農業検定試験問題を用いた大規模言語モデルの性能評価

Benchmarking Large Language Models Using Agricultural Certification Exam Questions

Yosuke Toda, Hiroki Kawai

*戸田陽介[1,2,5]　*河合 宏紀[3,4,5]

1 ㈱フィトメトリクス, 静岡県浜松市; 2 名古屋大学トランスフォーマティブ生命分子研究所, 愛知県名古屋市; 3 東京大学大学院工学系研究科, 東京都文京区; 4 エルピクセル㈱, 東京都千代田区; 5 ㈱ポケトク, 愛知県名古屋市

* 共同責任著者

戸田: yosuke{at}phytometrics.jp　河合: h.kawai{at}g.ecc.u-tokyo.ac.jp

農業分野における大規模言語モデル（LLM）の実用可能性を評価するため、日本語で記述された農業検定1級（2023年度）の4択問題全70問を対象に、近年のLLM 27種に対する予備的なベンチマーク評価を実施した。最高正答率は85.7%に達し、合格基準である70%を大きく上回るモデルも複数確認された。LLMが農業分野における専門的な知識理解・推論において実用域に到達しつつあることが示唆されたと同時に、知識の偏在や構文解釈の限界といった課題も浮き彫りとなった。分野特化型の知識強化や体系的な日本語ベンチマークの整備が、農業用LLMの高度化に向けた鍵となる。

背景

近年、大規模言語モデル(Large Language Model; LLM)の性能は飛躍的に向上しており、自動要約・文書分類・質問応答など、従来のモデルでは難しかった自然言語処理のタスクで高い精度を示すようになっている。こうした進展を背景に、LLMが特定の専門分野においても通用するかについての期待が高まりつつある。専門分野では一般的なコーパスに含まれない用語や領域特有の概念、あるいは最新知見に関する知識が必要とされることが多く、その網羅性や正確性は未知数である。また、評価そのものの難易度が高い要因として、ドメイン特化型のベンチマークデータセットが十分に整備されていない現状が挙げられる。例えば、医学分野ではMedQAベンチマーク(Jin et al., 2021)が評価基準として一定の地位を得ている一方で、金融や農業など他の専門分野では、研究者間で標準的に利用できるベンチマークの数がまだ限られている。このような背景から、特定分野でのLLM評価を正確に行うためのデータセットの整備や、専門知識の組み込み手法の開発が急務となっている。

農業分野においては、近年、試験的な取り組みがいくつか公開され、この分野におけるLLMの性能評価を推進する動きが始まっている。例えば、AgXQAは農業に特化した質問応答データセットであり、LLMの農業知識に関する理解力を評価することを目的としている(Kpodo et al., 2024)。一方で、AgEvalは、植

物のストレス表現型評価に焦点を当てたデータセットである (Arshad et al., 2024)。さらにまた、CROP Benchmarkは、作物科学分野でのLLMの有用性を包括的に評価するために設計されている(Zhang et al., 2024)。一方で、AgriBenchは、文章だけでなく、画像処理などを含む農業分野をテーマとしたマルチモーダルLLMを対象とした、データセットである。風景画像やセグメンテーションマスク、深度マップといった多様なデータ形式を用いて、現実世界における場面を想定したデータセットとして整備されている(Zhou & Ryo, 2024)。

もっとも、これらの既存ベンチマークの多くは、日本語圏以外の情報を対象とした英語での記述を基本とするデータセットであり、日本の農業知識や実務に即した日本語LLMの性能評価には十分でない。農業知識は国や地域によって制度・品目・栽培環境などが大きく異なるため、前提となるコーパスも異なる。したがって、日本語を用い、日本の農業実情に即したベンチマークの整備と評価が望ましい。例えば、J-AgriQAは、日本の農業農村工学分野に焦点をあてた、農業水利施設に関する日本語データセットとして整備が進められている(Hakoishi et al., 2024)。

本試行では、LLMの農業分野における専門知識活用能力を評価するため、日本独自の試験制度である「農業検定」の問題を活用した。農業検定は、栽培・農業全般・環境・食の4分野にまたがる基礎知識の習得を目的とした全国規模の検定試験であり、難易度別に1級から3級の試験区分が用意されている。そのうち、1級は農業の総合的な知識が問われる上級レベルで構成されている。本研究では、2023年度の1級試験問題70問をQAベンチマークとして採用し、27種のLLMに対して一律に評価を行った。このアプローチにより、一般的な自然言語処理タスクでは見えにくい、地域性・制度性・作物多様性といった日本の農業固有の文脈におけるLLMの実力を明らかにすることを試みた。

方法

本研究では、日本の農業知識に即したQAデータセットとして「日本農業検定」の試験問題（全国農協観光協会; https://nou-ken.jp/）を使用し、各LLMの性能を評価した。すべての設問は4択形式で、正答あるいは誤答を1つ選択する形式となっている。合格基準点は正答率70%であることが公表されている。

本試行では、2023年度の農業検定1級の全試験問題70問と正答のペアを抽出し、各LLMに対して一律の形式で評価を実施した。プロンプトとしては、「4択の中から該当する数字を1つ選び、続けて理由も述べよ」といった指示文を原則として与えた。ただし、モデルごとに応答形式の違いや制約があるため、プロンプト文はモデルごとに微調整した。生成される答えの再現性を高めるため、各LLMのtemperatureは0もしくは可能な範囲で最も低く設定した。

評価対象には、主要なAPI提供型LLM（クラウド型LLM）およびローカル実行可能なLLM（オンプレミス型LLM）を含めた。クラウド型LLMモデルは、無料・有料問わず、API経由で利用可能なものに限定した。オンプレミス型LLMは各種のLLMを簡便に実行することのできるtool/libraryであるOllama(*Ollama*, n.d.)を用いて実行し、モデルサイズに応じて量子化形式を選択した。具体的には、大規模モデルには4bitまたは8bitの量子化モデルを、それ以外のモデルにはfloat16形式を使用して推論を行った。

さらに、各モデル間の出力の類似性を評価するため、全モデルの選択肢出力（数値）をもとに、ペアごとの一致率を計算した。具体的には、全70問に対する出力結果を比較し、各モデルペア間において、同一の選択肢を選んだ問題の割合を「一致率」として定義した。ただし、モデル間の一致率は、単純に正答率の高い問題では一致しやすくなるというバイアスがある。そこで本研究では、各問題の正答率を

用いた重み付け補正を行い、特に誤答が分散しやすい難問での出力傾向の違いがより反映されるようにした。具体的には、各問題について、モデルペアの出力が一致した場合に1、異なった場合に0を与え、それを「1 + 正答率」で割ることで、難問ほど一致の寄与が大きくなるよう調整した。これにより、単なる多数派選択への収束ではなく、難問における出力の一致傾向をより正確に反映した補正後一致率マトリクス（adjusted similarity matrix）を作成した。最終的に、この補正後マトリクスをもとに、全モデル間の正答率補正付き出力類似度を可視化し、クラスタリングを通じて出力傾向の近いモデル群を探索した。

本論文の執筆にあたって文章の校正作業にChatGPT-4oを用いた。用語集の作成には、本原稿をChatGPT-4oに読み込ませた後、情報科学分野で使われる用語のうち、農学者や植物科学者が必要と想定されるものを抽出させ、その意味を出力させた後、執筆者らが校正を行った。

結果

全問題に対して、27種のLLMを用いた一律評価を実施し、モデルごとの正答率を算出した（図1）。結果として、最高正答率は85.7%を記録し、農業検定の合格基準である70%を上回るモデルが複数存在した。最も高い正答率を記録したのは Gemini 2.5 Pro (gemini-2.5-pro-preview-03-25) であり、85.7%の精度を示した。これに Claude 3.7 Sonnet (claude-3-7-sonnet-20250219)（82.9%）、DeepSeek R1（81.4%）、および別モデルの Gemini や Claude が続き、おおむね80%以上の正答率を示した。GPT-4o (GPT-4o-2024-08-06) や Qwen2.5 72B (qwen2_5_72b-instruct-q4_K_M)は70%前後と合格基準点付近の結果を示した。一方で、phi3.5-mini (phi3.5_3.8b_mini-instruct-fp16) やgemma 2 9b (gemma2:9b-instruct-fp16) など一部のモデルは30〜40%台にとどまり、そのままでは農業分野の知識応答には対応しきれていないことがわかった。

栄養繁殖に関する誤りを見出す設問を一例として、回答理由をLLMから抽出し、その思考過程を可視化した（補足図1）。Gemini 2.5 Pro、Claude 3.7 Sonnet、DeepSeek R1 などの上位モデルがいずれも正答である選択肢①（「栄養繁殖は有性生殖」とする誤記）を正確に指摘し、基本的な生物学的知識を踏まえて正しく応答した。一方で、phi3.5-mini のような正答率の低いモデルは④を誤って選択するなど正答には至らず、また、農業用語を踏まえた文章の生成に難を示した。それでもなお、「ウイルス感染のリスク」や「親株の性質の継承」など、関連するコンテクストを部分的に把握している様子が見受けられた。

次に、出題内容について、日本農業検定が提示する公式な知識領域区分に基づき、各設問に対して「主要作物・園芸作物の栽培知識」「作物生理と栽培環境」「環境」「農業全般」「食」の5種別タグを付与し、それぞれのカテゴリごとの正答率を算出した（図2）。興味深いことに、LLMを問わず、「主要作物・園芸作物の栽培知識」タグに該当する設問に対する正答率が、他カテゴリのそれと比べて一定程度（10%~20%）低い傾向が見られた。この結果は、特定作物の品種特性や栽培慣行など、地域性や実務性の高い知識領域において、現行のLLMが相対的に苦手とする傾向を示唆している。

特筆すべき点として、全27種のLLMにおいて正答率が0%であった設問が「主要作物・園芸作物の栽培知識」カテゴリにおいて1問存在した。これは、キュウリの特性に関する記述のうち、「華北型が黒いいぼ、華南型が白いいぼを有するキュウリである」という誤りを見出す問題である（補足図2）。正解選択肢1を選択するために必要な背景知識「華北型が白いぼ、華南型が黒いぼ」という記述が、日本および中国の農業知見として比較的よく知られている一方で、英語圏では一般的でない分類であり、学習コーパスにおいて十分にカバーさ

れていない可能性がある。言い換えると、知識として存在していても、それが明確に構造化され、言語的に関連付けられていなければ、LLMは正確な判断を下すことが難しいと考えられる。さらに、他の誤答理由を抽出すると、昼夜の温度差に関する記述に対して「温度差を大きくすると作物にとってストレスになる」という部分的かつ定性的な知識に基づき誤答選択肢3を選択したり、「キュウリは緑黄色野菜に分類される」といった単純なハルシネーションにより4を選択したケースも見られた。これらは、LLMが農業知識を単純な知識ベースとしてではなく、定性的な経験則や一般化された知識の干渉によって引き起こされたことが示唆された。

最後に、農業分野をテーマとした問に対するLLM同士の出力の類似性を調べるため、モデル間で同一の選択肢を選んだ割合を算出し、正答率に応じた重み付けを加えた一致度を導入した（正答率が高ければ一致度が高くなるということを避けるため）。図3に、補正後のモデル間一致度を可視化した。

gemini-2.5-pro、claude-3.7-sonnet、DeepSeek R1など、正答率上位のモデルは、正答率の補正を考慮しても、互いに高い一致度を示し、クラスタリングにおいても近い位置に配置された。難問における判断傾向にも共通性があり、具備する知識や推論アルゴリズムにおいて類似性を持つ可能性が示唆された。

また、Claude系（Claude 3.5, 3.7）やGemini系（Gemini 2.5 Pro, Gemini_exp_1206）など、開発系列が共通するモデル群では、系列内での一致度が特に高く、学習データやチューニング方針の類似性が応答傾向にも反映されていると考えられる。一方で、OpenAIのChatGPTモデル（4o, o1, 4.5）は、それらと比べ応答性の差異が際立った。この傾向はオンプレミス型LLMでも同様であり、qwen2.5-32b-instruct、qwen2.5-32bに対して日本語を強化する形で多段階の学習等が行われたqwq-bakeneko-32b、qwen2.5-bakeneko-32b-instruct-v2は高一致度を示しているが、同等の正答率であるgemma3-27bとの一致度は高くない。

興味深いことに、クラウド型LLMの中でDeepSeek R1やGPT-o1といった思考型のモデルとの一致率が最も高いオンプレミス型LLMはこちらも思考型モデルであるqwq-bakeneko-32bであり、正答率に開きはあっても近い回答傾向を示した。

まとめると、農業分野に特化した正答率補正後の一致度に基づくクラスタ分析は、LLMの系列的な類似性や、対応戦略の差異を定量的に捉える手法として有効であることが確認された。

考察

農業検定問題に対するLLMの正答率

本試行では、一般用途に設計された大規模言語モデル（LLM）であっても、農業分野において一定の専門知識を応用できる段階に到達しつつあることが示された。実際、最高正答率は85.7%に達し、複数のモデルが農業検定1級の合格基準である70%を上回った。この結果は、汎用的なLLMであっても、日本語による農業関連の問いに対しておおむね妥当な回答を生成できる能力を持っていることを示唆している。

一方で、「合格」という数値的基準を満たしているからといって、すべての場面においてLLMに農業判断を委ねられるかというと、慎重な評価が求められる。例えば、医療においては決して行ってはならない行為というものが存在する。そのため、医師国家試験においては、禁忌肢と呼ばれる、一定数選択すると他の成績によらず不合格となる選択肢が存在する。これによって数値的な基準だけではない評価を行う。実際に医療LLMにおいては成績が基準を越えていても、禁忌肢を選んでしまうこと

があることが確認されている(Kasai et al., 2023)。同様に、農業現場においても、判断の誤りが収量や品質、安全性に直結することも多く、LLMの一部の誤答や曖昧な理由説明は、実務にそのまま適用するには不十分な場合がある。一般に、判断に対応する結果が明らかとなるまでが長期間に及び、さらには、その結果が複合的な要因によることが多く、判断が適切であったかというフィードバックが困難である。一方で、人による判断も必ずしも正しいとは限らず、その違いは「誤った判断に対する責任の所在をどこに求めるか」という点に集約される。LLMによる判断は高精度であっても、最終的な意思決定において責任を持てる主体として扱うには、まだ越えるべき課題が多い。

現時点でのLLMは、農業における意思決定を支援する「補助的ツール」としての活用が現実的であり、人間による検証や判断を伴う形での導入が望ましいと考えられる。それでもなお、人によってキュレーションされた農業ナレッジベースを活用したRetrieval Augmented Generation（RAG）や、インターネット検索を通じて外部情報に基づいた推論を行うGrounded Generationなどの活用により、LLMの補助ツール（究極的には人間の判断の完全な代替）としての信用性・有用性はさらに高まると考えられる。今回合格点に到達しなかったモデルについても、基本的な農業用語を理解していたことを踏まえると、それら技術の適用による活用の可能性は十分に残されている。

現評価手法の技術的限界

本研究には、いくつかの技術的限界が存在する。1. 多くのLLMは訓練データや学習手法が非公開であり、モデル間の厳密な比較および性能の差の分析は困難である。さらには、2. 他モデルの出力を学習に利用した例も近年報告されており、性能評価における独立性が担保されていない可能性がある。3. 農業検定の過去問は問題と解答が原則として別PDFで提供され、解答理由の記載はないため、正答がそのまま事前学習データに含まれている可能性は低いものの、依然としてデータリークの可能性は否定できない。今後は、最新の過去問や2023年以前の問題を用いた正答率の経年変化の分析、ならびに自由記述形式や理由生成を含む多面的評価の導入が求められる。

最後に

近年、OpenAI社のChatGPT-o1を2025年の大学入学共通テストや東京大学の二次試験問題を解かせた結果、合格圏内の性能を有することが報告された(株式会社LifePrompt, 2025)。これは、「ロボットは東大に入れるか」プロジェクトにおいて、AIが東大合格は困難であると結論付けられた2016年の報告(国立情報学研究所, 2016)と対照的であるとともに、この1世紀における自然言語処理の進歩が、高度な知識や文脈理解を含む複雑なタスクにおいて、専門家の想定を超えるスピードで実用域に達しつつあることを示している。

より多様で高難度な問題群によってLLMの限界を検証する試みとして、100以上の分野から2,500問の難問を集めた新しいマルチモーダルベンチマーク「Humanity's Last Exam」が提案された(Phan et al., 2025)。数学や人文科学、自然科学など幅広いドメインをカバーする公開データと非公開テストデータセットが作成され、LLMの性能を見極める設計となっている。2025年4月時点では、最も成績が良いGemini 2.5 Proで18.2%の回答精度であり、これは現状のLLMと専門家レベルの学力との間に大きなギャップが未だあることを示している。ところで、本評価においてもGemini 2.5 Proが正答率1位であったことは注目に値する。この一致は、同モデルが非常に広範かつ困難な問題群に対しても、また特定言語・特定分野に特化した問題に対しても、一貫して高い汎用性と適応力を示していることを示唆している。

本評価では、日本語で記述された農業検定1級の問題を題材としてベンチマーキングを行った。特に農業は地域性が強く、作物や制度、気象条件が国・地域ごとに大きく異なるため、日本固有の農業知識に基づく、専門家によるキュレーションされた評価データセットの構築が重要である。農業検定問題に限らず、学会や自治体が保有する論文、技術報告、各種統計データなどのローカル知識資源の価値も、再評価されるべき段階に来ている。

用語集

用語	英語表記	解説
LLM	Large Language Model	大規模言語モデル。膨大なテキストデータをもとに自然言語を生成するAIモデル。
ベンチマーク	Benchmark	モデルの性能を定量的に比較・評価するための標準化されたテストや指標。
評価	Evaluation	モデルの性能を測定・比較するプロセス全体。
質問応答	Question Answering	ユーザーの問いに対して自動で適切な答えを返す自然言語処理の一種。
プロンプト	Prompt	モデルに与える入力文。質問や指示、文脈などを含む。
推論	Inference	学習済みモデルを使って新しいデータに対して予測を行う処理。
推論過程	Inference Process	モデルが答えを導出するまでの内部的な思考・処理の流れ。
temperature		LLMが出力を生成する際のランダム性を制御するパラメータ。低いほど決定的、高いほど多様性がある出力をLLMが行う。
コーパス	Corpus	モデル学習や評価に使われる大規模なテキストデータの集合。
RAG	Retrieval-Augmented Generation	外部知識を検索して活用する生成モデル手法。
ナレッジベース	Knowledge Base	外部に構築された事実・知識の集約。RAG等で使用される。
Grounded Generation		インターネット検索結果など「根拠に基づいた生成」を行う手法。
ドメイン特化	Domain-Specific	特定分野（農業、医学など）に特化した知識や処理。
マルチモーダル	Multimodal	テキストだけでなく画像・音声など複数の情報形式を扱うモデル。
クラウド	Cloud	外部サーバー上でサービスやモデルを提供する方式。
オンプレミス	On-Premise	クラウドではなくローカル環境で運用する方式。
API	Application Programming Interface	外部ソフトからモデルやサービスを操作するための仕組み。
float16		浮動小数点16ビット精度。軽量かつ高速な計算向き。
量子化	Quantization	モデルのサイズを圧縮する手法。bit精度を下げて軽量化。
4bit		量子化モデルの精度単位。低ビットほど小型・高速。
8bit		量子化モデルの精度単位。低ビットほど小型・高速。
GPT	Generative Pre-trained Transformer	OpenAIによるLLMシリーズ。GPT-4, GPT-4oなど。
Claude		Anthropic社によるLLMシリーズ。Sonnet, Haiku, Opusなど。
Gemini		Google DeepMindによるLLM。Gemini 1〜2.5など。
DeepSeek		DeepSeek社のLLM。DeepSeek R1やDeepSeek V3など。

Qwen		Alibaba製のLLM。Qwen2.5など。
phi		Microsoftの軽量系LLM。phi-3, phi-3.5など。
モデルサイズ	Model Size	モデルのパラメータ数（7B, 70Bなど）による規模の違い。
ハルシネーション	Hallucination	LLMが事実ではない誤情報をもっともらしく生成する現象。
ファインチューニング	Fine-tuning	事前学習済みモデルを特定タスクやデータに再学習させる工程。
チューニング	Tuning	モデルのパラメータやプロンプトを調整して性能を向上させる行為。
意思決定支援	Decision Support	モデルの出力をもとに人間の判断を助ける用途。
学習済みモデル	Pretrained Model	既に大規模データで学習されており、すぐ推論に使えるモデル。

引用文献

Arshad, M. A., Jubery, T. Z., Tirtho, R., Rim, N., Singh, A. K., Hegde, A. S. C., Baskar, G., Krishnamurthy, A. B. A., & Soumik, S. (2024). AgEval: A benchmark for zero-shot and few-shot plant stress phenotyping with multimodal LLMs. In *arXiv*. arXiv. https://arxiv.org/html/2407.19617v1

Hakoishi, K., Sugeta, D., Hitokoto, M., Shiono, T., Kudo, A., Omino, S., Ichihara, A., & Yokoyama, K. (2024). Creation of a Japanese Dataset on Agricultural Water Management Facilities and Future Prospects. 農業農村工学会大会講演会講演要旨集, *2024*年度（第*73*回）, 237–238.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences (Basel, Switzerland)*, *11*(14), 6421. https://doi.org/10.3390/app11146421

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., & Radev, D. (2023). Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2303.18027

Kpodo, J., Kordjamshidi, P., & Nejadhashemi, A. P. (2024). AgXQA: A benchmark for advanced Agricultural Extension question answering. *Computers and Electronics in Agriculture*,

225(109349), 109349. https://doi.org/10.1016/j.compag.2024.109349

Ollama. (n.d.). https://github.com/ollama/ollama

Phan, L., Gatti, A., Han, Z., & Li, N. (2025). Humanity's Last Exam. *SuperIntelligence - Robotics - Safety & Alignment*, *2*(1). https://doi.org/10.70777/si.v2i1.13973

Zhang, H., Sun, J., Chen, R., Liu, W., Yuan, Z., Zheng, X., Wang, Z., Yang, Z., Yan, H., Zhong, H.-S., Wang, X., Ouyang, W., Yang, F., & Dong, N. (2024). Empowering and assessing the utility of large language models in crop science. *Neural Information Processing Systems*, *37*, 52670–52722. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5e5783c673cf05cfd4b3ebf46e96abfc-Abstract-Datasets_and_Benchmarks_Track.html

Zhou, Y., & Ryo, M. (2024). AgriBench: A hierarchical agriculture benchmark for MultiModal Large Language Models. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2412.00465

国立情報学研究所大学共同利用機関法人 情報・システム研究機構. (2016). *NII*人工知能プロジェクト「ロボットは東大に入れるか」／センター試験模試*6*科目で偏差値*50*以上. https://www.nii.ac.jp/news/release/2016/1114.html

株式会社LifePrompt. (2025).【東大理*3*合格】*ChatGPT o1*と*DeepSeek R1*に*2025*年度東大受験を解かせた結果と答案分析【採点協力：河合塾】. https://note.com/lifeprompt/n/n0078de2ef36b

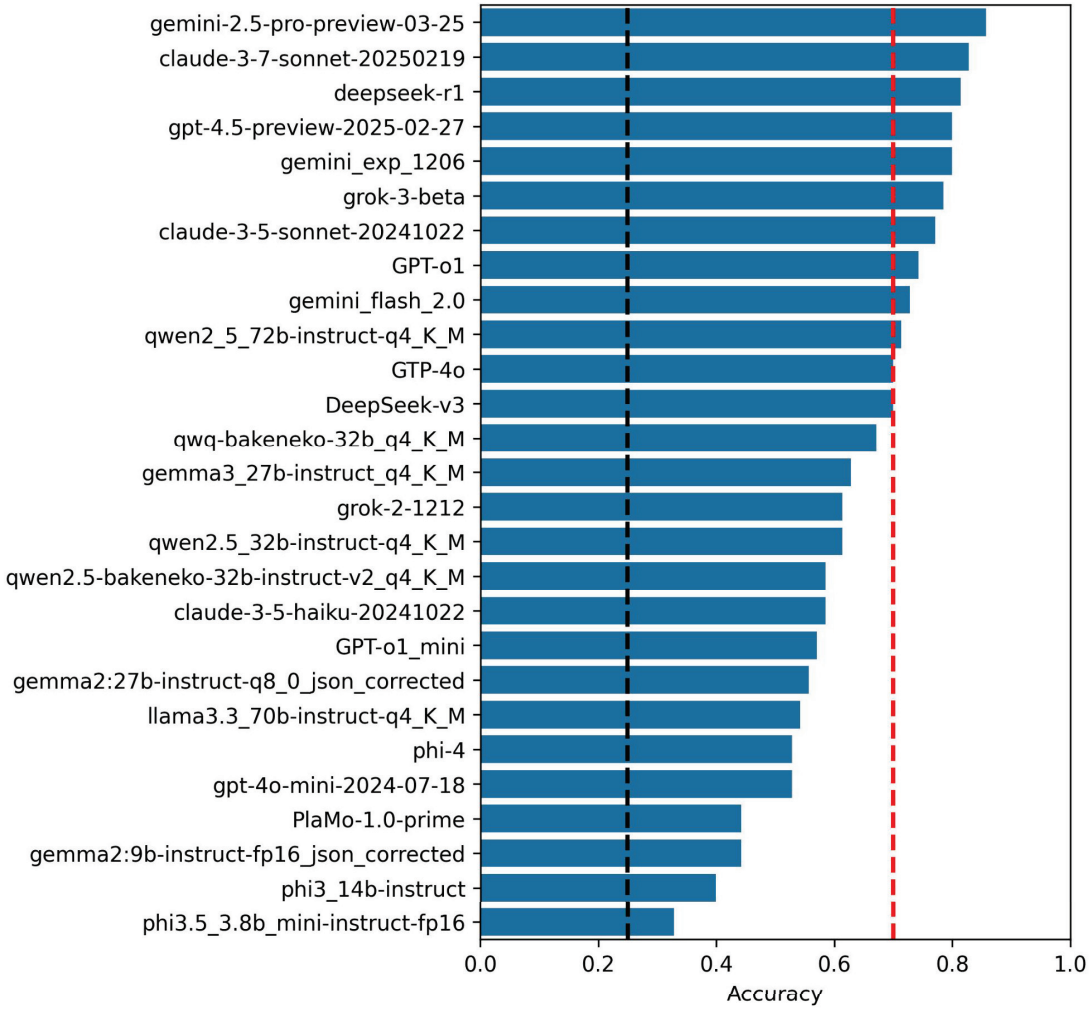

図1｜農業検定問題を対象とした LLMの正答率

各LLM（大規模言語モデル）について、農業検定1級70問に対する正答率を0～100%で棒グラフに示した。人間の合格基準である70%には赤の点線、ランダム選択時の期待値である25%には黒の点線をそれぞれ表示した。

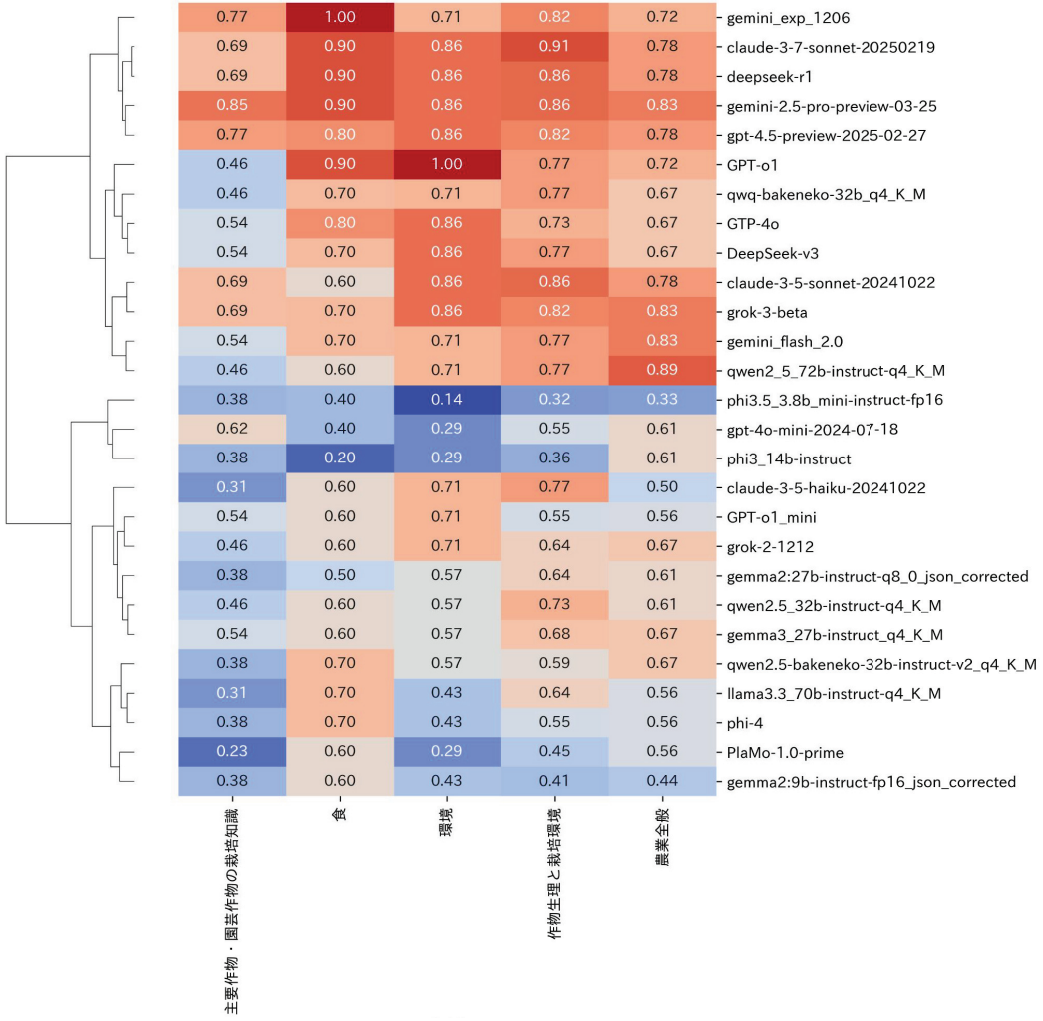

図2｜タグ別正答率ヒートマップ

主要な農業カテゴリ(「主要作物・園芸作物の栽培知識」「作物生理と栽培環境」「環境」「農業全般」「食」)ごとに、各LLMの正答率をヒートマップとして可視化した。階層クラスタリングにより、モデル間の傾向を視覚的に把握可能とした。カラーマップにはcoolwarmを使用し、正答率が高いほど赤、低いほど青に表示した

図3

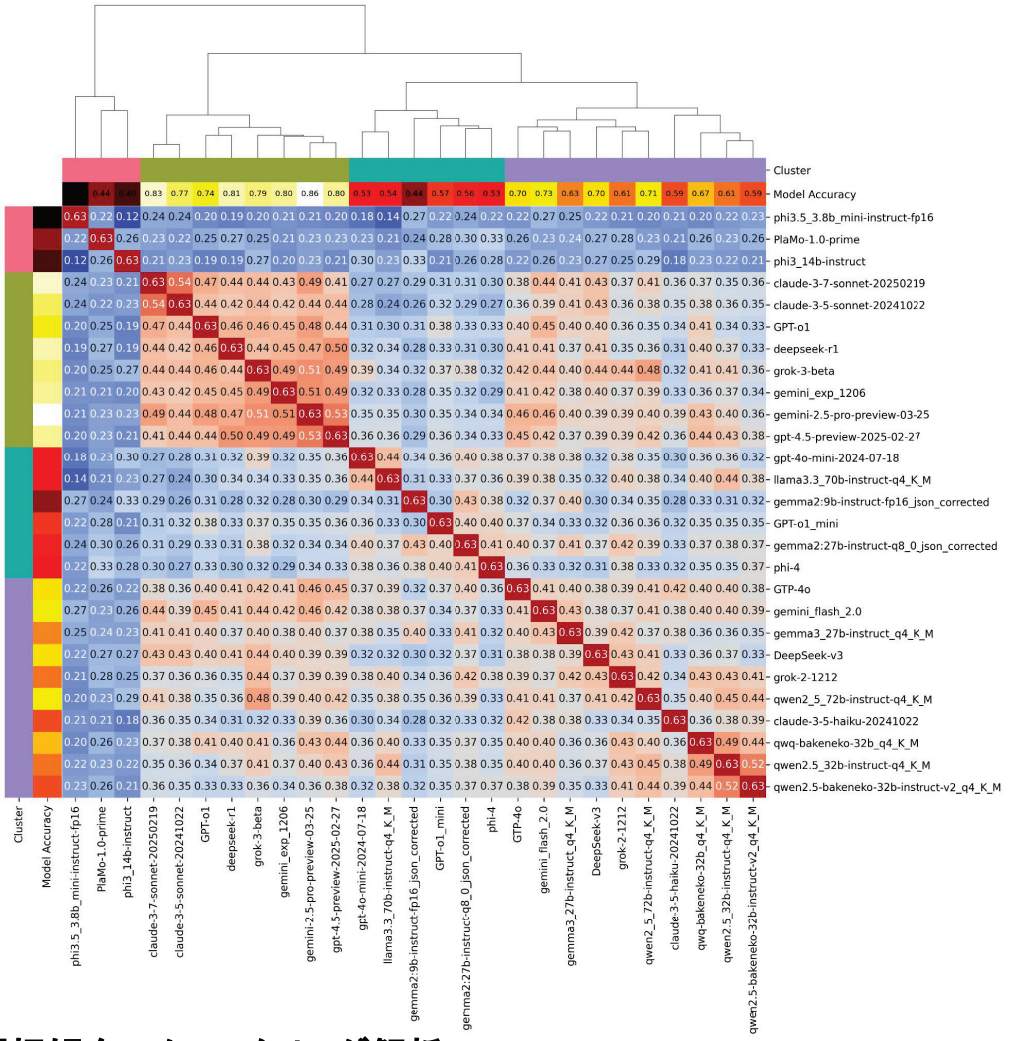

図3｜LLM同士の回答選択傾向のクラスタリング解析

各LLMが同一の選択肢を選んだ割合に基づいてモデル間の類似性を算出し、正答率による補正後の一致度行列としてヒートマップで示した。クラスタリングにより、類似した選択傾向を持つモデル群を色分けして表示した。また、各モデルの正答率に基づいたカラーバーを追加し、モデルの性能と回答傾向の関係を同時に視覚化した。

補足図1

問：栄養繁殖についての説明で、間違っているものは次のうちどれですか。①栄養繁殖は、種子を経由せずに、根・茎・葉などの栄養器官から、次の世代を有性生殖で繁殖させる方法である。②栄養繁殖の長所は、親株と同じ性質の株を早く育てられることであるが、ウイルス病に感染すると、そのウイルスを保持したまま増殖してしまう。③栄養繁殖による繁殖は、野菜・果樹・花き園芸で行われ、挿し木、取り木、接ぎ木、株分けなども栄養繁殖の方法である。④ウイルスフリー苗は、ウイルスに感染していない成長点近辺にある分裂組織細胞を無菌培養した、ウイルスに感染していない苗である。
正答：①栄養繁殖は有性生殖ではなく、無性生殖であるため。

モデル	回答	理由
gemini-2.5-pro-preview-03-25	1	栄養繁殖は無性生殖であり、有性生殖ではないため。
claude-3-7-sonnet-20250219	1	理由: 栄養繁殖は種子を経由せずに根・茎・葉などの栄養器官から次の世代を繁殖させる方法ですが、「有性生殖」ではなく「無性生殖」で繁殖させる方法です。有性生殖は雌雄の配偶子が合体して新しい個体を作る方法であり、種子繁殖がこれに当たります。栄養繁殖は親と同じ遺伝情報を持つクローンを作る無性生殖の一種です。
deekseekR1	1	①の説明は「有性生殖で繁殖させる方法である」とあるが、栄養繁殖は無性生殖に分類されるため。有性生殖では生殖細胞の融合を伴うが、栄養繁殖は親の体細胞から直接新個体が形成される無性生殖の一種である。
qwen2_5_72b-instruct-q4_K_M	1	栄養繁殖は有性生殖ではなく、無性生殖の一種です。種子を経由せずに、根・茎・葉などの栄養器官から次の世代を繁殖させる方法であるため、①が間違っています。
phi3.5_3.8b-instruct-fp16	4	ウイルスに感染すると、そのウイルスを保持したまま増殖することができる点は間違い。正確な情報では親株から同じ性質の子株を育てやすく、但ウイルスに関連した不正確さがあります。

補足図2

問:キュウリについての説明で、間違っているものは次のうちどれですか。①キュウリは品種によっていぼの色が異なり、黒いぼの華北型と白いぼの華南型が代表的である。②果面にブルーム(果粉)がついていると商品価値が低下するため、発生を抑えるためにブルームレス台木に接ぎ木すると良い。③ハウス栽培での温度管理は、昼間は25℃、夕方には15℃前後、夜は13℃くらいとし、昼夜の温度差を大きくすると良い。④果実中の95%以上が水分で、低カロリーのキュウリは果皮は緑だが、緑黄色野菜の部類には含まれていない。
正答:①華北型が白いぼ、華南型が黒いぼであり記述が逆であるため。

モデル	回答	理由
gemini-2.5-pro-preview-03-25	3	大きな昼夜温較差はキュウリの生育にとって最適とは言えず、むしろストレスとなる可能性があるため。
claude-3-7-sonnet-20250219	3	キュウリのハウス栽培における適切な温度管理は、昼夜の温度差を大きくするのではなく、むしろ温度差を小さくすることが望ましいです。昼夜の温度差が大きすぎると植物体に負担がかかり、生育不良や病害の発生リスクが高まります。キュウリは比較的温暖な環境を好み、夜温が低すぎると生育が抑制されます。適切な管理としては、昼間は25℃程度、夜間は18-20℃程度を維持し、急激な温度変化を避けることが推奨されています。
deekseekR1	2	②の説明ではブルームレス台木に接ぎ木するとあるが、ブルームの有無は接ぎ木の台木ではなく品種自体の特性による。ブルームレス品種は品種改良で生まれたもので、台木を変えてもブルーム発生は抑制されないため記述が誤っている。
qwen2_5_72b-instruct-q4_K_M	2	キュウリの果面にブルーム(果粉)がついていることは、商品価値を高める要素であり、発生を抑えるべきものではありません。逆に、ブルームレス台木を使用することで、キュウリ本来の特徴が失われる可能性があります。
Grok2 1212	4	キュウリは緑黄色野菜に分類されます。