

+ • [COMSE6998-015] Fall
2024

Introduction to Deep
Learning and LLM based
Generative AI Systems
Parijat Dube and Chen Wang

Lecture 11 11/12/24



Agenda

- LLM Evaluation Benchmarks
 - GLUE, SuperGLUE, HELM, MMLU, Big-Bench, LLMPERF
- LLM Evaluation Metrics

Evaluation benchmarks



MMLU (Massive Multitask
Language Understanding)

BIG-bench A small orange icon of a chair with a wavy line underneath it.

GLUE

- General Language Understanding Evaluation (GLUE) benchmark
- A collection of NLU tasks including question answering, sentiment analysis, and textual entailment
- An associated online platform for model evaluation, comparison, and analysis
- GLUE does not place any constraints on model architecture beyond the ability to process single-sentence and sentence-pair inputs and to make corresponding predictions
- Encourage development of models that can generalize across several linguistic tasks
- Centered on 9 English sentence understanding tasks

GLUE



The tasks included in SuperGLUE benchmark:

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

Source: Wang et al. 2018, “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”

GLUE Tasks: Single-Sentence Tasks

1. **CoLA:** The Corpus of Linguistic Acceptability consists of English acceptability judgments drawn from books and journal articles on linguistic theory. Each example is a sequence of words annotated with whether it is a grammatical English sentence.
2. **SST-2:** The Stanford Sentiment Treebank consists of sentences from movie reviews and human annotations of their sentiment. The task is to predict the sentiment of a given sentence.

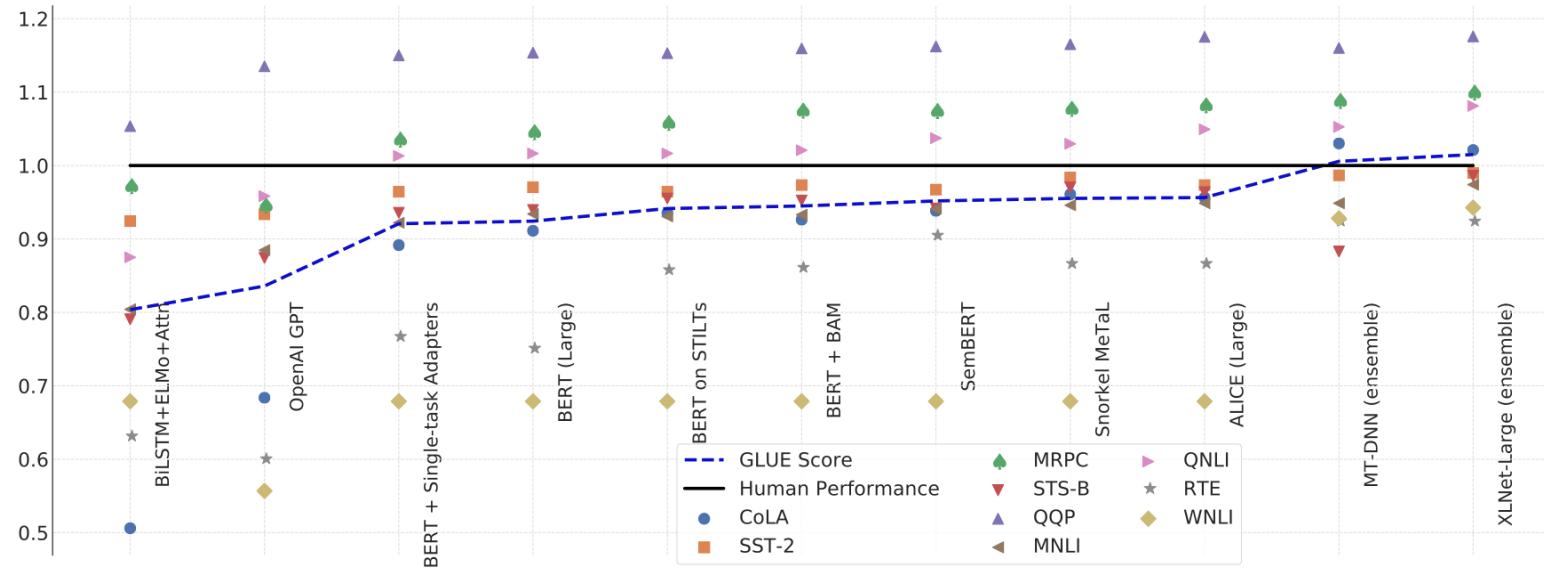
GLUE Tasks - Similarity and Paraphrase Tasks

3. **MRPC:** The Microsoft Research Paraphrase Corpus is a corpus of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent.
4. **QQP:** The Quora Question Pairs2 dataset is a collection of question pairs from the community question-answering website Quora. The task is to determine whether a pair of questions are semantically equivalent.
5. **STS-B:** The Semantic Textual Similarity Benchmark is a collection of sentence pairs drawn from news headlines, video and image captions, and natural language inference data. Each pair is human-annotated with a similarity score from 1 to 5; the task is to predict these scores.

GLUE Tasks: Inference Tasks

6. **MNLI:** The Multi-Genre Natural Language Inference Corpus is a crowd-sourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral).
7. **QNLI:** The Stanford Question Answering Dataset is a question-answering dataset consisting of question-paragraph pairs, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator).
8. **RTE:** The Recognizing Textual Entailment (RTE) datasets
9. **WNLI:** The Winograd Schema Challenge is a reading comprehension task in which a system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices.

2019 SoTA Performance on GLUE



GLUE Leaderboard

<https://gluebenchmark.com/leaderboard>

The screenshot shows the GLUE Leaderboard interface. At the top, there are navigation links for GLUE, SuperGLUE, Paper, Code, Tasks, Leaderboard (highlighted), FAQ, Diagnostics, Submit, and Login. Below the header is a search bar with placeholder text "Search GLUE Leaderboard". The main content area displays a table of the top 10 models, each with a link to its details page.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	Avg
1	Microsoft Alexander v-team	Turing ULR v6	View	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
2	JDExplore d-team	Vega v1	View	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
3	Microsoft Alexander v-team	Turing NLR v5	View	91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57
4	DIRL Team	DeBERTa + CLEVER	View	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
5	ERNIE Team - Baidu	ERNIE	View	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
6	AliceMind & DIRL	StructBERT + CLEVER	View	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
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	View	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
8	HFL iFLYTEK	MacALBERT + DKM	View	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
9	PING-AN Omni-Sinicic	ALBERT + DAAF + NAS	View	90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51
10	T5 Team - Google	T5	View	90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53

SuperGLUE

- Eight language understanding tasks
- Four categories of tasks
 1. QA: question answering
 2. NLI: Natural Language Inference
 3. WSD: Word Sense Disambiguation
 4. coref.: coreference resolution

SuperGLUE Tasks

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

Example of different SuperGLUE tasks

- **BoolQ (Boolean Questions):** QA task where each example consists of a short passage and a yes/no question about the passage

BoolQ *Passage: Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*

Question: *is barq's root beer a pepsi product* **Answer:** No

- **COPA (Choice of Plausible):** system is given a premise sentence and must determine either the cause or effect of the premise from two possible choices

COPA *Premise: My body cast a shadow over the grass. Question: What's the CAUSE for this?*
Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.*
Correct Alternative: 1

Example of different SuperGLUE tasks

- **MultiRC (Multi-Sentence Reading Comprehension):** each example consists of a context paragraph, a question about that paragraph, and a list of possible answers. The system must predict which answers are true and which are false.

MultiRC

Paragraph: Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week

Question: Did Susan's sick friend recover? **Candidate answers:** Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)

- **WiC (Word-in-Context):** is a word sense disambiguation task cast as binary classification of sentence pairs. Given two text snippets and a polysemous word that appears in both sentences, the task is to determine whether the word is used with the same sense in both sentences.

RTF

Text: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.

Hypothesis: Christopher Reeve had an accident. **Entailment:** False

Example of different SuperGLUE tasks

- **ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset)**: is a multiple-choice QA task. Each example consists of a news article and a Cloze-style question about the article in which one entity is masked out. The system must predict the masked out entity from a list of possible entities in the provided passage

ReCoRD

Paragraph: (*CNN*) *Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*

Query For one, they can truthfully say, “Don’t blame me, I didn’t vote for them,” when discussing the <placeholder> presidency **Correct Entities:** US

SuperGLUE



The tasks included in SuperGLUE benchmark:

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

Source: Wang et al. 2019, “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”

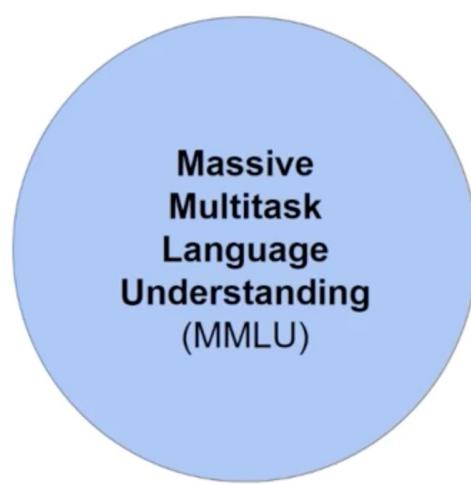
SuperGLUE Leaderboard

<https://super.gluebenchmark.com/leaderboard/>

The screenshot shows the SuperGLUE Leaderboard interface. At the top, there's a navigation bar with links for SuperGLUE, GLUE, Paper, Code, Tasks, Leaderboard (which is highlighted), FAQ, Diagnostics, Submit, and Login. Below the navigation bar, the page title is "Leaderboard Version: 2.0". The main content is a table listing the top 10 models, their names, and their performance scores across 14 different NLP tasks.

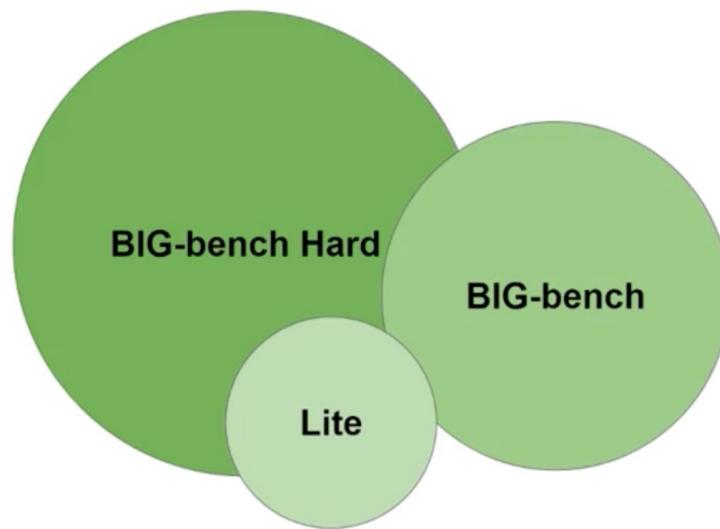
Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g	
+	1	Inspur Cloud	Hairuo		91.4	92.5	96.5/97.6	100.0	90.5/67.9	94.1/93.2	92.8	76.1	100.0	64.6	96.1/94.7
	2	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	3	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	4	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	5	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	6	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	7	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	8	DeBERTa Team - Microsoft	DeBERTa / TuringNLv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	9	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	10	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

Benchmarks for LLMs



2021

Source: Hendrycks, 2021. "Measuring Massive Multitask Language Understanding"



2022

Source: Suzgun et al. 2022. "Challenging BIG-Bench tasks and whether chain-of-thought can solve them"

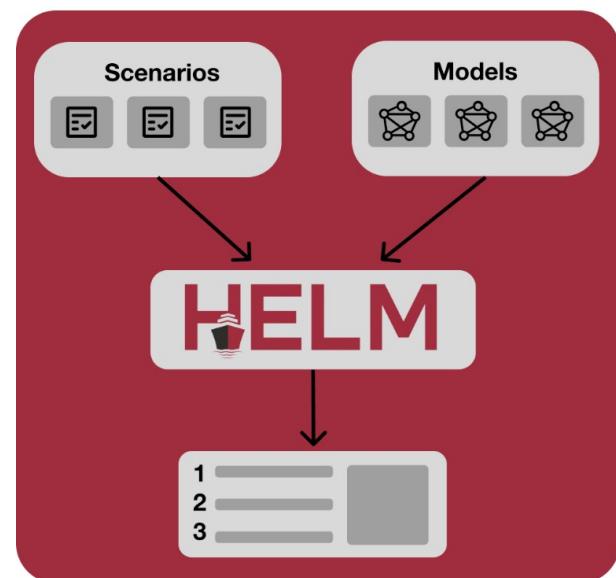
Holistic Evaluation of Language Models (HELM)

- A holistic framework for evaluating language models
- Leaderboards with many scenarios, metrics, models

Core scenarios

- NarrativeQA
- NaturalQuestions (open-book)
- NaturalQuestions (closed-book)
- OpenbookQA
- MMLU (Massive Multitask Language Understanding)
- MATH
- GSM8K (Grade School Math)
- LegalBench
- MedQA
- WMT 2014

<https://crfm.stanford.edu/helm/>



Holistic Evaluation of Language Models (HELM)



Metrics:

1. Accuracy
2. Calibration
3. Robustness
4. Fairness
5. Bias
6. Toxicity
7. Efficiency

Scenarios

	Models										
	J1-Jumbo	J1-Grande	J1-Large	Anthropic-LM	BLOOM	T0pp	Cohere-XL	Cohere-Large	Cohere-Medium	Cohere-Small	GPT-NeoX
NaturalQuestions (open)			✓	✓	✓	✓	✓	✓	✓	✓	✓
NaturalQuestions (closed)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BoolQ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NarrativeQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HellaSwag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OpenBookQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TruthfulQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMLU	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MS MARCO				✓	✓	✓	✓	✓	✓	✓	✓
TREC				✓	✓	✓	✓	✓	✓	✓	✓
XSUM				✓	✓	✓	✓	✓	✓	✓	✓
CNN/DM				✓	✓	✓	✓	✓	✓	✓	✓
IMDB				✓	✓	✓	✓	✓	✓	✓	✓
CivilComments				✓	✓	✓	✓	✓	✓	✓	✓
RAFT				✓	✓	✓	✓	✓	✓	✓	✓

<https://crfm.stanford.edu/helm/>

LLM Evaluation - Challenges

“Mike really loves drinking tea.”



=

“Mike adores sipping tea.”



“Mike does not drink coffee.”

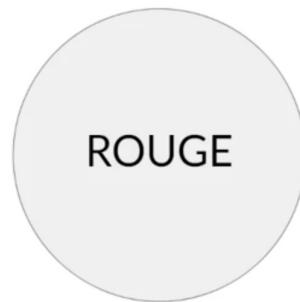


≠

“Mike does drink coffee.”



LLM Evaluation Metrics

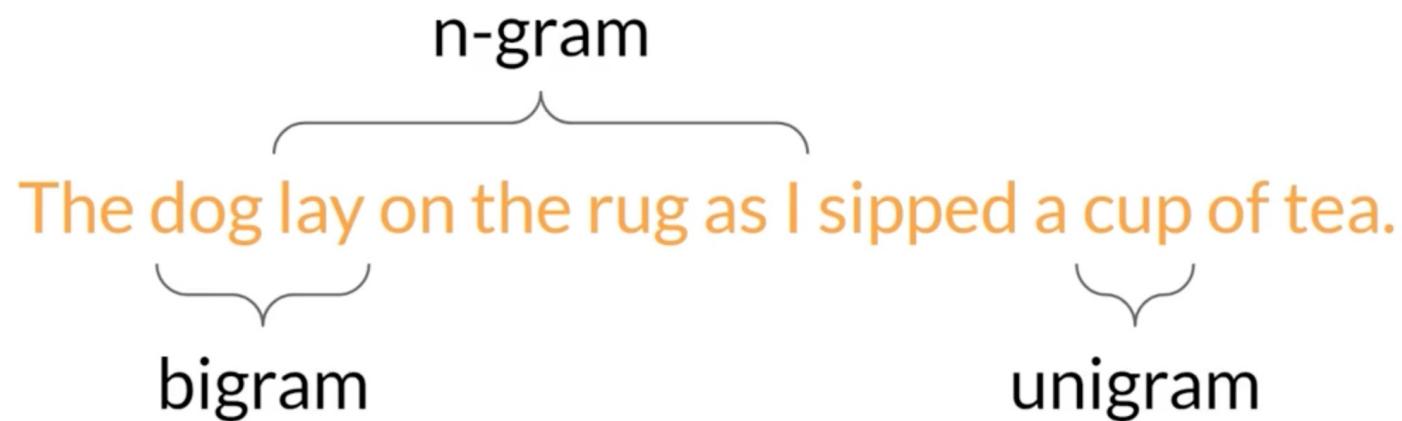


- Used for text summarization
- Compares a summary to one or more reference summaries



- Used for text translation
- Compares to human-generated translations

LLM Evaluation - Metrics - Terminology



LLM Evaluation - Metrics - ROUGE-1

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

Generated output:

It is not cold outside.
—

$$\text{ROUGE-1} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

$$\text{F1: } \text{ROUGE-1} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.8}{1.8} = 0.89$$

LLM Evaluation - Metrics - ROUGE-2

Reference (human):

It is cold outside.

It is is cold

cold outside

Generated output:

It is very cold outside.

It is is very

very cold cold outside

$$\text{ROUGE-2} = \frac{\text{bigram matches}}{\text{bigrams in reference}} = \frac{2}{3} = 0.67$$

$$\text{Precision: } \text{ROUGE-2} = \frac{\text{bigram matches}}{\text{bigrams in output}} = \frac{2}{4} = 0.5$$

$$\text{F1: } \text{ROUGE-2} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.335}{1.17} = 0.57$$

LLM Evaluation - Metrics - ROUGE-L

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

Longest common subsequence (LCS):

It is
cold outside

2

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

$$\text{ROUGE-L} = \frac{\text{LCS}(\text{Gen}, \text{Ref})}{\text{unigrams in reference}} = \frac{2}{4} = 0.5$$

$$\text{Recall: } \text{Precision: } \frac{\text{LCS}(\text{Gen}, \text{Ref})}{\text{unigrams in output}} = \frac{2}{5} = 0.4$$

$$\text{F1: } \text{ROUGE-L} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.2}{0.9} = 0.44$$

- ROUGE is a collection of ROUGE-1, ROUGE-2, ROUGE-L metrics
- ROUGE scores are only comparable for the same task

Problems with ROUGE

Reference (human):

It is cold outside.

Generated output:

cold cold cold cold

$$\text{ROUGE-1 Precision} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{4} = 1.0$$



$$\text{Modified precision} = \frac{\text{clip(unigram matches)}}{\text{unigrams in output}} = \frac{1}{4} = 0.25$$

Generated output:

outside cold it is

$$\text{Modified precision} = \frac{\text{clip(unigram matches)}}{\text{unigrams in output}} = \frac{4}{4} = 1.0$$



LLM Evaluation - Metrics - BLEU

BLEU metric = Avg(precision across range of n-gram sizes)

Reference (human):

I am very happy to say that I am drinking a warm cup of tea.

Generated output:

I am very happy that I am drinking a cup of tea. - BLEU 0.495

I am very happy that I am drinking a warm cup of tea. - BLEU 0.730

I am very happy to say that I am drinking a warm tea. - BLEU 0.798

Quantitative performance metrics for LLM Inference

- Performance metrics for LLM inference
 - Completed requests per minute
 - Time to first token (TTFT): how long before the LLM returns the first token
 - Inter-token latency (ITL): average time between consecutive tokens
 - End-to-end Latency: approximately the same as the average output length of tokens multiplied by the inter-token latency
 - Cost per typical request: API providers can usually trade off one of the other metrics for cost. For example, you can reduce latency by running the same model on more GPUs or using higher-end GPUs.

<https://www.anyscale.com/blog/reproducible-performance-metrics-for-lm-inference>

Metrics for LLM Reasoning

- Large language models show improved downstream task performance when prompted to generate step-by-step reasoning to justify their final answers (Chain of Thought, Tree of Thought, ReAct)
- Reasoning steps greatly improve model interpretability and verification,
- Objectively studying the correctness of reasoning is difficult
- How often the stated reasoning steps actually support the final end task predictions?

Example of 2-shot reasoning

2-shot examples

I will explain how to make logical inferences:

Context: John wakes to the sound of his children giggling outside the door. They bust in and say "Happy Father's Day!" with a plate full of food. 2022 has been a great year.

Question: Can we infer that John was born before the inauguration of Barack Obama?

Explanation: We can infer that John is a father because of "his children". We can infer that the current year is 2022. Barack Obama was inaugurated in 2009, so it has been about 13 years since Obama was inaugurated. John's children are old enough to speak, so if John were born after Obama's inauguration, he would have been younger than 10 when he had children. So, we can infer that John was born before Obama's inauguration. The answer is "yes".

Context: Sam sees a piano in the subway station. Laughing, his friends prod him to play, saying "Remember how amazing your rendition of Chopsticks was at the Christmas party?"

Question: Can we infer that John majored in piano at Juilliard?

Explanation: Juilliard is a top school for music, so someone who majored in piano at Juilliard would be an expert pianist. "Chopsticks" is a very simple song to play on the piano. His friends were therefore likely being sarcastic when they told him to play, which would imply that Sam was not an expert piano player. So it is unlikely that Sam majored in piano at Juilliard. The answer is "no".

Step-by-step reasoning for logical inference task generated from 2-shot GPT-3

Context: Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!

Question: Is it likely that Shelly will be near the Pacific Ocean this weekend?

GPT-3 Step-by-step reasoning:

- No, it is not likely that Shelley will be near the Pacific Ocean this weekend.

- Shelley is from Virginia, which is on the East Coast of the United States.

- The Pacific Ocean is on the West Coast.

- Shelley is visiting a city with a famous market, which could be anywhere in the world.

- Tuesday is her travel day, so she will likely not be traveling this weekend.

Taxonomy of Step-by-Step Reasoning Errors

Error Type	Description	Example
Grammar	Faulty, unconventional, or controversial grammar usage	Reference: He chews 4 pieces of gum a day so over 30 days he will chew $4*30 = \ll 4*30=120 \gg$ 120 pieces of gum. Model Expl: He eats 4 pieces of gum a day and lasts 30 days so he eats $4*30 = \ll 4*30=120 \gg$ 120 pieces of gum.
Factuality	Information about object (i.e. quantity, characteristics) or a personal named entity does not match information provided in the question	Context: Anakin caught 10 fish. Model Expl: Locsin caught 10 fish.
Hallucination	Information is not provided in the problem statement and is irrelevant or wrong	Context: The basketball team went to the steakhouse to eat dinner. The first player ate a 6-ounce steak. The second player ate beef tips, containing 8 beef tips, each an ounce in size. The third player ate a one-pound steak. And the fourth and fifth players ordered vegetarian meals. In total, how many ounces of meat were consumed by the team? Model Expl: The fourth and fifth players ordered vegetarian meals, for a total of 2 ounces of meat.
Redundancy	Explanation contains redundant information, which even though might be factual, is not required to answer the question	Context: Bob is cold. If something is quiet then it is blue. All blue things are round. If something is round and smart then it is cold. Question: Bob is cold. Model Expl: If something is round and not smart then it is blue. All blue things are round. Answer: true.
Repetition	Step paraphrases information already mentioned in previous reasoning steps	Context: Parker chews 4 pieces of gum a day. A pack of gum has 15 pieces of chewing gum per pack. How many packs of gum will he need to last him 30 days? Model Expl: He chews 4 pieces of gum a day and there are 30 days in a month so he chews $4*30 = 120$ pieces of gum a month. Parker chews 120 pieces of gum a month. He needs 120 pieces of gum a month and a pack of gum has 15 pieces of gum so he will need $120/15 = 8$ packs of gum.

[ROSCOE: A SUITE OF METRICS FOR SCORING STEP-BY-STEP REASONING](#)

Taxonomy of Step-by-Step Reasoning Errors

Missing step	The content of the generated reasoning is incomplete and lacks required information to produce the correct answer. If this missing steps are added, the model would produce the correct answer	Context: A farm has 10 2-legged animals and 15 4-legged animals. If all animals are healthy, how many pairs of animal legs in total are on the farm? Answer: 40 Model Expl. : 10 2-legged animals have $10*2=20$ legs. 15 4-legged animals have $15*4=60$ legs. In total on the farm, there are $20+60=80$ legs. Answer: 80 missing step: Total number of pairs is $80/2=40$
Coherency	Steps contradict each other or do not follow a cohesive story	Context: If the reactor is intact, then it is safe. If the reactor is safe, then no graphite is outside it. Model Expl. : The reactor is intact. All graphite is inside of it. There are some graphite outside of the the reactor.
Commonsense	Model lacks relations that should be known from general knowledge about the world (i.e. how to compute velocity, how many inches in one foot, all ducks are birds, etc.)	Question: Jordan has 2 children who wear diapers. Each child requires 5 diaper changes per day. Jordan's wife changes half of the diapers. How many diapers does Jordan change per day? Model Expl. : Jordan has $2*5=10$ diapers to change per day. Jordan's wife changes $10/2=5.0$ diapers per day. For both children, Jordan changes $10+5=15$ diapers per day.
Arithmetic	Error in math calculations	Reference: If a truck was carrying 20 tons of fertiliser packed in bags, the total number of bags in a truck is $20*20 = 400$ bags Model Expl: If a truck was carrying 20 tons of fertiliser packed in bags, the total number of bags in a truck is $20*20 = 40$ bags

[ROSCOE: A SUITE OF METRICS FOR SCORING STEP-BY-STEP REASONING](#)

Reasoning Scorer: ROSCOE

- Suite of metrics for scoring Step-by-Step reasoning
- **Source context** of T sentences indicating a problem statement followed by a question and is prompted to generate step-by-step reasoning

$$\text{source context } \mathbf{s} = \{s_1, \dots, s_T\}$$

- Hypothesis of N-steps, including a final answer as the last step

$$\text{hypothesis } \mathbf{h} = \{h_1, \dots, h_N\}$$

- Gold step-by-step reasoning of K steps (may or may not be available)

$$\text{references } \mathbf{r} = \{r_1, \dots, r_K\}$$

[ROSCOE: A SUITE OF METRICS FOR SCORING STEP-BY-STEP REASONING](#)

Reasoning Scorer: ROSCOE

- Fine-grained metrics under 4 categories
 1. Semantic-Alignment Metrics
 2. Semantic-Similarity Metrics
 3. Local Inference Metrics
 4. Language Coherence Metrics

[ROSCOE: A SUITE OF METRICS FOR SCORING STEP-BY-STEP REASONING](#)

Reasoning Alignment Vector

- Reasoning alignment vector from the N -step hypothesis \mathbf{h} to the source \mathbf{s} of length T

$$r\text{-align}(\mathbf{h} \rightarrow \mathbf{s}) = \{\alpha_1, \alpha_2, \dots, \alpha_N\}$$

$$\alpha_i = r\text{-align}(h_i \rightarrow \mathbf{s}) = [1 + \max_{j=1}^T (\cos(h_i, s_j))] / 2 \in [0, 1]$$

normalized cosine similarity between hypothesis step and most similar sentence in a context

- Measures the grounding of the step-wise reasoning with respect to the source text

Semantic Alignment Metrics

Table 3: Semantic alignment metrics (ROSCOE-SA).

Score	Description
Faithfulness-Step $(\mathbf{h} \rightarrow \mathbf{s})$	This step-level score is based on the alignment from the hypothesis steps to the source sentences, and is calculated as the mean reasoning alignment score over the steps of reasoning (see illustration in Appendix D, Figure 3): $(1/N) \sum_{i=1}^N r\text{-align}(\mathbf{h}_i \rightarrow \mathbf{s})$. Faithfulness measures if the model misinterpreted the problem statement, or the reasoning chain is too vague, irrelevant, or misuses information.
Informativeness-Step (Info-Step) $(\mathbf{h} \leftrightarrow \mathbf{s})$	Measures how well information present in the source is used in the reasoning steps: $[(1/T) \sum_{t=1}^T r\text{-align}(\mathbf{s}_t \rightarrow \mathbf{h}) + (1/N) \sum_{i=1}^N r\text{-align}(\mathbf{h}_i \rightarrow \mathbf{s})]/2$. Info-step gives a higher score to reasoning steps that are well-grounded with respect to the source, and identifies the degree of information from source that is covered by the generated hypothesis. A lower Info-Step score corresponds to the reasoning steps that are not related to the source sentences or have missed information provided in the context.
Hallucination $(\mathbf{h} \rightarrow (\mathbf{s}, \mathbf{r}))$	To find irrelevant reasoning steps, we use alignment score to identify steps that are both not related to the context and not in the reference chain (to avoid punishing for possibly relevant commonsense knowledge): $1 - \max_{i=1..N} ([\mathbf{1} - r\text{-align}(\mathbf{h} \rightarrow \mathbf{s})] \cdot [\mathbf{1} - r\text{-align}(\mathbf{h} \rightarrow \mathbf{r})])$. Here, $\mathbf{1}$ is an all-ones vector, and (\cdot) is the element-wise product.
Redundancy $(\mathbf{h} \rightarrow \mathbf{r})$	To find chains that contain information that is not required to solve the problem (i.e., redundant steps), we identify those hypothesis steps that are least aligned with the the reference steps: $\min_{i=1..N} r\text{-align}(\mathbf{h}_i \rightarrow \mathbf{r})$. This score punishes chains with steps that are not required for the correct solution.

Semantic Similarity Metrics

- Quantify the degree of semantic equivalence between pieces of text
- Considers text as a whole, rather than relying on text units comparisons

Table 4: Semantic similarity metrics (ROSCOE-SS).

Score	Description
Informativeness-Chain (Info-Chain) ($h \rightarrow s$)	Similar to Info-Step, this metric quantifies the degree of agreement between the hypothesis chain and the source and is calculated as $[1 + \cos(h, s)]/2$. We embed reasoning chain and source context as a whole, as opposed to using step-wise embeddings in *-Step types of metrics introduced in Table 3.
Repetition-Step ($h_i \leftrightarrow h_j$)	Measures repetition-related errors on the step level by checking if it paraphrases information already mentioned in the previous steps: $(1 - \max_{i=2..N} \max_{j=1..i-1} [\cos(h_i, h_j)])/2$. Unlike Repetition-Token, which is orderless and compares individual tokens in pairs of steps, Repetition-Step considers step embeddings similarity and is more robust to changing contexts.
Semantic Coverage- Chain ($r \leftrightarrow h$)	Reflects the overall degree of similarity between the reference and hypothesis chains, comparing reference and hypothesis embeddings as a whole: $[1 + \cos(r, h)]/2$.

Logical Inference Metrics

- Measure logical errors between pieces of text
- Use a Natural Language Inference (NLI) model trained to classify hypothesis-context pairs into entailment, neutral, and contradiction classes to infer the contradiction probability p_{contr}

Table 5: Logical inference metrics (ROSCOE-LI).

Score	Description
Self-Consistency $(h_i \leftrightarrow h_j)$	Measures logical entailment errors <i>within</i> the reasoning steps: $1 - \max_{i=2..N} \max_{j < i} p_{\text{contr}}(h_i, h_j)$. This metric will punish chains where there is a pair of steps that are likely to contradict each other.
Source-Consistency $(h \leftrightarrow s)$	Measures logical entailment errors between any generated reasoning h and the source context s : $1 - \max_{i=1..N} \max_{j=1..T} p_{\text{contr}}(h_i, s_j)$. Specifically, for each reasoning step we measure the probability that it contradicts any sentence in the context. We take the maximum probability of contradiction over all steps, following the logic that a contradiction anywhere in the reasoning chain signals a failure of the overall argument.

Language Coherence Metrics

- Use perplexity PPL as scored by the GPT2-Large model and English grammatical acceptability p_gram scores

Table 6: Language coherence metrics (RŌSCOE-LC).

Score	Description
Perplexity-Chain (h)	Average perplexity of all tokens in the generated reasoning steps: $1/\text{PPL}(h)$. The context used to score each token is the previous tokens in the current and all previous steps. Steps are joined with a space character. To keep the range and orientation consistent with the other scores we invert the perplexity.
Perplexity-Step (h_i)	Average perplexity of all tokens in the generated reasoning steps, where the context used to score each token is only the previous tokens within the current step: $1/[(1/N) \sum_{i=0}^N \text{PPL}(h_i)]$. To keep the range and orientation consistent with the other scores we invert the perplexity.
Grammar (h_i)	Probability of grammatical acceptability of each step, averaged over all steps: $(1/N) \sum_{i=0}^N p_{\text{gram}}(h_i)$.

LLMPerf

- Tool for evaluating performance of LLM APIs
- Two tests for evaluating LLMs:
 1. Load test – check for performance
 2. Correctness test – check for correctness

<https://github.com/ray-project/LLMPerf>

LLMPerf Load Test

- Spawns a number of concurrent requests to the LLM API and measures the inter-token latency and generation throughput per request and across concurrent requests
- Prompt send with each request

```
Randomly stream lines from the following text. Don't generate eos tokens:  
LINE 1,  
LINE 2,  
LINE 3,  
...  
...
```

- The lines are randomly sampled from a collection of lines from Shakespeare sonnets.
- Tokens are counted using the LlamaTokenizer regardless of which LLM API is being tested. This is to ensure that the prompts are consistent across different LLM APIs.

LLMPerf Correctness Test

- Spawns a number of concurrent requests to the LLM API with the following format:

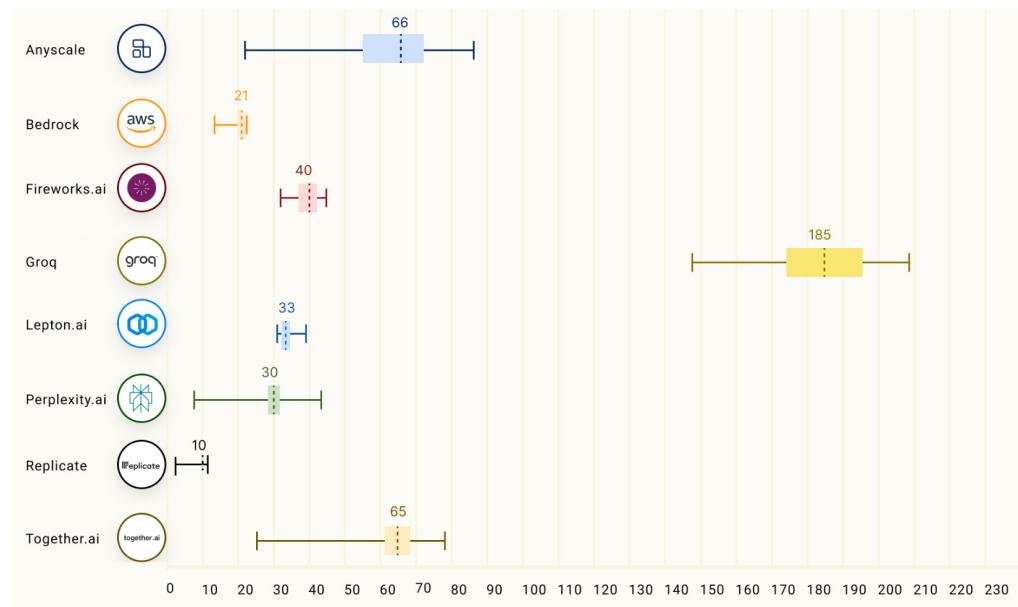
```
Convert the following sequence of words into a number: {random_number_in_word_format}. Output just your final answer.
```

- `random_number_in_word_format` could be for example "one hundred and twenty three". The test then checks that the response contains that number in digit format which in this case would be 123.
- The test does this for a number of randomly generated numbers and reports the number of responses that contain a mismatch.

LLMPerf Leaderboard

Output tokens throughput (tokens/s)

70B Models



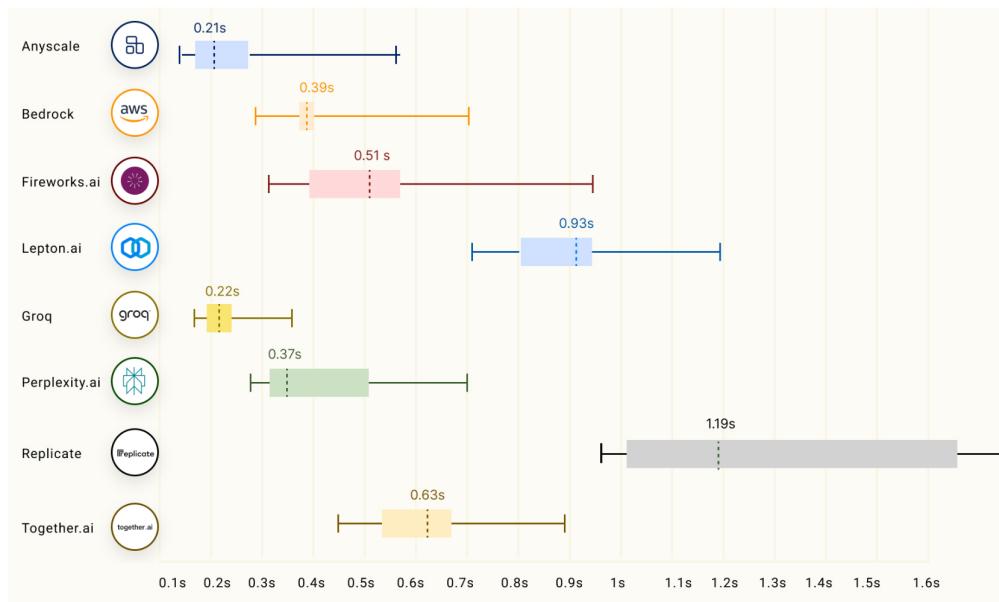
Framework	Model
anyscale	meta-llama/Llama-2-70b-chat-hf
bedrock	meta.llama2-70b-chat-v1
fireworks	accounts/fireworks/models/llama-v2-70b-chat
groq	llama2-70b-4096
lepton	llama2-70b
perplexity	llama-2-70b-chat
replicate	meta/llama-2-70b-chat
together	together_ai/togethercomputer/llama-2-70b-chat

<https://github.com/ray-project/llmperf-leaderboard>

LLMPerf Leaderboard

Time to first token (seconds)

70B Models



Framework	Model
anyscale	meta-llama/Llama-2-70b-chat-hf
bedrock	meta.llama2-70b-chat-v1
fireworks	accounts/fireworks/models/llama-v2-70b-chat
groq	llama2-70b-4096
lepton	llama2-70b
perplexity	llama-2-70b-chat
replicate	meta/llama-2-70b-chat
together	together_ai/togethercomputer/llama-2-70b-chat

<https://github.com/ray-project/llmperf-leaderboard>