大学の授業試験問題を用いた
大規模言語モデルの評価と解答精度向上の検討

坂　口　琢　哉*

（受付　2023 年 5 月 30 日）

1. は　じ　め　に

近年，AI の研究発展は目覚ましく，特に自然言語処理分野における大規模言語モデル（LLM: Large Language Model）に注目が集まっている。同分野における AI 研究は数年前まで，画像認識などと比較して出遅れていたが，2017年に発表された Tranformer［1］によってその精度が飛躍的に向上し，実用化が進められた。そして，同モデルを学習させた LLM として「ChatGPT」［2］が一般公開され，誰でも利用できるようになったことで，その実力が一気に注目を集めることとなった。現に同サービスは2022年11月に公開されて以来，僅か 5 日間で100万ユーザ登録を突破し，現在は 1 億人以上ものユーザに利用されている。その応用も様々で，この半年間だけでも翻訳からコーディングまで幅広い分野において成果が報告され，同サービスの可能性や課題について議論が始まっている。

こうした背景を踏まえ，本研究でも LLM の評価実験を行い，その実用性について分析と考察を行う。その際，実験には大学の授業で実際に出題された試験問題を使用する。また，それらの問題を様々な形式で入力し，結果を比較することにより，LLM の出力を最適化するための手法についても言及する。

2. 研　究　の　概　要

2.1　研究の目的

本研究の目的は，LLM の評価実験を行い，結果を分析すると共に，入出力の最適化手法について検討，提案することである。ここでは特に，実験を行う LLM として ChatGPT を対象とする。また評価実験の内容として，本研究では日本の大学の授業で使用される試験問題を扱う。具体的には，広島修道大学において筆者が担当する科目「総合教養講義 a（情報と

* 広島修道大学

社会)」において，2022年度の期末定期試験として実際に出題された問題を使用する。これらの問題文をChatGPTに入力し，各問題に対する解答を出力させた上で採点を行うことにより，システムの問題解答能力を評価する。また，問題文の入力方法について複数のフォーマットを試し，それぞれのフォーマットによって解答精度にどのような差があるのか，より良い結果を得るためにはどのような入力方法が適切であるかについて分析，検討を行う。

2.2 関 連 研 究

前章で述べたとおり，LLMの実力が一般社会に認知されたのはChatGPTの登場が大きいが，AIの研究分野ではそれ以前にも様々なモデルが発表され，議論され続けてきた。具体的には，事前学習とファインチューニングの組合せにより汎用性を高め，自然言語処理の多数のタスクにおいて当時のSoTA（State of the Art）を記録した「BERT」[3]や，約1750億パラメータの学習によって精度を大幅に向上させ，ChatGPTのベースにもなった「GPT-3」[4]などが挙げられる。

また，こうしたLLMが次々と登場するにつれ，それらの実用性を評価する実験や，その能力を最大限引き出し，出力を最適化するための研究なども増えつつある。このうち評価実験に関する研究としては，例えばChatGPTに対して米医師国家試験を解答させたもの[5]や，MBAの最終試験を受験させたもの[6]などが知られている。これらは実在する試験問題を使用した実践的な研究であり，評価の結果いずれも合格基準を満たす結果であったことが報告されている。一方，LLMの出力を最適化する研究について，特にユーザが入力するテキストのことを「プロンプト」と呼ぶが，このプロンプトの入力フォーマットを工夫することで出力精度を大きく向上できることが最近明らかになってきている。例えばWeiら[7]は，プロンプトの中で問題に対する具体的な思考過程を例示することにより，解答精度を向上できることを示している。またKojimaら[8]は，GPT-3に算術計算や記号推論といったタスクを解かせる際，プロンプト冒頭に“Let’s think step by step（一歩ずつ考えよう)”という一文を追加しただけで解答精度が大きく改善したと報告している。こうした研究は「プロンプトエンジニアリング」と呼ばれており，現在も新たな手法が次々と提案され，注目を集めている。

こうした一連の研究の流れを受け，本研究でも実在の試験問題を使用した評価実験を行うと共に，プロンプトエンジニアリングによる出力精度の向上を試みる。次章以降において，本研究で実際に使用する問題の内容や実験方法，実験結果に対する分析と考察，およびそれらに基づいたプロンプトエンジニアリングの検討と提案について詳述する。

3. 評 価 実 験

3.1 対象科目および試験問題

本節では，本研究で実験に使用した「総合教養講義 a（情報と社会）」の試験問題について説明する。まず，同科目は「知能」をテーマとした教養科目に属する授業であり，人間が持っている知能と，コンピュータで実現される知能，すなわち人工知能という 2 つの側面について学修する。学問分野としては認知科学と知能情報学の 2 分野にまたがっており，幅広い内容を扱う授業である。

また，学期末に課される定期試験ではこれらの内容について，知識や考え方を問う問題が出題される。このうち知識を問う問題では，各説明文に対して最適な単語を，全問題共通の語群から探して解答する「選択式」で出題される。一方考え方を問う問題では，図や表で与えられたデータを，正しいアルゴリズムに基づいて計算した上で解を示す「短答式」で出題される。本研究ではこれらのうち知識問題のみを対象とし，特に2022年度に実施した定期試験の問題をそのまま採用した。同年出題した知識問題は計40問あり，内訳は前半20問が認知科学に関する内容，後半20問が知能情報学に関する内容であった。また，解答候補として提示した語群は，ダミーも含め計60単語であった。図 1 ～図 3 に，実際に出題した問題文の一部および解答語群を示す。

(1)　[　Q.01　]は、脳の中でも最も外側にあり、思考や言語といった高度な知的活動を担う組織である。
(2)　大脳を 52 の領野に細分化し、脳の機能局在性を示したものは[　Q.02　]脳地図と呼ばれる。
(3)　神経細胞の出力部である軸索が、他の神経細胞と接合している部分を[　Q.03　]と呼ぶ。

図 1 ：認知科学に関する知識問題（一部抜粋）

(21)　AI には様々な目的があるが、特に知的処理の一部代行と目的としたものは「[　Q.21　]型 AI」と呼ばれる。
(22)　AI がゲーム木を探索する際、推移の良し悪しや解までの距離を予測した[　Q.22　]と呼ばれる指標が用いられる。
(23)　第 1 次 AI ブームが収束した理由の一つに、当時の AI が[　Q.23　]しか解決できなかったことが挙げられる。

図 2 ：知能情報学に関する知識問題（一部抜粋）

語　群　一　覧				(※五十音順に並べています)
01. BERT	02. BMI	03. CNN	04. DQN	05. ELIZA
06. GAN	07. ReLU	08. RNN	09. Transformer	10. word2vec
11. ウェルニッケ	12. エキスパートシステム	13. 海馬	14. ガヴァガイ	15. 過学習
16. 確証バイアス	17. カクテルパーティ	18. 確率荷重関数	19. 完全情報	20. 桿体
21. 強化学習	22. 教師あり学習	23. 形態素解析	24. ゲシュタルト	25. 勾配消失
26. 錯誤相関	27. 事前学習	28. シナプス	29. シミュラクラ	30. 深層化
31. シンボルグラウンディング	32. 錐体	33. 選択的注意	34. 大脳新皮質	35. 大脳辺縁系
36. ディープフェイク	37. トイ・プロブレム	38. 特徴抽出	39. 特化	40. ドロップアウト
41. ナッシュ均衡	42. バックプロパゲーション	43. パレート最適	44. 汎化	45. ヒューリスティック
46. 評価関数	47. ファインチューニング	48. 不気味の谷	49. 物体検出	50. プライミング
51. フレーム	52. ブローカ	53. ブロードマン	54. プロスペクト理論	55. ペンフィールド
56. マガーク	57. マッチングバイアス	58. モラベック	59. リハーサル	60. 連言錯誤

図 3：解答語群一覧

なお，本試験は2022年 7 月28日，同科目履修者を対象に実施された。その際，授業資料の持ち込みは許可し，試験時間は50分間とした。また，最終的な受験者数は大学生181名であった。表 1 に，同試験の採点結果について概要を示す。

表 1：学生を対象とした試験の解答結果

	認知科学分野	知能情報学分野	2 分野合計	計算問題（参考）
平均点	33.91	27.07	60.98	5.39
標準偏差	7.28	10.17	16.19	4.31
正答率	84.8%	67.7%	76.2%	27.0%

3.2　実験 1：問題文のみの提示

前節で示した問題文と解答語群のうち，問題文のみを ChatGPT に入力した上で解答させる実験を行った。まず，システムに入力するプロンプトについて，冒頭の指示文を設定し，これに続けて（1)～(40）の問題文の全テキストをそのままコピーしつつ 1 問ごとに改行して，以下の図 4 のような内容とした。

以下の説明文について，[　Q.01　]～[　Q.40　]の空欄に最適な語句を答えなさい。ただし，解答はすべて単語とし，余計な解説や文は書かないこと。

(1) [　Q.01　]は，脳の中でも最も外側にあり，思考や言語といった高度な知的活動を担う組織である。
(2) 大脳を 52 の領野に細分化し，脳の機能局在性を示したものは[　Q.02　]脳地図と呼ばれる。
:
～ 以下，同様に問題を(40)まで提示 ～

図 4：問題文のみの提示に使用したプロンプト

次に，上記のプロンプトを入力して解答を出力させた上で，それらが正解かどうかを手動で採点した。なお，本実験ではシステムに対し解答語群を提示しないため，本質的には正解であっても表記揺れや言い換えなどによる差異が生じる可能性がある。そこで採点の際は，模範解答と一字一句違わない完全な解答を「正答」，表現上の差異はあるものの正解に該当し得る解答を「準正答」と定義し，各問題に対して 2 種類の基準で採点を行った。

また，ChatGPT は確率モデルであるため，質問に対してその都度異なる返答を生成する。これを踏まえ，本研究では同じ内容の実験を計 5 回実施し，全ての実験を通じて「正答率」および「準正答率」を算出することにより評価を行った。その他，途中でエラー等によりシステムが停止した場合，その実験を中断してやり直すものとした。

以上の方法に基づき，システムからの出力を得た上で結果を採点し，各問題の正答率と準正答率を算出した。図 5 に，認知科学分野20問，知能情報学分野20問，およびそれらを統合した全体40問に対する，正答率と準正答率の集計結果を示す。

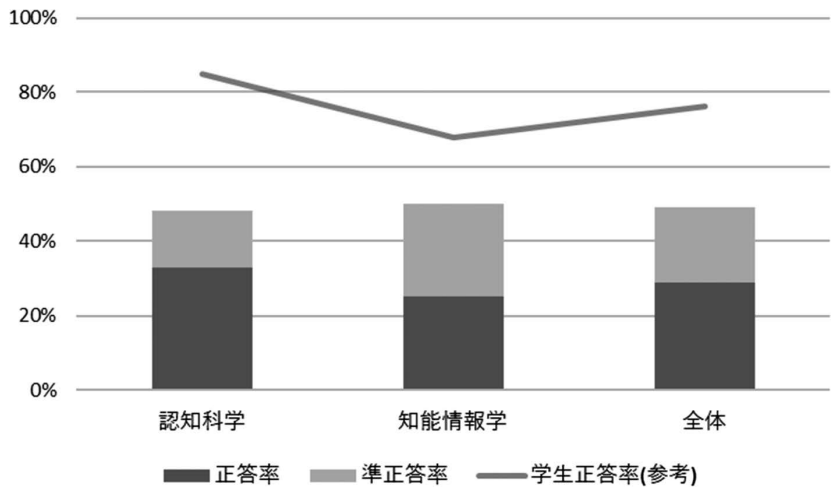

図 5：分野ごとの正答率および準正答率

図 5 から分かるとおり，問題全体に対する正答の割合は概ね30％前後，準正答も含めると約50％であった。この結果は表 1 でも示した人間の正答率には及ばないものの，そちらの出題形式は解答を語群一覧から探せる「選択式」であったことや，試験時間中も自由に持込資料を参照できたことなどを考慮すると，十分比較に値する結果と言える。また，認知科学分野と知能情報学部分野との間で，正答率の差はほとんど見られず，システムが幅広い知識を偏り無く保持し，解答能力を有していることが示唆された。

3.3 実験 2 ：問題文と解答語群の提示

次に，第3.1節で示した問題文と解答語群を両方とも提示した上で，解答語群から正しい単語を選択させる実験を行った。その際，プロンプトは以下の図 6 の内容で構成した。

以下の説明文について，[Q.01]～[Q.40]の空欄に最適な語句を【語群一覧】から選択して答えなさい。
ただし，解答はすべて番号もしくは単語とし，余計な解説や文は書かないこと。

(1) [Q.01]は，脳の中でも最も外側にあり，思考や言語といった高度な知的活動を担う組織である。
(2) 大脳を 52 の領野に細分化し，脳の機能局在性を示したものは[Q.02]脳地図と呼ばれる。
:
～ 以下，同様に問題を(40)まで提示 ～

【語群一覧】
01. BERT
02. BMI
:
～ 以下，同様に語群単語を 60.まで提示 ～

図 6 ：問題文と解答語群の提示に使用したプロンプト

上記のプロンプトから得られた出力に対し，前節同様「正答」「準正答」の 2 つの基準で採点を行った。図 7 に，本実験における正答率と準正答率を示す。

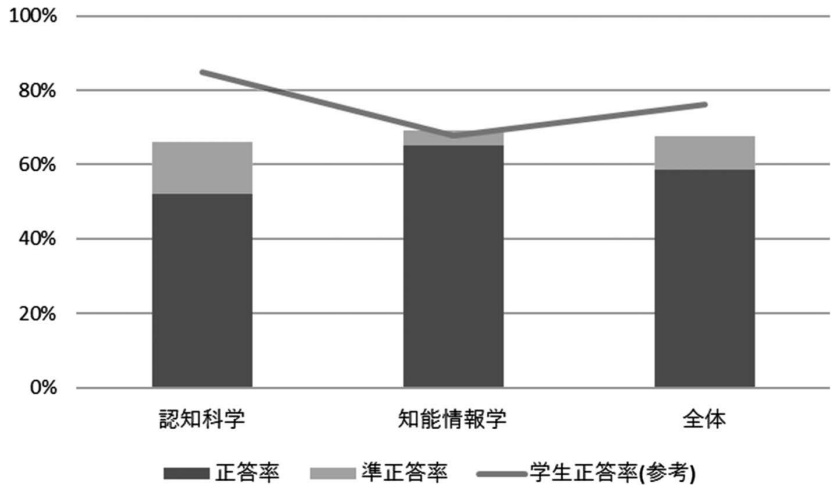

図 7 ：分野ごとの正答率および準正答率

実験 1 と比較すると，プロンプトで解答語群を提示したことにより正答率が大きく向上し，特に知能情報学分野に関しては，授業を受講した学生と比較しても遜色無いレベルで解答で

きていることが分かる。一般に，実験 1 のように解答候補が示されない課題は「再生課題」，実験 2 のように解答候補が示される課題は「再認課題」と呼ばれ，人間に対してこれらを実施した場合，前者より後者の方が高い結果になることが知られている。本実験を通じて，これと同様の傾向が LLM の問題解答においても見られることが明らかになった。

なお本実験では冒頭の指示文において，解答には語群一覧に示した番号もしくは単語を用いるよう指示していたが，結果的に 5 回の実験ではいずれも単語のみが出力され，番号による出力は得られなかった。また出力された単語の中には，例えば「円錐」「活性化関数」など，解答語群に含まれないものや，「コクリア効果」のように一見それらしいがそもそも実在しない単語なども含まれていた。これらのことから，システムは解答語群をあくまで参考程度にしか捉えておらず，相当数の解答は「選択」ではなく「生成」によって出力されている可能性が推察された。

そこで単語を解答語群の中から確実に選択させるために，解答の際は単語と番号の両方を出力するよう指示することとし，プロンプトを以下の図 8 のように修正した形で再度実験を行った。

以下の説明文について，[　Q.01　]～[　Q.40　]の空欄に最適な語句を【語群一覧】から選択して答えなさい。

(1) [　Q.01　]は，脳の中でも最も外側にあり，思考や言語といった高度な知的活動を担う組織である。
(2) 大脳を 52 の領野に細分化し，脳の機能局在性を示したものは[　Q.02　]脳地図と呼ばれる。
:
～ 以下，同様に問題を(40)まで提示 ～

【語群一覧】
01. BERT
02. BMI
:
～ 以下，同様に語群単語を 60.まで提示 ～

ただし，解答はすべて次の例のような形で番号と単語の両方を答えるものとし，余計な解説や文は書かないこと。
・解答例: (1) 01.BERT

図 8 ：番号 + 単語の出力を得るために使用したプロンプト

更に，採点に際しては番号と単語の両方が正しい場合のみを正解とし，単語が正確であっても番号がずれているものについては不正解とした。

以上の実験に対する結果を，図 9 に示す。

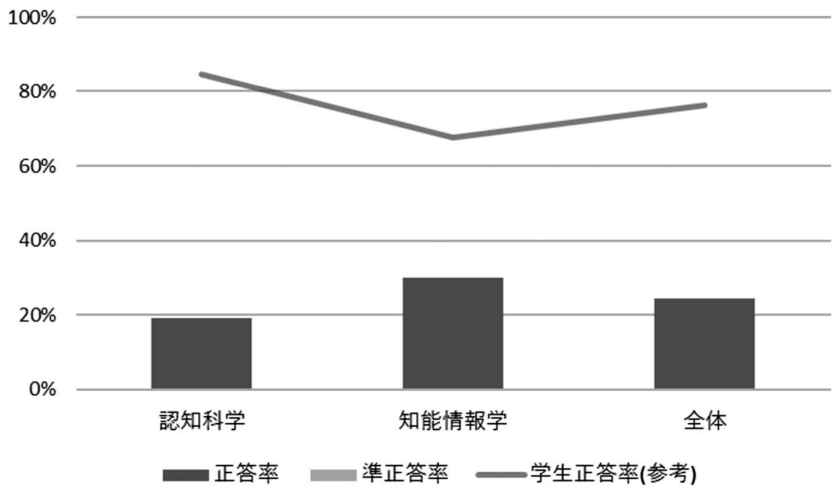

図 **9** ：分野ごとの正答率および準正答率

プロンプトを修正したことで，システムは解答語群で示した中から単語を選択して出力するようになり，準正答に該当する表記揺れなどは解消された。しかし一方で，純粋な正答率だけを比較してもスコアを大幅に落とす結果となった。そこで本研究ではこの要因について考察すると共に，正答率を向上させるためのプロンプトエンジニアリングについて検討，検証を行った。

3.4 正答率向上のためのプロンプトエンジニアリング

前節の実験で正答率が向上しなかった要因について，本研究では幾つかの予備実験を通じて分析と考察を行った。その結果，現状のプロンプトが長文になり過ぎており，その結果問題文から解答語群への参照が困難になっていることが原因の一つとして推察された。そこで本研究では，問題文を一括入力するのではなく，数回のプロンプトに分けて入力していくことで 1 回あたりのプロンプトを短縮させる方法を提案すると共に，その有効性について検証を行った。

提案手法ではまず，問題（1)～(40）を上から 5 問ずつ，計 8 グループに分割する。その上で，前節で使用したプロンプトの指示文や解答語群はそのままに，問題のみをグループ単位に限定した形で入力する。この結果，例えば問題（1)～(5）を入力する場合のプロンプトは，図10のような内容となる。

以下の説明文について，[　Q.01　]～[　Q.05　]の空欄に最適な語句を【語群一覧】から選択して答えなさい。

(1) [　Q.01　]は，脳の中でも最も外側にあり，思考や言語といった高度な知的活動を担う組織である。
(2) 大脳を 52 の領野に細分化し，脳の機能局在性を示したものは[　Q.02　]脳地図と呼ばれる。
:
～ 以下，同様に問題を(5)まで提示 ～

【語群一覧】
01. BERT
02. BMI
:
～ 以下，同様に語群単語を 60.まで提示 ～

ただし，解答はすべて次の例のような形で番号と単語の両方を答えるものとし，余計な解説や文は書かないこと。

図10：問題文の分割提示に使用したプロンプト

上記のプロンプトを入力し，システムからの出力が得られたら，このプロンプトの問題部分のみを次のグループに差し替えた形で再度入力を行う。これを繰り返すことにより，問題に対する解答を順次出力させ，最終的に全ての問題に対する出力を得ることとした。

更にこの提案手法に基づき，上記とは別に問題を10問ずつ 4 グループに分けた場合，および20問ずつ 2 グループに分けた場合についても実験を行い，それぞれについて正答率を算出した。図11に，各実験の結果を示す。

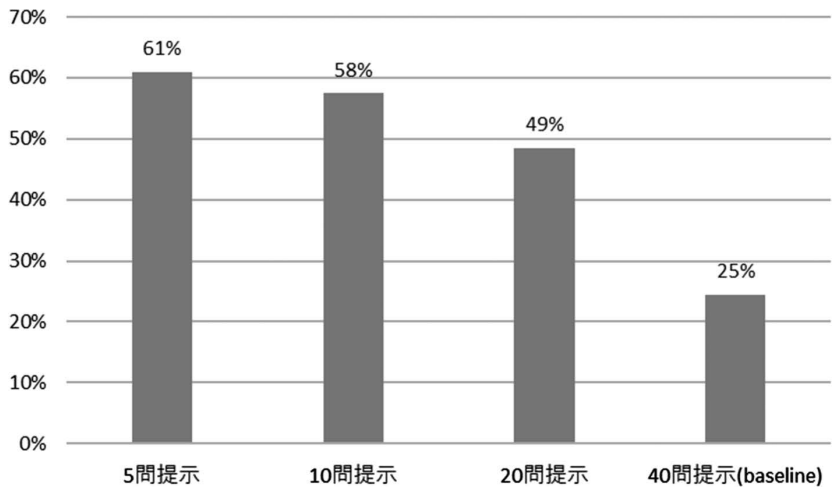

図11：問題文の分割提示による正答率の差異

実験の結果，基本的に問題数を減らしてプロンプトを短縮するほど，正答率が向上することが分かった。このことは，一度に入力する情報量がシステムの解答に影響を及ぼすことを裏付けるものであり，プロンプトが長すぎる場合は文章間の参照が困難になるという仮説を支持するものと考えらえる。同時に，プロンプトを構成する際は，その適切な情報量について考慮することが有効であることを示唆している。

また，本研究では上記の結果に対し，問題を一括提示した実験を baseline とした上で，各実験との間で Welch の t 検定を実施した。その結果，表 2 に示す通り，いずれの実験も baseline と比較して $p<0.05$の範囲で有意差が確認され，提案手法の有効性が確認された。

表 2 ：各実験に対する t 検定

	n	$\bar{x}$	σ^2	t	p（両側）
40問提示（baseline）	5	0.245	0.00575	—	—
20問提示	5	0.485	0.00175	6.19677	0.00081
10問提示	5	0.575	0.00281	7.97441	0.00009
5 問提示	5	0.610	0.00206	9.23385	0.00004

4. お わ り に

4.1 研究のまとめ

本研究では，近年急速に発達した LLM の中でも，とりわけ2022年11月に一般公開された ChatGPT に着目し，その問題解答能力について評価実験を行った。その際，実験には広島修道大学の教養科目である「総合教養講義 a（情報と社会）」の試験問題を使用し，特に2022年度の期末定期試験で実際に出題された認知科学分野の問題20問，知能情報学分野の問題20問の計40問の選択式問題を採用した。

実験では，これらの問題を含めたプロンプトを構成した上でシステムに入力し，その出力結果を手作業で採点した。その結果，認知科学分野と知能情報学分野，いずれも偏り無く解答できており，また人間の正答率には及ばないものの，十分な解答能力を示すことができた。また，問題の提示方法について，問題文のみの提示と，問題文と解答語群の両方の提示それぞれについて実験を行った結果，前者よりも後者の方が高い正答率を示し，人間が「再生課題」「再認課題」に解答する場合と同様の傾向が確認された。

一方で上記の実験において，プロンプトの中で提示された解答語群が正しく利用されず，システムが自ら単語を生成している可能性が示唆された。この問題について分析と考察を加えた結果，プロンプトに含まれる情報量が多くなりすぎた場合，問題文と解答語群との間の

参照が困難になる可能性が推察された。そこで，システムが解答を語群一覧の中から確実に選択するようプロンプトを修正し，更にその条件下であっても正答率を維持するための方法として，問題文を分割した上で繰り返し提示する手法を提案し，検証を行った。具体的には，問題文を 5 問ずつ分割提示する場合，10問ずつ分割提示する場合，20問ずつ分割提示する場合についてそれぞれプロンプトを構成して実験を行い，それらの結果について分析した。その結果，いずれの場合についても，問題文を一括提示した場合と比較して正答率が有意に改善した。このことから，プロンプト構成において適切な情報量に配慮することの重要性，および問題文を分割提示するという提案手法の有効性が示された。

4.2 今後の展望

本研究では，ChatGPT に対し「総合教養講義 a（情報と社会）」で出題した試験問題を使用して評価実験を行い，その結果認知科学分野と知能情報学分野の双方において，同システムが実用的な問題解答能力を有していることを確認した。今後の展望として，実験に使用する問題数を増やしつつ，他の分野の問題なども採用することで，分析を深めていくことが考えられる。また出題形式についても，本研究で使用した選択式問題の他に，短答式問題や論述式問題といった多様な形式について実験を行うことで，より多角的な分析が可能になると考える。

一方，本研究の中で，システムの解答精度向上には一度に入力する情報量の最適化が有効であることを示し，またその具体的なプロンプトの構成手法について提案した。今後は解答精度に影響する他の要因についても追究し，改善のためのプロンプトエンジニアリングについて検討，提案していきたい。

最後に，LLM が持つ汎用性と更なる可能性に着目し，今後は本研究で対象としたような試験問題の解答に留まらず，解答の採点や問題の生成といった新たな応用についても言及し，検証を進めていきたい。

LLM の応用については現在議論が始まったばかりであり，研究の余地がまだ多く残されている。こうしたフロンティアに対し，幅広い視点で様々な研究を重ねていくことが社会の発展に繋がり，また AI 自身の進化にも寄与するものと期待している。

参 考 文 献

［1］ Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention is all you need”, Advances in Neural Information Processing Systems, 30（NeurIPS2017）, pp. 5999−6009, 2017.

［2］ OpenAI, “Chat GPT”, https://openai.com/blog/chatgpt, 2022.

［3］ Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1（NAACL2019）, pp. 4171－4186, 2019.
［4］ Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, “Language Models are Few-Shot Learners”, Advances in Neural Information Processing Systems, 33（NeurIPS2020）, pp. 1877－1901, 2020.
［5］ Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo and Victor Tseng, “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models”, PLoS Digit Health, 2（2）:e0000198, 2023.
［6］ Christian Terwiesch, “Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course”, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023.
［7］ Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, Denny Zhou, “ Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, Advances in Neural Information Processing Systems, 35（NeurIPS2022）, 2022.
［8］ Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, “Large Language Models are Zero-Shot Reasoners”, Advances in Neural Information Processing Systems, 35（NeurIPS2022）, 2022.