Dify で自作したレファレンスチャットボットを用いた文章生成 AI の性能比較

－生成 AI は図書館サービスに使えるか？－

橋本 郷史
東邦大学医学メディアセンター 大橋病院図書室

【目的】多くの文章生成 AI(以下，LM(Language Model:言語モデル))は，図書館のサービス案内のような簡単な受け答えをこなす程度の能力をすでに有している。LM の応答はしばしば嘘をつくことが問題視されるが，RAG という情報参照技術を用いれば，各機関の情報に特化した嘘をつかない応答をさせることが比較的簡単に可能である。

しかし実際に実用することを考えた際に，RAG にどの程度の精度があるのか，またどの程度の規模・性能の LM であれば実用に耐えうるのかは不明である。そこで，これらを確認するために，自機関のサービス情報を学習させたチャットボットを作成し，パラメータ数等の異なる複数の LM を応答に用いて，その性能の比較を行うこととした。

【方法】チャットボットは AI ツール開発基盤「Dify」[1)]を用いて作成した。学習させる参照情報として，東邦大学医学メディアセンターの，研究費での資料購入に関する説明ページ[2)]を取り込んだ。

RAG のための Embedding Model として「text-embedding-3-large / OpenAI」[3)]と「snowflake-arctic-embed2 / snowflake」[4)]の 2 種類を用いた。応答用の LM には，「ChatGPT4o-mini / OpenAI」[5)]と，「Qwen3 / Alibaba Cloud」[6)]の 14B,8B,4B,1.7B の各モデルの計 5 種類を使用した。OpenAI 社のモデルは API 経由で使用し，snowflake と Qwen3 はローカルのパソコン上で動作させた。また Qwen3 は非推論モードで使用した。

参照情報を各 Embedding Model で取り込み，応答用の各 LM が参照情報に基づいて応答できるかどうかをテストした。”応答できるかどうか”は，「質問に答えているか」「嘘をつかないか」「関係ないことを答えないか」の 3 つの観点で評価した。各観点に対応する質問を各 Embedding Model と各応答モデルの組み合わせごとに 10 回ずつ行い，その応答内容を評価した。

【結果・考察】結果や考察については講演内で述べる。

[参考情報] ※all accessed 2025-May 12

1) LangGenious. dify. GitHub [internet].
2) 研究費購入資料について．東邦大学メディアセンター [internet].
3) OpenAI. text-embedding-3-large. OpenAI Platform [internet].
4) Snowflake. snowflake-arctic-embed-l-v2.0. Hugging Face [internet].
5) OpenAI. GPT-4o mini [internet].
6) Alibaba Cloud. Qwen3. GitHub [internet].