【2024年最新】共通テストを色んな生成AIに解かせてみた（ChatGPT vs Bard vs Claude2）

株式会社LifePrompt
2024年1月16日 19:53 フォローする

2023年の流行語大賞にも選ばれた「生成AI」。

ChatGPTだけでなく、Google BardやClaude2など似たようなAIチャットボットも登場し、性能も日に日に上がっている感覚がありますね。
しかし、結局どれが一番賢いんだろう？と思いつつひとまずChatGPTを使っている方も多いはず。

そこで、今どのチャットAIが一番頭良いのか白黒つけてしまおう！ということで、ちょうど週末に行われた大学入試共通テスト2024を使って学力テストを行いました！

参加する学生は、
①GPT-4くん
②Google Bardちゃん
③Claude2さん
の三名です

果たして誰が学力王の座に輝くのか・・・？

▼ 目次

すべて表示

選手入場

①GPT-4くん

一人目は、皆さんご存じChatGPTです。Open AI予備校に月額＄20の課金して学力武装しています。
特徴としては、プロンプトの研究が進んでおり、画像やPDFファイルの読み取りもできて、プラグインなど外部ツールとの連携もお手のものといった万能さがあります。
今回は、テキストでの読み込みをベースに、プロンプトとwebブラウジング、画像入力の機能を駆使して共通テストを解いてくれます！

②Google Bardちゃん

二人目はGoogleハイスクールの優等生、Bardちゃんです。
他の二人と違って、ファイルの読み込みや外部ツールとの連携などは今はまだできませんが、レスポンスの速さと画像読み込みの性能には定評があります。
今回は、得意の画像認識を武器に、可能な部分は全て画像のまま読み取りながらテストに立ち向かいます。

③Claude2さん

三人目はAnthropic塾からの刺客、Claude2さんです。ChatGPT超えの最強LLMが日本上陸！と話題になりました。
最大100,000トークン(=75,000語)という長いテキストに対応できる処理能力の高さと、嘘を言いにくいという安全性の高さを兼ね備えていると噂です。
今回は、テキスト入力に特化した回答を行ってくれるようです。

決戦の舞台：2024大学入試共通テスト

試験問題には、ちょうど週末に実施された2024年の大学入試共通テストを用います。

テストするのは、国語・英語（リーディング）・数学（1A, 2B）・社会（世界史・日本史）・理科基礎、のたっぷり5教科7科目です。

今回は、河合塾さんの「2024年 大学入試共通テスト速報」から問題と正答をお借りして検証させていただきました。みなさんもぜひ一度目を通してみてください！

2024年度 大学入学共通テスト速報 | 大学入試解答速報 | 大学受験の予備校・塾 河合塾

2024年度 大学入学共通テスト速報では、河合塾が作成した分析コメントを公開しています。大学受験の予備校・塾、学校法人河合

kaisoku.kawai-juku.ac.jp

実験方法

基本的には、テキストか画像で試験問題をAIに入力し、それに対するテキストでの出力内容をもとに答え合わせをする形式を採ります。

試験問題を読み取らせる上での工夫点は二つです。

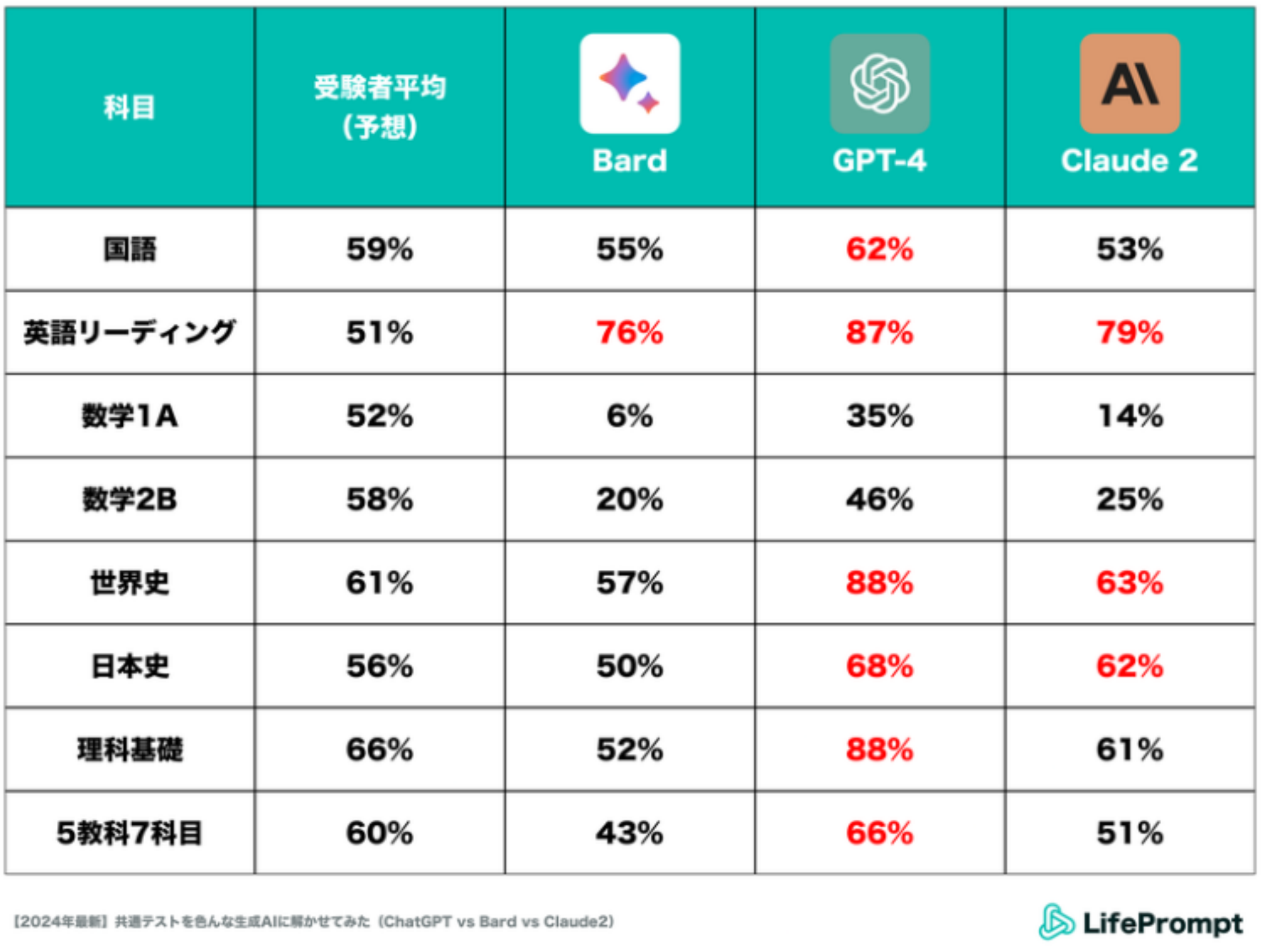
①文字と表はテキスト化する
Google Documentの文字起こし機能と、マークダウン形式による特殊記号や表をテキスト化を活用します。詳しくは以前のnote「ChatGPTに共通テスト（旧センター試験）を解かせてみた」をご覧ください
(https://note.com/lifeprompt/n/n75b6f4bf4e05)

②グラフや図は、テキスト化するor画像ファイルに変換する

画像はそのままPDFもしくはpng形式で貼り付けることも可能ですが、枚数制限がある場合などもあるため、「![画像](https:.....)」といった形でリンク化して入力するのとで適宜使い分けます。

結果発表：やはりGPT-4はバケモノだった

まずは試験の結果をご覧ください



赤字は受験者平均点(河合塾予想)を超えた箇所

ということで、**優勝は断トツでGPT-4くん**となりました！

一目見ただけでも面白そうな結果になりましたね、GPT-4が数学以外の科目で受験者平均を突き放してしまいました。Claude2もGPT-4には及びませんでしたが複数科目で受験者平均を上回っています。

この結果の表を眺めるだけでもいくつか傾向が読み取れそうです。
①GPT-4がすべての科目で他二つのツールを圧倒
②数学科目に関してはどのAIも全然点取れていない
③高得点を狙えている科目でも、満点は取れていない

共通テストの結果を深掘り

①GPT-4がすべての科目で他二つのツールを圧倒
こちらの理由に関しては、以下の理由が考えられます。

- GPT-4の生成AIとしての性能がシンプルに高い
- 他のAIに比べてプロンプトや効果的な活用方法が研究されているため、ポテンシャル発揮率が高かった

とりわけリンク化された画像を読み取る性能や、解釈が定まっている事実を的確に取り出す能力の高さは、社会や理科を回答させている中で実感できるレベルでした。

②数学科目に関してはどのAIも全然点取れていない
こちらの理由としては、AIたちが以下の二重苦を抱えていた説が有力です。

- 生成AIの計算スキルが高校数学の範囲を簡単に溶けるレベルまで進化していなかった
- 共通テスト数学の特殊な解答形式に対応できなかった

後者に関して。共通テストの数学は最も形式に癖がある科目であり、ただ計算をするだけでなく問題の空欄に書かれているカタカナに正しく数字や記号をあてはめなくてはなりません。今回の検証では、専用のGPTsを汲んだGPT-4はまだカタカナへの当てはめミスは少なかったですが、Google BardやClaude2については即興のプロンプトでは対応できずミスを連発してしまっていました。

数学Ⅱ・数学B　（注）この科目には，選択問題があります。（27ページ参照。）

第1問　（必答問題）（配点　30）

〔1〕

(1)　$k > 0$，$k \neq 1$ とする。関数 $y = \log_k x$ と $y = \log_2 kx$ のグラフについて考えよう。

(i)　$y = \log_3 x$ のグラフは点$\left(27,\ \boxed{\text{ア}}\right)$を通る。また，$y = \log_2 \dfrac{x}{5}$ のグラフは点$\left(\boxed{\text{イウ}},\ 1\right)$を通る。

(ii)　$y = \log_k x$ のグラフは，k の値によらず定点$\left(\boxed{\text{エ}},\ \boxed{\text{オ}}\right)$を通る。

(iii)　$k = 2$，3，4 のとき

$y = \log_k x$ のグラフの概形は $\boxed{\text{カ}}$

$y = \log_2 kx$ のグラフの概形は $\boxed{\text{キ}}$

である。

（数学Ⅱ・数学B第1問は次ページに続く。）

共通テストの数学科目の解答形式のイメージ
(引用：河合塾 2024年大学入試共通テスト解答速報 問題PDF)

③高得点を狙えている科目でも、満点は取れていない

満点が取れない理由はたまたまではなく、

問題の中に、一部(AIたちにとって)複雑な形式の問題が仕込まれており、それらを高確率で落としているから。

です。これについては後ほど詳しく説明します。

AIについてアレコレ

総論：明確化した現行AIの得意不得意

共通テストを解いてもらう中で、今で回っているAIには、ツール名に関係なく次のような特徴があることが見えてきました。

①単純な知識問題や読解問題は瞬殺で正解するようになった

②複数の処理を同時に求められると急激にパフォーマンスが悪くなる

①例えば英語については、どのAIも75%以上の得点率と非常に高得点を記録しています。今のAIは英語がベースのLLMを背景に持っていることもあるでしょうが、共通テストの英語は、語彙のレベルや文章構造の複雑さに関しては我々の母語である日本語の現代文よりかは大分易しいものとなっているので、AIにとって処理しやすいのだと思われます。

他にも、一見点数がパッとしない日本史や理科基礎の科目に関しても、失点源はおおよそ指示が複雑な問題となっており(後述)、語句の穴埋めやシンプルな正誤問題ではほとんど正解しています。

②逆にAIが苦手としているのが、「複数の処理を同時に求められる問題」です。満点を取ることができなかった原因もこれです。

実際に、AIたちが苦しんでいた問題パターンをいくつか紹介します。

1、順序並び替え問題

次の問題を見てください。日本の史実を年代順に並び替える問題で、答えは⑤Ⅲ－Ⅰ－Ⅱ。各文の内容がどの年代にあたるかや歴史の前後関係を追えれば他の問題と難易度はさほど変わらないように見えますが・・・

問 2　下線部ⓑに関して述べた次の文Ⅰ～Ⅲについて，古いものから年代順に正しく配列したものを，後の①～⑥のうちから一つ選べ。 2

Ⅰ　交易を目的として朝鮮の三浦に居住していた日本人らが，特権の縮小に反発して暴動を起こしたため，その後の通交は一時的に衰退した。

Ⅱ　全国を統一した人物が九州に本陣を置いて朝鮮半島に出兵した際に，朝鮮の金属活字が日本にもたらされた。

Ⅲ　倭寇の活動が活発化したため，朝鮮が倭寇の本拠地と考えた対馬を襲撃し，貿易が一時中断した。

① Ⅰ — Ⅱ — Ⅲ　② Ⅰ — Ⅲ — Ⅱ　③ Ⅱ — Ⅰ — Ⅲ
④ Ⅱ — Ⅲ — Ⅰ　⑤ Ⅲ — Ⅰ — Ⅱ　⑥ Ⅲ — Ⅱ — Ⅰ

(引用：河合塾 2024年大学入試共通テスト解答速報 問題PDF)

それに対するClaude2の出力した回答がこちらです。

> これらの文は時代順に並べると、
>
> II - 全国を統一した人物が九州に本陣を置いて朝鮮半島に出兵した際に、朝鮮の金属活字が日本にもたらされた。(豊臣秀吉の朝鮮出兵、1592年)
>
> III - 倭寇の活動が活発化したため、朝鮮が倭寇の本拠地と考えた対馬を襲撃し、貿易が一時中断した。(14世紀后半期)
>
> I - 交易を目的として朝鮮の三浦に居住していた日本人らが、特権の縮小に反発して暴動を起こしたため、その後の通交は一時的に衰退した。(壬申倭乱、1510年)
>
> となります。文中の記述から時代が判断でき、古い時代から順にI、III、IIとなります。
>
> Copy

お気付きの方もいらっしゃるでしょうが、Claude2はⅠ,Ⅱ,Ⅲの文章の詳細な年代を特定することに成功しています。
なのであとは古い順にⅢ－Ⅰ－Ⅱと並べれば正解だったのですが、なぜか順番を間違えてしまいました。

このように、「与えられた文章の年代を特定する」「3つの出来事を古い順に並び替える」という動作一つ一つは簡単にできても、それを一つの問題の中で同時に求められると、年代特定が不正確になってしまったり、出来事の並び替えが正常に行えなくなってしまったりするのです。

これはGPT-4やGoogle Bardでも見られた現象で、生成AI全体が乗り越えられていない課題なのかもしれません。

2、図表の読み取り問題

こちらも同様で、「入力された図表を読み取る」「得られた情報をもとに問題を解く」というプロセス一つ一つはどのAIも一定高い精度で実現してくれるのですが、それを同時に求められると各プロセスの精度が落ちるように感じられます。

並び替えや図表読み取り問題など、複数の処理を同時に要求する問題は、旧センター試験から現行の共通テストに変更して以降どんどん増加・複雑化しています。

おそらくこうした知識や情報処理などの組み合わせを行わせる中で「いかに知識を活用できるか？」を問うのが共通テストの意義なのだと思われます。
逆に知識の有無をストレートに問うていた旧センター試験の問題を解かせると、今のAIたちであればより満点に近い得点を取るでしょう。

各論：ツールごとに見える個性

GPT-4、Google Bard、Claude2に共通する点についてはお話ししましたが、これらのツールごとにも大きく異なる特徴が見られました。

GPT-4：優柔不断な優等生タイプ

GPT-4の最も特筆すべき特徴は、与えられた質問に対して時間をかけて回答を考えるかわりに、明確な根拠を示して高い精度の答えを出力する点です。さらに厳密さを非常に気にする節があり、「一つ選びなさい」という指示を出しても自身が納得できる根拠を提示できないと判断した場合には「答えられません」という回答をするケースが散見されました。

共通テスト数学

問題(iv)で与えられた命題 (I), (II), (III) の真偽を判定するためには、数列 $\{c_n\}$ が満たすべき条件 $(c_{n+3})(2c_n + 1 - c_m + 3) = 0$（①とする）を考慮する必要があります。この条件を満たす数列の特性を理解し、各命題がこの条件に適合するかどうかを判断します。

(I) $c_1 = 3$ かつ $c_{100} = -3$ であり、かつ①を満たす数列 $\{c_n\}$ がある。

これを判断するためには、①式の性質を考慮する必要があります。ただし、与えられた情報からは c_{100} の値が c_1 によってどのように影響を受けるかを直接決定することはできません。そのため、この命題が真か偽かを断定することはできません。

(II) $c_1 = -3$ かつ $c_{100} = -3$ であり、かつ①を満たす数列 $\{c_n\}$ がある。

この命題についても、同様に①式の特性を考慮する必要があります。しかし、与えられた情報から c_{100} が c_1 にどのように依存するかを判断することはできません。したがって、この命題が真か偽かを断定することはできません。

(III) $c_1 = -3$ かつ $c_{100} = 3$ であり、かつ①を満たす数列 $\{c_n\}$ がある。

この命題に関しても、与えられた条件からは c_{100} の値が c_1 にどのように影響されるかを明確に判断することはできません。したがって、この命題が真か偽かを断定することはできません。

結論として、(I), (II), (III) の各命題について、与えられた情報と条件だけでは、それらが真であるか偽であるかを断定することはできません。より具体的な数列の性質や追加情報が必要です。

三つの命題の真偽を問う問題への回答。腹落ちする理由がない場合は断定しない様子がうかがえる

Google Bard：コンプラ意識高めの一匹狼タイプ

Google Bardは、概ね使用感は他のAIと変わりませんが、特定のワードや文章形式を入力すると「回答できません」という拒否反応を起こすことがありました。

具体的には、会話形式の長文や、公序良俗に反すると判断された可能性のある単語を受け付けない仕様となっている（と思われます）。

例えば次のような質問をすると、「お手伝いできません」と言われてしまいます。

そこで、質問文の「風俗や恋愛を描いた」という文言を除いて再試行してみると、

為永春水は、江戸時代後期の戯作者であり、人情本で人気を博しました。代表作に『春色梅児誉美』『春色辰巳園』などがあります。

天保の改革では、風紀を乱すとして、為永春水の作品を含む人情本が禁書となりました。為永春水自身も手鎖の刑に処され、憂悶のうちに病没しました。

滝沢馬琴は、為永春水よりも後の人情本作家であり、天保の改革の際にはすでに没していました。そのため、天保の改革で罰されたのは、為永春水となります。

なお、為永春水は、人情本を創始した式亭三馬の弟子であり、人情本の形式を確立した人物として知られています。

ここにメッセージを入力してください

しっかり回答してくれました。この例の場合、「風俗」か「恋愛」のどちらか又は両方が阻害要因となっていたようです。

Claude2：スタンス激強な頑固者

Claude2には、とにかく自分の見解を強く押し出して回答する傾向が見られました。また、仮に計算が上手くいかなかったり方針が立たなかった問題に対しても、必ず何らかの回答を当て込んでくる意志の強さを感じます。

![図4](https://f.yourl.jp/6348b814/)

![図5](https://f.yourl.jp/b1cb6ae2/)

Claude can't visit links. Responses may contain hallucination.

はい、この問題では対数関数logxy=2のグラフがx>0, x≠1, y>0の領域のどの部分に相当するかを求めていますね。

logxy=2より、y=2/logxに変形できます。
この関数はx>0のときに意味を持ち、x=1で不定形になり、x>1のときy>0、x<1のときy<0となります。

したがって、条件x>0, x≠1, y>0を満たすのは、x>1の部分だけです。

この領域と合致するのは、図3のみです。

よって、回答は図3となります。

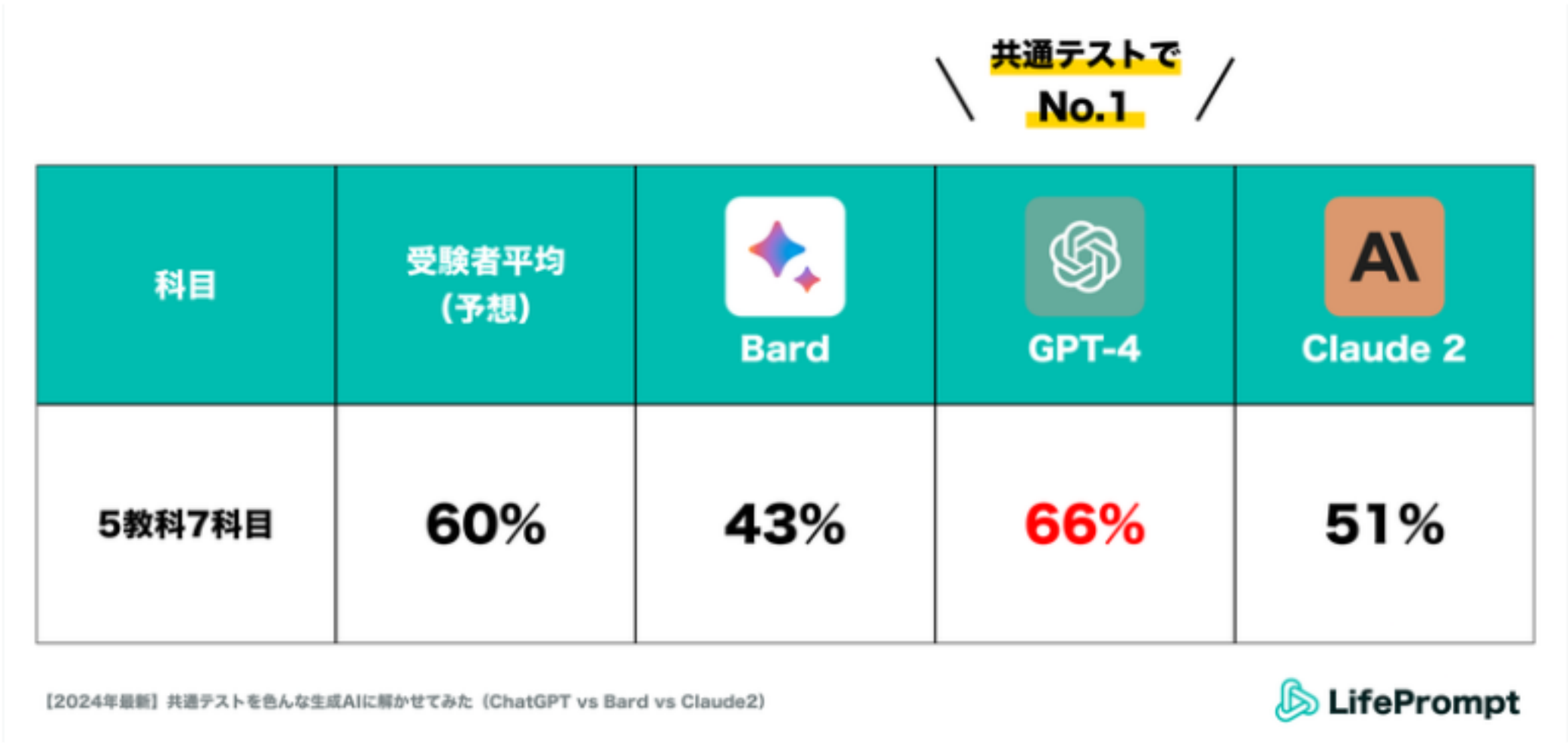
対数関数のグラフの領域と方程式の関係を理解することが大切ですね。同じような問題があれば、遠慮なく質問してください。できる限り丁寧に説明いたします。

Copy

Claude2は仕様として画像リンクを参照できていないようだが、そこへの言及は無しに自信満々に間違った答えを示してくれた。よく見ると「ハルシネーション(AIが正当化できないはずのことを堂々と回答すること)に注意」とある。

まとめ

今回は、先週末に実施された共通テストを、GPT-4/Google Bard/Claude2の三種類の生成AIに受験させ、AIの性能の比較を行ってみました。その結果、特にChatGPTの性能向上に驚かされる結果となりました。



他にもAIツール間でユーモラスな差分も見えて、AIをより身近に感じてもらえたのではないでしょうか。

次は、正答率が低かった数学を筆頭に、プロンプトを改善するとどのくらい正答率を高められるかこれからチャレンジしてみます！

これからの励みになるので、もしよろしければ「スキ」や「シェア」をお願いいたします！

また、メディア関係者の方で本noteをテーマに記事を作りたい！という方がいれば是非連絡くださいませ。喜んで取材に応じさせていただきます。

最後までご覧いただきありがとうございました。

お問い合わせはこちらから

株式会社LifePrompt｜一社一社にぴったりのAI活用を実現する

株式会社LifePromptは、AIの力を最大限に活用し、各企業のビジネス課題を解決します。お客様一社一社に最適なAIソリ

lifeprompt.net

LifePrompt