ChatGPT は半田高校に合格できるか？

Can ChatGPT pass Handa High School?

要旨　本研究では、ChatGPT が愛知県立半田高校の入試問題に対してどの程度正答できるかを検証した。実際の半田高校入試問題を用い、ChatGPT に英語・数学の問題を解かせ、得点を計算した。結果として、GPT-4o は英語で高い正答率を示した一方、数学では誤答が多く得点は低かった。一方 o4-mini は図形問題を含む数学で満点の成績を収めた。これらの結果から ChatGPT の成績には、プロンプトの調整とモデルの選択が影響することが示唆された。

Abstract　This study investigates how accurately ChatGPT can solve entrance exam questions from Aichi Prefectural Handa High School. Using actual exam questions, we evaluated ChatGPT's performance in English and mathematics by scoring its responses. The results showed that GPT-4o achieved a high accuracy rate in English but performed poorly in mathematics, with many incorrect answers. In contrast, o4-mini achieved a perfect score in mathematics, including geometry questions. These findings suggest that both prompt design and model selection significantly influence ChatGPT's performance.

1. 研究背景と研究目的・意義

1.1 研究背景

近年の大規模言語モデル（LLM)、特に OpenAI 社の GPT シリーズは、幅広い分野で高い性能を発揮している。GPT-4 は様々な学術・専門試験において人間並みの成績を達成しており、例えば模擬司法試験で上位 10%の成績を収めるなどの報告がある（OpenAI, 2023a)。また、多言語ベンチマーク（MMLU）でも従来モデルを大きく上回り、英語以外の言語でも高い性能を発揮している（OpenAI, 2023a)。こうした背景から、教育分野で AI の影響について関心が高まっており、既に ChatGPT を用いて英語読解の高校入試問題を解かせる研究などが進められている（de Winter, 2023)。日本国内では、GPT-4 を最適化プロンプトと組み合わせることで医師国家試験の合格ラインを超える成績を達成した事例も報告されている（Cheng, 2023)。

1.2 リサーチクエスチョンと先行研究・事例

本研究のリサーチクエスチョンは「ChatGPT は愛知県立半田高校の入試に合格できるか」である。この問いに対し、先行研究としてはオランダの高校入試での GPT の実績（d Winter, 2023）や、AI の数学的能力に関する課題などが報告されている（Cheng, 2023)。

例えば、ChatGPT は複雑な計算問題や文章題に誤答することが多いことが知られており（Cheng, 2023)、LLM の数学的推論能力は課題点として挙げられている。

1.3 研究の目的・意義

本研究では日本の高校入試（半田高校）の問題を対象に、ChatGPT の性能と限界を明らかにすることを目的とする。その意義は、AI の学習・教育現場への応用可能性や評価制度への影響を検証する点にある。

2. 研究方法１　入力形式による得点率の違い

2.1 研究の目的とリサーチクエスチョン・仮説との関係

研究 1 の目的は、「ChatGPT が愛知県立半田高校の入試問題にどの程度正答できるか」を明らかにすることである。この目的は、リサーチクエスチョン「ChatGPT は愛知県立半田高校の入試に合格できるか？」に直接対応している。また、本研究では「ChatGPT の得点は入力形式やプロンプトの工夫によって変化する」という仮説を立てた。これに基づき、実際の高校入試問題を用いて、入力方法・プロンプトの工夫が ChatGPT の回答精度に与える影響を検証する。

この研究方法では、得点の計算とプロンプトによる性能の変化を分析することで、AI が高校入試に対応可能かを検討し、教育分野での活用の可能性と限界について理解を深めることを目的とする

2.2 研究と分析方法

2.2.1 実施手順

1. 使用する試験問題
   - 愛知県公立高校入試問題（令和 6 年度 英語・数学）を使用した。
   - 数学では図形問題を含み、図版は画像データとして添付して提示した。
2. 使用するモデル
   - ChatGPT（GPT-4o）を対象とした。
   - 本研究ではすべて ChatGPT で利用可能な既定モデルを使用した。
3. プロンプトの条件
   - 以下の 4 通りの入力形式（プロンプト）を検証。
     1. 通常形式（プレーンテキスト）
     2. Markdown 形式での整形入力（見出しやリストを含む）
     3. 「ステップ・バイ・ステップで考えてみましょう」という指示付き入力
     4. 問題文を分割して入力

4. 繰り返し検証
   - 各プロンプト形式について、同一問題を 3 回ずつ入力し、回答の安定性を確認。
5. 採点方法
   - 各問題について、愛知県が公表している正答を基に採点。
   - 1 問ごとの正誤に基づき点数を計算し、得点率を記録。
   - 合格ラインは得点率 72%と設定した。

2.2.2 データの収集と分析

- 形式ごとの平均得点および得点分布を比較。
- 特定のプロンプトによる得点上昇の有無を統計的に検討。
- 問題ごとの正答・誤答の傾向（特に文章題や図形問題）を分析。

2.2.3 妥当性・適合性

- 入試問題は実際の高校入試で使用されたものであり、出題の難易度や構成は現実的である。
- ChatGPT はオンラインで誰でも再現可能なツールであり、使用環境も汎用的である。
- 入力形式を複数検証することで、AI 活用の具体的な指針を得ることができる。

2.2.4 限界

- 数学における図形問題など、画像を含む問題は現在の ChatGPT では読み取り精度が低く、回答が不安定となる場合がある。
- モデルのアップデートによって性能が変動するため、研究の再現性に注意が必要。

2.3 結果

本研究では、愛知県立半田高校の 2024 年度（令和 6 年度）入試問題（英語・数学）を用いて、ChatGPT（GPT-4o）に解答させた。各問題に対して 4 種類のプロンプト形式（①通常形式、②Markdown 形式、③「ステップ・バイ・ステップ」の指示、④段階的入力）を用いて 3 回ずつ試行し、得点率を計算した。その平均得点率は以下の通りである。

科目	通常形式	Markdown 形式	ステップ・バイ・ステップ	分割して入力
英語	70%	82%	82%	100%
数学	45%	59%	72%	77%

また、合格基準を得点率 72%と設定した場合、以下の通り合格ラインを超える形式が限定

された。
- 英語は「分割して入力」でのみ100%の正答率を示し、他の2形式でも合格ラインを上回った。
- 数学は「ステップ・バイ・ステップ」と「分割して入力」の形式でのみ合格ラインを超えた。

問題ごとの傾向として：
- 数学では応用問題において、「ステップ・バイ・ステップ」や「分割して入力」によって明らかに正答率が上昇した。一方、単純な計算問題ではプロンプトの影響が小さかった。
- 図形を含む問題については、画像認識がうまくいかず、正答には至らなかった。

2.4 考察

2.4.1 ChatGPTはなぜプロンプトによって得点が変わるのか？

今回の実験結果から、ChatGPTの得点率はプロンプトの工夫によって大きく変動することが明らかになった。特に数学では、通常形式と「分割して入力」では30%以上の得点差が見られた。これは、ChatGPTが複雑な問題に対して一度に全体を理解しようとすると誤りや思考の飛躍が生じやすい一方、分割して情報を与えることで論理の順序を保持しやすくなるためだと考えられる。

英語においても、「分割して入力」では文脈の誤解が減り、文法や語彙の選択がより安定・正確になったと考えられる。特に複数の選択肢を比較させるような問題では、選択肢を1つずつ検討することで正答率が向上した。

また、「ステップ・バイ・ステップで考えましょう」という指示は、ChatGPTに思考過程を明示させる点で有効であり、特に数学においては、計算の誤りや飛躍を防ぐ役割を果たした。これはCheng（2023）らの指摘とも一致し、LLMの数学的推論は明示的なステップ化によって補完可能であることが示唆される。

2.4.2 入力形式が英語・数学で異なる影響を与える理由

英語では、文章の意味を読む力や語彙知識が問われるが、ChatGPTの事前学習には英語データが大量に含まれており、言語的知識に関しては高い性能を持つ。そのため、形式の影響は限定的だった。しかし、問題の文脈や選択肢の違いを読み取るには、情報を整理した形式での提示が有効であり、Markdown形式や分割して入力することによってその精度が向上した。

一方、数学ではプロンプトの違いが極めて重要であった。これは、ChatGPT が数学的推論を「会話的に模倣する」能力に依存しており、問題の意図や計算手順の読み違いが起きやすいためである。入力が複雑すぎると誤った出力になりやすく、逆に分割して入力することによって、誤答が減少した。

2.4.3 今後の研究の問い

今回の結果から、以下の新たな問いが浮かび上がった：

- ChatGPT は、図形問題をどの程度正確に認識・処理できるか？ 特に図形問題では画像理解能力に限界が見られた。
- 複数の入力形式を組み合わせた場合（例：Markdown 形式＋段階的入力）、得点がさらに向上するのかどうか。この問いは本研究では十分に解明できなかったが、今後の研究で追究すべき重要なテーマである。

3. 研究方法 2　異なるモデルによる図形問題の正答率の違い

3.1 研究の目的とリサーチクエスチョン・仮説との関係

研究方法 2 では、ChatGPT は図形問題をどの程度正確に解答できるのか、そしてその正答率はモデル（GPT-4o と o4-mini）および画像の有無によってどのように変化するのかを明らかにすることを目的とする。

この目的は、リサーチクエスチョン「ChatGPT は半田高校に合格できるか？」のうち、特に数学の図形問題という難易度の高い課題への対応力を検証する要素にあたる。研究方法 1 で明らかになった「プロンプトの工夫によって得点が変化する」という結果に加え、研究方法 2 では「モデルの特性および図の提示の有無」という変数に着目し、さらなるパフォーマンスの違いを検証する。

仮説としては、「推論特化型モデル（o4-mini）は図形問題においてより安定して高得点を出す」こと、また「画像付きの問題提示が正答率を上げる」ことを設定した。

3.2 研究と分析方法

3.2.1 実験手順

項目	内容
使用問題	愛知県立半田高校 数学の図形問題（令和 6 年度入試より複数選定）
使用モデル	ChatGPT GPT-4o、および o4-mini（いずれも OpenAI 社製）

問題の提示形式	各問題を①画像なし（テキスト記述のみ）、②画像付き（図版添付）の 2 通りで提示
実施回数	各モデル・各条件（画像の有無）につき 7 題、各題 3 回ずつ回答を取得
環境	OpenAI ChatGPT デスクトップ版を使用
正答基準	愛知県が公表した正答例に基づき、1 問ごとに「正解：1 点」「不正解：0 点」で評価

3.2.2 データ収集と分析

- 回答中の言語的傾向（例：「～のように思われる」「～と仮定して考える」など）を分析し、誤答に至る過程を考察。
- 分析には基本的な記述統計を用い、図表で可視化。

3.3 結果

モデル	画像なし正答率	画像あり正答率	平均正答率
GPT-4o	29%	43%	36%
o4-mini	100%	100%	100%

- GPT-4o は画像提示によって正答率が若干上昇したが、依然として安定しない結果が見られた。
- o4-mini は画像の有無にかかわらず 100%の正答率を達成し、特に数学への適応力が高いことが示された。
- GPT-4o の回答では「途中で迷いが生じたような表現」や「根拠が曖昧なまま解答に至る」ケースが多かった。

3.4 考察

3.4.1 モデルによる性能差の背景

o4-mini が高い正答率を示したのは、数理的推論に特化した設計、図形構造を論理的に把握する能力が反映されたためと考えられる。これは、ChatGPT の中でもモデルによって得意分野が異なることを示す具体的な証拠である。

一方、GPT-4o では誤答が多かった。その原因として、以下の点が考えられる：

- 曖昧な図形の認識：画像を処理する能力はあるが、計算に即した形での解釈が不安定である。

- 中間的な推論エラー：途中で「仮定に基づく思考」を行うが、選択肢と合致しないまま誤答する傾向があった。

3.4.2 画像の有無による影響

画像付きでの提示は GPT-4o において正答率を改善したものの、その効果は限定的だった。これは、「画像認識」自体はできても、それを正確に数学的言語に変換して処理する力が十分でないことを示唆する。o4-mini は画像の有無にかかわらず同様の成績であり、内部的に図形構造を適切に処理する力があると考えられる。

3.4.3 今後の問い

- 図形問題において、より有効なプロンプト形式とは何か？：例えば、図形の構成要素をより詳細に言語化して分割入力することで、GPT-4o の精度が向上する可能性がある。

4. 結論と今後の展望

4.1 結論

本研究のリサーチクエスチョン「ChatGPT は愛知県立半田高校の入試に合格できるか？」に対し、実験の結果から得られた主な結論は以下の通りである。

まず、ChatGPT の得点は使用するプロンプト形式によって大きく変化することが明らかになった。特に「ステップ・バイ・ステップ」や「分割して入力」といった構造的なプロンプトは、英語・数学ともに得点の向上に有効であった。これにより、ChatGPT は適切な形式で入力すれば、合格ライン（得点率 72%）を超えることが可能であることが確認された。

また、モデル間の性能差も顕著に現れた。特に数学の図形問題において、GPT-4o は画像ありでも不安定な成績を示した一方、o4-mini はすべての条件下で 100%の正答率を記録した。この結果は、大規模言語モデルにおいても、用途や問題の性質に応じてモデルを選択することの重要性を示している。

したがって、「ChatGPT は半田高校に合格できるか？」という問いに対しては、「モデル選択とプロンプト設計を工夫すれば、十分に合格可能である」というのが本研究から導かれる結論である。ただし、すべての形式・条件で安定して合格できるわけではなく、特に図形問題の処理においてはモデルの限界やプロンプト工夫の余地が残されている。

4.2 今後の展望

今後の研究では、以下の 3 点を中心にさらなる検証を進めたい。

1. プロンプト形式の最適化と組み合わせの検証

本研究では各プロンプト形式を独立に検証したが、今後は「Markdown 形式」＋「ステップ・バイ・ステップ」＋「分割入力」など、複数の形式を組み合わせることで得点がさらに向上するかを調べる必要がある。

2. 図形問題における言語的記述の工夫

GPT-4o が図形問題に弱い原因の一つは、画像からの意味抽出が不安定である点にある。今後は、図形を言語化し、ChatGPT に「言葉で説明する」形で問題を与えることで精度が改善されるかを検証したい。

3. より高度な問題

今回の研究から、英語には GPT-4o、数学には 4o-mini を使用することによってどちらの試験においても満点を取れることが明らかになった。よって、さらなる研究のためには、より高難度の問題（数学オリンピックなど）を使用することが必要である。

5. 引用文献・参考文献

Benesse Corporation (2023). 「愛知県の高校受験の内申点・入試当日点（公立)」.https://czemi.benesse.ne.jp/open/nyushi/exam/23/naishin/. 2025 年 5 月 21 日アクセス.

Benesse Corporation (2023). 「愛知県の高校受験の内申点・入試当日点（公立）(は行)」.https://czemi.benesse.ne.jp/open/nyushi/exam/23/naishin/index_ha.html. 2025 年 5 月 21 日アクセス.

Joost C. F. de Winter (2023). ”Can ChatGPT Pass High School Exams on English Language Comprehension?”. https://link.springer.com/article/10.1007/s40593-023-00372-z. 2025 年 5 月 21 日アクセス.

OpenAI (2023). ”GPT-4 Technical Report”. https://arxiv.org/abs/2303.08774. 2025 年 5 月 21 日アクセス.

Vincent Cheng、Yu Zhang (2023). ”Analyzing ChatGPT’s Mathematical Deficiencies: Insights and Contributions”. https://aclanthology.org/2023.rocling-1.22/. 2025 年 5 月 21 日アクセス.

OpenAI (2023). ”GPT-4”. https://openai.com/index/gpt-4-research/. 2025 年 5 月 21 日アクセス.

de Winter, J. (2023). “Can ChatGPT pass high school exams in the Netherlands?”. https://arxiv.org/abs/2305.11738. 2025年5月21日アクセス.
Cheng, Y. (2023). “Performance of ChatGPT on Japan’s National Medical Licensing Examination”. https://arxiv.org/abs/2303.03120. 2025年5月21日アクセス.
愛知県教育委員会 (2024).「令和6年度 愛知県公立高等学校入学者選抜学力検査問題と正答」. https://www.pref.aichi.jp/kyoiku/gakko/hs/nyuushi/index.html. 2025年5月21日.