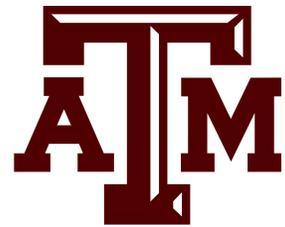


CSCSE 638 Natural Language Processing Foundation and Techniques

Lecture 12: Evaluation

Kuan-Hao Huang

Spring 2026



Project Proposal

- Due: Mar 6
- Page limit: 2 pages (excluding reference)
- Format: ACL style

Team Sign-Up

- <https://docs.google.com/spreadsheets/d/1qUZPFI4wciToJsXye8-WN4L7xVG38IWdS2GCCzmu84A/edit?usp=sharing>

Lecture Plan

- Evaluation for Classification
- Evaluation for Generation
- Evaluation Benchmarks

Evaluation for Classification

- Accuracy
 - Percentage of examples with correct predictions

Confusion Matrix, Precision, Recall, and F1 Score

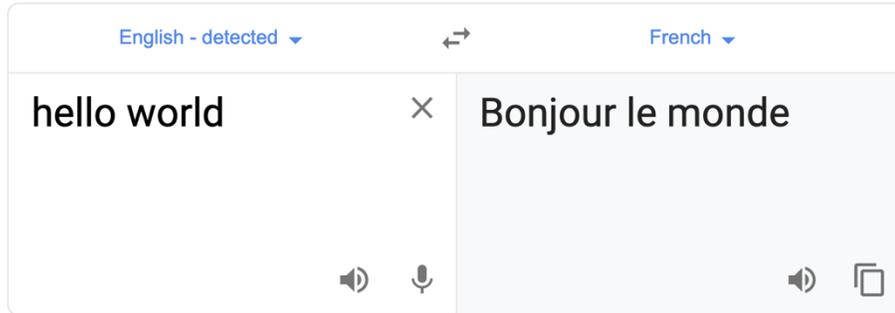
		Predicted condition	
		Positive (PP)	Negative (PN)
Actual condition	Total population = P + N		
	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

How About Generation Tasks?



I think I have an idea that should sort of improve campaign performance.

Tone Suggestion

Confident

I have an idea that should improve campaign performance.

Rephrase Dismiss

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
niki@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*¹
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin*¹
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

¹Work performed while at Facebook Brain



Summary

The document titled "Attention Is All You Need" introduces the Transformer model, a network architecture based solely on attention mechanisms, eliminating the need for recurrent or convolutional neural networks in sequence transduction tasks. The Transformer model achieves superior performance in machine translation tasks, demonstrating improved quality, parallelizability, and reduced training time compared to existing models. The key points and arguments presented in the document are as follows:

- The dominant sequence transduction models rely on complex recurrent or convolutional neural networks with an encoder-decoder structure and attention mechanisms.
- The Transformer model proposes a new architecture based solely on attention mechanisms, eliminating the need for recurrence and convolutions.
- Experiments show that the Transformer model outperforms existing models in machine translation tasks, achieving state-of-the-art results with reduced training time.
- The model utilizes self-attention to compute representations of input and output sequences, allowing for more parallelization and global dependencies.
- The Transformer model consists of stacked self-attention and fully connected layers for both the encoder and decoder, enabling efficient sequence transduction.
- Multi-Head Attention is employed to jointly attend to information from different representation subspaces at different positions, enhancing the model's performance.

Key Points:

- Transformer model introduces a network architecture based solely on attention

Perplexity

Hello world → Bonjour le monde

$$e^{-WLL(\mathcal{X})}$$

Perplexity

$$WLL(\mathcal{X}) = \frac{1}{W} \sum_{i=1}^n \log P(x_i)$$

Perplexity

Hello world → Bonjour le monde

$$e^{-WLL(\mathcal{X})}$$

Perplexity

$$WLL(\mathcal{X}) = \frac{1}{W} \sum_{i=1}^n \log P(x_i)$$

Bilingual Evaluation Understudy (BLUE)

Prediction: the cat is on mat

Reference: the cat is on the mat

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

Bilingual Evaluation Understudy (BLUE)

Prediction: the cat is on mat

Reference: the cat is on the mat

the | cat | is | on | mat

$$\text{1-gram precision} = \frac{5}{5} = 1.0$$

the cat | cat is | is on | on mat

$$\text{2-gram precision} = \frac{3}{4} = 0.75$$

the cat is | cat is on | is on mat

$$\text{3-gram precision} = \frac{2}{3} \approx 0.67$$

the cat is on | cat is on mat

$$\text{4-gram precision} = \frac{1}{2} = 0.5$$

Bilingual Evaluation Understudy (BLUE)

Prediction: the cat is on mat

Reference: the cat is on the mat

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

Brevity Penalty Factor

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\frac{1}{4} \times \log 1.0 + \frac{1}{4} \times \log 0.75 + \frac{1}{4} \times \log 0.67 + \frac{1}{4} \times \log 0.5$$

$$e^{(1-6/5)} \approx 0.82$$

$$BLEU = 0.82 \cdot \exp(-0.3465) \approx 0.579$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Prediction: the man is washing car

Reference: the man is fixing his car now

the | man | is | fixing | his | car | now

$$\text{1-gram recall} = \frac{4}{7} \approx 0.57$$

the | man | is | washing | car

$$\text{1-gram precision} = \frac{4}{5} = 0.8$$

$$\text{ROUGE-1} = \text{F1 Score} = 0.67$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Prediction: the man is washing car

Reference: the man is fixing his car now

the man | man is | is fixing | fixing his | his car | car now

$$\text{2-gram recall} = \frac{2}{6} \approx 0.33$$

ROUGE-2 = F1 Score = 0.4

the main | man is | is washing | washing car

$$\text{2-gram precision} = \frac{2}{4} = 0.5$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Prediction: the man is washing car

Reference: the man is fixing his car now

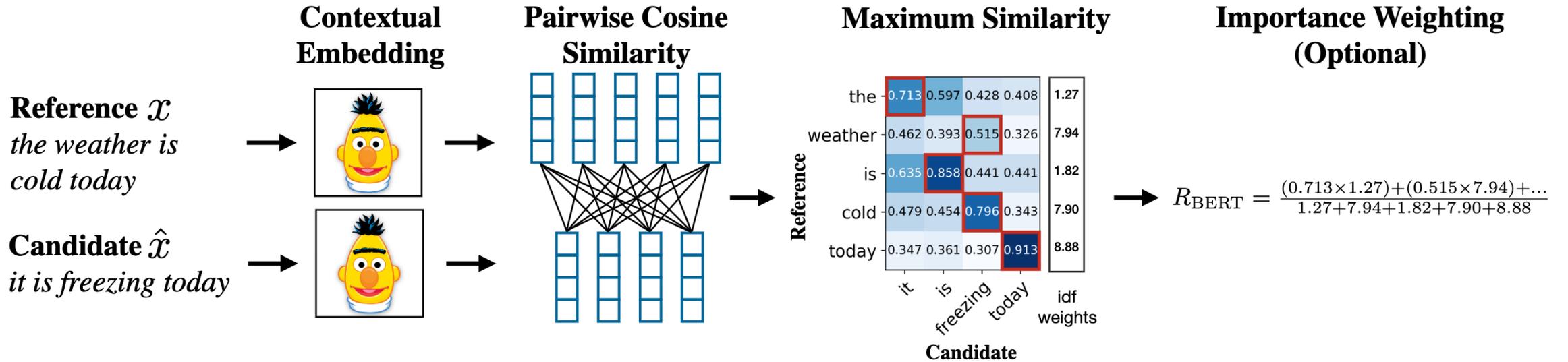
Longest Common Subsequence: **the man is car**

$$\text{LCS recall} = \frac{4}{7} \approx 0.57$$

$$\text{LCS precision} = \frac{4}{5} = 0.8$$

$$\text{ROUGE-L} = \text{F1 Score} = 0.67$$

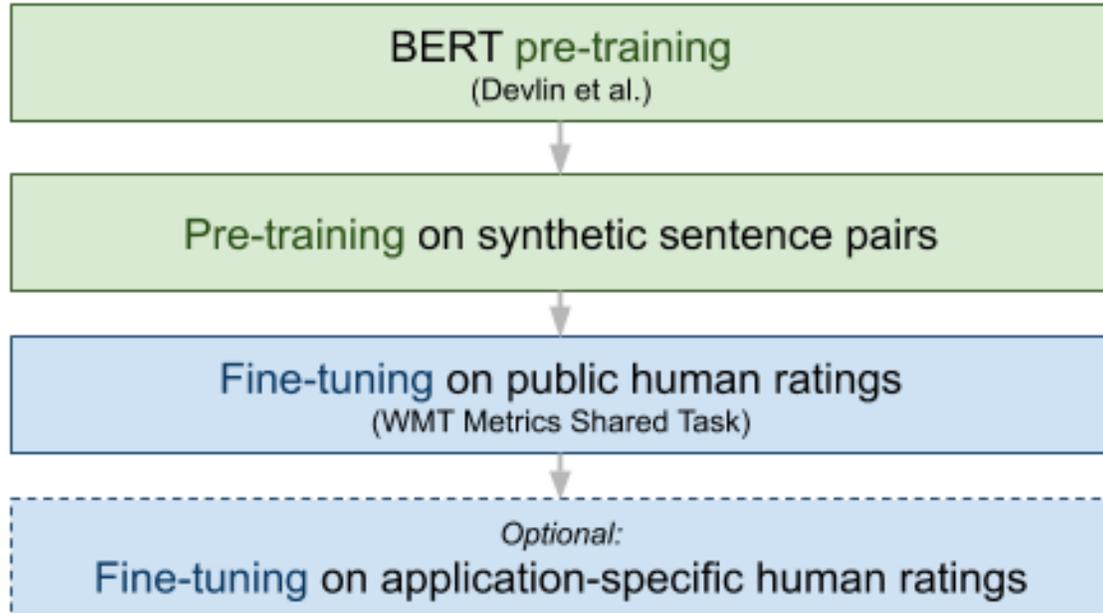
BERTScore



$$R_{\text{BERT}} = \frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} \max_{\hat{x}_j \in \hat{\mathcal{X}}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad P_{\text{BERT}} = \frac{1}{|\hat{\mathcal{X}}|} \sum_{\hat{x}_j \in \hat{\mathcal{X}}} \max_{x_i \in \mathcal{X}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

BLEURT

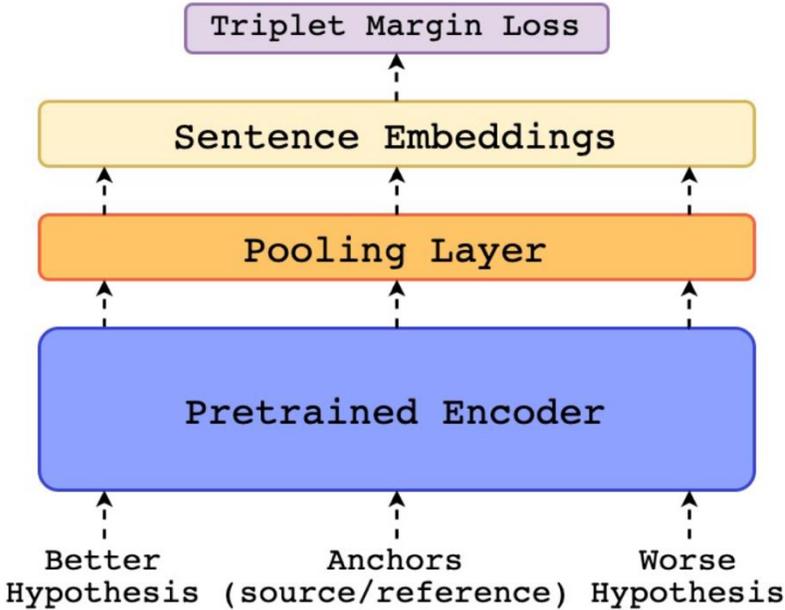
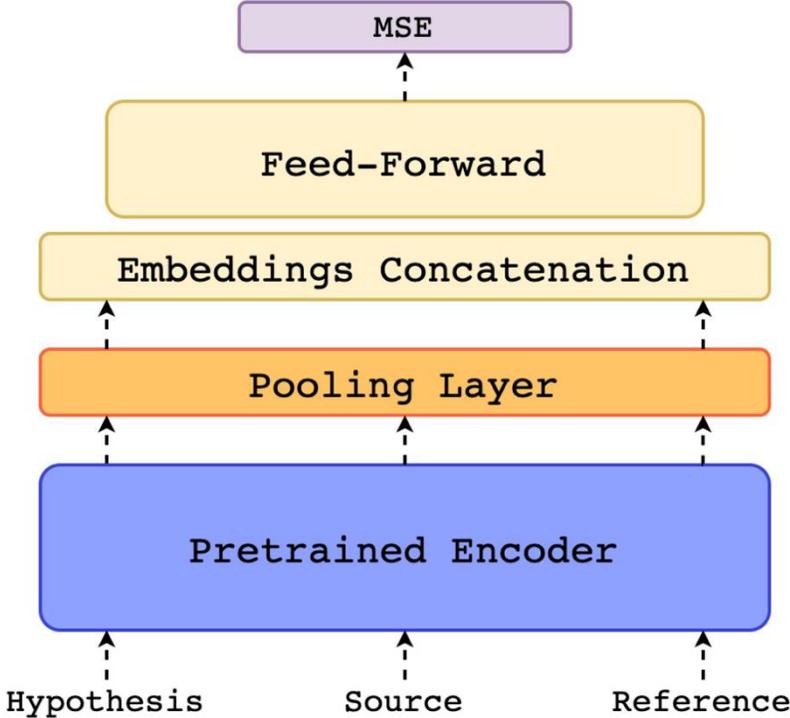
- Bilingual Evaluation Understudy with Representations from Transformers



model	cs-en τ / r	de-en τ / r	fi-en τ / r	lv-en τ / r
sentBLEU	29.6 / 43.2	28.9 / 42.2	38.6 / 56.0	23.9 / 38.2
MoverScore	47.6 / 67.0	51.2 / 70.8	NA	NA
BERTscore w/ BERT	48.0 / 66.6	50.3 / 70.1	61.4 / 81.4	51.6 / 72.3
BERTscore w/ roBERTa	54.2 / 72.6	56.9 / 76.0	64.8 / 83.2	56.2 / 75.7
chrF++	35.0 / 52.3	36.5 / 53.4	47.5 / 67.8	33.3 / 52.0
BEER	34.0 / 51.1	36.1 / 53.0	48.3 / 68.1	32.8 / 51.5
BLEURTbase -pre	51.5 / 68.2	52.0 / 70.7	66.6 / 85.1	60.8 / 80.5
BLEURTbase	55.7 / 73.4	56.3 / 75.7	68.0 / 86.8	64.7 / 83.3
BLEURT -pre	56.0 / 74.7	57.1 / 75.7	67.2 / 86.1	62.3 / 81.7
BLEURT	59.3 / 77.3	59.9 / 79.2	69.5 / 87.8	64.4 / 83.5

COMET

- Crosslingual Optimized Metric for Evaluation of Translation



LLM-as-a-Judge

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.

LLM-as-a-Judge: Pairwise Comparison

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

LLM-as-a-Judge: Single Answer Grading

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider factors
such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of
the response. Begin your evaluation by providing a short explanation. Be as objective as
possible. After providing your explanation, please rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".
```

```
[Question]
{question}
```

```
[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

LLM-as-a-Judge: Reference-Guided Grading

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer, assistant A's answer, and assistant B's answer. Your job is to evaluate which assistant's answer is better. Begin your evaluation by comparing both assistants' answers with the reference answer. Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Reference Answer]

{answer_ref}

[The End of Reference Answer]

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

Limitations of LLM-as-a-Judge

- Prompt Sensitivity
- Inconsistency
- Position bias
- Verbosity bias
- Self-enhancement bias
- Alignment with human?

Lecture Plan

- Evaluation for Classification
- Evaluation for Generation
- Evaluation Benchmarks

Why Benchmarks?

- Objective evaluation
- Comparison across models
- Tracking progress over time
- Reproducibility and fairness

Category Benchmark	Llama 3.1 70B	Llama 3.3 70B	Amazon Nova Pro	Llama 3.1 405B	Gemini Pro 1.5	GPT-4o	Claude 3.5 Sonnet
General							
MMLU Chat (0-shot, CoT)	86.0	86.0	85.9	88.6	87.1	87.5	88.9
MMLU PRO (5-shot, CoT)	66.4	68.9	-	73.4	76.1	73.8	77.8
Instruction Following							
IFEval	87.5	92.1	92.1	88.6	81.9	84.6	89.3
Code							
HumanEval (0-shot)	80.5	88.4	89.0	89.0	89.0	86.0	93.7
MBPP EvalPlus (base) (0-shot)	86.0	87.6	-	88.6	87.8	83.9	86.8
Math							
MATH (0-sho, CoT)	67.8	77.0	76.6	73.9	82.9	76.9	78.3
Reasoning							
GPQA Diamond (0-shot, CoT)	48.0	50.5	-	49.0	53.5	47.5	65.0
Tool use							
BFCL v2 (0-shot)	77.5	77.3	-	81.1	80.3	74.0	79.3
Long context							
NIH/Multi-needle	97.5	97.5	-	98.1	94.7	-	99.4
Multilingual							
Multilingual MGSM (0-shot)	86.9	91.1	-	91.6	89.6	90.6	92.8

GLUE and SuperGLUE

- GLUE: General Language Understanding Evaluation

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

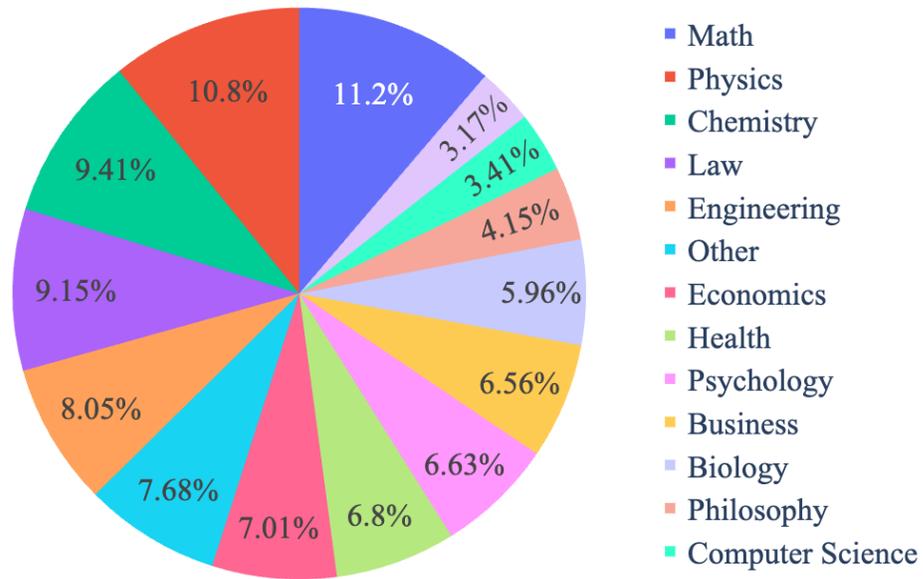
GLUE and SuperGLUE

- GLUE: General Language Understanding Evaluation

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
<i>Single-task single models on dev</i>										
BERT _{LARGE}	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
XLNet _{LARGE}	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	90.2/90.2	94.7	92.2	86.6	96.4	90.9	68.0	92.4	91.3	-
<i>Ensembles on test (from leaderboard as of July 25, 2019)</i>										
ALICE	88.2/87.9	95.7	90.7	83.5	95.2	92.6	68.6	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	96.8	93.0	67.8	91.6	90.4	88.4
RoBERTa	90.8/90.2	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89.0	88.5

Table 5: Results on GLUE. All results are based on a 24-layer architecture. BERT_{LARGE} and XLNet_{LARGE} results are from [Devlin et al. \(2019\)](#) and [Yang et al. \(2019\)](#), respectively. RoBERTa results on the development set are a median over five runs. RoBERTa results on the test set are ensembles of *single-task* models. For RTE, STS and MRPC we finetune starting from the MNLI model instead of the baseline pretrained model. Averages are obtained from the GLUE leaderboard.

MMLU and MMLU-pro



Find all c in \mathbb{Z}_3 such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.
 (A) 0 (B) 1 (C) 2 (D) 3

Figure 14: An Abstract Algebra example.

What is the embryological origin of the hyoid bone?
 (A) The first pharyngeal arch
 (B) The first and second pharyngeal arches
 (C) The second pharyngeal arch
 (D) **The second and third pharyngeal arches**

Figure 15: An Anatomy example.

Why isn't there a planet where the asteroid belt is located?
 (A) A planet once formed here but it was broken apart by a catastrophic collision.
 (B) There was not enough material in this part of the solar nebula to form a planet.
 (C) There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.
 (D) **Resonance with Jupiter prevented material from collecting together to form a planet.**

Figure 16: An Astronomy example.

Three contrasting tactics that CSO's can engage in to meet their aims are _____ which typically involves research and communication, _____, which may involve physically attacking a company's operations or _____, often involving some form of _____.
 (A) Non-violent direct action, Violent direct action, Indirect action, Boycott
 (B) Indirect action, Instrumental action, Non-violent direct action, Information campaign
 (C) **Indirect action, Violent direct action, Non-violent direct-action Boycott.**
 (D) Non-violent direct action, Instrumental action, Indirect action, Information campaign

Figure 17: A Business Ethics example.

BIG-Bench and BIG-Bench Hard

Model Input (Navigate)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.

Options:

- Yes
- No

A: Let's think step by step.

Model Output

We start at the origin (0, 0), facing the positive y-axis.

- (1) Turn left: (0, 0), facing the negative x-axis.
 - (2) Turn right: (0, 0), facing the positive y-axis.
 - (3) Take 5 steps: (0, 5), facing the positive y-axis.
 - (4) Take 4 steps: (0, 9), facing the positive y-axis.
 - (5) Turn around: (0, 9), facing the negative y-axis.
 - (6) Take 9 steps: (0, 0), facing the negative y-axis.
- Since (0, 0) is (0, 0), we are indeed where we started. So the answer is **Yes**. ✓

Model Input (Word Sorting)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: Sort the following words alphabetically: List: burley bela arapaho bacteria bock

A: Let's think step by step.

Model Output

The first letter: "burley": "b" (2). "bela": "b" (2). "arapaho": "a" (1). "bacteria": "b" (2). "bock": "b" (2).
We now have: (1) "arapaho" < (2) ["burley" ? "bela" ? "bacteria" ? "bock"]. Now let's sort this subpart ["burley" ? "bela" ? "bacteria" ? "bock"] by looking at their second letters.
The second letter: "burley": "u" (21). "bela": "e" (5). "bacteria": "a" (1). "bock": "o" (15). We now have: (1) "bacteria" < (5) "bela" < (15) "bock" < (21) "burley". Hence, we have "arapaho" < ["bacteria" < "bela" < "bock" < "burley"]. So the answer is **arapaho bacteria bela bock burley**. ✓

Math - GSM8K

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 $4 \times 2 = 8$ dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies
She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = 6$ 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 = 50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons.
She was able to sell 200 gallons - 24 gallons = 176 gallons.
Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \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 $3 \times 12 = 36$ sodas
6 people attend the party, so half of them is $6/2 = 3$ people
Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas
Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas
With one person drinking 5, that brings the total drunk to $5 + 9 + 8 + 3 = 25$ sodas
As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left

Final Answer: 11

Math - MathVista

ARI

silk scraps	\$9.08/lb
denim scraps	\$8.47/lb
canvas scraps	\$8.17/lb
felt scraps	\$7.29/lb
faux fur scraps	\$11.79/lb
lace scraps	\$6.37/lb

Question: Karen bought 4 pounds of silk scraps and 4 pounds of canvas scraps. How much did she spend? (Unit: \$)
Solution: Find the cost of the silk scraps. Multiply: $\$9.08 \times 4 = \36.32
 Find the cost of the canvas scraps. Multiply: $\$8.17 \times 4 = \32.68
 Now find the total cost by adding: $\$36.32 + \$32.68 = \$69$
 She spent \$69.
Answer: 69

STA

Question: How many sequences have negative Influence Scores?
Answer: 2

ALG

Question: The derivative of y at $x = 6$ is ___ that at $x = 8$.
Choices: (A) larger than (B) equal to (C) smaller than
Answer: (A) larger than

Question: How many zeros does this function have?
Answer: 1

Question: What is the value of y at $x = 1$?
Answer: 0

GEO

Question: \overline{AB} is a diameter, $AC = 8$ inches, and $BC = 15$ inches. Find the radius of the circle.
Diagram logic forms:
 PointLiesOnLine(D, Line(B, A))
 PointLiesOnCircle(B, Circle(D, radius))
 PointLiesOnCircle(A, Circle(D, radius))
 PointLiesOnCircle(C, Circle(D, radius))
Answer: (C) 8.5

NUM

Question: What is the age gap between these two people in image? (unit: years)
Named entities: Winston Churchill, Charles de Gaulle
Wiki caption: Winston Churchill and General de Gaulle at Marrakesh, January 1944
Answer: 16

SCI

Question: The graph of the concentration function $c(t)$ is shown after a 7-mg injection of dye into a heart. Use Simpson's Rule to estimate the cardiac output.
Answer: 5.77

LOG

Brain Teaser for IQ Test

Question: Find the value of the square in the figure.
Solution: Circle + Square = 5, Triangle + Triangle = 8, Triangle = 4. Circle + Triangle = 7, Circle = 3. Therefore Square = 2
Answer: 2

Code - HumanEval

```
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)
```

```
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)
```

Mostly Basic Python Problems Dataset

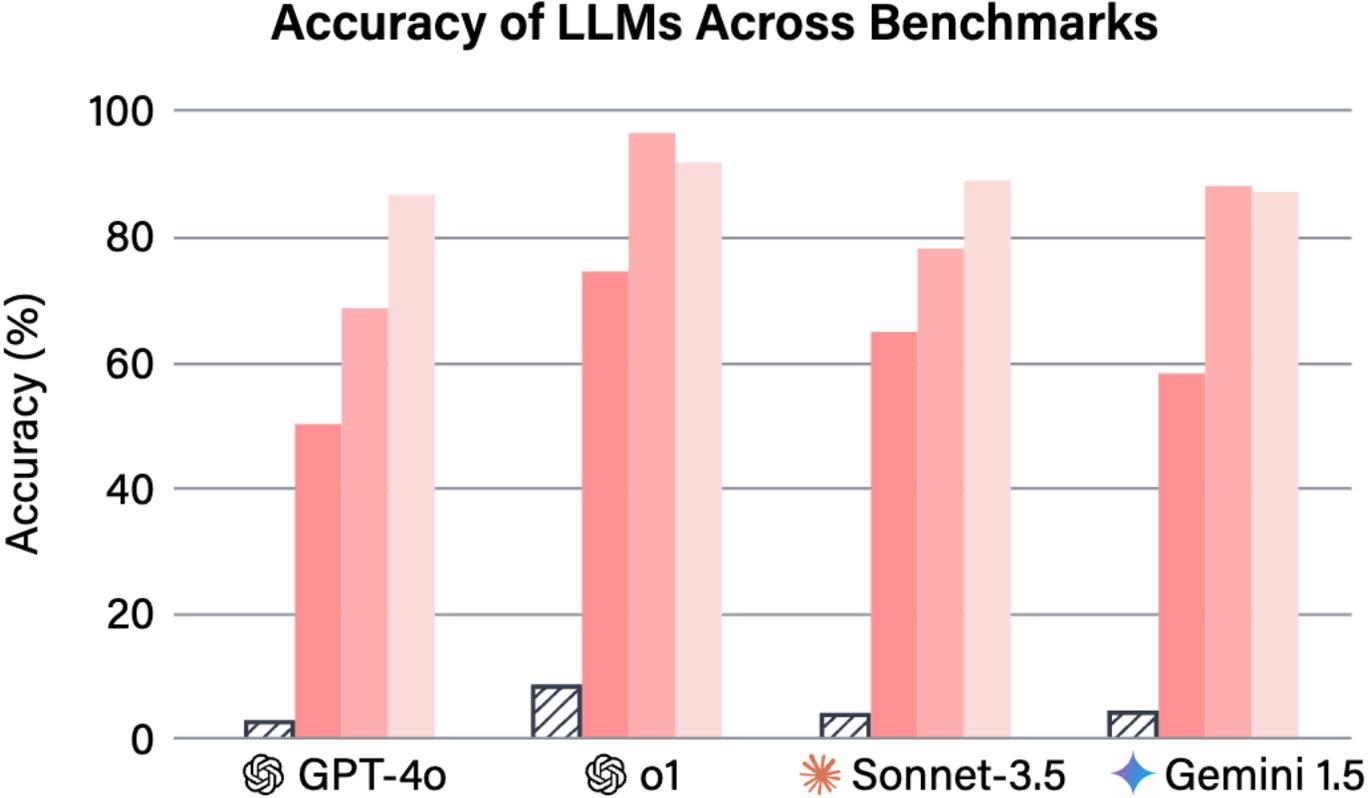
The benchmark consists of around 1,000 crowd-sourced Python programming problems, designed to be solvable by entry level programmers, covering programming fundamentals, standard library functionality, and so on. Each problem consists of a task description, code solution and 3 automated test cases.

As described in the paper, a subset of the data has been hand-verified by us. This data is `sanitized-mbpp.json`.

The dataset is in a `.jsonl` format (json per line).

Released as part of Program Synthesis with Large Language Models, Austin et. al., 2021.

Humanity's Last Exam



Humanity's Last Exam



Arena: LLM Battleground

Text 🕒 1 day ago				Code 🕒 1 day ago			
Rank	Model	Score	Votes	Rank	Model	Score	Votes
1	AI claude-opus-4-6-thinking	1506	4,745	1	AI claude-opus-4-6	1561	2,364
2	AI claude-opus-4-6	1503	5,540	2	AI claude-opus-4-6-thinking	1554	1,789
3	G gemini-3-pro	1486	36,354	3	AI claude-opus-4-5-20251101-...	1501	10,354
4	XI grok-4.1-thinking	1474	35,833	4	🌀 gpt-5.2-high	1471	1,693
5	G gemini-3-flash	1473	27,116	5	AI claude-opus-4-5-20251101	1469	10,491
6	📊 dola-seed-2.0-preview	1472 📄	3,164	6	Z glm-5	1458	2,048
7	AI claude-opus-4-5-20251101-...	1471	28,493	7	📊 minimax-m2.5	1453	2,604
8	AI claude-opus-4-5-20251101	1467	33,328	8	G gemini-3-pro	1444	16,457
9	XI grok-4.1	1463	39,912	9	🌐 kimi-k2.5-thinking	1442	3,342
10	G gemini-3-flash (thinking-...	1461	18,480	10	G gemini-3-flash	1441	12,079
View all				View all			

Limitations of Benchmarks

- Good performance on benchmarks = good model?
- Good human evaluation leaderboard = good model?
- Data contamination