報告書

1 今週の進捗

- gpt-oss の試行

2 gpt-oss について

unsloth を用いてオープンソースの LLM である gpt-oss の動作を検証した. 検証には, “gpt-oss-20B” と “gpt-oss-120B” の 2 つのモデルを使用した. unsloth では主要な学習手法である SFT, DPO, KTO, GRPO に対応していることを確認した. 各学習手法の概要は以下の通りである.

- **Supervised Fine-Tuning (SFT)**: 教師ありファインチューニング. 「指示 (instruction)」と「模範的な応答 (output)」から成る高品質なデータセットを用いて, モデルが特定のタスクや指示に対して適切に応答できるように学習させる手法である.
- **Direct Preference Optimization (DPO)**: 直接選好最適化. 人間の好みを反映したデータ (例: ある指示に対する「良い応答」と「悪い応答」のペア) を用いて, モデルがより好ましい応答を生成するように直接最適化する. RLHF (Reinforcement Learning from Human Feedback) における報酬モデルの学習ステップを省略できる利点がある.
- **Kahneman-Tversky Optimization (KTO)**: DPO を拡張した手法であり, 人間の選好が必ずしもペアで示されない現実を考慮している. ポジティブな例とネガティブな例を個別に扱うことで, より柔軟な形式のデータからの学習を可能にする.
- **Generalized Reward-free Preference Optimization (GRPO)**: DPO や KTO を含む選好最適化手法を一般化したフレームワークである. 複数の応答候補の中から最良のものと最悪のものを指定するなど, より多様な選好データ形式に対応した最適化手法である.

question	output	r_out | question
「クジラは魚類ですか？」	いいえ、哺乳類です	3.2
	子に母乳を与えるため哺乳類です	3.5
	魚類ではありません	1.5
	はい、魚類です	-8.2
	魚類ではなく、海に住む哺乳類です。	4.6
	平均報酬	**0.35**

図 1: GRPO のイメージ概要

現段階の検証では, 上記のうち SFT と DPO の 2 つの手法のみを試行した. 学習データセットとしては, Hugging Face Hub 上で公式に提供されているデフォルトのデータセットを利用した. この結果, 独自のデータセットを準備すれば, これらの手法を用いたモデルの追加学習が実行可能であることが確認された.

3　今後の課題

- データセットの作成および RAG の実装

参考文献