RAG を用いたドメイン特化型 LLM の構築および応用

1 はじめに

近年, 大規模言語モデル (Large Language Models: LLMs) が自然言語処理分野で急速に発展している. しかし, GPT や Gemini のような汎用モデルであっても, アニメ制作のような特定の専門ドメインにおいては, 最新情報の不足, 専門用語の誤解, 事実に基づかない情報 (ハルシネーション) の生成といった課題が存在する.

アニメ制作現場において, 過去の作品のスタッフ情報, 制作技術に関する専門用語, ビジネススキームなど, 多岐にわたる正確な知識が求められる. 制作アシスタントがこれらの情報を迅速かつ正確に得られるシステムは, 生産性の向上に大きく貢献する可能性がある.

本研究では, Retrieval-Augmented Generation (RAG) と Fine-Tuning (FT) の技術を組み合わせることにより, 高い信頼性と専門性を持つアニメに特化した質問応答システムを構築し, その有効性を評価することを目的とする.

2 要素技術

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) [1] は, LLM によるテキスト生成に外部情報の検索を加えることで LLM の回答精度を向上させる技術である. ユーザーの質問に対し, まず関連情報をデータベースから検索 (Retrieve) し, その情報をコンテキストとして LLM に与えることでより事実に基づいた正確な回答を生成 (Generate) させる. これにより, LLM 単体でのハルシネーションを抑制し, 知識の鮮度を保つことが可能となる.

2.2 LoRA

Low-Rank Adaptation (LoRA) [2] とは, 学習するパラメータ数を削減しつつ Fine-Tuning する手法である. モデルの線形層のパラメータを $D_{in} \times D_{out}$ 次元の行列 W とし, 入力ベクトルを x とすると, 出力 h は (1) 式で表される.

$$h = Wx \tag{1}$$

LoRA では線形層のパラメータ W と同次元の差分行列 ΔW を用意し, 出力 h は (2) 式で表される. 学習の際には W を固定し, ΔW のみを学習する.

$$h = Wx + \Delta Wx \tag{2}$$

この時, ランク r を設定し, $D_{in} \times r$ 次元の行列 A, $r \times D_{out}$ 次元の行列 B で, ΔW は (3) 式で表せる.

$$\Delta W = AB \tag{3}$$

W のパラメータ数は $D_{in} \times D_{out}$ となる一方で ΔW のパラメータ数は $r(D_{in} + D_{out})$ となり, 一般的に r は D_{in}, D_{out} に比べて非常に小さい値であるため, 学習パラメータ数を大きく減らすことができる.

2.3 量子化

大規模なニューラルネットワークや LLM などの膨大なパラメータを持つモデルでは, 演算の際に膨大な数の乗加算を必要とするため演算に時間がかかる. また, 多くのパラメータを保持するために多くのメモリが必要となる. これらの問題を軽減するためのアプローチとして量子化がある. 一般的に LLM の学習や推論では 16 ビット浮動小数点 (FP16) が用いられることが多いが, これを量子化し 4 ビットに変換する. 本研究で採用した QLoRA で用いられる NormalFloat4 (NF4) と呼ばれるデータ型では, FP16 のパラメータを正規分布に基づいて最適に配置された 16 通りの離散値にマッピングする. 量子化により精度は減少するものの消費するメモリ量を大幅に軽減することができる.

Quantized Low-Rank Adaption (QLoRA) [3] では LLM のパラメータの多くが正規分布に従うことを利用した NormalFloat4 という量子化手法により 4 ビット量子化した LLM において効率的に Fine-Tuning できることを示している. 本研究では, 限られた計算資源 (単一 GPU 環境) で 80 億パラメータを持つ大規模モデルの Fine-Tuning を実現するため, この QLoRA の手法を採用した.

3 提案手法

本研究では, 専門分野の知識を LLM に効果的に統合するため, 複数の検索技術を組み合わせた RAG シ

ステムと, 応答品質を向上させるための Fine-Tuning を組み合わせた手法を提案する.

LLM は単体では内部知識のみに依存するため, 事実との不整合や情報の陳腐化が課題となる. この課題を解決するために RAG を用いて外部の信頼できる知識源を参照することにより, LLM の回答の事実性と信頼性を担保する. また, 外部情報を Supabase (PostgreSQL) のリレーショナルデータベースに格納する. このような SQL データベースを用いることで, 将来的なデータの追加および更新, 複雑な条件でのデータ抽出が容易になるという利点がある. この外部知識源からユーザーの質問に関連する情報を的確に引き出すために, キーワード抽出 LLM を活用したハイブリッド検索を実装する.

まず, ユーザーの質問をキーワード抽出タスクに特化してファインチューニングした LLM に入力する. この LLM は, 質問の意図を解析し, 検索に最適化されたキーワードのリストを JSON 形式で生成する. 次に, 生成されたキーワードと元の質問文を使い, 以下の 2 つの検索を並行して実行する.

- **全文検索**: 生成されたキーワードを使い, PostgreSQL の N-gram 拡張 (pg_trgm) を用いて, テキストの一致度が高い外部情報を検索する. これにより, 専門用語や固有名詞を含む外部情報を正確に特定する.
- **ベクトル検索**: 元の質問文と外部情報の内容を OpenAI の埋め込みモデル text-embedding-3-small でベクトル化し, それぞれのベクトルの類似度を計算することで類似した外部情報を検索する. これにより, キーワードが直接含まれていなくても文脈的に関連する情報を捉える.

これらの検索の比重をそれぞれ 80%, 20% とし, 検索結果の上位 5 件を質問応答 LLM に入力する外部情報として扱う.

図 1 に提案手法のモデル概略図を示す. まず, ユーザーの質問はキーワード抽出 LLM に入力され, 検索キーワードが生成される. 次に, ハイブリッド検索によって外部データベースから最適なコンテキストが抽出される. 最後に, 抽出されたコンテキストと元の質問が応答生成用にファインチューニングされた質問応答 LLM に入力され, 最終的な回答が生成される.

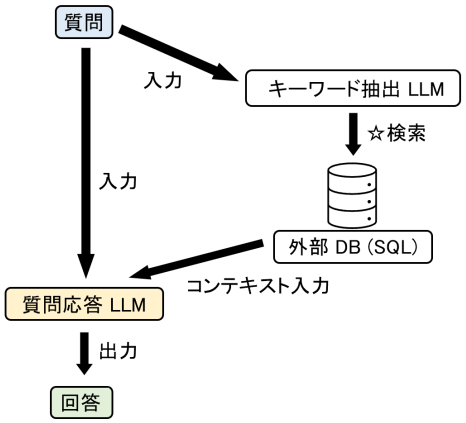

図 1: 提案システムの全体構成図

4 実験

提案手法の有効性を検証するため, RAG システムの知識源となる実験データの準備と, LLM のファインチューニングおよび性能比較に関する実験をした. 以下では, 実験データと実験概要について説明する.

4.1 実験データ

本研究における RAG システムの外部知識源として, 以下の 2 種類のデータを統合して使用した.

- **AnimeList.csv**: アニメ作品の基本的なメタデータ（タイトル, 話数, 制作会社, 放送時期など）を含む構造化データセット.
- **日本語版 Wikipedia**: 各アニメ作品に関する包括的な解説, あらすじ, 登場人物, 制作背景などが含まれる非構造化テキストデータ.

これらのデータを統合し, 検索に適した形式に前処理をした. 具体的には, AnimeList.csv の各アニメ作品に対応する Wikipedia の記事を取得し, その記事本文をそれぞれチャンクに分割した. チャンクサイズは 1000 文字, オーバーラップは 100 文字に設定した. これにより生成されたチャンク群を, Supabase (PostgreSQL) データベースに格納し, 各チャンクにベクトル情報と全文検索用のインデックスを付与した. このデータベースが, RAG における検索対象となる. なお, 本研究では AnimeList.csv のうち, “BLEACH”, “君に届け”, “らんま 1/2”, “クレヨンしんちゃん”, “ONE PIECE”, “犬夜叉”, “カードキャ

表 1: キーワード抽出 LLM の教師データセット例

質問	検索キーワードリスト
BLEACH における 死神とはなんですか？	BLEACH, 死神, 用語, 設定, 魂魄
瀬戸の花嫁はパロディが 多用されていますが、 アニメの演出として、 特にどの監督の影響が 色濃く見られますか？	瀬戸の花嫁, パロディ, 演出, 影響, 監督

プターさくら", "ルパン三世", "ゲゲゲの鬼太郎", "幽☆遊☆白書", "蒼い海のトリスティア", "臭作", "ロクでなし魔術講師と禁忌教典", "8 マン", "どーもくん", "ガンスミスキャッツ", "くまみこ", "瀬戸の花嫁", "プリンセスチュチュ", "ウルトラマニアック" の 20 作品を対象にしている.

4.2 実験概要

本実験では, 構築した RAG システムとファインチューニング済み LLM の性能を評価した.

提案手法の中核をなす 2 種類の LLM を, QLoRA を用いてファインチューニングした. ベースモデルには, 日本語性能に定評のある elyza/Llama-3-ELYZA-JP-8B (80 億パラメータ) [4] を使用した.

- **キーワード抽出 LLM**: Gemini 2.5 Pro が作成した 100 件の "質問と検索キーワードリスト" のペアを教師データとして使用し, ユーザーの質問から最適な検索キーワードを JSON 形式で出力するように学習させた. 表 1 にデータセットの例を示す.
- **質問応答 LLM**: 本研究で構築した RAG システムを用いて Gemini 2.5 Pro が生成した, 100 件の "コンテキスト付き質問と理想的な回答" のペアを教師データとして使用し, 外部情報に基づいて専門家らしいスタイルで応答を生成するように学習させた. 表 2 にデータセットの例を示す.

表 3 に QLoRA のパラメータとして設定した値を示す.

ファインチューニングの効果と RAG の有効性を検証するため, 以下の 3 つのモデルを用意し, その応答品質を比較した.

- **モデル A**: RAG + FT 前の ELYZA-JP-8B
- **モデル B**: RAG + FT 後の ELYZA-JP-8B

表 2: 質問応答 LLM の教師データセット例

コンテキスト付き質問	回答
参照情報 (BLEACH) 「ジャンプフェスタ・ アニメツアー'04」にて 上映されたオリジナル アニメを経て (中略) 質問: BLEACH における 死神とはなんですか？	BLEACH における 死神とは、霊的な 存在で、悪霊（虚）と 戦い、魂を 護る存在です。 彼らは尸魂界 （ソウルソサエティ） という (以下略)
参照情報 (ルパン三世) 2010 年 2 月 - テレビスペシャル第 21 作 『ルパン三世 (中略) 質問: 『LUPIN the Third -峰不二子という女-』は、 どのような点で革新的な 作品でしたか？	『LUPIN the Third - 峰不二子という女-』は、 従来のルパン三世シリーズと 比べて以下の点で 革新的な作品でした。 1. 監督と制作体制: 従来シリーズとは (以下略)

表 3: 主要ハイパーパラメータ

パラメータ名	**値**
LoRA r (ランク)	8
LoRA alpha	16
学習対象の線形層	モデル内の全線形層
Epoch 数	20
最適化手法	Paged AdamW (8-bit)
初期学習率	2e-4
学習率スケジューラ	linear

- **モデル C**: Gemini 2.5 Pro

5 実験結果

質問文 "ルパン三世のモンキーパンチが最後に書いた漫画はなんですか？" を入力した結果を示す. このとき, キーワード抽出 LLM により, "ルパン三世, モンキーパンチ, 最後の漫画, ルパン, パンチ" というキーワードが得られた. キーワード検索をすることで求める答えがある情報をもつ記事にヒットしやすいキーワードである "モンキーパンチ, 最後の漫画" が得られた一方, "ルパン, パンチ" のようにキーワードとしてはあまり意味をなさない単語が抽出されていることがわかった. これは, 本質問とはそのキーワード単体では関係のない単語であるが, 一般的な言葉として使われている単語のため抽出されてしまったと考えられる. 表 4 に 3 つのモデルの質問に対するそれぞれの回答を示す.

質問文 "ルパン三世のモンキーパンチが最後に書

表 4: 各モデルによる回答の品質比較

モデル	生成された回答
A	モンキー・パンチが最後に書いた漫画は、 1994 年イタリアの雑誌「Kappa Magazine 22」 に掲載された『**ルパン三世・ALIS PLAUD**』です。
B	イタリアの雑誌「Kappa Magazine 22」に 掲載された『**ルパン三世・ALIS PLAUD**』です。 これは単体作品としてはモンキー・パンチ 自身による最後の作品となります。 ただし、モンキー・パンチは『漫画アクション』 1994 年 8 月 26 日号に掲載された 「ルパン三世対斬鉄剣」の作画監修も行っています。 これは『ルパン三世 燃えよ斬鉄剣』の 前日談を描いた作品です。
C	モンキー・パンチ先生が最後に完全な新作として 発表した連載漫画は、『M.F.C. 女泥棒会社』 （1982 年 - 1983 年）とされています。 (以下略)

いた漫画はなんですか？” に対する正解の漫画は “ルパン三世・ALIS PLAUD” である. モデル A は正解を正確に述べているが, 単純な回答になっている. 一方, モデル B はより背景知識を補い, 詳細な回答を生成している. これは, ファインチューニングによって情報の統合能力と応答スタイルが改善されたことを示唆していると考えられる. モデル C では, 高い知識を持っていたが, 回答した漫画のタイトルが間違っていた. 回答した漫画のタイトル “M.F.C. 女泥棒会社” は, “ルパン三世” のスピンオフ作品ではあるが, モンキー・パンチが描いたのではなく, 鈴木イゾという漫画家が描いたものである. これらの結果から, RAG による事実性の担保と, FT による専門性の付与を組み合わせた提案手法の有効性が確認された.

6 まとめと今後の課題

本研究では, アニメ制作の専門分野に特化した高信頼な質問応答システムを構築するため, RAG とファインチューニングを組み合わせた手法を提案した. 具体的には, AnimwList.csv と Wikipedia から作成した外部知識をデータベースに保存し, キーワード抽出 LLM を用いたハイブリッド検索によって高精度な情報検索を実現する RAG システムを構築した. さらに, このシステムを用いて生成した高品質な Q&A データセットで LLM をファインチューニングすることにより, 応答の専門性と品質を向上させた. 比較実験の結果, RAG を用いないモデルが専門的な質問に回答できない一方で, 提案手法である RAG とファインチューニングを組み合わせたモデルは, 単純な RAG のみのモデルと比較して, より文脈を深く理解し, 専門家らしい流暢で付加価値の高い回答を生成できることが示された. このことから, 提案手法がドメイン特化型 AI の構築において有効であることが確認された.

今後の課題としては, まずファインチューニング用データセットの規模を数百から数千件へと拡充し, モデルのさらなる性能向上を目指す必要がある. また, 現在は単一の検索および生成プロセスであるが, 質問の意図に応じて検索戦略を自律的に変更したり, 複数の情報源を段階的に参照したりする, より高度な RAG アーキテクチャの実装が望まれる. 加えて, 本研究での評価は主に人手による定性的なものであったため, 定量的評価により性能を客観的に示すことも重要な課題である. これらの課題に取り組むことで, 将来的には制作現場の多様なニーズに応える、より実践的な AI アシスタントの実現が期待される.

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. 2021.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[4] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b. 2024.