

原著論文

生成AIを活用した学習システムの開発

西 原 秀 夫

北里大学海洋生命科学部

要旨

別稿の「生成AIの採点・評価での活用の可能性について」¹⁾の結果を受け、近年、急速にその性能の向上が見られる生成AIを利用した学習システムを開発した。開発したシステムでは、生成AIへのアクセスにAPIを利用することで生成AIをカプセル化（ブラックボックス化）し、操作性や利便性を向上させるとともに、生徒の解答を単に採点・評価するだけではなく、解答に向けた適切なヒントを提示したり、解法の詳細な解説を提供したりするなど、学習を支援するための多様なフィードバックを提供する。また、数式や化学式の入力に対応するための数・化学式エディタの新規開発や、今後の生成AIのさらなる高機能化を前提に、図表を伴う問題への対応や学習履歴の蓄積・提供にも対応した。

キーワード：ICT利活用、生成AI、自動採点・評価、ChatGPT、Gemini、学習システム

1 はじめに

開発システム（以下「本システム」という。）は、生徒の家庭等での学習を支援するため、練習問題に対する生徒の解答の正誤を生成AIで評価し、解答に対するアドバイスやヒント、AIによる模範解答を提供する。

生徒にとっては、解答入力後、直ちに評価結果が得られるので、効率的で効果的な学習が可能となる。

開発に当たっては、別稿の「生成AIの採点・評価での活用の可能性について」¹⁾で得た知見を元に、OpenAI²⁾とGoogle³⁾の各種の生成AIモデルのAPIを用い、生徒の解答に対する正誤判定の精度や応答時間、利用料金などを比較し、より正確なフィードバックをより迅速に、より低料金で得るという視点で、使用する生成AIを選定した。

一方で、生成AIによる正誤判定なので、現状では、必ずしもすべてのケースで正しいフィードバックが得られるとは限らない。生徒に対しては、生成AIの評価には誤りがあり得ることと、生成AIの評価が誤っていると思われる場合には、生成AIによる模範解答の間違いを指摘することにも大きな学習効果があることを、理解した上で学習に取り組む

よう指導する必要がある。

2 生成AIによる解答と正誤判定（生成AIのモデル選定）

使用する生成AIを選定するにあたって、採用候補として本稿執筆段階で比較的新しいモデルで、APIが一般的に利用できるOpenAIのgpt-4oとgpt-4o-mini、GoogleのGemini-1.5-Pro、Gemini-1.5-Flash、Gemini-1.5-Flash-8Bの5モデルを対象とした。性能的には、OpenAIでは、gpt-4o>gpt-4o-miniとなっており、Googleでは、Gemini-1.5-Pro>Gemini-1.5-Flash>Gemini-1.5-Flash-8Bとなっている。なお、本稿執筆時の最新モデルとして、OpenAIのo1やo1-mini、GoogleのGemini-2.0があるが、APIが一般には公開されていなかったり、利用制限があったりするので、検討対象からは外した。

検討方法としては、計算問題として単純計算問題と文章計算問題、記述式問題として穴埋め問題と論述問題の4つのカテゴリーの問題について、生成AIの解答と正誤判定の精度、応答時間、利用料金を比較検討した。そして、これらの結果から、開発システムで採用するモデルを選定した。

2.1 計算問題（単純計算問題と文章計算問題）

計算問題としては、表2-1-1に示す単純計算問題（4問）と表2-1-2に示す文章計算問題（4問）について検討した。なお、単純計算問題は、計算過程は求めず最終的な解答の値を評価し、文章計算問題は、計算過程も含めて評価した。

表2-1-1 検討対象の単純計算問題

①	$\frac{3x-y}{4} - \frac{5x+2y}{9}$	神奈川県公立 高校入試問題 (令和6年度)
	正解、評価に用いた生徒解答： $\frac{7x-17y}{36}$	
②	$\begin{cases} 0.2x + 0.8y = 1 \\ \frac{1}{2}x + \frac{7}{8}y = -2 \end{cases}$	神奈川県公立 高校入試問題 (令和5年度)
	正解、評価に用いた生徒解答： $x = -11, y = 4$	
③	$(x-5)(x+3) - 2x + 10$ の因数分解	神奈川県公立 高校入試問題 (令和5年度)
	正解、評価に用いた生徒解答： $(x-5)(x+1)$	
④	$\int_0^{\log \sqrt{3}} \frac{e^{3x} + 4e^{2x} + e^x}{e^{4x} + 2e^{2x} + 1} dx$	横浜国大 理系入試問題 (令和6年度)
	正解、評価に用いた生徒解答： $\frac{\pi}{12} + \frac{1}{2}$	

表2-1-2 検討対象の文章計算問題

⑤	<p>十の位の数が4である3桁の自然数がある。この自然数の、百の位の数と一の位の数の和は10であり百の位の数と一の位の数を入れかえた数はこの自然数より396大きい。このときこの自然数の一の位の数を求めなさい。</p> <p>正解： 7</p> <p>評価に用いた生徒解答：</p> <p>この自然数を、$a \times 100 + b \times 10 + c$ とする。</p> <p>問題文より、 $b = 4$ ……① $a + c = 10$ ……② $a \times 100 + b \times 10 + c + 396 = c \times 100 + b \times 10 + a$ ……③ ②より、 $a = 10 - c$ ……②’ ③に①と②’を代入 $(10 - c) \times 100 + 4 \times 10 + c + 396 = c \times 100 + 4 \times 10 + (10 - c)$ $1000 - 100c + 40 + c + 396 = 100c + 40 + 10 - c$ $-100c + c - 100c + c = 40 + 10 - 1000 - 40 - 396$ $-198c = -1386$ $\therefore c = 7$</p>	神奈川県公立 高校入試問題 (令和5年度)
⑥	<p>袋の中に、赤玉が1個、白玉が1個、青玉が4個、合わせて6個の玉が入っているこの袋の中から同時に2個の玉を取り出すとき2個とも青玉である確率を求めよ。ただし、どの玉が取り出されることも同様に確からしいものとする。</p> <p>正解： $\frac{2}{5}$</p> <p>評価に用いた生徒解答：</p> <p>全6個の玉から2個取り出す組み合わせは、 ${}_6C_2 = (6!)/2!(6-2)! = (6 \times 5)/(2 \times 1) = 15$ 青玉4個から2個の青玉を取り出す組み合わせは、 ${}_4C_2 = (4!)/2!(4-2)! = (4 \times 3)/(2 \times 1) = 6$ よって、求める確率は、 $6/15 = 2/5$</p>	東京都立 高校入試問題 (令和6年度)
⑦	<p>傾き 30° の斜面上に置かれた物体に、20Nの重力がはたらいている。この重力を斜面に平行な方向と垂直な方向に分解したとき、斜面に平行な方向の成分の大きさ[N]を求めよ。</p> <p>正解： 10[N]</p> <p>評価に用いた生徒解答： $20 \times \sin 30^\circ = 20 \times 1/2 = 10$ [N]</p>	東京都教員採用 試験問題 (令和6年度)

⑧	ある濃度の希硫酸 10mL を完全に中和するのに、0.10mol/L の水酸化ナトリウム水溶液を 9.8mL 要した。この希硫酸の濃度[mol/L]を求めよ。	東京都教員採用 試験問題 (令和6年度)
	正解： 4.9×10^{-2} [mol/L]	
	評価に用いた生徒解答： 硫酸と水酸化ナトリウムの反応式は、 $\text{H}_2\text{SO}_4 + 2\text{NaOH} \rightarrow \text{Na}_2\text{SO}_4 + 2\text{H}_2\text{O}$ 希硫酸のモル濃度を x とすると、 $0.10 \times (9.8/1000) = x \times (10/1000) \times 2$ よって、 $x = 0.049$ [mol/L]	

計算問題における、各モデルの解答精度を表2-1-3に、その際の応答時間（ミリ秒）を表2-1-4に示す。解答精度については、20回の試行における正解の回数を示している。また、応答時間は20回試行の応答時間の平均値を示している。応答時間については、ネットワークや生成AIの混雑状況等によって、大きく変動する値であると考えられるので、目安としてとらえるべきものである。なお、表中の丸数字は、問題番号を示している。

表2-1-3 計算問題における解答精度（正答回数／20試行）

生成 AI モデル	①	②	③	④	⑤	⑥	⑦	⑧
gpt-4o	20	19	20	0	19	20	20	20
gpt-4o-mini	20	16	20	1	16	20	20	20
Gemini-1.5-Pro	20	20	20	17	20	20	20	20
Gemini-1.5-Flash	20	20	20	12	20	20	20	20
Gemini-1.5-Flash-8B	20	19	20	0	17	20	17	18

網掛け：20回の試行すべてで正解

表2-1-4 計算問題における解答算出の平均応答時間（ミリ秒）

生成 AI モデル	①	②	③	④	⑤	⑥	⑦	⑧
gpt-4o	5688	6298	4524	10181	7654	5052	6356	7480
gpt-4o-mini	7456	9348	7012	12030	11847	8848	6612	10374
Gemini-1.5-Pro	7758	9070	4353	16223	9763	5444	5569	11102
Gemini-1.5-Flash	2510	2878	2224	5548	3361	2658	2296	3204
Gemini-1.5-Flash-8B	2127	2566	2099	4790	2666	2367	2088	3093

網掛け：応答時間3秒以内

解答精度については、やや難解な④の問題を除けば、どのモデルも80%以上の正答率であり、概ね良好な結果といえる。

応答時間においては、今回の5モデル中では、Gemini-1.5-FlashとGemini-1.5-Flash-8Bが良好な結果を示した。応答時間は、生徒が解答を入力して、その正誤評価を受けるまでの待ち時間に相当するもので、ストレスを感じないためには、個人差もあるが、一般的には、概ね3秒以内であることが望まれる⁴⁾。

次に正誤判定であるが、これは生成AIが直接解答を算出するのではなく、生徒の解答の正誤（解答過程を必須とする場合は、解答過程を含めて）を評価するもので、その精度や応答時間は、各モデルとも解答精度と同様の傾向を示すが、必ずしも一致するものではない。

計算問題におけるの正誤判定結果を表2-1-5に、その際の応答時間（ミリ秒）を表2-1-6に示す。正誤判定は、最終正解（計算結果の値）を与えた場合と与えない場合について、それぞれ10回試行し、正答に対する正しい判定（100点満点で80点以上の評価を正解と判定）の回数を、応答時間については、10回の試行の平均値を示している。

表2-1-5 計算問題における正誤判定精度（正しい判定回数／10試行）

生成 AI モデル	①	②	③	④	⑤	⑥	⑦	⑧
gpt-4o	10	10	10	10	7	10	10	10
	10	10	10	0	4	10	10	5
gpt-4o-mini	10	10	10	10	8	10	10	5
	10	10	10	0	4	0	1	0
Gemini-1.5-Pro	10	10	10	10	10	10	10	4
	10	10	10	9	10	9	0	0
Gemini-1.5-Flash	10	10	10	10	10	10	10	10
	10	10	10	5	7	10	10	10
Gemini-1.5-Flash-8B	10	10	10	10	10	10	10	0
	10	9	10	0	2	10	10	0

上段：最終正解を与えた場合、下段：与えない場合、網掛け：10回の試行すべてで正しい判定

表2-1-6 計算問題における正誤判定の平均応答時間（ミリ秒）

生成 AI モデル	①	②	③	④	⑤	⑥	⑦	⑧
gpt-4o	2685	2199	3135	2469	4153	3514	3323	3576
	8189	8469	7368	11943	4578	3463	3698	3229
gpt-4o-mini	3288	2613	2913	3252	3687	3689	3279	3046
	11493	13117	9610	16374	4087	4654	4165	3004
Gemini-1.5-Pro	2608	3026	2694	3241	4608	3870	4614	5839
	8059	10373	5785	19419	3971	4891	5378	4348
Gemini-1.5-Flash	1231	1230	1077	1224	1890	1876	1461	1907
	3263	3675	2958	6678	1741	1593	1402	1801
Gemini-1.5-Flash-8B	1075	1233	1122	1680	1566	1066	1045	2533
	2995	3225	2600	6257	1512	1205	958	1622

上段：最終正解を与えた場合、下段：与えない場合、網掛け：応答時間3秒以内

正誤判定においては、最終正解を与えた場合は、いずれのモデルも良好な結果となったが、最終正解を与えない場合は、文章計算問題における判定精度にモデル間のばらつきが見られ、Gemini-1.5-Flashが比較的良好な結果を出した。また、⑧の設問に対する結果が低い傾向が見られた。この誤判定の主要原因は、中和反応における酸・塩基の価数の取扱いによるもので、2価の酸と1価の塩基の中和の計算で、2倍する物質（酸or塩基）を誤判断したものである。

応答時間においては、解答応答時間と同様に、Gemini-1.5-FlashとGemini-1.5-Flash-8Bが良好な結果となった。

2.2 記述式穴埋め問題

記述式穴埋め問題としては、表2-1-1に示す、理科の問題を3問、法令文関係の問題を2問の合計5問について検討した。

記述式穴埋め問題の解答精度については、1つの設問中に複数個所の問いがあることを考慮し、それぞれの問いごとにも着目し、また、正誤判定については正答に加え、表2-2-1内に記載の誤答例に対する正誤判定の検討も行った。

なお、問題文中の「#hidden#」と「#/hidden#」で囲まれた部分（本システムでは不可視属性という。）は、本システムで独自に導入した機能で、生成AIの解答に対する指示や、正誤判定に対する評価基準などを記述するもので、解答入力時には表示されない。

表2-2-1 検討対象の記述式穴埋め問題

⑨	次の文章中の空欄ア～ウに当てはまる語句を答えなさい。 図は送電の仕組みを模式的に示している。発電所から一定の電力を送り出す場合、送電線で発生するジュール熱によって失われる電力を小さくするには、送電電圧を〔ア〕することで、送電線を通る電流を〔イ〕するとよい。そのため、発電所で発電された電気は、何度か変圧されたのちに家庭へ送られている。送電に〔ウ〕を用いると、このような変圧を容易に行える。 #hidden#（※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない。） 図は、発電所→変圧器→送電線→変圧器→家庭を模式的に示している。 #/hidden#			<p>大学入試センター問題(令和2年度)</p>
		〔ア〕	〔イ〕	〔ウ〕
	正答	高く	小さく	交流
	誤答例	高く	小さく	直流

⑩	<p>次の文章は、遺伝子の本体にせまる歴史的実験について述べたものである。空欄ア～エに当てはまる語句を答えなさい。</p> <p>肺炎双球菌（肺炎球菌）には病原性の S 型菌と非病原性の R 型菌がある。グリフィスは、R 型菌と加熱殺菌した S 型菌を混ぜてネズミに注射する実験を行った。すると、このネズミには病気の症状が現れ、その体内から生きた [ア] が見つかった。これは、死滅した S 型菌の中の物質が R 型菌の性質や特徴を変化させたために起こった現象であり、このような現象を [イ] という。また、エイブリーらは、S 型菌の抽出液からタンパク質を分解させたものと、DNA を分解させたものを作り、それぞれ R 型菌と混ぜて培養する実験を行った。この場合、[ウ] を分解させた抽出液を用いた実験では [ア] の出現が確認されたが、[エ] を分解させた抽出液を用いた実験では確認されなかった。</p>				大学入試センター問題(令和2年度)
		[ア]	[イ]	[ウ]	[エ]
	正答	S 型菌	形質転換	タンパク質	DNA
	誤答例	R 型菌	形質転換	タンパク質	DNA
⑪	<p>プレート境界に関する次の文章読み、空欄ア～ウに当てはまる語句を答えなさい。</p> <p>プレート境界には、発散（拡大）境界、収束境界、すれ違い境界の3種類がある。海底にある発散境界で見られる代表的な地形は [ア]、陸上の発散境界で見られる地形は地溝（リフト）帯である。地震はどの種類の境界でも起こるが、深発地震が起こるのは [イ] 境界である。また、[ウ] 境界では火山活動は見られない。</p> <p>#hidden#（※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない。） 採点基準：模範解答と同義語の場合は一致とみなし、減点はしない。 #/hidden#</p>				大学入学共通テスト(令和5年度)
		[ア]	[イ]	[ウ]	
	正答	海嶺	収束	すれ違い	
	誤答例	山脈	収束	すれ違い	
⑫	<p>次の記述は、「日本国憲法」の条文の一部である。空欄ア～エに当てはまる語句を答えなさい。</p> <p>第十五条 公務員を選定し、及びこれを [ア] することは、国民固有の権利である。 すべて公務員は、全体の [イ] であつて、一部の [イ] ではない。 公務員の選挙については、成年者による [ウ] 選挙を保障する。 すべて選挙における投票の [エ] は、これを侵してはならない。選挙人は、その選択に関し公的にも私的にも責任を問はれない。</p> <p>#hidden#（※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない。） 「法的な正確性」を重視すること。 #/hidden#</p>				神奈川県教員採用試験問題(令和5年度)
		[ア]	[イ]	[ウ]	[エ]
	正答	罷免	奉仕者	普通	秘密
	誤答例	罷免	奉仕者	直接	秘密

⑬	次の記述は、「教育基本法」(平成18年12月公布)及び「学校教育法」(令和4年6月改正)の条文の一部である。空欄ア～エに当てはまる語句を答えなさい。 教育基本法 第十三条 学校、家庭及び地域住民その他の関係者は、教育におけるそれぞれの[ア]を自覚するとともに、相互の[イ]に努めるものとする。 学校教育法 第四十三条 小学校は、当該小学校に関する保護者及び地域住民その他の関係者の[ウ]を深めるとともに、これらの者との[イ]の推進に資するため、当該小学校の[エ]その他の学校運営の状況に関する情報を積極的に提供するものとする。 #hidden# (※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない。) 「法的な正確性」を重視すること。 #/hidden#				神奈川県教員採用試験問題 (令和5年度)
		[ア]	[イ]	[ウ]	[エ]
	正答	役割と責任	連携及び協力	理解	教育活動
	誤答例	役割	連携及び協力	理解	教育活動

記述式穴埋め問題における、各モデルの解答精度を表2-2-2に、その際の応答時間（ミリ秒）を表2-2-3に示す。解答精度については、10回の試行における正解の回数を、穴埋めの空欄ごとに示している。また、応答時間は10回の試行における応答時間の平均値を示している。なお、表中の丸数字は問題番号で、ア・イ・ウまたはア・イ・ウ・エは空欄名である。

表2-2-2 記述式穴埋め問題における解答精度（正答回数／10試行）

生成 AI モデル	⑨			⑩				⑪			⑫				⑬			
	ア	イ	ウ	ア	イ	ウ	エ	ア	イ	ウ	ア	イ	ウ	エ	ア	イ	ウ	エ
gpt-4o	10	10	9	10	10	10	10	10	10	10	9	10	10	10	0	0	10	2
gpt-4o-mini	10	10	0	10	10	5	5	8	10	10	1	1	5	9	0	0	10	3
Gemini-1.5-Pro	10	10	10	10	10	10	10	10	10	10	9	9	6	10	0	0	10	4
Gemini-1.5-Flash	10	10	5	10	10	10	10	10	10	10	10	10	1	10	0	0	7	7
Gemini-1.5-Flash-8B	10	10	0	10	9	10	10	9	10	10	2	0	3	5	0	0	8	0

網掛け：10回の試行すべてで正解

表2-2-3 記述式穴埋め問題における解答算出の平均応答時間（ミリ秒）

生成 AI モデル	⑨	⑩	⑪	⑫	⑬
gpt-4o	6779	6839	4551	5938	5800
gpt-4o-mini	6161	5304	4348	5196	7818
Gemini-1.5-Pro	6188	9719	5750	5544	7881
Gemini-1.5-Flash	2449	3178	2036	2579	2861
Gemini-1.5-Flash-8B	1573	2342	1743	1801	2330

網掛け：応答時間3秒以内

解答精度については、法令文関係以外の問題（⑨～⑪）では、gpt-4o、Gemini-1.5-Pro、Gemini-1.5-Flashが良好な結果を出している。ここで、送電に関する⑨の問題の「ウ」での不正解が目立つが、これはすべて、『送電に「ウ」を用いると、このような変圧を容易に行える。』という設問に対して、「交流」を正解としているところ、「変圧器（トランス）」と解答しているもので、必ずしも誤りとまでは言えないものである。また、⑩の問題のgpt-4o-miniの「ウ」と「エ」については、5回の試行とも「タンパク質」と「DNA」を逆に解答するという誤りであった。

一方、法令文関係の問題（⑫と⑬）では、gpt-4o、Gemini-1.5-Pro、Gemini-1.5-Flashの3モデルが日本国憲法に関する⑫の問題で検討はしているものの、いずれのモデルとも不得意としていることが分かる。法令文の様に、文章中の文言の厳密性が問われる場合は、本システムで導入した不可視属性を利用し、解答のための補足情報を記述するなどの対応が必要となる。

応答時間については、計算問題同様に、Gemini-1.5-FlashとGemini-1.5-Flash-8Bが良好な結果を示した。

次に、記述式穴埋め問題における正誤判定結果を表2-2-4に、その際の応答時間（ミリ秒）を表2-2-5に示す。正誤判定は、模範解答を与えた場合と与えない場合について、それぞれ10回試行し、正答及び誤答に対する正しい判定（100点満点で60点以上の評価を正解と判定、60点未満の評価を不正解と判定）の回数を、応答時間については、10回の試行の平均値を示している。なお、誤答判定については、表2-2-1に示した問題中の誤答例で、誤りの部分を正しく評価したもののみをカウントしている。

表2-2-4 記述式穴埋め問題における正誤判定精度（正しい判定回数／10試行）

生成 AI モデル	⑨		⑩		⑪		⑫		⑬	
	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価
gpt-4o	10	10	10	10	10	10	10	10	6	10
	10	10	10	10	10	10	10	1	9	2
gpt-4o-mini	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	0	10	1	0	5	7
Gemini-1.5-Pro	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	0	0	10	0
Gemini-1.5-Flash	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	6	1	5
Gemini-1.5-Flash-8B	10	8	10	2	10	0	10	0	10	2
	10	0	10	2	10	0	0	6	10	0

上段：正解を与えた場合、下段：与えない場合、網掛け：10回の試行すべてで正しい判定

表2-2-5 記述式穴埋め問題における正誤判定の平均応答時間（ミリ秒）

生成 AI モデル	⑨	⑩	⑪	⑫	⑬
gpt-4o	3342	2085	1947	2233	2404
	2388	2390	2127	2210	2125
gpt-4o-mini	1976	3040	2655	3411	3051
	2283	2635	3333	2313	2828
Gemini-1.5-Pro	3086	3163	2837	3304	2975
	3592	3608	3296	2993	3230
Gemini-1.5-Flash	1107	1180	1128	1166	1141
	1346	1313	1320	1303	1548
Gemini-1.5-Flash-8B	1171	1039	1157	1159	1079
	1273	1215	1228	1203	1292

上段：正解を与えた場合、下段：与えない場合、網掛け：応答時間3秒以内

正誤判定については、法令文関係以外の問題（⑨～⑪）では、正答の付与の有無にかかわらず、Gemini-1.5-Flash-8Bを除いて良好な結果を出している。⑨と⑩と⑪の問題のGemini-1.5-Flash-8Bの誤答例の評価では、⑨の問題では「ウ」の誤答の「直流」を、⑩の問題では「ア」の誤答の「R型菌」を、⑪の問題では「ア」の誤答の「山脈」を指摘しているにもかかわらず、Gemini-1.5-Flash-8Bが評価基準（評価基準として『生徒の解答に誤りがある場合は得点を必ず最優先で50点より低くする。』と生成AIに指示）を認識できていないためか、正解の判定（100点満点で60点以上）とする結果となった。また、⑪の問いのgpt-4o-miniの正答評価では、「ウ」について、解答出力では正解（「すれ違い」）を出していたものの、正答を付与しない場合の判定では、「発散境界」を正解としたため、誤判定となった。

一方、法令文関係の問題（⑫と⑬）では、正答を付与した場合は、いずれのモデルも良好な結果となったが、正答を付与しなかった場合は、逆にどのモデルも正しい判定を出すことができなかった。

応答時間については、解答出力の場合のGemini-1.5-FlashとGemini-1.5-Flash-8Bに加え、gpt-4oも比較的良好な結果を示した。

2.3 論述式問題

論述式問題としては、表2-3-1に示す、理科の問題を2問、情報の問題を2問、法令文関係の問題を1問の合計5問について検討した。

論述式問題の正誤判定については正答に加え、表2-3-1に記載の誤答例に対する正誤判定の確認も行った。

表2-3-1 検討対象の論述式問題

⑭	顕微鏡で生物を観察する際、倍率を 40 倍から 100 倍に変えたときの視野の広さと明るさについて説明せよ。		神奈川県公立 高校入試問題 (令和 6 年度)
	正解基準	視野の範囲と像の明るさの変化の記載。	
	模範解答	視野は狭くなり、見える像は暗くなる。	
	正答例	視野は狭くなり、暗くなる。	
	誤答例	視野は広くなり、暗くなる。	
⑮	遺伝子を扱う技術である「ジデオキシ法 (サンガー法)」について説明せよ。		東京都教員採用試験問題(令和 6 年度)
	正解基準	DNA 合成を停止させ、断片を分離して DNA 配列を決定することの記載。	
	模範解答	DNA 鎖の合成反応中にジデオキシヌクレオチドを利用して選択的に DNA 合成を停止させる。これにより、異なる長さの DNA 鎖を生成し、それを電気泳動によって分離することで DNA の塩基配列を決定する方法。	
	正答例	ジデオキシヌクレオチドを用いて DNA 配列の合成を停止させ、生成する断片を電気泳動で分離して DNA 配列を決定する方法。	
	誤答例	核酸やたんぱく質などのように水溶液中で帯電する物質に電圧をかけて移動させながら、大きさで分離して分析する方法。	
⑯	情報技術に関する用語である「ユニバーサルデザイン」について説明せよ。		神奈川県教員採用試験問題(令和 3 年度)
	正解基準	あらゆる人が利用しやすいデザインをすることの記載。	
	模範解答	年齢や性別、言語などに関係なく、あらゆる人が利用できる製品情報の設計のこと。	
	正答例	障害の有無、年齢、性別、人種等にかかわらず多様な人々が利用しやすいよう都市や生活環境をデザインする考え方。	
	誤答例	体の不自由な人や高齢者などが、情報技術を利用する時に遭遇する様々な障害を除くための方策。	
⑰	疑似相関について、「気温」「かき氷の売上数」「熱中症患者数」「因果関係」を用いて説明せよ。 #hidden# (※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない) 評価基準：生徒の解答で「気温」「かき氷の売上数」「熱中症患者数」「因果関係」のキーワードの 1 つでも欠けている場合は、他の評価基準を優先して、例外なく得点を 0 点とする。 #/hidden#		大阪府教員採用試験問題(令和 6 年度)
	正解基準	すべてのキーワードが正しく使われている。	
	模範解答	かき氷の売上数が増えると熱中症患者数も増えるが、その背景には、「気温」という隠れた第三の要因 (交絡因子) があり、もとの 2 つの事象には因果関係はない。	
	正答例	かき氷の売上数と熱中症患者数は正の相関があるように見えるが、これは「気温」という交絡因子による疑似相関である。気温上昇が両方の増加を引き起こしており、かき氷の売上と熱中症患者数の間には直接的な因果関係はない。	
	誤答例	かき氷の売上と熱中症患者は、高温の時に増えるが、両者の間に関係はない。	

⑮	著作権法第 35 条は、学校その他の教育機関における複製等について規定している。学校その他の教育機関で、著作物を使用したオンデマンド授業を行う場合、誰が何をしなければならないと定められているか答えよ。 #hidden# (※ #hidden#と#/hidden#で囲まれた部分は、生徒画面には表示されない) 「法的な正確性」を重視すること。 「教育機関を設置する者」が行う行為であることに注意。 「相当な額の補償金の支払い」に注意。 #/hidden#		大阪府教員採用試験問題(令和 3 年度)
	正解基準	教育機関を設置する者が、補償金を支払うことの記載。	
	模範解答	教育機関を設置する者が、相当な額の補償金を著作権者に支払わなければならない。	
	正答例	教育機関（学校や大学など）の設置者は、オンデマンド授業で著作物を使用する場合、著作権者または著作権管理団体に相当な額の補償金を支払う義務がある。	
	誤答例	教育者や管理者がガイドラインを設定し、利用を管理しなければならない。	

論述式問題における、各モデルの解答精度を表2-3-2に、その際の応答時間（ミリ秒）を表2-3-3に示す。解答精度については、10回の試行における正解の回数を示している。ここでの正解の判定は、表2-3-1内に記載の正解基準に基づき筆者が行った。また、応答時間は10回の試行における応答時間の平均値を示している。なお、表中の丸数字は問題番号である。

表2-3-2 論述式問題における解答精度（正答回数／10試行）

生成 AI モデル	⑭	⑮	⑯	⑰	⑱
gpt-4o	10	10	10	10	10
gpt-4o-mini	10	10	10	10	10
Gemini-1.5-Pro	10	10	10	10	10
Gemini-1.5-Flash	10	10	10	10	10
Gemini-1.5-Flash-8B	9	10	10	9	10

網掛け：10 回の試行すべてで正解

表2-3-3 論述式問題における解答算出の平均応答時間（ミリ秒）

生成 AI モデル	⑭	⑮	⑯	⑰	⑱
gpt-4o	7121	8979	10181	8273	6525
gpt-4o-mini	7139	14500	11540	10851	8265
Gemini-1.5-Pro	10061	18042	15083	11530	9098
Gemini-1.5-Flash	4679	6923	4086	5015	5792
Gemini-1.5-Flash-8B	3484	4780	3297	3027	2818

網掛け：応答時間 3 秒以内

解答精度については、いずれのモデルも良好な結果となった。生成AIが大規模言語モデルであることを改めて感じるところである。

一方、応答時間については、⑬の問いのGemini-1.5-Flash-8B以外は、3秒以上を要したが、その中でも、Gemini-1.5-FlashとGemini-1.5-Flash-8Bは、比較的良好な結果を示した。

次に、論述式問題における正誤判定結果を表2-3-4に、その際の応答時間（ミリ秒）を表2-3-5に示す。正誤判定は、模範解答を与えた場合と与えない場合について、それぞれ10回試行し、正答及び誤答に対する正しい判定（100点満点で60点以上の評価を正解と判定、60点未満の評価を不正解と判定）の回数を、応答時間については、10回の試行の平均値を示している。

表2-3-4 論述式問題における正誤判定精度（正しい判定回数／10試行）

生成 AI モデル	⑭		⑮		⑯		⑰		⑱	
	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価	正答評価	誤答評価
gpt-4o	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	10	10	10
gpt-4o-mini	10	10	10	10	10	8	10	10	10	10
	5	9	10	10	10	2	10	10	10	10
Gemini-1.5-Pro	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	10	10	10
Gemini-1.5-Flash	10	10	10	10	10	6	10	10	10	10
	10	10	10	10	10	10	10	10	10	10
Gemini-1.5-Flash-8B	10	4	10	10	10	0	10	10	10	10
	10	3	10	10	10	1	10	10	10	2

上段：模範解答を与えた場合、下段：与えない場合、網掛け：10回の試行すべてで正しい判定

表2-3-5 論述式問題における正誤判定の平均応答時間（ミリ秒）

生成 AI モデル	⑭	⑮	⑯	⑰	⑱
gpt-4o	2650	3790	3303	3321	2967
	3104	3485	3157	3687	3261
gpt-4o-mini	2838	3093	2904	2675	2475
	3377	3312	2807	2786	2519
Gemini-1.5-Pro	4023	7514	5038	6211	5104
	5152	8512	5740	6851	6834
Gemini-1.5-Flash	1420	2935	2027	1905	1939
	1809	2891	2371	2243	2590
Gemini-1.5-Flash-8B	1585	1899	2002	1893	1888
	1684	2318	2128	2261	2084

上段：模範解答を与えた場合、下段：与えない場合、網掛け：応答時間3秒以内

正誤判定については、gpt-4oとGemini-1.5-Proで良好な結果が出た。次いで、Gemini-1.5-Flash、gpt-4o-miniそして、Gemini-1.5-Flash-8Bの順となり、大規模言語モデルとしてのOpenAI、Googleそれぞれのモデルの性能が現れる結果となった。

応答時間については、Gemini-1.5-FlashとGemini-1.5-Flash-8Bが良好な結果となり、次いでgpt-4o-miniが比較的良好な結果を示した。

人手による採点で比較的時間と手間を要する記述式問題での利用の可能性に期待が持てる結果となった。

2.4 生成AIの利用料金

生成AIの利用には、入出力するデータ量に応じた利用料金が発生する。利用料金は、OpenAI、Googleそれぞれが定義する、入出力のデータ量に応じたToken数という単位に基づいている。

2.4.1 テキストデータの利用料金

各モデルのテキストデータの入出力に係る利用料金（100万Tokens当たり）と、問い①～⑧の計算問題のToken数と料金の平均を表2-4-1に示す。

表2-4-1 各モデルのテキストデータの利用料金（154円/ドル）

生成 AI モデル	入力データ 100 万 Tokens	出力データ 100 万 Tokens	問い①～⑧の平均		
			入力 Token 数 利用料金	出力 Token 数 利用料金	合計利用料金
gpt-4o	385.000 円	1540.000 円	213 0.082 円	474 0.73 円	0.812 円
gpt-4o-mini	23.100 円	92.400 円	213 0.0049 円	509 0.047 円	0.0519 円
Gemini-1.5-Pro	539.000 円	1617.000 円	130 0.070 円	383 0.62 円	0.690 円
Gemini-1.5-Flash	11.550 円	46.200 円	130 0.0015 円	414 0.019 円	0.0205 円
Gemini-1.5-Flash-8B	5.775 円	23.100 円	130 0.00075 円	474 0.011 円	0.0118 円

100万Tokens 当たりの利用料金では、gpt-4o-miniはGemini-1.5-Flashの2倍であるが、実際の1問当たりの利用料金は2倍以上である。これは、5モデルとも入力データ（生成AIへの指示データ）は共通であるが、OpenAIとGoogle（Gemini）では、Token数のカウント値に差があり、Google（Gemini）の方が、4割ほど小さい値になっている。また、出力データについても、生成AIからのレスポンス（正解、アドバイス、得点など）はモデルごとに異なるが、Google（Gemini）の方が、レスポンスに詳細な解説が付加される

など、比較的大きなデータを返すのに対して、総じて少ないToken数である。つまり、実際の使用料金は、単に100万Tokens当たりの料金の比較で判断できないということである。

2.4.2 イメージデータの利用料金

各モデルのイメージデータの入力に係る利用料金（1イメージ当たり）のToken数と料金を表2-4-2に示す。

表2-4-2 各モデルのイメージデータの利用料金（154円/ドル）

生成 AI モデル	画像サイズ	1 画像の利用料金
gpt-4o	512×512 画素単位	0.09825200 円
gpt-4o-mini	512×512 画素単位	0.19635000 円
Gemini-1.5-Pro	全サイズ	0.13906200 円
Gemini-1.5-Flash	全サイズ	0.00297990 円
Gemini-1.5-Flash-8B	全サイズ	0.00148995 円

OpenAI のモデルは、512×512 画素単位の料金

Google のモデルは、画素数に関係なく 1 画像あたり 258Tokens の固定料金

利用料金は、Google（Gemini）の場合イメージの画素数に依存しないが、OpenAIの場合は、イメージの画素数に依存する。そこで、イメージの画素数と認識精度の関係を確認するため、図4-2-1に示す、⑨の問題の送電の図のイメージ（画素数は、240×90画素）について、gpt-4o-miniと Gemini-1.5-Flashに対して、説明を求めた結果の抜粋を、表2-4-3に示す。

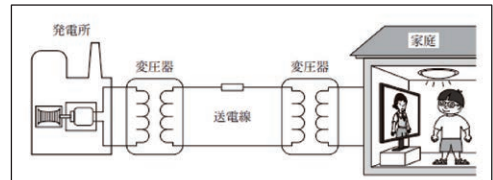


図4-2-1 ⑨の問題の送電の図

表2-4-3 生成AIによる図4-2-1の説明結果（抜粋）

gpt-4o-mini	<p>この図は、発電所から家庭への電力供給の過程を示しています。</p> <ol style="list-style-type: none"> 1. 発電所：発電所で電気が発電されます。通常、発電所では非常に高い出力が必要とされ、最初に生成される電圧は比較的低いです。 2. 変圧器（昇圧）：発電所の近くにある変圧器で、発電所で作られた電気の電圧を高くします。高い電圧にすることによって、送電中のエネルギー損失を減少させることができます。 3. 送電線：高い電圧に変換された電気は、長距離を送電線を通して移動します。送電線を使って、電力を都市や家庭まで効率よく届けることができます。 4. 変圧器（降圧）：送電線を通った高電圧の電気は、家庭やビルに到達する前に、近くの変圧器で電圧を低くします。家庭用の電化製品に適した電圧（例えば 100V）に降圧されます。 5. 家庭：最後に、低電圧に変換された電気が家庭に供給され、テレビや照明などの家電製品に使用されます。 <p>このように、電力供給は発電所から家庭に至るまで、電圧の変換（昇圧と降圧）を行いながら、効率的に電気を送る仕組みとなっています。</p>
-------------	---

Gemini-1.5-Flash	<p>この図は、発電所から家庭への電力の送電経路を簡略化して表したものです。</p> <p>○図の説明</p> <p>発電所：電力（電気エネルギー）を生み出す場所です。火力発電、水力発電、原子力発電など、様々な発電方法があります。</p> <p>変圧器：電圧を上げ下げする装置です。発電所から送電線までの長距離の送電では、電力を効率よく送るために高圧に変圧します。</p> <p>送電線：発電所から変電所、そして各家庭へと電力を送るための線です。</p> <p>変電所：高圧の電力を家庭で使える低圧に変圧する場所です。</p> <p>家庭：電力を消費する場所です。電灯、テレビ、冷蔵庫など、様々な電気製品が電力を利用しています。</p> <p>○まとめ</p> <p>この図は、発電所から家庭までの電力の流れを簡略化して示したものです。発電された電力は、高圧に変換され、送電線を通して変電所に送られます。変電所で低圧に変換された後、各家庭に配電され、電気製品によって消費されます。</p>
------------------	--

両モデルとも、OpenAIの料金算定の単位画素数である512×512画素以下でも、驚くほど完璧な認識精度を示している。gpt-4o-miniの場合、イメージを正確に読み取り、図中の家庭の電気製品であるテレビや照明器具まで認識していることには驚かされる。一方、Gemini-1.5-Flashは、降圧側の変圧器を変電所としたり、家庭の電気製品に冷蔵庫を入れたりなど、既存の知識を活用している点が特徴的である。

このことから、いずれのモデルにおいても、OpenAIの最低料金の画素数（512×512画素以内）の中で、十分な認識精度を得ることが可能と考える。

2.5 モデルの選定

本システムで採用する生成AIのモデル選定に当たっては、生徒に対して、より正しいフィードバックを、より短時間で、より低コストで返すことを条件とした。表2-5-1に、本稿で検証した18問の問題における、各モデル950回の試行（法令文関係の問題を除いた場合は740回の試行）での、正しい応答を返した割合とその際の平均応答時間を示す。

表2-5-1 生成AIによる正しい応答率と平均応答時間

生成 AI モデル	正しい応答率 (%)	平均応答時間 (ミリ秒)
gpt-4o	90.0 (93.6)	4991 (5207)
gpt-4o-mini	78.4 (82.8)	6077 (6394)
Gemini-1.5-Pro	90.2 (95.8)	6827 (7100)
Gemini-1.5-Flash	91.8 (96.6)	2559 (2599)
Gemini-1.5-Flash-8B	70.5 (78.1)	2162 (2234)

() 内は、法令文関係の問い (⑫と⑬と⑱) を除いた値
網掛け：90%以上の正しい応答率、応答時間 3 秒以内

表2-5-1の結果から、Gemini-1.5-Flashは、正しい応答率が91.8%（法令文関係の問いを除くと96.8%）と最も高く、かつ、応答時間も3秒未満と問題がない。また、コスト（利用料金）においても、比較的廉価（計算問題1回の正誤判定当たり約0.02円）であるので、本システムでは、生成AIとしてGemini-1.5-FlashのAPIを使用することとした。一方で、Gemini-1.5-Flashにおいても、8%強（法令文関係の問いを除いた場合で3%強）の誤ったフィードバックを返す。これは、小さい数字とは言えない。そこで、Gemini-1.5-Flashが誤ったフィードバックを返す問題におけるgpt-4oとGemini-1.5-Proの解答状況を確認した結果、表2-5-2及び表2-5-3の様であった。

表2-5-2 生成AIによる解答における正しいフィードバック数

生成 AI モデル	④正解	⑨正解	⑫正解	⑬正解			
				ア	イ	ウ	エ
gpt-4o	0	9	1	0	0	10	2
Gemini-1.5-Pro	9	10	0	0	0	10	4
Gemini-1.5-Flash	5	5	6	0	0	7	7

網掛け：再評価での正しい解答・解説の確率がより高い

表2-5-3 生成AIによる正誤判定における正しいフィードバック数

生成 AI モデル	④正誤判定 正答付与なし	⑫誤答評価 正答付与なし	⑬正答評価 正答付与なし	⑬誤答評価 正答付与なし	⑯誤答評価 正答付与あり
gpt-4o	0	1	9	2	10
Gemini-1.5-Pro	9	0	10	0	10
Gemini-1.5-Flash	5	6	1	5	6

網掛け：再評価での正しい応答の確率がより高い

これらの結果から、Gemini-1.5-Flashが誤ったフィードバックを返した場合でも、Gemini-1.5-Proが、それを、完全とまでは言えないが補い得ることが分かった。そこで、本システムでは、生徒が生成AIの正誤判定結果に対して、意義（いわゆる誤採点の申告）を申し出る機能を実装し、意義があった場合は、Gemini-1.5-Proを用いて再評価することとした。

3 生成AIによる解答や正誤判定の精度を上げる工夫とプロンプト

生成AIによる解答や正誤判定の精度を上げるためには、すでに述べてきたとおり、模範解答の付与が有効である。また、生成AIは、プロンプト（命令）の出し方によって、フィードバックの質が大きく異なることが知られており、意図した通りのフィードバックを得るためには、より適切なプロンプトを入力することが重要となる。

そこで、本システムではプロンプトに、生成AIの役割（教師であること）、処理内容（ヒント提供、解答作成、生徒解答の正誤判定などの別）、問題文、模範解答（必要に応じて）、

生徒解答（生徒解答の正誤判定の場合）、評価基準（生徒解答の正誤判定の場合）、フィードバックの内容について記述し、さらに、ヒント、AI解答・解説、正誤判定の場合は問題の種類別（計算問題、選択肢問題、記述問題の別）と、解答過程別（不要、必須の別）に記述している。これらによって、プロンプトの最適化を図っている。本システムで採用しているプロンプトの概略を表3-1に示す。

表3-1 プロンプトの概略

	算出過程	模範解答	プロンプト
ヒント	—	なし あり	あなたは【〇〇の教師】として、ステップ・バイ・ステップで、【問題種別】を解答するためのヒントを作成します。生徒の解答の助けになる様なヒントを提供してください。 問題：【問題文】 模範解答：【模範解答文】（模範解答ありの場合のみ） 「ヒント」のみを提供してください。その際、後処理を踏まえ、見出しを付けるようにし、最終的な答えは提供しないでください。なお、数式や化学式はLaTeX形式で提供してください。
AI 解答・解説	—	なし	あなたは【〇〇の教師】として、ステップ・バイ・ステップで【問題種別】の解答を作成します。 問題：【問題文】 模範解答：【模範解答文】（模範解答ありの場合のみ） 詳細な「解答過程」と「最終的な答え」を提供してください。後処理を踏まえ、見出しを付けるようにしてください。数式と化学式はLaTeX形式で提供してください。
		あり	模範解答なし時の処理によって算出される解答が、模範解答と一致するまで、最大 10 回繰り返す。
計算問題の 正誤判定	不要	なし	「AI 解答」を取得し、模範解答とする。 以下、模範解答あり時と同様の処理を行う。
		あり	あなたは【〇〇の教師】として、模範解答を参考に、生徒の解答の正誤を評価します。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠とあなたの考える正解を提供してください。 評価基準： 1. 解答過程は不要とする。解答過程が無くても減点しない。 2. 満点を 100 点とする。 3. 生徒の最終的な答えが正しい場合は得点を 80 点以上とする。 4. 生徒の解答に一部でも誤りがある場合は例外なく得点を 0 点とする。 5. 模範解答と生徒の解答が一致している場合は例外なく得点を 100 点とする。 問題：【問題文】 模範解答：【模範解答文】 生徒の解答：【生徒の解答文】 評価：【レスポンスの書式】の形式で提供してください。ただし、アドバイス内に模範解答は提供しないでください。なお、数式と化学式はLaTeX形式で提供してください。
	必須	なし	「AI 解答」を取得し、模範解答とする。 以下、模範解答あり時と同様の処理を行う。
		あり	あなたは【〇〇の教師】として、模範解答を参考に、生徒の解答の正誤をステップ・バイ・ステップで評価します。 評価基準に基づいて、生徒の解答を評価し得点を返してください。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠とあなたの考える正解を提供してください。 評価基準： 1. 解答過程は必須とする。 2. 満点を 100 点とする。 3. 解答過程がない場合は得点を 0 点とする。 4. 生徒の解答が誤りの場合は得点を 0 点とする。 5. 解答過程に誤りがある場合は得点を 50 点以下とする。 問題：【問題文】 模範解答：【模範解答文】 生徒の解答：【生徒の解答文】 評価：【レスポンスの書式】の形式で提供してください。ただし、アドバイス内に模範解答は提供しないでください。なお、数式と化学式はLaTeX形式で提供してください。

選択肢問題の 正誤判定	—	なし	<p>あなたは【〇〇の教師】として、あなたの考える正解を元に、生徒の解答の正誤を評価します。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠とあなたの考える正解を改行を含めずに提供してください。</p> <p>選択肢の記号の表記については、「①」「②」などの丸付き数字や「1」「2」、「1」「2」は、同じものとして扱い、表記の違いは無視してください。</p> <p>評価基準:</p> <ol style="list-style-type: none"> 1. 解答過程は不要とする。解答過程が無くても減点しない。 2. 満点を 100 点とする。 3. 生徒の解答が正しい場合は得点を 60 点以上とする。 4. 生徒の解答に一部でも誤りがある場合は例外なく得点を 0 点とする。 5. 生徒の解答が、あなたの考える正解と一致する場合は例外なく得点を 100 点とする。 <p>問題: 【問題文】</p> <p>生徒の解答: 【生徒の解答文】</p> <p>評価: 【レスポンスの書式】の形式で提供してください。ただし、アドバイス内にあなたの考える正解は提供しないでください。なお、数式と化学式は LaTeX 形式で提供してください。</p>
	—	あり	<p>あなたは【〇〇の教師】として、模範解答を参考に、生徒の解答の正誤を評価します。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠を提供してください。</p> <p>選択肢の記号の表記については、「①」「②」などの丸付き数字や「1」「2」、「1」「2」は、同じものとして扱い、表記の違いは無視してください。</p> <p>評価基準:</p> <ol style="list-style-type: none"> 1. 解答過程は不要とする。解答過程が無くても減点しない。 2. 満点を 100 点とする。 3. 生徒の解答が正しい場合は得点を 60 点以上とする。 4. 生徒の解答に一部でも誤りがある場合は例外なく得点を 0 点とする。 5. 模範解答と生徒の解答が一致している場合は例外なく得点を 100 点とする。 <p>問題: 【問題文】</p> <p>模範解答: 【模範解答文】</p> <p>生徒の解答: 【生徒の解答文】</p> <p>評価: 【レスポンスの書式】の形式で提供してください。ただし、アドバイス内に模範解答は提供しないでください。なお、数式と化学式は LaTeX 形式で提供してください。</p>
穴埋め問題の 正誤判定	—	なし	<p>あなたは【〇〇の教師】として、あなたの考える正解を元に、生徒の解答の正誤を評価します。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠とあなたの考える正解を改行を含めずに提供してください。</p> <p>評価基準:</p> <ol style="list-style-type: none"> 1. 満点を 100 点として評価する。 2. 生徒の解答に誤りがある場合は得点を必ず最優先で 50 点より低くする。 3. 生徒の解答に誤りがなく、内容的に正しい場合は得点を 60 点より高くする。 4. 生徒の解答が、あなたの考える正解と内容的に一致する場合は例外なく得点を 100 点とする。 <p>問題: 【問題文】</p> <p>生徒の解答: 【生徒の解答文】</p> <p>評価: 【レスポンスの書式】の形式で提供してください。ただし、アドバイス内に正解は記述しないでください。なお、数式と化学式は LaTeX 形式で提供してください。</p>
	—	あり	<p>あなたは【〇〇の教師】として、模範解答を元に、生徒の解答の正誤を評価します。そして、100 文字以内のアドバイスを与えてください。また、得点の算出根拠とあなたの考える正解を改行を含めずに提供してください。</p> <p>評価基準:</p> <ol style="list-style-type: none"> 1. 満点を 100 点とする。 2. 生徒の解答に誤りがある場合は得点を必ず最優先で 50 点より低くする。 3. 生徒の解答に誤りがなく、内容的に正しい場合は得点を 60 点より高くする。 4. 生徒の解答が、あなたの考える正解と内容的に一致する場合は例外なく得点を 100 点とする。 <p>問題: 【問題文】</p> <p>模範解答: 【模範解答文】</p> <p>生徒の解答: 【生徒の解答文】</p> <p>評価: 【レスポンスの書式】の形式で提供してください。ただし、アドバイス内に模範解答やあなたの考える正解は提供しないでください。なお、数式と化学式は LaTeX 形式で提供してください。</p>

(注) このプロンプトは、フィードバックの精度向上を目指して、日々改善を加えている。

本システムでは、APIアクセス段階で、それぞれのモデルに適応するようにプロンプトを調整することで、全モデルで共通のプロンプトを定義し、改善を容易にしている。

4 問題作成支援機能・問題入力支援機能

法令文問題など生成AIが不得意とする問題や図表等の読み取りを要する問題、また、共通の評価基準や生徒解答の揺らぎ（選択肢の記号の表記や漢字・カナの別など）等、問題によってはシステムに内蔵しているプロンプトのみでは、意図するフィードバックを得ることができない場合がある。これらに対応するため、既に述べた通り、本システムには、問題文中の「#hidden#」と「/#hidden#」で囲まれた部分を、生徒の問題文に表示させない「不可視属性」の機能を実装している。この「不可視属性」に、解答への補足事項や画像の説明、追加の評価基準などを記述することで、より適切なフィードバックを得ることが可能となる。とりわけ、教師による追加の評価基準は、問題の解釈、解答手順の評価、得点の算出の判断材料として使用され、これにより、採点は一貫性を持ち、客観的に行われるため、生徒は自分がどの部分で間違えたのか、何が足りなかったのかを明確に理解することが可能となる。

また、問題文の入力については、一般的な文字装飾や編集機能である太字、斜体、下線、上付き、下付き、表組の各機能、後述の数式・化学式エディタ機能、これらに加え、画像データからテキストを取り出すOCR機能を実装している。このOCR機能を利用することで、画面上に表示されている問題をスクリーンキャプチャ（Windowsの場合、Shift + Winキー+Sなど）し、テキスト化することができ、問題入力の強力な支援機能となっている。OCR機能については、GoogleのCloud Vision API（手書き文字認識対応）とGoogle Gemini-1.5-Flash API（数・化学式認識対応）を切り換えて使用する。

5 数式・化学式エディタ

本システムでは、数式や化学式は生成AIへの投入を考慮して、LaTeXで記述したものを、MathJaxライブラリを用いてブラウザ上に表示している。

しかし、LaTeXの文法に習熟するのは容易なことではなく、教師による作問、生徒の解答入力ともに、LaTeXでの記述を求めるのは現実的でない。そこで、本システムでは、数式や化学式の入力を補助するための数・化学式エディタを新たに開発し実装した。

教師の作問時の問題文入力や正解入力、生徒の解答時に〔数・化学式〕ボタンを押下すると、図5-1に示す数式・化学式エディタが起動する。



図5-1 数式・化学式エディタ

この数式・化学式エディタは、LaTeXの文法を意識することなく、タブメニューと数学・化学記号ボタンを操作することで、容易に数式や化学式を入力できるようになっている（現

在のバージョンでは、数式・化学式エディタ内で漢字や全角の英数記号を扱うことはできない。)。問題文等に配置した数式には、再編集のためのリンク情報を持たせているので、数式や化学式部分をダブルクリックすることで、数式・化学式エディタに読み込み、再編集できるようになっている。

数式や化学式を生成AIに投入する際には、入力トークン数を削減するため、図5-2に示すアルゴリズムによって、再編集のためのリンク情報やMathJaxの生成したSVGグラフィック情報などを除去し、LaTeXテキストのみを取り出してる。

```
function convMathDispToSimpleLatex(inputHtml) {
  var pattern = /<(div|img|span)
    class="mmleditor2_math"[^"]*ondblclick="mmleditor2_editMath%(event, this,
    &quot;([^&]+)&quot;;%)"[^"]*>([¥s¥S]*?<¥/¥1>)?/g;
  return inputHtml.replace(pattern, function(match, tag, p1) {
    var transformedString = p1.replace(/¥¥¥¥/g, '¥¥');
    return `<span>${transformedString}</span>`;
  });
}
```

図5-2 LaTeXテキスト取り出しアルゴリズム

6 学習履歴の蓄積と提供

本システムは、生徒の学習進捗を効果的にサポートするために、学習履歴の蓄積と提供に対応している。生徒が問題に取り組むたびに、解答結果をデータベースに記録する。具体的には、生徒が解答した内容とその解答に対する評価結果、そして採点基準に基づく採点が蓄積される。そして、生徒の要求に応じて、蓄積されている解答履歴情報が生徒に提供される。この提供情報は、単に点数を知らせるだけでなく、どのように改善すべきかのアドバイスを示している。例えば、間違えた問題については、その部分の解法や概念を再度確認するよう促したり、似たような問題を練習することを勧めたりする。これらにより、生徒は自分の学習スタイルや進捗を自己管理し、自分の強みや改善点を理解し、効率的に学習を進めることができる。

将来的には、単に学習履歴を提供するだけでなく、そのデータを分析して、生徒がどの教科・科目や単元（以下「学習領域」という。）が得意か、また苦手かを可視化することを目指したい。この可視化により、生徒は自分の強みや弱みを一目で理解できるようになる。例えば、数学の問題を解いた場合、生徒は代数、幾何、確率、統計など、学習領域ごとに評価され、その結果がグラフやチャートで表示される。これにより、生徒はどの学習領域で高いパフォーマンスを発揮しているのか、どの学習領域で改善が必要なのかを迅速に把握できる。また、システムは過去の学習履歴を元に、生徒の得意と苦手を分析し、学習をさらに効率的に進めるための指針を提供でき、これらにより、生徒は自身の得意な学習領域を伸ばし、苦手な学習領域を克服できるようになると考える。

さらに、学習履歴だけでなく、解答傾向を分析し、これに基づいて生徒に最適な問題を提供する機能を持つことも可能となる。具体的には、解答履歴から、生徒が何度も間違えた問題や解答に時間がかかる問題を特定し、その生徒に対してさらにターゲットを絞った問題を提供する。例えば、ある生徒が何度も特定の問題タイプ（例えば、因数分解の問題）で間違える場合、その生徒にはその学習領域に関連する問題を集中的に提示する。これにより、学習の焦点が絞られ、効率的に苦手な学習領域を克服することができる。逆に、生徒の解答速度や正答率が向上している場合には、少し難易度の高い問題を提供し、生徒の学力を引き上げるための挑戦を促すことも可能となる。このように、学習履歴の蓄積を通して、生徒の学力向上に繋がる仕組みの実現を目指したい。

7 システム概要

本システムは、本研究によって得られた知見に基づいて開発したもので、練習問題等の採点（正誤評価）を、解答者である生徒の端末で解答時に生成AIを用いて行うことで、生徒の学習効率（繰り返し頻度）を向上させることを目的としたシステムである。

本システムには、教師が作問や生徒管理を行うための「教師モード」と、生徒が学習を行うための「生徒モード」がある。

「教師モード」における作問では、イメージファイル内のテキストをOCR処理によりテキスト化する機能や、分数などの数式や化学反応式の入力のための数・化学式エディタ機能を利用することができる。

生徒は、教師から与えられる問題コードを用いて、「生徒モード」で本システムにアクセスし、提示される問題に、キーボードや数・化学式エディタを用いて解答を入力し、生成AIに対して解答の正誤の評価を問い合わせる。

生成AIによる評価結果が、選択肢問題の場合80/100以上、記述式問題の場合60/100の場合（設定により変更可能）は、合格となり、次の問題へ進める。生徒は、生成AIに対して、解答のためのヒントや、生成AIの考える解答を要求できる。

本システムでは、生成AIに、GoogleのGemini-1.5-FlashのAPIを使用している。なお、本システムの詳細は、本稿末に記載のURLにあるマニュアルを参照されたい。

8 まとめ

本研究では、システムの開発段階において、生成AIの評価精度やユーザーインターフェースの使い勝手などの課題が予想されたため、学習者からのフィードバックを収集することが重要であると考え、gpt-4o-miniを使用した試作システムにおいて、筆者が担当する教職課程の学生を対象に、教員採用試験の過去問題を中心に、システムを実際に使用してもらい、システムの利用体験に基づく使用感や効果を評価するためのアンケートを実施した。このアンケートでは、システムの機能が学習にどの程度貢献したか、また、学習者

が直面した課題や不便について意見を求めた。具体的には、生成AIによる評価の精度、アドバイスやヒントの有用性、学習過程における問題点を調査した。表8-1に調査結果を示す。

表8-1 学生による試作システム使用後のアンケート結果

質問	AI の評価は正確でしたか？		
回答	回答	回答数	代表的な理由
	とても正確	4	採点ミスがなかった。 とても正確だった。
	正確	16	最後の問題で、単位がないことで不正解だった。 正解の選択肢を選択しても不合格判定が出た。 ほとんど正確であったが2問ほど不正解であった。
	不正解	10	6問ほど合っているのに間違っている判定をされた。 数字に○がついていないと正解の数字を選んでいても不正解となる（複数あり）。 何度も正解の解答を不合格にしてくる。 たまに全ての選択肢が不正解になる。 正解なのに再確認しないと合格にならない場合がある。
	とても不正解	0	
	無回答等	1	
質問	AI のアドバイスやヒント、解説は、参考になりましたか？		
回答	回答	回答数	代表的な理由
	とてもなる	4	不合格になった時、合格するためのプロセス・ヒントを提示してくれたため。 ヒントの文章が簡潔でわかりやすかった。 わかりやすい。とてもタメになる。
	なる	20	とてもタメになる。 答える問題によって間違っているけど解説が変わるから。 必要な知識を提示してくれたから。 ヒントはあまり役に立たなかったけれど解説は良かった。 直接的に何の知識を見直すべきかをある程度伝えてくれるため。 間違っている時もあるが、ほとんどの的確な解説だったから。 回答に導くためのヒントが細かく書かれていて、分かりやすかった。
	ならない	4	ヒントがほぼ答えを指してしまっている内容があったため。 答えがそのまま書いてあったり、抽象的な表現だったりするから。
	全くならない	0	
	無回答等	3	

質問	このシステムでの学習についての感想		
回答	回答	回答数	代表的な理由
	肯定的回答	19	この学習は答えてすぐ解説してくれるからとてもわかりやすい。 解説見て自分で勉強できるのはいいと思った。 自習をする時にはとても助けになると感じた。 スマホで取り組めるので、移動時間など隙間時間で勉強に活用できる。 問題演習しやすく、是非利用したいと思う。 手元で短時間で学習できるため良いシステムだと感じた。 一通り勉強し終わったあとの確認テストという使い方もできると感じた。授業の復習課題として使ってみたい。
	否定的回答	8	文字数が多くて少し見づかった。 どのタイミングでも好きな問題に移動できればいいと思った。 不正解が正解になるのは危険。 図を見ながらの問題は非常にやりにくかった。
	無回答等	4	

アンケート結果は、システムの強みとともに改善が必要な部分を明確に示しており、システム改善に向けた貴重な指針となった。

生成AIによる評価の正確性については、半数以上の学生が「正確」と回答しているが、一定数の学生が精度に不満を抱いていることも明らかになった。具体的には、生成AIによる評価が「とても正確」と評価されたのは4名（約12%）で、「正確」と感じた学生は16名（約48%）であった。これらの学生は、生成AIの採点が概ね正確であると感じたものの、いくつかのケースでは不正確な判定が行われたと報告している。特に、「正しい解答でも不合格判定を受ける」、「数字に○がついていないと正解と認められない」などの細かな採点基準に不満を持つ学生が多く、これらの点について改善が求められる。一方で、「不正確」と感じた学生は10名（約30%）に上り、これらの学生は「正解にもかかわらず不合格となった」、「正しい選択肢を選んだのに不正解とされた」といった事例を挙げている。生成AIが誤判定を行うと、学習者のモチベーションに悪影響を及ぼし、正当な努力が無駄に感じられることがある。今後の改善点として、採点基準の柔軟化や誤判定を減らすための機能強化が求められる。例えば、数字や単位の違いによる不合格判定を緩和する方法や、採点結果に対する再確認（意義の申告）のプロセスを導入することが有効と考えられる。

生成AIのアドバイスやヒント、解説については、大多数の学生が有用だと感じており、特に「とてもなる」と評価した学生が4名（約12%）、次いで「なる」と感じた学生が20名（約60%）であった。これらの学生は、「間違えた時に具体的な改善方法を示してくれる」や、「問題を解くプロセスに関するヒントが有益である」と評価している。アドバイスや解説は、特に学習者が自分の理解を深めるために重要な役割を果たしており、効果的な学習を支援するためには、これらの機能の充実が不可欠であると考えられる。しかし、一部の学生からは「ヒントが答えをほぼ指摘してしまう」、「解説が抽象的で理解しにくい」という意見があり、これらの点については、改善が求められる。

学習システム全体に対する感想としては、19名（約58%）の学生が「とても良い」と感じた」と答えた。肯定的な意見では、「即座に解説が提供される」、「自習や復習に役立つ」、「隙間時間にスマホで学習できる」など、学習の効率を高める点が特に評価されている。また、「問題演習がしやすく、便利だと感じた」、「確認テストとして活用したい」といった意見もあり、学習者がこのシステムは積極的に活用できるという認識を持っていることが分かった。一方で、否定的な意見も8名（約24%）から寄せられており、「文字数が多くて見づらい」、「問題の移動に自由度がない」、「図を用いた問題がやりにくい」といった操作性や視覚的な問題が指摘された。特に、「問題間の移動に自由度がない」という点は、学習者が自分のペースで学習を進めるために重要な要素であり、この自由度を高めることが求められる。また、視覚的な要素については、情報の過剰やレイアウトの不備が学習の妨げになることがあり、インターフェースの整理や視覚的な工夫が必要となると考える。

アンケート結果を通じて、この学習システムは一定の評価を受けているものの、いくつかの改善点が浮き彫りになった。生成AIによる評価の精度向上や、アドバイスや解説の質の改善、ユーザーインターフェースの使いやすさ向上など、今後の改良が求められる。特に、採点基準やヒントの提供方法を柔軟にし、学習者が自分のペースで効果的に学習できるような環境を整えることが重要と考える。これらの改善により、学習者一人ひとりに、より適切な学習支援を提供し、より効果的な学習を実現することができると考える。

なお、このアンケートにより明らかとなった改善点の一部は、既にシステムに実装し、本研究での解答精度や正誤判定精度の検証に反映している。

本システムでは、生成AIが生徒の解答に対して即時にフィードバックを提供し、解答ミスの指摘や解答過程の改善策を提示する。誤答があった場合には、具体的なアドバイスや補完問題を提示し、理解を深めながら反復学習を促進できる。学習者が困難に直面した際には、適切なヒントを与えることで問題解決能力を育む支援を行う。これにより、学習者は積極的に学習に参加し、学習効果を最大化することができると考える。

将来的には、生成AIの高機能化と蓄積データの高度な活用により、学習者の解答履歴や学習スタイル、進捗状況をリアルタイムで解析し、強みや弱みを正確に把握することで、最適な問題を自動生成・提供することが可能となり、学習者一人ひとりに個別最適な学習を提供できるようになり、さらに、生成AIの音声認識や自然言語処理を活用すれば、学習者の疑問にリアルタイムで回答したり、学習の進行を確認したりすることも可能となる。これによって、学習者は従来の受動的な学習から、より能動的かつ対話的な学習へと移行し、自分のペースで学びを進められるようになると確信している。

教師にとっては、生成AIが提供する個別化された学習内容やフィードバックを活用することで、生徒一人ひとりの進捗や問題点を細かく把握し、指導に生かすことが可能となる。生成AIから得られるデータを基に、より効果的な支援を行うことで、学習環境をさらに強化できると期待される。

生成AIが登場して僅か2年、この間の生成AIの進化と、そのスピードには目を見張るものがあり、その進化は衰えを見せるどころか加速している。極めて近い将来、より広範かつ高度で難解な問題に対しても、高精度のフィードバックが得られるようになることは確実であると考ええる。その時に向け、本システムが生徒の学習を支援する一助となれば幸いである。

最後に、本システムを試用するためのURLを示す。

<https://edu-mat.sakuraweb.com/studywithai/>

参考文献

- 1) 宮野誠, 西原秀夫, 生成AIの採点・評価での活用の可能性について, 北里大学教職課程センター教育研究, 第10号, 2025.
- 2) <https://openai.com/ja-JP/chatgpt/overview/>
- 3) <https://gemini.google.com/app?hl=ja>
- 4) 丸岡邦明, 中村亜由美, 住谷香織, コンピュータの応答遅れ時間に対するユーザーの寛容度, 木更津工業高等専門学校紀要, 第44号, 2011.

本稿では、生成AIの解答評価のため、神奈川県公立高等学校入試、東京都立高等学校入試、京都大学入試、横浜国立大学入試、大学入試センター、大学入試共通テスト、神奈川県教員採用試験、東京都教員採用試験、大阪府教員採用試験の過去の問題を引用した。