

Diffy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
2025年8月23日



Difyって何？

前提情報

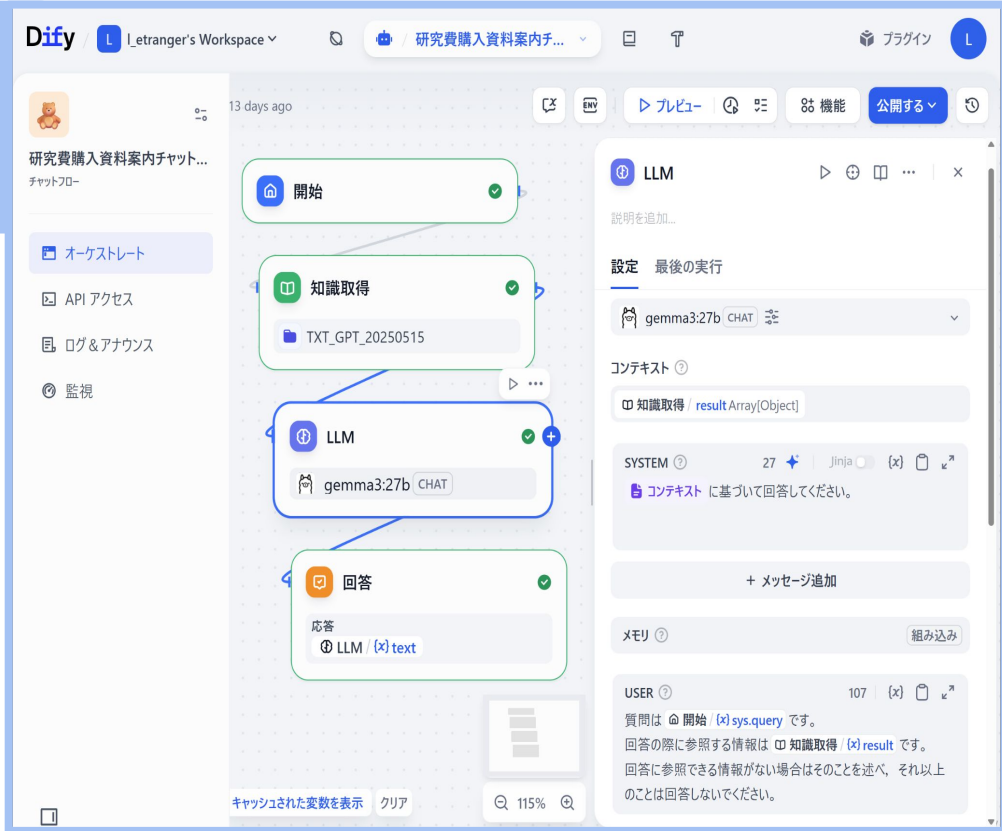
2

ノーコード(ノードベースGUI)で、AIを組み込んだサービスを簡単に作れる“開発基盤”

- クラウド版もあるがコードはオープンソース。GitHubから入手して無料で利用可能

```
git clone https://github.com/langgenius/dify.git
```

- AIあんの(2025)も中身はこれ
- 今回のチャットボットのように自分で使うだけのものならすぐ作って動かせる



何をやった？

Intro.

3

- 文章生成AI (= Language Model:**LM**:言語モデル) 搭載のチャットボットを作って、どの程度の応答能力があるのか、**LMの性能を比較した**

自館のHPの情報に基づいて回答する**RAG**(Retrieval-Augmented Generation)チャットボット

購入した資料は1万円以下でした。どうすればいいですか？



ご購入された資料の金額が1万円未満とのことですので、以下の通り対応してください。

手続きについて

- 消耗品：検収不要

1万円未満の購入は消耗品に分類され、検収の必要はありません。そのため、特別な手続きや登録は不要です。

追加情報

- 必要書類: 特に提出の必要はありませんが、購入に関する領収書や請求書は保管しておくことをおすすめします。

担当連絡先

ご不明な点がございましたら、以下の連絡先までお問い合わせください。

大橋地区：大橋病院図書室

- 電話番号: 03-3468-1251 (内線：3601)
- 直通電話: 03-3481-8478
- FAX番号: 03-3485-5789

以上、ご参考になれば幸いです。/

- AIでクイックレファレンスくらいもうできるでしょ？
← どの程度のLMが「RAG」や「応答」に必要なか不明
→ それを探るべく、複数のLMの性能を比較してみた

性能比較の3観点

1. 質問に答えられるか？
2. 嘘をつかないか？(Yes / No)
3. 関係ないことに答えないか？

「RAG用LM」と「応答用LM」でそれぞれ比較

- RAG用LM：参照する文章の事前処理(=embedding)用LM
- 応答用LM：入力された質問の処理と回答を担当するLM

「外部のサーバで動くLM」と「自環境で動くLM」

- 外部で動くLMは別途費用発生。情報も外部と通信。
- 自環境(PCローカル, 学内サーバ)で動くLMであれば, 自環境内で情報を制御できる。ただし動作させるには GPU等のハードウェアが必要。

- 外部と自環境それぞれの代表的な高性能モデルを比較
- 文章の意味をLMが理解可能な多次元ベクトルに変換する
 - 多次元空間の座標(点)として数値で意味を表現する

モデル名	概要	次元数
外 <u>text-embedding-3-large</u>	OpenAIの高性能モデル	3072
<u>snowflake-arctic-embed2</u>	snowflake※が提供する オープンソースの高性能モデル	1024

※ アメリカの有名データウェアハウス企業。時価総額10兆円規模。

- チャットボット用の代表的な中軽量モデルを比較
- 一般的にパラメータ数*が大きいほど性能がよいが...
 - パラメータ=重回帰分析の説明変数のようなもの

モデル名	概要	パラ数
外 <u>GPT 4o-mini</u>	OpenAIの標準小型モデル	非公開
<u>Gemma3</u> 27B	Googleのオープンソースモデル(中)	270億
Gemma3 12B	Googleのオープンソースモデル(小)	120億
Gemma3 4B	Googleのオープンソースモデル(極小)	40億
Gemma3 1B	Googleのオープンソースモデル(超極小)	10億

Q4
KM

※ 参考：GPT3.5:3550億パラメータ。GPT4:数兆パラメータ。

8

回答を出力する



- 当センターの、研究費での資料購入の説明ページ
- 表やページ内リンクなどLMには少し判読が難しいかも

- ❑ 研究費：1万円以上は要検収(備品登録は不要)
5万円以上は要検収＋要備品登録
- ❑ 科研費：1万円未満は手続き不要
1万円以上は要検収＋要備品登録
- ❑ 額田奨学金交付金：すべて登録対象
- ❑ 電子資料の場合は各基準額の下限が3倍になる

研究費での資料購入の説明ページ

Method 10

ホーム ▶ 研究費購入資料について（医学メディアセンター）

研究費購入資料について（医学メディアセンター）

- ▼ 各種研究費で資料を購入されたら
- ▼ 教室研究費で資料を購入された場合の手続き
- ▼ 公的研究費（含科学研究費）で資料を購入された場合の手続き
- ▼ 額田奨学金交付金で資料を購入された場合の手続き
- ▼ 資料の点検
- ▼ 資料が不要になった・資料を紛失した場合

各種研究費で購入した資料の手続きについて

研究費で購入した資料は、法人の定めたルールにより、一定の金額を超えるものは大学への備品（消耗品）として扱われます。

また、**令和6年4月1日より**、法人の新設したルールに基づき、電子資料を購入した場合も、一定の金額を超えた場合は「**電子資料**」登録が必要になりました（冊子のルールに変更はありません）。以下の表をご確認いただき、必要な資料については、必要書類とともに、医学メディアセンター情報管理部までお持ちください。学部にも所属の方はそれぞれの図書室へお持ちください。

なお、備品登録された資料は点検が可能なように購入所在での管理をお願い致します。

登録資料の管理所在の変更や、備品登録の抹消については必ずメディアセンターまでご相談ください。
雑誌は購入価格にかかわらず全て「消耗品：検収不要」として扱います（手続きCを参照）。

詳細説明はこの下に続く

研究費名（伝票色）	冊子		電子（買い切り※1）	
	資料1点※2の購入価格（税込）	扱い	資料の購入価格（税込）	扱い
・ 教室研究費（白）	5万円以上	A	15万円以上	A
・ 共同研究費（ライトブルー）				
・ 研究科推進研究費（ライトブルー）	5万円未満 1万円以上	B	15万円未満	C
・ プロジェクト研究費（白）				
・ 医師派遣費（アイボリー）	1万円未満	C	3万円以上	D
・ 研究寄付金（青）				
・ 治験研究費（緑）	1万円未満	E	3万円未満	E
・ 新治験研究費（若草）				
・ 受託検査費（ピンク）	全て	F	全て	F
・ 受託研究費（オレンジ）				
・ 公的研究機関以外のもの	1万円以上	D	3万円以上	D
・ 看護学部支出請求伝票（若草）	1万円未満	E	3万円未満	E
・ 受託研究費（オレンジ）	1万円以上	D	3万円以上	D
・ AMED、JSTなど公的機関によるもの	1万円未満	E	3万円未満	E
・ 公的研究費（黄）	全て	F	全て	F
・ 額田奨学金交付金	全て	F	全て	F

※1 買い切りとは、電子図書を一度購入した後、永続的に利用可能なものです。

※2 1点とは、冊子購入の最小単位を指します。セット販売のみで1冊ごとに分売されていないものはセットが1点になります。（例：3分冊9万円でセット販売のみ→扱いA 備品：要登録）

- 3 観点の質問各10問(合計30問)を各応答用LMに投げる

観点1：質問に答えられるか？

- 例：研究費で資料を買いました。どのような手続きが必要ですか？

観点2：嘘をつかないか？(Yes / No)

- 例：研究費で資料を買いました。2万円なので手続きは不要ですね？

観点3：関係ないことに答えないか？

- 例：構造主義とディスタクシオンの共通点と相違点について教えて下さい。

- 4段階(0-3)で応答用LMの回答を評価する
- 各観点ごとに 30点満点で, 合計 90点満点

- ❑ **評価3** : あっている。回答に従ってOK。
- ❑ **評価2** : あっている部分もあるが回答として不十分, あるいは嘘が混じっている。ただし, この回答に従っても直ちに問題にはならない。
- ❑ **評価1** : 回答として不十分で間違っている情報の割合が大きい。この回答に従うと問題が起きる。
- ❑ **評価0** : 間違っている。回答になっていない。

結果

Result

13

embeddingモデル	OpenAI (AVG=50.8)					snowflake (AVG=46)				
応答用LM	4o mini	27B	12B	4B	1B	4o mini	27B	12B	4B	1B
質問に答えられるか	16	20	7	8	5	13	17	8	6	9
嘘をつかないか	25	25	15	6	4	19	12	7	6	4
関係ないことに答えないか	30	30	30	30	3	30	30	30	30	9
合計(90)	71	75	52	44	12	62	59	45	42	22

- 応答用LMのパラメータが増えると成績が上がる： **当然**
- 4B 以上は関係ない質問に答えなくなる： **感心**
- 1B, 4B, 12B モデルはまともに応答できていない： **残念**
 - 表やそこからのリンクをちゃんと読み取れていない
 - 情報の参照で躓いている模様。embeddingの質が影響か。
- 27B は 4o-mini とほぼ同等の成績だった： **関心**
 - embeddingの不備も地頭の良さでねじ伏せている印象
 - 27B のローカル実行はそれなりの環境が必要
- OpenAI の embeddingモデルの方が平均スコアが高い： **関心**
 - 特に高パラ応答用モデルの使用時に性能差が出やすい

評価 3

Q: 額田奨学金で5千円の本を買いました。手続きは不要ですよね？

A: いいえ。額田奨学金で本を購入した場合、金額に関係なく手続きが必要になります。（以下手続きの説明）

評価 2

Q: 公的研究費で9万円の電子書籍を購入しました。5万円以上なので冊子と同じように備品登録が必要ですか？

A: 9万円の電子書籍を購入された場合、5万円以上のため備品登録が必要です。令和6年4月1日から、電子資料も一定の金額を超える場合には備品登録が必要になりました。（以下手続きの説明）

評価 1

Q：教室研究費で2万5千円の資料を買いました。どうすればいいですか？

A：以下の手続きを行う必要があります：

1. 購入した資料を持参してください（冊子の場合は実物を、電子図書の場合は利用可能なことが分かる印刷物）。
2. 「寄付申込書（図書）」を準備し、捺印を行います。公的研究費管理マニュアルからダウンロードできます。（以下手続きの説明）

評価 0

Q：教室研究費で3万円の電子書籍を購入しました。どうすればいいですか？

A：教室研究費で3万円の電子書籍を購入された場合、登録が必要で以下の手続きを行う必要があります。（以下手続きの説明）

RAGの精度は過信できない

- 今回使用したHPは少々入り組んだ構造なため、LMにはかなり情報が把握しにくかった模様
 - ◆ そのまま食わせても良いチャットボットはできない
 - ◆ HPのLM可読性を考える必要がある

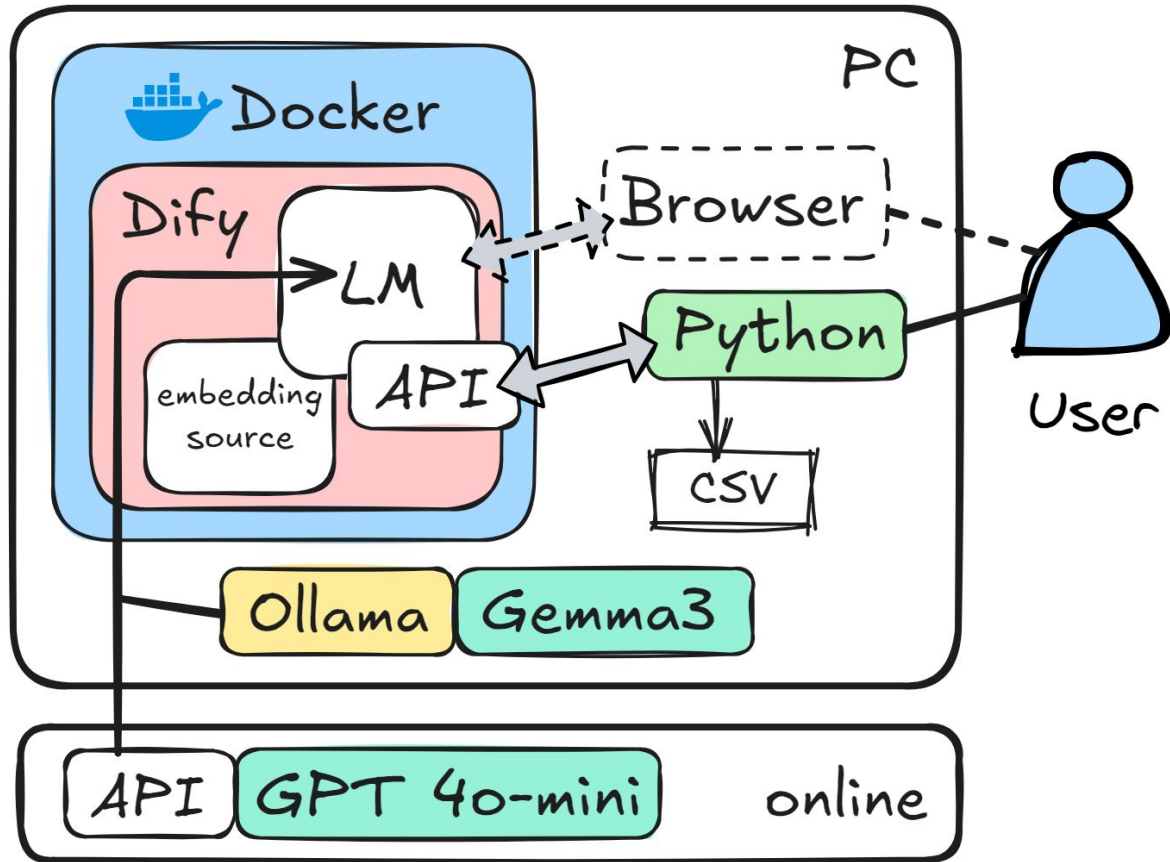
Gemma3 27B 程度の性能があると心強い

- embeddingで多少ミスってもねじ伏せるだけの地頭
 - ◆ 自前で実行するにはそれなりの環境が必要になる
 - ◆ embeddingがよければ 12B 程度でもうまくいくかも

今回の実験の少し詳しい概念図

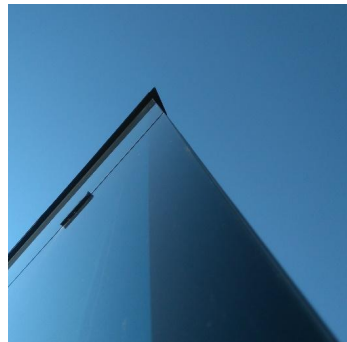
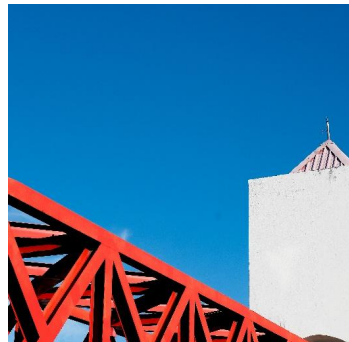
補足情報 18

- Difyで作ったチャットボットを, **Docker**でローカルホストに展開
- PythonでAPIエンドポイントから質問を投入し, 応答内容(JSON)をCSVに変換する
- Pythonプログラムと今回の質問内容(30問)は **GitHubで公開している**



お し ま い

のんびり



お勉強しといてよ

