生成 AI を使用した博物館ガイドの実装方法の比較検証

Comparison and Verification of Implementation Methods for AI Museum Guides Using Generative AI

西村 正迪[*1], 矢野 浩二朗[*2]
Masamichi NISHIMURA[*1], Kojiro YANO[*2]
[*1] 大阪工業大学大学院情報科学研究科
[*2] 大阪工業大学情報科学部
Email: m1m23a23@st.oit.ac.jp

キーワード：AI,博物館,fine-tuning, ChatGPT,GPT

1. はじめに

近年, VR 技術の進化により, 様々なバーチャル博物館が制作されている. バーチャル博物館は誰でも簡単に作れ, 物理的な制約を超えたアクセスと鑑賞を可能にする. 我々は, この利点を生かし彦根の歴史や地理を紹介する「バーチャル彦根かるた博物館」を作成中である.

しかしバーチャル博物館は, 説明を行うガイドや同伴者が不在であり, 展示ラベルや 3D モデルなどだけで情報を提示するため, 来館者が満足できる鑑賞体験を提供できていないのではないか, という懸念がある. そこで我々は, 生成 AI を活用したアプローチを探求することにした.

生成 AI は 2023 年から利用が爆発的に拡大し, それを博物館の展示に活用する試みに関する報告も既に存在し, 一定の有用性が認められている[1,2]. そこで我々も, 来館者が必要とする情報をリアルタイムで提示できる生成 AI を使ったガイドを配置することにした. しかし, 先行研究では Fine tuning による性能改善が必要, 反応速度が遅いなどの課題も指摘されており, その実用性の検証が必要と考えられたため, 本研究では彦根かるた, および彦根に関する一般的情報を用いて複数の実装方法による博物館ガイド AI の応答能力の比較検証を行った.

2. 実験手法

2.1 使用した生成 AI モデル

本研究では, OpenAI 社が提供している GPT-4-Turbo を使用した. ただし, Fine tuning は GPT-3.5 のみが利用可能であるため, そちらを使用した.

2.2 検証した実装方法

本研究で検討した実装方法は, 以下の 3 つである: プロンプト, Fine tuning, Assistants API

1 つ目の「プロンプト」では, system message に知識データを与え, user message に学芸員という役割, 「以下の質問に回答せよ」という指示, および質問を送信した.

2 つ目の Fine tuning では, GPT-3.5-Turbo にたいし, 必要な情報を与えて学習させたモデルを使用する. 今回は「プロンプト」の system message と同内容の Training ファイルを与え Tuning した. また, Training ファイルの内容を質問する Validation ファイルを用いて, Tuning の成否を検証した. Training の進行は Training loss, Validation loss によって確認した.

3 つ目の Assistants API は, 検索拡張生成(Retrieval Augmented Generation)を用いて, GPT に「プロンプト」の system message と同内容の情報を事前に与えて, 質問に対してリアルタイムで検索し, 回答できるようにした. Instruction には「プロンプト」の user message と同じ内容の指示を与え, そちらに質問も入力した.

2.3 与えたデータ

本研究では, 50 枚の彦根かるたの読みと解説および彦根に関する 50 の一般的な用語の解説を与えた. 総トークン数は 18, 311, 総字数は 15375 であった. 表 1 は, データの例である.

表 1 AI に与えたデータの例

質問	回答
「あ」の読み札は？	赤備え　直政～
ひこにゃんって？	彦根城に住む～

2.4 データ収集方法

Unity 上で OpenAI の API を呼び出すプログラムを作成実行し, Unity から検証用の質問の送信と回答の収集を行った. 質問は全モデル共通で, 20 問である. 回答の評価方法は, 以下の基準を用い○, △, ×で行った.

正答率は, ○の数を 20 で割ったものを 100 分率で算出する. 回答時間に関しては, 20 問の平均を回答時間とする.

表 2 評価基準

評価	評価基準
○	与えた通りのデータで回答できている
△	一部のデータが違う、部分的にしか回答できていない。
×	まったく違う回答をしている。
評価	解答例
○	彦根かるたの「あ」の読み札は「赤備え 直政武勇の関ヶ原」とあります
△	彦根かるたの「あ」の読み札は「赤備え 」とあります
×	彦根かるたの「あ」の読み札は「青い海たかすやまから絶景なり」とあります

3. 実験結果

以下に実験の結果をまとめた(表 3).

表 3 実験結果

	プロンプト	FT	RAG
正答率	100%	30%	75%
回答時間	26.4秒	24.3秒	96.3秒

・プロンプト

プロンプトは, 回答時間は平均 26.4 秒であり, 他の手法より圧倒的に速かった. 正答率も 100%と, 全ての質問に対して正しい回答ができていた.

・Fine tuning(FT)

まず, すべてのデータを用いて Training を行った時の結果を図 1 に示す。緑の線が Training データ, 赤が Validation データの Loss である.

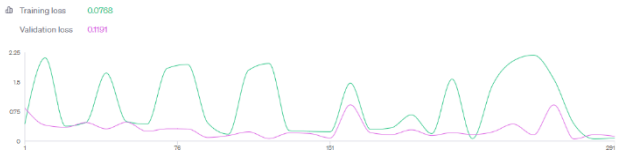

図 1 全てのデータを入力した場合の Loss の推移

Training データ, Validation データ共に波形のようになり, Loss の減少は確認できなかった. そこで, かるたのみのデータ, 一般用語のデータのみで分けて Fine tuning を行った. 結果は図 2, 3 の通りで, こちらではある程度の Loss の減少が確認できた.

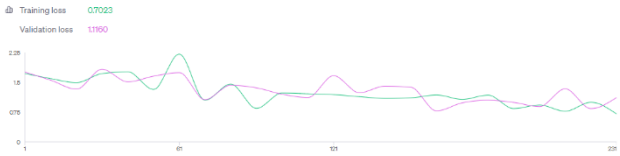

図 2 かるたデータのみの場合の Loss の推移

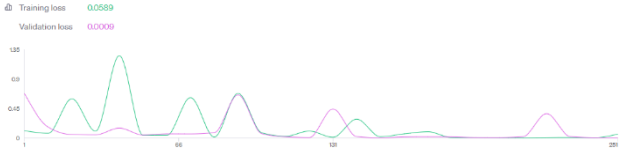

図 3 一般用語データのみの場合の Loss の推移

これらのモデルを用いて質問を行ったところ, 全体の正答率は 30%であった. 特に, かるたに関する質問は全て×で, 勝手に想像した内容（ハルシネーション）が多く見られた. 平均回答時間は, 24.3 秒とプロンプトより少し早い結果になった.

・Assistant API

Assistant API では, 正答率 75%となった. 評価も×はなく, △が多い結果となった. 一般質問よりかるたの正答率が低かった. ただし, かるたの読み札を間違えるのではなく,「わからない」と回答することが多く, ハルシネーションは見られなかった. 一方で回答時間が 96.3 秒と他のモデルよりかなり時間が必要であった.

4. 考察

プロンプトに関しては, 正答率, 回答時間ともに他の 2 つのモデルより良いことが, 実験からわかった. ただし, かるたのデータは今後増えることはないが, 一般データに関しては増える可能性がある. その場合トークン数の制限で全てのデータを入れることができなくなる恐れがある.

一方, Fine tuning はトークン数の消費が少ないという利点があるが, 今回は Training が上手く行かず, 特に, かるたに関しては全問不正解であった. かるたの場合, 同じような文章で質問が続くため, AI がそれらを区別できていない可能性がある. ただし, 一般知識だけの Training なら, GPT-4 であれば, ある程度の正答率を確保できる可能性はあると考える.

Assistant API に関しては, 正答率は高いがプロンプト程ではなかった. 特にカルタの正答率が低く, RAG の際に読み札を正確に区別して必要な情報を抽出できていないことが考えられた. ただし, わからない質問に対しては,「わからない」と返答し, 間違った知識を回答していないので, Fine tuning よりは良い性能である. しかし, 回答時間にかなりの時間が必要であったことは, 博物館 AI としては良いものではないだろう. 一方で, Assistant API はプロンプトに比べると AI に与えられるデータの量がプロンプトより多いため, 今後与える知識量が増えた場合には有用であると思われる.

今回の結果からは, 博物館 AI に最も利用可能なモデルはプロンプトであるということがわかったが, 今後のスケールアップを考えると, カルタの読み札の情報の抽出能力を向上させた RAG を独自に実装するなどの工夫が必要であると思われる.

参考文献

(1) Trichopoulos, Georgios, et al. “Crafting a museum guide using Chatgpt4.” Big Data and Cognitive Computing, vol. 7, no. 3, p. 148 (2023)

(2) Spennemann, Dirk HR. "Exhibiting the Heritage of COVID-19—A Conversation with ChatGPT." Heritage vol. 6, no. 8, pp. 5732-5749 (2023)