AJACS「ChatGPT等の生成AIツールを研究活動に活用する注意点を知って・学んで・使う」
講義中にいただいた質問への回答

No	内容	備考
1	「文献調査AIのOSS活動」に参加してみたいのですが、どうすればよいでしょうか？	紹介ページを作成しました。詳しくはこちらをご覧ください！ https://gensurv.notion.site/Home-080bd169f48849568ef001a4aa08ca1e
2	今のLLMで山田先生が「こんな事ができたらいいな」と思う、まだ実現していない事があればお聞きしたいです。	ユーザー個人の秘書のようなエージェントです。常にユーザーが見聞きする情報を共有し、同じコンテキストを共有してくれるようなものの登場に期待しています。
3	実験ノートをOneNoteなどにテキストで自由記述で書いた物から、生成AIで構造化データを作れますか？	非構造化データの構造化は生成AI活用のメジャーなトピックですね！LangChainを使う場合は、Pydanticというデータ型を定義するライブラリで期待する出力の型を設定できます。私は愛用しています。以下のページを参考にしてみてください。 https://python.langchain.com/v0.2/docs/concepts/#output-parsers プロンプトが長く煩雑にはなりますが、ChatGPTでも同様のことはできると思います。まずはそこからやってみて、対象ドキュメントが増えてきたりして自動化のニーズが高まったらシステム化すると良いでしょう。
4	文章添削をしてくれるAIはありますか？　主語や目的語、単語の説明不足、流れが散漫などの理由で読みづらい既に書かれた日本語の文章を、読みやすい文章に書き換える業務に利用したいです。	既存サービスは存じません。私ならChatGPTに「以下の文章は主語や目的語、単語の説明不足、流れが散漫などの理由で読みづらいです。読みやすい文章に書き換えてください」と入力してみます。まずはそこから試してみてはいかがでしょうか！
5	ダークデータを活用する際、AIのミスをどうやって確認したらよいのでしょうか？	定量的な評価指標を作成します。つまりは正解データですね。そうです、ダークデータの活用に希望が見えるかもしれませんが、結局のところ最後は構造化データを作って定量的評価をする必要があるのです…。ただ従来の深層学習では大量のデータを作らないとまずお試しもできませんでしたが、生成AIはひとまずクイックに定性的評価をできるところがミソです。この時点で無理そうならデータ作成にコストを割かなくてすみます。
6	AIのアルゴリズムの不明点について、ChatGPT4oに質問することが多いのですが、その回答が正しいか、ハルシネーションかを見分ける方法についてアドバイスをお願いします。	私は自分が知らないことをChatGPTに聞くときは、何で検索すれば良いのかのキーワードを得るため、調査の方向性を大まかに決めることを目的とします。最終的にはオリジナルのソースに当たります。AIのアルゴリズムについてであれば論文に辿り着くことが多いかと思います。そういうときはさらに論文を生成AIに読ませて詳しく掘り下げてみるという使い方をします。一回のやり取りで完全な回答を得ようとするのではなく、自分が1人でやっていた作業を伴走してもらいながら効率化してもらうという気持ちで使いこなすと良いのではないでしょうか。生成AIの回答に対して01で判断するのではなく、不正確なことがあると理解した上でうまく付き合うことをお勧めします。
7	公開されている画像生成モデル（img2img）を教育・研究用途で使う際、モデルがどのような学習データを使っているかをどこまで気にする必要があるでしょうか？　生成した画像は直接的には公開しない（論文等でもモザイクをかける）つもりではあります.	私個人の考えとして、著作権侵害だけを考えるなら、意図して他者の著作物に類似させていなければ問題ないと考えています（当然ですが専門家ではないので保証はできません。詳しくは法律の専門家にご相談ください）。 また、著作権侵害に関わらなくとも反感を買う可能性が高い領域には気をつけています。例えば画像生成AIのNovelAIはDanbooruというサイトの画像を学習に用いています。しかしDanbooruには著作者の許可を得ずにアップロードされた著作物が含まれており、結果的にNovelAIは著作者に許可を取らずに著作物を学習に利用してしまったことが明らかになっています。AIモデルの学習に使うことは著作権侵害ではないというのが一般的な見解ですが、それと感情は別物です。個人的には誰かの感情を逆撫でするようなものを使うのならリスクが許容範囲内か判断するよう気をつけています。例えば一部の研究者にしから見られない論文で研究目的で使う分にはリスクは低いでしょうが、そのようなモデルを使って教育をしていることが大々的にニュースなどで報道されればSNSで攻撃を受けるリスクが高いだろうなと思います。
8	LLMは大量のデータでトレーニングされた統計言語モデルとありますが、「大量のデータ」とはどのようなデータなのでしょうか？新聞や書籍でしょうか？	企業が開発するモデルの場合、学習データの全ては明らかにされていません。一つの競争優位性であり、それをわざわざ公開する義理もありません（大抵痛くもない腹を突かれることになるでしょうし）。 研究目的で公開されたデータセットは様々なものが存在するので「llm datasets」などで調べてみてください。Common Crawlなどが有名ですね。
9	AIを使いこなすためには使用しているPCの性能はどの程度影響するでしょうか？	ChatGPTなどウェブのサービスで使う分にはブラウザを立ち上げられれば十分です。計算はサービス提供者のサーバーで行われるため、手元のPCのスペックは必要ありません。スマホでもChatGPTアプリが出ていますよ。

No	内容	備考
10	今日、お話頂いたレベルにChatGPTを使うのに必要なCPUパワーはどの程度でしょうか？　普段持ちあるくレベルのPCでどのくらいまでできるのでしょうか？	（No. 9の回答を参照してください）
11	LLMへ入力した内容はどのように保存・蓄積されているのでしょうか？	メモリとストレージどちらかで回答が変わりますが、おそらく後者への質問でしょうか。それはサービスプロバイダーのみ知ることです…。
12	生成AIの電気使用量などの環境負荷と利用メリットとのバランスについてご意見を伺いたいです。	まさにそのような視点は本日紹介したカタストロフィに相当するような長期的なリスクですね。国内ではALIGN (https://www.aialign.net/) などが似たような関心を持っていると思います。一度ご覧ください。
13	最近の論文(Nat Commun, 2024, https://doi.org/10.1038/s41467-024-48005-w)のMethodsに"The prompting strategy"としてGPT-3.5をどう使ったか記載されていました。このような手法の記述のような流れは続くと思われますか？	さらっと拝見しました。道具としてLLMを使う以上、再現性の担保のためprompting strategyや実際のpromptを記載する必要があると思います。したがって、続くと思います。
14	Perplexityは、紹介されませんでしたが、他と比較してどうでしょうか？	個別のチャットボットサービスについて意見を聞かれることがしばしばあるのでまとめました。端的に言えば自社のモデルを持っているか（資本関係で取り込んでいるか）どうかに着目しています。 OpenAI (GPT), Google (Gemini), Anthropic (Claude)はいずれも自社モデルを持っており、MicrosoftはOpenAIに出資しています。 またその上で各企業が学習データや計算資源を持っているか（資本関係で取り込んでいるか）も重視しています。Google, Microsoftは言わずもがなクラウドサービスを運営していること、検索エンジンを持っていることが大きな強みです。OpenAIとAnthropicはそれぞれMicrosoft, Amazonから出資を受けています。 以上が私が上記4社の4つのサービスをおすすめする理由です。その観点から、他社のAPIを叩いて運営しているチャットボットサービスは余程のことがないと積極的に使い込むことはないです。 また特化型の生成AIツールに関しても、特定用途のプロンプトを書いてくれた、だけが価値なら使うツールを1つ増やすデメリットが勝ちます。特化型のもので唯一愛用しているのはGitHub Copilotですが、これも本業のサービスがソースコードという学習データを大量に保有していること。またMicrosoftから出資を受けていることなどが重要なポイントですね。 その観点から考えると、YouTubeという動画サービスを抑えているGoogleは生成AIの動画・音声方面の成長にアドバンテージがあると思っています。 また私の専門でもある論文関連に関してはGitHubがソースコードを押さえているように、出版社が大量の学習データ（＝論文）を保有しています。IT王者（Google, Microsoft, Amazon）かLLM強者（OpenAI, Anthropic）と組んで論文関連の特化型AIサービスが出るんじゃないかなぁと予想しています。
15	リートンってどうでしょうか？　AI検索とChatGPT4が使えるとされています。	(上記の回答をご参照ください。)
16	ChatGPTの無料版から有料版に切り替えるタイミングを教えてください。	無料版のフラグシップモデルの利用枠に不満を覚えたタイミングがおすすめです！
17	自分のGPTsを公開する際の注意点を教えてください。	GPTsを作成する際に個人情報や機密情報を入れないようお気をつけください。またあえて世界中に公開する目的がないのなら、公開範囲を利用者の範囲で留めるのがよろしいかと思います。
18	公開されているGPTsに微修正を加えて利用することはできますか？	GPTsの内部のプロンプトは外部から確認できない仕様です（GPTsが出始めた頃はプロンプトを引き出す攻撃が流行り、コピーGPTsがたくさん生まれるなどありましたが、現在簡単な穴は塞がれたと認識しています）。
19	RAGで追加した情報はChatGPTにも学習情報として取り込まれることになりますか？	RAGを実装する場合、API経由でLLMを利用します。もともと利用規約でOpenAIのAPI経由でのLLMの利用では学習情報が取り込まれない、ということにはなっていましたが、規約変更によりどうとでも読み取れる表現になっています。気にされる場合は別の手段を用いるのがよろしいかと思います。
20	論文や知財情報に対するRAG開発のコストを教えてください。	一般的なビジネスの文書はPowerPointなどのOffice製品を対象とすることがあり、それと比較すると構造化がされている論文は扱いやすいです。一方で特許などは論文と比べて文章量が多かったり、特殊な書き方で難読化されており一筋縄ではいかないと理解しています。

No	内容	備考
21	今後は LangChain よりも Dify がよさそうでしょうか？	いずれもLLMアプリケーションを開発するためのツールですが、LangChainはフレームワークであるのに対して、Difyはローコードツールです。言い換えるとLangChainはプログラマー向けのものであり、Difyは非プログラマー向けのものと言えるでしょう（やや乱暴な括りですが）。Difyは簡単に使える分、複雑なことや大規模なことには向きません。そういう場合にはLangChainを採用すると良いでしょう。
22	生成AI Agent 向けに、データベース側が用意・準備しておくとよい点はありますか？	素晴らしい質問ありがとうございます！テーブル名やカラム名は実態を反映した名前を使うことがおすすめです。エージェントがDBを扱う際にそれらの情報をもとに判断する設計にすることがあるからです。また、DBの内部にDBのメタ情報を書き込めない場合は別途READMEなどを丁寧に残していただけると助かる方が多いかと思います。