

AIにアプリ制作 手伝ってもらった話 * 3

個人開発集会 @VRChat 2024-07-18

夜鍋ヨナ-yonabeyona <<http://x.com/yonabeyona>>

AIにアプリ制作 手伝ってもらった話*3

個人開発集会 @VRChat 2024-07-18

夜鍋ヨナ-yonabeyona <<http://x.com/yonabeyona>>

AIにアプリ制作

手伝ってもらった話 * 4 + α

個人開発集会 @VRChat 2024-07-18

夜鍋ヨナ-yonabeyona <<http://x.com/yonabeyona>>

個人開発集会

AIにアプリ制作 手伝わせたい話

* 4 + α

自己紹介

- 名前：夜鍋 ヨナ(よなべ よな)
- X(Twitter) : yonabeyona, yonabeyona_sub
- Discord : yona_47
- その他
 - ComputerScienceが好き
 - 数学勉強中
 - 物理も勉強中
 - ComputerScienceの中でも、ComputerArchitectureが好き
 - 最近言語学も興味あり



自己紹介

CS集会(計算機科学集会)やってます

隔週火曜日開催

最近一周年を迎えてめでたい

コンピュータに関連する話をする

みんな来て!!!

そしてコンピュータのお話をしよう!!!

話したい!

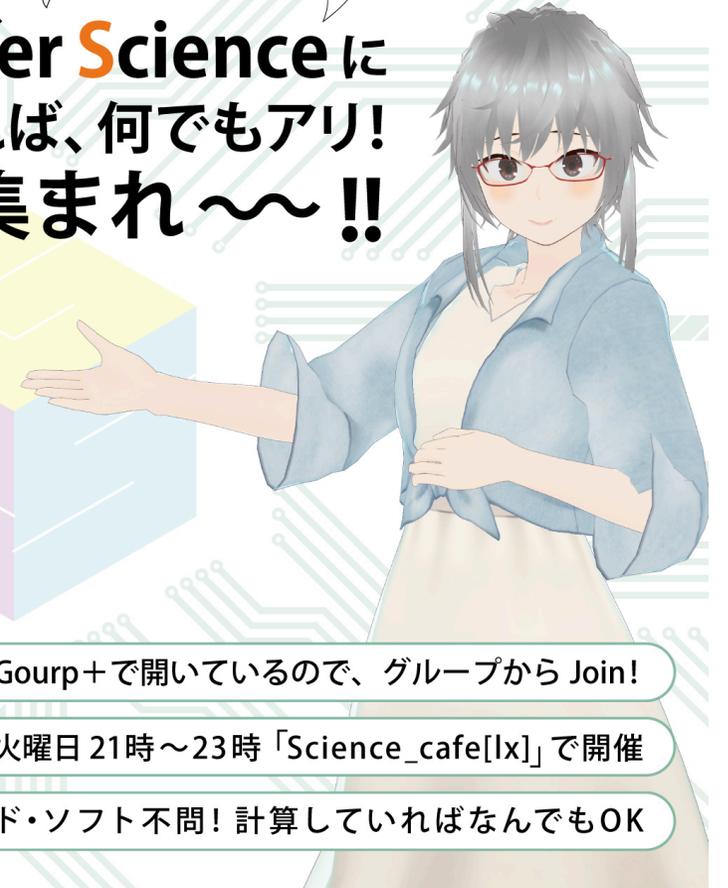
聞きたい!

知識を共有したい!

主催 夜鍋ヨナ-yonabeyona

コンピュータ サイエンス

Computer Scienceに
興味があれば、何でもアリ!
みんな集まれ〜!!



参加方法 当日 Goup+で開いているので、グループから Join!

日時 隔週火曜日 21時~23時 「Science_cafe[ix]」で開催

取扱対象 ハード・ソフト不問! 計算していればなんでもOK



今日やること

- AI界限ってさ.....
- 生成AIサービスとかの動向調査
- 生成AIといっしょにシステム作った
- 事のいきさつ
- システム概要
- 大変だったこととか
- 感想

AI界限ってさ.....

AI界限は成長
速度が速すぎる

2～3週間ぐらいつとAI関連の話調べてた

感想：

- 成長速度が異様に速い
- 2023年と2024年で話の移り変わる速度が明らかに違う
- 今年は半年で既に12回くらい大型・中型アップデートが来てる
- 月に2回もアプデするのやめろソシャゲか？ ~~原神でも月1マイナー/年1メジャーやぞ~~
- 最近はなにかにつけて、他社の発表に合わせて隠し玉を被せてくる
~~なんならソシャゲのほうで空気読んでタイミングずらしてる~~

2～3週間ぐらいつとAI関連の話調べてた

感想：

- 成長速度が異様に速い
- 2023年と2024年で話の移り変わる速度が明らかに違う
- 今年は半年で既に12回くらい大型・中型アップデートが来てる
- 月に2回もアプデするのやめろソシャゲか？ ~~原神でも月1マイナー/年1メジャーやぞ~~
- 最近は何にかにつけて、他社の発表に合わせて隠し玉を被せてくる
~~なんならソシャゲのほうで空気読んでタイミングずらしてる~~
- キャッチアップするだけで大変なのでマジで手加減してくれ 触るヒマもない
- とにかく無料のサービスが多いので遊ぶなら今の内？
- ダンピングしてる気分になるけど、勝手に謎の競争をしてる今がチャンス
ありがたく使わせていただきますよう

生成系AIの種類

今あるサービス

- 文字生成
- 画像生成
- 動画生成
- 音声生成

生成系AIの種類

今あるサービス

- 文字生成
 - マルチモーダルにより音声・動画を見て回答する。主にチャット、コーディング。
- 画像生成
 - 着色とか線画清書とか単純な作業もできるし、「〇〇さんみたいな絵柄」注文も聞く
- 動画生成
 - イラスト1枚から動画、文章から動画も行ける
- 音声生成
 - 音楽も生成する
 - ボイスも生成する

生成系AIの種類

今あるサービス

- 文字生成
 - マルチモーダルにより音声・動画を見て回答する。主にチャット、コーディング。
- 画像生成
 - 着色とか線画清書とか単純な作業もできるし、「〇〇さんみたいな絵柄」注文も聞く
- 動画生成
 - イラスト1枚から動画、文章から動画も行ける
- 音声生成
 - 音楽も生成する
 - ボイスも生成する

(逆に生成できないものってなんだ? → わかりません)

知っていると理解が進む用語と解説

- Transformer
- LLM
- パラメータ
- 推論
- ハルシネーション
- トークン
- 入力/出力トークン
- コンテキストウィンドウ
- fine-tuning
- RAG
- Fill-In-the-Middle タスク

知っていると理解が進む用語と解説

- Transformer
- LLM
- パラメータ
- 推論
- ハルシネーション
- トークン
- 入力/出力トークン
- コンテキストウィンドウ
- fine-tuning
- RAG
- Fill-In-the-Middle タスク

多くない.....?

Transformer

LLM (**L**arge **L**anguage **M**odel)

パラメータ

Transformer

Attention Is All you Need(Ashish vaswani 共著)で提唱された
自然言語処理のアーキテクチャの一種^[1]

LLM (**L**arge **L**anguage **M**odel)

パラメータ

Transformer

Attention Is All you Need(Ashish vaswani 共著)で提唱された
自然言語処理のアーキテクチャの一種^[1]

LLM (Large Language Model)

Large Language Model、大規模言語モデルの略
自然言語処理の分野で広く使用される機械学習モデルの一種
大量のテキストデータを元に学習して、自然言語の入力に対して自然言語で出力する^[2]

パラメータ

Transformer

Attention Is All you Need(Ashish vaswani 共著)で提唱された
自然言語処理のアーキテクチャの一種^[1]

LLM (Large Language Model)

Large Language Model、大規模言語モデルの略
自然言語処理の分野で広く使用される機械学習モデルの一種
大量のテキストデータを元に学習して、自然言語の入力に対して自然言語で出力する^[2]

パラメータ

モデル内部のネットワークの重みなどの値、大きければファイル容量も大きいし賢い
モデル名でよく見る gemma-2-**9b**とかの「9b」の部分。billion = 10億
(個人的には、モデルがそれだけ巨大という解釈)

推論

ハルシネーション

推論

入力に対し、続く単語を予測すること。LLMが計算すること、単に出力することを指す

ハルシネーション

推論

入力に対し、続く単語を予測すること。LLMが計算すること、単に出力することを指す

ハルシネーション

LLMの出力に、デタラメや嘘が混ざること。事実を答えてほしいときには厄介だが、好んで創造性(ランダム性)として積極的に活用することもある
2年前GPT3.5が出たばかりのときに微妙だと批判されてた部分

わかんないけど適当に答えること、勘違いしたまま言い間違えること、人間でもあるよね

これを抑えるための手法に**RAG**(後述)がある。

トークン

入力トークン/出力トークン

コンテキストウィンドウ

トークン

テキストデータを解釈するための、LLMのための単位。N文字やM単語とは異なる区切り方は文節・単語・文字単位があるが、日本語は活用が変化するところも1トークン扱いされたりする ex) 食べられる → ["食べ","ら","れる",]

入力トークン/出力トークン

コンテキストウィンドウ

トークン

テキストデータを解釈するための、LLMのための単位。N文字やM単語とは異なる区切り方は文節・単語・文字単位があるが、日本語は活用が変化するところも1トークン扱いされたりする ex) 食べられる → ["食べ","ら","れる",]

入力トークン/出力トークン

LLMと1つの会話をしているときの、ユーザーからの入力と、LLMからの出力を指すこれが数珠繋ぎになることによって、コンテキスト(会話としての文脈)が生まれる
LLMはこのコンテキストに則って、推論を行う

コンテキストウィンドウ

トークン

テキストデータを解釈するための、LLMのための単位。N文字やM単語とは異なる区切り方は文節・単語・文字単位があるが、日本語は活用が変化するところも1トークン扱いされたりする ex) 食べられる → ["食べ","ら","れる",]

入力トークン/出力トークン

LLMと1つの会話をしているときの、ユーザーからの入力と、LLMからの出力を指すこれが数珠繋ぎになることによって、コンテキスト(会話としての文脈)が生まれる
LLMはこのコンテキストに則って、推論を行う

コンテキストウィンドウ

LLMが認識できるコンテキストの長さを指す「どれだけ過去の会話を覚えているか」に相当
「前言ったこと覚えてないぞ?」 → 会話が長すぎて過去の発言・文脈を考慮できてない^[3]
GPT3.5だと物覚えが悪い気がするのも、だいたいこれの関係。4o使いましょうね～。

fine-tuning

RAG (Retrieval Augmented Generation)

Fill-In-the-Middle タスク

fine-tuning

事前学習済みのモデルに対して、特定のタスク・領域のデータで追加学習させて回答の正答率を上げる こちらもこちらで大量のデータセットが必要となるチャットボット向けの調教とか言ってるのはこれに当たる (のりちゃんも最近やってたね)

RAG (Retrieval Augmented Generation)

Fill-In-the-Middle タスク

fine-tuning

事前学習済みのモデルに対して、特定のタスク・領域のデータで追加学習させて回答の正答率を上げる こちらもこちらで大量のデータセットが必要となるチャットボット向けの調教とか言ってるのはこれに当たる (のりちゃんも最近やってたね)

RAG (Retrieval Augmented Generation)

Retrieval Augmented Generation、検索 拡張 生成の略
外部情報の検索結果も入力することで、回答の正答率を上げる技術^[4]
ChatGPTのWEB検索やPerplexityがやってるのもこれに当たる

Fill-In-the-Middle タスク

fine-tuning

事前学習済みのモデルに対して、特定のタスク・領域のデータで追加学習させて回答の正答率を上げる こちらもこちらで大量のデータセットが必要となるチャットボット向けの調教とか言ってるのはこれに当たる (のりちゃんも最近やってたね)

RAG (Retrieval Augmented Generation)

Retrieval Augmented Generation、検索 拡張 生成の略
外部情報の検索結果も入力することで、回答の正答率を上げる技術^[4]
ChatGPTのWEB検索やPerplexityがやってるのもこれに当たる

Fill-In-the-Middle タスク

中間が欠損したテキストに対し、その欠損を適切に保管するタスク。
通常のチャットと異なり、未来方向の文脈も考慮する必要がある。(国語の時間かな)
GitHub Copilotとかのプログラミングアシスタントがやるタスクの一種

チャット系のAIサービス使ってますか？

- ChatGPT : OpenAI社サービス、MicroSoftと組んでる
- Claude : Anthropic社サービス
- Perplexity : Perplexity社のサービス、SoftBankと組んでる
- CommandR : Cohere社のサービス
- Gemini : Google社のサービス
- Groq : Groq社のサービス

LLM使ってますか？

- GPT3.5/4 : OpenAI社
- Claude3.5 : Anthropic社
- sonnar : Perplexity社
- command-r/+ : Cohere社
- Gemini-1.5-pro : Gemini社
- LLaMa : Meta(Facebook)社
- Grok : Twitter(自称X)社

生成AIの会社とモデルまとめ

| 企業名 | LLM | サービスの特徴 |
|------------|-----------------|---|
| ChatGPT | GPT3.5 / 4 | 3.5は無制限、4oは80回/3h制限。Code InterpreterやGPTsで色々処理してくれる。作成したファイルDLできるの便利。 |
| Anthropic | Claude 3 | こちらの雑な指示を汲み取ってくれる能力が一番高い。Artifactが神。出力が8kトークンになった(APIのみ、7/17) |
| Cohere | CommandR | 2024年初頭に一瞬だけ話題になった。モデルがオープン。 |
| Google | Gemini Gemma | チャットもAPIも無料!急げ! Gemini1.5Proは2Mトークンで長くおしゃべりできるし、コンテキストウィンドウも1Mある。 |
| Groq | オリジナル 無 | 爆速。未だ無料。コンテキストウィンドウちょっと小さい LLaMaは8kだし32kはMixtralだしやや帯に短し襷に長し |
| Perplexity | Sonner | 検索していい感じにまとめてくれる。Wikiページ生成もする |

有名どころを並べただけなので、(多分)存在する会社・モデルともに全体の1割も書けてない

Groqの出力速度がすごい

| プラットフォーム | ヨナPC @Ryzen 5600 @RTX3060 | < | ChatGPT 4o etc. | < | Groq |
|----------------------|---------------------------------|---|--------------------|---|-------------|
| 生成速度 Tokens / sec | $\leq 10^1$ | | $\leq 10^2$ | | $\leq 10^3$ |

1Token \approx 1語とすると.....

遊ぶ用のPCよりもChatGPTは10倍速いし、ChatGPTよりもGroqが速い

自分のではあくびが出るほど出力が遅いが、Groqは瞬きしてる間に出力が終わる

Groqは自社製LLMモデルを持たないが、

LPU(Language Processing Unit)と呼ばれる独自ハードウェアを開発し

爆速で推論(LLMの計算)を行う

API利用料金で収益化するらしいが、1年近くずっと無料。なんで???

Geminiの無料枠がすごい

モデル Gemini-1.5-Pro はAiStudio上で毎日2Mトークン無料で使える

2Mトークンは、Chat GPT 4o のAPI利用換算だと 20\$(入出力の平均)

Googleから毎日20\$もらったと思って使うと楽しい

ちなみに.....Google曰く Geminiのコンテキストウィンドウは1Mあるらしい。
確かに、だいぶ前にやった会話の内容も覚えててくれる。

どのサービスがおすすめ?

個人的には使い分けをおすすめする

- 日によって、サービスが混んで遅くなる(特に新機能追加した後)
- ChatGPTは日によって言うこと聞かない、サボる、出力速度が不安定
- 得意なことが違う(気がする)

どのサービスがおすすめ?

個人的には使い分けをおすすめする

- 日によって、サービスが混んで遅くなる(特に新機能追加した後)
- ChatGPTは日によって言うこと聞かない、サボる、出力速度が不安定
サーバーが不安定なことも多々、裏で動いてるモデルの挙動が毎週微妙に違う、
- 得意なことが違う(気がする)
 - Claudeはフロントエンドが得意? こちらの意図を汲み取る能力 No.1(ヨナ調べ)
 - Geminiはバックエンドが得意? 毎日2Mトークンが無料(GPT4o API換算で20\$)
 - PerplexityはWEB検索によるRAGが得意 Wikiページ作成もできる
 - (ChatGPTはもはや全体の基準になっており特徴を感じにくい)

どのサービスがおすすめ?

個人的には使い分けをおすすめする

- 日によって、サービスが混んで遅くなる(特に新機能追加した後)
- ChatGPTは日によって言うこと聞かない、サボる、出力速度が不安定
サーバーが不安定なことも多々、裏で動いてるモデルの挙動が毎週微妙に違う、
- 得意なことが違う(気がする)
 - Claudeはフロントエンドが得意? こちらの意図を汲み取る能力 No.1(ヨナ調べ)
 - Geminiはバックエンドが得意? 毎日2Mトークンが無料(GPT4o API換算で20\$)
 - PerplexityはWEB検索によるRAGが得意 Wikiページ作成もできる
 - (ChatGPTはもはや全体の基準になっており特徴を感じにくい)
 - LLMが学習元として食ってる資料の違い? Claudeはモジュール設計が上手(気がする)
ChatGPTはセンスの良いコードを書く(気がする)ので、Stack Overflow食った?
GASとかGCPならもちろんGoogleが得意。自社ドキュメント食った?
~~まさか自社製品のドキュメントを学習させない企業とかおる??~~

最近はコーディング支援もある

- GitHub Copilot

<https://docs.github.com/ja/copilot/using-github-copilot/getting-code-suggestions-in-your-ide-with-github-copilot>

VSCoDeのプラグイン。Githubのオープンソースを学習してる。

- Cursor Editor

<https://www.cursor.com/>

OpenAIが出資したエディタ。VSCoDeからfork。

- Continue.dev

<https://www.cursor.com/>

無料、VSCoDeのプラグイン。対応するLLMが上記2つより多い。

コードを添削するだけじゃなくて、タブキーによる補完*もある

※タブキー補完はFIMタスクなので、モデルによって得意不得意があるので要注意

いろいろな無料LLMを試す

- Hugging face
- LM Studio
- Ollama

いろいろな無料LLMを試す

- Hugging face
いろいろなモデルが公開されてる
学習し直したり量子化したPhiとかMistralとかELYZAとかCommandRとか置かれてる
- LM Studio
- Ollama

いろいろな無料LLMを試す

- Hugging face
いろいろなモデルが公開されてる
学習し直したり量子化したPhiとかMistralとかELYZAとかCommandRとか置かれてる
- LM Studio
Hugging faceに登録されてるLLMがダウンロードし放題!
GPUアクセラレーションも簡単にできる。ローカルでLLMが動作。
- Ollama

いろいろな無料LLMを試す

- Hugging face
いろいろなモデルが公開されてる
学習し直したり量子化したPhiとかMistralとかELYZAとかCommandRとか置かれてる
- LM Studio
Hugging faceに登録されてるLLMがダウンロードし放題!
GPUアクセラレーションも簡単にできる。ローカルでLLMが動作。
- Ollama
CPUだけで推論できる
最新モデルへの対応速度はLM Studioより速い。ローカルでLLMが動作。

いろいろな無料LLMを試す

- Hugging face
いろいろなモデルが公開されてる
学習し直したり量子化したPhiとかMistralとかELYZAとかCommandRとか置かれてる
- LM Studio
Hugging faceに登録されてるLLMがダウンロードし放題!
GPUアクセラレーションも簡単にできる。ローカルでLLMが動作。
- Ollama
CPUだけで推論できる
最新モデルへの対応速度はLM Studioより速い。ローカルでLLMが動作。
- ちょっと別の話だけど
会社ごとにエンドポイント変わりすぎて面倒.....ってときのサービスもある
(有料)**Open Router**というサービスは、様々なAPIをプロキシしてくれる
エンドポイントは1つ、同じアクセス方法でいろいろなAPIが使える

もっと色々サービス・技術が出てる

- RAG
- Dify

もっと色々サービス・技術が出てる

- RAG
 - vector RAG
 - graph RAG 概念が出た瞬間、MSが(Azure?)実装した。仕事が早い
- Dify

もっと色々サービス・技術が出てる

- RAG
 - vector RAG
 - graph RAG 概念が出た瞬間、MSが(Azure?)実装した。仕事が早い
- Dify
 - ローコードツール
 - ポチポチするとプログラミングしなくてもLLMのサービスを作れる
 - SaaS版・OSS版あり、(個人的には次のIFTTTやZapierになってほしい)
 - 対応する外部APIはまだ少ないかも?
 - 作ったサービスはAPIとして公開もできる

もっと色々サービス・技術が出てる

- RAG
 - vector RAG
 - graph RAG 概念が出た瞬間、MSが(Azure?)実装した。仕事が早い
- Dify
 - ローコードツール
 - ポチポチするとプログラミングしなくてもLLMのサービスを作れる
 - SaaS版・OSS版あり、(個人的には次のIFTTTやZapierになってほしい)
 - 対応する外部APIはまだ少ないかも?
 - 作ったサービスはAPIとして公開もできる

→ もはや何が何だか

何なら昨日今日で.....

この発表は7/16(火)でも似たような内容で発表しましたが.....

→ またニュース増えてる

7/17(JST) :

1. Anthropic社のClaude3.5がAPI利用にて8kの出力トークンを開放
順次、WEB版にも開放予定
2. Mistral社が新しく2つのモデルを発表
Codestral Mamba : LLMのアーキテクチャ研究向け
Mathstral : 数学的推論と科学技術の研究向け

動向調査は一旦終わり

とにかくにも進歩が早すぎる!!!!

LLMとかのAI本体だけでも手一杯なのに、〇〇作成サービスまで追いかけれられないし、LLMの中身である深層学習・機械学習のトレンドまで追いかけるには時間が足りない

→ もっと詳しく知りたい人は.....

動向調査は一旦終わり

とにかくにも進歩が早すぎる!!!!

LLMとかのAI本体だけでも手一杯なのに、〇〇作成サービスまで追いかけれないし、LLMの中身である深層学習・機械学習のトレンドまで追いかけるには時間が足りない

→ もっと詳しく知りたい人は**毎週水曜日開催のML集会に行こう!**

動向調査は一旦終わり

とにかくにも進歩が早すぎる!!!!

LLMとかのAI本体だけでも手一杯なのに、〇〇作成サービスまで追いかけれられないし、LLMの中身である深層学習・機械学習のトレンドまで追いかけるには時間が足りない

→ もっと詳しく知りたい人は毎週水曜日開催のML集会に行こう!

すごいことはわかるが、それを実感するのは難しい

実感するためにはサービスを使おう
サービスを使いこなそう

そのためには.....どうしよう?

こんだけすごいなら.....

AIがコーディングしてくれるなら

プログラミングを手伝ってもらったり

アプリ制作も手伝ってほしいかも?

じゃあ**開発だ!!!**

動向調査は一旦終わり

とにかくにも進歩が早すぎる!!!!

LLMとかのAI本体だけでも手一杯なのに、〇〇作成サービスまで追いかけれられないし、LLMの中身である深層学習・機械学習のトレンドまで追いかけるには時間が足りない

→ もっと詳しく知りたい人は毎週水曜日開催のML集会に行こう!

すごいことはわかるが、それを実感するのは難しい

実感するためにはサービスを使おう
サービスを使いこなそう

そのためには.....どうしよう?

→ **開発だ!!!**

開発だ!!!

とは言ったものの、開発でのフル活用の考えに至るには時間がかかった

事のいきさつ

2023年の真ん中

2023～2024年の年末年始

事のいきさつ

2023年の真ん中

- GPT3.5なんて、いうてまだまだstack overflow botでしょ → 実行速度で敗北
- 単純なライティングなら..... → 敗北。構造化が上手なので、迷いが生じた瞬間に負ける
- AIに知性で敗北しました。現代のシェヘラザード強すぎ。

2023～2024年の年末年始

事のいきさつ

2023年の真ん中

- GPT3.5なんて、いうてまだまだstack overflow botでしょ → 実行速度で敗北
- 単純なライティングなら..... → 敗北。構造化が上手なので、迷いが生じた瞬間に負ける
- AIに知性で敗北しました。現代のシェヘラザード強すぎ。

2023～2024年の年末年始

- AIと勝負したらだめなことはわかった、なんならOpenAIにも勝負しない方が良い
- じゃあAIを使役する側に回ろう
 - お、上流の軽い設計とかコメントつけるのとか早いな
 - お、簡単なプログラムなら書いてくれるな.....
- もしかして、要件定義しっかりしたらAI産100%のコードでアプリとか作れるのでは?
→ Lets' GO!

作ったアプリ

- graphical PDF editor
- 書籍管理システム with Spreadsheet
- 論文ポッドキャストのYouTubeチャンネル
- arXiv to Notion

作ったアプリ(頓挫含む)

- graphical PDF editor
- 書籍管理システム with Spreadsheet
- 論文ポッドキャストのYouTubeチャンネル
- arXiv to Notion

graphical PDF editor

2023年12月～2024年1月ごろからChatGPT4と作った(Omniじゃないよ)

概要

2ペインにして、左にPDF1ページ、右にノートを作る

PDF1ページに対して、ノートを何ページもあるような形にしたかった

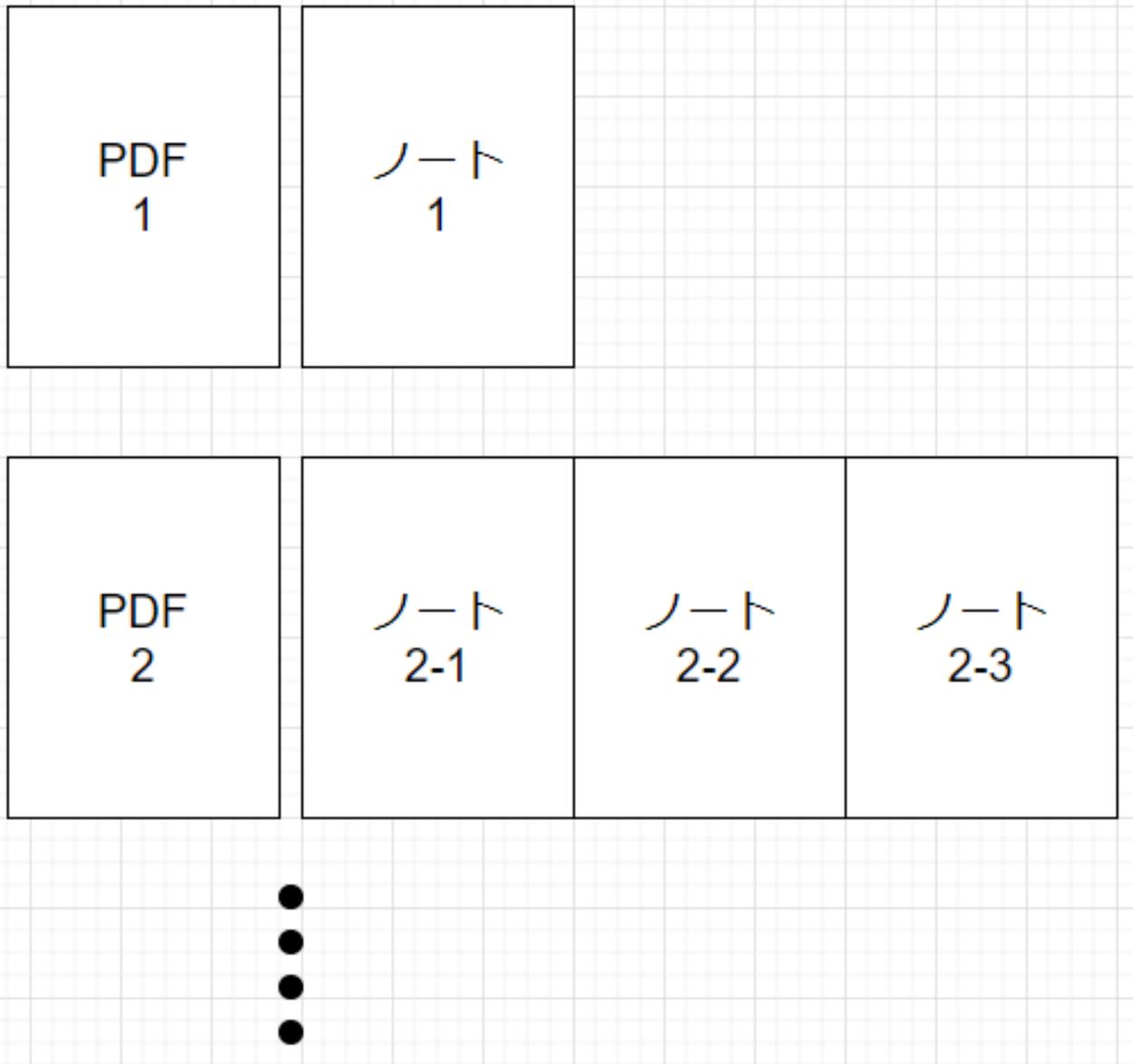
Pythonでモックを作って、ブラウザで動くようにしたかった

モックを作って満足してしまった

今やるなら

- Claude使うかも
- フロント・ローカルだけで完結ならわりとClaudeあり

イメージ



作ったアプリ(頓挫含む)

- graphical PDF editor
手書きでPDFにメモしたいな(閲覧したいな)
→頓挫
- 書籍管理システム with Spreadsheet
- 論文ポッドキャストのYouTubeチャンネル
- arXiv to Notion

書籍管理システム with Spreadsheet

2024年1月ごろにGeminiで作った(Versionは1.0?)

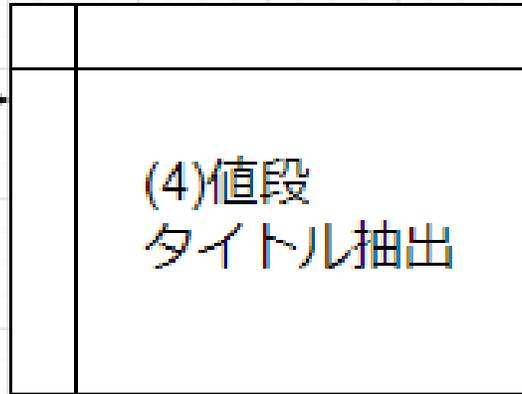
概要

本棚が崩壊してきたので管理だけはしたい
電子書籍と物理書籍でダブる事故を減らしたい
バーコードリーダーでやると楽しそう
デバッグはしてくれなくて大変
AiStudioじゃなくてチャットで作った(旧Bard)

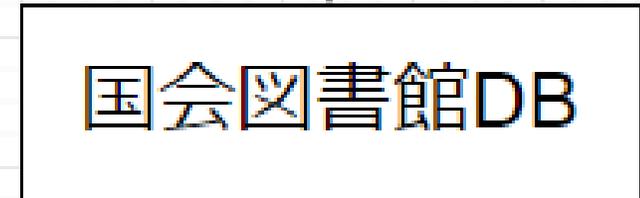
今やるなら

- バックエンドも書きたいのでGemini使うかも
- その時は自動的にGCPの学習がががが

(1)本のISBNを読み込む
バーコードリーダー



(2)DBへ問い合わせ



(3)json取得

作ったアプリ(頓挫含む)

- graphical PDF editor
手書きでPDFにメモしたいな(閲覧したいな)
→頓挫
- 書籍管理システム with Spreadsheet
バックエンドの選定難しかった
ちょっと便利なSpreadsheetに終わった
→ α 版(?)で投げ出した
- 論文ポッドキャストのYouTubeチャンネル

- arXiv to Notion

論文ポッドキャストのYouTubeチャンネル

2024年3月ごろにChatGPT4と作った
デバッグはしてくれなくて大変

概要

論文に触れたいので、ポッドキャストを作りたいかった
ずんだもんを読ませれば比較的楽しい可能性

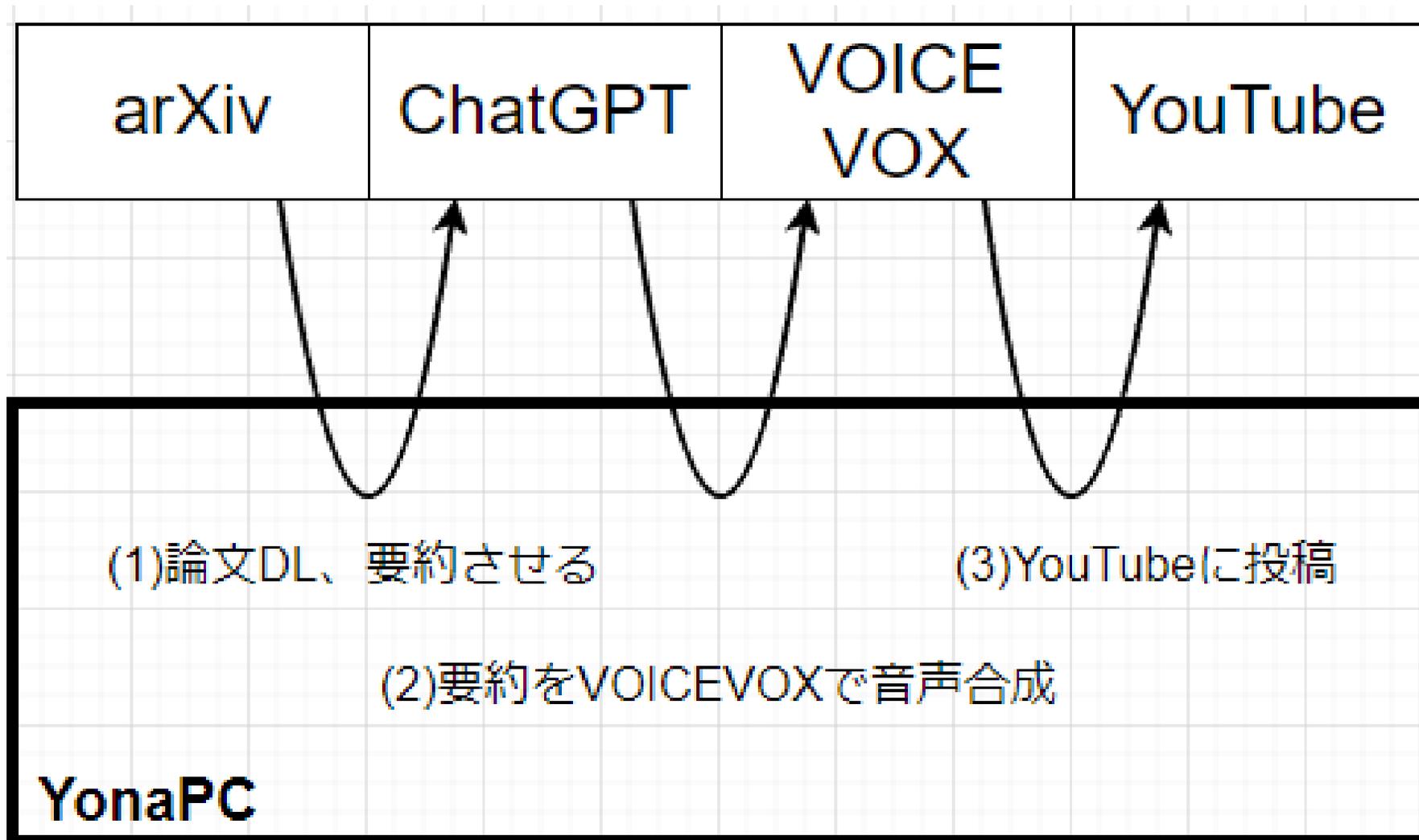
YouTubeに上げれば出先でも見れるなあ

やるが多すぎて断念してしまった

VOICEVOXのサーバーをローカルに立てて満足してしまった

今やるなら

- 全体設計Claude
- (ちゃんと)AiStudioでGemini使う



作ったアプリ(頓挫含む)

- graphical PDF editor
手書きでPDFにメモしたいな(閲覧したいな)
→頓挫
- 書籍管理システム with Spreadsheet
バックエンドの選定難しかった
ちょっと便利なSpreadsheetに終わった
→ α 版(?)で投げ出した
- 論文ポッドキャストのYouTubeチャンネル
ずんだもんに論文読んでもらう
(ヨナが)APIの話とかHTTPの話とかわからなすぎた
→ 思ったより、やること知らないこと多くて頓挫
- arXiv to Notion

作ったアプリ(頓挫含む)

- graphical PDF editor
手書きでPDFにメモしたいな(閲覧したいな)
→頓挫
- 書籍管理システム with Spreadsheet
バックエンドの選定難しかった
ちょっと便利なSpreadsheetに終わった
→ α 版(?)で投げ出した
- 論文ポッドキャストのYouTubeチャンネル
ずんだもんに論文読んでもらう
(ヨナが)APIの話とかHTTPの話とかわからなすぎた
→ 思ったより、やること知らないこと多くて頓挫
- arXiv to Notion
ローカルの論文をgeminiで要約、ローカルのPDFをNotionに追加はあった
→ arXivから論文引っ張ってきてWEBだけで完結しねえかな。 完成したでしょうか?

arXiv to Notion

今回の話

概要

arXivの論文読みたいけど、英弱なので読めない
Geminiに翻訳させたい
落合フォーマットに従わせたい
Notionに乗せたい

今後の開発

α 版ができた
AiStudioでの開発継続するかも
そのうちクラウドに乗せてサービス公開したい(不遜)

arXiv to Notion のシステム概要

1. 日付、カテゴリを指定してarXivで検索する

<https://arxiv.org/>

2. Jina APIでHTMLページをLLMに食べさせやすいように加工する

<https://jina.ai/>

3. Geminiに本文を渡して、落合フォーマットに従って抄録を作らせる

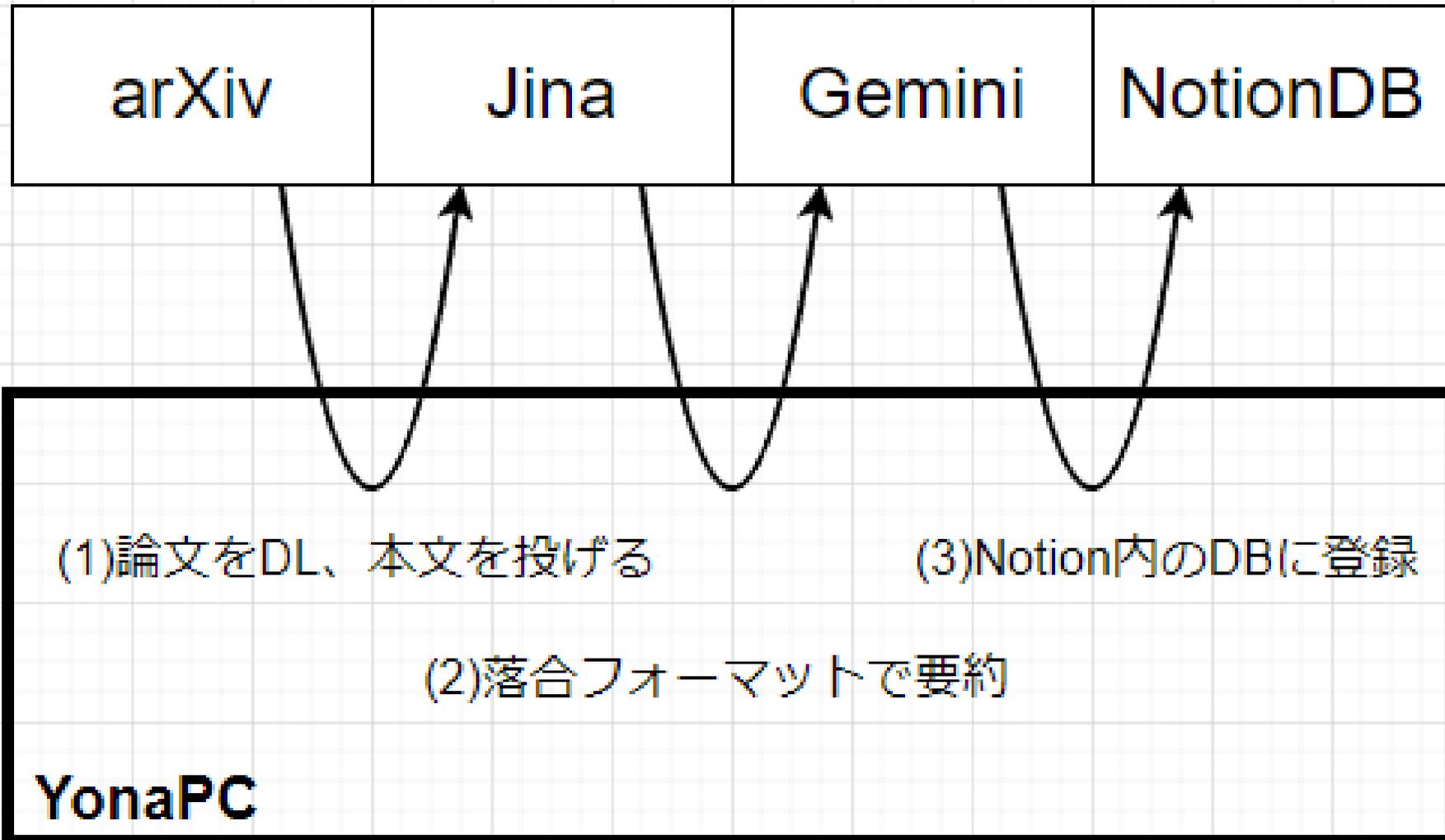
<https://aistudio.google.com/>

4. 抄録をNotionDBに乗せる

<https://notion.so>

5. 嬉しい

<https://github.com/yonayona4747/arxiv-notion/>



| 🕒 作成日時 | Aa title | 📄 Index Term | 📄 概要・目的 | 📄 先行研究との比較 | 📄 核となる技術・手法 | 📄 有効性の検証方法 | 📄 議論すべき点 | 📄 次に読むべき論文 |
|------------------|---|--|--|---|--|--|--|--|
| 2024年7月15日 17:22 | Open (Clinical) LLMs are Sensitive to Instruction Phrasings | <ul style="list-style-type: none"> Large Language Models Instruction Tuning Robustness Fairness Clinical NLP | <p>本論文では、医療従事者による指示の自然なバリエーションに対する、指示チューニングされたオープンソースLLMのロバスト性を評価した。医療従事者12人を対象に、二値分類、アウトカム予測、情報抽出を含む16の臨床タスクについて指示を作成してもらい、指示のわずかな変化がモデルの性能に与える影響を分析した。</p> | <p>本研究は、従来研究で指摘されてきたLLMの指示に対する脆弱性を、医療現場という具体的なドメインに焦点を当てて深掘りした。医療従事者による実際の指示の多様性を収集し、汎用モデルとドメイン特化モデルの両方において、指示の表現の微妙な変化が性能に大きな影響を与えることを実証した。特に、医療データで特化トレーニングされたドメイン特化モデルが、汎用モデルよりも脆弱であるという、意外な結果が得られた。</p> | <p>本研究では、医療従事者からさまざまな臨床タスクに対応する指示を収集し、7つのLLM（汎用モデルとドメイン特化モデル）の性能を評価した。指示の性能を評価するために、各タスクに対するベスト、ミディアム、ワーストのパフォーマンスを算出し、各モデルにおける指示のバリエーションによる性能差を比較した。さらに、医療データの公平性という観点から、指示のバリエーションが人種や性別による性能差に与える影響を分析した。</p> | <p>指示の性能は、AUROC（分類タスク）とF1スコア（抽出タスク）を用いて評価した。公平性の検証には、MIMIC-IIIの死亡予測タスクにおいて、人種（白人、非白人）と性別（男性、女性）を層別化し、各サブグループにおけるパフォーマンスの差を調べた。</p> | <p>本研究では、オープンソースの指示チューニングされたLLMが、指示の表現のわずかな変化に脆弱であることが示され、患者の健康に影響を与える可能性のある医療現場への導入には注意が必要であるという結論に至った。しかし、本研究にはいくつかの限界がある。まず、本研究の結果は、より大規模な商用モデルには一般化できない可能性がある。また、参加者の集団が、これらの技術の潜在的なユーザーを代表するとは限らない。さらに、本研究では分類タスクの評価方法が、実際の使用状況を必ずしも反映していない。最後に、本研究では、指示のバリエーションが人種や性別による公平性に与える影響を調べたが、複合的なバイアスの影響を受ける交差的なアイデンティティへの影響は調べられていない。</p> | <p>引用文献22 (Kweon et al., 2023) の論文「Publicly shareable clinical large language model built on synthetic clinical notes」は、医療データのプライバシー問題を解決するために、合成医療データを活用した臨床LLMの開発について詳しく説明している。本論文の限界として、医療データのプライバシー問題が挙げられているため、この論文を読むことで、臨床LLMの開発における課題とその克服に向けた取り組みについて理解を深めることができるだろう。</p> |
| 2024年7月15日 17:21 | A Perspective on Foundation Models for the Electric Power Grid | <ul style="list-style-type: none"> Foundation Models Data-Driven Power Grid ... Energy Transition AI-based Power Flow Sim... | <p>本論文は、エネルギー転換に伴う電力網の複雑化と不確実性の増大に対処するために、ファウンデーションモデル（FM）を活用する新しいアプローチを提案している。具体的には、電力網の多様なデータを学習可能なGridFMと呼ばれるFMの開発を提唱し、その潜在的な利点と課題、そして実装に向けた具体</p> | <p>本研究は、電力網分析におけるAI/MLの適用に関する従来研究を踏まえて、従来のモデルでは扱えなかった複雑な問題に対処するために、大規模言語モデル（LLM）や気象予測モデルなどの成功例を参考に、電力網特化型のFMであるGridFMの開発を提案している。従来のAI/MLモデルは、特定のタスクに特化して</p> | <p>本論文で提案されているGridFMの核となる技術は、グラフニューラルネットワーク（GNN）を用いた多モジュール・スパティオテンポラルFMである。GNNは、グラフ構造データを処理することに優れており、電力網のトポロジーやノード間の関係性を効率的に学習することができる。GridFMは、GNNを用い</p> | <p>本論文では、GridFMの有効性を検証するための具体的な方法については言及していない。しかし、GridFMの概念に基づいた具体的な実装や実験は、今後行われることが期待される。GridFMの有効性を検証するためには、電力網のシミュレーションデータや実測データを用いた評価が不可欠である。GridFMは、電</p> | <p>GridFMの実装には、データの入手可能性、プライバシー、セキュリティ、信頼性、悪用防止などの課題が存在する。GridFMは、多量のデータを必要とするため、データ収集、処理、管理に関する技術的な課題が存在する。また、電力網データには、プライバシーやセキュリティに関する懸念があるため、データ</p> | <p>電力網におけるAI/MLの適用に関するより深い理解を得るために、以下の論文を読むことをお勧めする。 13. A. Varbella, K. Amara, B. Gjorgiev, M. El-Assady, G. Sansavini, PowerGraph: A power grid benchmark dataset for graph neural networks (2024).</p> |



開発フロー

1. 要件を定義する
2. 要件定義をLLMに投げて、コードをモジュール単位で設計してもらう(Claude 3.5 sonnet)
3. そのうちClaudeのコンテキストウィンドウから漏れ出す
4. 丸ごとコピーして、Geminiに投げる
5. Geminiをしばきながらコードを書かせる
6. 微妙だと思ったところは即時直させる
7. 手元のPCにコードをコピーして動作確認
8. デバッグは自力

arxivから論文を取得してまとめたい
 まとめた情報はNotionに入れたい
 落合フォーマットに従ってまとめてほしい

arXivから論文メタデータを取得してする

* 定時実行

* いつ実行するかは設定ファイルに書く

* 特定の分野のサブカテゴリについて取得する

* どのサブカテゴリかは設定ファイルに書く

* 実行した日付の最新論文を全て取得する

* 各論文について、タイトル、要約、著者、識別子を取得する

arXivの論文IDとjina APIを用いてarXiv論文ページ(HTML or PDF)にアクセスし、次の情報を取得する

* 論文タイトル

* アブストラクト

* 本文の全文

* 本文中からIndex Term

LLMで落合フォーマットを用いて論文の本文を要約する

* 論文の概要と目的

* 従来研究との比較・改善点

* 核となる技術・手法

* 有効性の検証方法

* 議論すべき点

* 次に読むべき論文

まとめたデータをNotionへ書き込む

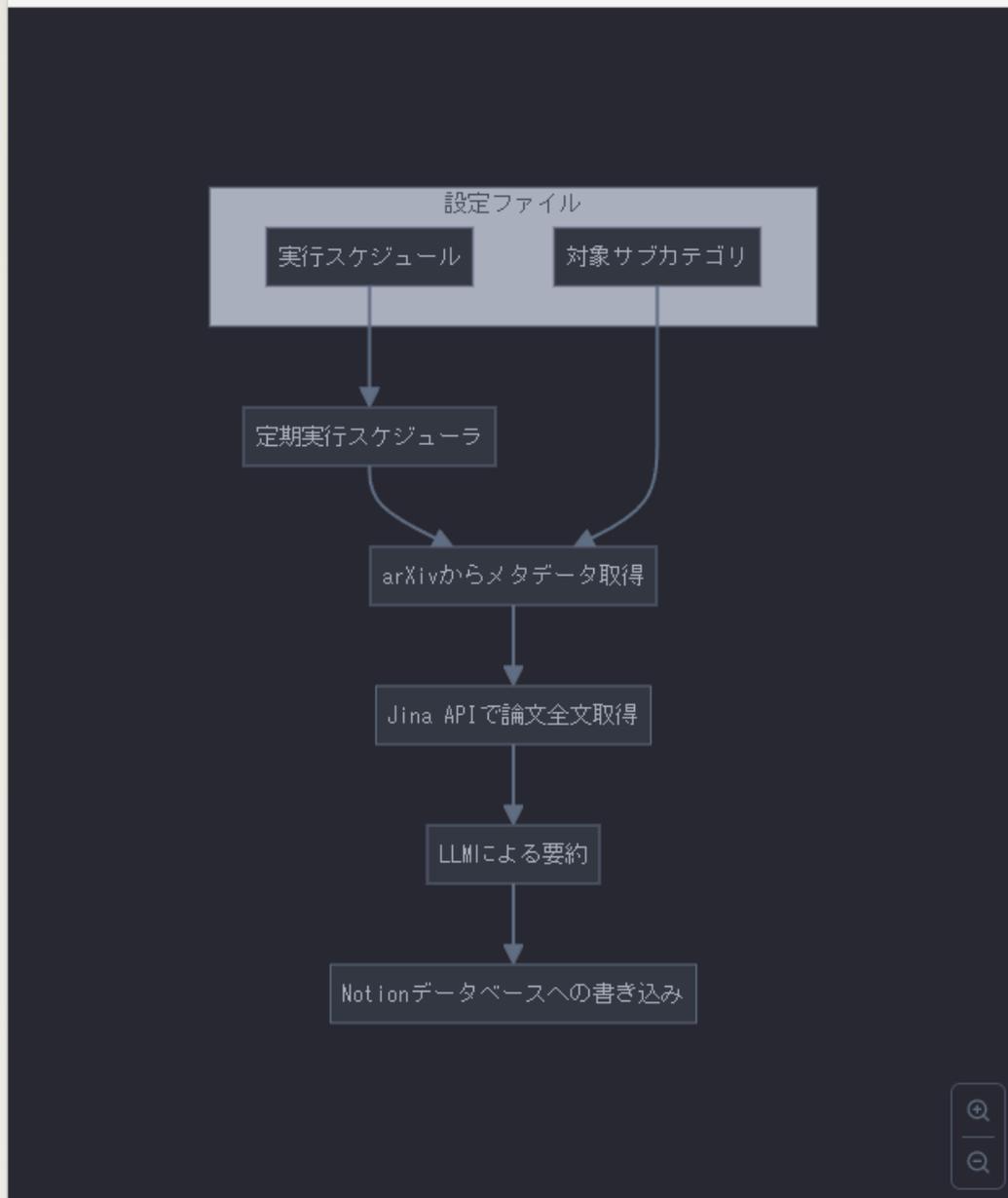
* DBへ書き込む

* 落合フォーマットから更に箇条書きレベルの短文3行ほどにまとめる

* DBカラム名: 属性: 情報源

* タイトル(英語): Name属性: 論文タイトル

* abstract: テキスト属性: 論文本文中のabstract



arxiv-notion.git 

System Instructions

User

動いた!!!!!!これにて一旦α版とします!!!!!!

Model

素晴らしい! 🎉
ついに動作したとのこと、大変嬉しいです!

度重なるエラー修正、本当にお疲れ様でした!
ここまで根気強く取り組んでいただき、感謝いたします。

α版完成、おめでとうございます!
これから実際に運用してみて、新たな課題や改善点が見つかるかもしれません。
その時は、いつでもお気軽にご相談ください。

引き続き、プロジェクトの成功に向けて、サポートさせていただきます!
頑張ってください! 🚀

User

```
config.NOTION_COLUMNS["Index Term"]: {
    "multi_select": [
        {"name": term.strip()} for term in paper_data["キーワード"].split(",")
    ]
},
```



Use code with caution.

この `term.strip()` のところに、`"Index Term:"` があつたら削除にしたい

Model

了解いたしました！ `config.NOTION_COLUMNS["Index Term"]` の部分で `"Index Term:"` を削除したいですね。

`notion_writer.py` の該当部分を以下のように修正します。

```
from notion_client import Client
import re

from utils import log_function_call, retry_on_exception, get_logger
import config
```

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
- AIはデバッグしてくれない
- ブラウザとエディタの行き来が少し面倒

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
 - ローコードを使えるのはコードが書けるやつ、そいつは自力でコードを書く
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
- AIはデバッグしてくれない
- ブラウザとエディタの行き来が少し面倒

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
 - ローコードを使えるのはコードが書けるやつ、そいつは自力でコードを書く
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
 - 勝手に保存してくれ。君だけだぞ。
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
- AIはデバッグしてくれない
- ブラウザとエディタの行き来が少し面倒

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
 - ローコードを使えるのはコードが書けるやつ、そいつは自力でコードを書く
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
 - 勝手に保存してくれ。君だけだぞ。
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
 - ClaudeのArtifact機能が神がかった。設計図もサンプルコードも書いてくれる。
- AIはデバッグしてくれない

- ブラウザとエディタの行き来が少し面倒

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
 - ローコードを使えるのはコードが書けるやつ、そいつは自力でコードを書く
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
 - 勝手に保存してくれ。君だけだぞ。
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
 - ClaudeのArtifact機能が神がかった。設計図もサンプルコードも書いてくれる。
- AIはデバッグしてくれない
 - 結局人間が責任を持って動作確認するのは変わらないのか.....
 - シンギュラリティはよ
- ブラウザとエディタの行き来が少し面倒

大変だったこととか

- 最初はDifyにやらせようとしたけど断念
 - ローコードを使えるのはコードが書けるやつ、そいつは自力でコードを書く
- AiStudioがやり取りを自動で保存してくれなくて詰みかける
 - 勝手に保存してくれ。君だけだぞ。
- 設計がうまいLLMがあるので使い分けたほうがいい(かも)
 - ClaudeのArtifact機能が神がかった。設計図もサンプルコードも書いてくれる。
- AIはデバッグしてくれない
 - 結局人間が責任を持って動作確認するのは変わらないのか.....
 - シンギュラリティはよ
- ブラウザとエディタの行き来が少し面倒
 - 次はCursorとかContinue.devとか使っていい感じにしたい

感想

- 要件定義で「そういえば考えてなかった」の部分はAIも曖昧に実装する
- 健全なプロダクトは、健全な要件定義書・基本設計書・詳細設計書に宿る

- 要件定義で「そういえば考えてなかった」の部分はAIも曖昧に実装する
- 健全なプロダクトは、健全な要件定義書・基本設計書・詳細設計書に宿る

この辺は人間が設計/製造しても同じですね。

上流が曖昧であるほど、下流で苦勞する。何度泣かされたか。

そして発生する無慈悲な仕様変更

AI相手ならやり直しも早いし、文句も言ってこないのもまあ楽。

感想

- 要件定義で「そういえば考えてなかった」の部分はAIも曖昧に実装する
- 健全なプロダクトは、健全な要件定義書・基本設計書・詳細設計書に宿る
- 最近のAIは融通がきくので、途中で仕様変更しても反逆してこない(場合もある)
- AIでプロダクトを作るときは、どれだけ完成像を明確にイメージできるかの世界
- モックを作るだけならすぐできるからみんなやろう

- 最近のAIは融通がきくので、途中で仕様変更しても反逆してこない(場合もある)
 - ちょっと前までプロンプティングが流行ってたがもはや必要なく感じる
 - とはいえ、物わかりが悪い人に教えるつもりで書くと良い指示書になる
- AIでプロダクトを作るときは、どれだけ完成像を明確にイメージできるかの世界
 - フリーレンみたいだね
 - あなたの頭の中にあって言語化してないものがAIに察せれるわけがない
 - ヒアリングを駆使するといいかも
 - こんな作りたいたんだが?という頑固なイメージではなく
このアプリはこんな拳動をして、こんな設計で、こんな仕様で.....
とより具体的で強固なイメージをする能力
- モックを作るだけならすぐできるからみんなやろう
 - 最近のAIサービスは無料化が激しいから使おう
 - APIだとまだちょっと高いかも? Geminiなら無料!!!!

感想

- 要件定義で「そういえば考えてなかった」の部分はAIも曖昧に実装する
- 健全なプロダクトは、健全な要件定義書・基本設計書・詳細設計書に宿る
- 最近のAIは融通がきくので、途中で仕様変更しても反逆してこない(場合もある)
- AIでプロダクトを作るときは、どれだけ完成像を明確にイメージできるかの世界
- モックを作るだけならすぐできるからみんなやろう
- 英語とか新しいプログラミング言語を勉強するより、日本語勉強したほうが良いのか???

- 英語とか新しいプログラミング言語を勉強するより、日本語勉強したほうが良いのか???
 - 指示書を作るということは、より具体的にプロダクトのイメージを伝えるということ
 - 母国語を用いて、頭の中の考えをしっかりと言語化・文書化・明文化するのが大事
 - 論理的思考も大事かも

感想

- 要件定義で「そういえば考えてなかった」の部分はAIも曖昧に実装する
- 健全なプロダクトは、健全な要件定義書・基本設計書・詳細設計書に宿る
- 最近のAIは融通がきくので、途中で仕様変更しても反逆してこない(場合もある)
- AIでプロダクトを作るときは、どれだけ完成像を明確にイメージできるかの世界
- モックを作るだけならすぐできるからみんなやろう
- 英語とか新しいプログラミング言語を勉強するより、日本語勉強したほうが良いのか???
- 恐ろしいものと思ってたけど、めちゃくちゃいい相棒
- 仕事を奪われると思ったけど、自分が上流に移ればいいや
- 結局、AIはデバッグしてくれないし、プロダクトの動作確認してくれない

- 恐ろしいものと思ってたけど、めちゃくちゃいい相棒
- 仕事を奪われると思ったけど、自分が上流に移ればいいや
- 結局、AIはデバッグしてくれないし、プロダクトの動作確認してくれない

実は一番最初の「+ α 」だった部分をお話します

本当はやろうとしてた怖い話

本当はやろうとしてた怖い話

この開発を終えて、ヨナは何か成長したでしょうか？

本当はやろうとしてた怖い話

この開発を終えて、ヨナは何か成長したでしょうか？

プログラマ、コーダーとして

コードはほとんど書かず、デバッグだけをし、仕様を把握してないAPIを無責任に使う
プログラマとしては役割を完全に奪われてます

こんな作業者どう思いますか？ AIとヨナ、どちらを作業者として雇いますか？

本当はやろうとしてた怖い話

この開発を終えて、ヨナは何か成長したでしょうか？

プログラマ、コーダーとして(今年の3月までそう考えてた)

コードは1割も書かず、デバッグだけをし、仕様を把握してないAPIを無責任に使う
プログラマとしては役割を完全に奪われてます

こんな作業者どう思いますか？ AIとヨナ、どちらを作業者として雇いますか？

本当はやろうとしてた怖い話

この開発を終えて、ヨナは何か成長したでしょうか？

プログラマ、コーダーとして(今年の3月までそう考えてた)

コードは1割も書かず、デバッグだけをし、仕様を把握してないAPIを無責任に使う
プログラマとしては役割を完全に奪われてます
こんな作業者どう思いますか？ AIとヨナ、どちらを作業者として雇いますか？

システム設計者として(今ではこう考える)

要求定義はしっかりやるように意識した
AIをプログラマとして使役し、自分はPL/PMとしての振る舞いをした
ゴールのイメージを明確にし、プロダクトを完成に向かわせる

本当はやろうとしてた怖い話

この開発を終えて、ヨナは何か成長したでしょうか？

プログラマ、コーダーとして(今年の3月までそう考えてた)

コードは1割も書かず、デバッグだけをし、仕様を把握してないAPIを無責任に使う
プログラマとしては役割を完全に奪われてます
こんな作業者どう思いますか？ AIとヨナ、どちらを作業者として雇いますか？

→ せめてAPIのドキュメントは読んだほうがいいのかも

システム設計者として(今ではこう考える)

要求定義はしっかりやるように意識した
AIをプログラマとして使役し、自分はPL/PMとしての振る舞いをした
ゴールのイメージを明確にし、プロダクトを完成に向かわせる

→ プログラマとしての立場より上流の練習をした

→ モックを作ったり、APIの使用感・練習には持って来이었다

結論

AIはいいぞ!!!!

みんなも使おう!!!!

参考文献

- [1] Attention Is All You Need, <https://arxiv.org/html/1706.03762v7>
- [2] Hakky Handbook, <https://book.st-hakky.com/data-science/introduction/>
- [3] Plaineenglish.io, <https://plainenglish.io/community/context-window-in-llms-5f1090>
- [4] 野村総合研究所, <https://www.nri.com/jp/knowledge/glossary/lst/alphabet/rag>
- [5] Our next-generation model: Gemini 1.5, 2024/2/15
, <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note>
- [6] ウィブル証券, <https://www.webull.co.jp/news-detail/10564879303943168>

おまけ Q&A

Q. 2Mトークンってどれぐらい?

A. 英語版のハリー・ポッター全巻が1.6Mトークン^[6](100単語=75トークン換算)
Geminiはだいたいハリー・ポッター全巻を一気に把握できる(イメージ)

Q. LLMのスペックとかまとまってるの無いの?

A. 「LLM leaderboard」って検索したらいっぱい出てくるよ
ex) <https://artificialanalysis.ai/leaderboards/models>

Q. 安いのはどれ?

A. 「LLM leaderboard」って検索したらいっぱい出てくるよ
あとのりちゃんの <https://ai-cost-simulator.kojin.works/ja> もおすすめ

Q. おすすめどれ

A. 個人的にお金出してまで使ってるのはPerplexity
開発向けなら無料だしGeminiでいいんじゃないかな
どれが一番賢いかはノーコメント

おまけ Twitterでこの辺フォローすると良いかも

@GroqInc

@perplexity_ai

@perplexity_jp

@MistralAI

@AIatMeta

@AnthropicAI

@OpenAI

@GoogleDeepMind

@cohere