仕訳問題文からの解答生成―ChatGPTを用いた予備的考察―

Answer Generation from Journal Entry Exercise Descriptions: A Preliminary Evidence using ChatGPT

中　島　隆　広
NAKAJIMA Takahiro

〈要　旨〉

OpenAIが開発したChatGPTの学習者支援への活用方法に関する検討や，様々な分野の試験問題をどの程度解けるのかを検証するなど生成AIに着目した研究が最近，相次いで登場している。そのような中，本稿では日商簿記検定3級レベルの仕訳問題を取り上げ，ChatGPTにより仕訳問題文から解答を生成できるかを検証した。主要な発見事項として，(1) Zero-shot学習よりもFew-shot学習を用いたときに正解率の水準が高くなること，(2) Few-shot学習に一部の単元までの仕訳問題と解答を例示するよりも，全単元における仕訳問題を一通り例示させることで正解率の水準が高くなること，(3) 解答すべき仕訳問題に勘定科目の候補を加えることで，Zero-shot学習や例示が不十分な場合のFew-shot学習において正解率の水準が高くなるが，全単元の仕訳問題を一通り例示させた場合では正解率の水準が必ずしも高くなるとはいえないことが分かった。

〈キーワード〉

ChatGPT，GPT-3.5，簿記，仕訳問題，解答生成，正解率

1　本稿の目的

本稿の目的は，日商簿記検定3級レベルの仕訳問題をChatGPTがどの程度解答できるかを調査することである。

OpenAIが開発したChatGPTの登場により，ビジネスの現場に留まらず，教育の現場にも大きな影響をもたらす可能性がある。たとえば，データ分析の演習時にどのようにコードを書けばよいか判断に迷う場合，従来は書籍や公式のドキュメントなどを参照し試行錯誤しながらコードを書くことになるが，ChatGPTを用いると対話形式で質問を投げかけることで出力としてコードが自動的に生成されるため，従来よりも調べる時間を短縮することができる。ただし，常に正しいコードが生成される保証はなく，出力結果が正しいかの判断は人間が行う必要がある。しかしながら，ChatGPTを活用することで調べる時間の短縮やソースコードから演習問題文を自動生成することで演習問題の作問など教材作成に役立つことが期待される。

このような生成AIは簿記教育にも役立つ可能性がある。簿記を始めて学ぶ学生にとって最初に立ちはだかる困難として，取引を借方と貸方の2つの要素に分割し記帳する複式簿記の考え方を理解することがあげられる。このようなハードルを乗り越えるためには仕訳問題を数多く解くことが求められ，学習者は教科書に加えて問題集を購入し，問題演習を行うことが推奨される。このとき，学習者支援に生成AIが活用できるかは重要な論点である。たとえば，問題を解くだけではなく自身で作問し，それが正しい問題文であるかを生成AIに判定してもらうことは深い学習に繋がる可能性もある（倉次野他，2023）。このような学習者支援などにChatGPTのような生成AIが活用できるのかについて議論する前段階として，そもそも簿記の仕訳問題をどの程度解くことができるかは現時点で必ずしも明らかにされていない。そのため，本稿ではChatGPTに仕訳問題をどの程度解くことができるのかについて検証し考察を加えることを目的とする。

2　関連研究

ChatGPTは質問を自然言語でプロンプト（prompt）に入力すると，自然言語で応答を出力してくれる。質の高い出力を得るためのプロンプト技法の1つとして，プロンプトに解きたいタスクをいくつか例示する方法がある。具体的には，翻訳タスクにおいては「こんにちわ//hello」という日本語と英語のペアをプロンプトに例示することであり，複数のペアを例示する方法をFew-shot学習（Few-shot Learning）といい，1つも例示しない方法をZero-shot 学習（Zero-shot Learning）という[1]。

OpenAIがChatGPTを2022年11月にリリースして以降，ChatGPTが試験問題をどの程度解けるのかを検証した研究がいくつか存在する。佐竹・大塚（2023）は脆弱性があるシステムやサービスを攻撃してフラグ（秘密の文字列）を入手するCTF（Capture The Flag）という情報セキュリティの知識を使うセキュリティコンテンストのうち，カーネギーメロン大学が主催する初学者向けのオンラインコンテストpicoCTF2022の問題にChatGPTを利用し，全64問中48問のフラグを獲得することに成功し，全参加者7,794人中575位という成績を収めた（佐竹・大塚，2023）。Bordt and Luxburg（2023）はドイツの大学でコンピュータ・サイエンスを学んでいる学部2年生のアルゴリズムとデータ構造の試験問題をChatGPTに取り組ませた。問題形式には多肢選択問題，証明問題，疑似コードの記述問題，グラフの描写問題[2]などが出題されている。採点を行うにあたり，どの答案用紙がChatGPTによるものかを採点者が判別できないようにするため，出力結果を手書きで答案用紙に書き写す工夫を行った。実験の結果，試験は40点満点中20点以上が合格となるところ，GPT-3.5でZero-shot学習を実施した場合に20.5点を獲得し試験に合格できることが分かった。さらに，GPT-3.5よりもパラメータ数が多くテキストに加えて画像などを含めたマルチモーダルなデータセットで事前学習したGPT-4でZero-shot学習を実施した場合は24点を獲得した。これは，試験を受けた200名の学生の平均点である23.9点をわずかに上回る結果である。さらに，OpenAIが公表したGPT-4のテクニカルレポートにおいても，様々な試験でGPT-3.5よりもGPT-4が高い点数を獲得したことが報告されている（OpenAI, 2023, Table1）。これらの結果は，GPT-3.5よりもGPT-4が高い性能を有する可能性を示唆している。

その他にも，日本や米国の医師免許試験に取り組んだ研究（Kasai et al., 2023; Kung et al., 2023）や米国司法試験に取り組んだ研究（Bommarito and Katz, 2022; Katz et al., 2023）が存在している。さらに，米国公認会計士試験を始めとする会計関連の資格試験[3]の模擬試験問題に取り組んだEulerich et al.（2023）では，表や図などの画像ファイルを使用していない多肢選択式問題を対象とし，ChatGPTがどの程度問題に解答できるかを検証している。Zero-shot学習時では，GPT-3.5を用いた場合の各試験における正解率を平均して53.1%の正解率が得られたのに対し，GPT-4を用いた場合では69.6%の正解率が得られることが分かった。さらに，GPT-4をFew-shot学習した場合には正解率の平均が76.2%とZero-shot学習時よりも高い水準であることを明らかにした。また，外部ツールとの連携を行うことで外部知識を検索する行動を実行し，そのような行動と推論を交互に繰り返すことで出力を生成するReActというプロンプト技法をFew-shot学習したGPT-4に用いた場合，各試験の正解率の平均が85.1%にまで達することを報告している。

以上のように，様々な分野においてChatGPTが試験問題等をどの程度の精度で解答できるかを検証した先行研究が存在しており，GPT-3.5からGPT-4にモデルを変更することや，Zero-shot学習よりもFew-shot学習を用いたときに正解率の水準が高まることが報告されている。さらに，ChatGPTをエージェントに見立てて外部知識の検索などの行動や推論を繰り返すプロンプト技法を利用することでさらに正解率が高くなることも報告されている。

---

[1] 本稿ではChatGPTのプロンプトにいくつかの見本を示すことをFew-shot学習としており，少量データでモデルのパラメータを訓練する意味で用いてない点に注意されたい。

[2] 試験問題はLaTexで作成されているため，問題のソースファイルをChatGPTのプロンプトに入力することでグラフの描写問題などに対応している。

[3] 具体的には米国内における公認会計士（Certified Public Accountant），公認管理会計士（Certified Management Accountant），公認内部監査人（Certified Internal Auditor），税理士（Enrolled Agent）の模擬試験を対象にしている。

3　データと検証方法

3.1　使用するデータ

本稿では日商簿記検定3級に準拠したTAC出版のテキスト『合格テキスト日商簿記3級Ver. 14.0』（以下，合格テキスト）の仕訳問題文と解答のペアをFew-shot学習で用いる見本とする。本テキストは大学で使用されている標準的な簿記のテキストであり，商品売買や手形取引など学習する単元（テーマ）ごとに解説と例題があり，各単元の最後には確認問題が配置されている。本稿では例題に着目し，証ひょうや伝票を除く各単元における単純な仕訳問題のみをFew-shot学習の対象とする[4]。

このような合格テキストの例題における仕訳問題文と解答のペアをプロンプトに例示したうえで，ChatGPTが解答すべき仕訳問題については，合格テキストの補助教材であるWebアプリケーション『仕訳猛特訓』を利用する。仕訳猛特訓は，商品売買や手形取引などのジャンルごとに仕訳問題が出題される「ジャンル別仕訳」と，ジャンル別仕訳の全123問から20問が出題される「とことん仕訳」がある。本稿では「とことん仕訳」の全20問のうち，伝票・証ひょう問題など画像ファイルが使用される問題を除いた仕訳問題をChatGPTに取り組ませてどの程度の正解率が得られるのかを検証する。ただし，1回の「とことん仕訳」で出題される20問だけでは出題される問題に重複や偏りが生じる場合があるため，1回の正解率を評価指標とするのではなく，10回繰り返した結果から考察を行う。

3.2　検証方法

本稿ではGPT-3.5-turbo[5]をChatGPT API経由で利用し，はじめにZero-shot学習とFew-shot学習で正解率が異なるかを記述統計量の比較により検証する。GPT-3.5-turboにおける入出力のトークン数[6]の上限が4,000程度であるため，合格テキストの最初の単元の例題から3,659トークンまでの単元（具体的にはクレジット売掛金まで）の例題と解答のペアを例示するFew-shot学習を行う。

次に，プロンプトに例示する仕訳問題数の違いが正解率に影響を及ぼすかを検証するにあたり，入出力における最大トークン数が16,000程度までに拡張されたGPT-3.5-turbo-16kモデルを利用する。後述する前処理を行った後の合格テキストの例題と解答のペアのトークン数は16,004であり，GPT-3.5-turbo-16kモデルを利用することで，合格テキストの各単元における仕訳問題を一通りカバーできる。

プロンプトに例示する仕訳問題文と解答のペアに関して，次のような前処理を実施した。仕訳問題文については，プロンプトに入力できる文字数には上限があるため，日付や企業名など解答するうえで必要でない仕訳問題の場合は文字数削減のために除外している[7]。また，解答が重複している類題についても文字数削減のために除外した。解答については，借方科目と借方金額，貸方科目と貸方金額の4要素を含める必要がある。本稿では自然言語で出力させるのではなく，一定の構造を持たせた出力のために解答をJSON形式で表現する。JSON（JavaScript Object Notation）とは汎用的なデータ形式であり，コンピュータのシステム間でデータをやり取りする際に利用される。金融庁による決算報告書の電子開示システムであるEDINETにおいて有価証券報告書などがXBRL（eXtensible Business Reporting Language）形式で提供されているが，これはXML（eXtensible Markup Language）形式を決算報告書に適用したものである。JSONはXMLよりも後の期間（2006年）に仕様が規定されたデータ形式であるが，XMLよりも記述が簡便であり，データをkeyとvalueの2つの要素として対応づけた辞書形式で表現し，辞書にリストを組み込むことで複雑なデータ構造であっても簡便に取

---

[4] 後述する仕訳猛特訓アプリケーションでは，証ひょうと伝票の仕訳問題はテキストだけでなく画像ファイルで情報が表示されるため，これらの単元は対象外とする。さらに，試算表，精算表，財務諸表の作成などの単元も，仕訳問題を対象とするため対象外とする。

[5] 2023年9月時点で利用可能なGPT-4のプロンプトの入出力の上限は8,000トークン程度である。本稿で利用する合格テキストの例題と解答のペアは16,000トークン近くあり，GPT-4モデルを採用するとFew-shot学習ですべての例題を利用できない。したがって，本稿ではGPT-4を検証対象とはせず，GPT-3.5-turboを採用する。なお，トークンについては脚注6を参照されたい。

[6] トークンとは部分文字列のことであり，1トークンは標準的な英文テキストの約4文字相当であるが，日本語の場合は必ずしもそのような関係になっていない。したがって，日本語で記述した仕訳問題文と解答のペアのトークン数をカウントするためにOpenAIのTokenizerを利用した（https://platform.openai.com/tokenizer）（2023年9月13日閲覧）。

[7] 日付や企業名以外にも仕訳問題文に不要な語句が含まれている場合は取り除いている。具体的には「7月20日，埼玉㈱は，山梨㈱に対する買掛金300円を，例題01の東京㈱振出の小切手で支払った」という問題文が存在しており，仕訳を解くにあたり例題01という文言が無くても問題ないため削除している。

り扱うことが可能である。このようなJSON形式の仕訳の具体例をあげると，「銀行より現金1,500円を借り入れた」という仕訳問題文が与えられた場合は図表1のように解答が表現される。

JSON形式で表現するにあたり，文字数を節約するため「借方科目」については「借」と短縮して表現し，「借方金額」は「金」，「貸方科目」は「貸」，「貸方金額」は「金額」と短縮して表現する。

最後に，ChatGPTをAIP経由で利用する場合のパラメータについて説明する。APIではプロンプトに相当するmessages部分にroleというパラメータがある。roleにはsystem，user，assistantの3種類があり，systemでChatGPTの役割を指定し，userでChatGPTへの指示（質問）を入力する。そして，出力結果がassistantに返ってきて，過去の会話を踏まえた出力を得たいときに利用する。本稿ではsystemとuserを図表2のように設定する。

その他のパラメータのうち，出力のランダム度を調整する温度（temperature）パラメータは0から2の範囲で設定でき，値が大きくなるほど出力のランダム度が高くなる。本稿では同じ質問に対しては毎回同じ解答が行われることが望ましいためtemperatureをゼロに変更し，その他はデフォルトの設定とする。

図表1　仕訳の表現

Panel A: 通常の仕訳

借方科目	金額	貸方科目	金額
現金	1,500	借入金	1,500

Panel B: JSONで表現した仕訳

```
[
{"借方科目":"現金", "借方金額":1500, "貸方科目":"借入金", "貸方金額":1500}
]
```

図表2　プロンプト（systemとuser）の設定例

```
#system
簿記の仕訳問題文とJSON形式での解答の対応関係を例示します。この例を参考にして質問に回答してください。
銀行より現金1,500円を借り入れた。¥¥[{"借":"現金","金":1500,"貸":"借入金","金額":1500}]
商品を900円で販売し，代金は現金で受け取った。¥¥[{"借":"現金","金":900,"貸":"売上","金額":900}]
…（省略）…
#user
以下の仕訳問題文の¥¥以降にJSON形式で答えを示してください。
クレジット売掛金490円について、信販会社から当社の当座預金口座に入金された。¥¥
```

4　検証結果と考察

4.1　Zero-shot学習とFew-shot学習における正解率の比較

図表3は，仕訳猛特訓アプリケーションからランダムに20問出題される「とことん仕訳」を10回繰り返したときの，Zero-shot学習とFew-shot学習における正解率の記述統計量である。なお，API経由で利用するモデルはGPT-3.5-turboである。1行目のZero-shotはプロンプトに仕訳問題文と解答のペアを全く例示しなかった場合の結果であり，2行目のFew-shot-4kはプロンプトに3,659トークンの例題と解答のペアを例示し，入出力を合わせて約4,000トークンに収まるよう調整したときの結果である。

図表3をみると，Zero-shotの平均値（中央値）は0.140（0.122）である。これは，20問出題された場合に平均して2，3問程度しか正解できなかったことを意味する。また，最小値は0であり，1問も正解できなかった場合も存在することが分かる。この結果はChatGPTに何も例示を行わなければ，仕訳問題に対する正しい解答が得られないことを示唆している。

その他にも，一般的でない勘定科目がFew-shot-4kの場合と比べて多くの問題で生成されていた。たとえば，「現金過

不足」勘定を使用すべき場合に「不明金」という勘定が生成されていた。これは，どのような勘定科目を使用すべきか例示されていないことが原因の1つとして考えられる。そのため，正解の仕訳で用いる勘定科目の候補を指定勘定科目として示すことで，正解率が高くなるかを検証する必要がある。仕訳猛特訓の仕訳問題では勘定科目の候補が6つ表示されており，このような指定勘定科目をプロンプトに追加した場合については第5節で追加検証として実施する。

図表3のFew-shot-4kの平均値（中央値）は0.340（0.323）であった。これはZero-shotの平均値（中央値）0.140（0.122）よりも高い値[8]を有しており，プロンプトに仕訳問題と解答のペアを例示することで正解率の水準が高くなることを示唆する。しかし，20問出題された場合に平均して6，7問程度しか正解できないことも意味するため，正解率の水準は必ずしも高くはない。このような結果が得られた理由として，単元の途中（クレジット売掛金）までしか例示されていないことが原因として考えられる。したがって，次項ではプロンプトに例示する仕訳問題文と解答のペアを増やした場合について検証を実施する。

図表3　Zero-shot学習とFew-shot学習における正解率の記述統計量

	Obs.	Mean	Std.Dev.	Min	25%	Median	75%	Max
Zero-shot	10	0.140	0.090	0.000	0.111	0.122	0.163	0.313
Few-shot-4k	10	0.340	0.080	0.222	0.294	0.323	0.388	0.500

4.2　例示する仕訳問題数の違いによる正解率の比較

本項ではFew-shot学習において，例示する仕訳問題文と解答のペアの数の違いにより正解率に差が生じるかを検証する。GPT-3.5-turbo-16kモデルを使用した場合だと合格テキストのすべての単元における仕訳問題の例題を一通り例示できるため，途中の単元までしか例示できないGPT-3.5-turboのモデルを使用した場合よりも正解率が高くなることが予想される。

図表4は4.1と同様に仕訳猛特訓の仕訳問題1回あたり20問を10回繰り返した正解率の記述統計量である。Few-shot-4kは図表3のGPT-3.5-turboを利用したときの正解率の記述統計量を再掲しており，Few-shot-16kはGPT-3.5-turbo-16kを利用し，16,004トークンの例題と解答のペアを例示したときの正解率の記述統計量である。

Few-shot-16kの平均値（中央値）は0.621（0.625）である。これは，Few-shot-4kの平均値（中央値）0.340（0.323）よりも高い値[9]を有しており，20問出題された場合に平均12問程度は正解できるようになった。この結果は，Few-shot-4kのように期中取引の途中までの単元だけを例示するのではなく，すべての単元における仕訳問題について一通り例示することで正解率の水準が高くなることを示唆している。

図表4　Few-shot学習における正解率の記述統計量

	Obs.	Mean	Std.Dev.	Min	25%	Median	75%	Max
Few-shot-4k	10	0.340	0.080	0.222	0.294	0.323	0.388	0.500
Few-shot-16k	10	0.621	0.074	0.467	0.611	0.625	0.684	0.688

4.3　小括

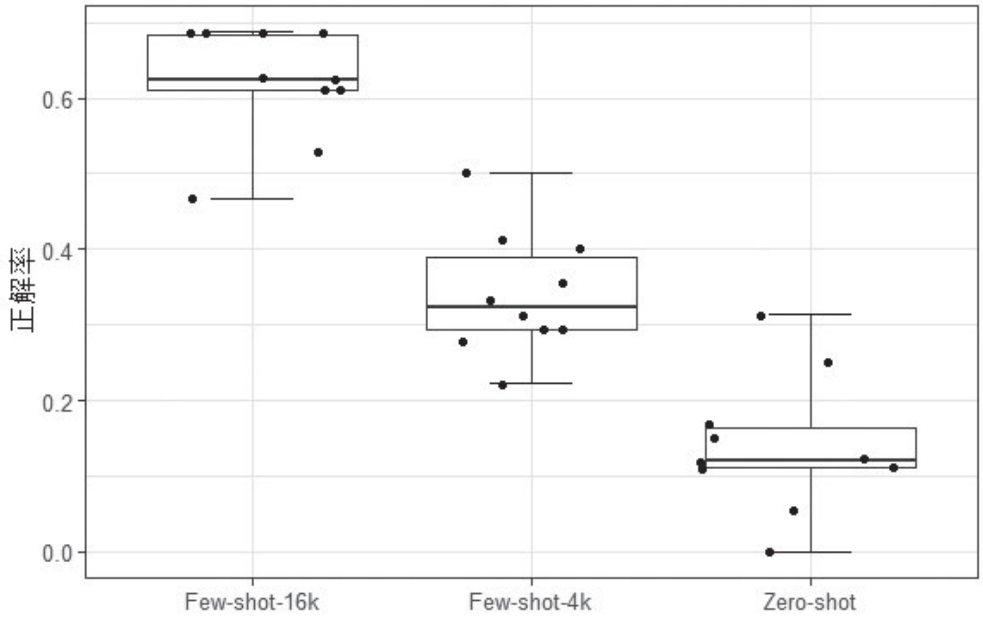
本節の検証結果を箱ひげ図としたのが図表5である。各箱ひげ図の下ひげが最小値，上ひげが最大値を示している。また，四角形のボックスの下側の線が25パーセンタイル点，上側の線が75パーセンタイル点を示しており，ボックス内の線が中央値を示している。さらに，データポイントについては仕訳問題1回20問あたりの正解率を示している。

Zero-shot学習時では中央値だけでなく25パーセンタイル点から75パーセンタイル点にかけても正解率が0.2以下であり，これは，10回仕訳問題に取り組んだときに半数程度で正解率が20%下回っていることを意味する。一方，Few-shot

[8] 平均値の差の検定に等分散性を仮定しないWelchのt検定を実施したところ，検定統計量の実現値が5.270と両側1%水準で有意である。さらに，中央値の差についても等分散性を仮定しないBrunner-Munzel検定を実施したところ，統計量の実現値が8.910と両側1%水準で有意である。

[9] 平均値の差の検定に等分散性を仮定しないWelchのt検定を実施したところ，検定統計量の実現値が8.150と両側1%水準で有意である。さらに，中央値の差についても等分散性を仮定しないBrunner-Munzel検定を実施したところ，統計量の実現値が34.648と両側1%水準で有意である。

図表5　Zero-shot学習とFew-shot学習における正解率



学習を行った場合，中央値近辺だけでなく広い範囲でZero-shot学習時よりも正解率の水準が高いことが分かる。とりわけ，一通りの単元の仕訳問題と解答のペアを例示しているFew-shot-16kでは25パーセンタイル点から60%を超える正解率を有している。この結果は，仕訳問題と解答のペアを例示することが重要であることに加えて，例示する仕訳問題は一部の単元だけでなく，全体の単元をまんべんなく例示することが重要であることを示唆している。

5　追加検証

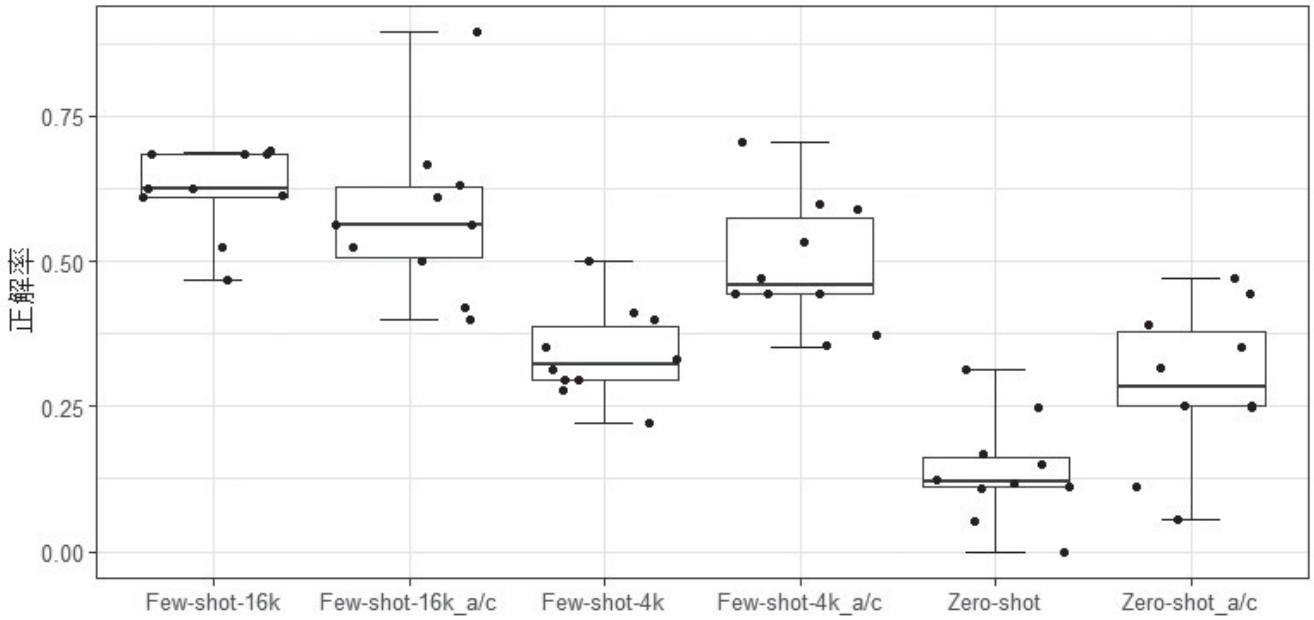
第4節の検証では，Zero-shot学習の正解率が最も低いことや，Few-shot学習と比べて独自の勘定科目が生成されることが明らかとなった。これはプロンプトに仕訳問題と解答のペアを例示していないことが原因として考えられる。本節では，勘定科目の候補をプロンプトに加えることで正解率が高くなるかを検証する。具体的には，アプリケーションの仕訳問題では勘定科目の候補が6つ表示されており，このような指定勘定科目をプロンプトのuserに追加した場合の正解率について比較を行う。なお，APIのsystemの内容や温度パラメータなどは第4節の検証と同様の設定であるが，user部分については，指定勘定科目を含めたうえで以下のような変更を加えている。

図表6　プロンプト（user）の設定例

以下の仕訳問題文の¥¥以降にJSON形式で答えを示してください。なお、勘定科目には以下の指定勘定科目の中から適切なものを選択してください。
###指定勘定科目
未収入金　未収地代　売掛金　土地　普通預金　売上
###仕訳問題文
８月10日 土地（帳簿価額100円）を100円で売却し、代金は月末に受け取ることとした。¥¥

指定勘定科目をプロンプトに加えたうえで，仕訳猛特訓アプリケーションの仕訳問題1回20問を10回繰り返したときの正解率の記述統計量について，指定勘定科目を加えない場合の正解率と比較した結果が図表7である。なお，Zero-shot_a/cはZero-shot学習時に勘定科目を指定するプロンプトを加えた場合の結果であり，Few-shot-4k_a/cはGPT-3.5-turboモデルのFew-shot学習時に勘定科目を指定するプロンプトを加えた場合，Few-shot-16k_a/cはGPT-3.5-turbo-16kモデルのFew-shot学習時に勘定科目を指定するプロンプトを加えた場合の結果を意味する。

図表7　指定勘定科目の有無による正解率



図表7をみると，Zero-shot学習時では，プロンプトに勘定科目の候補を追加することで正解率の水準が全体的に高いことが分かる。具体的には，Zero-shotの中央値が0.122であるのに対し，Zero-shot_a/cの中央値は0.283と高い値を有する[10]。この結果はFew-shot-4kでも同様に観察[11]されており，仕訳問題の例示が不十分な場合についても，勘定科目の候補をプロンプトに追加することで正解率の水準が高まることを示唆する。

しかしながら，Few-shot-16kの場合は，勘定科目の候補を加えることで正解率の水準が低くなった。具体的には，Few-shot-16kの中央値が0.625であるのに対し，Few-shot-16k_a/cの中央値は0.563と低い値を有するが，統計的に有意な水準で差があるとはいえないことが分かった[12]。他にも，Few-shot-16k_a/cの正解率の25パーセンタイル点は0.507であり，75パーセンタイル点は0.627であった。これは，問題演習を繰り返し実施したときに半数の正解率が5割から6割前半であることを意味している。Few-shot-16kの正解率については，25パーセンタイル点は0.611であり，75パーセンタイル点は0.684であった。これは問題演習時の半数で6割台の正解率を有することを意味しており，勘定科目の候補を追加した場合と比べると，安定して6割台の正解率を獲得できていることを意味する。

これらの結果に関して，今回の検証だけでは明確に理由を説明できないが，Few-shot学習では仕訳問題と解答のペアのみを例示しており，勘定科目の候補を含めていなかった。そのため，user部分に追加した6つの指定勘定科目がノイズとなったことで全体的に正解率が低くなり，正解率のバラつきが大きくなった可能性がある。そのため，systemに例示する仕訳問題と解答のペアに指定勘定科目を含めたFew-shot学習時の正解率についても検証する必要がある。しかしながら，現時点におけるOpenAIが公開しているGPT-3.5の入出力の上限は約16,000トークンが最大であるため，より多くのトークンをプロンプトに入力できるバージョンが利用可能となるのを待つ必要がある。

---

[10] 平均値も中央値と同様の傾向であり，Zero-shotの平均値は0.140に対して，Zero-shot_a/c平均値は0.289と高い値を有する。なお，平均値の差の検定に等分散性を仮定しないWelchのt検定を実施したところ，検定統計量の実現値が2.917と両側5%水準で有意である。さらに，中央値の差についても等分散性を仮定しないBrunner-Munzel検定を実施したところ，統計量の実現値が2.873と両側5%水準で有意である。

[11] Few-shot-4kの平均値（中央値）が0.340（0.323）に対して，Few-shot-4k_a/cの平均値（中央値）は0.496（0.458）と高い値を有する。なお，平均値の差の検定に等分散性を仮定しないWelchのt検定を実施したところ，検定統計量の実現値が3.631と両側1%水準で有意である。さらに，中央値の差についても等分散性を仮定しないBrunner-Munzel検定を実施したところ，統計量の実現値が5.440と両側1%水準で有意である。

[12] 中央値の差について等分散性を仮定しないBrunner-Munzel検定を実施したところ，統計量の実現値が-0.850と両側10%水準でも有意でない。また，平均値も中央値と同様の傾向である。Few-shot-16kの平均値0.621に対しFew-shot-16k_a/c平均値は0.578と低い値を有しており，等分散性を仮定しないWelchのt検定を実施したところ，検定統計量の実現値が-1.450と両側10%水準でも有意でなかった。

6　結論と今後の課題

生成AIが簿記の学習者支援のために活用できるかを議論する前段階として，OpenAIのGPT-3.5を利用して日商簿記3級レベルの仕訳問題をどの程度解けるのかについて，正解率の記述統計量の比較を行った。本稿の発見事項は次のとおりである。1つ目は，Zero-shot学習よりもFew-shot学習を用いたときに仕訳問題の正解率の水準が高いことが分かった。2つ目はFew-shot学習に一部の単元までの仕訳問題と解答を例示するよりも，すべての単元の仕訳問題を一通り例示させたほうが正解率の水準が高くなることを明らかにした。3つ目は，仕訳問題だけでなく勘定科目の候補をプロンプトに加えることにより，Zero-shot学習や例示が不十分な場合のFew-shot学習において正解率の水準が高くなることが分かった。一方で，全単元の仕訳問題を一通り例示させた場合では，プロンプトに勘定科目の候補を加えたとしても，必ずしも正解率の水準が高くなるとはいえないことが分かった。

これまでに，コンピュータセキュリティのコンテストやコンピュータ・サイエンスの学部レベルの試験，米国公認会計士試験の多肢選択問題など，様々な分野の試験問題に対してどの程度の精度でChatGPTが解答できるかを検証した先行研究が存在する。そのような中で，本稿では簿記の仕訳問題に着目し，例示方法の違いにより仕訳問題の正解率が異なるなどいくつかの点を明らかにした。このような結果は今後，簿記教育における学習者支援に生成AIの活用を検討する際に基礎となる情報を提供できた点で一定の貢献を有すると考えられる。

今後の課題については次のとおりである。OpenAIのGPT-3.5では入出力可能なトークンの上限が最も大きなモデルでも16,000程度のモデル（GPT-3.5-turbo-16k）しか現時点では公開されていない。そのため，本稿では全単元における仕訳問題を一通り例示させているが，類題などについてはトークンの上限により例示できなかった。したがって，今後は他の仕訳問題集の類題を含めた仕訳問題のバリエーションを多数含めた例示をプロンプトに含めることや，勘定科目の候補も併せて例示することで正解率が高くなるか検証する必要がある。また，GPT-4を利用したときにどの程度正解率の水準が高くなるのかや，GPT-3.5などの基盤モデルをファインチューニングすることで正解率が高くなるかを検証することも興味深い。OpenAIのGPT-3.5ではファインチューニング可能な16kモデルは現時点では公開されていないが，Meta社のLlama 2などOpenAI以外が提供する基盤モデルが存在しており，最近では日本語で学習を行った基盤モデルも他の会社により公開されている。そのようなモデルを利用することで正解率が高くなるかを検証することも今後の課題としてあげられる。

参考文献

Bordt, S., and U. von Luxburg. 2023. ChatGPT Participates in a Computer Science Exam. Working paper available at: https://arxiv.org/abs/2303.09461.

Bommarito, M. J., and D. M. Katz. 2022. GPT Takes the Bar Exam. Working paper available at: https://ssrn.com/abstract=4314839.

Eulerich, M., A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood. 2023. Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA? Working paper available at: https://ssrn.com/abstract= 4452175.

Kasai, J., Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. Working paper available at: https://arxiv.org/abs/2303.18027.

Katz, D. M., M. J. Bommarito, S. Gao, and P. Arredondo. 2023. GPT-4 Passes the Bar Exam. Working paper available at: https://ssrn.com/abstract=4389233.

Kung, T. H., M. Cheatham, and A. Medenilla. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education using Large Language Models. *PLOS Digit Health* 2 (2): e0000198.

OpenAI. 2023. GPT-4 Technical Report. Working paper available at: https://arxiv.org/abs/2303.08774.

倉次野恵・江口奈穂・林浩一. 2023.「対話型生成AIを解答者とする作問学習による生徒の知識の活用力向上の試み」情報処理学会情報教育シンポジウム2023論文集: 215-220.

佐竹達也・大塚玲. 2023.「Capture The Flagにおける大規模言語モデルの応用」2023年度人工知能学会全国大会論文集: 4N2-GS-10-01.