

Module 2

Evaluating LLMs

Wei Dong
wdong@aaalgo.com

M I N D
A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND
INTELLIGENCE

BY A. M. TURING



Alan Turing

Computer Scientist

Department of Mathematics

a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end

from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end

LLM Evaluation

- Natural Language Processing
 - Natural language understanding
 - Natural language generation
- Reasoning
- Coding
- Domain applications
 - Math
 - Medical application
 -
- Robustness/Ethics/Biases/Trustworthiness

Natural Language Understanding

- Popular in 1st generation of LLM (GPT1, BERT, ... ~2019)
- Evaluation by categorization
- Training sets provided but external training set is essential
 - Not zero-shot/few-shot
- Examples
 - Grammatical acceptability
 - Sentiment analysis
 - Equivalence test
 - Entailment test

2018-06 – GPT 1.0: Natural Language Understanding

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	93.2	-	-	-	-
TF-KLD [23]	-	-	86.0	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	45.4	91.3	82.3	82.0	70.3	72.8

- **Grammatical acceptability**
- 10657 sentences from 23 linguistics publications
- 90% training, 10% held out

```
gj04 1 Bill drank from the hose.
gj04 1 This metal hammers flat easily.
gj04 1 They made him president.
gj04 1 They made him angry.
gj04 0 * They caused him to become angry by making him.
gj04 0 * They caused him to become president by making him.
gj04 0 * They made him to exhaustion.
gj04 1 They made him into a monster.
gj04 1 The trolley rumbled through the tunnel.
gj04 1 The wagon rumbled down the road.
gj04 1 The bullets whistled past the house.
gj04 1 The knee replacement candidate groaned up the stairs.
gj04 0 * The car honked down the road.
gj04 0 * The dog barked out of the room.
gj04 1 The dog barked its way out of the room.
gj04 1 Bill whistled his way past the house.
gj04 1 The witch vanished into the forest.
gj04 1 Bill disappeared down the road.
gj04 0 * The witch went into the forest by vanishing.
gj04 1 The witch went into the forest and thereby vanished.
gj04 1 The building is tall and wide.
```

GPT-1	2018-06	45.4
BERT	2018-10	60.5
Kaggle Leaderboard As of 2024-08		0.74

SST2: Stanford Sentiment Treebank v2

- 11855 hand-annotated sentences from movie reviews
- GPT-1: 91.3, BERT: 94.9

1 a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films
0 apparently reassembled from the cutting room floor of any given daytime soap
0 they presume their audience wo n't sit still for a sociology lesson , however entertainingly presented , so
1 this is a visually stunning rumination on love , memory , history and the war between art and commerce
1 jonathan parker 's bartleby should have been the be all end all of the modern office anomie films
1 campanella gets the tone just right funny in the middle of sad in the middle of hopeful
0 a fan film that for the uninitiated plays better on video with the sound turned down
1 b art and berling are both superb , while huppert is magnificent
0 a little less extreme than in the past , with longer exposition sequences between them , and with fewer g
0 the film is strictly routine
1 a lyrical metaphor for cultural and personal self discovery and a picaresque view of a little remembered v
0 the most repugnant adaptation of a classic text since roland joff and demi moore 's the scarlet letter
0 for something as splendid looking as this particular film , the viewer expects something special but instead
1 this is a stunning film , a one of a kind tour de force
1 may be more genial than ingenious , but it gets the job done

MRPC: Microsoft Research Paraphrase Corpus

- Whether two sentences are semantically equivalent
- 5801 sentence pairs

```
{'sentence1': 'Martin , 58 , will be freed today after serving two thirds of his five-year sentence for the manslaughter of 16-year-old Fred Barras .', 'sentence2': 'Martin served two thirds of a five-year sentence for the manslaughter of Barras and for wounding Fearon .', 'label': 0, 'idx': 4072}
{'sentence1': '" We have concluded that the outlook for price stability over the medium term has improved significantly since our last decision to lower interest rates , " Duisenberg said .', 'sentence2': 'In a statement , the ECB said the outlook for price stability over the medium term had " improved significantly " since its last decision to lower interest rates in March .', 'label': 1, 'idx': 4073}
{'sentence1': 'The notification was first reported Friday by MSNBC .', 'sentence2': 'MSNBC.com first reported the CIA request on Friday .', 'label': 1, 'idx': 4074}
{'sentence1': '"The 30-year bond US30YT = RR rose 22 / 32 for a yield of 4.31 percent , versus 4.35 percent at Wednesday 's close .", 'sentence2': 'The 30-year bond US30YT = RR grew 1-3 / 32 for a yield of 4.30 percent , down from 4.35 percent late Wednesday .', 'label': 0, 'idx': 4075}
```

GPT-1	2018-06	82.3
BERT	2018-10	89.3
Glue Leaderboard As of 2024-08		94.2

MNLI: The Multi-Genre NLI Gorpus NYU 2017

- 43K crowd-sourced sentence pairs
- 1st => 2nd: entailment 0, neutral 1 or Contradiction 2, accuracy
- GPT-1: 82.1/81.4; BERT: 86.7/85.9; SOTA: 92.5/92.1

```
{'premise': 'Postal Service were to reduce delivery frequency.',  
, 'hypothesis': 'The postal service could deliver less frequently.', 'label': 0, 'idx': 29}  
{'premise': "there's a uh a couple called um oh i'm going to forget his name now uh Dirkson", 'hypothesis': "I can't remember their name", 'label': 0, 'idx': 36}  
{'premise': 'She smiled back.', 'hypothesis': "She was so happy she couldn't stop smiling.", 'label': 1, 'idx': 48}  
{'premise': 'Were they in there?', 'hypothesis': 'Were they supposed to be in there?', 'label': 1, 'idx': 33}  
{'premise': 'but that takes too much planning', 'hypothesis': "It doesn't take much planning.", 'label': 2, 'idx': 52}  
{'premise': "it's slow it's uh there are many better machines on the market right now for", 'hypothesis': "This is the fastest machine, you won't find a better machine.", 'label': 2, 'idx': 56}
```

Genre	Train
<i>SNLI</i>	550,152
FICTION	77,348
GOVERNMENT	77,350
SLATE	77,306
TELEPHONE	83,348
TRAVEL	77,350
9/11	0
FACE-TO-FACE	0
LETTERS	0
OUP	0
VERBATIM	0
MultiNLI Overall	392,702

GLUE: A Multi-Task Benchmark for Natural Language Understanding

- 9 NLU tasks
- Single-Sentence: CoLA, SST-2
- Similarity: MRPC, QQP, STS-B
- Inference: MNLI, QNLI, RTE, WNLI
- GPT-1: 72.8

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6	🔗	91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5	🔗	91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	97.9	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE	🔗	91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER	🔗	91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	🔗	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6

HellaSwag: An Early Benchmark of Reasoning Still Used

```
{
  "activity_label": "Removing ice from car",
  "ctx": "Then, the man writes over the snow covering the window of a car,
and a woman wearing winter clothes smiles. then",
  "endings": [
    ", the man adds wax to the windshield and cuts it.",
    ", a person board a ski lift, while two men supporting the head of the
person wearing winter clothes snow as the we girls sled.",
    ", the man puts on a christmas coat, knitted with netting.",
    ", the man continues removing the snow on his car."
  ],
  "label": "3"
}
```

40K train
10K validation
10K test

		gpt-4o	Meta-Llama-3.1-405B	Meta-Llama-3.1-70B	Meta-Llama-3-70B	Meta-Llama-3.1-8B	Meta-Llama-3-8B
"activity_label": "A female large kitchen pre	boolq	0.905	0.921	0.909	0.892	0.871	0.82
"endings": ["contain egg "are then spr "are placed i "are filled w	gsm8k	0.942	0.968	0.948	0.833	0.844	0.572
], "label": "3"	hellaswag	0.891	0.92	0.908	0.874	0.768	0.462

The Llama 3 Herd of Models

Llama Team, AI @ Meta¹

¹A detailed contributor list can be found in the appendix of this paper.

Category	Benchmark	Llama 3 8B	Gemini 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 _(or12S)	GPT-4o	Claude 3.5 Sonnet
General	MMLU _(5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU _(0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [◊]	85.4	88.7	88.3
	MMLU-Pro _(5-shot, CoT)	48.3	–	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval _(0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus _(0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K _(8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH _(0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge _(0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA _(0-shot, CoT)	32.8	–	28.8	46.7	33.3	30.8	51.1	–	41.4	53.6	59.4
Tool use	BFCL	76.1	–	60.4	84.8	–	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	95.2	–	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	83.4	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	100.0	100.0	90.8
Multilingual	MGSM _(0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	–	85.9	90.5	91.6

IFEval: Instruction Following Google & Yale, 2023

- 25 verifiable instructions, 541 prompts
- Each instruction is tested with its own heuristics in Python

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}.

IFEval, More Examples

		... of square brackets, such as <<poem>>.
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.

IFEval Leaderboard

Models	Prompt-level strict-accuracy (%)	Inst-level strict-accuracy (%)	Prompt-level loose-accuracy (%)	Inst-level loose-accuracy (%)
GPT-4	76.89	83.57	79.30	85.37
PaLM 2 S	43.07	55.76	46.95	59.11

Rank	Model	Params (B)	Repo	Quantization	Final Score	Strict Prompt Score	Strict Inst Score	Loose Prompt Score	Loose Inst Score
1	LLama 3 70B	70	perplexity/llama-3-sonar-large-32k-chat (Openrouter)	FP16	0.8331	0.7671	0.8381	0.8373	0.8897
2	LLama 3 70B	70	llama3-70b-8192 (Groq)	FP16	0.8175	0.7431	0.8213	0.8244	0.8813
3	Qwen 2 72B	72	qwen/qwen-2-72b-instruct (Openrouter)	FP16	0.8056	0.7579	0.8357	0.7781	0.8513
4	Llama 3 8B	8	QuantFactory/NeuralDaredevil-8B-abliterated-GGUF	Q8_0	0.7884	0.7357	0.8058	0.7763	0.8357

Measuring Massive Multitask Language Understanding Berkeley 2021

- 57 subjects: STEM, humanities, social science ...;
- 15908 questions in all, test set is 88.5%
- Elementary level to advanced professional level
- Multiple choice (A/B/C/D)

```
abstract_algebra_test.csv      high_school_chemistry_test.csv      management_test.csv
anatomy_test.csv              high_school_computer_science_test.csv  marketing_test.csv
astronomy_test.csv          high_school_european_history_test.csv  medical_genetics_test.csv
business_ethics_test.csv     high_school_geography_test.csv        miscellaneous_test.csv
clinical_knowledge_test.csv  high_school_government_and_politics_test.csv  moral_disputes_test.csv
college_biology_test.csv     high_school_macroconomics_test.csv    moral_scenarios_test.csv
college_chemistry_test.csv   high_school_mathematics_test.csv      nutrition_test.csv
college_computer_science_test.csv  high_school_microeconomics_test.csv  philosophy_test.csv
college_mathematics_test.csv  high_school_physics_test.csv          prehistory_test.csv
college_medicine_test.csv    high_school_psychology_test.csv       professional_accounting_test.csv
college_physics_test.csv     high_school_statistics_test.csv       professional_law_test.csv
computer_security_test.csv   high_school_us_history_test.csv       professional_medicine_test.csv
conceptual_physics_test.csv  high_school_world_history_test.csv    professional_psychology_test.csv
econometrics_test.csv       human_aging_test.csv                 public_relations_test.csv
electrical_engineering_test.csv  human_sexuality_test.csv            security_studies_test.csv
elementary_mathematics_test.csv  international_law_test.csv          sociology_test.csv
formal_logic_test.csv       jurisprudence_test.csv              us_foreign_policy_test.csv
global_facts_test.csv       logical_fallacies_test.csv          virology_test.csv
high_school_biology_test.csv  machine_learning_test.csv          world_religions_test.csv
```

MMLU Examples

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. ✗

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 ✗

MMLU Examples

- 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. ✔

Figure 3: Examples from the Microeconomics task.

- 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.

Human Performance

- Unspecified humans from Amazon Mechanical Turk: 34.5%
- US Medical Licensing Examination 95th percentile : 87%
- Educated guess of overall 95th percentile: 89.8%

MMLU Leaderboard

Model	MMLU All Subjects - EM	Abstract Algebra - EM	Anatomy - EM	College Chemistry - EM	Computer Security - EM	Econometrics - EM	Global Facts -
Claude 3.5 Sonnet (20240620)	0.865	0.75	0.844	0.59	0.89	0.807	0.72
Claude 3 Opus (20240229)	0.846	0.64	0.8	0.6	0.85	0.789	0.66
Llama 3.1 Instruct Turbo (405B)	0.845	0.7	0.822	0.6	0.81	0.746	0.71
GPT-4o (2024-05-13)	0.842	0.66	0.911	0.61	0.85	0.693	0.64
Gemini 1.5 Pro (001)	0.827	0.75	0.83	0.62	0.83	0.728	0.66
GPT-4 (0613)	0.824	0.63	0.8	0.55	0.86	0.684	0.62
Qwen2 Instruct (72B)	0.824	0.67	0.793	0.65	0.85	0.737	0.58
GPT-4 Turbo (2024-04-09)	0.813	0.56	0.822	0.53	0.83	0.675	0.58
Gemini 1.5 Pro (0409 preview)	0.81	0.6	0.77	0.58	0.81	0.737	0.66
Llama 3.1 Instruct Turbo (70B)	0.801	0.55	0.8	0.59	0.8	0.675	0.61
Mistral Large 2 (2407)	0.8	0.7	0.785	0.52	0.81	0.693	0.56
GPT-4 Turbo (1106 preview)	0.796	0.53	0.807	0.47	0.86	0.675	0.58

CPQA: A Graduate-Level Google-Proof Q&A Benchmark 2021

- Difficult domain questions in physics, chemistry & biology
- Human expert 65%, non-expert 34%
 - Non-expert (PhD in other domains) spent 37min/problem w/ Internet
- GPT4o: 53.6
- Llama 3.1: 32.8/46.7/51.11

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?

- A) A metal compound from the fifth period.
 - B) A metal compound from the fifth period and a non-metal compound from the third period.
 - C) A metal compound from the fourth period.
 - D) A metal compound from the fourth period and a non-metal compound from the second period.
-

More GPQA Examples

Genetics

If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality?

- A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species.
 - B) Epistatic interactions between the genes of different species
 - C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote.
 - D) Chromosomal recombination will not occur in different species.
-

Molecular Biology

A scientist studies the stress response of barley to increased temperatures and finds a protein which contributes to heat tolerance through the stabilisation of cell membrane. The scientist is very happy and wants to create a heat-tolerant cultivar of diploid wheat. Using databases, they find a heat tolerance protein homologue and start analysing its accumulation under heat stress. Soon enough, the scientist discovers this protein is not synthesised in the wheat cultivar they study. There are many possible reasons for such behaviour, including:

- A) A miRNA targets the protein, which makes exonucleases cut it immediately after the end of translation and before processing in ER
 - B) Trimethylation of lysine of H3 histone in position 27 at the promoter of the gene encoding the target protein
 - C) A stop-codon occurs in the 5'-UTR region of the gene encoding the target protein
 - D) The proteolysis process disrupts a quaternary structure of the protein, preserving only a tertiary structure
-

Astrophysics

Astronomers are studying a star with a T_{eff} of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines ($E_W < 100 \text{ mÅ}$) of two chemical elements, E11 and E12. Given the atmospheric temperature of the star, E11 is mostly in the neutral phase, while E12 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?

- A) E12 I (neutral)
 - B) E11 II (singly ionized)
 - C) E12 II (singly ionized)
 - D) E11 I (neutral)
-

HumanEval OpenAI 2021

- 164 handwritten programming problems
- Relatively easy tasks.

```
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]
```

```
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)
```

HumanEval Leaderboard

LLama3.1

405B 88.6
70B 86.0
8B 72.8

#	Model	pass@1
1	GPT-4-Turbo (April 2024) 🌟	≤86.6
2	DeepSeek-Coder-V2-Instruct 🌟	≤82.3
3	GPT-4-Turbo (Nov 2023) 🌟	≤81.7
4	GPT-4 (May 2023) 🌟	≤79.3
5	CodeQwen1.5-7B-Chat 🌟	≤78.7
6	claude-3-opus (Mar 2024) 🌟	≤77.4
7	DeepSeek-Coder-33B-instruct 🌟	≤75
8	OpenCodeInterpreter-DS-33B 🌟❤️	≤73.8
9	WizardCoder-33B-V1.1 🌟	≤73.2
10	Artigenz-Coder-DS-6.7B 🌟	≤72.6
11	Llama3-70B-instruct 🌟	≤72
12	Mixtral-8x22B-Instruct-v0.1 🌟	≤72
13	OpenCodeInterpreter-DS-6.7B 🌟❤️	≤72
14	speechless-codellama-34B-v2.0 🌟❤️	≤72
15	DeepSeek-Coder-6.7B-instruct 🌟	≤71.3

MBPP: Mostly Basic Python Programs [Google 2021](#)

- 1000 python programs with test cases [LINK](#)
- Entry-level programming

```
- {  
  source_file: "Benchmark Questions Verification V2.ipynb",  
  task_id: 235,  
  prompt: "Write a python function to set all even bits of a given number.",  
  code: "def even_bit_set_number(n):  
    count = 0; res = 0; temp = n  
    while(temp > 0):  
      if (count % 2 == 1):  
        res |= (1 << count)  
      count+=1  
      temp >>= 1  
    return (n | res) ",  
  test_imports: [ ],  
  test_list: [  
    "assert even_bit_set_number(10) == 10",  
    "assert even_bit_set_number(20) == 30",  
    "assert even_bit_set_number(30) == 30"  
  ]  
},
```

MBPP Leaderboard

LLama3.1

#	Model	pass@1
1	DeepSeek-Coder-V2-Instruct 🌟	<75.1
2	GPT-4-Turbo (Nov 2023) 🌟	<73.3
3	claude-3-opus (Mar 2024) 🌟	<73.3
4	DeepSeek-Coder-33B-instruct 🌟	<70.1
5	GPT-3.5-Turbo (Nov 2023) 🌟	<69.7
6	Artigenz-Coder-DS-6.7B 🌟	<69.6
7	claude-3-sonnet (Mar 2024) 🌟	<69.3
8	CodeQwen1.5-7B-Chat 🌟	<69
9	Llama3-70B-instruct 🌟	<69
10	Magicoder-S-DS-6.7B 🌟❤️	<69
11	claude-3-haiku (Mar 2024) 🌟	<68.8
12	OpenCodeInterpreter-DS-33B 🌟❤️	<68.5
13	WhiteRabbitNeo-33B-v1 🌟	<66.9
14	OpenCodeInterpreter-DS-6.7B 🌟❤️	<66.4
15	DeepSeek-Coder-6.7B-instruct 🌟	<65.6

405B
70B
8B

89
80.5
72.6

GSM 8K: Grade School Math 2021

- 8.5K problems, 2-8 steps to solve, answers are mostly integer
- Llama 3.1: 84.5/95.1/96.8

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 $4 \times 2 = \ll 4 \times 2 = 8 \gg 8$ dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = \ll 12 \times 8 = 96 \gg 96$ cookies

She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = \ll 96/16 = 6 \gg 6$ 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 = $\ll 68 - 18 = 50 \gg 50$ gallons this morning.

So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = $\ll 68 + 82 + 50 = 200 \gg 200$ gallons.

She was able to sell 200 gallons - 24 gallons = $\ll 200 - 24 = 176 \gg 176$ gallons.

Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \$\ll 3.50 \times 176 = 616 \gg 616$.

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 $3 \times 12 = \ll 3 \times 12 = 36 \gg 36$ sodas

6 people attend the party, so half of them is $6/2 = \ll 6/2 = 3 \gg 3$ people

Each of those people drinks 3 sodas, so they drink $3 \times 3 = \ll 3 \times 3 = 9 \gg 9$ sodas

Two people drink 4 sodas, which means they drink $2 \times 4 = \ll 4 \times 2 = 8 \gg 8$ sodas










With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = \ll 5 + 9 + 8 + 3 = 25 \gg 25$ sodas

As Tina started off with 36 sodas, that means there are $36 - 25 = \ll 36 - 25 = 11 \gg 11$ sodas left

Final Answer: 11

2024 AI Math Olympiad

- 50 AIME-level problems
- Hardware limitation: 2x16G GPU
- Most teams use 8B model based on Deepseek-Math

#	Team	Members	Score	Entries	Last	Solution
1	Numina		29	2	2mo	
2	CMU_MATH		22	2	2mo	
3	after exams		21	2	3mo	
4	codeinter		21	2	2mo	
5	Conor #2		20	2	3mo	

Formal Mathematical Proving

- Presented in formal languages that be automatically verified
 - Lean, metamath, Isabelle, Hol Light
- MiniF2F
 - High School competition level, 244 validation and 244 test
- ProofNet
 - 371 undergraduate-level in lean only

MiniF2F and ProofNet Examples

```
theorem amc12_2000_p11
  (a b : ℝ)
  (h₀ : a ≠ 0 ∧ b ≠ 0)
  (h₁ : a * b = a - b) :
  a / b + b / a - a * b = 2 :=
begin
  field_simp [h₀.1, h₀.2],
  simp only [h₁, mul_comm, mul_sub],
  ring,
end
```

Two non-zero **real numbers**, a and b , satisfy $ab = a - b$. Which of the following is a possible value of $\frac{a}{b} + \frac{b}{a} - ab$?

- (A) -2 (B) $-\frac{1}{2}$ (C) $\frac{1}{3}$ (D) $\frac{1}{2}$ (E) 2

ProofNet dataset (ours)

Formal theorem statement:

```
theorem exercise_4_5_14 {G : Type*}
  [group G] [fintype G]
  (hG : card G = 312) :
  ∃ (p : ℕ) (P : sylow p G),
  P.normal
```

Natural language theorem statement:

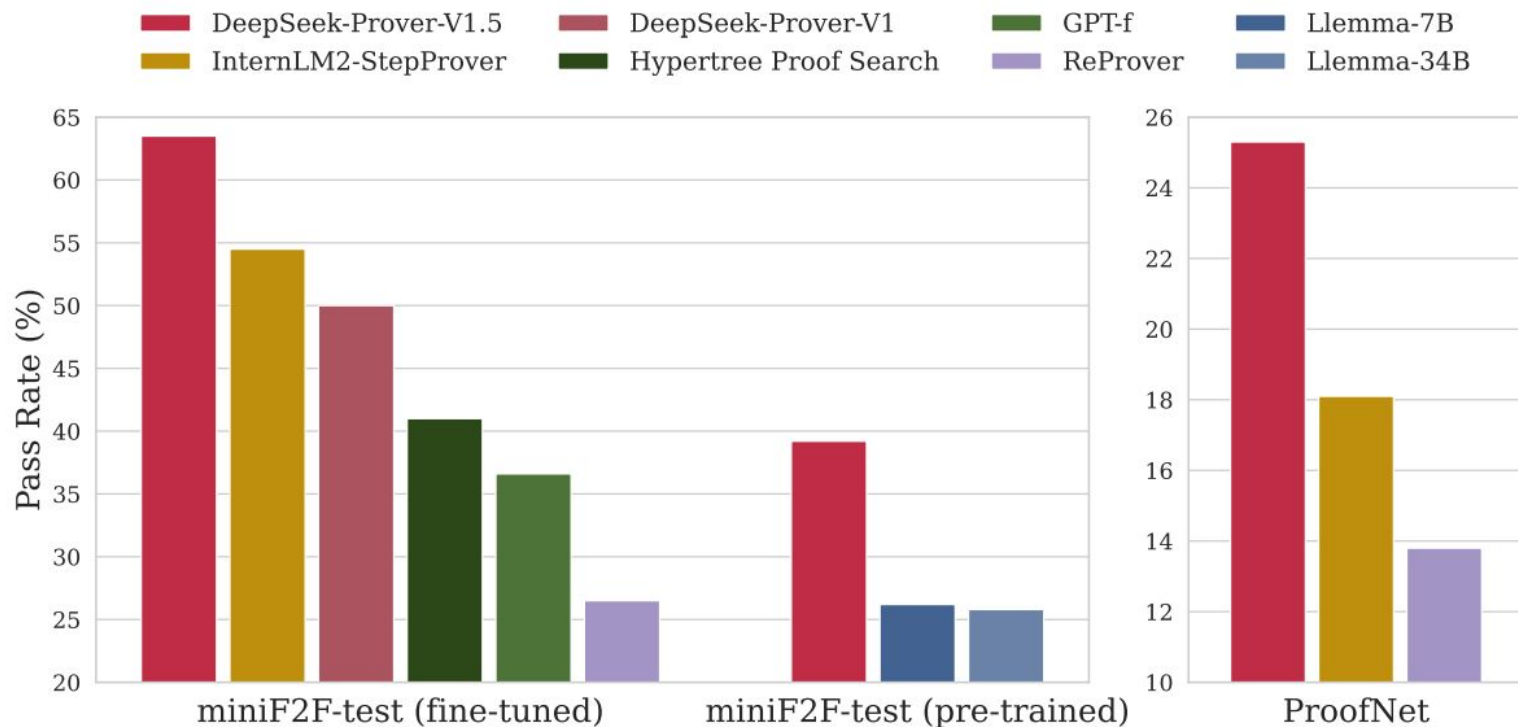
Prove that a group of order 312 has a normal Sylow p -subgroup for some prime p dividing its order.

Natural language proof:

Proof. Let n_{13} be the number of Sylow 13-subgroups of G . Then by Sylow's Theorem, $n_{13} \equiv 1 \pmod{13}$ and n_{13} divides $2^3 \cdot 3 = 24$. This implies $n_{13} = 1$, so that there is only one Sylow 13-subgroup, which is consequently normal. The last assertion follows from the fact conjugation preserves the order of a subgroup. So if there is only one subgroup H of order 13, then for any $g \in G$, we have $|gHg^{-1}| = |H| = 13$, so $gHg^{-1} = H$, i.e. H is normal. \square

MiniF2F and ProofNet Results

[DeepSeek-Prover paper, Aug 15, 2024](#)



ARC: Abstraction & Reasoning Corpus 2018

- 7787 science questions, 4-way multiple choice
- Llama 3.1: 83.4/94.8/96.9

Knowledge Type	Example
Definition	What is a worldwide increase in temperature called? (A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating
Basic Facts & Properties	Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon
Structure	The crust, the mantle, and the core are structures of Earth. Which description is a feature of Earth's mantle? (A) contains fossil remains (B) consists of tectonic plates (C) is located at the center of Earth (D) has properties of both liquids and solids
Processes & Causal	What is the first step of the process in the formation of sedimentary rocks? (A) erosion (B) deposition (C) compaction (D) cementation
Teleology / Purpose	What is the main function of the circulatory system? (1) secrete enzymes (2) digest proteins (3) produce hormones (4) transport materials
Algebraic	If a red flowered plant (RR) is crossed with a white flowered plant (rr), what color will the offspring be? (A) 100% pink (B) 100% red (C) 50% white, 50% red (D) 100% white
Experiments	Scientists perform experiments to test hypotheses. How do scientists try to remain objective during experiments? (A) Scientists analyze all results. (B) Scientists use safety precautions. (C) Scientists conduct experiments once. (D) Scientists change at least two variables.
Spatial / Kinematic	In studying layers of rock sediment, a geologist found an area where older rock was layered on top of younger rock. Which best explains how this occurred? (A) Earthquake activity folded the rock layers...

BFCL: Berkeley Function-Calling Leaderboard 2024

- Evaluating tool using capability
 - Input 1: description of problem.
 - Input 2: specification of function.
 - Output: specification of function call.
 - Evaluation: AST matching
- Dataset Construction
 - 2251 ((question, function), answer) pairs
 - Sources: partnerships and website access logs

BFCL Demo

Model: Gorilla OpenFunctions-v2

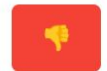
Temperature: 0.7

I'm planning a camping trip and I need to know the weather forecast. Can you fetch me the weather data for the campsite located at latitude 35.68 and longitude -121.34 for the next 10 days including daily temperature and precipitation forecasts? Also, I prefer the temperature 2 minute max in Fahrenheit and sum of precipitation in inches.

```
},
"params": {
  "latitude": {
    "type": "string",
    "description": "Geographical WGS84 coordinates of the location.
Multiple coordinates can be comma separated. E.g.,
&latitude=52.52,48.85&longitude=13.41,2.35. To return data for multiple
locations the JSON output changes to a list of structures. CSV and XLSX
formats add a column location_id. N is positive, S is negative",
    "required": true
  },
  "longitude": {
    "type": "string",
```

```
requests.get(url='https://api.open-meteo.com/v1/forecast', params=
{'latitude': '35.68', 'longitude': '-121.34', 'daily':
'temperature_2m_max,precipitation_sum', 'temperature_unit': 'fahrenheit',
'precipitation_unit': 'inch', 'forecast_days': 10})
```

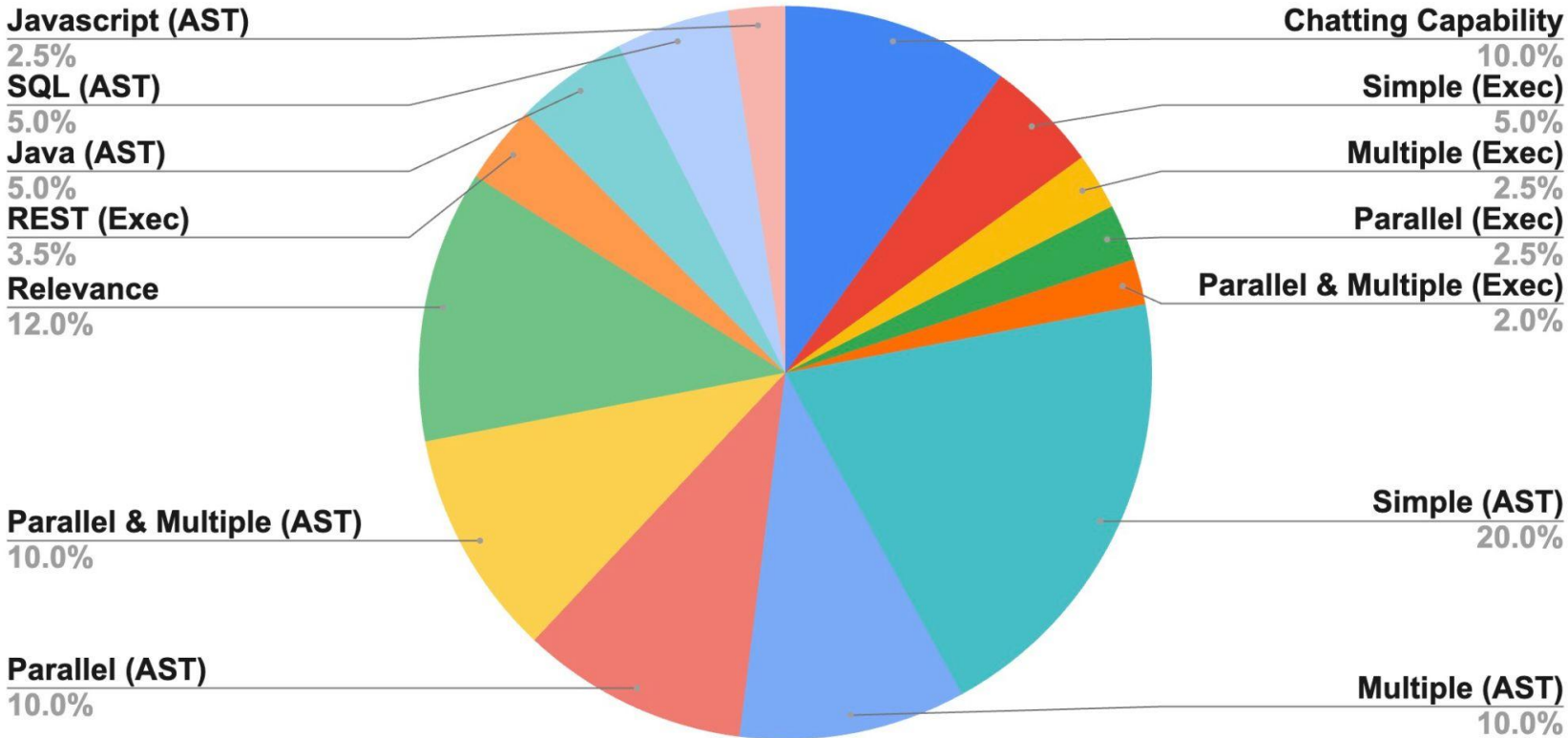
```
{
  "name": "requests.get",
  "arguments": {
    "url": "https://api.open-meteo.com/v1/forecast",
    "params": {
      "latitude": "35.68",
      "longitude": "-121.34",
      "daily": "temperature_2m_max,precipitation_sum",
      "temperature_unit": "fahrenheit",
      "precipitation_unit": "inch",
      "forecast_days": 10
    }
  }
}
```



Report Issue

BFCL Data Composition

Berkeley Function-Calling Leaderboard Evaluation Data Composition



BCF Leaderboard

Rank ▲	Overall Acc	Model
1	85.65	GPT-4-1106-Preview (Prompt)
2	84.35	GPT-4o-mini-2024-07-18 (Prompt)
3	84.23	GPT-4-0613 (Prompt)
4	84.11	GPT-4-turbo-2024-04-09 (Prompt)
5	84.09	GPT-4-0125-Preview (Prompt)
6	81.73	Functionary-Medium-v3.1 (FC)
7	80.52	Claude-3-Sonnet-20240229 (Prompt)
8	80.34	Claude-3.5-Sonnet-20240620 (Prompt)
9	79.75	Claude-3-Opus-20240229 (Prompt)
10	79.65	GPT-4-1106-Preview (FC)

Evaluation of Summarization

- [HELM: Holistic Evaluation of Language Models](#) Stanford 2023
- CNN/DailyMail
 - News articles
 - Highlights acting as summary
 - 287K (train) + 13K (val) + 11K (test) = 311K
- XSUM
 - BBC articles
 - 1st sentence is treated as summary.

Scenario: CNN/DailyMail

Input: *Two years ago, the storied Boston Marathon ended in terror and altered the lives of runners, ... Many bombing survivors... celebrating "One Boston Day," which was created to recognize acts of valor and to encourage kindness among Bostonians. ...*

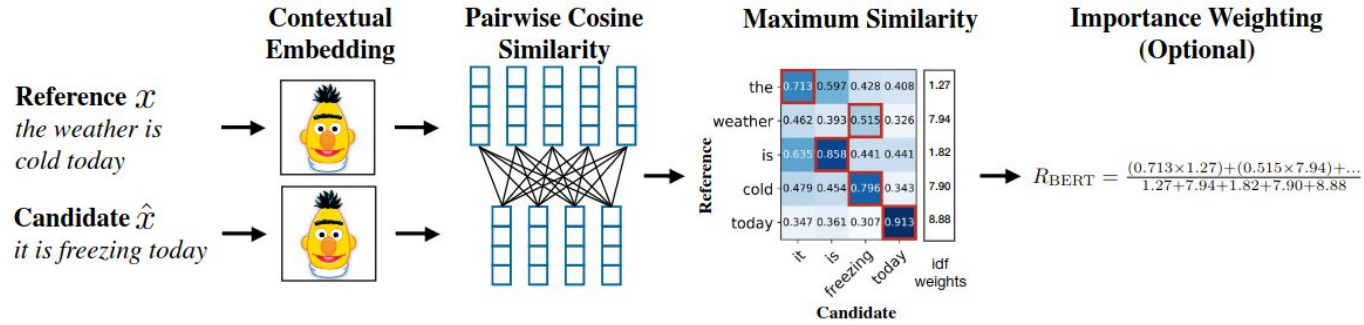
Reference: *Citizens gather to honor victims on One Boston Day, two years after the marathon bombings.*

Summarization Metrics

- ROUGE-N (acronym)

$$\begin{aligned}
 & \text{ROUGE-N} \\
 &= \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}(gram_n)}
 \end{aligned}$$

- BERT-Score



Summarization Benchmark [SOURCE](#)



Center for
Research on
Foundation
Models

HELM

Models Scenarios Results Raw runs

Summarization

In text summarization, given a piece of text (paragraph or document), the goal is to produce a much shorter summary.

[[Accuracy](#) | [Summarization metrics](#) | [Bias](#) | [Toxicity](#) | [Efficiency](#) | [General information](#) | [JSON](#)]

Accuracy

Model/adaptor	Mean win rate † [sort]	CNN/DailyMail - ROUGE-2 † [sort]	XSUM - ROUGE-2 † [sort]
Cohere Command beta (52.4B)	0.957	0.161	0.152
TNLG v2 (530B)	0.986	0.161	0.169
text-davinci-003	0.786	0.156	0.124
Anthropic-LM v4-s3 (52B)	0.857	0.154	0.134
GLM (130B)	0.829	0.154	0.132
text-davinci-002	0.857	0.153	0.144
Cohere xlarge v20221108 (52.4B)	0.886	0.153	0.153
Cohere Command beta (6.1B)	0.671	0.153	0.122
text-curie-001	0.514	0.152	0.076
text-babbage-001	0.471	0.151	0.046
Luminous Supreme (70B)	0.771	0.15	0.136
TNLG v2 (6.7B)	0.6	0.146	0.11
OPT (175B)	0.814	0.146	0.155
J1-Grande v2 beta (17B)	0.757	0.146	0.152
J1-Jumbo v1 (178B)	0.671	0.144	0.129
Cohere xlarge v20220609 (52.4B)	0.643	0.144	0.129
J1-Grande v1 (17B)	0.557	0.143	0.122

Text Summarization Benchmark from Another Source

Method	Model	Year	ROUGE-1	ROUGE-2	ROUGE-L
Extractive(unsup.)	TextRank [148]	2004	33.2	11.8	29.6
Extractive	SUMMARUNNER [154]	2017	39.60	16.20	35.30
Extractive	REFRESH [158]	2018	40.00	18.20	36.60
Extractive	BANDITSUM [50]	2018	41.50	18.70	37.60
Extractive	NEUSUM [284]	2018	41.59	19.01	37.98
Extractive (unsup.)	PACSUM [278]	2019	40.7	17.8	36.9
Extractive	HIBERT [272]	2019	42.37	19.95	38.83
Extractive	BERTSUMEXT [133]	2019	43.85	20.34	39.90
Extractive	DISCOBERT [240]	2019	43.77	20.85	40.67
Extractive	HSG [219]	2020	42.95	19.76	39.23
Extractive	MATCHSUM [279]	2020	44.41	20.86	40.55
Extractive	DiffuSum [260]	2023	44.83	22.56	40.56
Abstractive	PTGEN [187]	2017	36.44	15.66	33.42
Hybrid	BOTTOMUP [65]	2018	41.22	18.68	38.34
Abstractive	DCA [22]	2018	41.69	19.47	37.92
Abstractive	BERTSUMABS [133]	2019	41.72	19.39	38.76
Abstractive	BART [111]	2019	44.16	21.28	40.90
Abstractive	PEGASUS [266]	2020	44.17	21.47	41.11
Hybrid	GSum [51]	2020	45.94	22.32	42.48
Abstractive	SimCLS [134]	2021	46.67	22.15	43.54
Abstractive	SummIt (ChatGPT) [262]	2023	37.29	13.60	26.87
Abstractive	Element (GPT3) [224]	2023	37.75	15.20	34.25
Abstractive	SliSum (Claude2) [115]	2024	47.75	23.16	44.26
Abstractive	HADAS [231]	2024	-	-	20.12