

Evaluation

Sangdae Nam & Xutao Ma

09/23/2025

Why Evaluation matters

Evaluation sets the right target.

- Defining scenarios or datasets that mirror real use let evaluation as the North Star -> align model optimization with actual case
- Quantify improvement
- Hyperparameter selection

What is good eval dataset?

- clear purpose and scenarios
- broad coverage with difficulty
- contamination control
- reliable sources and labels

NLP benchmarks before MMLU

- **GLUE/SuperGLUE**

- A benchmark suite (2018) that combines multiple English sentence/sentence-pair classification and similarity tasks
- An upgraded SuperGLUE, more difficult successor introduced in 2019, includes tasks that demand deeper reasoning, commonsense etc
- *CoLA (Corpus of Linguistic Acceptability)*
 - Input:** "The books on the table is red."
 - Task:** Decide whether the sentence is grammatically acceptable or not. (Here: unacceptable)

- **SQuAD**

- A reading comprehension benchmark in which models are given a paragraph (from Wikipedia) and questions whose answers are spans of text in the paragraph
- **Context passage excerpt:** *"Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ ..."*
 - **Question:** "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?"
 - **Answer:** "Saint Bernadette Soubirous"

Modern NLP benchmarks

- **MMLU (Measuring Massive Multitask Language Understanding):**
 - multiple-choice benchmark covering ~57 academic subjects (STEM, humanities, social sciences, etc.)
 - test an LLM's zero-shot or few-shot knowledge and reasoning ability
- **LiveBench:**
 - newer benchmark with frequently updated questions from recent sources (math contests, arXiv papers, news, etc.),
 - aiming to avoid test set contamination, with automatic scoring using objective ground truth
- **GPQA (Graduate-Level Problems in Quantitative Analysis / Google-Proof Q&A):**
 - A benchmark of challenging multiple-choice problems in graduate-level physics, mathematics, chemistry
- **MT-Bench:**
 - multi-turn conversational benchmark that tests LLMs on coherent, instruction-following dialogue over multiple turns
 - comparing responses to approximate human preference

Data collection

Collect from existing natural sources

MMLU: “collected by graduate and undergraduate students from freely *available sources online*”.

SWE-bench: “drawn from real *GitHub* issues and corresponding pull requests”

AIME: American Invitational Mathematics Examination

Experts manually create dataset

GPQA: “We hire *61 contractors (PhD)* through Upwork to write and validate the dataset.”

SimpleQA: “We asked *AI trainers* to create knowledge-seeking questions”

LLM assisted dataset curation

ToolLLM: “We prompt *ChatGPT* to generate diverse instructions involving these APIs”

Data filtering

Error detection

- Incorrect question & answer

- Syntax errors

Quality filtering

- Debias

- Deduplication

- Remove too easy data

Multitask Language Understanding

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Multitask Language Understanding

Rephrasing the Web: A Recipe for Compute & Data-Efficient Language Modeling

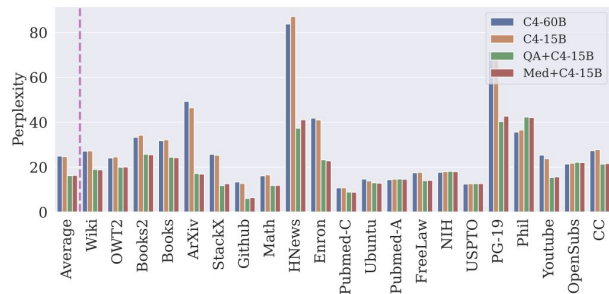


Figure 12: Perplexity across all domains of the Pile comparing combining multiple styles of synthetic data. Models are 350M parameters trained for a total of 75B tokens.

Dataset (Real Tok.)	ARC-C	SciQ	PubMedQA	MathQA	MMLU	Avg
C4-15B	21.2	77.1	50.6	22.2	23.1	38.8
C4-60B	23.4	76.2	46.4	22.0	23.0	38.2
QA+C4-15B	24.4	79.8	56.0	21.7	22.9	41.0
Med+C4-15B	22.7	74.5	53.6	22.0	23.1	39.2

Skill	Benchmark _(eval)	Tülu 3 8B	Qwen 2.5 7B Instruct	Llama 3.1 8B Instruct	Tülu 3 70B	Qwen 2.5 72B Instruct	Llama 3.1 70B Instruct	GPT-3.5 Turbo	GPT-4o Mini	Claude 3.5 Haiku
	Avg.	65.1	66.5	62.9	76.2	72.8	74.1	64.7	69.6	75.3
Knowledge	MMLU _(0 shot, CoT)	68.2	76.6	71.2	83.1	85.5	85.3	70.2	82.2	81.8
	PopQA _(15 shot)	29.1	18.1	20.2	46.5	30.6	46.4	45.0	39.0	42.5
	TruthfulQA _(6 shot)	55.0	63.1	55.1	67.6	69.9	66.8	62.9 [°]	64.8 [°]	64.9 [°]
Reasoning	BigBenchHard _(3 shot, CoT)	69.0	70.2	71.9	85.0	80.4	83.0	66.6 [†]	65.9 [°]	73.7 [†]
	DROP _(3 shot)	62.6	54.4	61.5	74.3	34.2	77.0	70.2	36.3	78.4
Math	MATH _(4 shot CoT, Flex)	43.7	69.9	42.5	63.0	75.9	56.4	41.2	67.9	68.0
	GSM8K _(8 shot, CoT)	87.6	83.8	83.4	93.5	89.5	93.7	74.3	83.0	90.1
Coding	HumanEval _(pass@10)	83.9	93.1	86.3	92.4	94.0	93.6	87.1	90.4	90.8
	HumanEval+ _(pass@10)	79.2	89.7	82.9	88.0	90.8	89.5	84.0	87.0	88.1
IF & chat	IFEval _(prompt loose)	82.4	74.7	80.6	83.2	87.6	88.0	66.9	83.5	86.3
	AlpacaEval 2 _(LC % win)	34.5	29.0	24.2	49.8	47.7	33.4	38.7	49.7	47.3
Safety	Safety _(6 task avg.)	85.5	75.0	75.2	88.3	87.0	76.5	69.1	84.9	91.8

Multitask Language Understanding

Motivations

- Previous benchmarks, like GLUE were not enough to measure the performance of up-to-date LLM, since top models already achieved superhuman performance.
- Also these benchmarks evaluated linguistic skills more than overall language understanding
- Need to bridge the gap between wide-ranging knowledge and existing measures across diverse set of subjects that humans learn.

Multitask Language Understanding

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?

(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.

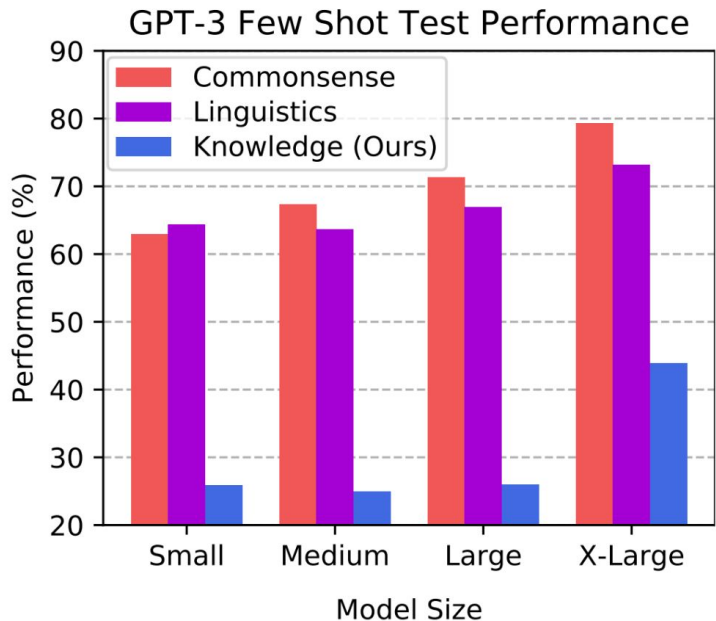
(A) -1 (B) 1 (C) i (D) $-i$

Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?

(A) 28 (B) 21 (C) 40 (D) 30

Answer: C



Multitask Language Understanding

Multi Task

- **57 tasks** in total
- questions and datasets manually collected by graduate and undergraduate students
 - from GRE or United States Medical Licensing Examination(USMLE) etc
- collected 15908 questions
 - split into few-shot development set, validation set, and test set.
 - few shot dev set: 5Q per subject
 - valid set: used for hyperparameter tuning, 1540Q
 - test set: 14079Q

Human-level Accuracy

- Unspecialized humans from Amazon Mechanical Turk: **34.5%**
- Real-world test-taker human's 95th percentile: **87% for USMLE**

Why task and domain diversity?

- The benchmark emphasizes **breadth and depth**
- Not just language skills but **real-world academic/professional knowledge** across domains—revealing **lopsided performance**

Multitask Language Understanding


Multi Task


- Humanities


- Law, philosophy, history
- Sources: ETHICS dataset etc


Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. 

(B) Yes, if Hermit was responsible for the explosive charge under the driveway. 

(C) No, because Seller ignored the sign, which warned him against proceeding further. 

(D) No, if Hermit reasonably feared that intruders would come and harm him or his family. 


- Social Science


- Economics, sociology, politics, geography, psychology etc
- Sources: High school questions with AP style.


Examination for Professional Practice in Psychology (EPPP)


Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

(A) producer surplus is lost and consumer surplus is gained. 

(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. 

(C) monopoly firms do not engage in significant research and development. 

(D) consumer surplus is lost with higher prices and lower levels of output. 

Multitask Language Understanding

Multi Task

- STEM

- Physics, computer science, math etc
- Sources: GRE, College mathematic questions etc

Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) 9.8 m/s^2	✓
	(B) more than 9.8 m/s^2	✗
	(C) less than 9.8 m/s^2	✗
	(D) Cannot say unless the speed of throw is given.	✗
College Mathematics	In the complex z -plane, the set of points satisfying the equation $z^2 = z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

- Others

- Medicine, finance, accounting, marketing etc
- Sources: Business course and questions
USMLE questions etc

Professional Medicine	A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL , albumin concentration of 4 g/dL , and parathyroid hormone concentration of 200 pg/mL . Damage to which of the following vessels caused the findings in this patient?	
	(A) Branch of the costocervical trunk	✗
	(B) Branch of the external carotid artery	✗
	(C) Branch of the thyrocervical trunk	✓
	(D) Tributary of the internal jugular vein	✗

Figure 5: A question from the Professional Medicine task.

Multitask Language Understanding

Experiments: Setup

- Models
 - Random Baseline: 25% since $\frac{1}{4}$ choice problem
 - RoBERTa + **finetune** on UnifiedQA training data + dev&val set
 - ALBERT + **finetune** on UnifiedQA training data + dev&val set
 - GPT-2 + **finetune** on UnifiedQA training data + dev&val set
 - UnifiedQA (already finetuned, evaluate **without any further tuning**)
 - GPT3 Small(2.7B) + **few shot**
 - GPT3 Medium(6.7B) + **few shot**
 - GPT3 Large(13B) + **few shot**
 - GPT3 X-Large(175B) + **few shot**

Multitask Language Understanding

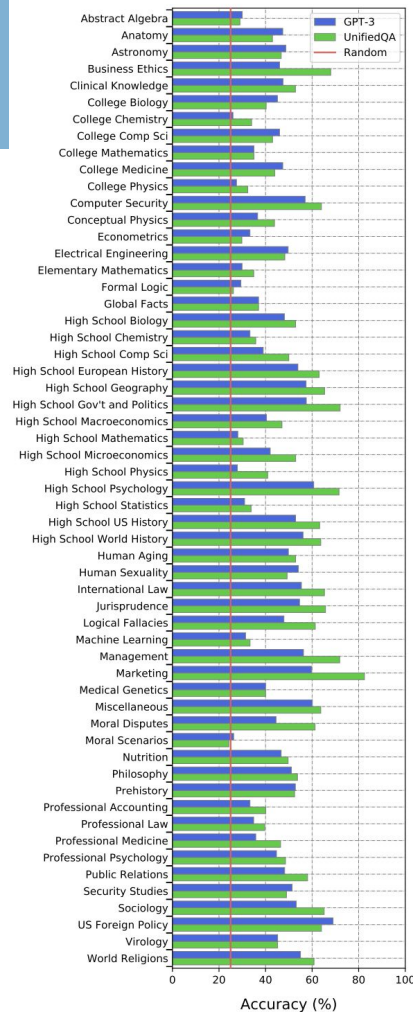
Experiments: Results

Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
RoBERTa	27.9	28.8	27.0	27.7	27.9
ALBERT	27.2	25.7	27.7	27.9	27.1
GPT-2	32.8	33.3	30.2	33.1	32.4
UnifiedQA	45.6	56.6	40.2	54.6	48.9
GPT-3 Small (few-shot)	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium (few-shot)	26.1	21.6	25.6	25.5	24.9
GPT-3 Large (few-shot)	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large (few-shot)	40.8	50.4	36.7	48.8	43.9

GPT-3 Results



UnifiedQA Results



Multitask Language Understanding

Experiments: Results

Declarative vs. Procedural Knowledge

Prompt and Completion:

The order of operations or PEMDAS is
Parenttheses Exponents Multiplication
Division Addition Subtraction

Prompt and Completion:

$(1 + 1) \times 2 =$ 3

Figure 7: GPT-3's completion for two prompts testing knowledge of the order of operations. The blue underlined bold text is the autocompleted response from GPT-3. While it *knows about* the order of operations, it sometimes does not *know how* to apply its knowledge.

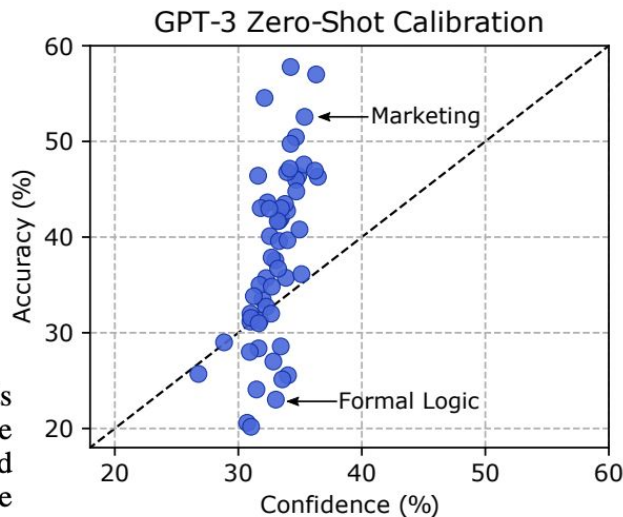


Figure 8: GPT-3's confidence is a poor estimator of its accuracy and can be off by up to 24%.

Multitask Language Understanding

Experiments: Results

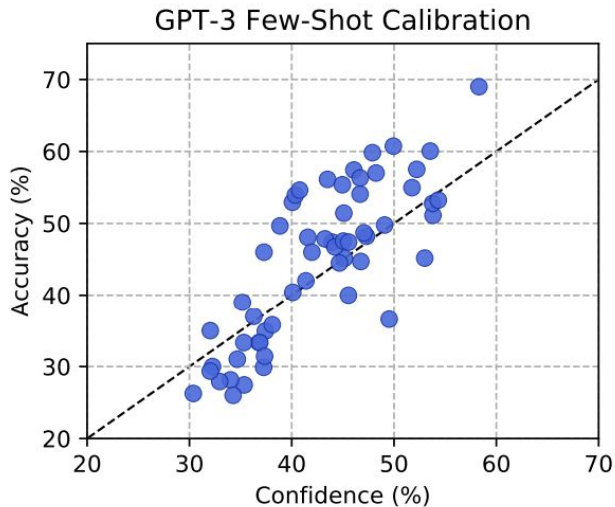


Figure 11: While models are more calibrated in a few-shot setting than a zero-shot setting, they are still miscalibrated, with gap between accuracy and confidence reaching up to 14%. Here the correlation between confidence and accuracy is $r = 0.81$, compared to $r = 0.63$ in the zero-shot setting.

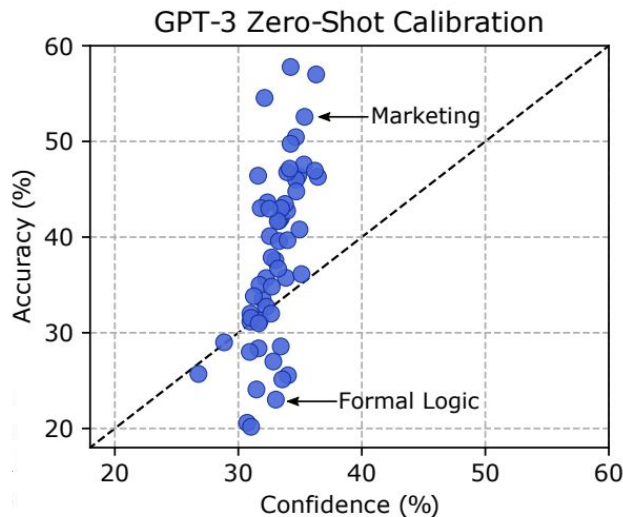


Figure 8: GPT-3's confidence is a poor estimator of its accuracy and can be off by up to 24%.

Multitask Language Understanding

Experiments: Results

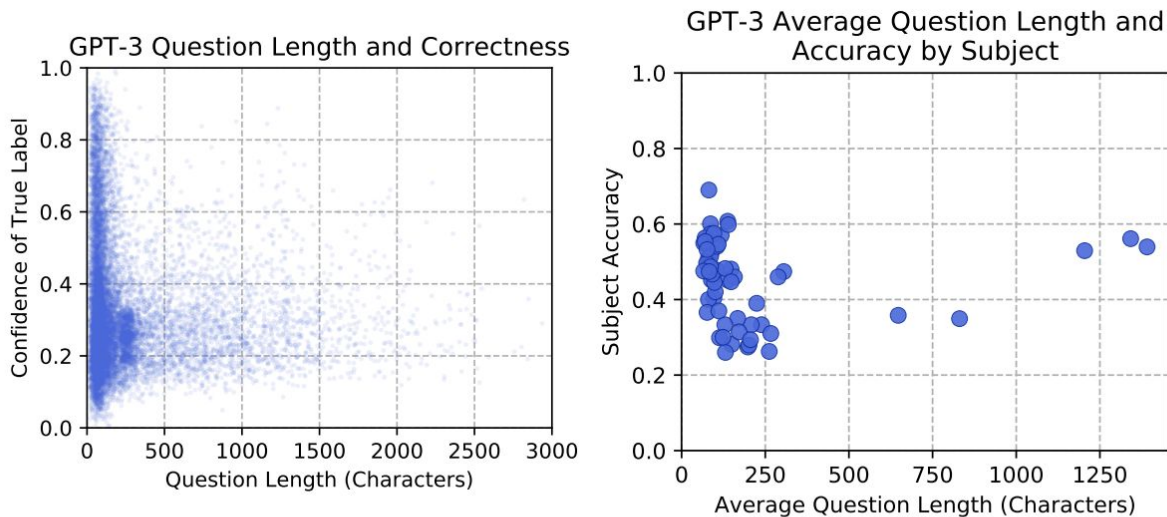
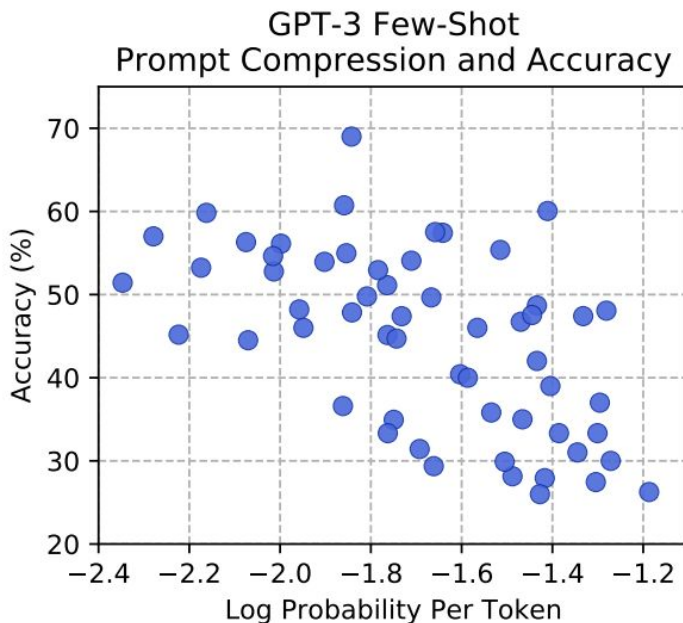
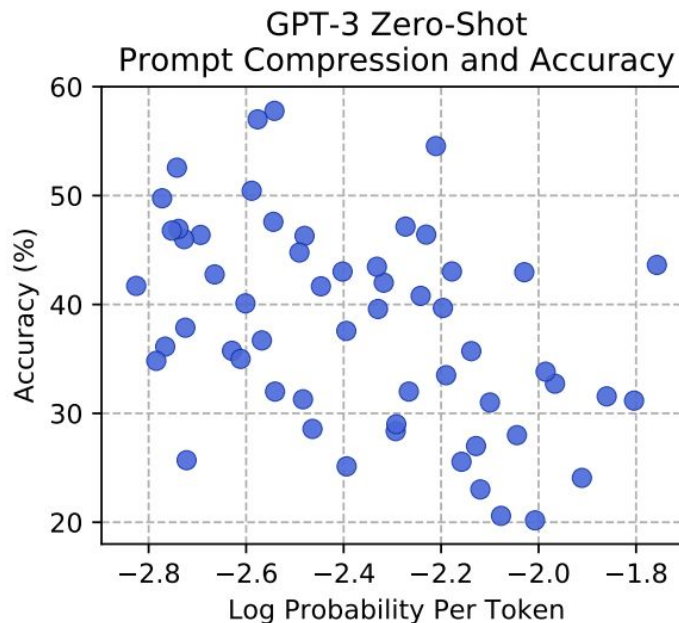


Figure 12: Figures on the relation between question difficulty and question length. For questions longer than a tweet (280 characters), the correlation between question length and true label confidence is slightly positive. This shows that longer questions are not necessarily harder.

Multitask Language Understanding

Experiments: Results



Measuring short-form factuality in large language models

Jason Wei*

Nguyen Karina*

Hyung Won Chung

Yunxin Joy Jiao

Spencer Papay

Amelia Glaese

John Schulman

William Fedus

OpenAI

7-Nov-2024

*"We present **SimpleQA**, a benchmark that evaluates the ability of language models to **answer short, fact-seeking questions**."*

Factual QA benchmarks before SimpleQA

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?

Answer: Fitness

Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Figure 1: Question-answer pairs with sample excerpts from evidence documents from TriviaQA exhibiting lexical and syntactic variability, and requiring reasoning from multiple sentences.

TriviaQA

Input: Question+(Excerpt)

Output: Answer

data collection: *"we gathered question-answer pairs from 14 trivia and quiz-league websites"*

Data set size: over **650K** question-answer-evidence triples

Factual QA benchmarks before SimpleQA

Example 1

Question: what color was john wilkes booth's hair

Wikipedia Page: John_Wilkes_Booth

Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astounding memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2

Question: can you make and receive calls in airplane mode

Wikipedia Page: Airplane_mode

Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Example 3

Question: why does queen elizabeth sign her name elizabeth r

Wikipedia Page: Royal_sign-manual

Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

Natural Questions (Google)

Input: Question+(Wiki pages)

Output: Answer

Data collection: *"The questions consist of real anonymized, aggregated **queries issued to the Google search engine.**"*

Data set size: *"The public release contains **307,373 training examples with single annotations**"*

Figure 1: Example annotations from the corpus.

SimpleQA

	Benchmark (Metric)	# Shots	DeepSeek-V2 Base	Qwen2.5 72B Base	LLaMA-3.1 405B Base	DeepSeek-V3 Base
	Architecture	-	MoE	Dense	Dense	MoE
	# Activated Params	-	21B	72B	405B	37B
	# Total Params	-	236B	72B	405B	671B
English	Pile-test (BPB)	-	0.606	0.638	0.542	0.548
	BBH (EM)	3-shot	78.8	79.8	82.9	87.5
	MMLU (EM)	5-shot	78.4	85.0	84.4	87.1
	MMLU-Redux (EM)	5-shot	75.6	83.2	81.3	86.2
	MMLU-Pro (EM)	5-shot	51.4	58.3	52.8	64.4
	DROP (F1)	3-shot	80.4	80.6	86.0	89.0
	ARC-Easy (EM)	25-shot	97.6	98.4	98.4	98.9
	ARC-Challenge (EM)	25-shot	92.2	94.5	95.3	95.3
	HellaSwag (EM)	10-shot	87.1	84.8	89.2	88.9
	PIQA (EM)	0-shot	83.9	82.6	85.9	84.7
	WinoGrande (EM)	5-shot	86.3	82.3	85.2	84.9
	RACE-Middle (EM)	5-shot	73.1	68.1	74.2	67.1
	RACE-High (EM)	5-shot	52.6	50.3	56.8	51.3
	TriviaQA (EM)	5-shot	80.0	71.9	82.7	82.9
	NaturalQuestions (EM)	5-shot	38.6	33.2	41.5	40.0
	AGIEval (EM)	0-shot	57.5	75.8	60.6	79.6

Liu, Aixin, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao et al. "**Deepseek-v3 technical report.**" *arXiv preprint arXiv:2412.19437* (2024).

Question	Answer
Who received the IEEE Frank Rosenblatt Award in 2010?	Michio Sugeno
On which U.S. TV station did the Canadian reality series *To Serve and Protect* debut?	KVOS-TV
What day, month, and year was Carrie Underwood's album "Cry Pretty" certified Gold by the RIAA?	October 23, 2018
What is the first and last name of the woman whom the British linguist Bernard Comrie married in 1985?	Akiko Kumahira

Table 1: Four example questions and reference answers from SimpleQA.

Motivation:

1. **Challenge parametric knowledge retrieval ability (Older benchmarks are saturated)**
2. **High correctness**

Data collection and verification

Creation step: **AI trainers** (i.e., human annotators) **created** question and answer pairs.

Validation step: Questions were independently answered by **another AI trainer** and **only kept if answers from both trainers matched**

Data collection and verification

Creation step: **AI trainers** (i.e., human annotators) **created** question and answer pairs.

Criteria:

1. Must have a **single answer**.
2. Reference answers should **not change over time**.
3. Reference answers must be **supported by evidence**. (provide a **link** to the webpage that supports the reference answer to the question)
4. Must be **challenging**.
four GPT models answer → at least 1 incorrect
5. The question must be **answerable as of 2023**.

Quality control:

run a series of few-shot-prompted ChatGPT classifiers to **detect criteria violations**

Data collection and verification

Validation step:

Another trainer answer → Must match the reference answer

Another trainer → Judge if problem-answer complies with criteria

How about using LLM+search_tools to do validation?

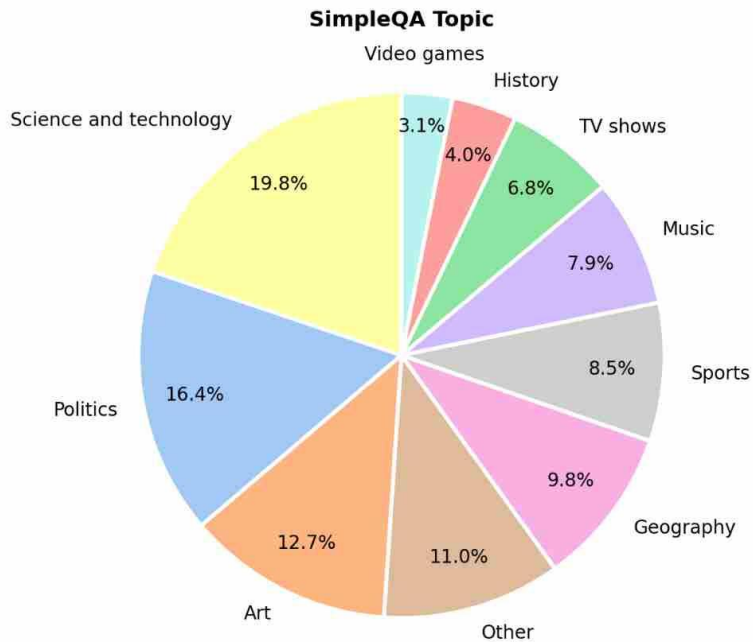
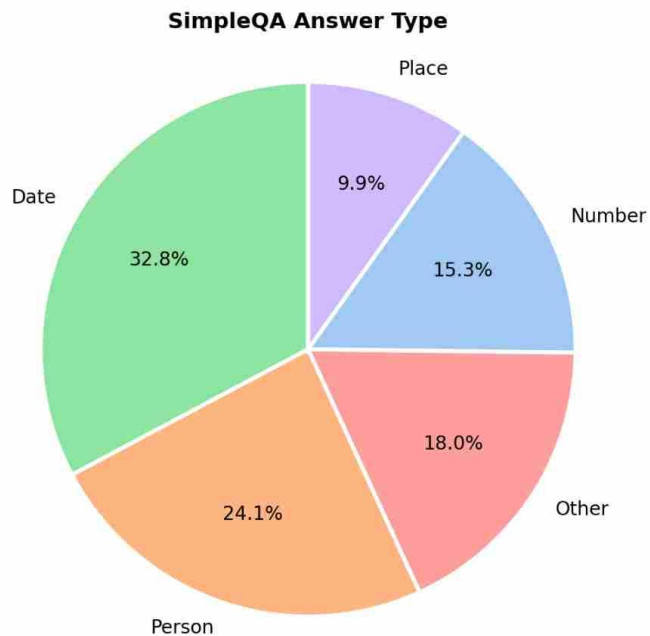
Final quality check:

Sample 1000 problems → Third trainer answer

Error rate of the benchmark is around 3%

why not test all the problem (4326 in total)?

Data diversity



Model evaluation

Grade	Definition	Example responses
Correct	The predicted answer fully contains the reference answer without contradicting the reference answer.	“Wout Weghorst”, “Wout Weghorst scored at 83’ and 90+11’ in that game”
Incorrect	The predicted answer contradicts the reference answer in any way, even if the contradiction is hedged.	“Virgil van Dijk”, “Virgil van Dijk and Wout Weghorst”, “Wout Weghorst and I think van Dijk scored, but I am not totally sure”
Not attempted	The reference answer is not fully given in the answer, and there are no contradictions with the reference answer.	“I don’t know the answer to that question”, “To find which Dutch player scored in that game, please browse the internet yourself”

Table 2: Grading categories with examples completions. The question here is “Which Dutch player scored an open-play goal in the 2022 Netherlands vs Argentina game in the men’s FIFA World Cup?” (Answer: Wout Weghorst).

Metric:

1. Overall correct (or “correct”):

$$\frac{\text{correct}}{\text{correct} + \text{incorrect} + \text{no_attempt}}$$

2. Correct given attempted:

$$\frac{\text{correct}}{\text{correct} + \text{incorrect}}$$

F-score:

$$\frac{1}{\text{overall_correct}} + \frac{2}{\frac{1}{\text{correct_given_attempted}}}$$

Model evaluation

Correct	No attempt	Incorrect	Overall correct	Correct given attempted	F-score
30%	0%	70%	30%	30%	30%
19%	76%	5%	19%	80%	30%

Towards a single number metric

$r(\text{correct}) = 1$

$r(\text{no_attempt}) = a, a \geq 0$

$r(\text{incorrect}) = 0$

Model would give an answer only **when the correct probability $\geq a$**

Model evaluation














Model	Correct	Not attempted	Incorrect	Correct given attempted	F-score
Claude-3-haiku (2024-03-07)	5.1	75.3	19.6	20.6	8.2
Claude-3-sonnet (2024-02-29)	5.7	75.0	19.3	22.9	9.2
Claude-3-opus (2024-02-29)	23.5	39.6	36.9	38.8	29.3
Claude-3.5-sonnet (2024-06-20)	28.9	35.0	36.1	44.5	35.0
GPT-4o-mini	8.6	0.9	90.5	8.7	8.6
GPT-4o	38.2	1.0	60.8	38.0	38.4
OpenAI o1-mini	8.1	28.5	63.4	11.3	9.4
OpenAI o1-preview	42.7	9.2	48.1	47.0	44.8

Table 3: Performance of various models on SimpleQA. F-score is the harmonic mean between correct and correct given attempted; see Appendix B for discussion.

SimpleQA leaderboard

Results independently reproduced by Kaggle. [Learn more](#)

Last updated 2025年9月2日

#	Model	↓	Score
1	 Gemini-2.5-Pro-Preview-06-05		55.1% ±1.5%
2	 Gpt-5-2025-08-07		51.1% ±1.5%
3	 O3-2025-04-16		50.5% ±1.4%
4	 Qwen3-235b-A22b-Thinking-2507		50.4% ±1.4%
5	 Grok-4-0709		50.3% ±1.5%
6	 O1-2024-12-17		47.2% ±1.5%
7	 Grok-3		41.3% ±1.4%
8	 Gpt-4.1-2025-04-14		40.9% ±1.5%
9	 Gpt-4o-2024-08-06		38.4% ±1.5%
10	 Claude-Opus-4-1-20250805		37.6% ±1.5%

Measuring calibration

Prompt:

Please give your best guess, along with your **confidence** as a percentage that that is the correct answer

Perfect calibrated model:

Have the **same actual accuracy as stated confidence**.

ie. collect all prompts with confidence p . the accuracy of these prompt should be p .

Measuring calibration

Perfect calibrated model:

Have the **same actual accuracy as stated confidence**.

ie. collect all prompts with confidence p . the accuracy of these prompt should be p .

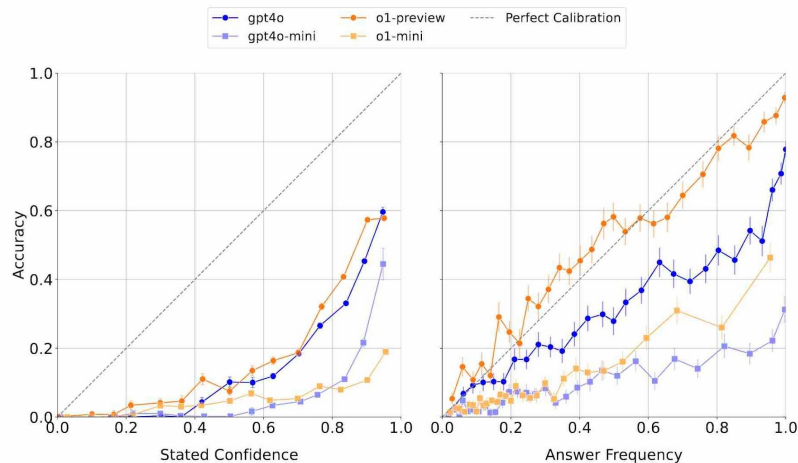


Figure 2: Left: Calibration of language models based on their stated confidence, uniformly binned into 15 intervals. Right: Calibration assessed by asking models the same question 100 times, quantile-binned into 30 intervals.

Impact

MMLU

- Filled a gap by testing knowledge during pretraining with 57 tasks from various subjects and level.
- Before MMLU, mainly targeted linguistic skills and saturated quickly.

SimpleQA

- Showed that modern large models have many shortcomings in factual knowledge retrieval & calibration (SOTA: 55.1% by gemini 2.5 pro).
- Setted a successful example of building a human manually written benchmark.

Discussion

More diverse methods to evaluate

- Benchmarks
 - **lm-eval harness**: Framework to run multiple benchmarks or tasks
 - Response accuracy, log likelihood => Also heavy job. How can we deal with it
- **LLM-as-judge** => Why we use LLM-as-judge at the industry level? How is it different?
 - Using external LLM as judgement
 - Providing context, criteria, rubric etc
- **Chatbot arena(lmsys)**: Human Eval
 - web platform run by LMSYS that lets users compare responses from two anonymous LLMs driven by the same prompt and vote which response they prefer.

Pros and cons of manually written queries

- High quality (eg. SimpleQA, GPQA)
- Hard to scale (eg. 4k in SimpleQA compared to 650k in TriviaQA)
- Diversity issue

Critic

Huanzhi Mao & Colin Wang

9/23

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Multiple-choice format can mask shallow shortcuts

- MMLU turns everything into four-option classification because open-ended NLG is hard to score consistently.
- Multiple choice enables elimination and surface-pattern cues; it rarely forces reasoning or stepwise computation.

“Unusual learning order” suggests corpus exposure over mastery

- GPT-3 does better on College Medicine (47.4%) and College Mathematics (35.0%) than on Elementary Mathematics (29.9%)
- Accuracy can reflect where text exposure is richest rather than competence progression.

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Prompt/format sensitivity can swing scores

- Some systems (notably UnifiedQA) change by several percentage points with small formatting tweaks; shots/order matter.
- Leaderboard positions can become prompt-engineering artifacts.
- Removing a single terminator token (“</s>”) drops UnifiedQA accuracy by several points; more few-shot exemplars steadily increase accuracy.

Content mix and human baselines

- Many items come from US exams (GRE, USMLE, AP), with US-centric subjects (e.g., US History, US Foreign Policy). Expert-level accuracy is partly estimated from 95th-percentile test takers.
- Cultural skew and uneven difficulty normalization make cross-domain comparisons noisier.

SimpleQA

- I like the spirit: renovate “toy” tasks into harder, real-world versions
- But benchmarks should be proactive, not just harder and even adversarial versions of old ones
- For simpleQA: **why emphasize long-tail recall when models now use agentic search?**
- Reinventing old tasks = fitting yesterday’s paradigm, not today’s models
- BrowseComp (first-authored again by Jason who created SimpleQA but at a later time) shows a forward-looking alternative

But why do people keep making this mistake?

- Researchers ask: “Why did models saturate old benchmarks?”
 - Because they’ve been evaluating their models on these benchmarks
 - This cycle creates researcher inertia → retrofitting instead of rethinking
- Rarely ask: “What benchmarks match current capabilities?”
- This drives adversarial tweaks of past tasks, not new ones
- Evaluation then feels misaligned with how models actually operate
- Benchmarks should be designed to anticipate how models are used now

Also, an emerging trend in benchmark curation...

- Past benchmarks usually involve taking exam questions from

Evaluation

Proponents

Bhavya Chopra, Qiuyang Mang

09/23

Benchmarks are imperfect, but necessary

Benchmarks and rigorous evaluation reveal hidden weaknesses:

- Failures in math, law, ethics, STEM reasoning
- Overconfidence: Models do not know when they're wrong, lack calibration
- Provide shared baselines for comparison
- Objectively measure performance of new capabilities

In the absence of evaluation, there is a false perception of human-level intelligence based on cherry-picked examples and demos.

Benchmarks are imperfect, but necessary

Benchmarks add dimensions to evaluate orthogonal aspects:

- MMLU, ARC, AGI: Reasoning tasks and subjects
- SimpleQA, FACTS: Factuality and Calibration
- SWE-Bench: Coding
- EQ-Bench: Emotional and Social abilities
- ChatBot Arena, LiveBench: Interaction, Fluency, User Satisfaction
- AIME: Math

Collectively, these enable measured progress and accountability—making ablations and experiments possible.

Short Testing Time (Simple Q&A)

- **Good researcher UX.** SimpleQA is fast and simple to run, as questions and answers are very short. Grading is also fast to run via the OpenAI API (or another frontier model API). Additionally, with 4,326 questions in the dataset, SimpleQA should have relatively low run-to-run variance.
 1. **Quick prompt & hyperparameter tuning**
 2. **Easy to reproduce**
 3. **More important in the agent era**

Check LLMs know what they know

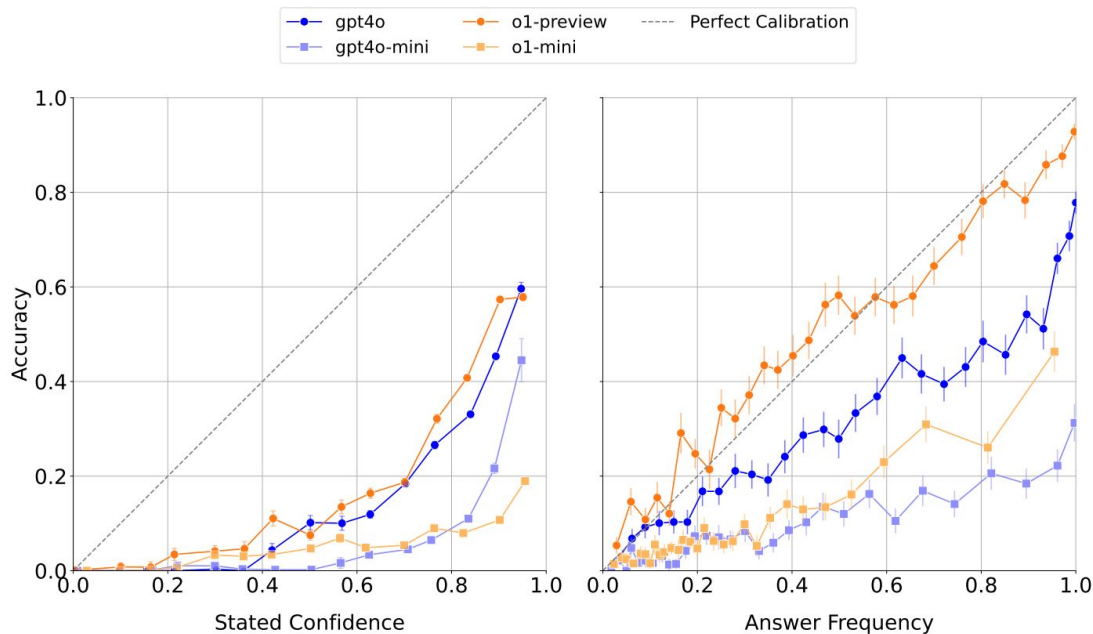


Figure 2: Left: Calibration of language models based on their stated confidence, uniformly binned into 15 intervals. Right: Calibration assessed by asking models the same question 100 times, quantile-binned into 30 intervals.

Diversity of Tasks

Evaluates the **general** ability of language models to answer short, fact-seeking questions

Not domain-specific knowledge

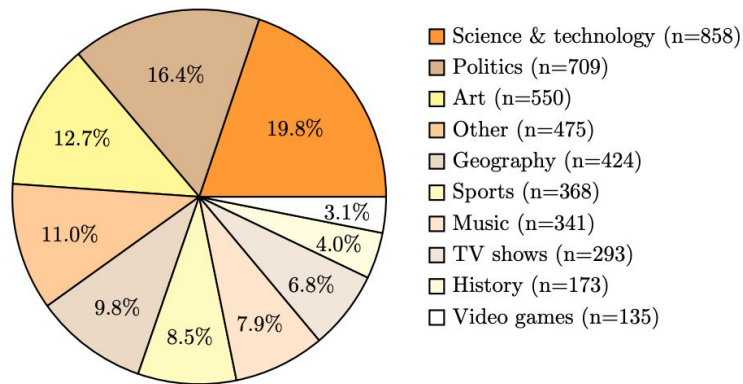


Figure 1: Distribution of topics in SimpleQA. The topic for each question was classified via a prompted ChatGPT model.

Driving Progress with Evolving Benchmarks

Benchmarks can quickly become saturated, yes.

LLMs quickly achieved human-level performance with GLUE and SuperGLUE.

Each saturation shows what is solved, and pushes the community to set harder goals and tasks.

Need for new, evolving, forward-looking benchmarks like:

- GPQA (targeting graduate level expertise, questions are “Google-proof”)
- Unsolved Questions (asking truly open-ended research questions)
- LongFact / FreshQA (long-form & fast-changing knowledge)

Tackling Standardization

Need to push for standardized evaluations, and push back on cherry-picking of evaluation benchmarks and methodologies, e.g. OLMES paper pushes for reproducible details:

- **Instance formatting:** How Q/A are presented to LLM
- **Few-shot examples:** Which in-context examples to use, how many
- **Probability normalization:** How to handle token probabilities for scoring
- **Task formulation:** When to use multiple-choice versus completion/cloze format
- **Implementation details:** Computational and processing specifications

E.g. OLMES considers both MCF and CF formats, and selects the better-performing one→giving equal leverage to small and big models.

Evaluation: Follow-up

UQ: Assessing Language Models on Unsolved Questions

Siddharth Gollapudi

09/23/2025

Key Takeaway

As models get better, benchmarks must keep pace with current capabilities

2018

- grammar, comprehension
- simpler data collection
- easy-to-quantify evaluation metrics
- models have surpassed humans in these benchmarks



Today

- Knowledge testing
- Hallucinations
- Spicier metrics of accuracy
- Models are getting there (?)

What's Next?

1. How to check if subjective/complex answers from a model are good?
 2. Are models actually being evaluated on a level playing field?
 3. How do we know models are actually applying what they know?
 4. Are the problems these models tackling realistic and important?
-

UQ: Assessing Language Models on Unsolved Questions

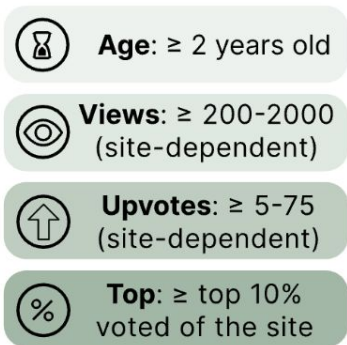
UQ Dataset

Unsolved Questions Raw Crawl



3,000,000+ candidates
from 80+ sites

Rule-based Filtering



33,916 candidates
(1.13% of original)

LLM-based Filtering



7,685 candidates
(0.26% of original)

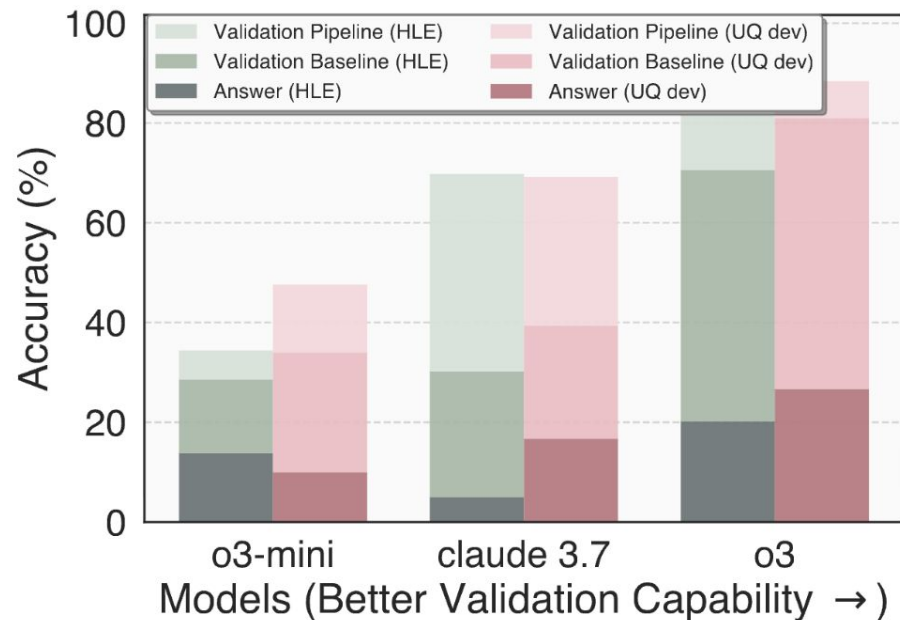
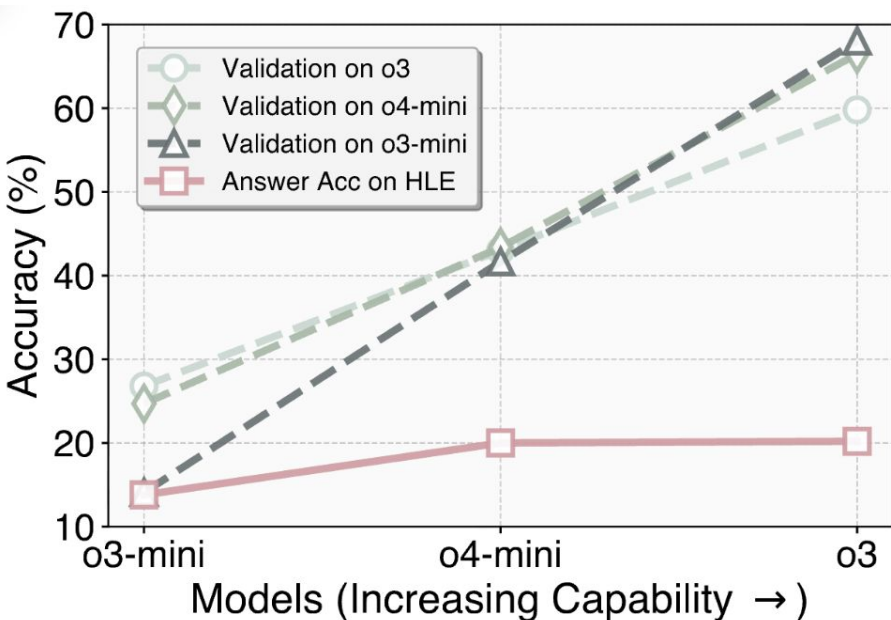
Human Review



500 final questions
(25 diamond set)

UQ Verifiers: Why?

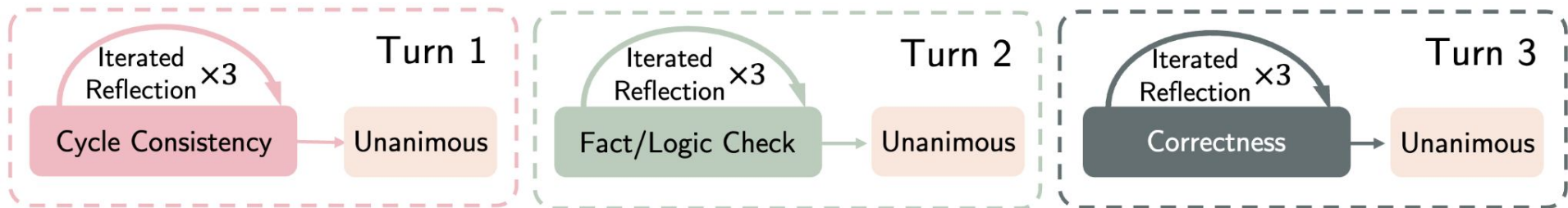
- No answers available, at least try and filter out incorrect answers
- For eval. purposes use Humanity's Last Exam, which transfers well



UQ Verifiers: How?

“Multi-shot” prompting strategies for improving verifier recall

1. Single prompt to verify for correctness/logic/question
2. Take a bunch of samples, have models “re-verify” their answer
3. Use a voting scheme on the samples to get a final answer



UQ Verifiers Evaluation

- Accuracy is percent of verifications correct

- Precision is number of false positives

- Recall is number of false negatives

Model	Strategy	Accuracy (%)	Precision (%)	Recall (%)
Claude Sonnet 3.7	Vanilla Prompt (Baseline)	21.60	13.26	90.77
	Correctness	30.20	14.85	92.31
	Correctness $\times 5$ Majority	29.40	14.53	90.77
	Correctness $\times 5$ Unanimous	41.20	15.82	81.52
	Correctness $\cap 5$ Unanimous	54.32	23.08	56.25
	3-Iter Pipeline	73.20	20.00	16.00
o3-mini	Vanilla Prompt (Baseline)	24.00	14.29	96.92
	Correctness	28.60	15.24	98.46
	Correctness $\times 5$ Majority	29.20	15.18	96.92
	Correctness $\times 5$ Unanimous	33.00	15.56	93.85
	Correctness $\cap 5$ Unanimous	30.00	15.16	95.38
	3-Iter Pipeline	34.40	15.84	93.85
o3	Vanilla Prompt (Baseline)	58.12	20.73	78.46
	Correctness	70.60	22.00	50.00
	Correctness $\times 5$ Majority	73.15	25.87	56.92
	Correctness $\times 5$ Unanimous	83.77	26.47	13.85
	Correctness $\cap 5$ Unanimous	78.60	28.57	43.08
	1-Iter Pipeline	75.40	24.00	42.00
	3-Iter Pipeline	81.65	30.99	34.38
	5-Iter Pipeline	81.50	26.23	25.40
Multi-model ensemble	Correctness (5 Models) Majority	45.00	17.99	90.77
	Correctness (5 Models) Unanimous	78.60	25.00	32.31
	3-Iter Pipeline (2 Models) Unanimous	85.40	40.00	24.62