LLM Evaluation

Emine Yilmaz

Professor of Computer Science, UCL

Amazon Scholar, Amazon

Emine.Yilmaz@ucl.ac.uk

Why is Evaluation so Important?

"What you can't measure you can't improve"

Lord Kelvin

Evaluation metrics serve as objectives LLM based models are optimized for

Why is Evaluation so Important?

- Poor evaluation can lead to:
  - Misinformation
  - Biased or harmful outputs
  - Safety risks
- Rigorous evaluation:
  - Ensures quality and fairness
  - Builds user trust
  - Informs responsible AI development
- Particularly important in critical domains (health, law, education)

Which Evaluation Metric?

- Quantitative Metrics
  - Use **numerical values** and **statistical methods** to evaluate performance
  - Most of them are **reference-based**.
  - Examples: Accuracy, BLEU, ROUGE, Perplexity, etc.
- Qualitative Metrics
  - Often consider aspects that are hard to quantify, such as **relevance**, **fluency**, or **coherence**.
  - Better suited for evaluating **open-ended text generation** or **conversational** AI use cases.
  - Could be **reference-free**.

Example Quantitative Metrics

- *Accuracy*: Fraction of correct responses
- *BLEU (Bilingual Evaluation Understudy)* [1]
  - Given a candidate response C and a reference text R,
    - What fraction of n-grams in C are present in R?
    - **Precision** for each matching n-gram between C and R
    - Compute geometric mean of all n-gram precisions
      - Best correlation with humans observed when max n = 4
    - Could also apply brevity penalty (BP)
      - $BP = e^{(1-|R|/|C|)}$ if $|C|<|R|$, 1 otherwise

[1] BLEU: A Method for Automatic Evaluation of Machine Translation. ACL 2022

Example Quantitative Metrics

- *ROUGE (*Recall-Oriented Understudy for Gisting Evaluation) [1]
  - Given a candidate response C and a reference text R,
    - Calculate **recall** by comparing the overlap of n-grams between C and R
    - What fraction of n-grams in R are present in C?
- *Perplexity* [2]
  - Quantifies **uncertainty** via **inverse probability** of the reference text R
  - Let R consist of words $w_1 \ldots w_N$. Perplexity can be computed as:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{t=1}^{N}\log p(w_t \mid w_1, \ldots, w_{t-1})\right)$$

  - **Lower perplexity** indicates **better** predictive performance, meaning **higher probabilities to correct words** in a sentence

[1] ROUGE: A Package for Automatic Evaluation of Summaries. ACL 2004
[2] Perplexity—A Measure of the Difficulty of Speech Recognition Tasks. The Journal of the Acoustical Society of America. 1977

Qualitative Metrics

- **Core Dimensions**
  - Fluency / Language Quality
    - Measures how natural, grammatical, and human-like the text is.
    - Focus: syntax, readability, style.
  - Coherence / Consistency
    - Logical flow of ideas across sentences/paragraphs.
    - Internal consistency (no contradictions within the same output).
  - Relevance
    - How well the output matches the input query/task.
    - Especially important in search, retrieval, and summarization.
  - Factuality / Faithfulness
    - Accuracy of statements with respect to the real world or provided context.
    - Guards against hallucinations.
  - Completeness / Coverage
    - Whether the answer covers all key aspects of the query.
    - Avoids missing important details.

Qualitative Metrics

- **Social & Ethical Dimensions**
  - Fairness / Bias
    - Avoiding stereotypes, discrimination, and skewed perspectives.
  - Toxicity / Harmfulness
    - Preventing offensive, hateful, or unsafe content.
    - Different from bias: toxicity is about *harmful expression*, not imbalance.
  - Safety / Alignment
    - Does the model refuse harmful requests (e.g., instructions for violence)?
    - Compliance with ethical use.

Qualitative Metrics

- **Task-Specific Dimensions**
  - Helpfulness / Utility
    - How useful is the response for the user's goal?
  - Conciseness/Verbosity
    - Does the model provide overly short or overly long answers?
    - Important in summarization, QA, and dialogue.
  - Creativity / Originality
    - Relevant for open-ended tasks like storytelling or ideation.
  - Reasoning / Faithful Chain of Thought
    - Does the model follow logical reasoning steps?
    - Accuracy of intermediate reasoning (esp. in math/code).

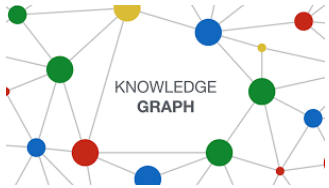
Other Dimensions in LLM Evaluation

- Robustness
  - Stability under paraphrased or adversarial inputs.
- Calibration / Uncertainty
  - Does the model know when it doesn’t know?
  - Confidence estimation
- Efficiency / Latency
  - Response time, computational cost.
  - Important for real-world applications
- Controllability
  - Ability to follow style, persona, or format instructions.

Example: Evaluating Factuality

- Human annotation based 
  - Human annotators check whether each statement is true, false, unverifiable, or partially correct

- Fact-checking with External Knowledge
  - Each claim is verified against reliable sources.
  - Use retrieval-based tools to cross-check claims:
    - Search engines (RAG-style evaluation).
    - Wikipedia, knowledge graphs (e.g., Wikidata), structured databases.



- LLM-as-a-judge
  - Another LLM is prompted to act as a fact-checker.
  - Example prompt: “Given this claim and this source text, is the claim factually correct? Answer YES, NO, or UNKNOWN.

Example: Evaluating Fairness/Bias

- Human annotation based
  - Annotators rate output for different dimensions related to bias
    - Presence of stereotypes
    - Imbalance in political or cultural viewpoints
- Counterfactual evaluation
  - Swap sensitive attributes and check consistency
- LLM-as-a-Judge
  - Another LLM (or multiple) is asked to rate the fairness of a response.
    - Example prompt: *“Does this output rely on stereotypes or show unfair treatment of a group?”*

Need for Annotations in Evaluation

- Some sort of annotation is necessary for most evaluations
- Human annotation is expensive and time consuming
  - Expert annotators require extensive training and guidelines
  - Could be a major bottleneck in system development
- Crowdsourcing could be an alternative
  - Much cheaper, does not require as much training
  - Quality could be an issue
- Different approaches developed to reduce judgment effort
  - Intelligent sampling approaches, automated evaluation based on user behavior, …
  - More recently: LLM-as-a-judge

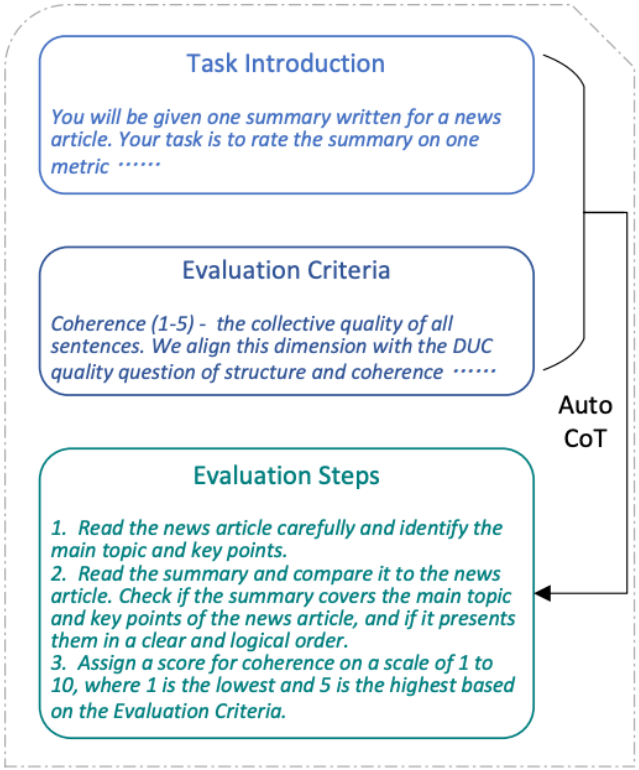
Humans vs. LLM-as-a-Judge

- Advantages of using LLM-as-a-judge
  - Scalable: Can evaluate thousands of outputs quickly
  - Flexible: Can adapt judging criteria via prompts or fine-tuning
  - Avoids fatigue and human subjectivity (to some extent)
  - Could be cheaper

- LLMs have shown very good performance when used as judges
  - Some research claimed LLMs to exhibit similar performance to expert annotators for tasks such as relevance assessment [1,2]
  - LLMs shown to perform better than crowd judges on certain tasks [3]

[1] Perspectives on Large Language Models for Relevance Judgment. ICTIR 2023
[2] A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look. ICTIR 2025
[3] Large Language Models can Accurately Predict Searcher Preferences. SIGIR 2024

Example LLM-as-a-Judge: G-EVAL



Better correlation with human judgments than BLEU/ROUGE in several dimensions such as Naturalness, Coherence, Relevance [1].

[1] G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP 2023.

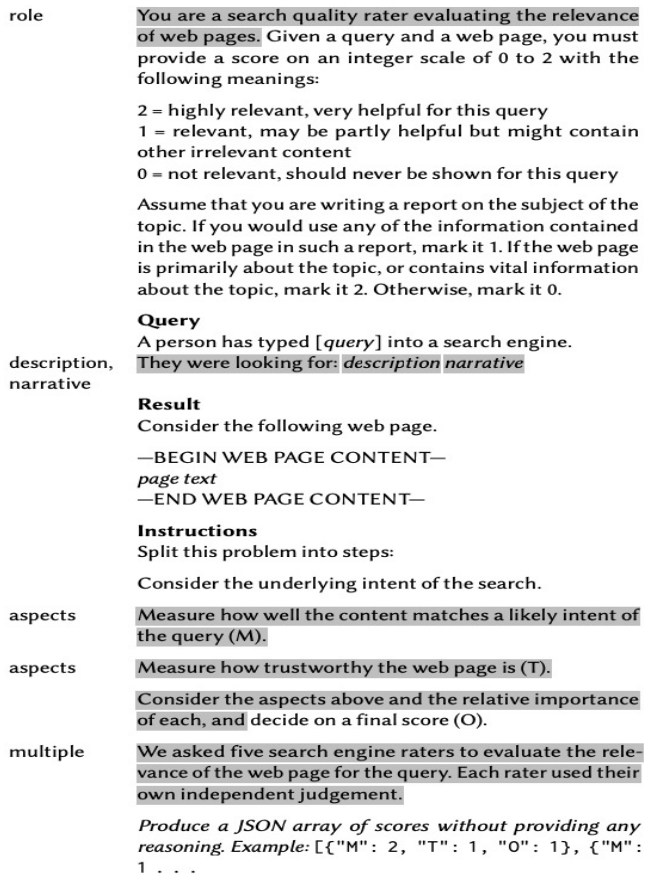
What to Include in an LLM-as-a-Judge Prompt?

Possible Dimensions of a Prompt [1]:

- Role
- Description
- Narrative
- Aspects
- Multiple Judgements

- What dimensions to include?

- Analyse the effect of different dimensions
  - Relevance annotation task



role — You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

2 = highly relevant, very helpful for this query
1 = relevant, may be partly helpful but might contain other irrelevant content
0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

Query
A person has typed [*query*] into a search engine.

description, narrative — They were looking for: *description narrative*

Result
Consider the following web page.

—BEGIN WEB PAGE CONTENT—
page text
—END WEB PAGE CONTENT—

Instructions
Split this problem into steps:

Consider the underlying intent of the search.

aspects — Measure how well the content matches a likely intent of the query (M).

aspects — Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

multiple — We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: `[{"M": 2, "T": 1, "O": 1}, {"M": 1 . . .`

[1] Large Language Models can Accurately Predict Searcher Preferences. SIGIR 2024.

What to Include in an LLM-as-a-Judge Prompt?

- Evaluating every component of your prompt could help.
- Paraphrasing parts of your prompt could have significant effect to output.
- Aspects give a substantial improvement.
- Role and multiple judge components could have a negative effect.
- Description and narrative add a little boost

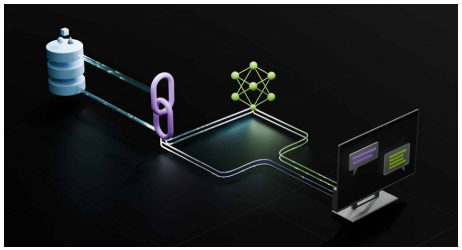
[8] Large Language Models can Accurately Predict Searcher Preferences. SIGIR 2024.

LLM-as-a-Judge with RAG

Key Benefit: Moves from generic evaluation to **contextually-aware judgment** using domain-specific knowledge and examples.

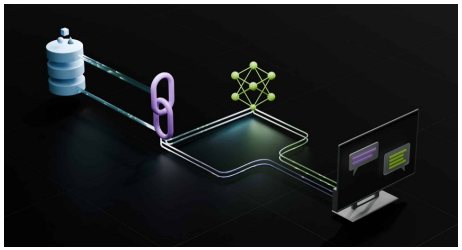
- Retrieve relevant context/examples or evaluation criteria during evaluation
- Provides evaluator with domain knowledge and precedents

Particularly useful for tasks requiring domain-specific knowledge



LLM-as-a-Judge with RAG

Input : Dialog turn to evaluate
↓
Retrieve: Relevant examples + evaluation criteria/data
↓
Prompt: "Given these examples of good/bad responses, evaluate the dialog turn for [helpfulness/appropriateness/etc.]"
↓
Output: Score + reasoning based on retrieved context



Lately: Agents as Judges

- **Tool use**: Access databases, APIs, calculators during evaluation
- **Multi-step reasoning**: Break complex evaluations into subtasks
- **Memory**: Maintain context across long evaluation sessions

Plain LLM vs. Agents as Judges

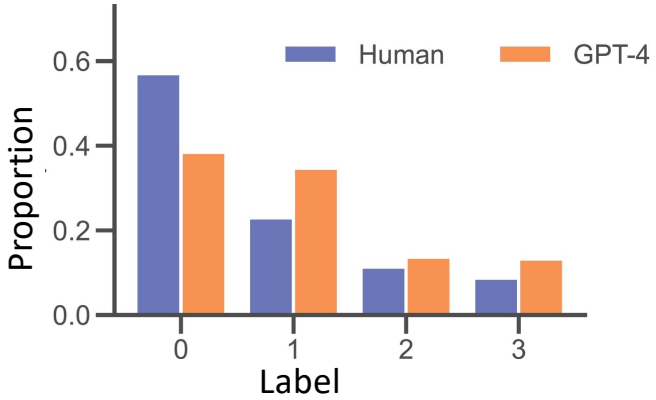
- Easy multi-criteria evaluation:
  - **Plain LLM:** Single-shot evaluation of all criteria
  - **Agent:** Step 1: Check grammar → Step 2: Verify facts → Step 3: Assess relevance
- Dynamic retrieval augmentation:
  - **RAG:** Fixed retrieval based on query
  - **Agent:** Iterative retrieval → Refine search → Gather additional context → Evaluate
- Consistency across scoring:
  - **Plain LLM:** May drift in scoring standards
  - **Agent:** Maintains evaluation criteria in memory → Consistent application

Could LLMs replace Human Judges?

- LLMs have shown remarkable performance as judges
- Are LLMs capable of truly replacing human judges?
- Possible limitations of using LLM-as-a-judge
  - Quality of judgments
  - Significant need for improvement for multi-modal tasks
  - Bias in judgments
  - Bias in evaluation results

Limitations of Using LLM-as-a-Judge: Quality of Judgments

- LLMs tend to assign higher scores to ”bad” responses [1]
  - Human assessors tend to apply stricter criteria than LLMs
  - Could cause evaluations to have less discriminative power

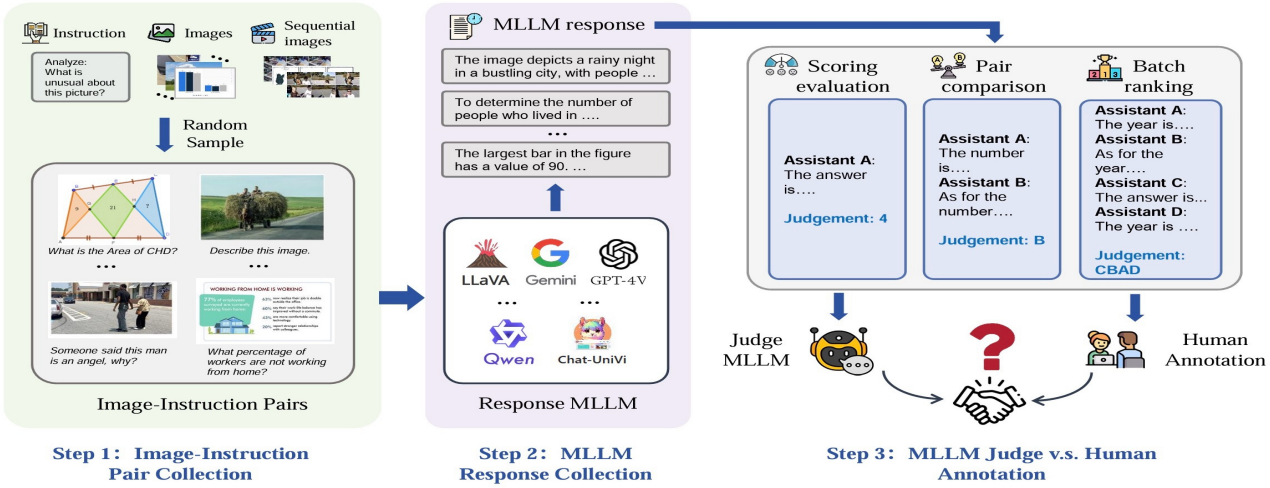


- Most LLMs exhibit susceptibility to judging non-relevant documents as relevant if query words are inserted at random positions [2]

[1] Towards Understanding Bias in Synthetic Data for Evaluation. CIKM 2025
[2] LLMs can be Fooled into Labelling a Document as Relevant. SIGIR 2024

Limitations of Using LLM-as-a-Judge: Multi-Modal Tasks

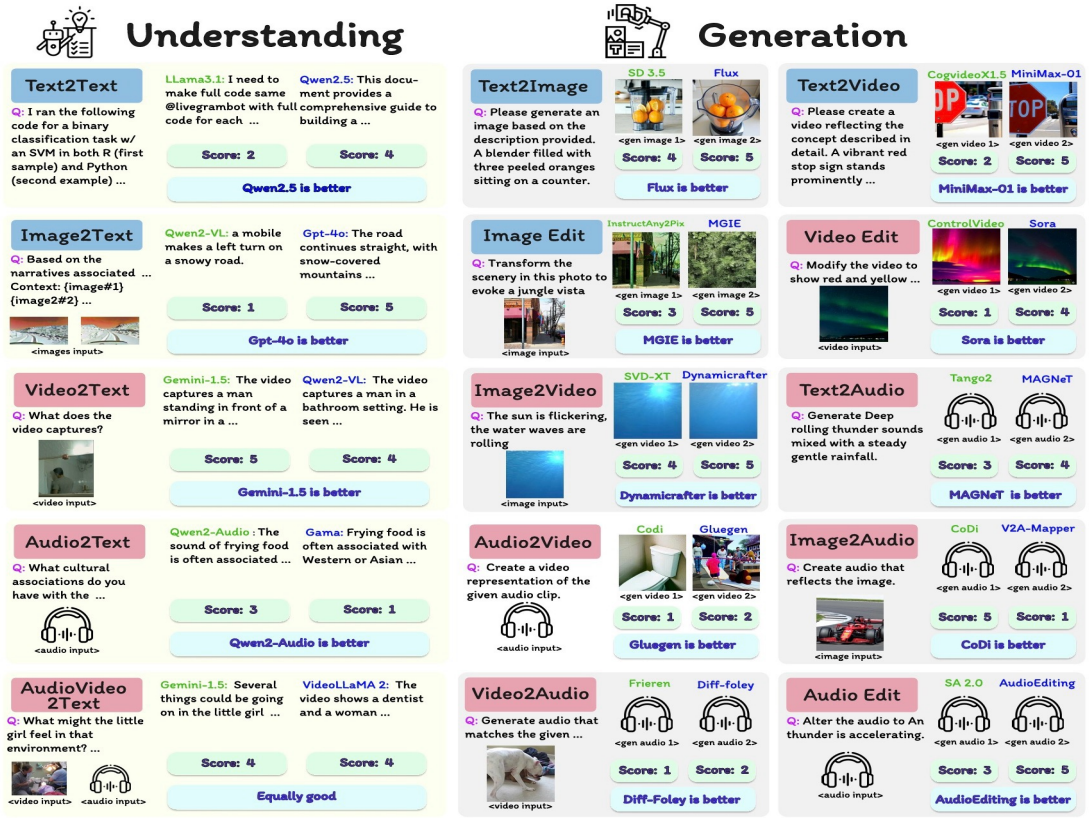
- Multimodal Large Language Models (MLLMs) demonstrate remarkable performance as judges on pair comparison
- Significant divergence from human preferences in scoring evaluation and batch ranking [1]
  - 79% agreement in pair comparison vs. 70% agreement for scoring vs. 42% in ranking on image to text tasks



[1] MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. ICML 2024.

Limitations of Using LLM-as-a-Judge: Any to Any Modality

Multi-modal understanding vs. multi-modal generation tasks with any-to-any modality [1]



[1] Judge Anything: MLLM as a Judge Across Any Modality. KDD 2025.

Limitations of Using LLM-as-a-Judge: Any to Any Modality

- Models tend to perform much better in multi-modal understanding (MMU) compared to multi-modal generation (MMG)
  - Average correlation of 66.55% on pair comparison, 42.79% on score evaluation for MMU
  - Average correlation of 53.37% on pair comparison, 30.05% on score evaluation for MMG
- Models tend to perform much better on pairwise comparison compared to scoring an individual item
  - Similar findings to image to text
- Correlations are quite low, could have a significant effect on model development
- There is need for significant improvement and further research when using MLLM-as-a-judge

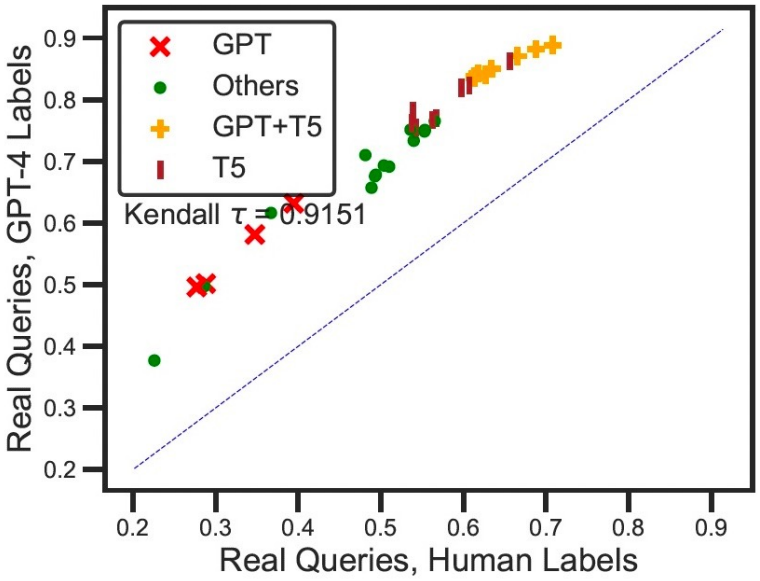
Judge Anything: MLLM as a Judge Across Any Modality. KDD 2025.

Limitations of Using LLM-as-a-Judge: Bias in Judgments

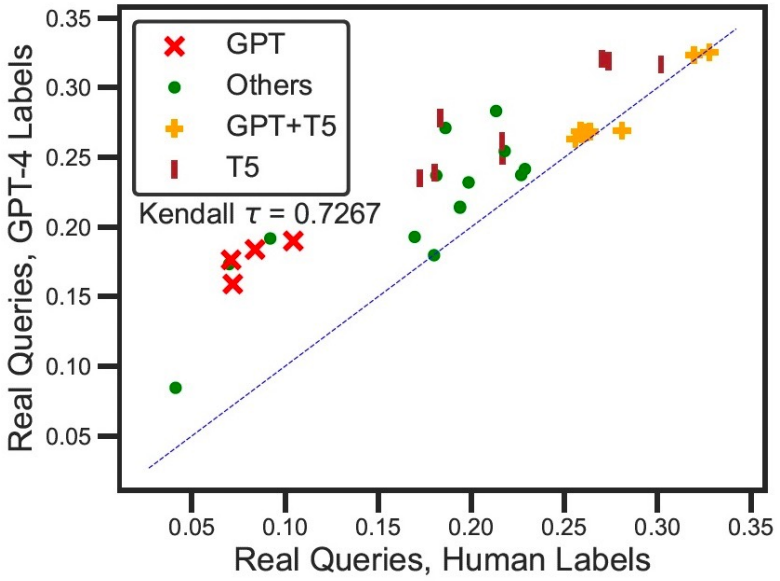
- LLM judges tend to prefer responses generated by LLMs [8, 9]
  - LLMs tend to be trained using similar approaches, using similar datasets
  - This bias could become worse when identical models are used for response generation and judging
- LLMs tend to favour longer responses (length bias) [10]
  - Not good at judging verbosity
- Similar biases observed when using MLLM-as-a-judge for multi-modal tasks

[8] Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. COLM 2024
[9] Towards Understanding Bias in Synthetic Data for Evaluation. CIKM 2025
[10] Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. SIGIR 2025

Limitations of Using LLM-as-a-Judge: Bias in Evaluation Results



NDCG@10



MAP

Towards Understanding Bias in Synthetic Data for Evaluation. CIKM 2025

Limitations of Using LLM-as-a-Judge: Bias in Judgments

- Linear Mixed-Effects Model to assess differences in metric scores when using evaluations obtained using human judges as a reference

- Three types of systems:
  - GPT, T5, GPT+T5
  - Evaluated using GPT-4 as a judge

$$
\begin{aligned}
Y_{ij} = {} & \beta_0 + \beta_1 \,\mathrm{JudgeGPT4}_{ij} \\
& + \beta_2 \,\mathrm{SysType}_{ij,\mathrm{GPT}} + \beta_3 \,\mathrm{SysType}_{ij,\mathrm{T5}} + \beta_4 \,\mathrm{SysType}_{ij,\mathrm{T5+GPT}} \\
& + \beta_5 \,\mathrm{QDR}_{ij} + \beta_6 \,\mathrm{QW}_{ij} + \beta_7 \,\mathrm{DL}_{ij} + \beta_8 \,\mathrm{isGPT4}_{ij} + \beta_9 \,\mathrm{MN}_{ij} \\
& + \beta_{10} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{SysType}_{ij,\mathrm{GPT}}\right) \\
& + \beta_{11} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{SysType}_{ij,\mathrm{T5}}\right) \\
& + \beta_{12} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{SysType}_{ij,\mathrm{T5+GPT}}\right) \\
& + \beta_{13} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{QDR}_{ij}\right) \\
& + \beta_{14} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{QW}_{ij}\right) \\
& + \beta_{15} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{DL}_{ij}\right) \\
& + \beta_{16} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{isGPT4}_{ij}\right) \\
& + \beta_{17} \left(\mathrm{JudgeGPT4}_{ij} \times \mathrm{MN}_{ij}\right) \\
& + \alpha_j + \varepsilon_{ij}
\end{aligned}
$$

Towards Understanding Bias in Synthetic Data for Evaluation. CIKM 2025

Limitations of Using LLM-as-a-Judge: Bias in Evaluation Results

- Factors playing a significant role
  - Judged by GPT-4
  - System type GPT
  - System type GPT + T5
  - Query difficulty
  - Document length
  - Interactions:
    - Judged by GPT-4, System type GPT
    - Judged by GPT-4, System type GPT+T5
    - Judged by GPT-4, Query length
    - Judged by GPT-4, Document length

Towards Understanding Bias in Synthetic Data for Evaluation. CIKM 2025

Implications for System Development

- Deploy “bad” systems just because
  - They use the same model as the one used as LLM-as-a-judge
  - They are based on LLMs
- Results presented to users could get dominated with responses generated by AI
  - More prone to hallucinations and certain biases
- Problems with ”model collapse”
  - Issues faced could become more severe over time
  - Responses become more verbose, etc.

Ways to Improve Limitations of LLM-as-a-Judge

- Use LLMs in combination with human assessors
  - Human assessments to continuously monitor LLM-as-a-judge
  - Monitor not just the overall quality of the judgments, but also for possible biases in the judgments

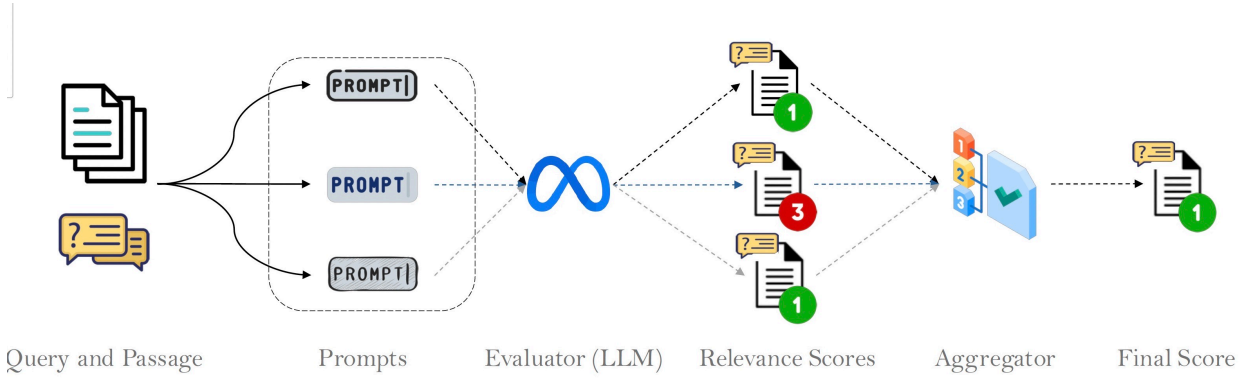
- Utilise the power of multiple models to improve problems with quality and bias

Utilising the Power of Multiple Models

- Different prompts or models could bring in different perspectives to evaluation, which can help to
  - Reduce quality issues
  - Bias in judgments and evaluation scores

- Aggregating multiple models could also help us utilise multiple smaller models to outperform the performance of larger models

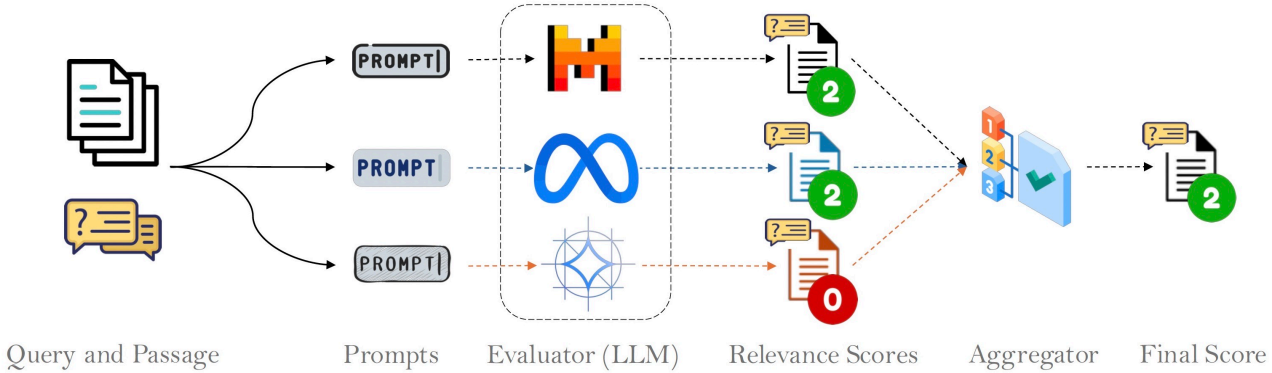
Utilising the Power of Multiple Models: Blending Multiple Prompts



Promptblender: Blending different Prompts

JudgeBlender: Ensembling Automatic Relevance Judgments. The Web Conference (WWW) 2025.

Utilising the Power of Multiple Models: Blending Multiple LLMs



LLMBlender: Blending different LLMs

JudgeBlender: Ensembling Automatic Relevance Judgments. The Web Conference (WWW) 2025.

Utilising the Power of Multiple Models

Method	Model	Cohen's Kappa
Best performing baseline (larger models)	GPT-4o	0.2519
Fine-tuned methods	Llama-3-8B	0.1823
PromptBlender	Llama-3-8B	0.2436
LLMBlender	Mistral 7B Gemma 7B Llama-3-8B	0.2619

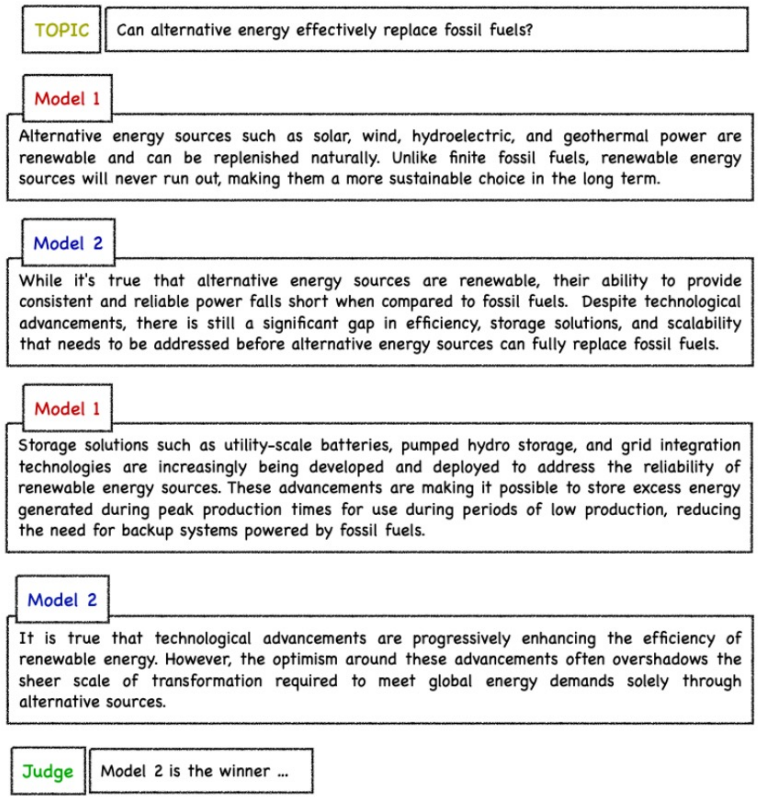
[11] JudgeBlender: Ensembling Automatic Relevance Judgments. The Web Conference (WWW) 2025

Utilising the Power of Multiple Models : LLMs Debating

Multiple LLMs debate to reach consensus on evaluation.

Each LLM is assigned to support one side of the argument.

Final decision made by a separate judge.



Evaluating the Performance of Large Language Models via Debates. ACL 2025.

Utilising the Power of Multiple Models : LLMs Debating

Since LLMs are typically not directly trained to debate, specifying the rules of the task is crucial via system prompts

- ask the models to support a side;
- ask that the arguments and rebuttals should be backed by logic, facts, and evidence;
- and that the answers should be convincing, factual and concise.

Judge model could be asked to state the main reason for its choice such as *clarity, factuality, counterarguments, persuasiveness, conciseness and coherence*.

Evaluating the Performance of Large Language Models via Debates. ACL 2025.

Utilising the Power of Multiple Models : LLMs Debating

Limitations:

An LLM judge "*consistently favors the side with the same LLM backbone*" in a debate.

- Either use completely homogenous or heterogenous models.

Without proper adversarial setup, agents might converge too quickly or implicitly collude.

- Design prompts to encourage genuine disagreements (not too much).

Cost and scalability

Summary

- Why is evaluation important?
- Commonly used metrics and dimensions for evaluation
- LLM-as-a-judge examples
- Limitations of using LLM-as-a-judge
  - Quality, multi-modal evaluation, bias in judgments and bias in evaluations
- Ways to improve these limitations
  - Use LLMs in combination with human assessors
  - Utilise the power of multiple models to improve problems with quality and bias