Using Natural Language Explanations to Rescale Human Judgments

TAUR LAB @UT Austin

Manya Wadhwa, Jifan Chen, Jessy Li, Greg Durrett

manya.wadhwa@utexas.edu

Task: Rate LLM outputs on a continuous scale

Use case: document-grounded QA. Annotators evaluate **answer completeness** with Likert judgments and explanations.

Source Document

Question

Why was the entire climbing season in doubt?

LLM Response

The Sherpas had walked out in protest of the deaths of 16 of their colleagues in an avalanche, and their demands for better pay, treatment and benefits.

Label: missing minor information

Explanation: *It is also unknown if the Sherpas would accept and return.* Sentences missing: 8

Label: missing minor information

Explanation: *The machine response answered the question correctly but missed some relevant information that clarifies the Sherpas' demands. Sentences 10, 15, 40 are missing*

Label: missing major information

Explanation: *Important information was neglected; sentences 11 and 15 are missing*

Our contribution: a method to turn judgments of LLM responses into *consistent numeric scores* based on labels and explanations.

Why labels and explanations?

- Labels alone are not fine-grained
- Explanations indicate different information annotators consider when giving judgment

Why numeric scores?

- Enable precise system evaluation
- Fine-grained rewards for RLHF

Method: Use an LLM to rescale label and explanation

Source Document

Why was the..

The article says…

Answer to be judged

missing all info — *missing major info* — *missing minor info* — *complete*

This answer should've given more detail…it is missing the following..

LLM Scorer

Takes label, explanation, and Rubric

(derived from explanations & can be hand-crafted)

Explanation- Based Rescaled Score (0-100)

- Reflects annotator's "true" assessment
- Rubric makes consistent with other annotators

Dataset

Step 1:

We collect 12.6k judgments from multiple human annotators on "answer completeness" of LLM response on task of document-grounded QA

Step 2:

Out of the 12.6k judgments, we sample 145 instances and get expert rescaling.

Expert rescaling

LLM rescaling

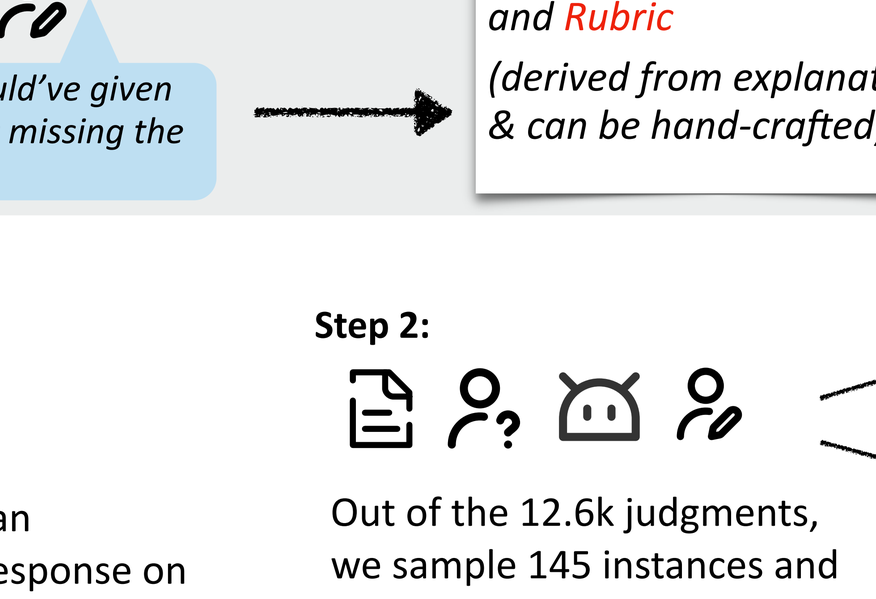
Compare!
- Kendall's Tau
- Mean Absolute Error

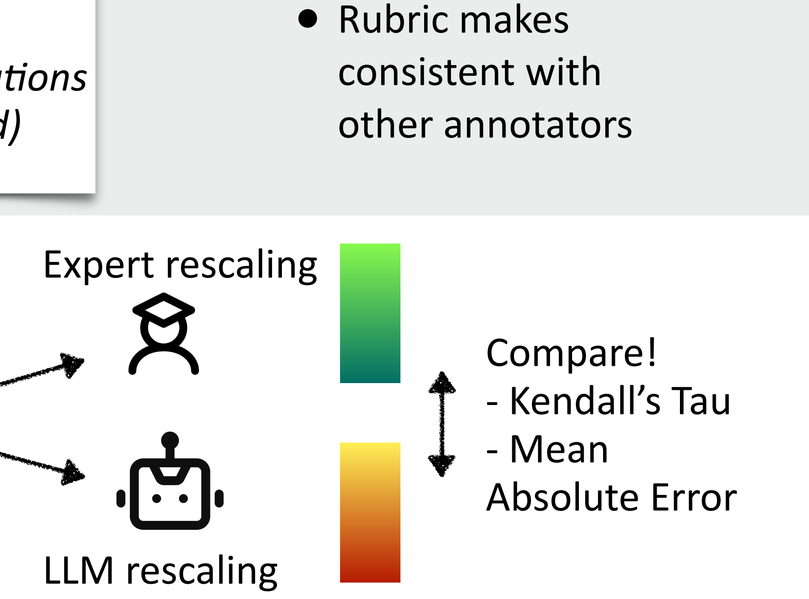
Results: Can LLMs do the rescaling?

Yes! Given a rubric*

We measure how *gpt-4-0613 rescaling* compares to reference rescaling (expert rescaled scores).

1. Heuristics like number of missing sentences do not rescale well - poor correlation and high MAE
2. *Rubric is* very important for the task - provides a grounding for rescaling the values





Takeaways

- LLMs rescale annotator judgments effectively
- Rescaled annotations align with expert reference rescaling
- Rescaling preserves correlation, capturing nuances, subjectivity and scale use differences

More Results and Analysis in the Full Paper