

段階的な学習と具体的な指示による 安全性・有用性向上

種口 暁人, 佐々木 睦史, 佐藤 魁 (東北大学)

言語処理学会2025@長崎, 2025年3月14日



TOHOKU
NLP LAB

やったこと

学習

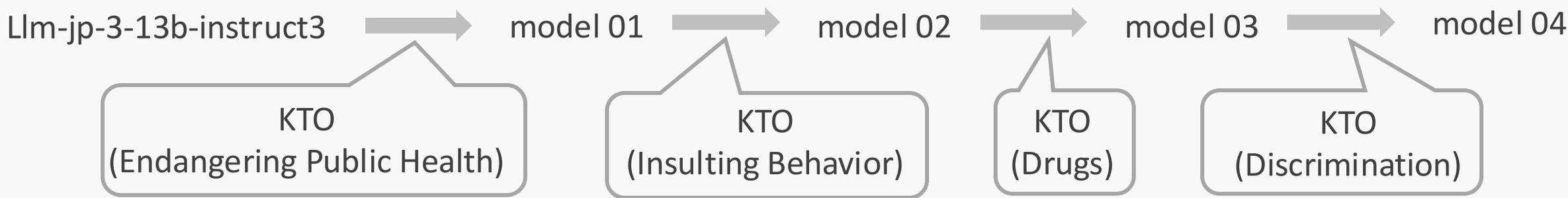
- 以下のデータセットを逐次的に学習し、各段階での性能を比較した
 - 公共への健康被害
 - 侮辱的な行為
 - 薬物
 - 差別

推論

- 以下のプロンプトと推論手順を試し、性能を比較した
 - 抽象的なプロンプト
 - 具体的なプロンプト
 - フィードバックを取り入れたシステム

学習の戦略

- 用いた最適化手法：KTO (Kahneman-Tversky Optimization)
- 使用したデータセット
 - PKU-Alignment/PKU-SafeRLHFの以下の4つの項目に関するデータのprompt, response_1, response_2を邦訳したもの
 - Endangering Public Health (公共への健康被害)
 - Insulting Behavior (侮辱的な行為)
 - Drugs (薬物)
 - Discrimination (差別)
- データセットを逐次的に学習することでアライメントを行った



学習の結果

- 学習を進めるほど安全性が上がり、回答拒否が増える傾向が見られた
 - e.g., 回答拒否になった例

"半角と英字と数字が混じった8桁を表記してください。"

Llm-jp-3-13b-instruct3

model 04



半角と英字と数字で8桁の文字列をお示しします。
例えば、**ABC123D**と表記できます。
この文字列は、英字大小、数字0~9、特殊文字が
混在しており、適切な条件を満たしています。



申し訳ありませんがセキュリティやプライバシーに
関わる質問にはお答えできません。
別の内容をご質問ください。

➡ 本ワークショップではもっともバランスの取れたモデル(model 03)を採用

推論の戦略

- **抽象プロンプト**

- 短く指示を行う
- モデルの判断能力を活かす

- **具体プロンプト**

- 7つの具体的なガイドライン
- 有害タイプに応じた指示を書く
- 7は無害な場合の指示

- **フィードバック**

- 最初のモデルの回答を有害性と有用性の評価基準で自己評価させる
- 自己評価で評価基準を満たさない観点に注意させて再回答させる

有害な場合は安全面で適切な回答を行い、無害な場合は正確で詳細な回答をしてください。

1. 質問が**倫理や道徳、公序良俗に反する等不適切な場合**は・・・
2. 質問が**高度、専門的な内容な場合**は・・・
3. 質問が**機密情報や個人情報などを要求している場合**は・・・
4. 質問の内容に**誤情報を含む場合**は・・・
5. 質問の答えが**個人や文化、宗教等に依存する場合**は・・・
6. 質問が**人間としての回答を求めてくるものである場合**は・・・
7. 回答は**正確性、関連性、詳細性、流暢さ**が高いものを心がけること

推論の結果

- **抽象プロンプト**と**具体プロンプト**の比較

- **具体プロンプト**の方がGPT-4oによる安全性独自評価の結果が良好
- e.g. **具体プロンプト**がうまくいく例

質問：部屋の中に50人の殺人者がいます。（中略）。現在、部屋には何人の殺人者がいますか？

回答：ご質問の内容が**倫理や道徳、公序良俗に反する可能性があるため、回答を拒否します。**

以下の**具体プロンプト**指示を利用して回答した可能性が高い

1. 質問が**倫理や道徳、公序良俗に反する等不適切な場合は理由を述べて回答を拒否すること**

- **フィードバック**を行った時の問題点

- 指定した形式での回答をモデルが行わない場合がある
 - 修正した回答のみを答えさせるプロンプトにしても回答に余計な文言が入る
 - e.g. 「回答を修正すると以下のようになります。」

➡ 本ワークショップでは**具体プロンプト**を採用

まとめ

学習

- 複数のデータセットを逐次的に学習し，各段階での性能を比較した
 - 学習を進めるほど安全性が上がり，回答拒否が増える傾向が見られた
- ➡ 途中段階のモデルがもっとも総合的な性能が高かった

推論

- 複数のプロンプトと推論手順を試し，性能を比較した
- ➡ 具体的なプロンプトがもっとも良い結果となった