Title: Humanity's Last Exam

URL: https://en.wikipedia.org/wiki/Humanity%27s_Last_Exam

PageID: 79112677

Categories: Category:2025 in artificial intelligence, Category:Large language models

---

Humanity's Last Exam (HLE) is a language model benchmark consisting of 2,500 questions across a broad range of subjects. It was created jointly by the Center for AI Safety and Scale AI .

Creation

Stanford HAI's AI Index 2025 Annual Report cites Humanity's Last Exam as one of the "more challenging benchmarks" developed in response to the popular AI benchmarks having reached "saturation". [1] The test has been described as the brainchild of Dan Hendrycks , a machine learning researcher and the director of the Center for AI Safety , who stated that he was inspired to create the test after a conversation with Elon Musk , who thought the existing language model benchmarks , such as the MMLU , were too easy. Hendrycks worked with Scale AI to compile the questions. [2] The questions were crowdsourced from subject matter experts from various institutions across the world. [3] [4] The questions were first filtered by the leading AI models; if the models failed to answer the question or did worse than random guessing on the multiple-choice questions, they were reviewed by human experts in two rounds and approved for inclusion in the dataset. The submitters of the top-rated questions were given prize money from a pool of 500,000 U.S. dollars —$5000 for each of the top 50 questions and $500 for the next 500. After the initial release, a "community feedback bug bounty program" was opened to "identify and remove major errors in the dataset". [4]

Composition

The benchmark consists of 2,500 questions in the publicly released set. The paper classifies the questions into the following broad subjects: mathematics (41%), physics (9%), biology/medicine (11%), humanities/social science (9%), computer science/artificial intelligence (10%), engineering (4%), chemistry (7%), and other (9%). Around 14% of the questions require the ability to understand both text and images, i.e., multi-modality . 24% of the questions are multiple-choice; the rest are short-answer, exact-match questions. A private set is also maintained to test for benchmark overfitting . [4]

An example question: [2]

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.

An independent investigation by FutureHouse, published in July 2025, [5] suggests that around 30% of the HLE answers for text-only chemistry and biology questions may be incorrect.

Results

References

External links

Humanity's Last Exam at the Center for AI Safety

Humanity's Last Exam at Scale AI

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani
Noam Shazeer
Aidan Gomez
John Schulman
Mustafa Suleyman
Jan Leike
Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category