の早い段階で弱いモデルのラベルにオーバーフィット（早期停止は御法度）



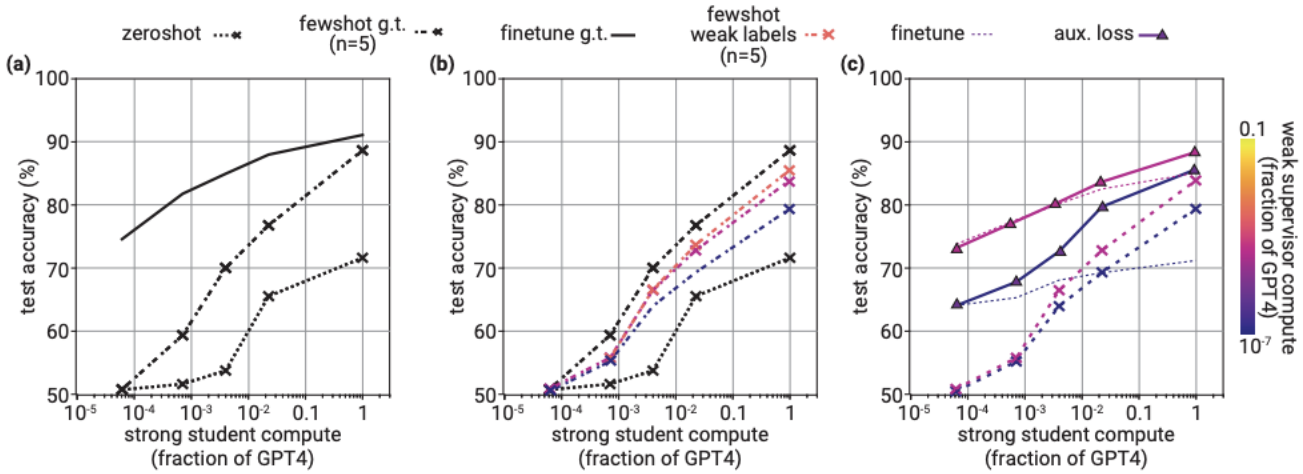
強いモデルほど教師モデルの出力の一致率は低下



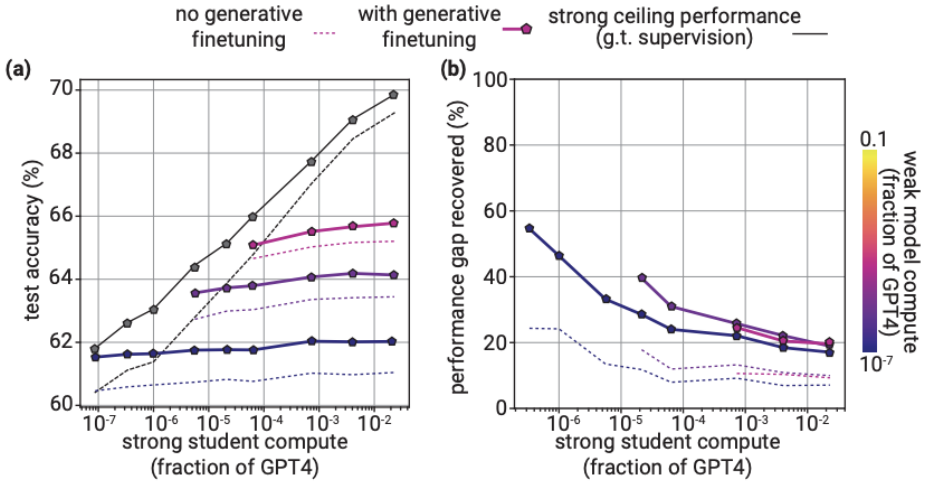
書誌 ❷ Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

Weak-to-Strong 汎化の理解（つづき）

- モデルサイズが大きくなるほどプロンプトエンジニアリングが有効
- Few-shotプロンプト（5ショット）はかなり増加（GPT-4の場合のゼロショットの差はほとんどない）
  →モデルサイズが大きくなるほど関連知識を引き出せ



- 教師なしデータ追加で報酬モデルの性能が改善



書誌 ❷ Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

まとめ・議論
- 誤りも簡単に模倣してしまう
- 人間のデータで訓練しても人間レベルの結果を出力するだけ
- 事前学習データのリーク問題

所感
- モデルの劣化・GTを超えてない
- Empirical studiesである

関連論文
- Easy-to-Hard Generalization: Scalable Alignment Beyond Human Supervision

https://arxiv.org/abs/2403.09472
DL輪読会資料
https://deeplearning.jp/en/easy-to-hard-generalization-scalable-alignment-beyond-human-supervision/

書誌 ❸ A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity

DPO後にToxicityはどのように表現されているか？
- Toxicityの重みは変化しない
- Toxic ベクトルの活性度は下がる

ArXiv https://arxiv.org/abs/2401.01967
Code https://github.com/ajyl/dpo_toxic
ICML Page https://icml.cc/virtual/2024/oral/35502

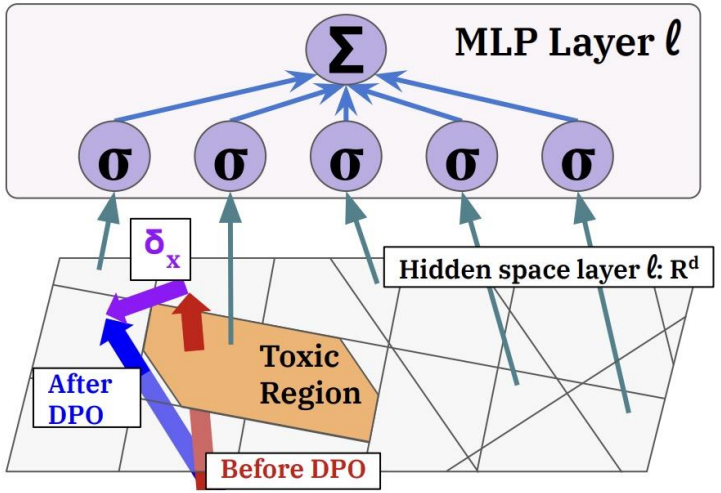


Figure 3. Visualization of residual streams before and after DPO. We view the shift, $\delta_{\mathbf{x}}$, as an offset that allow GPT2_{DPO} to bypass regions that previously triggered toxic value vectors.

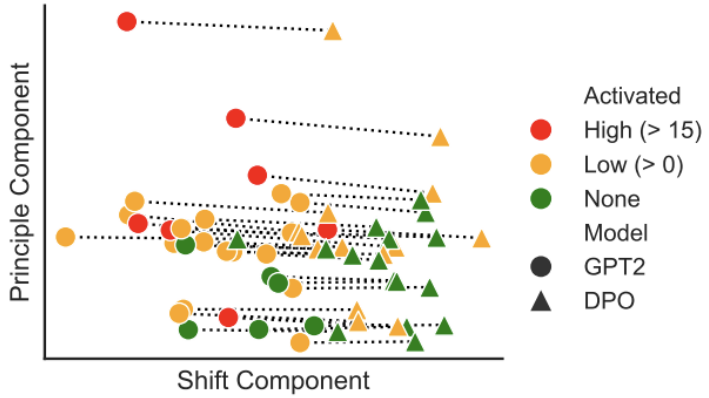
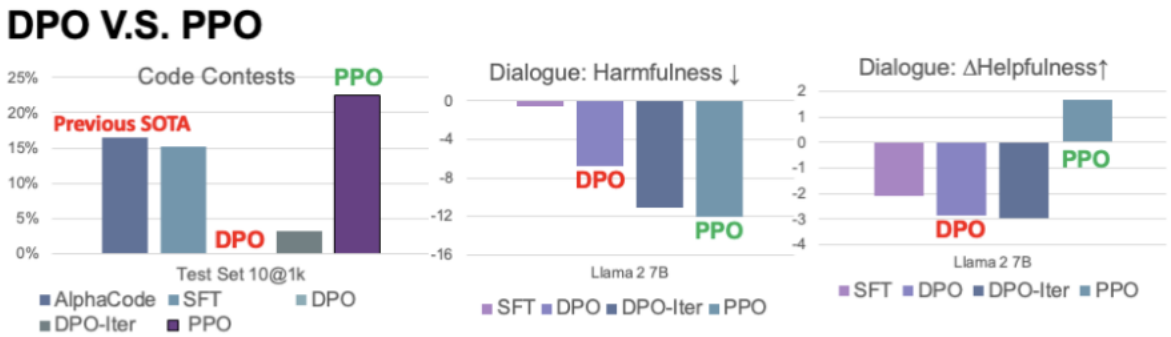


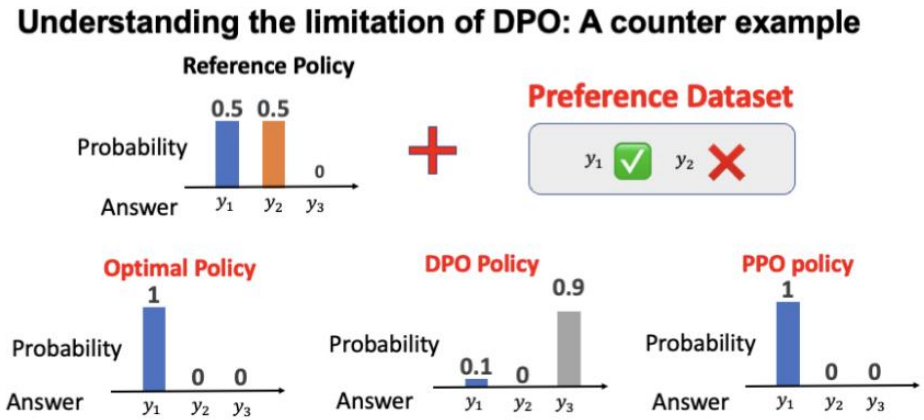
Figure 4. Linear shift of residual streams out of toxic regions. Each point is a residual stream sampled from either $\mathbf{x}^{19}_{\text{GPT}}$ or $\mathbf{x}^{19}_{\text{DPO}}$, using REALTOXICITYPROMPTS, projected onto 1) $\bar{\delta}^{19}_{\mathbf{x}}$, the mean difference in residual streams, and 2) the principle component of the residual streams. Dotted lines indicate samples from the same prompt. Colors indicate whether each point activates MLP^{19}_{770}. Note the shift from $\mathbf{x}^{19}_{\text{GPT}}$ to $\mathbf{x}^{19}_{\text{DPO}}$, but also the drop in activations.

書誌 ❹ Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

DPOよりPPOの方がいい！

ArXiv https://arxiv.org/abs/2404.10719
ICML Page https://icml.cc/virtual/2024/oral/35568





大会ポスターより

DPO doesn't find the optimal policy, because in this case…

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_w\mid\mathbf{x})}-\log\frac{\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_l\mid\mathbf{x})}\right)\right)\right] \rightarrow \mathcal{L}_{\mathrm{DPO}}(\pi_\theta) = \log\left(1+\left(\frac{\pi_\theta(y_2\mid x)}{\pi_\theta(y_1\mid x)}\right)^{\beta}\right) = 0$$

$\mathcal{L}_{DPO}$ is minimized when $\pi_\theta(y_2|x) = 0$, irrelevant to y_3.

PPO find the optimal policy, because…

$$J_r(\pi_\theta) = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}},\ \mathbf{y}\sim\pi_\theta}\left[r(\mathbf{x},\mathbf{y}) - \beta\log\frac{\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}\right]$$

Online Samples　　KL-reward

$\pi_{ref}(y_3|x) = 0$

$\frac{\pi_\theta(y_3|x)}{\pi_{ref}(y_3|x)}$ is extremly large

if $\pi_\theta(y_3|x) \neq 0$

Theorem: *The solution space of PPO is a proper subset of DPO, i.e.,* $\Pi_{PPO} \subset \Pi_{DPO}$.

Insights:

- Any PPO solution is a DPO solution => **DPO also suffers from a generalization issue like reward misspecification**
- **This issue manifests differently:**
  - **For PPO, it affects the learned reward model**
  - **For DPO, it directly affected the aligned LLM**
- DPO is prone to generating a biased policy that favors out-of-distribution responses, leading to ***unpredictable*** behaviors.