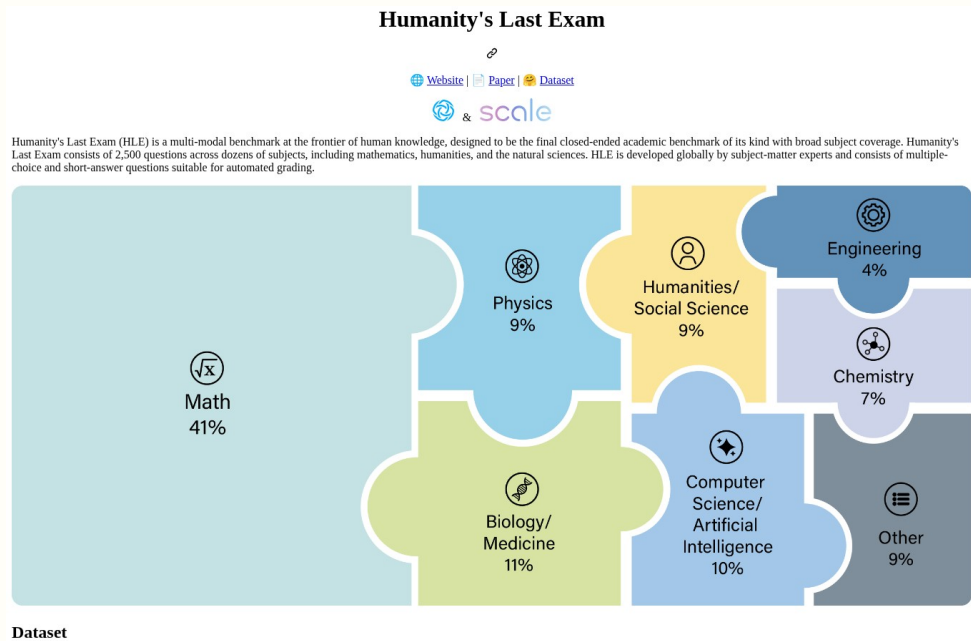


HLE

Humanity's Last Exam

専門知識を問う短答・選択式の評価データ

公式リポジトリ



1. 専門家の難問と実例
2. 採点方法とモデルの成績
3. 作成方法とデータの構成
4. 評価での利用と入手方法

1. 専門家の難問と実例

HLE は答えを検証できる専門分野の難問集

- Humanity's Last Exam (HLE) は、専門家が作った学術的な質問で、モデルの知識と推論を測る評価データである。
- 原著の公開版は 2,500 問。文章だけの問題に加え、図や画像を読む問題を含む。
- 入力は設問と必要な画像、出力は短い答えか選択肢である。

既存の試験が高得点になり、難問を集めた

- 作成時には、幅広い分野の選択式試験 MMLU で主要モデルの正解率が 90% を超え、モデル間の差を測りにくくなっていた。
- そこで、当時のモデルが解けず、人間の専門家が答えを確認できる質問を世界各地から集めた。

短答式が中心で、画像付き問題も含まれる

切り口	原著の構成
答えの形式	短答式 約 76% / 選択式 24%
入力形式	文章のみ 約 86% / 画像付き 約 14%
専門分野	数学・自然科学・人文社会など 100 超
公開版の規模	2,500 問

公開問題と非公開テストを用意している

- Phan ら（2026）は、共通の評価に使える 2,500 問を公開し、それとは別に非公開のテスト集合を保持している。
- 非公開集合は、公開問題に対する過学習や、ベンチマークに合わせた対策の影響を確認するために使う。
- 本資料の実例は、論文で一般公開された 6 問から選んでいる。本編で 3 問、Appendix で残りの 3 問を扱う。

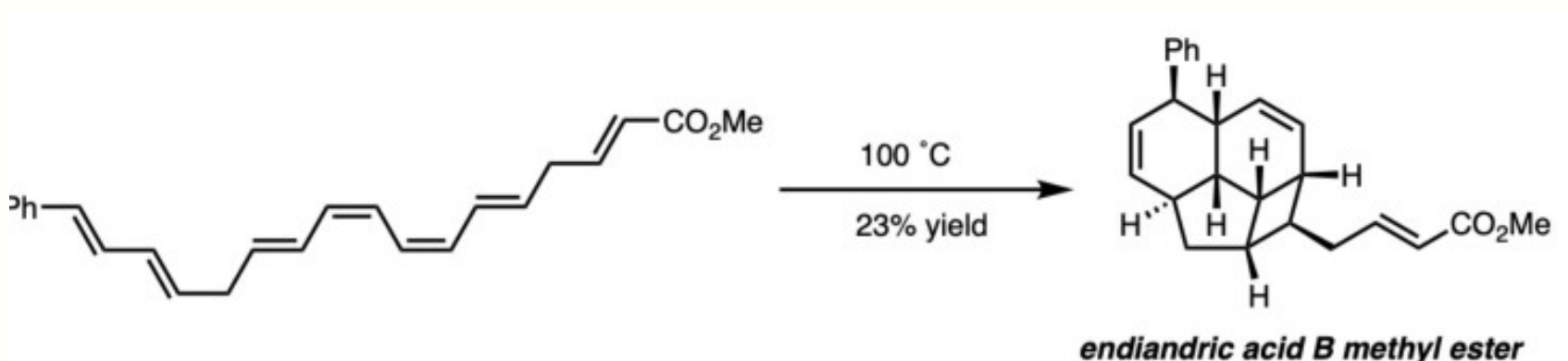
実例 1：ハチドリの骨が支える腱の数

原文（公開論文の設問）

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

質問の意味：ハチドリの尾の筋にある特定の種子骨について、その骨が支える「対になった腱」の数を数字で答える。

実例 2 : 連続反応の種類を答える



実例 2 では、反応図と化学の規則を結び付ける

- 出発物質のヘプタエンから、endiandric acid B methyl ester へ至る熱反応を読む。
- 二つの電子環状反応と、その後の環化付加について、電子数と回転方向、反応する成分の原子数を答える。
- 指定形式は「 $[n\pi]$ -con / $[n\pi]$ -dis」 「 $[m+n]$ 」。

実例 3 : 碑文に含まれる文字を訳す

Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script.

A transliteration of the text is provided: RGYN^o BT HRY BR ^cT^o HBL

実例 3 では、文字の認識と古典語の知識を使う

- ローマ時代の墓碑の図を見て、パルミラ文字の部分を翻訳する。設問にはその転写も与えられる。
- 転写を手掛かりに、語の意味、文法、人名の関係を解釈する。

2. 採点方法とモデルの成績

答えの同値性を判定して正解率を求める

- モデルは説明、最終回答、0～100%の確信度を出力する。採点用モデルが最終回答を取り出す。
- 採点用モデルは、数式の同値な表現や数値の小さな誤差も扱い、参照解答との一致を判定する。
- 原著の評価では o3-mini-2025-01-31 を採点に使った。

正解率は採点された正解数を全問数で割る

説明用の問題	参照解答	モデルの最終回答	得点
問題 A	0.5	1/2	1
問題 B	7	8	0
問題 C	D	D	1

3 問中 2 問が正解。正解率は約 66.7%。

確信度が正解率に見合うかも調べる

- 「90% の確信がある」と答える問題群で実際の正解が10% なら、強く過信している。
- 原著は、正解率に加えて RMS calibration error を報告する。予測した確信度と実際の正解率のずれを測る指標で、低いほどよい。

登場時のモデルはツールなしで低得点だった

モデル（公開年）	正解率	較正誤差
GPT-4o（2024）	2.7%	89%
Claude 3.5 Sonnet（2024）	4.1%	84%
Gemini 1.5 Pro（2024）	4.6%	88%
o1（2024）	8.0%	83%

Phan ら（2026）の評価。較正誤差：RMS calibration error。

2026 年の最新モデルも HLE で評価される

モデル	正解率	条件
Claude Opus 5.5	67.7%	ツールあり
Claude Fable 5.1	65.6%	ツールあり
GPT-6 Astra	57.2%	ツールあり

Opus 5.5 : adaptive thinking、max effort。

ツール利用の有無で正解率は変わる

Gemini 3.1 Pro （ High ）	HLE 正解率
ツールなし	44.4%
検索（ブロックリスト）＋コード	51.4%

評価対象： full set （ text + multimodal ）。

3. 作成方法とデータの構成

世界各地の専門家が質問を投稿した

- 原著では、約 1,000 人の専門家が、50 か国・500 超の機関から参加したと報告している。
- 参加者は主に大学教員、研究者、大学院修了者で、各分野の専門性を使って難問を作成した。
- 質問は英語で、明確で解けること、答えを短く確認できること、検索だけで容易に見つからないことが求められた。

投稿には解答と理由、作成者情報を付ける

提出項目	目的
設問・必要な画像	モデルへ与える入力を明確にする
参照解答・解答形式	何を正解とするか決める
詳しい解答の根拠	専門家が妥当性を確認する
専門分野・氏名・所属	担当分野と責任の所在を示す

モデルによる難度確認と人手審査を重ねた

段階	原著の規模・処理
投稿前の難度確認	70,000 回超の試行
専門家レビューへ進んだ質問	約 13,000 問
レビュー・公開後の改訂	設問と解答を修正・選別
公開版として説明される集合	2,500 問

投稿時に当時のモデルが解けないか確かめる

- 初期の難度確認には GPT-4o、 Gemini 1.5 Pro、 Claude 3.5 Sonnet、 o1 などを使った。文章だけの問題には o1-mini と o1-preview も加えた。
- 短答式は全モデルが解けないこと、選択式は偶然の正解を考慮して一つのモデル以外が解けないことを投稿条件とした。

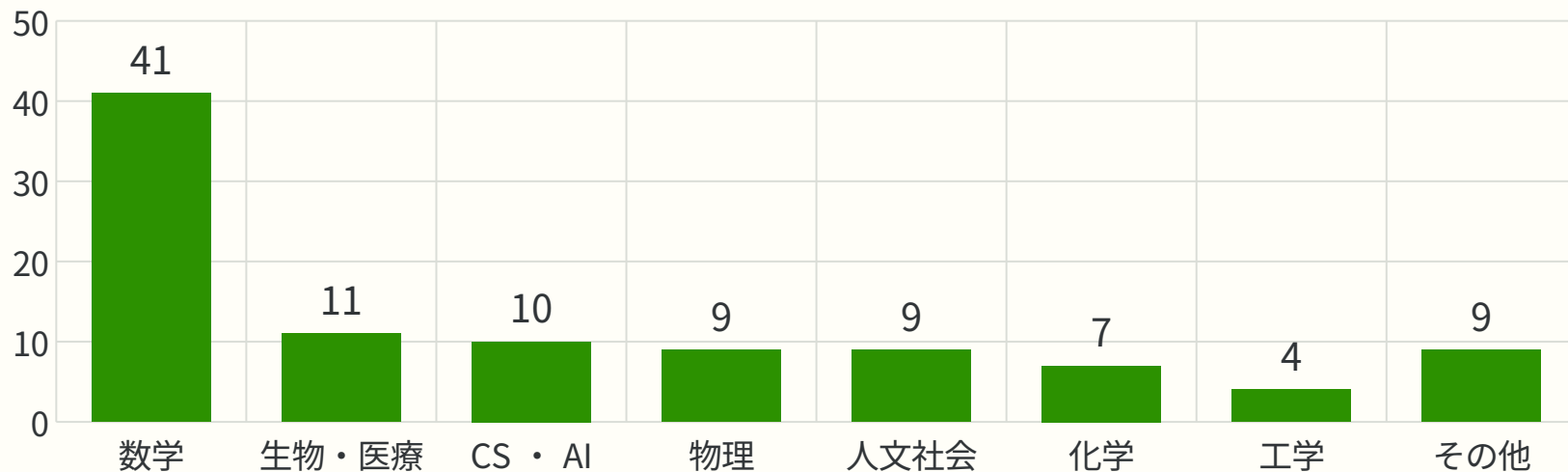
専門家が二段階でレビューした

- 第一段階では、関連分野の大学院学位を持つ査読者が、設問ごとに1～3件のレビューを行い、作者が修正する。
- 第二段階では、主催者や訓練された専門家が、基準を満たす良質な問題を最終承認する。
- 公開後も、コミュニティの指摘、著者への照会、追加の監査によって誤りや曖昧さを修正した。

賞金と共著資格で専門家の参加を募った

仕組み	原著の内容
賞金総額	500,000 米ドル
上位 50 問	各 5,000 米ドル
続く 500 問	各 500 米ドル
採択された作成者	論文の共著者となる機会

数学が 41% を占め、分野ごとの重みは違う



Phan ら（2026）。分野別の構成比（％）。

レビューを経ても専門家の不一致は残る

- 原著の付録では、200 問ずつの二つの監査と修正を経た公開版の専門家不一致率を 15.4% と推定している。
- 生物・化学・健康の部分集合を対象にした査読では、不一致率は約 18% だった。

4. 評価での利用と入手方法

HLE-Rolling は質問を継続的に更新する

- 原著の公開版は 2,500 問で、HLE-Rolling は問題を継続的に追加・修正する動的な版である。
- 公式リポジトリには、2026-02-20 の再追加・削除・修正と、2026-07-27 の追加・削除の履歴がある。

公式データは利用条件を確認して読み込む

```
from datasets import load_dataset

dataset = load_dataset("cais/hle", split="test")
```

Hugging Face で利用登録が必要な状態。許可されたアカウントで認証してから実行する。

学習への混入と問題の再配布を避ける

- 公式データカードは、評価の完全性を守るため、問題集合を公開共有・再アップロード・配布しないよう求めている。
- 公式 README も、ベンチマークの問題を学習コーパスへ入れないよう明記している。
- 配布物には、学習コーパスから識別して除外するための canary と呼ばれる特有の文字列も付いている。

HLE は専門知識の難問への応答を測る

- 答えが検証できる専門的な質問を集め、短答式と選択式、画像付き問題で知識と推論を測る。
- 当時のモデルが解けない問題を、専門家の投稿と複数段階のレビューによって作った。
- 現在のフラッグシップモデルでは、検索やコード実行を使う評価にも採用されている。

Appendix

1. 専門分野の追加例
2. 評価条件とデータの版
3. 制作機関と参考文献

1. 専門分野の追加例

追加例：余終で定義した変換を数える

Question:

The set of natural transformations between two functors $F, G : C \rightarrow D$ can be expressed as the end

$$\text{Nat}(F, G) \cong \int_A \text{Hom}_D(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$\text{CoNat}(F, G) \cong \int^A \text{Hom}_D(F(A), G(A)).$$

追加例：二つの圏の間の個数を求める

Let:

- $F = B_{\bullet}(\Sigma_4)_{*/}$ be the under ∞ -category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0-simplex $*$ of $B_{\bullet}\Sigma_4$.
- $G = B_{\bullet}(\Sigma_7)_{*/}$ be the under ∞ -category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0-simplex $*$ of $B_{\bullet}\Sigma_7$.

How many natural cotransformations are there between F and G ?

数学の例は、定義を理解して個数を答える

- 自然変換を表す `end` に対応させ、`coend` によって「natural cotransformations」を定義する問題である。
- 対称群に由来する二つの圏 $F \cdot G$ を与え、定義された変換がいくつあるかを尋ねる。

追加例：グラフ上のマルコフ連鎖を定義する

Question:

Let G be a graph. An edge-indicator of G is a function $a : \{0, 1\} \rightarrow V(G)$ such that $\{a(0), a(1)\} \in E(G)$.

Consider the following Markov Chain $M = M(G)$:

The statespace of M is the set of all edge-indicators of G , and the transitions are defined as follows:

Assume $M_t = a$.

1. pick $b \in \{0, 1\}$ u.a.r.
2. pick $v \in N(a(1 - b))$ u.a.r. (here $N(v)$ denotes the open neighbourhood of v)
3. set $a'(b) = v$ and $a'(1 - b) = a(1 - b)$
4. Set $M_{t+1} = a'$

追加例：一様な定常分布を持つ条件を選ぶ

We call a class of graphs $\mathcal{G}$ well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

Which of the following graph classes is well-behaved?

Answer Choices:

- A. The class of all non-bipartite regular graphs
- B. The class of all connected cubic graphs
- C. The class of all connected graphs
- D. The class of all connected non-bipartite graphs
- E. The class of all connected bipartite graphs.

計算機科学の例は、収束するグラフ類を選ぶ

- 各頂点に隣の頂点を一つ選ばせ、その選択を局所的に更新するマルコフ連鎖を定義する。
- どのグラフ類なら、唯一の定常分布へ収束し、その分布が一様になるかを、五つの候補から選ぶ。

追加例：ヘブライ語の閉音節を判定する

I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim

追加例：歴史的な発音研究を参照する

Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

מִן־גַּעְרָתְךָ יְנוּסִים מִן־קוֹל־רָעָמָךְ יִחְפְּזוּ (Psalms 104:7) ?

言語学の例では、ティベリア式発音を扱う

- 詩篇 104:7 のヘブライ語本文について、子音音で終わる閉音節を列挙する。
- ティベリア式の聖書ヘブライ語の発音、shewa の機能、音節末の子音などに関する近年の研究を踏まえるよう指定されている。

公開 6 実例は、専門性の異なる要求を示す

分野	必要な操作	求める答え
生態・解剖	特定の骨と腱の関係を調べる	数値
化学	構造式と反応規則を読む	指定された反応表記
古典学	碑文と転写を解釈する	翻訳
数学	圏と変換の定義を使う	個数
計算機科学	確率過程の条件を考える	選択肢
言語学	歴史的発音に基づき分ける	音節の列挙

2. 評価条件とデータの版

文章のみの問題に対する登場時の成績

登場時のモデル	文章のみ正解率	較正誤差
GPT-4o	2.3%	88%
Claude 3.5 Sonnet	4.3%	83%
Gemini 1.5 Pro	4.6%	87%
o1	7.8%	84%
DeepSeek-R1	8.5%	73%
o3-mini （ high ）	13.4%	80%

Phan ら （ 2026 ） による文章のみの問題での評価。

Phan et al. (2026). Humanity's Last Exam. arXiv:2501.14249v11. <https://arxiv.org/abs/2501.14249>

モデルには回答と確信度を別の欄で求める

Explanation: {your explanation for your answer choice}

Answer: {your chosen answer}

Confidence: {your confidence score between 0% and 100%}

公式の応答形式。採点モデルは Answer を抽出し、参照解答と比較する。

採点では最終回答と参照解答を比較する

- 公式の採点プロンプトは、問題を独立に解かず、抽出した最終回答と参照解答の意味のある差だけを見るよう指示する。
- 回答がない、不明確、非同値などの場合は不正解と判定する。数値は小さな誤差を許す設定になっている。

確信度のずれは、同じ正解率でも異なる

説明用の問題群	平均確信度	実際の正解率	差
10 問の問題群 A	90%	20%	70pt
10 問の問題群 B	30%	20%	10pt

配布データには問題以外の情報も付く

列名	内容
id / question	識別子／問題文
image / image_preview	画像の表現／確認用画像
answer / answer_type	参照解答／解答形式
rationale / rationale_image	解答の根拠／対応画像
raw_subject / category	元の専門分野名／集約カテゴリ
author_name / canary	作成者／学習混入の検出用文字列

公開集合・非公開集合と HLE-Rolling

集合・版	役割と更新
公開集合	Phan ら（2026）が公開する 2,500 問
非公開テスト集合	公開集合への過学習や評価対策の影響を確認
HLE-Rolling	質問の追加・削除・修正を継続する版

Phan et al. (2026). Humanity's Last Exam. arXiv:2501.14249v11. <https://arxiv.org/abs/2501.14249>

Center for AI Safety & Scale AI (2026). HLE rolling changes. <https://github.com/centerforaisafety/hle/blob/main/hle-rolling-changes.txt>

HLE-Diamond のツール付き評価環境

経路	公式手順の指定
実行エージェント	Claude Code、Codex CLI、Gemini CLI など
コード実行環境	タスク用の直接インターネット接続なし
Web 検索	web_search。最大 20 件からフィルター
Web 取得	web_fetch。本文を Markdown で取得
問題の漏えい対策	ブロックリストで検索・取得を制限

3. 制作機関と参考文献

CAIS と Scale AI、各分野の専門家が制作した

- 主催は Center for AI Safety （ CAIS ） と Scale AI 。 Long Phan 、 Alice Gatti 、 Ziwen Han 、 Nathaniel Li らが運営チームに名を連ねる。
- 質問の作成には多数の専門家が参加した。原著は運営チームとデータ作成への貢献者を区別して列挙する。
- Nature 掲載時の書誌では、CAIS 、 Scale AI 、 HLE Contributors Consortium の団体名で引用されている。

参考文献

A benchmark of expert-level academic questions to assess AI capabilities

Center for AI Safety, Scale AI & HLE Contributors Consortium / 2026 / Nature 649, 1139–1146

HLE の収集・審査・構成・登場時のモデル評価を報告する原著。

<https://doi.org/10.1038/s41586-025-09962-4>

Humanity's Last Exam

Phan et al. / 2026 改訂 / arXiv:2501.14249v11

公開実例と、投稿・監査・採点に関する詳細な付録を収録する。

<https://arxiv.org/abs/2501.14249>

参考文献

Humanity's Last Exam: official repository

Center for AI Safety & Scale AI / 2026 / GitHub

推論と採点の実装、版の更新履歴、ツール付き評価の手順を公開する。

<https://github.com/centerforaisafety/hle>

HLE dataset card

Center for AI Safety & Scale AI / 2026 / Hugging Face

配布形式、列定義、利用登録と共有に関する注意を確認できる公式データカード。

<https://huggingface.co/datasets/cais/hle>

参考文献

Claude Opus 5.5

Anthropic / 2026-09-22 / Model announcement

最新の Opus 系列の性能・計算量・評価条件を公表し、HLE のツール付き成績を掲載する。

<https://www.anthropic.com/news/claude-opus-5-5>

GPT-6 Astra: A new generation of intelligence

OpenAI / 2026 / Model announcement

Astra の能力と評価結果を公表し、HLE のツール付き正解率 57.2% を報告する。

<https://openai.com/index/gpt-6-astra/>

参考文献

Gemini 3.1 Pro

Google DeepMind / 2026 / Model performance

HLE のツールなし・検索とコード付きの成績、および評価条件を掲載する。

<https://deepmind.google/models/gemini/pro/>

Evaluating HLE-Diamond with tools

Center for AI Safety & Scale AI / 2026 / Evaluation documentation

コード実行環境、Web ツール、ブロックリスト、提供元の実行環境の指定を記録する。

<https://github.com/centerforaisafety/hle/blob/main/docs/evaluation-with-tools.md>