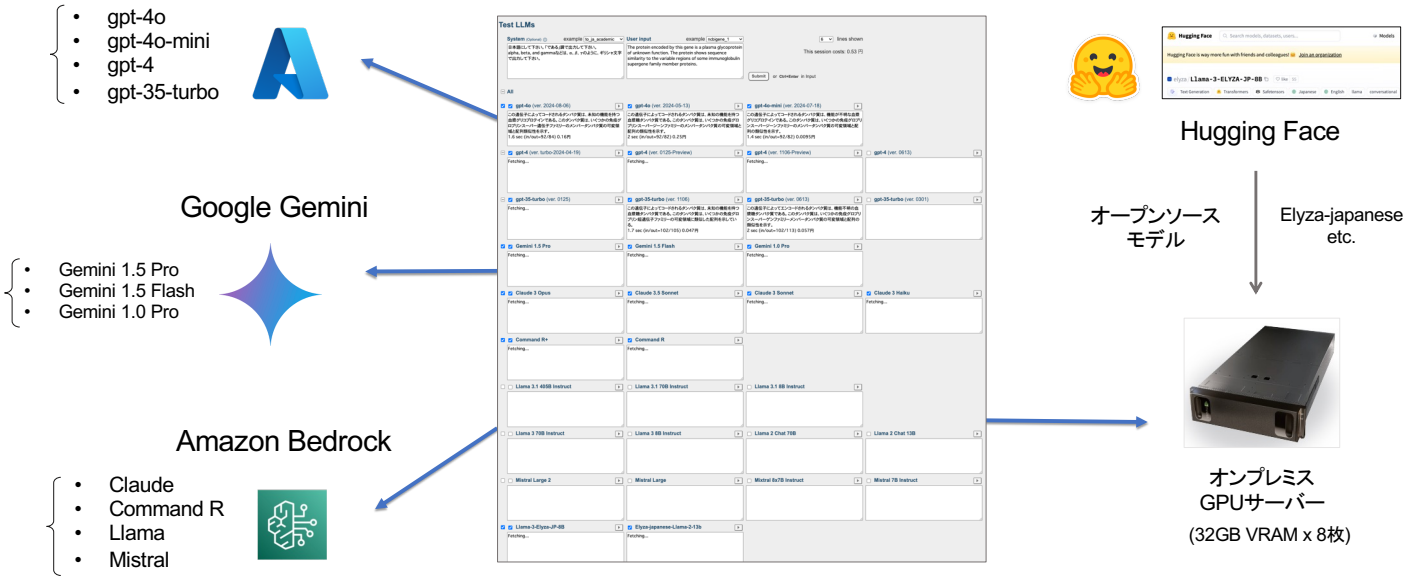
様々な大規模言語モデル (LLM) を実行・評価するための環境構築

千葉啓和 （ライフサイエンス統合データベースセンター）

Methods

クラウド環境、およびオンプレミス環境において様々なLLMを一度に実行し、結果を比較できる環境を構築した。フロントエンドはJavaScript、バックエンドはPython (Flask)で実装した。プログラムからAPIにアクセスすることも可能になっている。

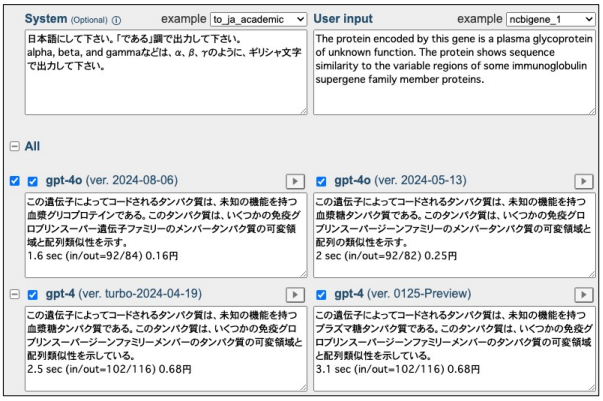


ここ1〜2年で、実に多くの種類の大規模言語モデル (LLM) が発表された。LLMを活用することによって、生命科学分野におけるデータ処理が加速されると考えられる。しかしながら、実際に各種のLLMを利用してみると、モデルやバージョンによって出力が異なり、どのLLMを用いればよいか分からないことも多い。そこで今回、様々なLLMを一度に実行し、出力を比較できるような環境の構築を行った。LLMの実行環境は、クラウドとGPUサーバーを併用した。より具体的には、OpenAIのモデルに関してはMicrosoft Azureを利用した。Anthropic、Cohere、Meta、Mistral AIのモデルに関しては、Amazon Bedrockを利用した。 Geminiに関しては、 Googleにアクセスしている。それ以外のモデルに関しては、Hugging Faceからモデルをダウンロードして、GPUサーバー上で実行した。現状では、33種類のモデル（バージョンの違いも含む）を試した結果を一度に表示することができている。ここでは、実行結果とその評価についても共有し議論したい。

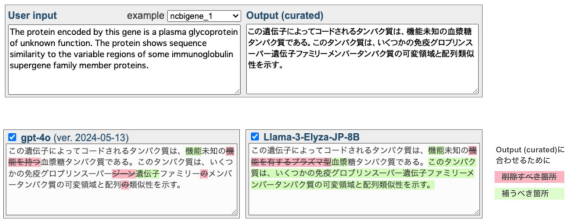
Results

ヒトのNCBI GeneのSummaryを、各LLMを用いて和訳した。一方で、10個の遺伝子について、キュレーションされた和訳も作成し、各モデルの出力との違いを定量的に評価した。差分の可視化にはJavaScriptのjsdiffを用い、距離の計算にはPythonのLevenshteinライブラリを用いた。

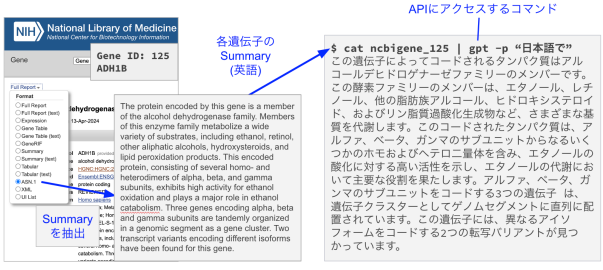
実行結果画面の例



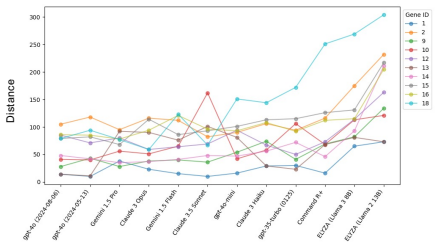
キュレーションされた和訳との差分の可視化



コマンドラインからAPIにアクセスし、遺伝子のSummaryを和訳する例



キュレーションされた和訳と各モデルの出力とのレーベンシュタイン距離



Model	Avg. distance	Avg. similarity
gpt-4o (2024-08-06)	57.9	0.774
gpt-4o (2024-05-13)	59.6	0.775
Gemini 1.5 Pro	64.7	0.720
Claude 3 Opus	68.1	0.715
Gemini 1.5 Flash	74.3	0.698
Claude 3.5 Sonnet	76.4	0.696
gpt-4o-mini	76.9	0.687
Claude 3 Haiku	78.4	0.684
gpt-35-turbo (0125)	79.5	0.681
Command R+	94.5	0.628
ELYZA (Llama 3 8B)	123.8	0.468
ELYZA (Llama 2 13B)	173.3	0.306

$$similarity = 1 - \frac{distance(output,\ curated)}{max(len(output),\ len(curated))}$$