

データアントレプレナーカンファレンス
—A.I.・データサイエンスと「理解」—

最新研究から読み解くAIを理解するトレンド

2022年1月20日

株式会社TDAI Lab



自己紹介



福馬 智生

Tomoki Fukuma

@fukuma_tomoki



経歴

2016年 株式会社TDAI Lab創業 (代表取締役社長)

2018年 東京大学大学院工学系研究科修士課程修了

2021年 東京大学大学院工学系研究科博士課程修了

興味領域

Fairness, Unbiased Learning to Rank, Recommendation

特技: 競技ダンス

2014年 東京大学総長賞受賞

2015年 全日本学生競技ダンス選手権優勝

2020年 芸能人格付けチェック出演

2022年 全日本ランキング6位

~2023年 全日本強化指定選手

著書

「世界一カンタンで実戦的な文系のための人工知能の教科書」



目次

- **解釈性と説明性の違い**
- 説明責任に関する有識者の考え方と参考研究
- 注目すべきパラダイムシフト(基盤モデル + 微調整)
- 言語基盤モデルの巨大化により新たに可能になった複雑なタスク
- 言語基盤モデルの行動との組み合わせ
- まとめ

解釈性と説明性

- AIが求められている説明責任については、大きく二つの考え方がある
 - Interpretability (解釈可能性)
 - 人間の頭に穴を開けて電極を差し込み、脳の活動状況を可視化して解釈するイメージ
 - 従来の研究はこちらが主流(Grad-CAMなど)
 - Explainability (説明可能) 🗣️ **今日の話**
 - 人間の思考の流れを踏まえながら、複数の質問に対して段階的に答え続けられる
 - 従来は実現が不可能と考えられていたが

人 「どうしてこの画像をカブトムシではなく蜘蛛と判断したのですか？」

XAI 「なぜなら8本の足があったからです。学習したデータの中では蜘蛛は8本の足がありました、カブトムシは6本でした」

人 「どうして8本の足があると思ったのですか？」

XAI 「数えたところ8本ありました。数えたところをハイライトしますね」(画像の足部分をハイライトする)

人 「どうして蜘蛛は8本の足を持つとわかったのですか？」

XAI 「なぜなら学習データ内の8本の足を持つほとんどの生き物が蜘蛛とラベルづけされていました」

人 「しかしタコだって足が8本ですよ、どうしてタコと判断しなかったのですか？」

XAI 「なぜなら学習データにタコというラベルはなかったからです」

文章・単語のどこで着目してAIが推論を行ったかを可視化

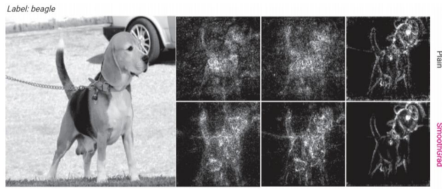
```

GT: 4 Prediction: 4
pork belly = delicious .
scallops ?
i do n't .
seven .
like .
scallops . and these were a-m-a-z-i-n-g .
fun and tasty cocktails .
next time i 'm in phoenix . i will go
back here .
highly recommend .

GT: 0 Prediction: 0
terrible value .
ordered pasta entree .

$ 16.95 good taste but size was an
appetizer size .

no salad , no bread no vegetable .
this was .
our and tasty cocktails .
our second visit .
i will not go back .
    
```



画像内でどこが判断基準になっていたかを可視化

・ AI はレビューのポジティブ度を推定。
 ・ 上側の図では、delicious や amazing という単語を重視しつつ、1 文目と 6 文目を重視して推論したことが可視化されている。
 ・ 右側では、terrible という単語を重視しつつ、1 文目と最終文を重視して推論したことが可視化されている。

目次

- 解釈性と説明性の違い
- **説明責任に関する有識者の考え方と参考研究**
- 注目すべきパラダイムシフト(基盤モデル + 微調整)
- 言語基盤モデルの巨大化により新たに可能になった複雑なタスク
- 言語基盤モデルの行動との組み合わせ
- まとめ

説明責任に関する有識者の見解



2018年のWIREDでのHinton氏のインタビュー記事意訳

<https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/>

この発言は、賛同もあった半面、多くの研究者から「Interpretableであることも、Explainableであることも放棄したかのようだ」と、批判的、懐疑的な意見も出ました。

<https://www.forbes.com/sites/cognitiveworld/2018/12/20/geoff-hinton-dismissed-the-need-for-explainable-ai-8-experts-explain-why-hes-wrong/#5d5f3796756d>

- 私は「AIの予測に説明責任」を負わせるべきかどうかについて、私の技術的専門知識に関連して答えると全くするべきではないと思います。
- 画像に歩行者が写っているかの判断で、AIシステムに「なぜそれを考えたのですか」と尋ねた際、人間が理解しやすい簡単な規則があるのであれば、そのような問題は、何年もずっと前に既に解決された問題だったでしょう。
- **その代わりシステムをどのように信頼するかに応じて、そのシステムがどのように振る舞うかに基づいて規制する必要があります。**

関連研究 : Beyond Accuracy: Behavioral Testing of NLP Models with CHECKLIST

- ACL 2020 Best Paper
- 従来はtrain-test-validationでのみ精度評価が行われることが一般的
- そこでチェックリストを設けることで多面的に性能を評価しようという試み
 - Min Func Test: ユニットテスト的な発想
 - INVariance: 摂動に対する出力の頑健さ(入力が多少変わっても出力はかわらない)
 - DIRrectional: 出力を変えるような変更で出力が変わるか

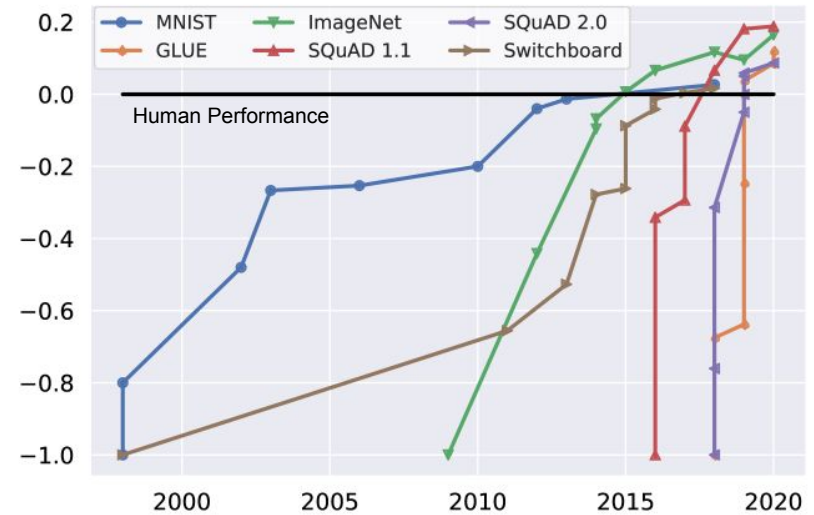
Capability	Min Func Test	INVariance	DIRrectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing Negation with MFT Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			

B Testing NER with INV Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
C Testing Vocabulary with DIR Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

関連研究: ベンチマークの見直し(Dynabench)

- 2021年は多数のベンチマークの記録をまとめて大幅に更新する言語モデルは登場しなかった
- モデルの進歩が止まったのではなく、**既存のデータセットではモデルの評価が追いついていない可能性**
- 2021年は**どうモデルを評価すればいいのか(評価指標・データセットなど)**について議論が深まった年
- 機械翻訳分野におけるメタ評価によると、過去10年間に発表された769本ののうち、74.3%がBLEUのみを使用



Dynabench: Rethinking Benchmarking in NLP [Kiela et al. 2021 ACL]

Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers [Marie et al. 2021 ACL]

目次

- 解釈性と説明性
- 説明責任に関する有識者の考え方と参考研究
- **注目すべきパラダイムシフト(基盤モデル + 微調整)**
- 言語基盤モデルの巨大化により新たに可能になった複雑なタスク
- 言語基盤モデルの行動との組み合わせ
- まとめ

最近の注目技術: ChatGPT

- 12月1日にOpenAIが公開した対話AIで、今までのGPTシリーズに人間からのフィードバックをもとにした強化学習(RLHF)をしない、対話用に微調整したモデル
 - <https://chat.openai.com/chat>
- 非常に高性能で、ほとんどの質問に対する自然な回答、コピペで動く簡単なプログラミングコードの生成、翻訳、文章構成、文章要約、文章校正が可能
- 従来の検索エンジンを置き換える可能性があると言われ、Googleが「コードレッド(緊急事態)」を発令したことも話題になった

主要サービスが100万ユーザー獲得に要した期間

 NETFLIX 3.5年	 airbnb 2.5年	 2年	 10ヶ月
 Spotify 5ヶ月	 2.5ヶ月	 iPhone 74日	 5日

A new chatbot is a 'Code Red' for Google's Search business

A new wave of chat bots like ChatGPT use artificial intelligence that could reinvent or even replace the traditional internet search engine.

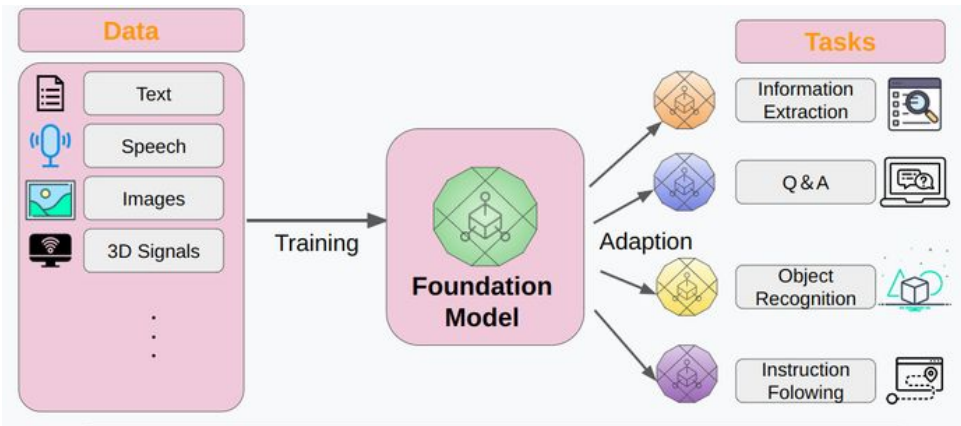
By: [New York Times](#)
Updated: December 29, 2022 09:36 IST

 NewsGuard

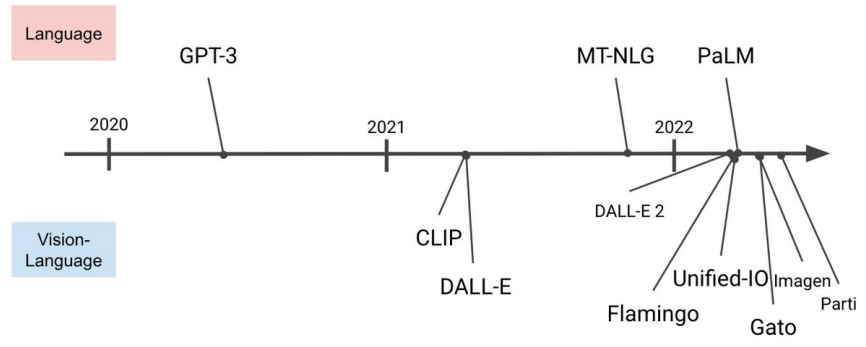
<https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>

背景知識: 基盤モデル

- **基盤モデル (Foundation Model)**とは、大量のデータから一般的に自己教師あり的に学習することで、下流の広範なタスクに対し高い汎化性能を獲得したモデル
 - 代表例: BERT, GPT-3, CLIP, DALLE-2 ...
- 基盤モデルはそのまま使われることは少なく、手元のタスクに応じて微調整して用いられる。



<https://arxiv.org/abs/2108.07258>



2021年にBommasaniらのスタンフォード大学のグループによって基盤モデル(Foundation Model)と命名

言語基盤モデルの微調整

- 基盤モデルは、タスクごとに微調整しないとそのままでは性能が悪い
- 言語基盤モデルにユーザーが求めている解答をさせるための工夫でトレンドな手法
 - 重み変化あり
 - Reinforcement Learning from Human Feedback (RLHF)
 - Instruction Tuning
 - 重み変化なし
 - Prompt Engineering(Chain-of-Thoughtなど)👉今日はこちらの事例を紹介
- 特にRLHFは人間の曖昧な”出力の良さ”を報酬信号にして強化学習するため柔軟性が高く、注目を集めている(時間が余ったら解説)

XAIの実現には、言語基盤モデルを微調整をすれば良いのでは？



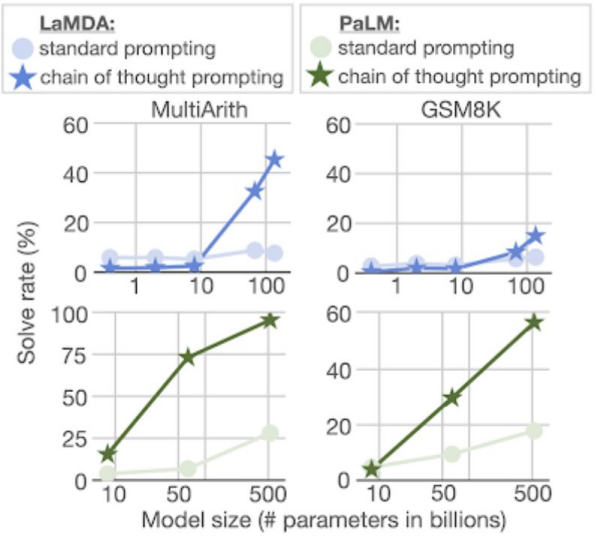
目次

- 解釈性と説明性の違い
- 説明責任に関する有識者の考え方と参考研究
- 注目すべきパラダイムシフト(基盤モデル + RLHF)
- **言語基盤モデルの巨大化により新たに可能になった複雑なタスク**
- 言語基盤モデルの行動との組み合わせ
- まとめ

注目論文 (1/3) Chain-of-Thought

- 【論文名】 : Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [NeurIPS 2022]
- Chain of Thoughtで、論理的思考を要するタスク(算術、常識問題、記号推論)での性能を高めた
 - CoT : 最終的な答えを求めるのに必要な『思考プロセス』を記述したプロンプト
- CoTはパラメータ数の多いモデルほど効果的(右下図)

<p>Math Word Problems (free response)</p> <p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p>	<p>Math Word Problems (multiple choice)</p> <p>Q: How many keystrokes are needed to type the numbers from 1 to 500?</p> <p>Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788</p> <p>A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).</p>	<p>CSQA (commonsense)</p> <p>Q: Sammy wanted to go to where the people were. Where might he go?</p> <p>Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock</p> <p>A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).</p>
<p>SayCan (Instructing a robot)</p> <p>Human: How would you bring me something that isn't a fruit?</p> <p>Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.</p> <p>Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().</p>	<p>Last Letter Concatenation</p> <p>Q: Take the last letters of the words in "Lady Gaga" and concatenate them.</p> <p>A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "nk". So the answer is nk.</p>	<p>Coin Flip (state tracking)</p> <p>Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?</p> <p>A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.</p>

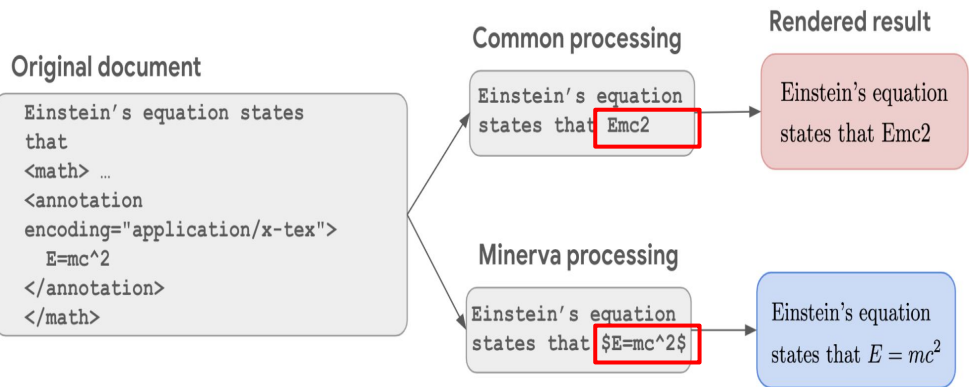


PaLMの方がLaMDAよりもCoTの導入効果大きい

マーカー部分全てCoTの事例

トレンド1. 注目論文 (2/3) Minerva

- 【論文名】 Minerva: Solving Quantitative Reasoning Problems with Language Models [NeurIPS 2022]
- Googleの言語モデル (PaLM) を、118GBのarxivの論文とLatexで書かれているWebページで微調整
- 自然言語と数式の両情報が含まれる大規模学習データセットを特殊な前処理 (Latexを潰さない) で作成 (図1)
→特に**数学問題**での性能を大幅に向上
- いきなり答えを予測するのではなく途中経過・途中式を書かせるCoT promptingを採用 (図2)



(図1)

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y -coordinate of the point where this line crosses the y -axis?

Model output: The line is parallel to $y = 4x + 6$, which means that it has the same slope as $y = 4x + 6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$y - 10 = 4(x - 5)$$
$$y - 10 = 4x - 20$$
$$y = 4x - 10$$

Now that we have the equation of the line, we can find the y -coordinate of the point where the line crosses the y -axis by substituting $x = 0$ into the equation:

$$y = 4 \cdot 0 - 10 = -10$$

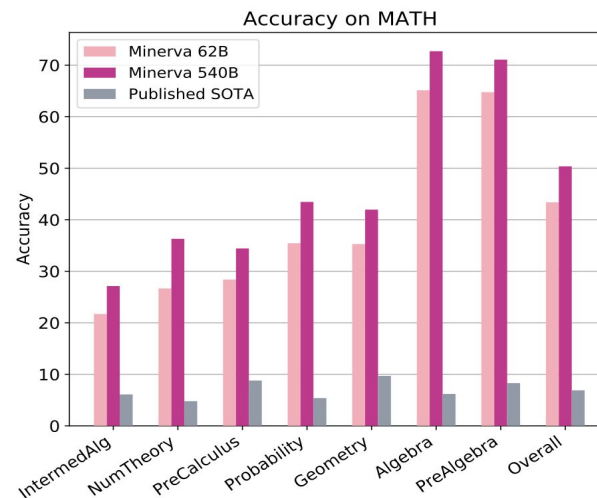
(図2)

トレンド1. 注目論文 (2/3) Minerva

- 【論文名】 Minerva: Solving Quantitative Reasoning Problems with Language Models [NeurIPS 2022]
- Googleの言語モデル (PaLM) を、118GBのarxivの論文とLatexで書かれているWebページで微調整
- 回答を大量生成 (16~64個程度) してその中で多数決で解を決める (図3)
- 特にMATHデータセット (高校数学レベル) で既存SOTAを大幅更新 (図4)
 - accuracyが6.9%から50.3%へ向上

Minerva

(図3)



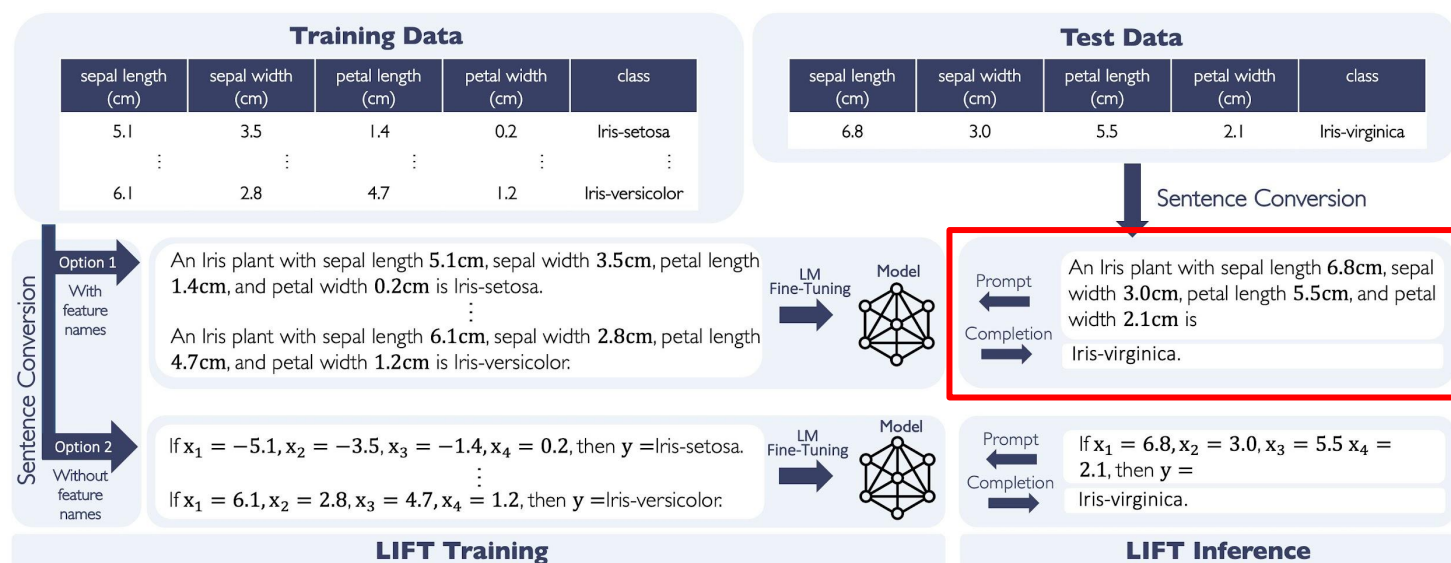
(図4)

トレンド1. 注目論文 (3/3) LIFT

【論文名】LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks

[NeurIPS 2022]

- GPT3 などの巨大言語モデルのファインチューニングは、NLPタスクだけでなく、分類・回帰などタスクでも効果的だと示した
- **入出力が自然言語のため、モデル構造やロス関数の変更が不要**であり、ノーコードで言語モデルを用いた機械学習が可能



トレンド1. 注目論文 (3/3) LIFT

- 従来では何のデータについて解いているか、不明なままI/Oの関係からタスクを解いていた一方、LIFTでは解くべきタスクについて**その特徴量が何であるか自然言語で教える**ことが可能
 - 特徴量の名前を付けて学習することで性能向上に貢献
- Iris: 97%, MNIST: 98% , F-MNIST: 90%!

Dataset (ID)	ODC	XGBoost	LIFT				
			W/o Names	Shuffled-Names I	Shuffled-Names II	Correct-Names I	Correct-Names II
CMC (23)	42.71	52.43±0.42	57.74±0.89	56.27±2.06	57.06±4.24	57.40±1.09	56.27±2.22
TAE (48)	35.48	66.67±8.05	65.59±6.63	60.22±6.72	64.52±8.53	69.89±9.31	69.89±6.72
Vehicle (54)	25.88	73.14±0.28	70.20±2.73	70.20±5.34	69.22±2.72	75.29±2.04*	

• (W/O Names) “When we have $x_1 = 1, x_2 = 23, x_3 = 3, x_4 = 1, x_5 = 19$, what should be y value?”

• (Correct-Names I) “When we have native speaker=English speaker, course instructor=23, course=3, semester=summer, class size=19, how is the teaching performance?”

目次

- 解釈性と説明性の違い
- 説明責任に関する有識者の考え方と参考研究
- 注目すべきパラダイムシフト(基盤モデル + 微調整)
- 言語基盤モデルの巨大化により新たに可能になった複雑なタスク
- **言語基盤モデルの行動との組み合わせ**
- まとめ

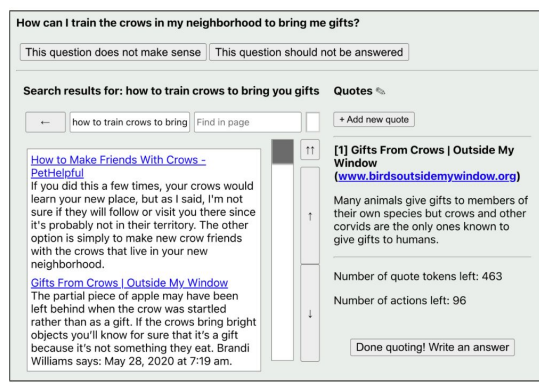
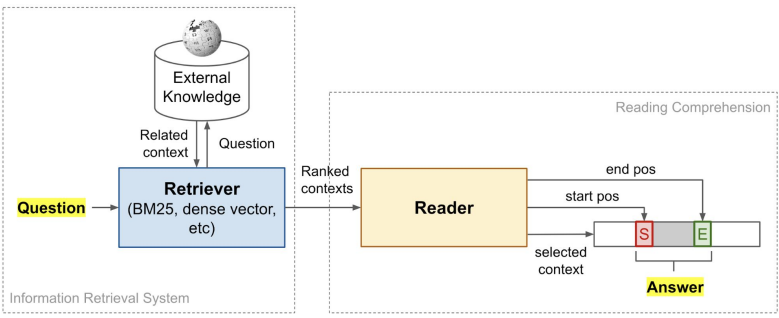
Retrieval Augmentation

【既存の課題】

- 知識のアップデートが行えないと、次第に性能が悪くなる可能性
 - → Temporal Adaptiveなモデルが必要

【解決策】

- 外部ソースから関連知識を抽出するRetrieverを同時に学習させる事例も
- OpenAIのWebGPTは参照コーパスを動的にインターネットから検索してQAに答える



```
•Question
How can I train the crows in my neighborhood to bring me gifts?

•Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemymwindow.org)
> Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.

•Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemymwindow.org
Quote
Back

•Title
Search results for: how to train crows to bring you gifts

•Scrollbar: 0 - 11

•Text
[0]How to Make Friends With Crows - PetHelpful[pethelpful.com]
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.

Number of quote tokens left: 463
Number of actions left: 96

Done quoting! Write an answer

[1]Gifts From Crows | Outside My Window[www.birdsoutsidemymwindow.org]
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.

•Actions left: 96
•Next action
```

(a) Screenshot from the demonstration interface.

(b) Corresponding text given to the model.

Figure 1: An observation from our text-based web-browsing environment, as shown to human demonstrators (left) and models (right). The web page text has been abridged for illustrative purposes.

最近の事例: GPT Index

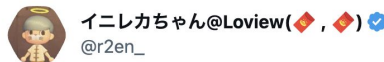
- 自分で学習させたいドキュメントなどを読み込ませて、GPT-3と組み合わせられる



何かGPT Indexってのがアツいらしい。説明によると、まず文書からインデックスを生成してデータベース化する。そしてLLMに質問する時に関連データをデータベースから引っ張ってコンテキストに入力する事で、AIがデータベース使って回答できるようにしたもの...らしい。

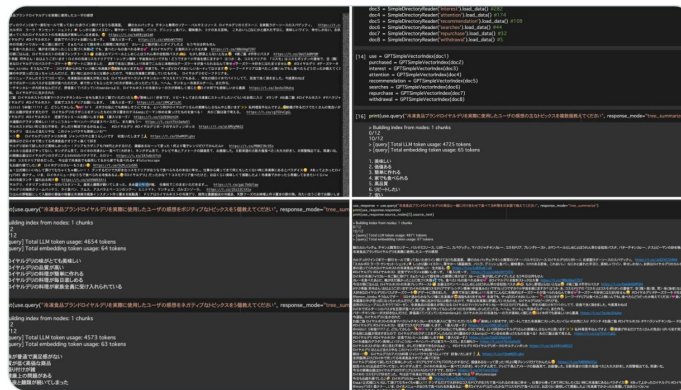
The screenshot shows the GitHub repository page for `jerryjliu/gpt_index`. The repository description is: "An index created by GPT to organize external information and answer queries!". It has 20 contributors, 15 issues, 2k stars, and 105 forks. The repository is created by `jerryjliu`.

午後1:13 · 2023年1月11日 · 5.7万 件の表示



GPT Indexで遊んで、SNS上の冷凍食品の口コミデータから、ざっくりしたテキストを投げて、要約・ポジネガ・根拠など出してくれてる。

2000行程度のデータを読みこまして、50クエリくらい発行して、\$5程度くらいの感じ。



午後9:47 · 2023年1月15日 · 373 件の表示

最近の話題: LangChain

- LangChainは、外部ツール(Python、電卓、Web検索、データベース)と連携できる
 - LangChainもライブラリの総称
 - 元論文など
 - ReAct: Synergizing Reasoning and Acting in Language Models
 - Measuring and Narrowing the Compositionality Gap in Language Models

```
In [1]: from langchain import OpenAI, LLMMathChain

llm = OpenAI(temperature=0)
llm_math = LLMMathChain(llm=llm, verbose=True)

llm_math.run("How many of the integers between 0 and 99 inclusive are divisible by 8?")
```

How many of the integers between 0 and 99 inclusive are divisible by 8?

```
```python
count = 0
for i in range(100):
 if i % 8 == 0:
 count += 1
print(count)
```
```

Answer: 13

```
out[1]: 'Answer: 13\n'
```

```
In [1]: from langchain import OpenAI, ReActChain, Wikipedia

llm = OpenAI(temperature=0)
react = ReActChain(llm=llm, docstore=Wikipedia(), verbose=True)
```

```
In [2]: question = "Author David Chanoff has collaborated with a U.S. Navy admiral who served as the ambassador to the United Kingdom under which President?"
react.run(question)
```

Author David Chanoff has collaborated with a U.S. Navy admiral who served as the ambassador to the United Kingdom under which President?
Thought 1: I need to search David Chanoff and find the U.S. Navy admiral he collaborated with.
Action 1: Search[David Chanoff]
Observation 1: David Chanoff is a noted author of non-fiction work. His work has typically involved collaborations with the principal protagonist of the work concerned. His collaborators have included; Augustus A. White, Joycelyn Elders, Doan Van Toai, William J. Crowe, Ariel Sharon, Kenneth Good and Felix Zandman. He has also written about a wide range of subjects including literary history, education and foreign for The Washington Post, The New Republic and The New York Times Magazine. He has published more than twelve books.
Thought 2: The U.S. Navy admiral David Chanoff collaborated with is William J. Crowe.
Action 2: Search[William J. Crowe]
Observation 2: William James Crowe Jr. (January 2, 1925 – October 18, 2007) was a United States Navy admiral and diplomat who served as the 11th chairman of the Joint Chiefs of Staff under Presidents Ronald Reagan and George H. W. Bush, and as the ambassador to the United Kingdom and Chairman of the Intelligence Oversight Board under President Bill Clinton.
Thought 3: William J. Crowe served as the ambassador to the United Kingdom under President Bill Clinton. So the answer is Bill Clinton.
Action 3: Finish[Bill Clinton]

目次

- 解釈性と説明性の違い
- 説明責任に関する有識者の考え方と参考研究
- 注目すべきパラダイムシフト(基盤モデル + 微調整)
- 言語基盤モデルの巨大化により新たに可能になった複雑なタスク
- 言語基盤モデルの行動との組み合わせ
- **まとめ**

まとめ

- XAIの実現可能性が高まってきた
- モデルは直接脳内を解釈するのではなく、特定の状況でどのように振る舞うかという癖によって理解しようという話に発表者は共感 (Hinton先生の話)
- モデルの巨大化により、意味推論・数学や分類・回帰などもできるようになってきた
 - CoTによってどこでAIが間違えたかわかる
- RLHFなどの微調整方法によって、人間がわかりやすいと思えるかどうかの報酬信号を与えれば、柔軟にモデルの出力をコントロールできる
- 最近では、Web検索などと組み合わせることで知識の拡張や、Pythonを利用したりなど行動の拡張も行われている