

チャット GPT による数学の単元ごとの正答率の調査と比較

大庭 百恵[†] 青山 大地[‡] 白木 厚司[‡]

東京学芸大学附属国際中等教育学校[†] 千葉大学大学院工学研究院[‡]

1. はじめに

文部科学省が公開している令和5年度学校基本統計結果の概要[1]によると、高等学校等の卒業者の大学等への進学率は過去最高を記録している。一方で、文部科学省が令和4年度に実施した児童・生徒の長期欠席（不登校等）に関する調査[2]によると、小・中学校、高等学校の長期欠席者の割合は増加傾向にあり、学校以外で学習できる環境が求められるようになりつつあると推測できる。

また、2022年11月に ChatGPT(Generative Pre-trained Transformer)が公開されたことをきっかけに、利用者の問いかけに対し AI(Artificial Intelligence)が回答してくれる生成 AI が話題となっている。ChatGPT を始めとする生成 AI が教育の分野で活用できる性能を持っていれば、不登校児童・生徒の学習の支援に役立つ。そこで本研究では、大学入学共通テストの数学を対象とし、ChatGPT がどの程度の正答率を得られるのかを検証する。また、問題を単元ごとに分けて比較・分析することで、ChatGPT の学習支援活用方法について検討する。

2. 関連研究

AI に入試問題を解かせる研究として、新井紀子教授らの「ロボットは東大に入れるか」プロジェクト[3]が挙げられる。このプロジェクトは自然言語処理や機械学習、確率・統計などの技術により開発されたロボットに模擬試験の問題を解かせ、ロボットの限界を再認識するというプロジェクトである。ただし、このロボットは一般の方が自由に利用することはできず、登録すれば誰でも利用できる Chat GPT を学習支援に活用する本研究とは目的が異なる。

3. ChatGPT

ChatGPT は 2022 年 11 月 30 日、AI 研究機関である「OpenAI」が公表した文章生成 AI である。情報収集に使える便利なツールだが、いくつか

問題点がある。例えば、必ず回答が正確であるとは限らず、中には不正解な情報が含まれている可能性がある。さらに、無料版の ChatGPT は 2021 年 9 月までの情報を基に開発されているため、最新の情報は含まれていない。そして機密情報が流出する可能性もある。このような ChatGPT の弱点の影響で、私の学校では ChatGPT で情報収集する際にはフェイクニュースの可能性に気を付けることを指示されている。

4. 研究手法

4.1 研究手順

本研究では 2021 年から 2023 年までに実施された大学入学共通テストの本試験のうち、数学 I・A と数学 II・B の問題を ChatGPT へ質問し、正しい解答を得られるかを検証する。しかし、ChatGPT は図形を認識できないため、図形が問題文に含まれる問題は省く。

手順の流れは次の通りである。

1. 本試験の数学の問題の問題文を ChatGPT の無料版、バージョン 3.5 に入力する
2. 再生成というボタンを 2 回押し、合計 3 個の解答を出力する
3. 3 回中 3 回正解だった場合、ChatGPT は正答したと見なす

なお、以下に示す条件の下で検証する。

- 数値の解答がなかった場合、不正解と見なす
- ②, ①3, ②4 のように解答群があった場合、ChatGPT の解答に混乱が生じるため②, ①, ②の番号は省略する。

4.2 分析手法

代表的な分析手法として、t 検定、分散分析、およびカイ二乗検定が挙げられる。t 検定は平均値の差を観測する分析方法であり、二項目のみに対応する。分散分析は t 検定と同様に平均値の差を分析するが、三項目以上の比較もできる。カイ二乗検定は、2 つのカウントデータの差を分析し、二群間が独立かどうかを調べる時に用いる。t 検定と分散分析は観測値を分析するが、カイ二乗検定は度数を分析する。

本研究では ChatGPT の正答率という度数の単元ごとの差を調べたいため、二群間のカウント

Research and Comparison of the Correct Response Rate of ChatGPT in Math Problems by Units
Momoe Oba[†] Daichi Aoyama[‡] Atsushi Shiraki[‡]
Tokyo Gakugei University International Secondary School[†]
Graduate School of Engineering, Chiba University[‡]

データを調べることに適しているカイ二乗検定を選んだ。

なお、正答数の期待値が5未満の場合は Fisher の正確検定を用いた。

5. 研究結果と考察

5.1 研究結果

数学I・A および数学II・Bの単元ごとの正答率を表1, 2にそれぞれ示す。また、2つの単元間の有意差の有無を表3, 表4にそれぞれ示す。ここで表中の△は有意傾向を表している。

表1. 数学I・Aの単元ごとの正答率

	数と式	図形の性質	整数の性質	場合の数と確率	合計
2021年	0/5	N/A	N/A	N/A	0/5
2022年	1/4	0/9	1/8	1/10	3/31
2023年	2/3	0/10	4/7	2/5	8/25

表2. 数学II・Bの単元ごとの正答率

	三角関数	指数対数	微分積分	ベクトル	合計
2021年	2/8	2/2	3/5	1/8	8/23
2022年	1/8	1/6	2/9	1/10	5/33
2023年	2/7	5/5	7/9	N/A	14/21

表3. 数学I・Aの2つの単元間の有意差の有無

	数と式	図形の性質	整数の性質	場合の数と確率
数と式	N/A	△	×	×
図形の性質	△	N/A	○	×
整数の性質	×	○	N/A	×
場合の数と確率	×	×	×	N/A

表4. 数学II・Bの2つの単元間の有意差の有無

	三角関数	指数関数	微分積分	ベクトル
三角関数	N/A	×	×	×
指数関数	×	N/A	×	△
微分積分	×	×	N/A	○
ベクトル	×	△	○	N/A

5.2 考察

表1と2から、全体的に正答率が低いことが見て取れる。年度別で見ると、2021年と2022年の間で正答率が明らかに低下している。この結果について Fisher の検定を用いて分析したが有意差は見られなかった。2022年の数学の共通テストで、数学I・Aは前年を19.72点下回り数学II・Bは前年から16.87点下回ったことから難化傾向が確かめられる。難易度が高いほど ChatGPT の正答率が下がる可能性がある。

表3と表4から単元ごとに ChatGPT の正答率には差があることが言える。例えば、整数の性質と図形の性質の p は 0.047124047 であり、有意水準を 0.05 とした場合、 $p < 0.05$ となるため有意差が確認できた。さらに、微分・積分とベクトルの p 値は 0.047 であり、有意差が確認できた。

6. まとめ

研究目的は ChatGPT の数学の問題の正答率を調べることで、学習支援としての可能性を調べることである。ChatGPT に大学入学共通テストの数学の問題を答えさせ、解答結果に対してカイ二乗検定で分析を行った。整数の性質と図形の性質、微分積分とベクトルに有意差が確認できたが、全体的に正答率が低く、現時点で学習支援としては信用できないと言える。単元ごとと年度ごとに正答率の差が確認できたため、問題の難易度によって正答率は変わるかもしれない。

謝辞

本研究は、国立研究開発法人科学技術振興機構 (JST) の支援を受けて実施するグローバルサイエンスキャンパス事業 (GSC) である千葉大学 ASCENT Program において実施されました。

参考文献

- [1] 文部科学省, “令和5年度学校基本統計結果の概要,” 2023, https://www.mext.go.jp/content/20230823-mxt_chousa01-000031377_001.pdf, (最終閲覧日: 2023年12月20日).
- [2] 文部科学省, “令和4年度児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査結果,” 2023, https://www.mext.go.jp/content/20231004-mxt_jidou01-100002753_1.pdf, (最終閲覧日: 2023年12月20日).
- [3] 新井紀子, “AI VS. 教科書が読めない子どもたち,” 東洋経済新報社, 2018.