

# LLM-RUBRIC: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie



## Motivation

- Increasingly, practitioners are turning to LLMs to assess large numbers of text documents.  
*But can LLM evaluation be trusted?*
- Humans are the gold standard, but they may not agree on complex/subjective tasks.  
*How do we align a single LLM and with multiple human judges?*

## Experiments

We model a human judge pool on a dialogue system evaluation task.

- Each judge  $a$  evaluates a user/system transcript  $T$  for overall user satisfaction ( $Q_0$ ).
- Each judge's rating  $y_0^a$  of  $T$  is on a 4-point Likert scale (i.e.,  $y_0^a \in \{1, 2, 3, 4\}$ ).

## Main Findings

- Eliciting Likert scale ratings from an LLM either via classification (Argmax LLM  $Q_0$ , Fig. 1)

$$y_0^a = \operatorname{argmax}_{y \in \{1, 2, 3, 4\}} p_{LLM}(y|T; Q_0)$$

or regression (Expected LLM  $Q_0$ , Fig. 1)

$$y_0^a \neq \frac{y p_{LLM}(y|T; Q_0)}{\sum_{y \in \{1, 2, 3, 4\}} y p_{LLM}(y|T; Q_0)}$$

performs worse than a constant baseline.

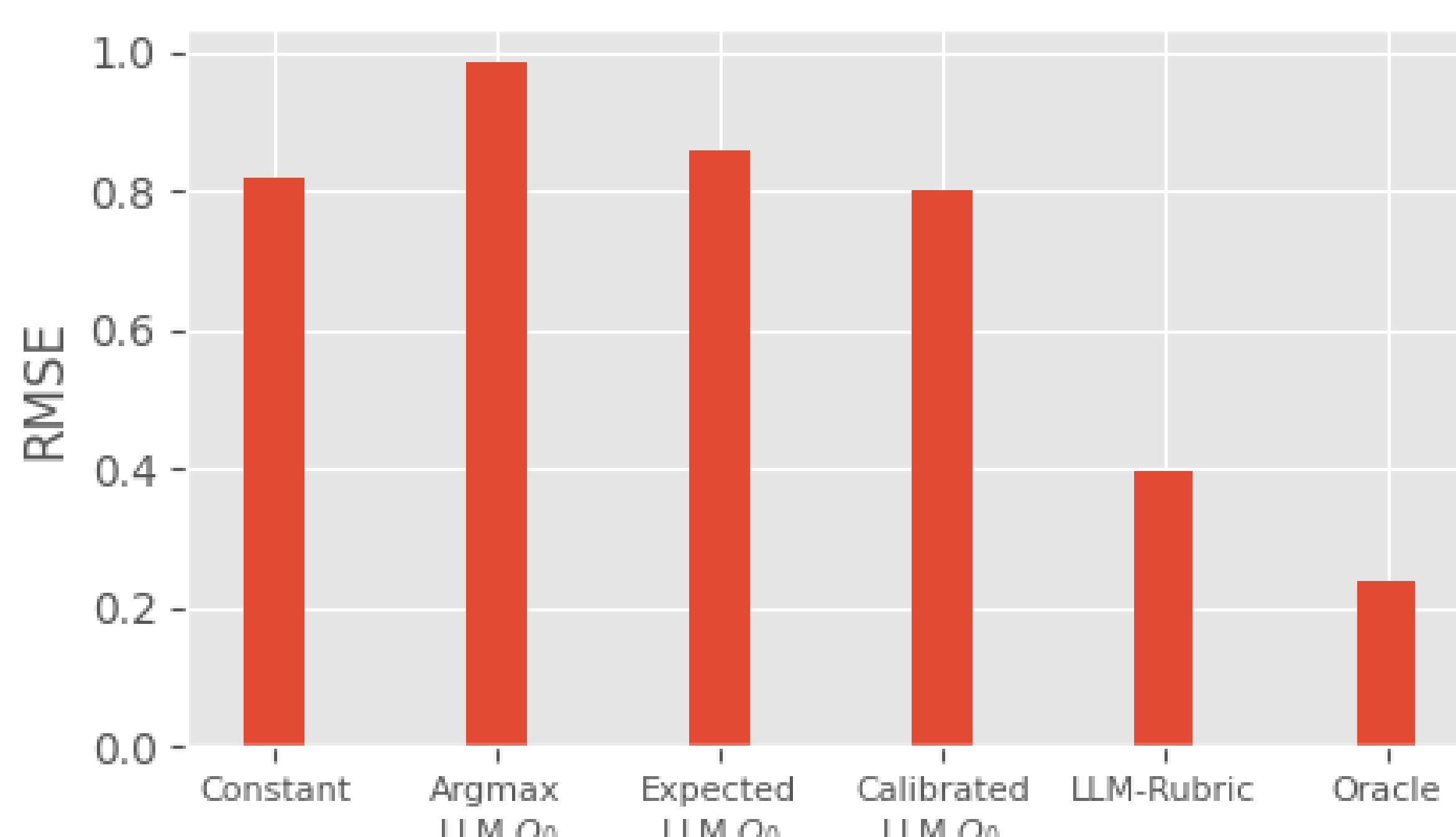


Figure 1. Evaluation results on dialogue evaluation task. Constant is the training set mean of  $y_0^a$ . Oracle uses ground truth  $y_1^a; \dots; y_8^a$  to predict  $y_0^a$ .

## Main Findings (continued)

- Learning to map  $p_{LLM}(y|T; Q_0)$  to each judge  $a$ 's distribution  $\hat{p}_a(y|T; Q_0)$  via a small feed-forward network (**Personalized Calibration Network (PCN)**) works better (Calibrated LLM  $Q_0$ , Fig. 1).

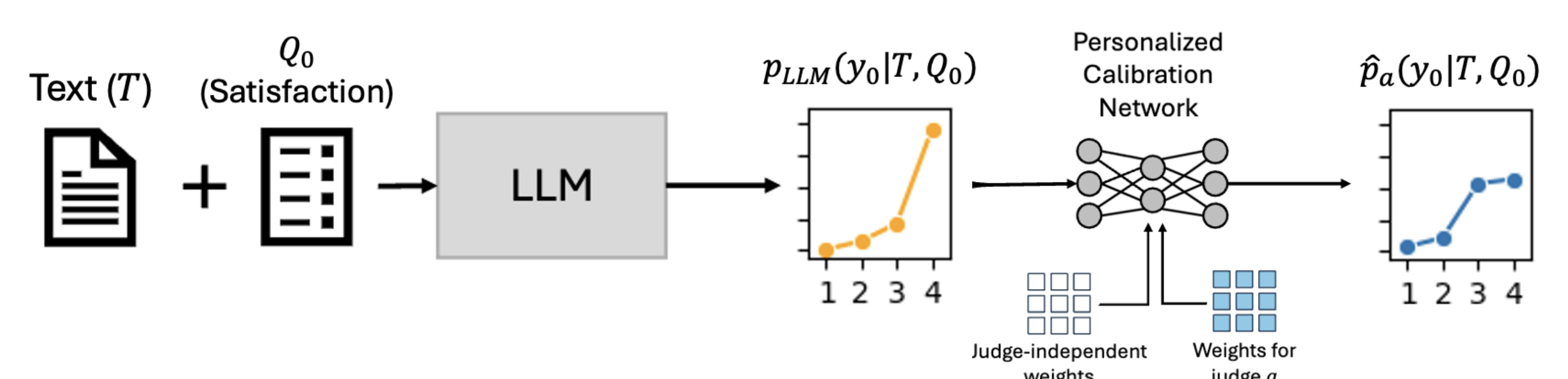


Figure 2. A high-level view of the Personalized Calibration Network.

- LLM-RUBRIC** Even better, learning to map  $p_{LLM}$  on related evaluation criteria ( $Q_1; \dots; Q_n$ ) to  $\hat{p}_a$  via the PCN further improves accuracy of  $\hat{p}_a(y|T; Q_0)$  (LLM-Rubric, Fig. 1).

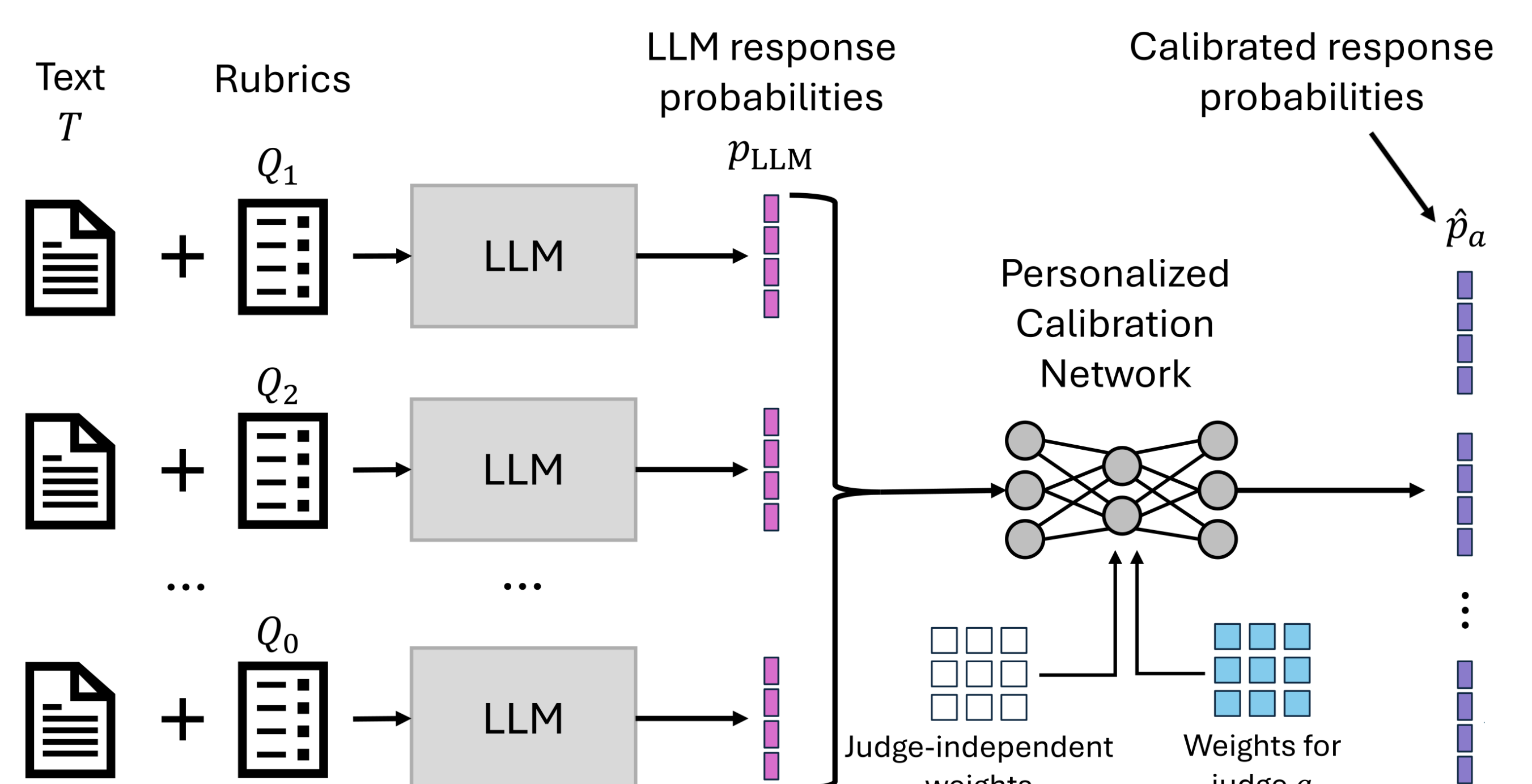


Figure 3. A high-level view of LLM-RUBRIC.

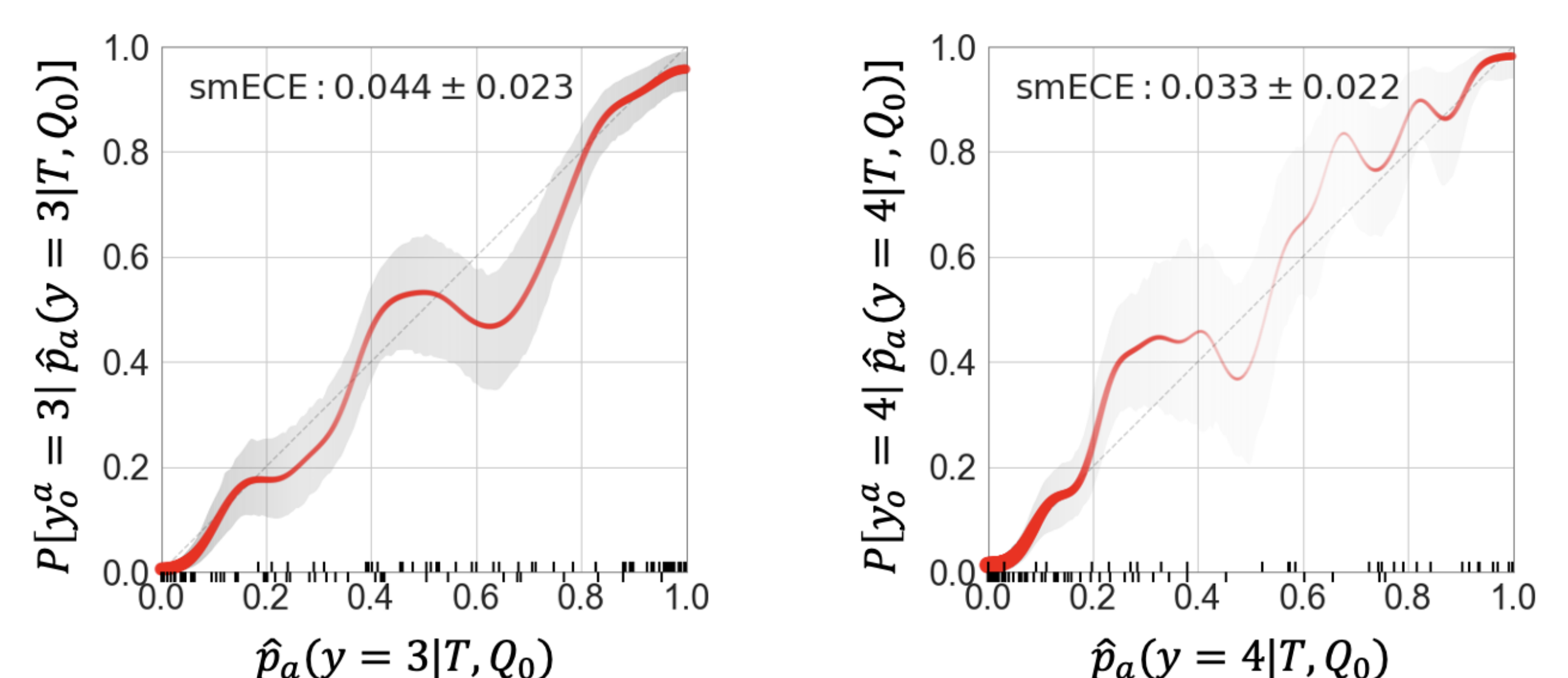


Figure 4. Reliability diagrams show that LLM-RUBRIC is well-calibrated.

- LLM-RUBRIC is well-calibrated, enabling future work on adaptive rubric item selection, identifying disagreement among different judge sub-populations, and more. See the paper for more details.

Code and data at: <https://github.com/microsoft/llm-rubric>