

LLM Evaluation

CS 5624: Natural Language Processing

Spring 2025

<https://tuvllms.github.io/nlp-spring-2025>

Tu Vu



Logistics

- Quiz 2 will be released this Friday
- Homework 2 will be released sometime next week
- We are sending out feedback on final project proposals
- Please email us at cs5624instructors@gmail.com

LLM News: Gemini 2.0 Flash native image generation

-

LLM News: Gemini 2.0 Flash native image generation

-

LLM News: LLM Agents

Chatbot Arena



Human evaluation

- To evaluate a 250-word generation:
 - Crowdworkers take ~10 sec
 - English teachers take ~70 sec
- High variability across runs, low agreement



BLUE & ROUGE

- N-gram overlap between the machine output and reference

BLEURT: Learning Robust Metrics for Text Generation

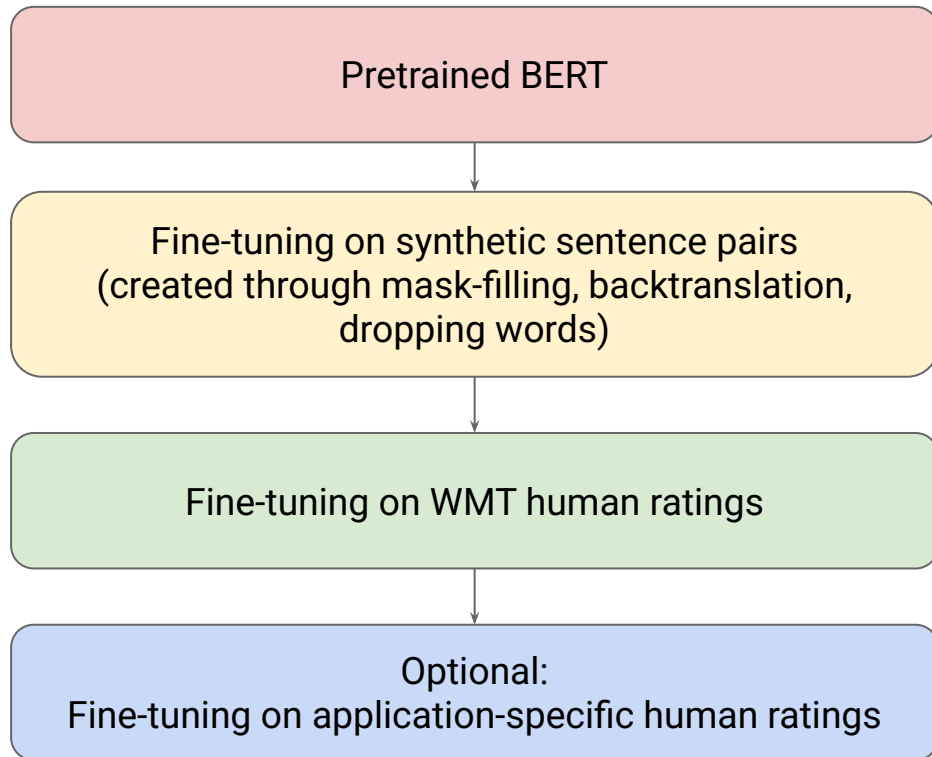
Thibault Sellam Dipanjan Das Ankur P. Parikh

Google Research

New York, NY

{tsellam, dipanjand, aparikh}@google.com

BLEURT (BLUE + BERT)



COMET: A Neural Framework for MT Evaluation

Ricardo Rei

Craig Stewart

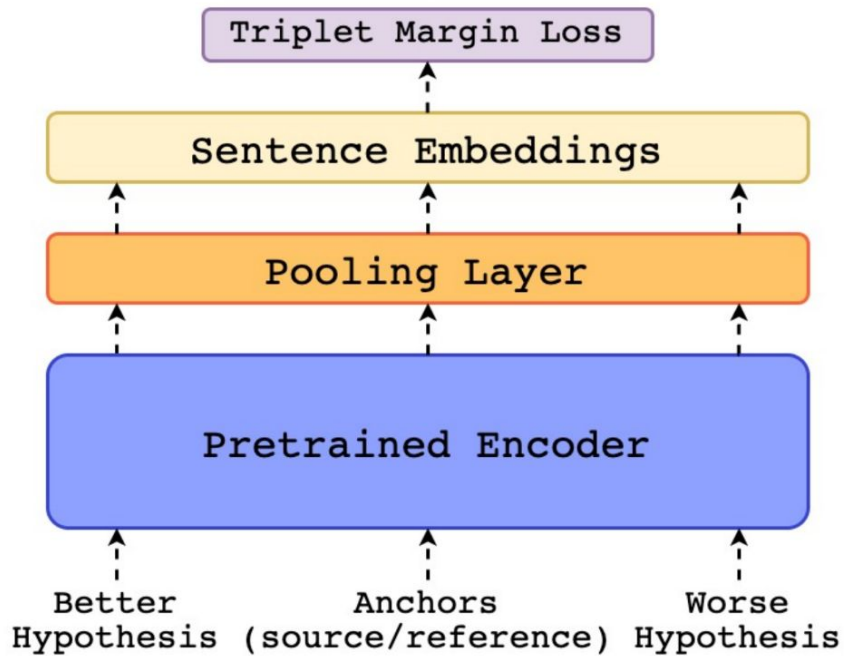
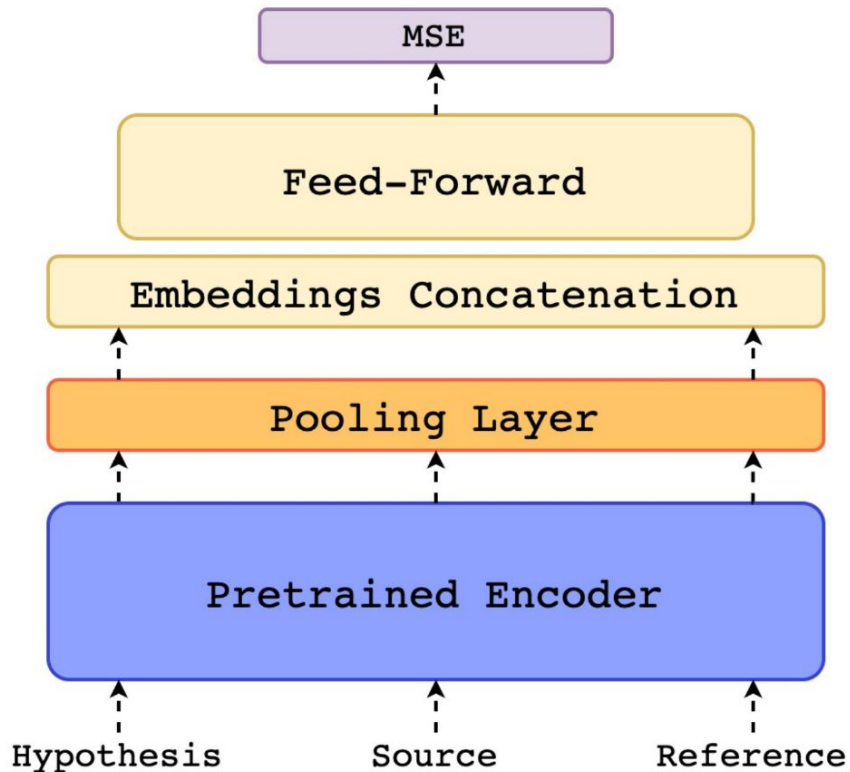
Ana C Farinha

Alon Lavie

Unbabel AI

{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

COMET



FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

**Sewon Min^{†1} Kalpesh Krishna^{†2} Xinxu Lyu¹ Mike Lewis⁴ Wen-tau Yih⁴
Pang Wei Koh¹ Mohit Iyyer² Luke Zettlemoyer^{1,4} Hannaneh Hajishirzi^{1,3}**

¹University of Washington ²University of Massachusetts Amherst

³Allen Institute for AI ⁴Meta AI

{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu

{kalpesh,miyyer}@cs.umass.edu {mikelewis,scotttyih}@meta.com

FactScore

breaks a generation into a series of atomic facts and computes the percentage of atomic facts supported by a reliable knowledge source.

Chat
GPT

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...



Tell me
a bio of
Bridget
Moynahan.



- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...



66.7%



LLM-as-a-Judge / LLM auto-rater

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

Lianmin Zheng^{1*} Wei-Lin Chiang^{1*} Ying Sheng^{4*} Siyuan Zhuang¹

Zhanghao Wu¹ Yonghao Zhuang³ Zi Lin² Zhuohan Li¹ Dacheng Li¹³

Eric P. Xing³⁵ Hao Zhang¹² Joseph E. Gonzalez¹ Ion Stoica¹

¹ UC Berkeley ² UC San Diego ³ Carnegie Mellon University ⁴ Stanford ⁵ MBZUAI

An LLM is given context
and tasked with
evaluating response
quality or selecting the
better response among
candidates

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.

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

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

Types of LLM-as-a-Judge

- Pairwise comparison
 - determine which response is better or declare a tie
- Single answer grading (pointwise evaluation)
 - directly assign a score to a single response
- Reference-guided grading
 - a reference solution is provided

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]

Pointwise evaluation

[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]

Reference-based evaluation

[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]

Advantages of LLM-as-a-Judge

Advantages of LLM-as-a-Judge

- **Scalability**
 - reduces the need for human involvement, enabling scalable benchmarks and fast iterations
- **Explainability**
 - provides not only scores but also explanations, making its outputs interpretable

Limitations of LLM-as-a-Judge

Position (Order) bias

- The LLM judge favors a particular response position.

Question: What are some business etiquette norms when doing business in Japan?

Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. **Bowing:** Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. **Business cards:** Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. **Punctuality:** Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. **Dress code:** Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. **Language:** Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. **Gift giving:** Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. **Dining etiquette:** When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. **Greetings:** A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. **Business cards:** In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. **Meetings:** Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. **Punctuality:** Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. **Gift giving:** Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. **Dining:** Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. **Communication:** Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

Table 2: Position bias of different LLM judges. Consistency is the percentage of cases where a judge gives consistent results when swapping the order of two assistants. “Biased toward first” is the percentage of cases when a judge favors the first answer. “Error” indicates wrong output formats. The two largest numbers in each column are in bold.

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	75.0%	0.0%	1.2%
	rename	56.2%	11.2%	28.7%	3.8%
GPT-3.5	default	46.2%	50.0%	1.2%	2.5%
	rename	51.2%	38.8%	6.2%	3.8%
GPT-4	default	65.0%	30.0%	5.0%	0.0%
	rename	66.2%	28.7%	5.0%	0.0%

Verbosity bias

- The LLM judge favors longer, verbose responses, even if they are not as clear, high-quality, or accurate as shorter alternatives.

Question: What is the central dogma of molecular biology? What processes are involved? Who named this?

Assistant A:

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription rephrased: The act of copying genetic information from DNA to RNA.
2. Translation rephrased: The act of using genetic information in RNA to create proteins.
3. Transcription: The process by which the genetic information in DNA is copied into RNA.
4. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

Assistant B:

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription: The process by which the genetic information in DNA is copied into RNA.
2. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

GPT-3.5 Judgment:

Assistant A provides a more detailed and accurate response to the user's question. While both assistants correctly state the central dogma of molecular biology and who coined the term, Assistant A provides a more comprehensive explanation of the processes involved in the central dogma, including the rephrased definitions of transcription and translation. Therefore, I choose A as the better assistant.

Claude-v1 Judgment:

Both assistants answer the question by defining the central dogma and detailing the two processes involved, transcription and translation. However, assistant A provides a more in-depth and detailed explanation by briefly rephrasing and elaborating on these processes. Assistant A also mentions who coined the term "central dogma", which provides useful context and shows a more comprehensive understanding of the topic. Overall, assistant A gives a fuller and more insightful response, so I have chosen it as the better assistant for this question.

GPT-4 Judgment:

Both Assistant A and Assistant B provide accurate and relevant information about the central dogma of molecular biology, the processes involved, and who named it. However, Assistant A's answer contains a minor redundancy by rephrasing transcription and translation twice. Assistant B's answer is more concise and avoids this redundancy. Therefore, Assistant B's answer is better.

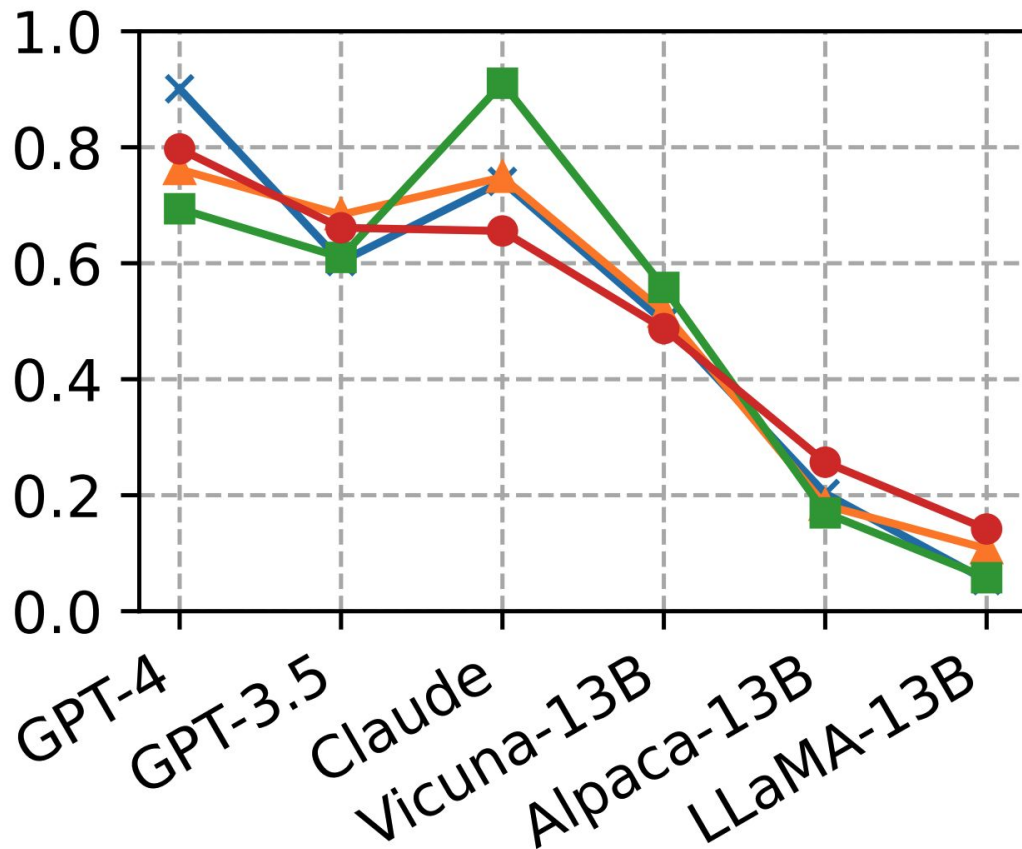
Table 3: Failure rate under “repetitive list” attack for different LLM judges on 23 answers.

Judge	Claude-v1	GPT-3.5	GPT-4
Failure rate	91.3%	91.3%	8.7%

Self-enhancement (egocentric) bias

- The LLM judge favors the responses generated by itself

—x— GPT-4 Judge —▲— GPT-3.5 Judge —■— Claude Judge —●— Human —◆— Human (first turn)



Limited capability in grading math and reasoning questions

Question: Benjamin went to a bookstore and purchased a variety of books. He bought 5 copies of a sci-fi novel, each priced at \$20, 3 copies of a history book priced at \$30 each, and 2 copies of a philosophy book for \$45 each. What was the total cost of his purchases?

Assistant A:

The total cost of Benjamin's purchases was \$115.

Assistant B:

The total cost of his purchases is \$295.

GPT-4 Judgment:

Assistant B provided the correct answer to the user's question. Assistant A's response was incorrect. To calculate the total cost of Benjamin's purchases, we need to multiply the number of books by their respective prices and then add the results:

5 sci-fi novels * \$20 = \$100

3 history books * \$30 = \$90

2 philosophy books * \$45 = \$90

Total cost = \$100 + \$90 + \$90 = \$295

Assistant B's response is accurate and directly answers the user's question. Therefore, my final verdict is: [[B]]

GPT-4's Answer (if asked separately):

The cost of the sci-fi novels is $5 * \$20 = \100 .

The cost of the history books is $3 * \$30 = \90 .

The cost of the philosophy books is $2 * \$45 = \90 .

The total cost of his purchases is $\$100 + \$90 + \$90 = \280 . Answer: $\boxed{280}$.

Figure 13: With a default prompt, GPT-4 shows limited capability in grading math questions. Despite being able to answer the question itself, its judgment was influenced by the given answers, leading to arithmetic mistakes highlighted in yellow.

Question: Each problem consists of three statements. Based on the first two statements, the third statement may be true, false, or uncertain.

1. Oranges cost more than apples.
2. Oranges cost less than bananas.
3. Bananas cost more than apples and bananas cost more than orange.

If the first two statements are true, then the third statement is

Assistant A:

false.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), then it is not possible for bananas to cost less than oranges and also cost more than both apples and oranges (statement 3). Therefore, statement 3 is false.

Assistant B:

False.

If the first two statements are true, then the third statement is false. The first statement says that oranges cost more than apples, and the second statement says that oranges cost less than bananas. This means that oranges cost less than bananas, which contradicts the third statement that bananas cost more than apples and more than oranges. Therefore, the third statement must be false.

GPT-4 Judgment:

Both Assistant A and Assistant B provided accurate and relevant responses to the user's question. They both concluded that the third statement is false based on the first two statements. However, Assistant B's response is slightly clearer and more concise in explaining the contradiction between the statements. Therefore, I choose Assistant B as the better response.

[[B]]

GPT-4's Answer (if asked separately):

True.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), it logically follows that bananas cost more than apples and bananas cost more than oranges, which is the third statement. So, if the first two statements are true, the third one is also true.

Figure 14: An example of GPT-4's limited capability in grading reasoning question. Despite GPT-4 knows how to solve the question (if asked separately), it made a wrong judgement saying both assistants' wrong answers are correct.

Addressing limitations

Addressing limitations

- Swapping positions
- Few-shot judge
- Chain-of-thought and reference-guided judge
- Fine-tuning a judge model

Table 12: Improvements of the few-shot judge on consistency for position bias.

Model	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	zero-shot	23.8%	75.0%	0.0%	1.2%
	few-shot	63.7%	21.2%	11.2%	3.8%
GPT-3.5	zero-shot	46.2%	50.0%	1.2%	2.5%
	few-shot	55.0%	16.2%	28.7%	0.0%
GPT-4	zero-shot	65.0%	30.0%	5.0%	0.0%
	few-shot	77.5%	10.0%	12.5%	0.0%

Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation

Tu Vu^{*,1}, Kalpesh Krishna^{*,2},

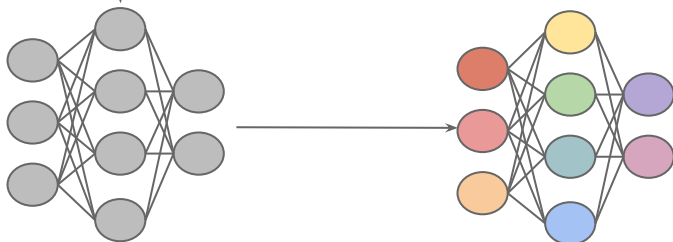
Salaheddin Alzubi³, Chris Tar¹, Manaal Faruqui² and Yun-Hsuan Sung¹

*Co-lead (equal contribution), ¹Google DeepMind, ²Google, ³UMass Amherst

From FLAN to FLAMe

NLP tasks

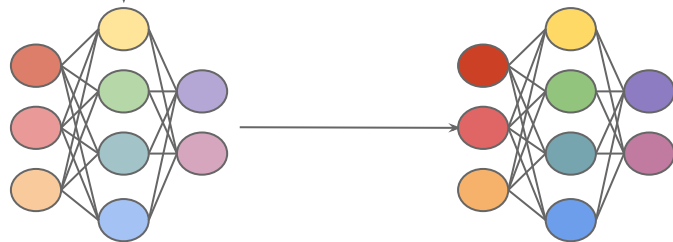
FLAN collection



Designing Data and Methods for Effective Instruction Tuning
(Longpre & Hou & Vu et al., **ICML 2022**)

human evaluations

FLAMe collection



Curating and standardizing existing human evaluations



Source: Gold Miner Adventure

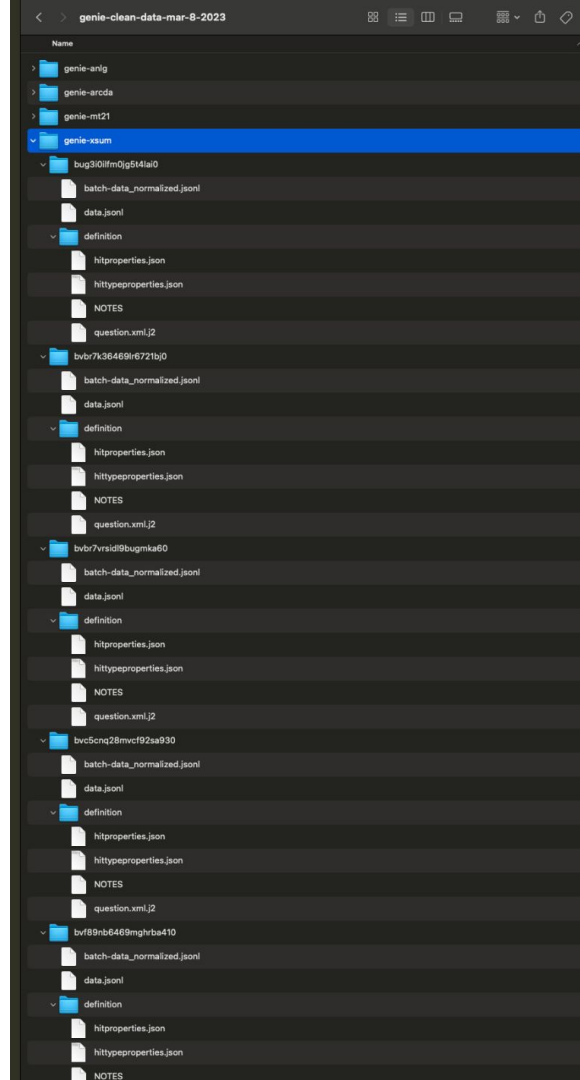
The FLAMe collection

- 102 quality assessment tasks comprising 5M+ total human judgments
- spans a wide variety of task types, from assessing summarization quality to evaluating how well AI assistants follow user instructions
- *all datasets are publicly available and under permissive licenses*

Collecting existing human evaluations is challenging

- Lack of standardization
- Diverse evaluation criteria
- Inadequate documentation
- Data privacy and proprietary concerns
- ...

Data preprocessing took 3-4 hours per dataset!



Unified task format

"""Input format."""

INSTRUCTIONS:

"""Task definition and evaluation instructions."""

title: Is all of the information in the summary fully attributable to the source article?

description: In this task, you will be shown a summary and a source news article on which the summary is based. Your task is to evaluate whether the summary is attributable to the source article. Answer 'Yes' if all the information in the summary is fully supported by the source article, or 'No' if any information in the summary is not supported by the source article. Provide an explanation for your answer.

output_fields: answer, explanation

CONTEXT:

"""Input fields for context, each starting with a label indicating its type or purpose and is separated by a newline, for example:

'article': <article>

'summary': <summary>

"""

article: *Tower Hamlets Council said it would sell Draped Seated Woman after "unprecedented" budget cuts. The work has not yet been valued but a Moore sold for £17m earlier this year. The council said the rising threat of metal theft and vandalism made it too expensive to insure if it was on show. The sculpture was bought by the former London County Council for £6,000 in 1960. The bronze sculpture, nicknamed Old Flo, was installed on the Stifford council estate in 1962 but was vandalised and moved to the Yorkshire Sculpture Park in 1997. A council spokesperson said: "With unprecedented cuts to council budgets, the council finds itself in a difficult situation and being forced to make hard decisions."*

summary: *A Moore sculpture of a woman sitting on a concrete plinth is to be sold.*

Unified task format (cont.)

```
"""Target format."""
```

EVALUATION:

```
"""Target fields, each starting with a label indicating its type or purpose and is separated by a newline, for example:
```

```
'choice': <choice>
```

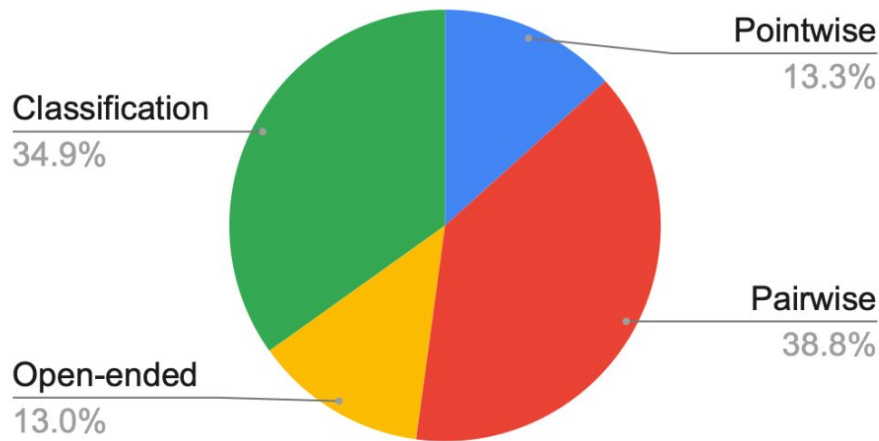
```
'explanation': <explanation>
```

```
"""
```

```
answer: No
```

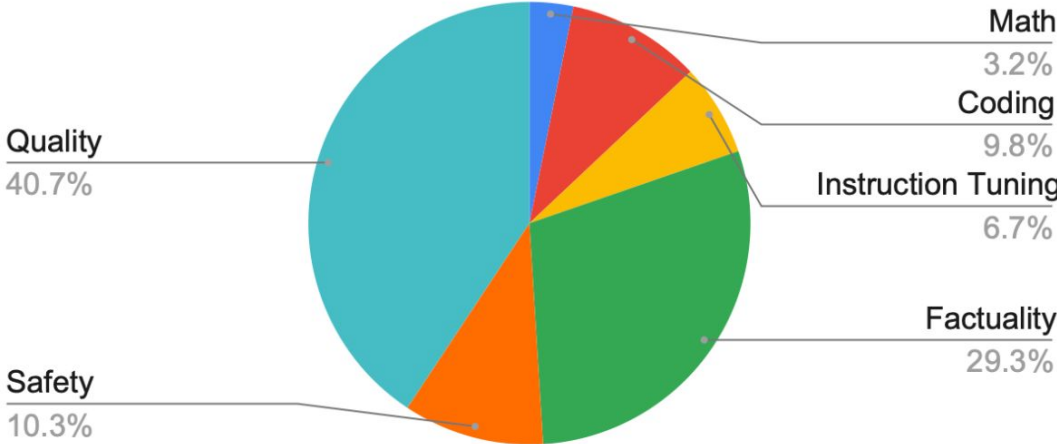
```
explanation: The detail that the woman is "sitting on a concrete plinth" is not in the article.
```

The FLAMe collection covers diverse task types



FLAMe data collection breakdown by task type, showing the percentage of datapoints (out of 5.3M) for each task type. Over half of FLAMe is dedicated to standard pairwise (“Which response is better?”) and pointwise (“Rate the response on a Likert scale.”) evaluation. The remainder includes classification (e.g., “Is the summary fully attributable to the source article? (Yes/No)”) and open-ended evaluation (e.g., “Explain why response A is better than response B.”).

The FLAMe collection encompasses key LLM capabilities



FLAMe data collection breakdown by LLM capability, showing the percentage of datapoints (out of 5.3M) for each LLM capability. We focus on standard LLM evaluation pillars: general response quality, factuality, safety, coding, and math.

FLAMe models

- **FLAMe**

- PaLM-2-24B trained on the FLAMe collection (examples-proportional mixing) for 30K steps

- **FLAMe-RM**

- FLAMe fine-tuned on a balanced mixture of four preference datasets (covering chat, reasoning, and safety) for 50 steps

- **FLAMe-Opt-RM**

- PaLM-2-24B trained on the FLAMe collection with RewardBench-optimized mixture weights (determined by tail-patch ablations) for 5K steps

FLAMe variants outperform all LLM-as-a-Judge autoraters on 8 out of 12 autorater evaluation benchmarks

Model	Reward Bench	LLM AggreFact	Summ Feedback	Alpaca Farm	Rank Gen	Co Poet	Contr Search	HHH	Dipper	Lit Trans	LFQA Eval	Help Steer
Llama-3-70B-Instruct	76.1	76.1	50.8	53.9	65.6	53.6	53.1	91.9	42.8	60.5	71.1	39.7
Mixtral-8×7B	77.8	73.8	43.8	55.1	63.3	52.9	56.6	90.0	42.2	61.7	71.5	34.0
GPT-3.5-turbo-0125	64.5	70.0	15.6	55.5	58.2	49.0	57.5	85.5	45.0	54.3	69.9	32.0
Claude-3-Opus	80.7	79.2	31.6	49.6	55.1	49.0	45.1	94.6	50.6	71.1	71.1	41.3
GPT-4-0125	85.9	80.6	46.5	49.6	62.5	56.9	55.8	94.6	45.0	67.6	77.0	37.9
GPT-4o	84.7	80.2	30.9	50.4	66.0	55.6	57.5	92.3	45.6	72.7	75.0	40.1
<i>our models</i>												
PaLM-2-24B	62.9	54.8	13.3	52.3	58.2	54.2	46.0	85.5	48.3	62.5	70.3	20.0
FLAMe-24B	86.0	81.1	48.0	58.2	62.1	53.6	69.9	91.4	48.3	67.2	74.2	48.4
FLAMe-RM-24B	87.8	80.8	53.1	57.8	65.2	57.5	57.5	91.0	47.8	67.6	72.7	46.6
FLAMe-Opt-RM-24B	87.0	80.2	52.3	53.1	69.5	52.9	48.7	89.1	48.3	69.5	69.5	35.9

Performance of FLAMe compared to popular LLM-as-a-Judge autoraters across various autorater benchmarks.

FLAMe-RM-24B was the top-performing generative model only on permissively licensed data

RewardBench: Evaluating Reward Models

Evaluating the capabilities, safety, and pitfalls of reward models

[Code](#) | [Eval. Dataset](#) | [Prior Test Sets](#) | [Results](#) | [Paper](#) | Total models: 108 | * Unverified models



RewardBench Leaderboard

Model Search (delimiter with ,)

Seq. Classifiers DPO Custom Classifiers Generative Prior Sets

▲	Model	▲	Model Type	▲	Score	▲	Chat	▲	Chat Hard	▲	Safety	▲	Reasoning	▲
1	google/gemini-1.5-pro-0514 *		Generative		88.1		92.3		80.6		87.5		92.0	
2	google/flame-1.0-24b-july-2024		Generative		88.1		92.2		75.7		90.7		93.8	
3	openai/gpt-4-0125-preview		Generative		85.9		95.3		74.3		87.2		86.9	
4	openai/gpt-4-turbo-2024-04-09		Generative		85.1		95.3		75.4		87.1		82.7	
5	openai/gpt-4o-2024-05-13		Generative		84.7		96.6		70.4		86.7		84.9	
6	Anthropic/claude-3.5-sonnet-20240620		Generative		83.8		96.4		74.0		80.1		84.7	
7	google/gemini-1.5-flash-001		Generative		82.1		92.2		63.5		87.7		85.1	
8	Anthropic/claude-3-opus-20240229		Generative		80.7		94.7		60.3		89.1		78.7	
9	meta-llama/Meta-Llama-3-70B-Instruct		Generative		76.0		97.6		58.9		69.2		78.5	
10	Anthropic/claude-3-sonnet-20240229		Generative		75.7		93.4		56.6		83.7		69.1	
11	Poll/gpt-3.5-turbo-0125_claude-3-sonnet-20240229...		Generative		75.6		95.3		54.1		79.5		73.5	
12	prometheus-eval/prometheus-8x7b-v2.0		Generative		75.3		93.0		47.1		83.5		77.4	

July 15, 2024

FLAMe is significantly less biased than other popular LLM-as-a-Judge models

Autorater	Avg. (↓)	Order (↓)	Compassion (↓)	Length (↓)	Egocentric (↓)	Bandwagon (↓)	Attention (↓)
Random	0.30	0.50	0.50	0.00	0.25	0.25	0.25
<i>baselines reported in Koo et al. (2023)</i>							
Falcon-40B	0.31	0.77	0.27	0.09	0.05	0.28	0.40
Cohere-54B	0.41	0.50	0.65	0.10	0.27	0.82	0.14
Llama-2-70B	0.19	0.61	0.26	0.12	0.06	0.04	0.03
InstructGPT	0.45	0.38	0.48	0.16	0.28	0.85	0.54
ChatGPT	0.45	0.41	0.66	0.13	0.58	0.86	0.06
GPT-4	0.31	0.23	0.79	0.06	0.78	0.00	0.00
<i>our models</i>							
FLAMe-24B	0.13	0.08	0.09	0.03	0.38	0.18	0.00
FLAMe-RM-24B	0.13	0.11	0.08	0.02	0.40	0.17	0.00
FLAMe-Opt-RM-24B	0.15	0.15	0.14	0.00	0.41	0.17	0.00

Autorater bias analysis on the CoBBLer bias benchmark from [Koo et al. \(2023\)](#). **Lower values indicate better or less biased autoraters** across all columns. Overall, FLAMe variants exhibit significantly less bias compared to popular LLM-as-a-Judge autoraters like GPT-4. Compared to Table 2 in [Koo et al. \(2023\)](#), we combine first/last numbers for Order/Compassion, report $|\text{bias} - 0.5|$ for Length, and only report the order setup in Egocentric.

Learning to Plan & Reason for Evaluation with Thinking-LLM-as-a-Judge

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, Tianlu Wang

FAIR at Meta

LLM-as-a-Judge models generate chain-of-thought (CoT) sequences intended to capture the step-by-step reasoning process that underlies the final evaluation of a response. However, due to the lack of human-annotated CoTs for evaluation, the required components and structure of effective reasoning traces remain understudied. Consequently, previous approaches often (1) constrain reasoning traces to hand-designed components, such as a list of criteria, reference answers, or verification questions and (2) structure them such that planning is intertwined with the reasoning for evaluation. In this work, we propose EvalPlanner, a preference optimization algorithm for Thinking-LLM-as-a-Judge that first generates an unconstrained evaluation plan, followed by its execution, and then the final judgment. In a self-training loop, EvalPlanner iteratively optimizes over synthetically constructed evaluation plans and executions, leading to better final verdicts. Our method achieves a new state-of-the-art performance for generative reward models on RewardBench (with a score of 93.9), despite being trained on fewer amount of, and synthetically generated, preference pairs. Additional experiments on other benchmarks like RM-Bench, JudgeBench, and FollowBenchEval further highlight the utility of both planning and reasoning for building robust LLM-as-a-Judge reasoning models.

Thank you!