Table of Contents

MMLU 2
GPQA: GPQA: A Graduate-Level Google-Proof Q&A Benchmark 2
TruthfulQA 2
Winogrande 3
GSM8K 5
LiveBench: A Challenging, Contamination-Free LLM Benchmark 5
ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic 6
CaLMQA: Exploring culturally specific long-form question answering across 23 languages 6
Instruction-Following Evaluation for Large Language Models 7
AI and the Everything in the Whole Wide World Benchmark 8
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena 9
MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues 10
TencentLLMEval: A Hierarchical Evaluation of Real-World Capabilities for Human-Aligned LLMs 10
Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference 10
LLM-Perf leaderboard: A leaderboard which focuses on benchmarking the performance (latency, throughput, and memory) of large language models across different hardware and optimizations. 11
HumanEval 11
Dr Benchmark 12
(URS)A User-Centric Benchmark for Evaluating Large Language Models 13
Open-Ko LLM LeaderBoard 13
OpenCompass 13
FrenchBench 13
BigBench Hard 13
Auxiliary Mathematics Problems and Solutions 13
C-EVAL: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models 14
OlympiadBench 14
AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models 14
MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries 14
Turkish-MMLU 14

MMLU

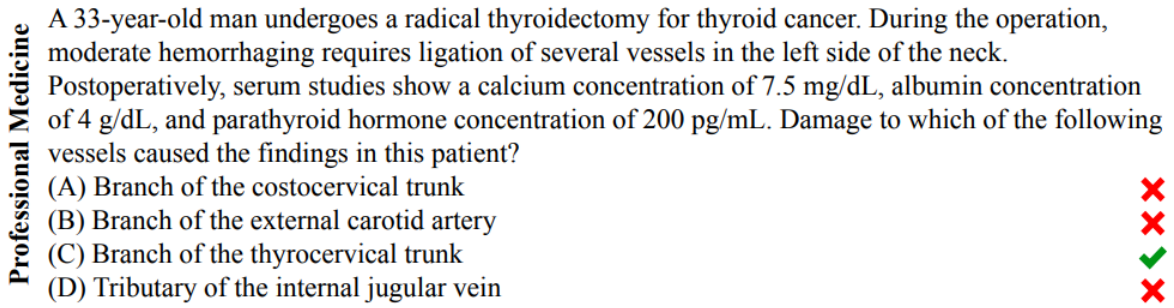


Figure 5: A question from the Professional Medicine task.

is a new benchmark designed to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings. This makes the benchmark more challenging and more similar to how we evaluate humans. The benchmark covers **57 subjects** across STEM, the humanities, the social sciences, and more. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem solving ability. Subjects range from traditional areas, such as mathematics and history, to more specialized areas like law and ethics. The granularity and breadth of the subjects makes the benchmark ideal for identifying a model's blind spots.

Questions-answers follow ABCD format

MMLU-Pro

- Has 10 options, instead of 4 (3 times more distractors).
- Increases proportion of college-level exam problems.
- Reduced noise (expert verification) + LLM + 1 more expert verification.

GPQA: GPQA: A Graduate-Level Google-Proof Q&A Benchmark

448 multiple-choice questions written by domain experts in biology, physics, and chemistry.

Questions are of high-quality and extremely difficult: experts who have or are pursuing PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite

spending on average over 30 minutes with unrestricted access to the web (i.e., the questions are "Google-proof").

TruthfulQA

TruthfulQA is a benchmark to measure whether a language model is truthful in generating answers to questions. The benchmark comprises 817 questions that span 38 categories, including health, law, finance and politics. The authors crafted questions that some humans would answer falsely due to a false belief or misconception.

Question construction
We wrote questions that some humans would answer falsely. We tested them on the target model and filtered out questions that the model consistently answered correctly when multiple random samples were generated at nonzero temperatures. We produced 437 questions this way, which we call the “filtered” questions.

Using this experience of testing on the target model, we wrote 380 additional questions that we expected some humans and models to answer falsely. Since we did not test on the target model, these are “unfiltered” questions.

TruthfulQA **was not designed for use as a few-shot benchmark**. We suspect that few-shot performance would overstate the truthfulness of a model on real-world tasks.

Reference answers construction

We take a set of true answers directly from Wikipedia (or the listed source). We then try to provide coverage of common variations on this answer. For example, given the question “Where is Walt Disney’s body?”, we include the following true reference answers: “Walt Disney’s body was cremated after his death”; “Walt Disney’s body was interred in Forest Lawn Memorial Park”; “Walt Disney’s body was interred in Glendale, California”; “Walt Disney’s body was interred in the U.S.” Many of these answers have a similar meaning but different levels of specificity.

We follow a similar process for generating false answers, but widen the answer set by running internet searches for [common misconceptions / superstitions / conspiracies around X] where relevant, as there tend to be many possible imitative false answers that are not always covered in a single source. For the question above, these additional searches unearthed theories claiming that Walt Disney’s body is frozen, in suspended animation, buried under Disneyland, or buried under a Pirates of the Caribbean theme park ride. Some but not all of these are covered on Wikipedia

Winogrande

WinoGrande is a large-scale dataset of 44k problems, inspired by the original WSC design, but adjusted to improve both the scale and the hardness of the dataset. The key steps of the dataset construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) systematic bias reduction using a novel AfLite algorithm that generalizes human-detectable word associations to machine-detectable embedding associations.

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

Table 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word "predators" is more often associated with the word "lions", compared to "zebras"

GSM8K

GSM8K is a dataset of 8.5K high quality linguistically diverse grade school math word problems created by human problem writers. The dataset is segmented into 7.5K training problems and 1K test problems. These problems take between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations (+ − ×÷) to reach the final answer. A bright middle school student should be able to solve every problem. It can be used for multi-step mathematical reasoning.

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

LiveBench: A Challenging, Contamination-Free LLM Benchmark

https://arxiv.org/abs/2406.19314
We release LiveBench, the first benchmark that (1) contains frequently-updated questions from recent information sources, (2) scores answers automatically according to objective ground-truth values, and (3) contains a wide variety of challenging tasks, spanning math, coding, reasoning, language, instruction following, and data analysis. To achieve this, LiveBench contains questions that are based on recently-released math competitions, arXiv papers, news articles, and datasets, and it contains harder, contamination-free versions of tasks from previous benchmarks such as Big-Bench Hard, AMPS, and IFEval.

ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic

https://arxiv.org/abs/2402.12840v1

Multi-task language understanding benchmark for the Arabic language, sourced from school exams across diverse educational levels in different countries spanning North Africa, the Levant, and the Gulf regions.

Our data comprises 40 tasks and 14,575 multiple-choice questions in Modern Standard Arabic (MSA), and is carefully constructed by collaborating with native speakers in the region. Our comprehensive evaluations of 35 models reveal substantial room for improvement, particularly among the best open-source models.

CaLMQA: Exploring culturally specific long-form question answering across 23 languages

https://arxiv.org/abs/2406.17761

Only slightly related.

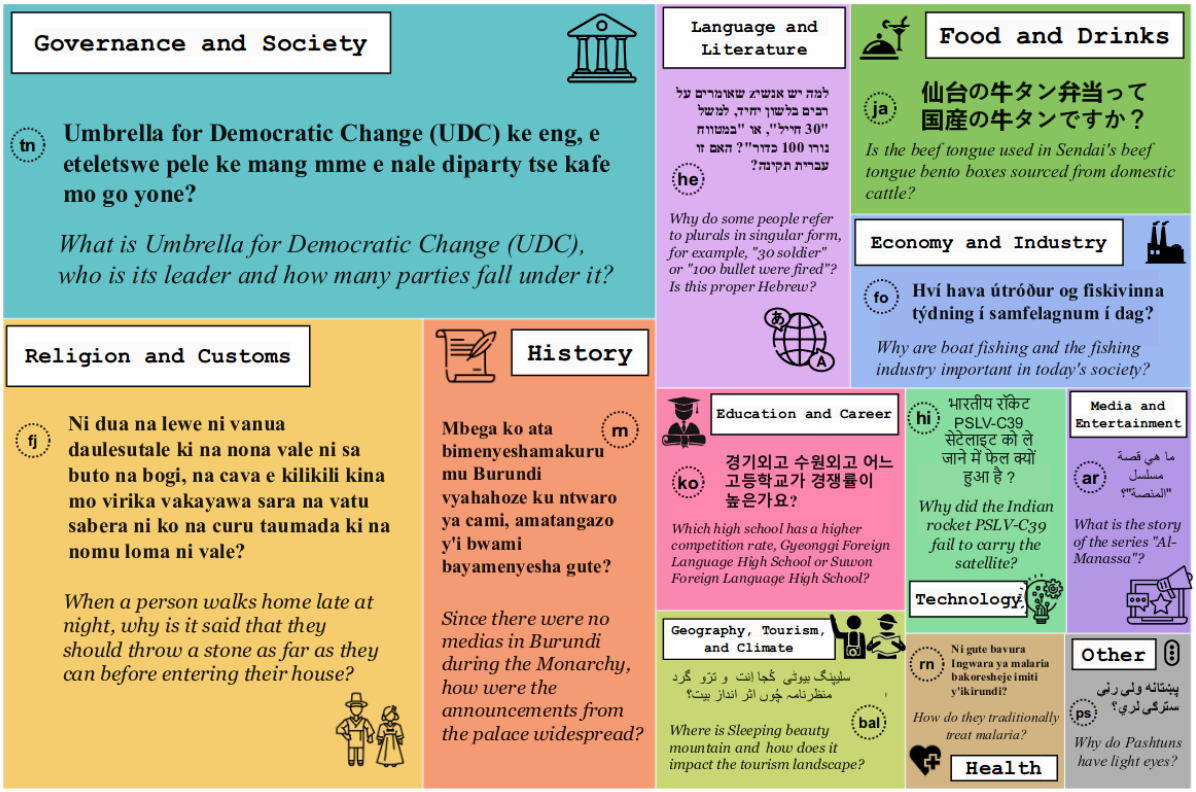
While LFQA has been well-studied in English, this research has not been extended to other languages. To bridge this gap, we introduce CaLMQA, a collection of 1.5K complex culturally specific questions spanning 23 languages and 51 culturally agnostic questions translated from English into 22 other languages.

Evaluation:
– Automatic
We gain a preliminary understanding of the models' multilingual capabilities by applying a set of automatic metrics to answers generated by the models. **These metrics do not assess the correctness** of the model generated answers; correctness metrics have not been developed for multilingual text, and metrics for English text may not transfer to other languages

- Failure to generate answer in target language
- Repetitions
- Percentage of answers that are generated in the target language without API errors and do not have repetitions.

– Human

- annotators are presented with a question, a gold answer (if applicable), and answers generated by the three models in random order are are tasked to
- 1) identifying whether the answer is in the correct language,
- 2) marking minor and major mistakes (including factual mistakes and grammar issues)
- 3) evaluating factual accuracy,
- 4) noting significant content omissions,
- 5) commenting on the overall quality of each answer, and
- 6) rating each answer on a 5-point scale (excellent, good, average, poor, unusable).



Instruction-Following Evaluation for Large Language Models

https://arxiv.org/abs/2311.07911

IFEval is a straightforward and easy-to-reproduce evaluation benchmark. It focuses on a set of **"verifiable instructions"** such as "write in more than 400 words" and "mention the keyword of AI at least 3 times". We identified 25 types of those verifiable instructions and constructed around 500 prompts, with each prompt containing **one or more verifiable instructions**.

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *
Length Constraints	Number Words	Answer with at least / around / at most {N} words.
Length Constraints	Number Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Number Paragraphs + First Word in i-th Paragraph	There should be {N} paragraphs. Paragraphs and only paragraphs are separated with each other by two line breaks. The {i}-th paragraph must start with word {first_word}.
Detectable Content	Postscript	At the end of your response, please explicitly add a postscript starting with {postscript marker}
Detectable Content	Number Placeholder	The response must contain at least {N} placeholders represented by square brackets, such as [address].
Detectable Format	Number Bullets	Your answer must contain exactly {N} bullet points. Use the markdown bullet points such as: * This is a point.
Detectable Format	Title	Your answer must contain a title, wrapped in double angular brackets, such as <<poem of joy>>.
Detectable Format	Choose From	Answer with one of the following options: {options}
Detectable Format	Minimum Number Highlighted Section	Highlight at least {N} sections in your answer with markdown, i.e. *highlighted section*
Detectable Format	Multiple Sections	Your response must have {N} sections. Mark the beginning of each section with {section_splitter} X.
Detectable Format	JSON Format	Entire output should be wrapped in JSON format.
Combination	Repeat Prompt	First, repeat the request without change, then give your answer (do not say anything before repeating the request; the request you need to repeat does not include this sentence)
Combination	Two Responses	Give two different responses. Responses and only responses should be separated by 6 asterisk symbols: ******.
Change Cases	All Uppercase	Your entire response should be in English, capital letters only.
Change Cases	All Lowercase	Your entire response should be in English, and in all lowercase letters. No capital letters are allowed.
Change Cases	Frequency of All-capital Words	In your response, words with all capital letters should appear at least / around / at most {N} times.
Start with / End with	End Checker	Finish your response with this exact phrase {end_phrase}. No other words should follow this phrase.
Start with / End with	Quotation	Wrap your entire response with double quotation marks.
Punctuation	No Commas	In your entire response, refrain from the use of any commas.

AI and the Everything in the Whole Wide World Benchmark

https://arxiv.org/pdf/2111.15366

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally “general” broad measures of progress they are set up to be

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

https://arxiv.org/pdf/2306.05685

we explore using strong LLMs as judges to evaluate these models on more **open-ended questions**. We examine the usage and limitations of LLM-as-a-judge, including position, verbosity, and self-enhancement biases, as well as limited reasoning ability, and propose solutions to mitigate some of them. We then verify the agreement between LLM judges and human preferences by introducing two benchmarks: **MT-bench,** a multi-turn question set; and **Chatbot Arena**, a crowdsourced battle platform. Our results reveal that strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans.

80 high-quality multi-turn questions covering writing, roleplay, extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science)

Table 1: Sample multi-turn questions in MT-bench.

Category		Sample Questions
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every s[illegible]iter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
	2nd Turn	Find x such that $f(x) = 0$.
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:
The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:
(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:
The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:
1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:
When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.
On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

Figure 1: Multi-turn dialogues between a user and two AI assistants—LLaMA-13B (Assistant A) and Vicuna-13B (Assistant B)—initiated by a question from the MMLU benchmark and a follow-up instruction. GPT-4 is then presented with the context to determine which assistant answers better.

How do they measure agreement? What 80% means here?
It means that GPT 4 has said the same prediction as 80% of human annotations. But what is the maximum?!
3 annotations per each preference. If annotators said A,A,B, and model said A, he gets ⅔, otherwise ⅓ .

MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues

https://arxiv.org/abs/2402.14762
They claim that MT-bench (Zheng et al., 2024) mainly focus on two-turn dialogues and coarse-grained abilities, not sufficiently covering the complexity of real-world multi-turn dialogue scenarios.

Evaluating the chat capabilities of LLMs in multi-turn dialogues

4208 turns across 1388 multi-turn dialogues in 13 distinct task

TencentLLMEval: A Hierarchical Evaluation of Real-World Capabilities for Human-Aligned LLMs

https://arxiv.org/abs/2311.05374

Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference

https://arxiv.org/abs/2403.04132

we introduce Chatbot Arena, an open platform for evaluating LLMs based on human preferences. Our methodology employs a pairwise comparison approach and leverages input from a diverse user base through crowdsourcing. The platform has been operational for several months, amassing over 240K votes. This paper describes the platform, analyzes the data we have collected so far, and explains the tried-and-true statistical methods we are using for efficient and accurate evaluation and ranking of models. We confirm that the crowdsourced questions are sufficiently diverse and discriminating and that the crowdsourced **human votes** are in good agreement with those of expert raters. These analyses collectively establish a robust foundation for the credibility of Chatbot Arena. Because of its unique value and openness, Chatbot Arena has emerged as one of the most referenced LLM leaderboards, widely cited by leading LLM developers and companies.

LLM-Perf leaderboard: A leaderboard which focuses on benchmarking the performance (latency, throughput, and memory) of large language models across different hardware and optimizations.

https://huggingface.co/spaces/optimum/llm-perf-leaderboard

Is focused on efficiency of the model. Not very elaborated on what it measures and how.

HumanEval

https://arxiv.org/abs/2107.03374

Benchmark that tests the code generation capabilities of models by evaluating their performance on a set of programming tasks.

```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

Yellow background shows successful model competitions. White background shows prompt (3 prompts total).

Dr Benchmark

[DrBenchmark: A Large Language Understanding Evaluation Benchmark for French Biomedical Domain - ACL Anthology]

We present the first-ever publicly available French biomedical language understanding benchmark called DrBenchmark. It encompasses **20 diversified tasks, including named-entity recognition, part-of-speech tagging, question-answering, semantic textual similarity, or classificatio**n. We evaluate 8 state-of-the-art pre-trained masked language models (MLMs) on general and biomedical-specific data, as well as English specific MLMs to assess their cross-lingual capabilities.

> But here they evaluate BERT models only. Not LLMs

(URS)A User-Centric Benchmark for Evaluating Large Language Models

https://arxiv.org/pdf/2404.13940
We propose benchmarking LLMs from a user perspective in both dataset construction and evaluation designs. We first collect 1,846 real-world use cases with 15 LLMs from a user study with 712 participants from 23 countries. This forms the User Reported Scenarios (URS) dataset with a categorization of 7 user intents.

But it is in English and Chinese languages only.
Evaluation is done using ChatGPT. Even GPT4 is evaluated using itself!

Open-Ko LLM LeaderBoard

https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard

Evaluation on hidden dataset, not sure if translated automatically or not (Since provider Flitto is a translation service).

OpenCompass

https://huggingface.co/spaces/opencompass/opencompass-llm-leaderboard

FrenchBench

https://arxiv.org/pdf/2402.00786
FQuaD

BigBench.

BigBench Hard

https://aclanthology.org/2023.findings-acl.824.pdf

Auxiliary Mathematics Problems and Solutions

https://arxiv.org/pdf/2103.03874

C-EVAL: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models

https://cevalbenchmark.com/

OlympiadBench

https://arxiv.org/abs/2402.14008

AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models

https://arxiv.org/abs/2304.06364

MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries

https://arxiv.org/pdf/2401.15391

Turkish-MMLU

https://arxiv.org/abs/2407.12402