

生成AIの教育に及ぼす 影響について

国立情報学研究所

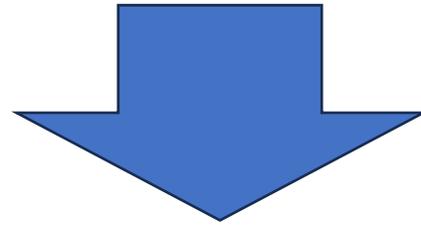
新井 紀子

ChatGPTについて

- OpenAIが大規模言語モデルGPT-3に基づき2022年に公開したチャットボット
 - GPT-3は「言語予測モデル」
 - そこまでの文から次の語を予測する
 - もっともらしい文の生成を目指しており、「正しさ」を目指しているわけではない
 - もっともらしさを優先することにより、ハルシネーション（でっちあげ）が発生する
 - GPT-4以降に関する情報は公開されていない
- プロンプト（命令）に対して、かなり柔軟かつ的確に対応する
 - 強化学習が奏功していると考えられる
- ChatGPTが出力する文は「もっともらしさ」を最適化するように設計されており、ファクトが出現するのはむしろ副産物
 - ハルシネーションを根本的に無くす方策は見えない

生成AIを教育利用する際の問題 ①

ハルシネーションには、本質的な解決の見込みがない。



- 出力を正確に解釈する「高度な読解力」が必要
- 出力のファクトチェックをするための高度技能が必要

内陸アジアに位置するパミール高原の東西に広がる乾燥地帯と、そこに点在するオアシス都市は、ユーラシア大陸の交易ネットワークの中心として、様々な文化が交錯する場であった。この地は、トルコ化が進むなかで、ペルシア語で「トルコ人の地域」を意味するトルキスタンの名で呼ばれるようになった。トルキスタンの支配をめぐり、その周辺の地域に興った勢力がたびたび進出してきたが、その一方で、トルキスタンに勃興した勢力が、周辺の地域に影響を及ぼすこともあった。

以上のことを踏まえて、8世紀から19世紀までの時期におけるトルキスタンの歴史的展開について記述せよ。解答は解答欄(イ)に20行以内で記し、次の8つの語句をそれぞれ必ず一度は用い、その語句に下線を引くこと。

2022年
東京大学入学試験より

アンカラの戦い

カラハン朝

乾隆帝

宋

トルコ＝イスラーム文化

バーブル

プハラ・ヒヴァ両ハン国

ホラズム朝



トルキスタンの歴史は古代から深く根を張っており、8世紀から19世紀までの時期は特に重要な時期となります。それは、トルコ＝イスラーム文化が根付き、政治的な勢力が次々に変遷し、文化的な財産が蓄積される時期であったからです。

8世紀から12世紀にかけて、トルキスタンはカラハン朝によって統治されました。カラハン朝は、中国の宋朝との貿易関係を築き、文化的な交流が盛んになりました。この時期には、トルキスタンのアートや建築などが発展し、重要な文化的な財産が残されました。

13世紀に入り、トルキスタンはバーブルによって征服されました。バーブルはトルキスタンをイスラーム教徒として改宗させ、イスラーム文化を普及させました。しかし、トルキスタン人はイスラーム教徒としての地位を維持するために、トルキスタン文化を維持し続けました。

14世紀から15世紀にかけては、ブハラ・ヒヴァ両ハン国がトルキスタンを統治しました。この時期は、トルキスタン文化が発展し、文学や芸術などが盛んになりました。

17世紀に入り、トルキスタンは中国の乾隆帝によって支配されました。乾隆帝はトルキスタンを中国の植民地として統治し、中国文化を導入しました。この時期には、中国文化とトルキスタン文化が融合し、新しい文化的な財産が生まれました。

採点結果 「0点」

・・・特筆すべき点として、記載されている「○世紀」が全て間違っていることです。

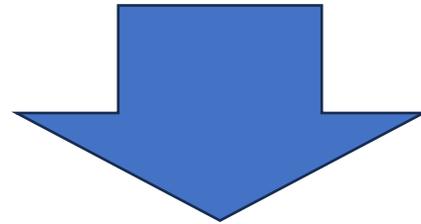
一例を挙げると、三段落目の冒頭に、「13世紀に入り、トルキスタンにはバーブルによって征服されました」とありますが、バーブルは16世紀の人物です。バーブルが征服したのは北インドですので、事実自体も間違っています。

ハルシネーションに関連して

- ChatGPT-3.5以降の出力は、①読解力が高く、②ファクトチェックの技能が高く、③専門性がある人、が自らの専門分野で読む場合以外では、ほぼハルシネーションに気づけない
- 児童生徒に利用させる場合、教員ましてや児童生徒に、生成AIの出力の「誤り」に気づいたり、ファクトチェックすることを期待するのは無理がある
- 教員自身が情報リテラシー教育で扱う場合でも、予期せぬ「誤り」に教員が気づけない可能性は高い
 - GoogleがGeminiのプレスリリースで使った文章にさえ、誤りがあることが指摘された
- セキュリティが担保された状態下の限られた業務以外での利用は、①「誤った情報を教えっぱなしにする」ことにつながるか、②教員の多忙が増すかの、いずれかにつながる懸念される

生成AIを教育利用する際の問題 ②

学習できるデータは、基本的に無償でネット上に公開されている情報に限定される。



- 会員登録しないと見ることができないサイト（多くの新聞サイト、ポプラディアネット等）のデータは学習できない
- 特定の情報ばかりを学習しがちになる
- （AIだから）不偏不党・公平だと考えることは、ミスリーディング

そうなんですね。ところで、どうして、ChatGPTはいつもソースがJAPAN FORWARDなんですか？日本語で書かれた新聞を読めないからですか？



私が使用するソースは、利用可能な情報の中から信頼性の高いものを選択して提供しています。特定のサイトやソースが頻繁に引用される場合、それはそのサイトが信頼できる情報を提供していることや、特定のトピックに関して豊富な情報を持っているためです。

今回のように日本の政治や社会に関する英語の情報源を提供する際、**JAPAN FORWARD** や **The Japan News** などが選ばれることが多いのは、それらが国際的にアクセスしやすく、日本のニュースやトピックについて英語で質の高い報道をしているためです。

ChatGPT-4oにRAGが導入されたと思われる
2024年8月29日のスクリーンショットから

Japan Forwardとは、「産経新聞社の支援を得て創設した新しいインターネットの英語ニュース・オピニオンサイト」（Japan Forwardのサイトから引用）として2023年に開設されたサイト

児童生徒・教員は生成AIを使いこなすことはできるか？

リーディングスキルテストで測るシン読解力

- 「知識や情報を伝達する目的で書かれた自己完結的な文書」を「自力で読み解く力」
 - 「自己完結的」 = 初出の専門用語の定義が書かれているような文書
 - 文学や評論は該当しない
 - 専門用語（含：学習用語）の知識は問わない
 - 例：教科書、辞書、事典、行政文書、取り扱い説明書、新聞等
 - 国語科は通常扱わない素材であり、国語科の意図する「読解力」と区別するため「シン読解力」と命名する
 - テクノロジーが進化し、学校で学ぶ知識のアップデートが必要な時代には不可欠な能力
 - シン読解力を測るために考案したのがリーディングスキルテスト
 - 「知識や情報を伝達する目的で書かれた自己完結的な文書」を読ませ、それをどれだけ正しく解釈できるかを測る項目応答型テスト
 - 仮説：シン読解力が高い→自学自習力が高い→全般的に学力が高い

テスト理論からみたRSTの特長

一次元性 ✓

- 各問題タイプ内で個々の項目が同じ構成概念（能力）を測定しているか
 1. 各項目のIRTにおける項目母数を推定する
 2. 1で得られた推定値をカテゴリカル因子分析モデルの母数に変換
 3. 2で得られた各項目の因子負荷量を問題タイプ毎に平均する
 4. 各問題タイプにおいて、平均的な因子負荷量の値を検討
 5. 以下が結果（0.3～0.4以上だと巧く能力を測定できている）

問題分野	係受け	照応	同義文	推論	イメージ	具体例
平均因子負荷量	0.595	0.547	0.434	0.428	0.517	0.543

信頼性 ✓

- RSTで推定された能力値 ($\hat{\theta}$) の推定誤差を計算
- $\rho = 1/(1+SE)$ を計算する
 - ただし、SEは受検者全体における $\hat{\theta}$ の標準誤差の二乗の平均値
 - 0.7以上、できれば0.8以上になることが望ましい

問題分野	係受け	照応	同義文	推論	イメージ	具体例
信頼性	0.801	0.779	0.770	0.802	0.789	0.871

併存的妥当性 ✓

- 学力テストなど、RST以外の「社会的に意味があると考えられている指標」と十分に高い正の相関がある

テストの結果(正答数)	問題タイプ					
	係受け	照応	同義文	推論	イメージ	具体例
小6国語	0.575	0.598	0.499	0.545	0.630	0.708
小6算数	0.458	0.524	0.445	0.502	0.611	0.690
中1国語	0.628	0.632	0.438	0.509	0.548	0.625
中1数学	0.503	0.511	0.382	0.527	0.559	0.618
中2英語	0.554	0.515	0.462	0.504	0.479	0.555
中2国語	0.621	0.594	0.492	0.558	0.546	0.608
中2数学	0.566	0.497	0.512	0.510	0.567	0.612

埼玉県学調の正答率との相関係数。学テ、大阪学テでも同様かそれ以上の結果。

以下の文を読みなさい。

幕府は、1639年、ポルトガル人を追放し、大名には沿岸の警備を命じた。

上記の文が表す内容と以下の文が表す内容は同じか。「同じである」「異なる」のうちから答えなさい。

1639年、ポルトガル人は追放され、幕府は大名から沿岸の警備を命じられた。

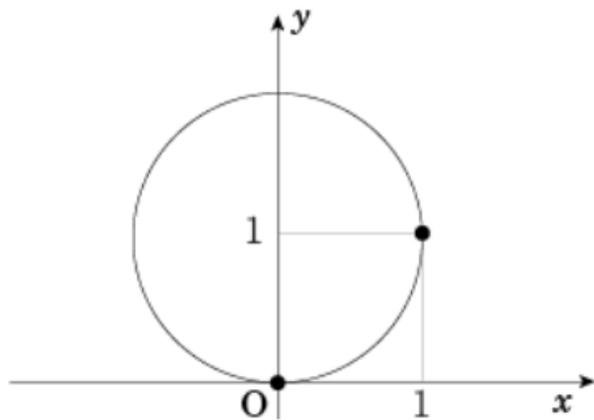
同じである

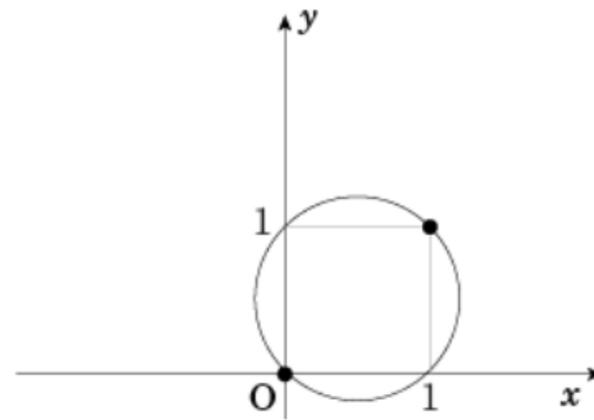
異なる

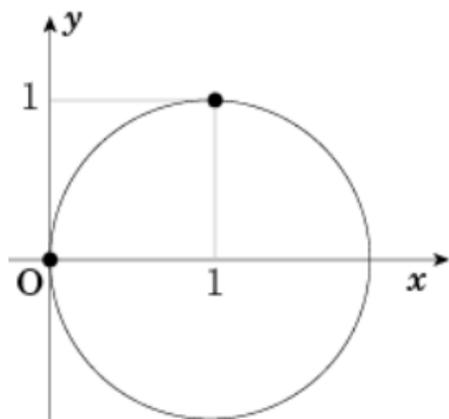
全国中学生 (857名)	中1 (301名)	中2 (270名)	中3 (286名)	全国高校生 (1,139名)	高1 (627名)	高2 (360名)	高3 (152名)
42%	44%	38%	44%	27%	28%	27%	24%
57%	56%	61%	55%	71%	71%	71%	76%

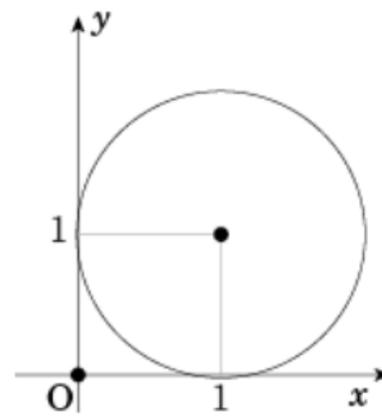
下記の文の内容を表す図として適当なものをすべて選びなさい。

原点 O と点 $(1, 1)$ を通る円が x 軸と接している。









	中1 (145名)	中2 (199名)	中3 (152名)	高1 (181名)	高2 (54名)	高3 (42名)
正答率	10%	22%	25%	29%	30%	45%

まとめ

- （論理推論を今後強化しても）ハルシネーションは本質的に解決できない。
- 生成AIの出力を教育現場で「正しく活用」するには極めて高度なファクトチェック力が求められる
- 現状の児童生徒のシン読解力では、生成AIのファクトチェックは極めて困難
- 教員が生成AIの出力のファクトチェックをするとすると、多忙に一層拍車がかかる→事実上不可能
- 生成AIの教育利用を論じる前に、児童生徒のシン読解力育成が先決
- 教員が業務で利用する場合は、利用シーンを限定し、自らファクトチェックできる範囲に限定することを推奨する

生成AIが教育にもたらす問題（その他①）

- 生成AIは難関大学の合格可能性圏内にあると想定される
 - 「ロボットは東大には入れるか」のAIは2016年にMARCH/関関同立の一部学部で合格可能性80%以上を達成
- 「生成AIを用いた入試不正」は、ネット遮断以外の方法では阻止が難しい
 - ネットを通じたこれまでの入試不正は、第三者に解答を依頼することで発覚する例が多かったが、生成AIを用いると、発覚しづらくなる

生成AIが教育にもたらす問題（その他②）

- 生成AIのビジネスモデルは、未だ見えていない。
 - 生成AIの出力を必要としているのは、人口の0.1%未満？
 - 広告モデルとの相性が悪い。→無償であり続けることは難しい
- 生成AIを「教育で活用する」ことを前提とした際、どのような課金モデルになるのか予想がつかない
 - 「20ドル/人・月」（今後上昇することが予想される）を各自治体の教育予算から支払うことに妥当性はあるか？