

# LLM-as-a-Judge in Human Evaluation of Model Performance

May 28, 2025



# You might be familiar with this

MT Evaluation Project								Share
File Edit View Insert Format Data Tools Extensions Help								
125% \$ % 0.00 123 Default... 10 B I A								
H12								Summarize this tab
	A	B	C	D	E	F	G	H
1	Task ID	Language	Source Text	MT Output	Annotator	Translation Quality Score (1-5)	Error Categories	Comments
2	001	EN->FR	That is why Mr Savary's report deserves our full support.	C'est pourquoi le rapport de M. Savary mérite tout notre soutien.	Sara	5		
3	002	EN->FR	I think that is important.	C'est important.	John	5	accuracy	Missing context: The original sentence implies "I think", which expresses the speaker's subjective opinion.
4	003	EN->FR	A difficult issue was that of personnel.	Une question difficile était celle du personnel.	John	5		
5	004	EN->FR	Mr Savary, I welcome the compromise that we were able to reach in that respect.	Monsieur Savary, je me félicite du compromis auquel nous avons pu parvenir dans ce respect.	David	5		
6	004	EN->FR	Mr Savary, I welcome the compromise that we were able to reach in that respect.	Monsieur Savary, je me félicite du compromis auquel nous avons pu parvenir dans ce respect.	Sara	4	fluency	Using "dans ce respect" instead of "à cet égard". While "dans ce respect" is understandable, "à cet égard" is more natural and idiomatic in French.
7	005	ES->EN	Gracias.	Thank you very much.	Jose	2	accuracy	"Gracias" simply means "Thank you". The addition of "very much" is an expansion of the original meaning.
8	006	ES->EN	Señora Presidenta, ¿se ha contabilizado mi voto, que no ha podido ser realizado electrónicamente, porque no tengo la tarjeta?	Madam President, has my vote been counted? I was unable to vote electronically, since I do not have a card.	Ana	5		
9	007	ES->EN	Mi voto era "a favor" .	My vote was "in favour" .	Ana	4	locale	Favour vs. Favor, is a regional variation

# What if it could look like this?

### MT Evaluation

Source	MT Output
That is why Mr Savary's report deserves our full support.	C'est pourquoi le rapport de M. Savary mérite tout notre soutien.
-----	

Error 1

#### Your Score of Translation Quality

☆ ☆ ☆ ☆ ☆

\* Catastrophic

the translation is incomprehensible or contains mistakes that are life-threatening or could damage the author's reputation.

\*\* Inadequate

the translation is hard to understand or use and cannot be relied on.

\*\*\* Passable

the translation is overall comprehensible, though it contains some errors or it's not fluent.

\*\*\*\* Good

Most of the meaning is present in the translation and the language is fluent. Only few minor errors are allowed.

\*\*\*\*\* Perfect

no errors and native fluency

Comment (optional)

↶ ↷ ✕ ≡

Submit

humansignal.com

# Using an LLM to Judge the Output of Other Models



# LLM-as-a-Judge



<https://www.deepchecks.com/what-is-llm-as-a-judge-strategies-impact-and-best-practices/>

## Pros:

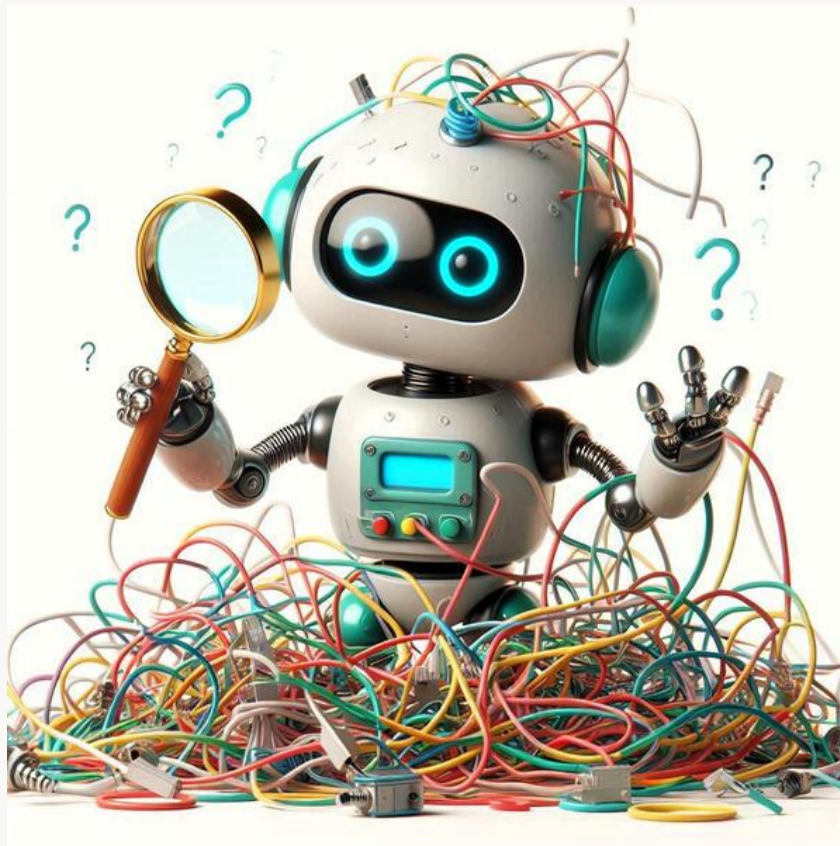
- Faster and cheaper than humans
- Can handle cases where a rule-based approach for evaluation won't work
- Scalable and explainable



# The problem with LLM-as-a-Judge

## Bias

- Position Bias – the first thing is always the best
- Verbosity Bias – the longest thing is the best
- Self Enhancement Bias – what I wrote is best



# The problem with LLM-as-a-Judge



- How well do they mimic humans?
- What context does the LLM have for *your* needs?



# How do we audit LLM-as-a-Judge?





# Human-created test set

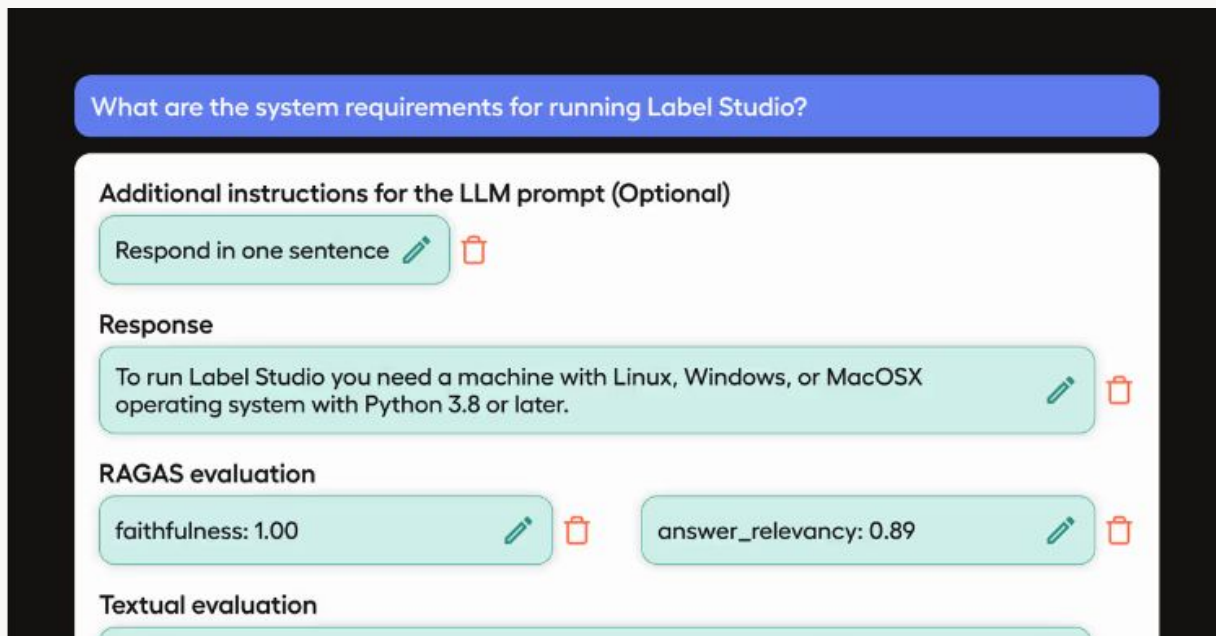


- Allows us to compare LLM-as-a-judge with our human raters
- Can aid in prompt tuning
- Helps us understand where our models do well (and where they can improve)





# Check for multiple types of “correctness”

Relevancy, accuracy, clarity, and more!





What are the system requirements for running Label Studio?





Additional instructions for the LLM prompt (Optional)

Respond in one sentence  

Response

To run Label Studio you need a machine with Linux, Windows, or MacOSX operating system with Python 3.8 or later.  

RAGAS evaluation

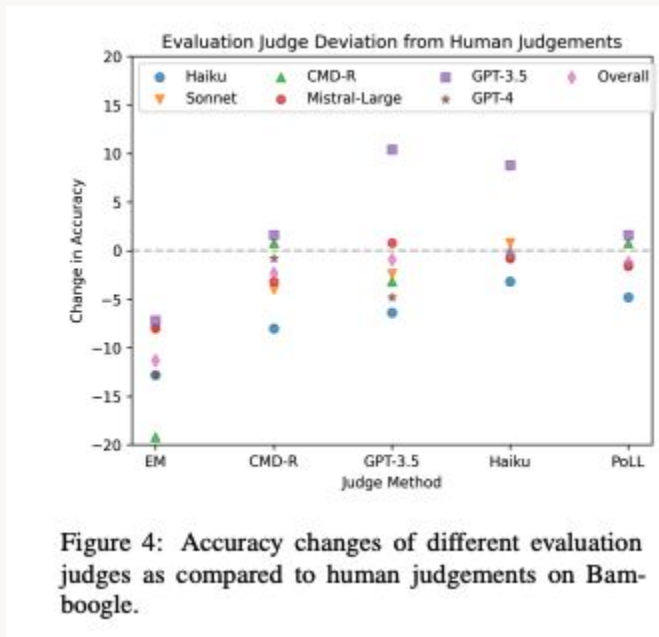
faithfulness: 1.00  	answer_relevancy: 0.89  
--	--

Textual evaluation



# LLM-as-a-Jury

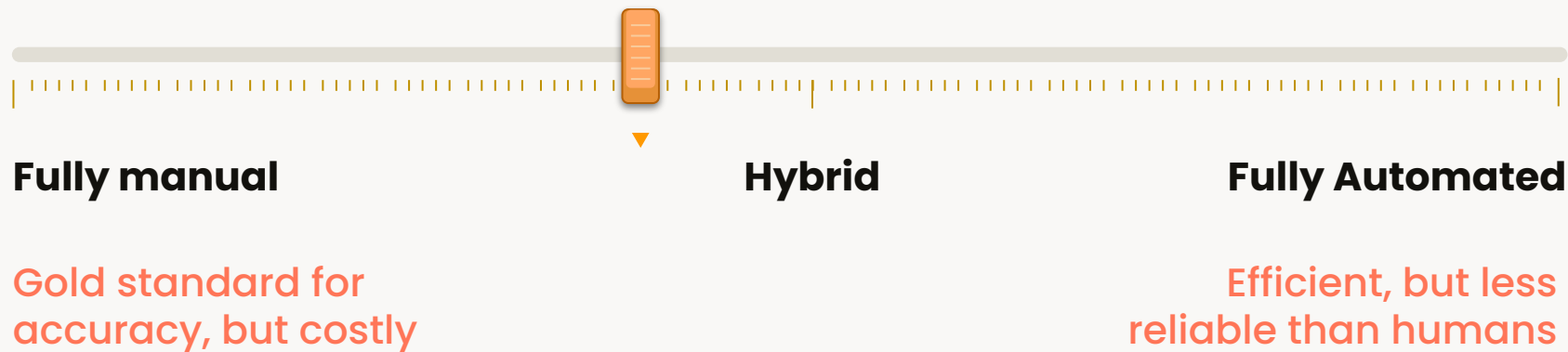
Two heads are better than one



Verga et al 2024: Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models



# The challenges inherent in AI model evaluation



The best solution: a hybrid approach that effectively balances cost with reliability

# Label Studio

[labelstud.io](https://labelstud.io)

## Open Source Data Labeling Platform

The most flexible data labeling platform to fine-tune LLMs, prepare training data or validate AI models.

Quick Start

Compare Versions

LAST COMMIT: MAY 12, 2025

LATEST VERSION: NIGHTLY

### Quick Start

PIP BREW GIT DOCKER

```
1 # Install the package
  # into python virtual environment
2 pip install -U label-studio
3
4 # Launch it!
5 label-studio
```



# Sources

Zheng et. al. 2023, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, <https://arxiv.org/pdf/2306.05685>

Verga et al 2024: Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models, <https://arxiv.org/abs/2404.18796>



# Questions?

Micaela Kaplan

[micaela@humansignal.com](mailto:micaela@humansignal.com)

