053. RAGによる既存情報をLLMに活かす試み

樋口千洋[1,2]、夏目やよい[1,3,4]

[1]国立研究開発法人 医薬基盤・健康・栄養研究所、[2]国立大学法人 東京科学大学、
[3]国立大学法人 徳島大学、[4]国立大学法人 大阪大学

モチベーション

医薬基盤・健康・栄養研究所(NIBIOHN)はこれまで4省連携の枠組みのもと、バイオサイエンスデータベースセンター(NBDC)と生命情報データのカタログ化や横断検索について取り組み、創薬・疾患研究のためのデータベースSagaceの開発に取り組んできた。本研究は令和4年度をもって終了したが、置き換わるように出現したchatGPTに代表される大規模言語モデル(LLM)が急速に浸透した。LLMはクエリに対する文章生成や要約翻訳などので様々な機能を有していて、その有用性は誰もが認めるが、これまでキュレートして培ってきたデータベースとは性質が異なることやや、LLMの普及でこれまで培ってきた資産が意味をなさなくなるのではという不安が生じた。そこで、従来の資産を有効に活かしてのLLM活用形態について検討した。

LLMとその特徴

LLMは、非常に多くのパラメータとデータを用いて訓練された人工知能モデルで、自然言語処理(NLP)のタスクに非常に優れた性能を発揮する。これらのモデルは、大量のテキストデータを学習し、文章の生成や質問応答、翻訳など、幅広いタスクに対応できるようになっている。2022年11月30日に公開されるや否や爆発的に普及した。LLMは様々な場面で活用が進んでいるが、ハルシネーションと呼ばれる存在しない事実を生成してしまい、これをいかに回避するかあるいは検知することが課題となっている。なお、近年の研究でこれを完全に排除するのは困難とされている[1]。本事象を軽減するための方策としてファインチューニングとRAGが挙げられるが、前者は多くのGPUリソースが必要とされ、一般的な利用環境での実施は容易ではない。

RAG

RAG (Retrieval-Augmented Generation)は、情報検索と生成モデルを組み合わせた手法である。まず、ユーザーのクエリに対して外部の知識ベースやドキュメントから関連情報を検索し、その情報を基に生成モデルが回答を生成する。これにより、モデルが訓練データに依存するだけでなく、最新の情報や信頼性の高いデータを取り込むことができます。特に、事実ベースの応答が求められる場面で効果的で、生成結果の精度向上や知識補完に役立つ。RAGは、LLMの知識限界を補う実用的なアプローチとして広く研究されており、様々な改善手法が提案されている[3, 4]。GoogleのNotebookLM[5]は最もよく知られたRAGシステムと思われる。

ローカルLLM

従来、LLMの利用はインターネット経由に限られていた。公開されたLLMは自由に利用することができたが、LLMの構築はもちろんのこと利用するだけでも高性能な環境が必要であった。LLMの活用は多岐にわたり、個人情報の匿名化にも応用されつつある。また、個人情報データ解析の場では外部インターネットへの接続は容易ではないため、ローカルLLM利用の需要が高まったが、先述の性能という課題があった。ところが最近になってこの制限がなくなりつつある。例えばOllama[6]を使えば、数十GBのメモリを搭載したノートPCでもLLMが使える。今春開催されたハッカソンイベントでは64GBのメモリを搭載したMacBookProでCohereのcommandR LLMを使い、Dify[7]というフレームワークを使って実用的なRAGシステムの提案ができた[8]。

RAGシステム作成のためのフレームワーク

先述のDifyはDockerを用いて簡単に構築できるGUIベースのノーコードツールである。構築したワークフローはYAML形式でインポートおよびエクスポートができる。DifyではRAGのソースとしてシンプルなテキストファイルを与えることができる他、Nortonで作成したドキュメントを用いることもできる。Llamaindex[9]はPythonのライブラリであるが、NBDCの横断検索のエンジンはElasticSearchに直接接続しシームレスなRAGソースの供給ができる。

現在の取り組み

NIBIOHN で公開をおこなってきたSagace(Fig.1)[10] をRAGとLLMで実装を検討している。ローカルLLMを使用して非インターネット環境での利用を想定している。もし近い将来BitNet[11]でのLLMが実用的になれば、スマホでのスタンドアローン利用形態も現実味を帯びてくる。

Fig. 1: 国内の医学／生命科学データと生物資源をさがす Sagace

(関連情報)GENIAC松尾研LLM開発プロジェクトと成果

これまでさまざまなLLMが開発され、そのモデルもHugging Faceなどのリポジトリに公開され誰でもが利用可能な状態であったが、例えばモデルを構築するのにどれくらいのGPUリソースが必要だとか、どれくらいの費用が掛かったかといった情報は断片的に存在したが、LLM構築にどのようなノウハウが必要かと言った細かな技術情報は共有されていなかった。今春経済産業省とNEDOが進める「GENIAC」プロジェクトの一環としてGENIAC松尾研LLM開発プロジェクトが開始され、非専門家が集まりLLM構築に関わった。その結果GPT3.5-turboやllama3.Xに匹敵する「Tanuki-8x8B」が誕生した(Fig.2)。

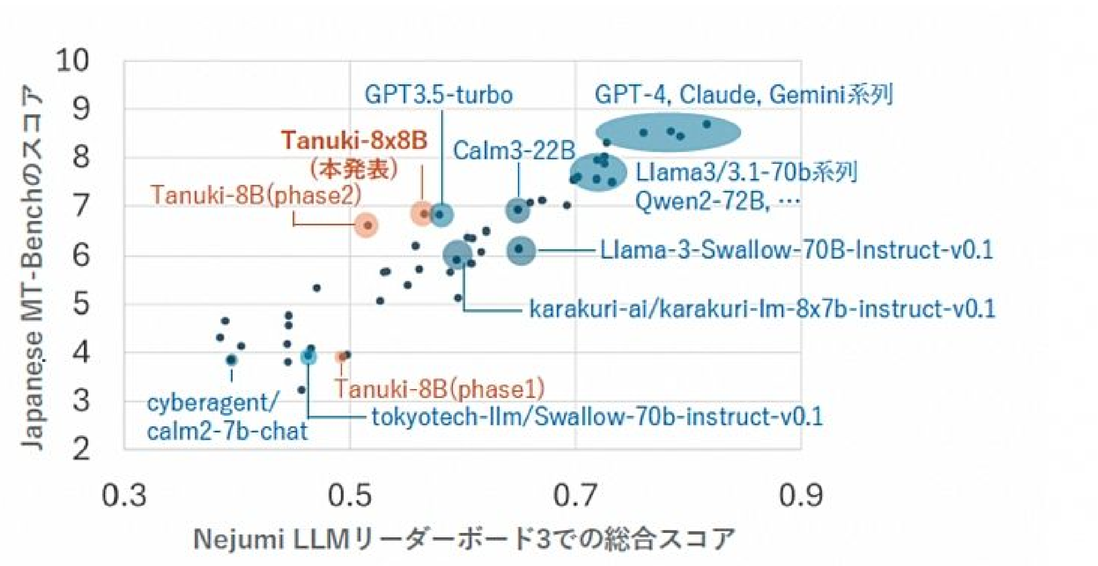

Fig. 2: Tanuki-8x8Bモデルの既存LLMとの比較

本プロジェクトの真の成果は、LLM構築を通じてたくさんの知見が見出され、Zennなどの技術ブログを通じて今後の課題や展望が広く国内外の技術者に公開されたことにあると考える。学習のための日本語コーパスが不足しているのではという仮設も本プロジェクトの過程で明らかになり、そのため信頼性の高い合成データを学習に使った結果、性能向上に寄与した。
本発表ではハルシネーション抑制にファインチューニングが挙げられるもののリソースの問題で一般的な利用環境では実施できないために、既存のキュレートされた情報をRAGのソースに使うことを検討したが、LLMのファインチューニングの機会があれば、これらをLLM構築時の学習データに使うことで、より正確な事実の生成が期待される。

References

[1] Sourav Banerjee et al. LLMs Will Always Hallucinate, and We Need to Live With This. https://arxiv.org/abs/2409.05746.

[2] 江上周作他 大規模言語モデルを用いたSPARQLクエリ生成の予備的実験 https://www.jstage.jst.go.jp/article/jsaisigtwo/2023/SWO-060/2023_04/_pdf.

[3] Darren Edge et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. https://arxiv.org/abs/2404.16130.

[4] Hongjin Qian et al. MemoRAG: Moving towards Next-Gen RAG Via Memory-Inspired Knowledge Discovery. https://arxiv.org/abs/2409.05591.

[5] Google NotebookLM. https://notebooklm.google.com/.

[6] Ollama. https://github.com/ollama/ollama.

[7] Dify. https://dify.ai/.

[8] Shi3z. テーマは「夢」大阪 24時間 AI ハッカソンレポート https://note.com/shi3zblog/n/n1c5313748849.

[9] Llamaindex. https://www.llamaindex.ai/.

[10] Sagace. https://sagace.nibiohn.go.jp/.

[11] Jacob Nielsen et al. BitNet b1.58 Reloaded: State-of-the-art Performance Also on Smaller Networks. https://arxiv.org/abs/2407.09527.

[12] GENIAC 松尾研 LLM開発プロジェクト https://weblab.t.u-tokyo.ac.jp/geniac_llm/.

国立研究開発法人
医薬基盤・健康・栄養研究所
National Institutes of Biomedical Innovation, Health and Nutrition