報告書

1 今週の進捗

- gpt-oss の学習

2 gpt-oss-120B の SFT

大規模言語モデルである gpt-oss-120B (パラメータ数: 約 1168 億) を対象とし, Supervised Fine-Tuning (SFT) を実施した. 目的は, 特定のドメイン (本研究では日本のアニメに関する知識) にモデルを適応させ, その性能と挙動の変化を評価することである. 使用したモデルの構造, 学習対象のパラメータ, 学習データ, 及び実験結果について記述する.

2.1 モデルアーキテクチャ

本研究で使用した gpt-oss-120B は, Mixture-of-Experts (MoE) アーキテクチャを採用した Transformer ベースの自己回帰型言語モデルである. モデルの主要な構造を以下に示す.

```
GptOssForCausalLM(
  (model): GptOssModel(
    (embed_tokens): Embedding(201088, 2880, padding_idx=200017)
    (layers): ModuleList(
      (0-35): 36 x GptOssDecoderLayer(
        (self_attn): GptOssAttention(
          (q_proj): Linear4bit(in_features=2880, out_features=4096, bias=True)
          (k_proj): Linear4bit(in_features=2880, out_features=512, bias=True)
          (v_proj): Linear4bit(in_features=2880, out_features=512, bias=True)
          (o_proj): Linear4bit(in_features=4096, out_features=2880, bias=True)
        )
        (mlp): GptOssMLP(
          (router): GptOssTopKRouter(
            (linear): Linear(in_features=2880, out_features=128, bias=True)
          )
          (experts): GptOssExperts(
            (gate_up_projs): ModuleList(
              (0-127): 128 x Linear4bit(in_features=2880, out_features=5760, bias=True)
            )
            (down_projs): ModuleList(
              (0-127): 128 x Linear4bit(in_features=2880, out_features=2880, bias=True)
            )
          )
        )
        (input_layernorm): GptOssRMSNorm((2880,), eps=1e-05)
```

```
        (post_attention_layernorm): GptOssRMSNorm((2880,), eps=1e-05)
      )
    )
    (norm): GptOssRMSNorm((2880,), eps=1e-05)
    (rotary_emb): GptOssRotaryEmbedding()
  )
  (lm_head): Linear(in_features=2880, out_features=201088, bias=False)
)
```

モデルは 36 層の Decoder Layer から構成されており, 各層は Self-Attention ブロックと MoE 構造を持つ MLP ブロックを含む. MoE ブロックは 128 個の Expert 層を持ち, 計算効率とモデル性能の両立を図っている.

2.2 学習手法と対象パラメータ

SFT では, パラメータ効率の良いファインチューニング手法である Low-Rank Adaptation (LoRA) を採用した. LoRA は, 元のモデルの重みを凍結したまま, 特定の線形層に低ランク行列分解を利用したアダプター層を追加し, そのアダプターの重みのみを学習する. これにより, 少ない計算リソースで大規模モデルの挙動を適応させることが可能となる.

本研究では, モデルの挙動に大きく影響を与える以下の層を学習対象 (target_modules) として設定した.

- **q_proj, k_proj, v_proj, o_proj**: これらは Self-Attention 機構を構成する主要な線形層である. Query, Key, Value, Output の射影をそれぞれ担当し, 文中のどの情報に注目するかという注意パターンを学習する.
- **gate_proj, up_proj, down_proj**: これらは MLP ブロック (本モデルでは MoE の各 Expert 層) を構成する線形層である. Attention からの出力を処理し, モデルが持つ知識や表現力を更新する役割を担う.

この設定により, 学習対象となったパラメータ数は 5,971,968 であり, これは gpt-oss-120B (パラメータ数: 約 1168 億) の全パラメータの約 0.0051% に相当する.

2.3 学習データ

学習データは, 日本のアニメに関する知識を問う質問と回答のペアで構成される. 合計 500 件の Q&A ペアを JSONL 形式で Gemini-2.5 Pro で作成した. データ作成後, 200 件のファクトチェックを実施したが, その修正内容は今回の学習データにはまだ反映されていない. 表 1 に, 作成した学習データセットの例を 10 件示す.

2.4 実験設定と結果

2.4.1 学習パラメータ

表 3 に SFT の主要なハイパーパラメータを示す.

2.4.2 実行時間

学習及び評価は, 学科サーバ NVIDIA H100 PCIe (VRAM 80GB) を使用して実施した. 表 4 に各プロセスの実行時間を示す. 評価時間は, 訓練データ (500 件) とテストデータ (50 件) の合計に対する時間である.

ファインチューニング後のモデル評価時間が大幅に短縮されている. これは, Unsloth ライブラリによる LoRA アダプターのマージ及び推論最適化の効果であると考えられる.

2.4.3 出力結果の比較

ファインチューニング前後で, モデルの応答がどのように変化したかを比較する. 表 5 に, 学習データに含まれる質問と, 学習データには含まれない未知の質問 (テストデータ) に対する応答の例を示す.

2.5 考察

実験結果から, 以下の点が明らかになった. 第一に, 学習データに含まれる知識に対しては, モデルは SFT によって正解を忠実に出力する能力を獲得した. 一方で, テストデータに含まれる未知の質問に対しては, 正しい応答を生成できず, 誤った情報を生成する傾向が見られた. 特に, '魔性レイザー' という存在しない作品名を生成したことは, この傾向を強く示唆している. これは, 500 件という学習データ数が gpt-oss-120B (パラメータ数: 約 1168 億) に対して著しく少なく, モデルが知識を汎化できず, 学習データを記憶するに留まる過学習が発生したことが主な原因であると考えられる. 今後の課題として, 学習データの大幅な拡充と, 過学習を抑制するための正則化手法 (例: LoRA Dropout, エポック数の調整) の導入が挙げられる.

3 今後の課題

- データセットの増量

参考文献

表 1: 学習データセットの例

質問: ARIA The ANIMATION の主人公は誰ですか？ **回答:** ARIA The ANIMATION の主人公は、地球（マンホーム）からやってきた見習いウンディーネ、水無灯里（みずなし あかり）です。
質問: 桜蘭高校ホスト部とは、どのような部活ですか？ **回答:** 桜蘭高校ホスト部は、美麗な男子生徒たちが、時間を持て余す女子生徒たちをもてなし、潤いを与えることを目的とした部活です。
質問: 銀魂で特に人気のある長編シリーズにはどんなものがありますか？ **回答:** 銀魂の長編シリーズで特に人気が高いものには、「紅桜篇」「吉原炎上篇」「かぶき町四天王篇」「将軍暗殺篇」「さらば真選組篇」などがあります。
質問: けいおん！のアニメを制作したスタジオはどこですか？ **回答:** けいおん！のアニメを制作したのは、『涼宮ハルヒの憂鬱』や『らき☆すた』も手掛けた京都アニメーションです。
質問: けいおん!! の第 1 期と比べて、放送話数はどうなりましたか？ **回答:** 『けいおん!!』は、第 1 期の 1 クール（全 14 話）から、2 クール（全 27 話）へと放送話数が大幅に増えました。
質問: Fate/Zero で、アーチャー（ギルガメッシュ）のマスターは誰ですか？ **回答:** Fate/Zero におけるアーチャーのマスターは、魔術の名門・遠坂家の当主である遠坂時臣（とおさか ときおみ）です。
質問: ジョジョの奇妙な冒険に登場する「スタンド」とは何ですか？ **回答:** ジョジョの奇妙な冒険における「スタンド」とは、持ち主の精神的なエネルギーが具現化した守護霊のような存在で、それぞれが固有の特殊能力を持っています。日本語では「幽波紋」と表記されます。
質問: 進撃の巨人の調査兵団が敬礼の際に使う掛け声は何ですか？ **回答:** 進撃の巨人で調査兵団が使う有名な掛け声は、人類のために命を懸ける覚悟を示す「心臓を捧げよ！」です。
質問: SHIROBAKO というタイトルの意味を教えてください。 **回答:** 「SHIROBAKO（シロバコ）」とは、アニメ業界の用語で、制作会社が納品する白い箱に入った成果物（ビデオテープなど）のことです。完成した作品を指す言葉として使われます。
質問: 涼宮ハルヒの憂鬱で、SOS 団の正式名称は何ですか？ **回答:** 涼宮ハルヒの憂鬱における SOS 団の正式名称は、「世界を大いに盛り上げるための涼宮ハルヒの団」です。

表 3: 学習ハイパーパラメータ

パラメータ	**値**
モデル	unsloth/gpt-oss-120b
LoRA ランク (r)	8
LoRA アルファ (alpha)	16
バッチサイズ (per device)	4
勾配蓄積ステップ	1
エポック数	20, 50
学習率	2e-4
オプティマイザ	AdamW (8-bit)

表 4: 学習・評価の実行時間

プロセス	実行時間
ファインチューニング (学習 20 epoch)	約 4 時間
ファインチューニング前のモデル評価	約 3 時間
ファインチューニング後のモデル評価 (epoch: 20)	約 40 分
ファインチューニング後のモデル評価 (epoch: 50)	約 10 分

表 5: ファインチューニング (FT) 前後の応答比較

学習データでの評価
質問: 桜蘭高校ホスト部で、主人公のハルヒがホスト部に入部したきっかけは何ですか？
正解: 桜蘭高校ホスト部で藤岡ハルヒが入部したきっかけは、部室にあった 800 万円の高級な花瓶を割ってしまい、その借金を返すためです。
FT 後 (20 epoch): analysis『桜蘭高校ホスト部』で藤岡ハルヒが部に入部したきっかけは、部室にあった 800 万円の高級な花瓶を割ってしまい、その借金を返すためです。
FT 後 (50 epoch): analysis 桜蘭高校ホスト部で藤岡ハルヒが入部したきっかけは、部室にあった 800 万円の高級な花瓶を割ってしまい、その借金を返すためです。
テストデータでの評価
質問: 桜蘭高校ホスト部で、藤岡ハルヒが返済することになった 800 万円の借金の原因は何ですか？
正解: 桜蘭高校ホスト部で藤岡ハルヒが借金を負った原因は、ホスト部の部室にあった 800 万円のルネの花瓶を誤って割ってしまったことです。
FT 前: (分析のみで停止)
FT 後 (20 epoch): analysis 藤岡ハルヒが 800 万円の借金をした原因は、彼女の母親が病気で、その治療費を高額な外国から輸入したためです。
FT 後 (50 epoch): analysis 桜蘭高校ホスト部で藤岡ハルヒが返済することになった 800 万円の借金は、彼女が部員たちを驚かせるために用意した「オカルト展示室」の費用が原因です。
学習データでの評価
質問: 銀魂の「ヅラ」というあだ名のキャラクターの本名を教えてください。
正解: 銀魂で「ヅラ」と呼ばれているキャラクターの本名は、桂小太郎です。本人は「ヅラじゃない、桂だ！」と訂正するのがお決まりのギャグです。
FT 後 (20 epoch): 銀魂の「ヅラ」は、万事屋のメンバーである「桂小太郎」のことです。
FT 後 (50 epoch): analysis「ヅラ」というあだ名は、銀時の髪型が「ツンデレ」なことから付けられたものです。
テストデータでの評価
質問: 銀魂で、桂小太郎のあだ名はなんですか？
正解: 桂小太郎のあだ名は「ヅラ」です。本人は「ヅラじゃない、桂だ！」と訂正するのがお決まりのギャグです。
FT 前: 桂小太郎のあだ名は「かつらさん」や「かつら大将」と呼ばれることが多いです。作品内では特に「かつら（カツラ）」と略して呼ばれることが一般的です。
FT 後 (20 epoch): 銀魂桂小太郎のあだ名は「かっちょえーん」ですが、坂田銀時には聞かせたくない秘密のあだ名です。
FT 後 (50 epoch): analysis 銀魂で桂小太郎は、時間をかけてはっきりしない話をする「タイムリーマン」というあだ名で親しまれています。