Title: Language model benchmark

URL: https://en.wikipedia.org/wiki/Language_model_benchmark

PageID: 79100184

Categories: Category:Benchmarks (computing), Category:Datasets in machine learning, Category:Natural language processing, Category:Software comparisons

Source: Wikipedia (CC BY-SA 4.0).

---

Language model benchmark is a standardized test designed to evaluate the performance of language model on various natural language processing tasks. These tests are intended for comparing different models' capabilities in areas such as language understanding , generation , and reasoning .

Benchmarks generally consist of a dataset and corresponding evaluation metrics . The dataset provides text samples and annotations, while the metrics measure a model's performance on tasks like question answering, text classification, and machine translation. These benchmarks are developed and maintained by academic institutions, research organizations, and industry players to track progress in the field. In addition to accuracy, the metrics can include throughput, energy efficiency, bias, trust and sustainability. [ 1 ]

Overview

Types

Benchmarks may be described by the following adjectives, not mutually exclusive:

Classical : These tasks are studied in natural language processing, even before the advent of deep learning. Examples include the Penn Treebank for testing syntactic and semantic parsing, as well as bilingual translation benchmarked by BLEU scores.

Question answering : These tasks have a text question and a text answer, often multiple-choice. They can be open-book or closed-book . Open-book QA resembles reading comprehension questions, with relevant passages included as annotation in the question, in which the answer appears. Closed-book QA includes no relevant passages. Closed-book QA is also called open-domain question-answering . [ 2 ] [ 3 ] Before the era of large language models, open-book QA was more common, and understood as testing information retrieval methods. Closed-book QA became common since GPT-2 as a method to measure knowledge stored within model parameters. [ 4 ]

Omnibus : An omnibus benchmark combines many benchmarks, often previously published. It is intended as an all-in-one benchmarking solution.

Reasoning : These tasks are usually in the question-answering format, but are intended to be more difficult than standard question answering.

Multimodal : These tasks require processing not only text, but also other modalities, such as images and sound. Examples include OCR and transcription .

Agency : These tasks are for a language-model–based software agent that operates a computer for a user, such as editing images, browsing the web, etc.

Adversarial : A benchmark is "adversarial" if the items in the benchmark are picked specifically so that certain models do badly on them. Adversarial benchmarks are often constructed after SOTA models have saturated a benchmark, to renew the benchmark. A benchmark is "adversarial" only at a certain moment in time, since what is adversarial may cease to be adversarial as newer SOTA models appear.

Public/Private : A benchmark might be partly or entirely private, meaning that some or all of the questions are not publicly available. The idea is that if a question is publicly available, then it might be used for training, which would be "training on the test set" and invalidate the result of the

benchmark. Usually, only the guardians of the benchmark has access to the private subsets, and to score a model on such a benchmark, one must send the model weights, or provide API access, to the guardians.

The boundary between a benchmark and a dataset is not sharp. Generally, a dataset contains three "splits": training, test, validation . Both the test and validation splits are essentially benchmarks. In general, a benchmark is distinguished from a test/validation dataset in that a benchmark is typically intended to be used to measure the performance of many different models that are not trained specifically for doing well on the benchmark, while a test/validation set is intended to be used to measure the performance of models trained specifically on the corresponding training set. In other words, a benchmark may be thought of as a test/validation set without a corresponding training set.

Conversely, certain benchmarks may be used as a training set, such as the English Gigaword [ 5 ] or the One Billion Word Benchmark, which in modern language is just the negative log likelihood loss on a pretraining set with 1 billion words. [ 6 ] Indeed, the distinction between benchmark and dataset in language models became sharper after the rise of the pretraining paradigm.

Lifecycle

Generally, the life cycle of a benchmark consists of the following steps: [ 7 ]

Inception: A benchmark is published. It can be simply given as a demonstration of the power of a new model (implicitly) that others then picked up as a benchmark, or as a benchmark that others are encouraged to use (explicitly).

Growth: More papers and models use the benchmark, and the performance on the benchmark grows.

Maturity, degeneration or deprecation: A benchmark may be saturated, after which researchers move on to other benchmarks. Progress on the benchmark may also be neglected as the field moves to focus on other benchmarks.

Renewal: A saturated benchmark can be upgraded to make it no longer saturated, allowing further progress.

Construction

Like datasets, benchmarks are typically constructed by several methods, individually or in combination:

Web scraping: Ready-made question-answer pairs may be scraped online, such as from websites that teach mathematics and programming.

Conversion: Items may be constructed programmatically from scraped web content, such as by blanking out named entities from sentences, and asking the model to fill in the blank. This was used for making the CNN/Daily Mail Reading Comprehension Task.

Crowd sourcing: Items may be constructed by paying people to write them, such as on Amazon Mechanical Turk . This was used for making the MCTest.

Evaluation

Generally, benchmarks are fully automated. This limits the questions that can be asked. For example, with mathematical questions, "proving a claim" would be difficult to automatically check, while "calculate an answer with a unique integer answer" would be automatically checkable. With programming tasks, the answer can generally be checked by running unit tests, with an upper limit on runtime.

The benchmark scores are of the following kinds:

For multiple choice or cloze questions, common scores are accuracy (frequency of correct answer), precision, recall , F1 score , etc.

pass@n: The model is given n {\displaystyle n} attempts to solve each problem. If any attempt is correct, the model earns a point. The pass@n score is the model's average score over all problems.

k@n: The model makes n {\displaystyle n} attempts to solve each problem, but only k {\displaystyle k} attempts out of them are selected for submission. If any submission is correct, the model earns a point. The k@n score is the model's average score over all problems.

cons@n: The model is given n {\displaystyle n} attempts to solve each problem. If the most common answer is correct, the model earns a point. The cons@n score is the model's average score over all problems. Here "cons" stands for "consensus" or "majority voting". [ 8 ]

The pass@n score can be estimated more accurately by making N > n {\displaystyle N>n} attempts, and use the unbiased estimator 1 − ( N − c n ) ( N n ) {\displaystyle 1-{\frac {\binom {N-c}{n}}{\binom {N}{n}}}} , where c {\displaystyle c} is the number of correct attempts. [ 9 ]

For less well-formed tasks, where the output can be any sentence, there are the following commonly used scores: BLEU ROUGE , METEOR , NIST , word error rate , LEPOR , CIDEr, [ 10 ] SPICE, [ 11 ] etc.

Issues

error: Some benchmark answers may be wrong. [ 12 ]

ambiguity: Some benchmark questions may be ambiguously worded.

subjective: Some benchmark questions may not have an objective answer at all. This problem generally prevents creative writing benchmarks. Similarly, this prevents benchmarking writing proofs in natural language, though benchmarking proofs in a formal language is possible.

open-ended: Some benchmark questions may not have a single answer of a fixed size. This problem generally prevents programming benchmarks from using more natural tasks such as "write a program for X", and instead uses tasks such as "write a function that implements specification X".

inter-annotator agreement: Some benchmark questions may be not fully objective, such that even people would not agree with 100% on what the answer should be. This is common in natural language processing tasks, such as syntactic annotation. [ 13 ] [ 14 ] [ 15 ] [ 16 ]

shortcut: Some benchmark questions may be easily solved by an "unintended" shortcut. For example, in the SNLI benchmark, having a negative word like "not" in the second sentence is a strong signal for the "Contradiction" category, regardless of what the sentences actually say. [ 17 ]

contamination/leakage : Some benchmark questions may have answers already present in the training set. Also called "training on the test set". [ 18 ] [ 19 ] Some benchmarks (such as Big-Bench) may use a "canary string", so that documents containing the canary string can be voluntarily removed from the training set.

saturation: As time goes on, many models reach the highest performance level practically possible, and so the benchmark can no longer differentiate these models. For example, GLUE had been saturated, necessitating SuperGLUE.

Goodhart's law : If new models are designed or selected to score highly on a benchmark, the benchmark may cease to be a good indicator for model quality. [ 7 ]

cherry picking : New model publications may only point to benchmark scores on which the new model performed well, avoiding benchmark scores that it did badly on.

List of benchmarks

General language modeling

Essentially any dataset can be used as a benchmark for statistical language modeling , with the perplexity (or near-equivalently, negative log-likelihood and bits per character, as in the original Shannon 's test of the entropy of the English language [ 20 ] ) being used as the benchmark score. For example, the original GPT-2 announcement included those of the model on WikiText-2, enwik8, text8, and WikiText-103 (all being standard language datasets made from the English Wikipedia ). [ 4 ] [ 21 ]

However, there had been datasets more commonly used, or specifically designed, for use as a benchmark.

One Billion Word Benchmark: The negative log likelihood loss on a dataset of 1 billion words. [ 6 ]

Penn Treebank : The error or negative log likelihood loss for part-of-speech tags on a dataset of text.

Paloma (Perplexity Analysis for Language Model Assessment): A collection of English and code texts, divided into 546 domains. Used to measure the perplexity of a model on specific domains. [ 22 ]

General language understanding

See [ 23 ] for a review of over 100 such benchmarks.

WSC ( Winograd schema challenge ): 273 sentences with ambiguous pronouns. The task is to determine what the pronoun refers to. [ 24 ]

WinoGrande: A larger version of WSC with 44,000 items. Designed to be adversarial to 2019 SOTA, since the original had been saturated. This dataset consists of fill-in-the-blank style sentences, as opposed to the pronoun format of previous datasets. [ 25 ] [ 26 ]

CoLA (Corpus of Linguistic Acceptability) : 10,657 English sentences from published linguistics literature that were manually labeled either as grammatical or ungrammatical. [ 27 ] [ 28 ]

SNLI (Stanford Natural Language Inference : 570K human-written English sentence pairs manually labeled for balanced classification with 3 labels " entailment ", "contradiction", and "neutral". [ 29 ] [ 30 ]

WMT 2014 (Workshop on Statistical Machine Translation): a collection of 4 machine translation benchmarks at the Ninth Workshop on Statistical Machine Translation. The Attention Is All You Need paper used it as a benchmark. [ 31 ]

MultiNLI (Multi-Genre Natural Language Inference): Similar to SNLI, with 433K English sentence pairs from ten distinct genres of written and spoken English. [ 32 ]

CNN/Daily Mail Reading Comprehension Task: Articles from CNN (380K training, 3.9K development, 3.2K test) and Daily Mail (879K training, 64.8K development, 53.2K test) were scraped. The bullet point summaries accompanying the news articles were used. One entity in a bullet point was replaced with a placeholder, creating a cloze-style question. The goal is to identify the masked entity from the article. [ 33 ]

SWAG (Situations With Adversarial Generations): 113K descriptions of activities or events, each with 4 candidate endings; the model must choose the most plausible ending. Adversarial against a few shallow language models ( MLP , bag of words , one-layer CNN , etc). [ 34 ]

HellaSwag (Harder Endings, Longer contexts, and Low-shot Activities for SWAG): A harder version of SWAG. Contains 10K items. [ 35 ] [ 36 ]

RACE (ReAding Comprehension Examinations): 100,000 reading comprehension problems in 28,000 passages, collected from the English exams for middle and high school Chinese students in the age range between 12 and 18. [ 37 ]

LAMBADA: 10,000 narrative passages from books, each with a missing last word that humans can guess if given the full passage but not from the last sentence alone. [ 38 ]

General language generation

NaturalInstructions: 61 distinct tasks with human-authored instructions, and 193k task instances (input-output pairs). The instructions are obtained from crowdsourcing instructions used to create existing NLP datasets and mapped to a unified schema. [ 39 ]

Super-NaturalInstructions: 1,616 diverse NLP tasks and their expert-written instructions, and 5M task instances. [ 40 ]

IFEval (Instruction-Following Eval): 541 instructions to be followed, each containing at least one verifiable constraint, such as "mention the keyword of AI at least 3 times". [ 41 ]

LMArena (formerly Chatbot Arena): Human users vote between two outputs from two language models. An Elo rating for each language model is computed based on these human votes. [ 42 ]

MT-Bench (multi-turn benchmark): An automated version of Chatbot Arena where LLMs replace humans in generating votes. [ 42 ]

MultiChallenge: 273 instances. Each instance is a multi-turn (up to 10 turns) conversation history between two parties, ending with a final user turn containing a requirement/question. Designed to test for instruction-following, context allocation, and in-context reasoning at the same time. Scored by LLM as judge with instance-level rubrics. [ 43 ]

CharXiv: 9292 descriptive questions (examining basic chart elements) and 2323 reasoning questions (synthesizing information across complex visual elements) about 2323 charts from scientific papers. [ 44 ]

Open-book question-answering

MCTest (Machine Comprehension Test): 500 fictional stories, each with 4 multiple-choice questions (with at least 2 requiring multi-sentence understanding), designed to be understandable by a 7-year-old. The vocabulary was limited to approximately 8,000 words probably known by a 7-year-old. The stories were written by workers on Amazon Mechanical Turk . [ 45 ]

SQuAD (Stanford Question Answering Dataset): 100,000+ questions posed by crowd workers on 500+ Wikipedia articles. The task is, given a passage from Wikipedia and a question, find a span of text in the text that answers the question. [ 46 ]

SQuAD 2.0: 50,000 unanswerable questions that look similar to SQuAD questions. Every such unanswerable question must be answered with an empty string. Written by crowd workers. [ 47 ]

ARC (AI2 Reasoning Challenge): Multiple choice questions, with a Challenge Set (2590 questions) and an Easy Set (5197 questions). Designed specifically to be adversarial against models that had saturated SNLI and SQuAD. [ 48 ]

CoQA (Conversational QA): 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. [ 49 ]

WebQuestions: 6,642 question-answer pairs designed to be answerable with knowledge present in the 2013 version of Freebase . [ 50 ]

Natural Questions: 323045 items. Each containing a question that had been searched on Google, a Wikipedia page relevant for answering the question, a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or "null" if no long/short answer is present. [ 51 ]

TriviaQA: 650K question-answer-evidence triples. Includes 95K question-answer pairs scraped from 14 trivia and quiz-league websites, and (on average 6) evidence documents for each pair, gathered by searching with Bing and Wikipedia. [ 52 ]

OpenBookQA: 5960 multiple choice questions, each coming with an elementary level science fact (the "open book"). There are 1329 such facts in total. [ 53 ]

SearchQA: 140,461 question-answer pairs from the J! Archive , with each pair augmented with (on average 50) snippets and urls obtained by searching the question on Google. [ 54 ]

HotpotQA: 113K multi-hop questions that require reading multiple Wikipedia-based passages to answer. They were produced by showing crowd workers multiple supporting context documents and asking them to produce questions that requiring reasoning about all of the documents. [ 55 ]

StrategyQA: 2,780 questions annotated with relevant passages from Wikipedia, such that the question require multi-hop reasoning over the passages to answer. For example, "Did Aristotle use a laptop?" is annotated with passages from the Wikipedia pages for "laptop" and "Aristotle". [ 56 ]

DROP (Discrete Reasoning Over the content of Paragraphs): 96,567 questions along with Wikipedia passages, especially from narratives rich in numerical information (like sports summaries and history), often involving multi-step numerical reasoning over several text spans. Adversarial against 2019 SOTA. [ 57 ]

GRS-QA: Graph Reasoning-Structured Question Answering Dataset. A dataset designed to evaluate question answering models on graph-based reasoning tasks. [ 58 ]

ChartQA: 32,719 questions about 20,882 charts crawled from four diverse online sources ( Statista , Pew Research Center , Our World In Data , OECD ). Of these, 9,608 were human-written (in ChartQA-H), and 23,111 were machine-generated (in ChartQA-M). The answers are either verbatim texts from the chart or integers calculated based on the chart's data. [ 59 ]

DocVQA: multimodal, 50,000 questions on 12,767 document images, sectioned from 6,071 distinct documents. The documents were sourced from 5 industries (tobacco, food, drug, fossil fuel, chemical) of the UCSF Industry Documents Library, mostly from the 1940-2010 period. Documents with structured elements like tables, forms, lists, and figures were prioritized. The answers are verbatim extracts from the document text. [ 60 ] [ 61 ] [ 62 ]

Closed-book question-answering

C-Eval (Chinese Eval): 13948 multiple choice questions about in 52 subjects at 4 levels of difficulty. In Chinese. [ 63 ]

TruthfulQA: 817 questions in health, law, finance and politics with common misconceptions. Adversarial against GPT-3 and T5 . [ 64 ]

PIQA (Physical Interaction QA): 17951 two-choice questions. Each question gives a goal (like separating egg yolk from egg white with a water bottle), and 2 choices for accomplishing it. [ 65 ]

MedQA: 61097 questions from professional medical board exams, in English, Simplified Chinese, Traditional Chinese. [ 66 ]

ScienceQA: 21208 multiple choice questions in natural science, social science, and linguistics, with difficulty level from grade 1 to grade 12, sourced from elementary and high school science curricula. Some questions require reading a diagram. Most questions are annotated with lecture textual lectures and explanations. [ 67 ]

SimpleQA: 4,326 short questions that are answerable with knowledge as of 2023. Each answer is graded as either "correct", "incorrect", or "not attempted". Adversarial against GPT-4 specifically. [ 68 ]

RealWorldQA: 765 multimodal multiple-choice questions. Each containing an image and a question. Designed to test spatial understanding. Images are drawn from various real-world scenarios, including those captured from vehicles. [ 69 ]

OpenEQA (Open Embodied QA): over 1600 questions accompanying about videos, scans of real-world environments, and simulations. [ 70 ]

Omnibus

Some benchmarks are "omnibus", meaning they are made by combining several previous benchmarks.

GLUE (General Language Understanding Evaluation): collection of 9 benchmarks designed for testing general language understanding. The tasks are in the format of sentence- or sentence-pair. There are over 1M items. [ 71 ] [ 72 ]

SuperGLUE: An update to GLUE. Designed to be still challenging to the SOTA models of the time (2019) since the original had been saturated. Includes 8 additional tasks (e.g. logical reasoning, commonsense inference, coreference resolution). [ 73 ]

Big-Bench (Beyond the Imitation Game): A benchmark collection of 204 tasks. [ 74 ] A particular subset of 23 tasks is called BBH (Big-Bench Hard). [ 75 ] An adversarial variant of BBH is called BBEH (Big-Bench Extra Hard), made by replacing each of the 23 tasks from BBH with a similar but

adversarial variant. [ 76 ]

MMLU (Measuring Massive Multitask Language Understanding): 16,000 multiple-choice questions spanning 57 academic subjects including mathematics, philosophy, law, and medicine. [ 77 ] Upgraded to MMLU-Pro which increases the number of choices from 4 to 10, eliminated the trivial and noisy questions from MMLU, and added harder problems. [ 78 ]

MMMLU (Multilingual MMLU): The test set of MMLU, translated into 14 languages by professional human translators. [ 79 ]

CMMLU (Chinese MMLU): 1,528 multiple-choice questions across 67 subjects, 16 of which are "China-specific", like Classical Chinese . Some data collected from non-publicly available materials, mock exam questions, and questions from quiz shows to avoid contamination. More than 80% of the data was crawled from PDFs after OCR. [ 80 ]

Multimodal

Some benchmarks specifically test for multimodal ability, usually between text, image, video, and audio.

MMMU (Massive Multi-discipline Multimodal Understanding): A vision-language version of MMLU. 11550 questions collected from college exams, quizzes, and textbooks, covering 30 subjects. The questions require image-understanding to solve. Includes multiple-choice questions and open-ended QA (which are scored by regex extraction). Human expert baseline is 89%. [ 81 ] [ 82 ]

VideoMMMU: Like MMMU, but with videos. Contains 300 college-level lecture videos in 30 subjects in 6 disciplines (Art, Business, Science, Medicine, Humanities, and Engineering), with 900 questions. [ 83 ] [ 84 ]

MMMU-Pro: 1730 multiple-choice multimodal questions in the same format as MMMU, designed to be adversarial against text-only models. Some problems in MMMU turned out to be answerable without looking at the images, necessitating MMMU-Pro. Each question has 10 choices, and presented in both text-image format, and screenshot/photo format. [ 85 ]

Vibe-Eval: 269 visual understanding prompts, with standard responses written by experts. Of these, 100 were "hard" meaning they could not be solved by an LLM (Reka Core) at the time of publication. Automatic scoring by LLMs. [ 86 ]

MMT-Bench is designed to assess LVLMs performance on massive multimodal tasks that involve expert knowledge, visual recognition, localization, reasoning, and planning. The test bench includes 31,325 multi-choice questions from visual multimodal scenarios (like driving and navigation) that cover 32 core meta-tasks and 162 subtasks. [ 87 ]

Agency

GAIA: 450 questions with unambiguous answers that require information that can be obtained by browsing the Internet, requiring different levels of tooling and autonomy to solve. Divided into 3 difficulty levels. [ 88 ]

WebArena: 241 mock-up websites based on real-world websites ( Reddit , GitLab , Magento 's admin portal, etc), and 812 tasks to be performed on the websites. The tasks include information-seeking, site navigation, and content and configuration operation. [ 89 ]

Mind2Web: 2,350 tasks collected from 137 websites, and crowdsourced action sequences. The task is to reproduce the action sequence. [ 90 ]

OSWorld: 369 multimodal computer-using tasks, involving multiple real web and desktop apps and OS file I/O. In both Windows and Ubuntu . Each task includes an initial state setup configuration, and is tested by an execution-based evaluation script. [ 91 ]

Windows Agent Arena: 154 multimodal tasks with the same format as OSWorld. Only in Windows. [ 92 ]

WebVoyager: 643 multimodal tasks based on 15 popular websites. Evaluation is by screenshotting the action sequence and asking a vision language model to judge. [ 93 ]

BFCL (Berkeley Function-Calling Leaderboard): The task is to write API calls according to a specification. Released in 3 versions, with 1760, 2251, and 1000 items respectively. Some calls are evaluated by parsing into an AST and comparing against the reference answer, while others are evaluated by calling and comparing the response against the reference response. Includes Python , Java , JavaScript , SQL , and REST API . [ 94 ]

TAU-bench (Tool-Agent-User benchmark, also written as τ -bench): Two environments (retail, airline booking) that test for an agent to fulfill user instructions, interactively over multiple turns of dialogue. The user is simulated by a language model. [ 95 ] Updated to TAU2-bench ($\tau^2$-bench), which focuses on telecom applications. Tasks are synthesized by LLM-generated product requirements document , agent database schema, agent tools, and user environments ("mocked phone"). [ 96 ]

terminal-bench: A collection of complex tasks in the Linux terminal . [ 97 ]

BrowseComp: 1,266 questions that requires internet browsing for finding a short factual answer. Adversarial against GPT-4o with and without browsing, OpenAI o1, and an early version of the Deep Research model. [ 98 ]

Context length

Some benchmarks were designed specifically to test for processing continuous text that is very long.

Needle in a haystack tests (NIH): This is not a specific benchmark, but a method for benchmarking context lengths. In this method, a long context window is filled with text, such as Paul Graham's essays, and a random statement is inserted. The task is to answer a question about the inserted statement. [ 99 ]

Long Range Arena: 6 synthetic tasks that required 1K to 16K tokens of context length to solve. [ 100 ]

NoLiMa: Long-Context Evaluation Beyond Literal Matching. The benchmark assesses long-context models beyond simple keyword matching. Specifically, the words in the question have minimal or no direct lexical overlap with the words in the "needle" sentence. The "haystacks" are 10 open-licensed books. [ 101 ]

L-Eval: 2,000+ human-labeled query-response pairs over 508 long documents in 20 tasks, including diverse task types, domains, and input length (3K—200K tokens). [ 102 ]

InfiniteBench: 3946 items in 12 tasks from 5 domains (retrieval, code, math, novels, and dialogue) with context lengths exceeding 100K tokens. [ 103 ]

ZeroSCROLLS: 4,378 items in 6 tasks. Includes 6 tasks from SCROLLS and introduces 4 new datasets. Named "zero" because it was designed for zero-shot learning during the early days of pretraining paradigm, back when zero-shot capability was uncommon. [ 104 ]

LongBench: 4,750 tasks on 21 datasets across 6 task categories in both English and Chinese, with an average length of 6,711 words (English) and 13,386 characters (Chinese). [ 105 ] Updated with LongBench v2 that contained 503 more tasks, that require a context length ranging from 8K to 2M words, with the majority under 128K. [ 106 ] [ 107 ]

RULER: 13 tasks in 4 categories (retrieval, multi-hop, aggregation, question answering). Each task is specified by a program which can generate arbitrarily long instances of each task on demand. [ 108 ]

LOFT (Long-Context Frontiers): 6 long-context task categories (text retrieval, visual retrieval, audio retrieval, retrieval-augmented generation , SQL -like dataset query, many-shot in-context learning ) in 35 datasets and 4 modalities. Up to 1 million tokens. [ 109 ]

MTOB (Machine Translation from One Book): translate sentences between English and Kalamang after reading a grammar book of Kalamang (~570 pages), [ 110 ] a bilingual word list (2,531 entries, with Part-of-Speech tags ) and a small parallel corpus of sentence pairs (~400 train sentences, 100 test sentences, filtered to exclude examples from the book), both published on Dictionaria . [ 111 ] [

112 ]

FACTS Grounding: 1,719 items divided into a public set (860) and a private held-out (859) set. Each contains a document, a system instruction requiring the LLM to exclusively reference the provided document, and a user request that requires understanding of the document. Answers are scored by frontier LLMs. [ 113 ] [ 114 ]

Michelangelo: 3 tasks generated programmatically, and can be arbitrarily long. They are Multi-Round Co-reference Resolution (MRCR, track identities and references in adversarial conversation histories up to 1M tokens), Latent List, I Don't Know (IDK). [ 115 ]

Reasoning

Mathematics

Alg514: 514 algebra word problems and associated equation systems gathered from Algebra.com. [ 116 ] [ 117 ]

Math23K: 23,164 elementary school Chinese mathematical word problems, collected from various online educational websites. [ 118 ]

AQuA-RAT (Algebra Question Answering with Rationales): Also known as just "AQuA". 100,000 algebraic word problems with 5 choices per problem, and an annotation for the correct choice with natural language rationales. 34,202 "seed problems" were collected from many sources, such as GMAT and GRE, which were then expanded to the full dataset with Amazon Turk. [ 119 ]

GSM8K (Grade School Math): 8.5K linguistically diverse elementary school math word problems that require 2 to 8 basic arithmetic operations to solve. [ 120 ] Contains errors that had been corrected with GSM8K-Platinum. [ 121 ]

GSM1K: 1205 items with the same format and difficulty as GSM8K. More securely contained to avoid the data contamination concerns with the previous GSM8K. [ 122 ]

MATH: 12,500 competition-level math problems divided into difficulty levels 1 to 5 (as the Art of Problem Solving ), with AIME problems being level 5. There are 1,324 level 5 items. [ 123 ] An adversarial version is MATH-P, obtained by modifying a few characters in the original questions. [ 124 ]

MathQA: 37,200 word problems in English. Each problem came from AQuA-RAT, and annotated with an "operation program" which exactly specifies the mathematical operations required to solve the problem, written in a domain-specific language with 58 operators. [ 125 ] Has a variant, MathQA-Python, consisting of 23,914 problems, produced by taking the solutions to a subset of the MathQA dataset, and rewriting into Python. [ 126 ]

MathEval: An omnibus benchmark that contains 20 other benchmarks, such as GSM8K, MATH, and the math subsection of MMLU. Over 20,000 math problems. Difficulty ranges from elementary school to high school competition. [ 127 ]

TheoremQA: 800 questions that test for the use of 350 theorems from math, physics, electric engineering, computer science, and finance. [ 128 ]

ProofNet: 371 theorems in undergraduate-level mathematics, each consisting of a formal statement in Lean, a natural language statement, and a natural language proof. There are two tasks: given an informal (formal) statement, produce a corresponding formal (informal) statement; given an informal theorem statement, its informal proof, and its formal statement, produce a formal proof. [ 129 ] Originally was in Lean 3, [ 130 ] but the original authors deprecated it in favor of the Lean 4 version. [ 131 ]

miniF2F (mini formal-to-formal): 488 Olympiad-level mathematics problems from AIME , AMC , and IMO , stated in formal languages ( Metamath , Lean , Isabelle (partially) and HOL Light (partially)). The task is to formally prove the formal statement, which can be verified automatically. [ 132 ]

U-MATH: 1100 math problems sourced from real-world university curricula, balanced across six subjects with 20% of problems including visual elements. [ 133 ]

MathBench: 3709 questions in English and Chinese, divided into 5 difficulty levels (basic arithmetic, primary school, middle school, high school, college). Divided into 2,209 questions of MathBench-T (theoretical) and 1,500 questions of MathBench-A (applied). [ 134 ]

PutnamBench: 1709 formalized versions of Putnam competition questions during 1962 - 2023. The task is to compute the numerical answer (if there is a numerical answer) and to provide a formal proof. The formalizations are in Lean 4 , Isabelle , and Rocq (then: Coq ). [ 135 ] [ 136 ]

Omni-MATH: 4428 competition-level math problems with human annotation. [ 137 ]

FrontierMath: Several hundred questions from areas of modern math that are difficult for professional mathematicians to solve. Many questions have integer answers, so that answers can be verified automatically. Held-out to prevent contamination. [ 138 ]

MathArena: Instead of a purpose-built benchmark, the MathArena benchmark simply takes the latest math competitions (AIME and HMMT ) as soon as possible and uses those to benchmark LLMs, to prevent contamination. [ 139 ]

Programming

APPS: 10,000 problems from Codewars , AtCoder, Kattis, and Codeforces . [ 140 ]

MBPP (Mostly Basic Programming Problems): 974 short Python functions designed to be solved by entry-level programmers. Each comes with a text description and unit tests. They were written by an internal pool of crowdworkers who have basic knowledge of Python. [ 126 ]

DS-1000: 1000 data science problems obtained by reformulating 451 unique StackOverflow problems, requiring the use of 7 Python libraries, such as NumPy and Pandas. The responses are scored by running test cases and comparing outputs, and checking for the presence/absence of specific APIs or keywords. [ 141 ] [ 142 ]

HumanEval: 164 problems where the solution is always a python function, often just a few lines long. [ 9 ]

CodeElo: 387 contest problems from Codeforces during 2024, annotated with metadata such as contest divisions, problem difficulty ratings, and problem algorithm tags. Benchmarking is run by directly submitting to Codeforces, resulting in an Elo rating . Limited to 8 submissions per problem. [ 143 ]

Aider Polyglot: 225 of the hardest coding exercises from Exercism , in languages of C++, Go, Java, JavaScript, Python and Rust. [ 144 ]

BigCodeBench: 1140 tasks that requires multiple function calls. The benchmark involves 139 libraries and 7 domains. A subset BigCodeBench-Hard involves just a 148-task subset of the full benchmark. [ 145 ] [ 146 ]

SWE-bench: 2,294 software engineering problems drawn from real GitHub issues and corresponding pull requests across 12 popular Python repositories. Given a codebase and an issue, the task is to edit the codebase to solve the issue. [ 147 ] There are 2 subsets: Lite (300 problems that are faster to run), Verified (human-validated subset of 500 problems reviewed by software engineers). [ 148 ]

Multi-SWE-bench: 1,632 problems across 7 languages: Java, TypeScript, JavaScript, Go, Rust, C, and C++. Similar to SWE-bench. [ 149 ]

SWE-bench Multimodal: a variant of SWE-bench, with 619 task instances from 17 popular JavaScript repositories, each featuring images that are required for solving the task. [ 150 ]

SWE-Lancer: 1,488 freelance software engineering tasks from Upwork . Includes implementation tasks (from $50 bug fixes to $32,000 feature implementations), called "IC" (for "Individual Contributor"), and "Management" tasks, where the model must choose between technical implementation proposals. [ 151 ] [ 152 ]

KernelBench: 250 PyTorch machine learning tasks, for which a CUDA kernel must be written. [ 153 ]

Cybench (cybersecurity bench): 40 professional-level Capture the Flag (CTF) tasks from 4 competitions. Tasks are broken down into subtasks for more fine-grained scoring. At least one professional-level human team at each competition was able to solve each of the tasks. The time it took the fastest team to solve each task ranged from 2 minutes to 25 hours. [ 154 ]

HCAST (Human-Calibrated Autonomy Software Tasks): 189 tasks in machine learning, cybersecurity, software engineering, and general reasoning. Each task has a "baseline", the measured average time required for a human skilled in the task domains, working under identical conditions as AI agents. The baseline ranges from 1 minute to 8+ hours. [ 155 ]

PaperBench: 8,316 individually gradable tasks that would be necessary for replicating 20 Spotlight and Oral papers from ICML 2024 from scratch. The human baseline of ML PhDs (best of 3 attempts) at 48 hours of effort is 41.4%. [ 156 ]

ScienceAgentBench: 102 multimodal data science tasks, each being a real scientific data-driven discovery problem reframed as a code-generation task. Agents must produce a complete Python program file that implements the task, can run in isolation, and saves its outputs. In the domains of Bioinformatics , Computational Chemistry , Geographical Information Science , and Psychology & Cognitive Neuroscience . Sourced from 44 peer-reviewed publications that have released their code and data under permissive licenses. Each task is validated by domain experts. [ 157 ]

DSBench: 466 data analysis tasks and 74 data modeling tasks sourced from Kaggle and ModelOff competitions, spanning exploratory analysis, multi■table joins, and predictive modeling with large CSVs and multimodal prompts. [ 158 ]

SpreadsheetBench: 912 real-world spreadsheet manipulation tasks scraped from public Excel help forums, spanning formula writing, data cleaning, filtering and layout edits in diverse formatting. Scored automatically on 2729 test cases at cell-, sheet- and overall levels. [ 159 ]

General

GPQA (Google-Proof Q&A;): 448 multiple-choice questions written by domain experts in biology, physics, and chemistry, designed to be PhD-level. OpenAI found that human experts achieve an average score of 69.7% on the Diamond subset. [ 160 ] It is composed of 3 sets: "Extended" with 546 problems, containing all problems solicited from writers; "Main" with 448 problems, which were an expert-validated subset from "Extended"; "Diamond" with 198 problems, which were the hardest problems from "Main". In the dataset, there is also an anonymized list of 60 experts who validated the dataset, and their qualifications. [ 161 ] [ 162 ] The inter-expert agreement on the Extended set is only 74%. The construction of the dataset cost ~$120K. Each question cost an average of 2 expert-hours. Each expert was paid $100/hr. [ 163 ]

SuperGPQA: 26,529 multiple-choice questions collected by domain experts in 285 graduate-level disciplines. The questions were collected by individuals with or pursuing a PhD and then refined and inspected with the help of large language models. [ 164 ]

MathVista: 6,141 questions involving quantitative reasoning that requires reading a picture to solve. [ 165 ]

AGIEval: questions from 20 official, public, and high-standard admission and qualification exams, such as SAT , Gaokao , law school admission tests, math competitions, lawyer qualification tests, and national civil service exams . [ 166 ]

OlympicArena: 11,163 problems from 62 distinct Olympic competitions. [ 167 ]

OlympiadBench: 8,476 math and physics problems in English and Chinese, sourced from International Olympiads, Chinese Olympiads, and Gaokao. [ 168 ]

ARC-AGI (Abstraction and Reasoning Corpus for Artificial General Intelligence): Given three pairs of before-and-after diagrams of applying a rule, apply the same rule to the fourth before-diagram. It is similar to a Raven's Progressive Matrices test. [ 169 ]

LiveBench: A series of benchmarks released monthly, including high school math competition questions, competitive coding questions, logic puzzles, and other tasks. [ 170 ]

Humanity's Last Exam : 3,000 multimodal questions across over a hundred academic subjects, with a held-out private dataset left unreleased to prevent contamination. 10% of questions requires both image and text comprehension and the rest are fully text-based. 80% of questions are scored by exact string matching, and the rest are multiple-choice. [ 171 ]

SimpleBench: A multiple-choice text benchmark with over 200 questions covering spatio-temporal reasoning, social intelligence, and linguistic adversarial robustness (or trick questions). It is designed to test "everyday human reasoning". [ 172 ]

See also

List of large language models

List of datasets for machine-learning research

References

Sources

Hodak, Miro; Ellison, David; Van Buren, Chris; Jiang, Xiaotong; Dholakia, Ajay (2024). "Benchmarking Large Language Models: Opportunities and Challenges". Performance Evaluation and Benchmarking . Vol. 14247. Cham: Springer Nature Switzerland. pp. 77– 89. doi : 10.1007/978-3-031-68031-1_6 . ISBN 978-3-031-68030-4 . Retrieved 2025-09-12 .

External links

Epoch AI - AI Benchmarking Hub

v

t

e

AI-complete

Bag-of-words

n -gram Bigram Trigram

Bigram

Trigram

Computational linguistics

Natural language understanding

Stop words

Text processing

Argument mining

Collocation extraction

Concept mining

Coreference resolution

Deep linguistic processing

Distant reading

Information extraction

Named-entity recognition

Ontology learning

Parsing Semantic parsing Syntactic parsing

Semantic parsing

Syntactic parsing
Part-of-speech tagging
Semantic analysis
Semantic role labeling
Semantic decomposition
Semantic similarity
Sentiment analysis
Terminology extraction
Text mining
Textual entailment
Truecasing
Word-sense disambiguation
Word-sense induction
Compound-term processing
Lemmatisation
Lexical analysis
Text chunking
Stemming
Sentence segmentation
Word segmentation
Multi-document summarization
Sentence extraction
Text simplification
Computer-assisted
Example-based
Rule-based
Statistical
Transfer-based
Neural
BERT
Document-term matrix
Explicit semantic analysis
fastText
GloVe
Language model ( large )
Latent semantic analysis
Seq2seq
Word embedding

Word2vec
Corpus linguistics
Lexical resource
Linguistic Linked Open Data
Machine-readable dictionary
Parallel text
PropBank
Semantic network
Simple Knowledge Organization System
Speech corpus
Text corpus
Thesaurus (information retrieval)
Treebank
Universal Dependencies
BabelNet
Bank of English
DBpedia
FrameNet
Google Ngram Viewer
UBY
WordNet
Wikidata
Speech recognition
Speech segmentation
Speech synthesis
Natural language generation
Optical character recognition
Document classification
Latent Dirichlet allocation
Pachinko allocation
Automated essay scoring
Concordancer
Grammar checker
Predictive text
Pronunciation assessment
Spell checker
Chatbot
Interactive fiction

Question answering
Virtual assistant
Voice user interface
Formal semantics
Hallucination
Natural Language Toolkit
spaCy
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