

# LLM-RUBRIC: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts<sup>†</sup>

**Helia Hashemi\*** **Jason Eisner\*** **Corby Rosset** **Benjamin Van Durme** **Chris Kedzie**  
Microsoft

{hhashemi,jeisner,corbyrosset,ben.vandurme,chriskedzie}@microsoft.com

## Abstract

This paper introduces a framework for the automated evaluation of natural language texts. A manually constructed rubric describes how to assess multiple dimensions of interest. To evaluate a text, a large language model (LLM) is prompted with each rubric question and produces a distribution over potential responses. The LLM predictions often fail to agree well with human judges—indeed, the humans do not fully agree with one another. However, the multiple LLM distributions can be *combined* to *predict* each human judge’s annotations on all questions, including a summary question that assesses overall quality or relevance. LLM-RUBRIC accomplishes this by training a small feed-forward neural network that includes both judge-specific and judge-independent parameters. When evaluating dialogue systems in a human-AI information-seeking task, we find that LLM-RUBRIC with 9 questions (assessing dimensions such as naturalness, conciseness, and citation quality) predicts human judges’ assessment of overall user satisfaction, on a scale of 1–4, with RMS error < 0.5, a 2× improvement over the uncalibrated baseline.

## 1 Introduction

Many fields that must assess large numbers of short documents have turned to NLP-assisted workflows. For example, lawyers conducting legal discovery must identify all relevant documents (Quartararo et al., 2019)—a task also faced by journalists and historians. Social scientists and market researchers must code survey responses (Mellon et al., 2024; enumerate.ai; ATLAS.ti). Teachers or examiners must evaluate student writing (Page, 1968; Ramesh and Sanampudi, 2022) and provide feedback (Meyer et al., 2024). Doctors, social workers, or public health agencies or researchers may assess an individual’s mental health or safety from their social media posts (Chancellor and De Choudhury, 2020; Xu et al., 2024; Al-Garadi et al., 2022) or

from clinical interviews and assessments (Galatzer-Levy et al., 2023).

The above settings evaluate human-authored texts. In addition, NLP developers must assess the quality of their machine-generated texts—texts that are consumed by end users, but also hidden intermediate steps in agentic workflows (such as chains of thought, tool calls, and revisions). With the recent commercialization of conversational AI, for example, it is crucial to evaluate dialogue systems during development and monitor them after deployment. Special care is needed in high-stakes settings like medical dialogue (Huang et al., 2024).

Manual evaluation has long been the gold standard for assessing text, including generated text (Saphra et al., 2023; van der Lee et al., 2021). Humans are often asked to consider multiple criteria and then provide a final assessment (Hosking et al., 2023). Humans may also be asked to produce reference answers to which other humans can compare the target text. Yet manual evaluation is expensive, time-consuming, and not without its own quality and reliability issues (Hosking et al., 2023; Liu et al., 2016; Smith et al., 2022). Because of these challenges, and the increasing abilities of large language models (LLMs) (Brown et al., 2020), experimenters have recently been eliciting ratings directly from an LLM (Chiang and Lee, 2023; Fu et al., 2023; Liu et al., 2023a; Thomas et al., 2024; ChainForge; and others). *But can LLM evaluation be trusted?* It solves the time, scaling, and possibly cost issues, but leaves open the problem of aligning these LLM ratings with human judgments.

We present a general approach to this alignment problem. We demonstrate its value for the evaluation and comparison of LLM-powered dialogue systems, in an information-seeking dialogue task (Zamani et al., 2023) similar to Lowe et al. (2015). Evaluation in this setting is complex owing to competing factors that might affect a human judge’s assessment of the dialogue. These may include correctness of responses, accuracy and helpfulness of citations, length and complexity of responses, and more (Smith et al., 2022).

\*Equal contribution.

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

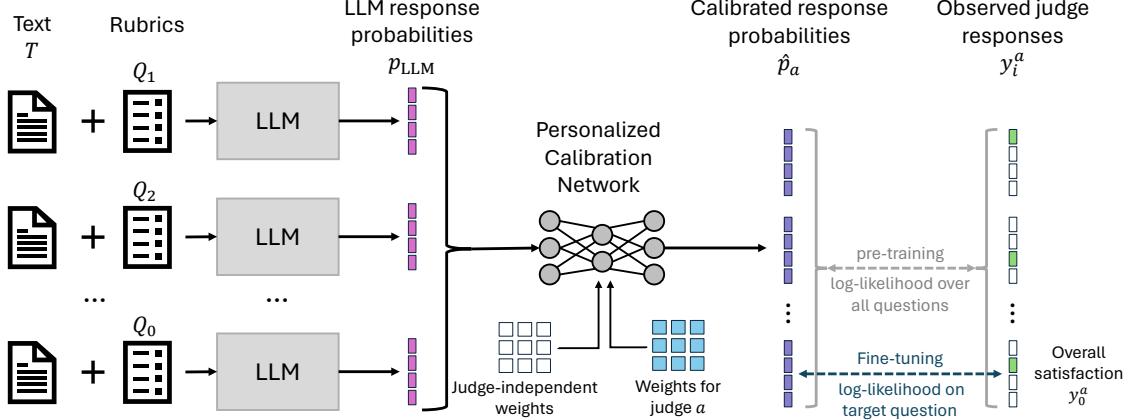


Figure 1: An overview of the LLM-RUBRIC framework. The LLM and its prompts are fixed across texts and judges, but the calibration network weights are trained to predict the responses of various human judges.

Our LLM-RUBRIC approach begins with a manually authored evaluation rubric. The rubric’s multiple-choice questions cover various evaluation dimensions, and it may also include a question that assesses *overall* quality or relevance. Evaluating a text, such as a dialogue, then consists of two main steps: (1) for each rubric question we elicit the LLM’s probability distribution over possible responses, by prompting it with the text and the rubric question, and (2) we aggregate and calibrate these distributions with a small feed-forward network that has been trained to match the individual preferences of human judges. A high-level overview of LLM-RUBRIC is shown in Figure 1.

For research in generative NLP, once the rubric and LLM are fixed, LLM-RUBRIC can be used like other metrics (BLEU, ROUGE, etc.) to drive system development, monitor quality, demonstrate the value of a new technique, and conduct competitions. In our dialogue evaluation experiments, each user–AI dialogue is evaluated by 3 trained annotators (randomly drawn from a larger pool) who each answered the same 9 rubric questions. Our method uses these data to train an automatic LLM-based evaluator, without treating the 24 human annotators as interchangeable. Overall, we find<sup>1</sup> that

- Personalized calibration of an LLM evaluator of overall satisfaction on < 750 synthetic dialogues significantly improves its prediction of human judgments and correlations with human judgments, but still works poorly.
- Incorporating LLM evaluations of 8 additional criteria (LLM-RUBRIC) improves these metrics by over 2× over the uncalibrated LLM.

<sup>1</sup>See Table 1, right side, rows 3, 4, and 6.

Accurate automated text assessment could replace human assessment in many other settings, such as those reviewed at the start of this paper. It could also be used in new settings where human assessment was never feasible. In AI-powered user interfaces, instantaneous scoring of user-written text can feed into downstream decisions such as providing writing feedback or deciding how to proceed with a dialogue. An AI reasoning engine may internally apply a rubric to assess the validity of a proposed natural-language reasoning step (Weir et al., 2024). When processing a large document collection, an LLM can be used to assess the compatibility of two text passages (Zhang et al., 2023; Viswanathan et al., 2023; Choi and Ferrara, 2024), potentially in a more nuanced way than vector similarity; this problem arises in workflows for matching, routing, clustering, and fact-checking (Charlin and Zemel, 2013; Harman, 1996; and the papers just mentioned). Finally, automated assessments could provide signals for *training* text generation (Keskar et al., 2019; Tambwekar et al., 2019; Bai et al., 2022).

To allow LLM-RUBRIC to support all of these use cases, we release general code along with the datasets we created for this paper (see URL on page 1). We discuss limitations at the end of the paper.

## 2 The LLM-RUBRIC Method

It is challenging to model human preferences in a combinatorial space such as text. Reasonable human judges may differ (Aroyo and Welty, 2015) on (1) what textual properties they happen to prefer (e.g., concise vs. detailed, formal vs. informal, novice vs. expert audience), (2) how they combine multiple preferences into an overall assess-

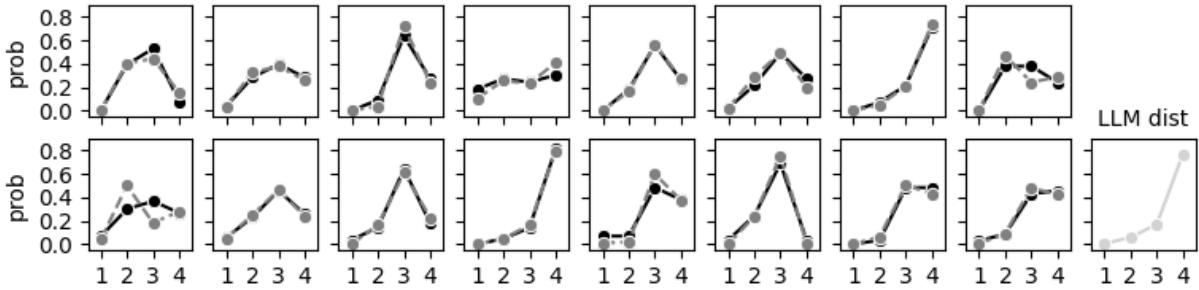


Figure 2: Our calibration network learns how different human judges use the response range 1–4. Each black curve shows a different judge’s distribution of responses to the “overall satisfaction” question  $Q_0$  on our synthetic conversation dataset. (We show the judges who evaluated  $\geq 30$  conversations.) The corresponding gray curve shows the average distribution predicted for that judge on the same dialogues by LLM-RUBRIC (using cross-validation). The final curve in light gray shows the original uncalibrated distribution of responses to  $Q_0$  by the LLM (gpt-3.5-turbo-16k).

ment, and (3) how they convey that assessment through a numerical score. Figure 2 shows that in our dataset (§3), different human judges indeed have very different marginal distributions of overall score. Clearly these cannot all be matched by a judge-independent system (e.g., the LLM shown at the lower right of Figure 2).

To expose the different properties and preferences at play, we ask the human judges a series of finer-grained questions about different evaluation criteria. It is already common in practical settings (§1) to at least mention such criteria in instructions to human judges. We use the same questions to query an LLM,<sup>2</sup> and train a calibration network to jointly adjust the LLM’s scores to match the scores of any given human judge. We refer to this methodology as LLM-RUBRIC. The gray curves in Figure 2 show that on held-out dialogues, the calibrated overall score is now distributed like that of the given judge. We will see later that these scores are also more accurate on the individual dialogues.

In this section, we present LLM-RUBRIC in a general way, but for concreteness, we also introduce details of our specific experimental setup.

**Evaluation Rubric Construction.** We wrote 8 dialogue evaluation questions ( $Q_1, \dots, Q_8$ ) inspired by the NLG evaluation literature (Zhou et al., 2022; van der Lee et al., 2021). These questions are shown in Appendix C. They address various dimensions such as naturalness, relevance, attribution, citation quality, and conciseness. Our final

<sup>2</sup>It is convenient to use the *same* questions, as we have already crafted them. However, different or additional questions could in principle be used—or multiple variants of each question, or multiple LLMs. This could potentially provide more useful evidence to the calibration network below, at the cost of slowing down evaluation and at the risk of overfitting.

question ( $Q_0$ ) asked the judge to assess the overall quality of the dialogue (in this case, focusing only on whether the user would be satisfied), on a Likert scale of 1–4. Each question stated its allowed multiple-choice responses (usually scores 1–4, with a meaning provided for each score).

#### Multi-Dimensional Evaluation with LLMs.

We use an LLM to evaluate a given text  $T$  (in our case, a dialogue transcript). For each question  $Q_i$  ( $0 \leq i \leq 8$  in our case), we instruct the LLM to generate a label  $y_i \in \mathcal{Y}_i$ , where  $\mathcal{Y}_i$  is the set of allowed responses to  $Q_i$  (e.g., {“1”, “2”, “3”, “4”}). Specifically, we prompt it with a preamble, the text  $T$ , and the question  $Q_i$ , where  $Q_i$  also specifies the allowed responses  $\mathcal{Y}_i$  (see Appendix D). We chose to do this independently for each question  $Q_i$  to avoid confounding the LLM’s responses. We thus obtain  $p_{\text{LLM}}(y_i | T, Q_i)$  for all questions  $Q_0, \dots, Q_8$  and each possible response  $y_i \in \mathcal{Y}_i$ .<sup>3</sup>

**Aggregated Evaluation with Personalized Calibration.** We then use a small feed-forward *calibration network* (Figure 1 and equations (3)–(5) below) to map this collection of LLM probabilities  $p_{\text{LLM}}(y_i | T, Q_i)$  to a collection of adjusted probabilities  $\hat{p}_a(y_i | T, Q_i)$  that predict the responses of a particular judge  $a$ . Note that each  $\hat{p}_a(y_i | T, Q_i)$  is predicted from the LLM’s behavior on *all questions* about  $T$ , not just  $Q_i$ . This design lets the calibration network inspect some additional proper-

<sup>3</sup>The LLM also allocates some probability to responses outside  $\mathcal{Y}_i$ , so  $Z_i \stackrel{\text{def}}{=} \sum_{y_i \in \mathcal{Y}_i} p_{\text{LLM}}(y_i | T, Q_i) < 1$ . We do not normalize the probabilities by  $Z_i$  before presenting them to the calibration network. This allows our calibration network, in principle, to notice when  $Z_i \ll 1$  and to learn not to rely on the LLM’s answer to  $Q_i$  in such cases. In practice, however, our prompts result in  $Z_i$  being very close to 1.

ties of  $T$  that might influence  $a$ 's response to  $Q_i$ .<sup>4</sup> This design also extends to the case where the LLM was not asked the specific question  $Q_i$  for which we are predicting  $a$ 's response (see footnote 2).

We train the calibration network by maximum likelihood (regularized by early stopping). That is, given a dataset  $\mathcal{D}$  of annotations, we maximize<sup>5</sup>

$$\sum_{(T,i,a,y_i^a) \in \mathcal{D}} \log \hat{p}_a(y_i^a | T, Q_i) \quad (1)$$

where  $(T, i, a, y_i^a) \in \mathcal{D}$  means that judge  $a$  answered  $Q_i$  on  $T$  with response  $y_i^a$ .

**Decoding.** Given a new text  $T$ , the trained calibration network predicts any judge  $a$ 's possible responses to question  $Q_i$  via the *distribution*  $\hat{p}_a(y_i | T, Q_i)$ . If we wish to output a *single* predicted value  $\hat{y}_i^a$  for downstream use, then we also need a *decoding principle* that extracts  $\hat{y}_i^a$  from  $\hat{p}_a$ . In our experiments, actual responses  $y_i^a$  are integers, predictions  $\hat{y}_i^a$  are real numbers, and we will be evaluating the predictions by  $L_2$  loss,  $(\hat{y}_i^a - y_i^a)^2$ .<sup>6</sup> Thus, our principle is to minimize the *expected*  $L_2$  loss (our “Bayes risk”). This is accomplished simply by predicting the mean of distribution  $\hat{p}_a$ ,

$$\hat{y}_i^a = \sum_{y_i \in \mathcal{Y}_i} \hat{p}_a(y_i | T, Q_i) \cdot y_i \quad (2)$$

We remark that we could have constructed a network that directly predicted the  $\hat{y}_i^a$  values, and trained it to minimize  $L_2$  loss on training data—a regression problem. However, by modeling the entire distribution  $\hat{p}_a$  and not just its mean, we make fuller use of the training data for representation learning—our representations are trained to be able to predict the full distribution. Indeed, we found in pilot experiments that our method slightly outperforms the regression method. Furthermore, modeling  $\hat{p}_a$  lets us report our predictive uncertainty—e.g., the entropy or variance of  $\hat{p}_a(y_i | T, Q_i)$  and not just its expectation  $\hat{y}_i^a$ . Finally, equation (2) nicely guarantees that  $1 \leq \hat{y}_i^a \leq 4$  on any example.

<sup>4</sup>In the future, for this reason, the calibration network's input could also include an embedding of the full text  $T$ .

<sup>5</sup>This formula models the  $y_i^a$  for different  $i$  as conditionally independent given  $T$ . This assumption could be relaxed. For example, perhaps all of the  $y_i^a$  should be made to also depend on a latent variable, e.g., judge  $a$ 's mood while annotating  $T$ .

<sup>6</sup>This setup treats the integers as falling on an interval scale, not just an ordinal scale. For example, outputting 1.4 when the true answer is 1 is considered exactly as bad as outputting 2.6 when the true answer is 3. This is not always appropriate.

## Calibration Network Architecture.

Our network's input is a feature vector  $\mathbf{x} = [p_{\text{LLM}}(y_i | T, Q_i) : i \in \{0, \dots, 8\}, y_i \in \mathcal{Y}_i]$ . These are already *extremely* high-level text features, extracted by the LLM. We next use a feed-forward neural net to transform  $\mathbf{x}$  into a representation  $\mathbf{z}_2 \in \mathbb{R}^{h_2}$ :

$$\mathbf{z}_1 = \sigma((W_1 + W_1^a)[1; \mathbf{x}]) \in \mathbb{R}^{h_1} \quad (3)$$

$$\mathbf{z}_2 = \sigma((W_2 + W_2^a)[1; \mathbf{z}_1]) \in \mathbb{R}^{h_2} \quad (4)$$

Here  $W_1, W_1^a \in \mathbb{R}^{h_1 \times (1+9)}$  and  $W_2, W_2^a \in \mathbb{R}^{h_2 \times (1+h_1)}$ . The parameters  $W_k$  are shared across all judges while  $W_k^a$  are judge-specific.

The learned representations  $\mathbf{z}_2$  are shared across all questions. For each  $i \in \{0, \dots, 8\}$ , we obtain  $\{\hat{p}_a(y_i | T, Q_i) : y_i \in \mathcal{Y}_i\}$  as a probability vector

$$\text{softmax}((V_i + V_i^a)[1; \mathbf{z}_2]) \in \mathbb{R}^{|\mathcal{Y}_i|} \quad (5)$$

The collection of matrices  $V_i \in \mathbb{R}^{|\mathcal{Y}_i| \times (1+h_2)}$  can be implemented as a 3D tensor  $V$  (padding  $V_i$  with extra rows when  $|\mathcal{Y}_i|$  is small).

**Multi-Task Learning.** Our calibration network performs multi-task learning: each rubric question is a different task. When the accurate prediction of  $y_0^a$  is our *main task*, the other tasks serve only as regularizing *auxiliary tasks*, which help training to discover useful hidden features  $\mathbf{z}_2$ . The weighting of the auxiliary tasks could be dynamically adapted during training (using a validation set), for example with the AuxiNash training algorithm (Shamsian et al., 2023). However, we currently use a simpler, faster shortcut that divides training into two phases. In the **pre-training** phase, we optimize the full log-likelihood objective (1). This learns useful initial representations.<sup>7</sup> In the **fine-tuning** phase, we continue training with a modified objective that sums over only the tuples in  $\mathcal{D}$  with  $i = 0$ . This adjusts the parameters to focus on the main task—predicting responses  $y_0^a$  to  $Q_0$ . In both phases, we use early stopping to avoid overfitting.<sup>8</sup>

<sup>7</sup>However, in contrast to AuxiNash, this shortcut does not try to identify and favor more useful auxiliary tasks. Equation (1) simply weights each question  $Q_i$  in proportion to its number of annotated answers in the training dataset  $\mathcal{D}$ . (In our experiments, all questions are equally represented in  $\mathcal{D}$ .)

<sup>8</sup>We also tried a variant where pre-training was itself divided into two stages and we fixed  $W_k^a = 0$  and  $V_i^a = 0$  during the first stage. This was intended to prevent overfitting of these judge-specific parameters, but we observed no improvement compared to the simpler method.

**Using the Predictions.** Since LLM-RUBRIC can predict any judge’s scores on a new text  $T$ , how should it be used in practice? In Appendix A, we propose approaches to score aggregation, system quality monitoring, and other practical issues.

**Future Extensions.** The idea of a calibration network is quite general and can easily be extended to various types of human and LLM ratings. In Appendix B, we sketch some natural extensions that were not needed in this paper’s experiments.

### 3 Data

Conversational AI systems are now being widely deployed. To test our methods on dialogue evaluation, we invest in developing both synthetic and real datasets of human–AI conversations.

We focus on English information-seeking dialogues in the “IT help” (enterprise information technology) domain (Lowe et al., 2015; Carletta et al., 2005). As in many real world domains, dialogue data here is often proprietary to the system owner and/or private to the user. Acquiring experimental access to live systems for evaluation is even more difficult. Thus, we build and evaluate several LLM-powered dialogue systems, which differ in their ability to search a corpus of websites related to Microsoft Azure<sup>9</sup> help topics.

For *training* data, we generate a corpus of *synthetic* dialogues with simulated users, and have human judges rate them. Collecting these artificial dialogues is efficient, since judges only have to annotate conversations and not interact with the systems first. For our final *test* data, we have our judges actually interact with the live systems as users and then annotate their own dialogues. All of our judges are professional annotators.

To mine the topics for both synthetic and live evaluation, we use real user queries and click data from a large commercial web search engine, which further increases the realism of our experiments.

Below, §3.1 explains how we compile a corpus of background documents and how we select topics to ensure that the generated and collected conversations are diverse and are indeed information-seeking, rather than navigational or transactional. §3.2 and §3.3 explain our approaches to synthetic dialogue generation and real dialogue collection.

<sup>9</sup><https://azure.microsoft.com/>

#### 3.1 Mining Topics for RAG

To simulate or collect diverse information-seeking dialogues, we need to know what information our users will seek. We picked an arbitrary IT help topic, Azure, for which many answers can be found on the subreddit [r/azure](#). We hypothesize that search queries are enterprise information-seeking topics related to Azure if they lead to satisfactory clicks on the Azure subreddit.<sup>10</sup> Using this heuristic to help filter query logs obtained from the Bing search engine, we construct a set  $\mathcal{S}$  of 2275 common English queries about Azure. We will use these as *topics* to prompt the creation of realistic and diverse conversations.

Some of our dialogue systems will condition their responses on relevant documents, as in retrieval-augmented generation (RAG) (Lewis et al., 2020). To build a corpus of potentially relevant documents, we mined and crawled all 37,982 clicked URLs in the web search engine’s results to the queries in  $\mathcal{S}$ . This includes but is not limited to the Azure subreddit URLs. We discard the ones that require login, are behind a paywall, or are no longer available (broken links). To ensure that the URLs are of high quality, we also make sure they exist in Clueweb 2022 Set B (Overwijk et al., 2022) top 200M most popular URLs. After filtering, we arrived at 23,243 unique webpages. We used BeautifulSoup to convert each webpage’s title and body into a plain text document, without any truncation. The mean document length is  $1246 \pm 1651$  words (denoting mean  $\pm$  standard deviation).

#### 3.2 Synthetic Dialogue Generation

To generate synthetic dialogues in English of varying quality, we use 5 different LLM-based approaches (DS1–DS5), described in Appendix F. These approaches have different levels of access to the document corpus. Also, the true topic (which is always provided to the simulated user) is only revealed to the dialogue system in DS1–DS3.

We use gpt-3.5-turbo-16k with its default parameters (OpenAI, 2024) for all of our data generation (§3.2, §3.3) and rubric-based evaluation (§4).

We randomly selected 50 topics, and used each of the systems DS1–DS5 to generate a synthetic conversation on that topic, resulting in 250 unique dialogues of varying quality. Each dialogue was evaluated by 3 judges (randomly assigned from

<sup>10</sup>A satisfactory click in a search engine is defined as a click that leads to a dwell time longer than a given threshold (Jiang and Allan, 2016). Here we use a threshold of 30 seconds.

a pool of 24 judges), resulting in 741 personalized data points for dialogue evaluation after some guardrail quality checks (see Appendix G). The average judge annotated  $31 \pm 13$  dialogues.

### 3.3 Real Dialogue Collection and Evaluation

To obtain more realistic data for evaluation, we collect conversations with DS1–DS3 where the user turns are not generated by the LLM but by a real human. The assistant in these three systems may be summarized as “no RAG” (DS1), “oracle RAG based on the topic” (DS2), and “BM25 RAG based on the topic” (DS3).

The human who plays the user role in the dialogue then also serves as the judge for that dialogue, making them particularly well qualified to judge overall user satisfaction  $Q_0$ . Details about the web interface and instructions to the humans can be found in Appendix H.

We collected a total of 223 evaluated human conversations by having 13 of the original 24 judges converse with systems DS1–DS3 (some judges were no longer available). Each judge engaged in and annotated  $17.2 \pm 3.4$  dialogues on average. The evaluations are summarized in Appendix I.

## 4 Experiments

We will evaluate how well LLM-RUBRIC can predict individual judges’ assessments  $y_0^a$  of our  $Q_0$  (overall user satisfaction). We evaluate predictions  $\hat{y}_0^a$  both in absolute terms (whether they achieve low root-mean-squared error, or RMSE) and in relative terms (how well  $\hat{y}_0^a$  correlates with  $y_0^a$ , i.e., whether  $\hat{y}_0^a$  can be used to rank  $(T, a)$  pairs).

We train our calibration networks on synthetic dialogues. We then evaluate them not only on held-out synthetic dialogues but also on real dialogues, to demonstrate that the LLM scoring and its calibration can generalize from synthetic to real data.

**Hyperparameter Selection.** To train a system on a given training set, we evaluate hyperparameter settings from a grid by 5-fold cross-validation on the training set, and then use the selected hyperparameters to train on the entire training set. We select the hyperparameters that maximize the main task objective, namely the log-likelihood of (held-out) annotations  $y_0^a$ . The hidden layer sizes  $h_1, h_2$  each range over  $\{10, 25, 50, 100\}$ , the batch size ranges over  $\{32, 64, 128, 256\}$ , the learning rate of the Adam optimizer ranges over  $\{0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005,$

$0.01\}$ , and the numbers of epochs for pre-training and fine-tuning each range over  $\{5, 10, 20, 30, 40, 50\}$ .<sup>11</sup>

**Synthetic Data Evaluation.** We test our calibration network on all 741 synthetic dialogues, using 5-fold cross-validation; the dataset is split at the dialogue level so that each dialogue appears in only one fold. Different folds may select different evaluation hyperparameters, resulting in different architectures for the calibration network.<sup>12</sup>

**Real Data Evaluation.** We test our calibration network on all 223 real dialogues, after training on all of the synthetic dialogues (again selecting hyperparameters by 5-fold cross-validation).

**Baseline Methods.** As Table 1 shows, we compare LLM-RUBRIC to these 5 baselines:

1. **Random.** For each dialogue independently, we produce 1, 2, 3, or 4 uniformly at random.
2. **Argmax LLM  $Q_0$ .** We use the top LLM prediction for  $Q_0$ :  $\text{argmax}_{y_0 \in \mathcal{Y}_0} p_{\text{LLM}}(y_0 | T, Q_0)$ . Note that this system always produces an integer.<sup>13</sup>
3. **Expected LLM  $Q_0$ .** We use the expected value of the LLM’s prediction for  $Q_0$ :  $\sum_{y_0 \in \mathcal{Y}_0} y_0 \cdot p_{\text{LLM}}(y_0 | T, Q_0) / Z_0$  (where  $Z_0$  normalizes the probabilities over  $\mathcal{Y}_0$ —see footnote 3).
4. **Calibrated LLM  $Q_0$ .** An ablated version of LLM-RUBRIC that uses only  $Q_0$ , i.e., the feature vector  $\mathbf{x} = [\tilde{p}(y_0 | T, Q_0) | y_0 \in \mathcal{Y}_0]$  is restricted to the  $Q_0$  answer probabilities. We train and evaluate the calibration network just as for LLM-RUBRIC, including cross-validation and hyperparameter selection.
5. **FActScore (Min et al., 2023).** This is a recent retrieval-based automatic evaluator<sup>14</sup> that predicts the percentage of factually correct sentences as the overall evaluation score. We use the Azure corpus described in §3.1 as the retrieval corpus in FActScore, which performs better than the default Wikipedia corpus.

<sup>11</sup>Instead of including the number of epochs in the hyperparameter grid search, an alternative would be to use a standard early stopping heuristic at each phase, by evaluating that phase’s training objective periodically on held-out data.

<sup>12</sup>When training on 4 folds to evaluate the 5th, we select the hyperparameters by an inner 5-fold cross-validation on this training set of about 593 examples, as explained above.

<sup>13</sup>In a pilot experiment, we found no significant improvement from few-shot prompting.

<sup>14</sup><https://github.com/shmsw25/FActScore>

Model	Synthetic Conversations			Real Human-Agent Conversations				
	RMSE ↓	P's $\rho \uparrow$	S's $\rho \uparrow$	K's $\tau \uparrow$	RMSE ↓	P's $\rho \uparrow$	S's $\rho \uparrow$	K's $\tau \uparrow$
1 Random Eval	1.499	0.002	-0.003	-0.003	1.427	0.011	0.006	0.005
2 Argmax LLM $Q_0$	0.984 <sup>1</sup>	0.153 <sup>1</sup>	0.161 <sup>1</sup>	0.147 <sup>1</sup>	1.186 <sup>1</sup>	0.106 <sup>1</sup>	0.123 <sup>1</sup>	0.120 <sup>1</sup>
3 Expected LLM $Q_0$	0.856 <sup>12</sup>	0.182 <sup>1</sup>	0.217 <sup>1</sup>	0.168 <sup>1</sup>	0.901 <sup>12</sup>	0.143 <sup>1</sup>	0.141 <sup>1</sup>	0.138 <sup>1</sup>
4 Calibrated LLM $Q_0$	0.801 <sup>123</sup>	0.198 <sup>12</sup>	0.196 <sup>1</sup>	0.193 <sup>12</sup>	0.784 <sup>123</sup>	0.211 <sup>123</sup>	0.218 <sup>123</sup>	0.192 <sup>123</sup>
5 FActScore (Min et al., 2023)	—	0.204 <sup>12</sup>	0.211 <sup>1</sup>	0.200 <sup>12</sup>	—	0.216 <sup>123</sup>	0.218 <sup>123</sup>	0.207 <sup>123</sup>
6 LLM-RUBRIC	0.396 <sup>1234e</sup>	0.401 <sup>1234e</sup>	0.398 <sup>1234e</sup>	0.393 <sup>1234e</sup>	0.422 <sup>1234</sup>	0.350 <sup>12345</sup>	0.347 <sup>12345</sup>	0.331 <sup>12345</sup>
a Oracle	0.237 <sup>bcd</sup>	0.611 <sup>bcd</sup>	0.626 <sup>bcd</sup>	0.605 <sup>bcd</sup>	0.289 <sup>bcd</sup>	0.717 <sup>bcd</sup>	0.711 <sup>bcd</sup>	0.675 <sup>bcd</sup>
b w/o LLM probs	0.276 <sup>*cef</sup>	0.551 <sup>*cef</sup>	0.548 <sup>*cef</sup>	0.533 <sup>*cef</sup>	0.357 <sup>*c</sup>	0.625 <sup>*c</sup>	0.629 <sup>*c</sup>	0.599 <sup>*c</sup>
c w/o Personalized Calibration	0.401 <sup>e</sup>	0.476 <sup>*e</sup>	0.471 <sup>*e</sup>	0.468 <sup>*e</sup>	0.389 <sup>*</sup>	0.582 <sup>*</sup>	0.587 <sup>*</sup>	0.565 <sup>*</sup>
d ↳ + Personalized isotonic regress	0.273 <sup>*cef</sup>	0.521 <sup>*cef</sup>	0.526 <sup>*cef</sup>	0.519 <sup>*cef</sup>	0.302 <sup>*bc</sup>	0.650 <sup>*bc</sup>	0.653 <sup>*bc</sup>	0.644 <sup>*bc</sup>
e Depersonalized Oracle	0.492	0.362	0.355	0.338	—	—	—	—
f ↳ + Personalized isotonic regress	0.321 <sup>*ce</sup>	0.482 <sup>*e</sup>	0.485 <sup>*e</sup>	0.477 <sup>*e</sup>	—	—	—	—

Table 1: Performance on predicting human judges’  $Q_0$  (overall quality). We report root mean squared error (RMSE) and, more important, correlations with human judges’ responses (Pearson’s  $\rho$ , Spearman’s  $\rho$ , Kendall’s  $\tau$ ). Results on the synthetic conversation dataset are based on 5-fold cross-evaluation; results on the real conversations are based on training on all synthetic conversations. The superscripts denote statistically significant improvements according to a paired permutation significance test ( $p < 0.05$ ). The asterisk \* means all methods in rows 1–6.

**Oracle Methods.** Table 1 also shows upper bounds on performance. The **Oracle** system is the same as LLM-RUBRIC, but the calibration network’s input  $x$ —at both training and test time—includes the judge’s *actual* response to each question  $Q_i$  (except for  $Q_0$ , which we aim to predict!) as a four-dimensional one-hot vector, in addition to the LLM response vector  $p_{\text{LLM}}(y_0 | T, Q_i)$ .

We ablate different components of the Oracle model by withholding the LLM response vector from the model input and by depersonalizing the calibration network (**Oracle w/o Personalized Calibration**) by dropping  $W_k^a$ . To restore a judge  $a$ ’s idiosyncratic distribution of  $Q_0$  scores (Figure 2), without restoring their idiosyncratic computation of  $Q_0$  from other dimensions, we try correcting the output of the depersonalized calibration network using an  $a$ -specific isotonic regression model.

Our **Depersonalized Oracle** is similar to the Oracle, but instead of using the responses of the actual target judge  $a$ , it uses the distribution of responses of all other judges (averaging their one-hot vectors), holding out the target judge.<sup>15</sup> It also drops the personalized weights  $W_k^a$ .

Thus, the Oracle provides a rough upper bound on LLM-RUBRIC. The Depersonalized Oracle provides a rough upper bound on a version of LLM-RUBRIC that produces  $a$ -independent results.

## 5 Results

A trivial baseline of predicting a *constant*  $Q_0$  (the overall mean from training data) achieves an RMSE of 0.82 on both synthetic and real conversations. LLM-RUBRIC roughly halves this (row 6 of Ta-

<sup>15</sup>We cannot run this on the real conversation dataset, where each dialogue is annotated only by a single judge.

ble 1), so it explains  $\approx \frac{3}{4}$  of the variance in human judgments of  $Q_0$  across judges  $a$  and texts  $T$ . Its predictions of  $Q_0$  have tolerably low error and correlate reasonably well with those of human judges.

In sharp contrast, the LLM’s direct response to  $Q_0$  (row 2 or 3) does *worse* than the constant baseline. Even calibrating its response distribution for each judge (row 4) barely improves on the baseline, explaining only 5–10% of the variance in human judgments and achieving only  $\approx 0.2$  correlation with them. This suggests that the LLM cannot help assess  $Q_0$  (user satisfaction) until we ask it about the finer-grained dimensions  $Q_1$ – $Q_8$ .

The results obtained by FActScore (row 5) do not correlate any better with overall satisfaction, so percentage of factually correct sentences is also not a good indicator of overall user satisfaction. Moreover, Liu et al. (2016) showed that dialogue systems were poorly evaluated by simple metrics of lexical overlap with human responses.

## 6 Analysis

**Calibration.** Does our trained LLM-RUBRIC produce well-calibrated probability distributions for  $Q_0$  (as one would expect from maximum-likelihood training)? We checked on synthetic data. It obtained excellent smECE values of  $< 0.05$  for each  $y_0 \in \mathcal{Y}_0 = \{1, 2, 3, 4\}$ , where smECE is the smoothed expected calibration error (Błasiok and Nakkiran, 2023). Informally, this means that for each  $y_0 \in \mathcal{Y}_0$ , when we examine the held-out examples  $(T, 0, a, y_0^a)$  with  $\hat{p}_a(y_0 | T, Q_0) \approx p$ , the fraction where  $y_0^a = y_0$  was in fact  $\approx p$ . Appendix K shows calibration plots and discusses how to use calibrated probabilities for downstream decisions.

Model	RMSE ↓	P's $\rho$ ↑
LLM-RUBRIC	0.422	0.350
w/o fine-tuning	0.493 <sup>▽</sup>	0.249 <sup>▽</sup>
w/o pre-training	0.525 <sup>▽</sup>	0.226 <sup>▽</sup>
w/o personalization	0.601 <sup>▽</sup>	0.198 <sup>▽</sup>
w/o $Q_0$ (Satisfaction)	0.554 <sup>▽</sup>	0.287 <sup>▽</sup>
w/o $Q_1$ (Naturalness)	0.463 <sup>▽</sup>	0.313 <sup>▽</sup>
w/o $Q_2$ (Grounding Sources)	0.471 <sup>▽</sup>	0.279 <sup>▽</sup>
w/o $Q_3$ (Citation Presence)	0.573 <sup>▽</sup>	0.075 <sup>▽</sup>
w/o $Q_4$ (Citation Suitability)	0.497 <sup>▽</sup>	0.311 <sup>▽</sup>
w/o $Q_5$ (Citation Optimality)	0.506 <sup>▽</sup>	0.192 <sup>▽</sup>
w/o $Q_6$ (Redundancy)	0.424	0.348
w/o $Q_7$ (Conciseness)	0.532 <sup>▽</sup>	0.254 <sup>▽</sup>
w/o $Q_8$ (Efficiency)	0.510 <sup>▽</sup>	0.161 <sup>▽</sup>

Table 2: Predicting  $Q_0$ : Ablation study on real conversation data for each design decision in our calibration network (top) and each rubric dimension (bottom). <sup>▽</sup> denotes a statistically significant performance drop from the full LLM-RUBRIC ( $p < 0.05$ ).

Expected LLM $Q_i$	LLM-RUBRIC	
	RMSE ↓	P's $\rho$ ↑
$Q_0$	0.901	0.143
$Q_1$	1.033	0.177
$Q_2$	0.799	0.140
$Q_3$	0.796	0.347
$Q_4$	0.919	0.166
$Q_5$	1.104	0.191
$Q_6$	1.726	0.030
$Q_7$	1.240	0.057
$Q_8$	0.981	0.059

Table 3: How well can LLM-RUBRIC predict the response  $y_i$  to question  $Q_i$ ? For each row, we fine-tune LLM-RUBRIC on the target rubric dimension and compare to Expected LLM  $Q_i$  on the real conversation data. Superscript \* indicates statistically significant improvement with 95% confidence ( $p < 0.05$ ).

**Ablation Studies.** §5 showed that LLM responses on 8 additional questions were useful, but was our calibration network the best way to incorporate them into our prediction of  $Q_0$ ? To justify each design decision, we try omitting pre-training, fine-tuning, and personalized weighting from our calibration network. The results on the real conversation data in Table 2 show that predictions were improved by each step. In particular, it was indeed useful to do multi-task pre-training of the calibration network (which required human judgments on all questions) and to then fine-tune on the main task. Personalized weighting had the greatest impact.

Also, were all 8 questions useful? We measured

the impact of each question by omitting it from the evaluation rubric for the LLM-RUBRIC model (bottom half of Table 2). All rubric dimensions contributed significantly to the  $Q_0$  prediction, except for  $Q_6$ , which focuses on redundancy in the dialogue. Using even more rubric dimensions might improve performance further (footnote 2 and Appendix B). That said, considering more rubric dimensions would mean more human annotations at pre-training time and/or more LLM computation.

**Oracle study.** Giving LLM-RUBRIC access to a judge’s true responses to  $Q_1$ – $Q_8$  lets us see how well the judge’s overall quality score  $Q_0$  is predictable from our particular rubric dimensions. This gets rather better results, including an excellent 0.72 Pearson’s  $\rho$  correlation between predicted and actual satisfaction scores on real dialogues (row ‘a’ in Table 1). Almost all of this performance can be obtained from *only* the judge’s responses, without access to the  $p_{LLM}$  score distributions (row ‘b’).

This suggests a future strategy (discussed below) of improving the *input* to LLM-RUBRIC by getting the LLM to better predict the judge-specific human rubric responses that were available to the Oracle (row ‘a’), or at least judge-independent versions (rows ‘e’–‘f’). Once such responses are available, the ensembling is still best done by a calibration network that understands an individual judge’s preferences—though under oracle conditions and with our population of judges, dropping that personalization would not be dramatically worse (row ‘c’), and a fraction of the difference can be made up simply by adjusting the predicted scores  $\hat{y}_0$  with personalized isotonic regression (row ‘d’).

**On which dimensions do zero-shot LLMs need improvement?** Table 3 shows these results. Redundancy ( $Q_6$ ), Conciseness ( $Q_7$ ), and Efficiency ( $Q_8$ ) were especially difficult for the LLM to predict—it showed close to zero correlation with human annotators. LLM-RUBRIC much better predicted these scores, as well as overall Satisfaction  $Q_0$ , by exploiting the full response vector  $\mathbf{x}$ : e.g., it improved RMSE by  $> 0.5$  in all of these cases.

The LLM’s errors on a difficult question  $Q_i$  could potentially be reduced through prompt engineering, few-shot prompting, fine-tuning the LLM, or calling a larger LLM. Is that worth it? To assess the potential improvement to  $Q_0$  prediction from better answering  $Q_i$ , one could use cross-validation to evaluate the benefit to  $Q_0$  from replacing just  $Q_i$  with oracle scores before training LLM-RUBRIC.

**How much human judge data is needed to train calibration?** See Appendix J for learning curves.

## 7 Related Work

**LLM Evaluation** Zero-shot or few-shot LLM evaluators have been shown to have higher agreement with human annotators than traditional lexical overlap or even earlier transformer embedding models, across a variety of natural language generation (NLG) tasks (Fu et al., 2023; Lin et al., 2024). Furthermore, when compared to crowd-workers, LLMs can have higher agreement with expert annotators (Gilardi et al., 2023; Chiang and Lee, 2023). Additional techniques like chain-of-thought prompting and auto-prompt engineering can also further improve alignment with human ground truth (Liu et al., 2023a,b; Lin et al., 2024). It seems that LLMs are capable of measuring an increasing range of evaluation dimensions including factuality (Min et al., 2023; Gekhman et al., 2023; Yue et al., 2023), interpretability (Lu et al., 2023), and relevance (Saad-Falcon et al., 2023). These works generally focus on average judge preferences on individual evaluation attributes, while we focus on using LLMs to capture the interplay of individual attributes to better predict all judgments (particularly of overall text quality) for a given judge.

**Calibration of LLM evaluators.** Zhao et al. (2023) develop a Pareto-optimal method for estimating the error rate of an LLM-based predictor by combining both LLM and heuristic predictions, which can in turn be used to correct the initial LLM prediction. While they similarly take advantage of an ensemble of predictors, they assume specific ground-truth answers, whereas LLM-RUBRIC produces distributions over reasonable answers.

**Subjectivity in Evaluation.** While LLMs can agree with expert judges, in cases where experts have low agreement, LLMs tend to have low agreement with the judges as well (Chiang and Lee, 2023). It is increasingly acknowledged that accounting for subjectivity (as opposed to collapsing or removing disagreements) in NLP evaluation is a key part of evaluation design (Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Uma et al., 2021a,b; Plank, 2022; Plepi et al., 2022; Sandri et al., 2023). By training a single network to model all judges, we take the view that “disagreement is not noise but signal” (Aroyo and Welty, 2015). Baan et al. (2022) put it more starkly: without

modeling the judge distribution, metric calibration is itself nonsensical on subjective tasks. Making downstream use of these disagreeing judges—or rather LLM-RUBRIC’s simulation of them on new texts—is discussed by Appendix A, Gant et al. (2020), and Uma et al. (2021b).

While our work is similar conceptually to Gant et al. (2020) in that we include judge-specific parameters to predict each human judge’s responses, we show that this can be further improved by predicting responses to multiple questions (our auxiliary tasks  $Q_1$ – $Q_8$  along with our main task  $Q_0$ ).

Xiao et al. (2023) analyze common NLG evaluation dimensions and metrics using the concepts of reliability and validity from measurement theory. They find that while manual judges may rate generated texts by different dimensions like ‘coherence’ or ‘relevance,’ these dimensions can exhibit poor validity structure. In their case, this means that they find that an individual judge’s correlation with their own ratings across coherence and relevance can be as high or higher than correlation between other judges within each dimension, supporting the idea individual judges may have idiosyncratic or conflated mappings of different evaluation criteria. Xiao et al. (2023) suggest several ways to improve the dimensions to account for this. We did not perform a similar analysis on our judges and rubric dimensions, although improvements here would be orthogonal to the benefits of LLM-RUBRIC, since judges may reasonably disagree even in the absence of validity structure issues.

## 8 Conclusions

This work proposed LLM-RUBRIC—a rubric-based framework for automatic evaluation of text. We trained and tested it on novel datasets of information-seeking dialogues. LLM-RUBRIC performs multidimensional evaluation using a black-box LLM, then aggregates and calibrates these multidimensional responses for each human judge.

Although the LLM’s raw responses do not highly correlate with human judgments in such a complex task, we found that combining its response distributions on all questions can predict each human judge’s responses, including overall satisfaction. We obtained substantial improvements on RMSE and on both linear and rank-based correlation metrics, on held-out synthetic conversations (development data) and real ones (test data). Below, we discuss limitations, ethics, uses, and extensions.

## Acknowledgments

We thank Val Ramirez and the data specialists who contributed to the creation of this work.

## Limitations

**Robustness.** In general, one might hope that the trained LLM-RUBRIC can successfully predict human scores even in *new* test domains—at least when it is given a broadly competent LLM, a broadly worded rubric, and training examples that exhibit a variety of score profiles on that rubric. However, we did not evaluate this, other than showing that our trained LLM-RUBRIC worked well when applied to a *slightly* different test distribution (real rather than synthetic dialogues) on the same topics (information-seeking Azure queries).

Robustness is particularly important when privacy rules prevent having human judges evaluate real examples from the test distribution, as in some deployed dialogue systems or when coding medically or legally sensitive data. Even when training examples can be drawn from the true test distribution, it may be hard to find judges who are competent to annotate the full range of topics and styles in such examples. For example, judges may be unavailable for low-resource languages—and it is not necessarily true that LLM scores bear the same relation to human scores for texts in those languages, since the LLM may be less competent to judge such texts (Ahuja et al., 2024), or the texts themselves may have different quality issues.<sup>16</sup>

Robustness is also needed when the test distribution shifts over time—either for exogenous reasons such as new topics or user populations, or because the metric has become a target (Goodhart’s Law) so that the texts are increasingly designed to score well on predicted  $Q_0$ . The latter case includes NLG engineering, as well as adversarial settings like essay grading or legal discovery, where test-takers or email conspirators have an incentive to write their texts so as to fool the evaluation system.

**Efficiency.** We used a large pretrained LLM to answer each rubric question. It would be cheaper to use smaller models where possible, perhaps fine-tuned on specific questions. One could also decide

<sup>16</sup>For example, when a multilingual dialogue system is used in a low-resource language, user satisfaction  $Q_0$  may be lower because of language-specific problems such as formality that did not arise in LLM-RUBRIC’s training, or were not as highly weighted, or were not directly assessed by the rubric at all.

which questions are worth asking (and which models to ask) by using an *adaptive* rubric: e.g., choose the next evaluation question to maximize the expected information gain, and stop at the point of diminishing returns, so that it is not necessary to ask all questions. An adaptive rubric could in principle be quite large, with only a small portion of it used on any particular text  $T$ . This direction and other possible extensions are discussed in Appendix B, but we did not try them.

**Downstream Evaluation.** Although we measured overall correlation between predicted and human scores on each rubric question, we did not evaluate the usefulness of our predicted scores for difficult downstream tasks such as choosing among similar candidate answers or dialogue systems. More rubric questions might be needed for sufficiently accurate evaluation (see footnote 2 and Appendix B).

A particularly challenging but important downstream use is to *improve* natural language generation. We have not addressed this. However, a metric such as our predicted overall quality  $\hat{y}_0$  (averaged over a set of judges as in Appendix A) could be used as a reward signal, for example to improve an LLM by proximal policy optimization (Schulman et al., 2017). More ambitiously, one could train the LLM using multi-objective reinforcement learning (e.g., Yang et al., 2019; Abels et al., 2019; Ramé et al., 2023; Wu et al., 2023) to consider idiosyncratic preferences *at runtime* and generate text that achieves a high predicted *user-specific* reward. For example, one could use  $\hat{y}_0^a$  as the runtime reward function if one modified our calibration network to do regression (§2) via  $\hat{y}_i^a = (\mathbf{v}_i + \mathbf{v}_0^a) \cdot [1; \mathbf{z}_2]$  where  $\mathbf{z}_2$  is judge-independent (compare equation (5)). Then  $\mathbf{z}_2$  serves as a multi-objective reward vector, and  $\mathbf{v}_0 + \mathbf{v}_0^a$  is the preference weighting that linearly scalarizes this reward at runtime, where  $\mathbf{v}_0^a$  may be regarded as a preference embedding of the user  $a$  (possibly computed from features of  $a$ ).

**Fine-Grained Evaluation.** We only considered evaluating entire texts. However, humans often perform finer-grained evaluation tasks—such as highlighting *problematic* spans in human- or machine-written text (e.g., to provide feedback and opportunities for revision), or highlighting *relevant* spans (e.g., to call a human or machine’s attention to them). We have not presented methods for automating or calibrating fine-grained evaluation.

## Ethics Statement

**Beyond User Satisfaction.** Evaluation metrics drive engineering and so have real-world consequences. Our experiments focused on predicting overall user satisfaction (our choice of  $Q_0$ ), but we do not in fact recommend this as the actual goal of dialogue system development. In practice, quality evaluation of a dialogue agent should also assess potential harms to the user (e.g., false information), to the dialogue system owner (e.g., reputational harm through violating policies on content or style), and to third parties (e.g., encouraging violence or outputting private or copyrighted information).

**Fairness Auditing.** Our aligned LLM’s ability to approximately match human judges does not answer the question of whether the unaligned LLM, the aligned LLM, the human judges, or the manually constructed rubrics are fair or unbiased. Even when our system does achieve low total error at matching fair judgments, it is not guaranteed that its errors or their downstream harms are evenly distributed. Thus, accuracy (Table 1), calibration (Appendix K), and rubric validity should be checked for various important subsets of the data. For example, in essay grading, does the calibrated LLM systematically underestimate the quality of the ideas of speakers of a particular dialect? In dialogue system evaluation, is a particular user population frustrated with a certain kind of error that they experience heavily, yet this type of error is underdiagnosed?<sup>17</sup>

**Human Data.** LLM-RUBRIC requires collecting data from human judges that reveal their personal preferences, such as their affinity for specific textual passages. Such data should always be carefully safeguarded. In certain cases it may even be appropriate to train the calibration network using differential privacy, to make it impossible to

<sup>17</sup>Or, going beyond auditing, one could try to learn a *multicalibrated* model in the first place (Hébert-Johnson et al., 2018). Such a model’s *average* rating over a subset of texts  $S$  will be approximately correct, for every  $S$  in a given large family of subsets that are computationally identifiable and not too small. This ensures that the errors are in a sense fairly distributed: the model cannot systematically underestimate or overestimate texts written by any particular subpopulation of authors, preferred by particular judges, having particular linguistic features, etc. Typically, a multicalibration algorithm builds up a complex model (without sacrificing accuracy): each step augments the current model with a learned post-correction step that adjusts the outputs on some subset of inputs. Such algorithms exist for regression (e.g., Globus-Harris et al., 2023) as well as classification, and have recently been applied to LLM evaluation (Detommaso et al., 2024).

guess information about particular judges from the network weights.

**Harmful Uses.** LLM-RUBRIC may enable generating or choosing content that appeals to a specific human’s preferences. This could improve their satisfaction with the NLG output, but it could also be used to optimize for their engagement—even when this is harmful (for example, confirming biases, spreading misinformation, provoking outrage, swindling, or recommending antisocial actions or self-harm).

**Environmental Costs.** LLM-RUBRIC is compute-intensive, as it involves calling an LLM several times for each NLG output. On a small evaluation dataset, the compute cost may be modest, but LLM-RUBRIC will add to the environmental footprint of a system if it is applied to a substantial fraction of user traffic, or is called many times during a hyperparameter tuning loop or to compute the reward signal for reinforcement learning. Costs might be reduced through distillation or an adaptive rubric, as discussed in the Limitations section.

## References

- Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. 2019. [Dynamic weights in multi-objective deep reinforcement learning](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 11–20.
- Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. [MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks](#). In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. [Natural language model for automatic identification of intimate partner violence reports from Twitter](#). *Array*, 15.
- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Magazine*, 36(1):15–24.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United

- Arab Emirates. Association for Computational Linguistics.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. **Constitutional AI: Harmlessness from AI feedback**. *Computing Research Repository*, arXiv:2212.08073.
- Dwight Barry. 2017. **Do not use averages with Likert scale data**. Online monograph.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. **We need to consider disagreement in evaluation**. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners**. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Jarosław Błasik and Preetum Nakkiran. 2023. **Smooth ece: Principled reliability diagrams via kernel smoothing**. *Computing Research Repository (CoRR)*, arXiv:2309.12236.
- Jean Carletta, Simone Ashby, Sébastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaikos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnieszka Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. **The AMI meeting corpus: A pre-announcement**. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction, MLMI'05*, page 28–39, Berlin, Heidelberg. Springer-Verlag.
- Stevie Chancellor and Munmun De Choudhury. 2020. **Methods in predictive techniques for mental health status on social media: A critical review**. *NPJ Digital Medicine*, 3(1).
- Laurent Charlin and Richard S. Zemel. 2013. **The Toronto Paper Matching System: An automated paper-reviewer assignment system**. In *Proceedings of the ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. **Can large language models be an alternative to human evaluations?** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Eun Cheol Choi and Emilio Ferrara. 2024. **FACT-GPT: Fact-checking augmentation via claim matching with LLMs**. *Computing Research Repository (CoRR)*, arXiv:2402.05904.
- Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. 2023. **Learning to maximize mutual information for dynamic feature selection**. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6424–6447.
- Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. 2024. **Multicalibration for confidence scoring in LLMs**. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 10624–10641.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. **BERT: Pre-training of deep bidirectional transformers for language understanding**. *Computing Research Repository*, arXiv:1810.04805.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. **GPTScore: Evaluate as you desire**. *arXiv preprint arXiv:2302.04166*.
- Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Nataraajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. **The capability of large language models to measure psychiatric functioning**. *Computing Research Repository (CoRR)*, arXiv:2308.01834.
- William Gant, Lelia Glass, and Aaron Steven White. 2022. **Decomposing and recomposing event structure**. *Transactions of the Association for Computational Linguistics*, 10:17–34.
- William Gant, Benjamin Kane, and Aaron Steven White. 2020. **Natural language inference with mixed effects**. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 81–87, Barcelona, Spain (Online). Association for Computational Linguistics.
- Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. **TrueTeacher: Learning factual consistency evaluation with large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. **ChatGPT outperforms crowd workers for text-annotation tasks**. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

- Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. 2023. **Multicalibration as boosting for regression**. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492.
- Yury Gorishniy, Ivan Rubachev, and Artem Babenko. 2022. **On embeddings for numerical features in tabular deep learning**. In *Advances in Neural Information Processing Systems*, volume 35, pages 24991–25004.
- Donna K. Harman. 1996. **Overview of the Fourth Text REtrieval Conference (TREC-4)**. Special Publication (NIST SP) 500-236, National Institute of Standards and Technology, Gaithersburg, Maryland.
- He He, Hal Daumé III, and Jason Eisner. 2012. **Cost-sensitive dynamic feature selection**. In *ICML Workshop on Inferning: Interactions between Inference and Learning*, Edinburgh. 6 pages.
- Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. **Multicalibration: Calibration for the (computationally-identifiable) masses**. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948.
- Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. **Human feedback is not gold standard**. *ArXiv*, abs/2309.16349.
- Andy S. Huang, Kyle Hirabayashi, Laura Barna, Deep Parikh, and Louis R. Pasquale. 2024. **Assessment of a Large Language Model’s Responses to Questions and Cases About Glaucoma and Retina Management**. *JAMA Ophthalmology*, 142(4):371–375.
- Jiepu Jiang and James Allan. 2016. **Reducing click and skip errors in search result ranking**. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 183–192. ACM.
- Mohammad Kachuee, Sajad Darabi, Babak Moatamed, and Majid Sarrafzadeh. 2019. **Dynamic feature acquisition using denoising autoencoders**. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2252–2262.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. **CTRL: A conditional transformer language model for controllable generation**. *Computing Research Repository*, arXiv:1909.05858.
- Diederik P. Kingma and Max Welling. 2019. **An introduction to variational autoencoders**. *Foundations and Trends in Machine Learning*, 12(4):307–392.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. **Retrieval-augmented generation for knowledge-intensive NLP tasks**. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. **AutoConv: Automatically generating information-seeking conversations with large language models**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.
- Ying-Chun Lin, Jennifer Neville, Jack W. Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. **Interpretable user satisfaction estimation for conversational systems with large language models**. *arXiv preprint arXiv:2403.12388*.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. **How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. **G-Eval: NLG evaluation using GPT-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023b. **Calibrating LLM-based evaluator**. *ArXiv*, abs/2309.13308.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. **The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems**. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.
- Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. **Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT**. *ArXiv*, abs/2303.13809.
- Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. **Do AIs know what the most important issue is? using language models to code open-text social survey responses at scale**. *Research & Politics*, 11(1).

- Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. [Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions](#). *Computers and Education: Artificial Intelligence*, 6:100199.
- Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Alexandru Niculescu-Mizil and Rich Caruana. 2005. [Predicting good probabilities with supervised learning](#). In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*.
- OpenAI. 2024. OpenAI GPT-3.5 Turbo 16K [gpt-3.5-turbo-16k-0613]. Available at: <https://platform.openai.com/docs/models/gpt-3-5-turbo>.
- Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. [ClueWeb22: 10 billion web documents with rich information](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 3360–3362, New York, NY, USA. Association for Computing Machinery.
- E. B. Page. 1968. [The use of the computer in analyzing student essays](#). *International Review of Education*, 14(3):253–263.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. [Unifying data perspectivism and personalization: An application to social norms](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mike Quartararo, Matt Poplawski, Adam Strayer, et al. 2019. [Technology Assisted Review \(TAR\) guidelines](#). Technical report, Bolch Judicial Institute of Duke Law School.
- Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. [Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Dadi Ramesh and Suresh Kumar Sanampudi. 2022. [An automated essay scoring systems: A systematic literature review](#). *Artificial Intelligence Review*, 55(3):2495–2527.
- S. T. Roweis and L. K. Saul. 2000. [Nonlinear dimensionality reduction by locally linear embedding](#). *Science*, 290(5500):2323–2326.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. [ARES: An automated evaluation framework for retrieval-augmented generation systems](#).
- Marta Sandri, Elisa Leonardielli, Sara Tonelli, and Elisabetta Jezek. 2023. [Why don't you do it right? analysing annotators' disagreement in subjective tasks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.
- Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2023. [First tragedy, then parse: History repeats itself in the new era of large language models](#). *ArXiv*, abs/2311.05020.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *ArXiv*, abs/1707.06347.
- Burr Settles. 2012. [Active Learning](#). Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer.
- Aviv Shamsian, Aviv Navon, Neta Glazer, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. 2023. [Auxiliary learning as an asymmetric bargaining game](#). In *Proceedings of the 40th International Conference on Machine Learning*.
- Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. [Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.
- Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2019. [Controllable neural story plot generation via reward shaping](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5982–5988.
- J. B. Tenenbaum, V. D. Silva, and J. C. Langford. 2000. [A global geometric framework for nonlinear dimensionality reduction](#). *Science*, 290(5500):2319–2323.

- Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. [Large language models can accurately predict searcher preferences](#). In *2024 International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. [SemEval-2021 task 12: Learning with disagreements](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. [Learning from disagreement: A survey](#). *J. Artificial Intelligence Research*, 72:1385–1470.
- Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. 2016. [Neural autoregressive distribution estimation](#). *Journal of Machine Learning Research*, 17(1):7184–7220.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. [Human evaluation of automatically generated text: Current trends and best practice guidelines](#). *Computer Speech & Language*, 67:101151.
- Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. [Large language models enable few-shot clustering](#). *Computing Research Repository*, arXiv:2307.00524.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. [Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration](#). In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Nathaniel Weir, Kate Sanders, Orion Weller, Shreya Sharma, Dongwei Jiang, Zhengping Jiang, Bhavana Dalvi Mishra, Oyvind Tafjord, Peter Jansen, Peter Clark, and Benjamin Van Durme. 2024. [Enhancing systematic decompositional natural language inference using informal logic](#). *Computing Research Repository (CoRR)*, arXiv:2402.14798.
- Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. [Fine-grained human feedback gives better rewards for language model training](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. [Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghazemi, Anind K. Dey, and Dakuo Wang. 2024. [Mental-LLM: Leveraging large language models for mental health prediction via online text data](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 8(1).
- Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. [A generalized algorithm for multi-objective reinforcement learning and policy adaptation](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14636–14647.
- Hiyori Yoshikawa, Tomoya Iwakura, Kimi Kaneko, Hiroaki Yoshida, Yasutaka Kumano, Kazutaka Shimada, Rafal Rzepka, and Patrycja Swieczkowska. 2021. [Tell me what you read: Automatic expertise-based annotator assignment for text annotation in expert domains](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 1575–1585.
- Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. [Automatic evaluation of attribution by large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.
- Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. [Conversational information seeking](#). *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.
- Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. [ClusterLLM: Large language models as a guide for text clustering](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore.
- Theodore Zhao, Mu Wei, J. Samuel Preston, and Hoifung Poon. 2023. [Automatic calibration and error correction for large language models via Pareto optimal self-supervision](#). *CoRR*, abs/2306.16564.
- Ruiqi Zhong, Charlie Snell, Dan Klein, and Jason Eisner. 2023. [Non-programmers can label programs indirectly via active examples: A case study with text-to-SQL](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5126–5152.
- Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. [Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

## A Aggregating Predicted Scores

Our use of judge-specific distributions  $\hat{p}_a$  can be regarded principally as a technique to improve training on human annotations. Judges are heterogeneous (Figure 2) and a training document will only be judged by some of them (§3), as discussed in Appendix B below. Knowing who the judges were can help us model the training data. For example, suppose  $T$  got mostly low LLM scores, yet all judges randomly assigned to  $T$  in training data gave it a high overall score. The model might “explain away” the high scores if it knows that those particular judges are generous or are focused on dimensions where  $T$  did well—and thus could still predict low overall scores from the remaining judges.

However, this means that our trained calibration network does not produce ground truth. It only models the idiosyncrasies of individual judges  $a$  (successfully, as shown in Figure 2 and Table 1). We do not even suggest that purely objective scores exist (see §7), except on extremely precise rubric questions. So which judge should we use in the end? That is, after training LLM-RUBRIC, how should we practically obtain a final assessment of a new text  $T$ ?

We might use the mean predicted overall quality,  $\hat{y}_0$ , where  $\hat{y}_i = \text{mean}_{a \in \mathcal{A}} \hat{y}_i^a$  for a fixed set  $\mathcal{A}$  of *trusted judges*.<sup>18</sup> This assumes that  $Q_0$  calls for numerical responses on an interval scale (see footnote 6), so that the mean is defined and meaningful. An unweighted mean also assumes that we equally want to please all judges in  $\mathcal{A}$  (see the start of §2). The benefit of LLM-RUBRIC is that we do not actually query these judges—we predict how each of them *would* respond by querying an LLM and calibrating *its* response distributions.

What makes a judge “trusted”? The judges in  $\mathcal{A}$  might have had additional training, insight, information, or time. For example, Thomas et al. (2024) distinguish between trained assessors and third-party crowdworkers. If LLM-RUBRIC scores are used to nominate interesting documents for more careful manual review, for example in a legal document review workflow, then  $\mathcal{A}$  might consist of the experienced lawyers or paralegals who perform the manual review (and who will continue to add to the training set by answering at least  $Q_0$  on newly nominated documents). Alternatively, a trusted judge, rather than being a single human, might correspond to the result of a discussion and reconciliation process among multiple untrusted human judges.

The various applications in §1 might call for other ways to aggregate the predicted judgments (or the resulting document rankings). E.g., to be safe, lawyers may want to replace mean with max in the definition of  $\hat{y}_0$  to review any document that at least one judge in  $\mathcal{A}$  would have deemed relevant. The predicted judgments can also be used *without* aggregation (Uma et al., 2021b; Plank, 2022; Gant et al., 2022) to train or evaluate other systems for generating or scoring text.

**Dashboards.** In our setting of dialogue evaluation (or NLG evaluation), the mean predicted score  $\hat{y}_0$  for a given text  $T$  can be used as a target metric for system development and monitoring.

To aid system developers, we can go beyond  $\hat{y}_0$  and compute  $\hat{y}_i$  on  $T$  for *each*  $Q_i$  (using a version of the network that has been re-fine-tuned to predict  $\hat{y}_i^a$  as its main task). We can also quantify the importance of improving  $T$  to raise its mean human  $Q_i$  rating:  $\frac{\nabla_{\mathbf{x}} \hat{y}_0 \cdot \nabla_{\mathbf{x}} \hat{y}_i}{\nabla_{\mathbf{x}} \hat{y}_i \cdot \nabla_{\mathbf{x}} \hat{y}_i}$  estimates the improvement in the prediction  $\hat{y}_0$  per unit of improvement in the prediction  $\hat{y}_i$ , if one could change  $T$  so as to change  $\mathbf{x}$  in the direction of steepest ascent of  $\hat{y}_i$ .<sup>19</sup>

A dashboard for the system developers could show how all of the above quantities are distributed over a representative set of texts  $\mathcal{T}$ , using kernel density estimation (or a histogram). The dashboard could also *display* these distributions for different subsets of  $\mathcal{T}$  representing specific topics or groups of users, could *compare* them across different versions of the system, and could *track their means or quantiles* over time. Uncertainty bands around each density curve can be found by computing it many times, each time

<sup>18</sup> Any judges not in  $\mathcal{A}$  still help regularize the training. They might be omitted during fine-tuning (just as  $Q_i$  was for  $i \neq 0$ ).

<sup>19</sup> Of course, it will not usually be possible to change  $T$  in quite this way: the desired direction  $\nabla_{\mathbf{x}} \hat{y}_i$  may send  $\mathbf{x}$  out of the feasible space of texts. Thus, a more sophisticated approach is to estimate the manifold of plausible  $\mathbf{x}$  vectors from known training texts (including desirable texts), so that each  $\mathbf{x}$  can be represented in terms of underlying manifold coordinates  $\mathbf{w}$  and residuals. Now  $\nabla_{\mathbf{x}}$  may be replaced with  $\nabla_{\mathbf{w}}$  throughout. This constrains the steepest-ascent direction to point along the manifold. The manifold may be estimated with methods such as Isomap (Tenenbaum et al., 2000), LLE (Roweis and Saul, 2000), or VAE (Kingma and Welling, 2019). Less ambitiously, one could merely represent the  $p_{\text{LLM}}$  distributions within  $\mathbf{x}$  using softmax parameters  $\mathbf{w}$ , so that steepest-ascent using  $\nabla_{\mathbf{w}}$  will at least constrain these distributions to the probability simplex.

substituting bootstrap replicates of  $\mathcal{A}$  and  $\mathcal{T}$  and—in the case of the density of  $\hat{y}_i^a$ —replacing each  $\hat{y}_i^a$  for each text  $T$  with a sample from  $\hat{p}_a(y_i | T, Q_i)$ .<sup>20</sup> Thus, small  $\mathcal{A}$ , small  $\mathcal{T}$ , and high-variance distributions  $\hat{p}_a$  for  $a \in \mathcal{A}$  will all lead to wider uncertainty bands. This procedure also yields confidence intervals on the statistics (means, differences of means, etc.).

Each of the above distributions over  $\mathcal{T}$  could optionally be disaggregated into a distribution over  $\mathcal{T} \times \mathcal{A}$ . Suppose  $\mathcal{Y}_i$  is a 1–4 Likert scale of “strongly disagree, disagree, agree, strongly agree” and  $|\mathcal{A}| = 2$ . If one judge probably disagrees and the other probably strongly agrees with  $Q_i$  for a given text ( $\hat{y}_i^a \approx 2.0$ ,  $\hat{y}_i^{a'} \approx 4.0$ ), then these two opinions would be recorded separately in the disaggregated view, rather than being averaged into “agree” ( $\hat{y}_i \approx 3.0$ ). Averaging Likert responses is often discouraged because it homogenizes diverse opinions and because it treats the Likert scale as an interval scale rather than an ordinal scale (Barry, 2017).<sup>21</sup> We suspect, however, that both aggregated and disaggregated views are useful in practice. Clicking on the lower tail of an *aggregated* distribution will display problematic dialogues that are predicted to have a low *average* score on  $Q_i$ . For a *disaggregated* distribution, the same click displays dialogues that are predicted to be problematic for *specific* judges, according to their idiosyncratic interpretations of  $Q_i$ .

## B Handling Other Types of Datasets

Our experimental datasets used a certain kind of rubric and a simple data collection mechanism. However, the idea of predicting human judgments with a calibration network is quite general and can be extended to a variety of practical settings. We discuss some useful examples in this appendix.

**Additional Features.** Our calibration network’s input is only  $\mathbf{x}$ , the vector of LLM responses on  $T$ . To give it access to other predictive features,  $\mathbf{x}$  or  $\mathbf{z}_2$  could also be augmented with a fixed-dimensional embedding of  $T$  (as already noted in footnote 4). The embedding function could be pretrained, or it could be fine-tuned jointly with the calibration network.

To avoid overfitting if the embeddings are high-dimensional, the embeddings can be replaced with  $\mathbf{0}$  during initial training. When the embeddings are revealed in the next phase of training, they may reveal properties of the text that systematically cause the calibrated LLM to overestimate or underestimate the human judges’ answers to certain questions. The calibration network can then learn to use them to further correct its estimates.

This is an example of the general principle that regression can be more accurate with more regressors. For the same reason, it may be useful for  $\mathbf{x}$  to include additional LLM questions (see footnote 2), which might cover additional criteria or use variant prompts. Ambitious questions might potentially ask the LLM to think step-by-step about the user’s goals and whether they are achieved (chain of thought), or to take on specific personas (Wang et al., 2024) that might reflect the values and needs of some human judges. If internal states of the (Transformer) LLM are available,  $\mathbf{x}$  can be further enriched with information about how it computed each distribution  $p_{\text{LLM}}(y_i | T, Q_i)$ , such as a high-layer encoding of the final token of the  $T, Q_i$  prompt, which strongly influences this distribution.<sup>22</sup> Similarly,  $\mathbf{x}$  could include other features of  $T$  that are extracted by manual code or trained classifiers rather than by prompted LLMs. It could even include features of the judge  $a$ , which allows sharing parameters across similar judges—especially useful when end users are enlisted as judges (discussed later in Appendix B). Finally, it may improve accuracy to include available *metadata* about  $T$ , such as its domain, date, and author—but such metadata should be

<sup>20</sup>As a caveat, this procedure assumes that the true  $y_i^a$  values actually follow this joint distribution, i.e., that the calibration network is correct. To incorporate uncertainty about the network’s parameters as well, we would also have to retrain them each time on a bootstrap replicate of the training set. Then a small training set would also lead to wider uncertainty bands. We would also likely get wider uncertainty bands by modeling and sampling the judgments  $y_i^a$  jointly (for each  $i$ ). We currently model them as independent (see footnote 5), but this assumes that the errors  $\hat{y}_i^a - y_i^a$  are uncorrelated. In fact they are likely to be positively correlated across judges  $a$  on the same text  $T$  and also across similar texts, since they are derived from the same or similar LLM response vectors  $\mathbf{x}$ .

<sup>21</sup>Disaggregation therefore avoids averaging over judges. Even then, however, each  $\hat{y}_i^a$  is still itself a weighted average over possible responses by  $a$ . This inner average may be problematic as well (footnote 6). Still, it elides only uncertainty, not disagreement, so disaggregating it seems less useful.

<sup>22</sup>Thanks to Zhichu (Brian) Lu for this observation.

masked for predictions that will be used to compare performance on different domains, dates, or authors, so that the predicted scores are fair in the sense that they depend only on the text  $T$ .

**Missing Features.** The Limitations section suggested using an “adaptive rubric” to reduce the number of queries to the LLM at test time. An adaptive rubric would query the LLM dynamically to ask the most useful questions first and to ask only as many questions as are needed to predict target quantities such as  $\hat{y}_0$ .

However, this requires being able to predict  $y_i^a$  values even when some of  $\mathbf{x}$  is missing.<sup>23</sup> If we train LLM-RUBRIC with dropout, then it will be able to handle this case.

Furthermore, we can extend the LLM-RUBRIC output so that it predicts not only distributions over the human responses  $y_i^a$ , but also distributions over the missing parts of  $\mathbf{x}$  (Uria et al., 2016; Devlin et al., 2018; Kachuee et al., 2019; Covert et al., 2023)—that is, over what the LLM might say if asked. This can be used for dynamically choosing the next LLM question. Dynamic feature selection dates back at least to He et al. (2012). We envision an approach similar to that of Covert et al. (2023) and Zhong et al. (2023), which greedily selects the next LLM question  $Q_i$  based on information gain—essentially, based on how much the variance of the predicted  $\hat{y}_0$ , for example, is expected to decrease after observing the LLM’s distributional answer to  $Q_i$ ,  $p_{\text{LLM}}(y_i \mid T, Q_i)$ . Computing this requires us to guess how the LLM is likely to respond to  $Q_i$ , given its responses to previous questions (i.e., we consider how it might fill in the missing part of  $\mathbf{x}$  given the part of  $\mathbf{x}$  that has been observed so far, and average over these possibilities).

Dealing with missing features is also necessary if the input feature set evolves over time. We may not wish to compute old features on new data, or new features on old data. Indeed, we may not be able to do so, if the feature has changed because the underlying LLM has been replaced with a new version.

**Irregular Datasets.** Our training objective (1) tries to predict  $y_i^a$  for each  $(T, i, a, y_i^a)$  tuple in the training dataset  $\mathcal{D}$ . Any collection of tuples can be used. That is, it is not necessary to obtain answers to all human questions for every text, or to use the same judge for all questions on a text. This is often useful.

First, in practice, new texts  $T$  or questions  $Q_i$  may periodically be added to the training dataset  $\mathcal{D}$  to better cover the observed distribution of test data and to track newly identified issues. The set of judges may also change over time due to staff turnover. As new tuples are collected, LLM-RUBRIC can simply be retrained on the growing, heterogeneous dataset.

Second, perhaps not every question is applicable to every text, and not every human judge has the same expertise. Thus, each text  $T$  might select a different set of questions  $Q_i$ , and might route different questions to different judges. A manually written policy can rule out inapplicable questions (for both human judges and LLM evaluators) by consulting text classifiers or the results of earlier questions. The applicable questions should be routed to judges with appropriate expertise,<sup>24</sup> which—depending on the question—may include familiarity with  $T$ ’s topic, dialect, or type of user. Yoshikawa et al. (2021) review and propose methods for routing texts to judges—a problem that is closely related to dynamic feature selection above. Some questions may require special expertise independent of  $T$ , e.g., questions that assess the harmfulness or inappropriateness of a dialogue system’s response according to the policies of the system’s owner.

Third, even when it is reasonable to ask a particular question of a particular judge, doing so may not be the best use of a limited annotation budget. One may use an active learning workflow (Settles, 2012) that prioritizes annotations that are not already predictable by LLM-RUBRIC—that is, where  $\hat{p}_a(y_i \mid T, Q_i)$  still has high variance after  $\hat{p}_a$  has been trained on previously collected data.

Fourth, in a dialogue system, we may be able to enlist our end users as additional judges, which is especially useful on private dialogues that only the users can see. For example, it is currently common to give users the opportunity to click or (which may in turn trigger followup questions). We regard this

---

<sup>23</sup>We can represent a missing LLM response in  $\mathbf{x}$  by having  $p_{\text{LLM}}(y_i \mid T, Q_i)$  put all of its probability on a special value  $y_i = \text{MASK} \notin \mathcal{Y}_i$ .

<sup>24</sup>We remark that to complement the judges’ own expertise, one might equip them with information beyond  $T$ . That is, for some  $(a, i)$  pairs, the judge  $a$  could consistently be shown additional information, such as the output of a fact-checker or machine translation system, or the user’s reaction to the system response. The trusted judges  $\mathcal{A}$  of overall quality ( $Q_0$ ) could be shown expert judges’ responses to other rubric questions, or their response distributions as predicted by LLM-RUBRIC.

click or non-click as just another human judgment  $y_i$  that we wish to predict.<sup>25</sup> Note that this question is distinct from the question  $Q_0$  that asks for the overall quality of the text (which usually uses a Likert scale and which may ask about aspects of the dialogue beyond user satisfaction). The calibration network can be used to predict the user’s response—that is, a click or non-click<sup>26</sup>—from various LLM questions. Some of these LLM questions may be designed to detect various kinds of verbal feedback from the user, such as praise or specific complaints, rather than assessing the system’s responses directly. In fact, Lin et al. (2024) present a method for *automatically* creating questions of this sort from private dialogues. Questions about verbal feedback may also be presented to additional human judges—though only on synthetic or other non-private dialogues—so that they contribute to multi-task regularization of LLM-RUBRIC and so that calibrated versions can be shown on a dashboard (Appendix A).

**Heterogeneous Response Types.** Equation (5) constructed a softmax distribution  $\hat{p}_a$  over a small finite response set  $\mathcal{Y}_i$ . But if some  $Q_i$  demands real-valued responses (e.g.,  $\mathcal{Y}_i = \mathbb{R}$ ), then  $\hat{p}_a(y_i | T, Q_i)$  for that  $i$  can simply be changed to a density model, where the calibration network predicts the parameters of some parametric density function from  $\mathbf{z}_2$ . Similarly, if some  $Q_i$  demands textual responses (e.g.,  $\mathcal{Y}_i = \Sigma^*$ ), then  $\hat{p}_a(y_i | T, Q_i)$  can be changed to an autoregressive language model conditioned on  $\mathbf{z}_2$ .

Next, consider the case where  $\mathcal{Y}_i$  is finite but large. Here the matrices in equation (5) are large, so generalization might be improved by smoothing them. This can be done by parameterizing  $V_i = Y_i U_i$  and  $V_i = Y_i U_i^a$ , where the rows of matrix  $Y_i$  serve as embeddings of the various responses  $y_i \in \mathcal{Y}_i$ . Similar responses should have similar embeddings. The unsmoothed case takes  $Y_i$  to be the identity matrix (yielding one-hot embeddings), but using a learned matrix  $Y_i$  with fewer columns can reduce the number of parameters. In some cases,  $Y_i$  does not even need to be learned: pre-trained word embeddings can be used if  $\mathcal{Y}_i$  is a natural-language vocabulary, and systematic number embeddings (Gorishniy et al., 2022) can be used if (e.g.)  $\mathcal{Y}_i = \{1, \dots, 100\}$ .

The preceding paragraph treats large response sets for human judges, which are predicted by the *output* of the calibration network. If the LLMs are also permitted to use large response sets, which appear in the *input* of the calibration network, a similar solution applies: premultiply the vector  $p_{\text{LLM}}(y_i | T, Q_i)$  by  $Y_i^\top$  to reduce its dimensionality before including it in  $\mathbf{x}$ . For infinite response sets as in the first paragraph, standard schemes can be used to embed numbers (Gorishniy et al., 2022) or text (Devlin et al., 2018).

Finally, returning to the setting of our own experiments, we observe that when  $\mathcal{Y}_i$  is an ordinal scale with  $n$  possible responses, such as a Likert scale, it is not strictly necessary to use a flexible softmax distribution as we did in equation (5). Instead,  $y_i^a$  could be modeled with fewer parameters as a quantized version of an *underlying real value*  $r_i^a$  whose distribution is predicted by the calibration network.<sup>27</sup> Furthermore, we could then (if desired) evaluate the text  $T$  using our best prediction of the underlying  $r_i^a$  rather than of the observed  $y_i^a$  (e.g., using the expected value as before, if we wish to minimize expected  $L_2$  loss). The intuition is that the reconstructed unquantized  $r_i^a$  values contain more information than their quantized versions—and that they might also be more comparable across judges  $a$ , if their different judgments (e.g., in Figure 2) mainly arise from different quantization boundaries.<sup>28</sup>

---

<sup>25</sup>Similarly, we may treat “Did the user choose to visit the system again the next day?” or “How long before the user’s next visit?” as a more implicit human judgment  $y_i$  that we wish to predict.

<sup>26</sup>“No click” will usually have probability close to 1. To avoid a large number of low-information training examples, one can downsample the “no click” examples in training data from this domain, provided that  $\mathbf{x}$  indicates whether the example comes from a downsampled domain (since this kind of downsampling will shift the priors on many questions  $y_i$  toward more extreme responses, and thus should shift the hidden features  $\mathbf{z}_1, \mathbf{z}_2$  guessed from an ambiguous example  $\mathbf{x}$ ). Also, to control the number of parameters in a system with many users, a reasonable simplification is to fix  $W_k^a = 0$  when  $a$  is a user. Then for each user  $a$ , we only have to learn a matrix  $V_i^a$  with two non-zero rows (for  $\square$  and  $\checkmark$ ; the row for no response can be fixed at  $\mathbf{0}$ , WLOG). Note that in the common case where user  $a$  has never provided any explicit feedback, so that  $V_i^a = 0$ , the backoff matrix  $V_i$  still ensures a reasonable prediction—particularly if  $a$ ’s demographics and/or user behavior are represented in  $\mathbf{x}$  when predicting the answer to this question, allowing the network to share statistical strength with similar users.

<sup>27</sup>That is,  $\hat{p}_a(y_i | T, Q_i) = \hat{p}_a(r_i \in \text{bin}_{y_i} | T, Q_i)$  where  $r_i^a$  has a normal (or logistic) distribution whose 2 parameters are predicted from  $T$  by the calibration network for  $a$ , and where the  $n$  bins are a partition of  $\mathbb{R}$  by  $n - 1$  learned thresholds that are specific to  $a$  or to  $(a, i)$ . This gives a discrete distribution over  $y_i^a$ , which can be used in the log-likelihood objective (1). This is a nonlinear, heteroskedastic version of ordered probit (or logit) regression.

<sup>28</sup>However, a trick is needed to ensure that  $r_i^a$  values are interpretable and comparable across judges  $a$ . The issue is that the method in the previous footnote does not identify the position or scale of  $r_i^a$ . (If we adjusted our model to double the predicted means, predicted standard deviations, and thresholds for judge  $a$ , we would get exactly the same distribution over observables  $y_i^a$ )

**Comparative Judging.** Our maximum-likelihood training principle (1) can be extended to other question types. In particular,  $Q_i = \text{“Does } T \text{ or } T' \text{ score higher on criterion } i\text{?”}$  can be interpreted as a comparison of underlying real values as in the preceding paragraph: the human judge is being asked whether  $r_i > r'_i$ . The calibration network could predict the probability of a “yes” response either directly, or else by integrating over a latent distribution  $\hat{p}_a(r_i, r'_i | T, T', Q_i)$  (perhaps modeling it as  $\hat{p}_a(r_i | T, Q_i) \cdot \hat{p}_a(r'_i | T', Q_i)$  via an independence assumption).<sup>29</sup>

## C LLM-RUBRIC Questions

These were the questions in our evaluation rubric. The human and LLM prompts in which these questions appeared are given in Appendix D and Appendix E respectively. When presenting the questions to the LLM (Appendix D), boldface was omitted. When presenting them to human judges on real data (Appendix I), boldface was again omitted, and the response choices were not numbered; instead, radio buttons were used (Figure 3b).

Question instances where the correct answer was “NA” were not included in our training dataset  $\mathcal{D}$  and were not used for evaluation.

*Q<sub>1</sub>*— In terms of naturalness and tone of the **assistant utterances**, to what degree are they likely to be produced by an **intelligent human** in a conversation? Disregard whether they are grounded in the search results.

1. Unlikely.
2. Somewhat unlikely.
3. Somewhat likely.
4. Likely.

*Q<sub>2</sub>*— If the references are provided, to what degree user’s questions can be answered or resolved using the references? The assistant’s responses should not impact your response to this question. **If no references are provided in the conversation, please write “NA” for this question.**

1. None of the questions that user has asked could be answered using the reference documents.
2. Less than half of documents that user has asked could be answered using the reference document.
3. Half or more than half of the questions that user has asked could be answered using the reference documents.
4. All the questions the user has asked could be answered with the reference documents.

*Q<sub>3</sub>*— Independent of what sources are cited in the conversation, to what degree the claims made by the assistant are followed by a citation. **If no references are provided in the conversation, please write NA.**

1. None of the claims are followed by a citation.
2. Less than half of the claims are followed by a citation.
3. Half, or more than half of the claims are followed by a citation.
4. All claims are followed by a citation.

*Q<sub>4</sub>*— What percentage of citations accurately support the claims made in the conversation? **If no references are provided in the conversation, please write NA.**

1. None of the citations accurately support the provided claims.
2. Less than half of citations accurately support the provided claims.
3. Half, or more than half of citations accurately support the provided claims.

and achieve the same log-likelihood. But  $r_i^a$  would now have twice the range and so would count more in a mean over judges  $A$ .) To break this tie, we can augment the log-likelihood objective with a second term (perhaps with infinitesimal weight) that does care about position and scale. Assuming that our ordinal scale  $\mathcal{Y}_i$  is numeric, a natural choice for this second term is the *unquantized* log-likelihood: that is, we ask the normal curve to assign a high log-density to the exact value  $y_i^a$  and not just a high log-probability to its bin. This ties  $r_i$  to the  $y_i$  scale, making it interpretable.

<sup>29</sup>Unfortunately, any reporting of the predicted  $r_i^a$  values (e.g.,  $\hat{r}_i^a$ ) as quality metrics runs into the same non-identifiability problem as in the previous footnote. We cannot know the position or scale of a judge’s  $r_i^a$  values if we only observe the results of  $>$  comparisons. A simple fix is to apply an affine transform to each judge’s  $r_i^a$  values so that on a given reference set of texts, the transformed values have mean 0 and variance 1. Then report these transformed values.

4. All citations accurately support the provided claims.

*Q<sub>5</sub>*— To what degree the cited sources are the best candidates among all the provided sources? **If no references are provided in the conversation, please write NA.**

1. For all citations, there is a better source to be cited.
2. For more than half of the citations, there is a better source to be cited.
3. For half or less than half of the citations, there is a better source to be cited.
4. The best sources are cited in all cases.

*Q<sub>6</sub>*— To what degree the content of the assistant utterances is free of redundant elements, such as **repetition, overspecification**, etc.

1. The conversation has **a large number** of redundant elements.
2. The conversation has **some** redundant elements.
3. The conversation has **a few** redundant elements.
4. The conversation is **completely free** of redundant elements.

*Q<sub>7</sub>*— To what degree the assistant responses are concise?

1. In all assistant utterances, the responses could have been shorter.
2. In more than half of the assistant utterances, the responses could have been shorter.
3. In half, or less than half of the assistant utterances, the responses could have been shorter.
4. In all assistant utterances, the responses are concise and the utterance length is appropriate.

*Q<sub>8</sub>*— Do you think the number of exchange turns or back and forth is appropriate given the complexity of the user information need?<sup>30</sup>

1. No, fewer interactions would be sufficient and would make this conversation more pleasant.
2. No, more interactions are needed for a better conversation experience.
3. Yes, the rate of exchanges between the user and the assistant is reasonable.

---

*Q<sub>0</sub>*— Imagine you are the user who had this conversation with the assistant. All in all, how you would rate your overall satisfaction while interacting with the assistant? The higher the rating, the better the experience.

1. 1
  2. 2
  3. 3
  4. 4
- 

## D Evaluation Prompt for LLM

In our LLM-RUBRIC experiments (§4), we use the following prompt template to ask the LLM an evaluation question  $Q_i$  about a conversational text  $T$ .

The variable  $\{\text{conversation}\}$  is the complete dialogue between the user and the assistant, and the variable  $\{\text{question}\}$  is one of the questions from the evaluation rubric presented in Appendix C.

The citation-related questions  $Q_2$ ,  $Q_3$ ,  $Q_4$ , and  $Q_5$  are not presented to the LLM if no references are provided in the conversation. In this case, we simply pretend that the LLM would have correctly answered “NA,” which means that the probability vector over the responses 1–4 is  $[0, 0, 0, 0]$  (see footnote 3).

---

<sup>30</sup>For  $Q_8$ , the numeric responses unfortunately do not form an ordinal scale. Response “3” should reasonably be considered closer to “1” than it is to “2”. Thus,  $L_2$  is not an appropriate loss function here. However, for simplicity we did still use  $L_2$  when decoding  $\hat{y}_8^a$  (it motivates equation (2)) and when evaluating the quality of  $\hat{y}_8^a$  (it motivates RMSE). This affects only the  $Q_8$  line of Table 3, all of whose metrics would presumably be improved if we fixed the problem by swapping “2” and “3” in both the human data and the LLM data. All of our other results would be unaffected by this relabeling.

You are given a conversation between a user and an intelligent assistant for an enterprise chat scenario. In some cases, some references and citations are provided to back up the claims made by the intelligent assistant. Your primary job is to evaluate the quality of the conversation based on a criterion. To do so, read the conversation and references, and answer the followed question, by selecting only one of the choices.

Conversation: {conversation}

Question: {question}

Only print '1', '2', '3', or '4'.

---

## E Evaluation Prompt and Preliminary Data Quality Questions for Humans

Below are the instructions we gave to human judges with the main questions in Appendix C.

The preliminary questions DQQ0–DQQ2 are used only to screen for problems with the generated synthetic dialogues of §3.2 (see Appendix G). They are not included when the human is judging the real dialogues of §3.3. Note that if the answer to DQQ is “No,” then the remaining questions are not answered, which is why our synthetic training dataset  $\mathcal{D}$  had only 741 examples rather than 750.

---

You are given a **conversation** between a user and an intelligent assistant for an enterprise chat scenario. You are also given **an information need that the user wants to fulfill through the course of the conversation (e.g., a problem the user faces and wants to resolve)**. In some cases some references and citations are provided to back up the claims made by the intelligent assistant. Each assistant utterance can only cite the references listed in the adjacent cell in the table.

**Your primary job is to evaluate the quality of the conversation through a series of criteria that we define later in the document. To evaluate the conversation, you need to answer a questionnaire. Each question captures one evaluation criteria that we care about.**

Read about the definition of labels criteria below:

**Naturalness (both content and form):** The degree to which the form and content of the conversation is realistic, and likely to happen in real-world. To measure naturalness you should answer below questions:

DQQ0- Is this a conversation between a user and an assistant?

1. Yes
2. No (if you select ‘No’, you can skip the rest of the questions)

DQQ1- To what degree the user tries to fulfill the information need during the course of conversation?

1. The conversation is not about the user information need at all.
2. The conversation does not exactly address the user information need, but it is somewhat related.
3. The conversation addresses the user information need but it also talks about other topics.
4. The conversation only addresses the user information need.

DQQ2- To what degree the form and content of the **user utterances** are likely to be produced by a **human in a conversation**?

1. Unlikely.
2. Somewhat unlikely.
3. Somewhat likely.
4. Likely.

{ $Q_1$ }

**Citation quality:** To what degree the claims made by the **assistant** are backed by reliable sources. Note that not all the sentences in a conversation require citation; only facts and claims need to be cited. To measure citation quality answer the following questions:

$\{Q_2\}$   
 $\{Q_3\}$   
 $\{Q_4\}$   
 $\{Q_5\}$

**Dialogue efficiency:** To what degree the dialogue has been conducted in an cost effective manner. To measure the dialogue efficiency answer the following questions:

$\{Q_6\}$   
 $\{Q_7\}$   
 $\{Q_8\}$

**User Satisfaction:** Answer the following question to rate the overall user experience with the assistant.

$\{Q_0\}$

---

## F Synthetic Dialogue Generation

This section describes the 5 approaches that we used in §3.2 to generate a variety of synthetic dialogues.

**DS1: LLM-Only Assistant with Simulated User.** In our baseline, the dialogue system has no access to external documents and can only answer the user from its internal knowledge. In this setting, the assistant cannot provide citations for its claims.

**DS2: Oracle RAG Assistant with Oracle Simulated User.** In this variant, the prompt includes highly relevant documents: the 5 documents that were most frequently clicked when the given topic appeared as a query in the real logs of §3.1. Thus, the assistant is essentially a RAG system with unrealistically good retrieval. In addition, the simulated user is unrealistically knowledgeable, having full access to the same documents for the initial question and all followup questions.

**DS3: RAG Assistant with Oracle Simulated User.** This variant resembles DS2, except that it uses the 5 documents that are most similar to the topic string according to the BM25 metric. We use the ElasticSearch<sup>31</sup> implementation of BM25.

**DS4: RAG Assistant with Simulated User.** This variant resembles DS3, but the topic is included in the prompt only when generating simulated user turns, and the 5 documents are included in the prompt only when generating assistant turns. In addition, the BM25 query is not the topic string but rather the dialogue history (all past utterances); thus, each assistant turn may be prompted using a different set of 5 documents.

**DS5: Retrieval-Augmented Dialogue Generation + Query Generation with Simulated User.** This variant resembles DS4, but the BM25 query is not the dialogue history. Instead, it is derived from the dialogue history by a separate prompt to the LLM (also shown in Table 6). This may be required as calling a *query generation tool*.

The prompts used for synthetic dialogue generation (DS1–DS5) are presented in Table 6.

## G Quality of the Generated Synthetic Dialogues

As mentioned in §3.2, each of the systems DS1–DS5 (Appendix F) was used to generate 50 synthetic dialogues, each of which was evaluated by 3 human judges, resulting in  $5 \times 50 \times 3 = 750$  completed questionnaires. The first question we asked (DQQ0) was “Is this a conversation between a user and an assistant?” As expected based on the findings presented in (Li et al., 2023), the answers to this question were vastly positive: 98.8% of the dialogues received a positive answer.

<sup>31</sup><https://www.elastic.co/>

	<b>DS1</b>	<b>DS2</b>	<b>DS3</b>	<b>DS4</b>	<b>DS5</b>
DQQ1	$3.8 \pm 0.5$	$3.6 \pm 0.7$	$3.6 \pm 0.8$	$3.6 \pm 0.8$	$3.5 \pm 0.9$
DQQ2	$3.6 \pm 0.6$	$3.4 \pm 0.8$	$3.3 \pm 0.9$	$3.3 \pm 0.9$	$3.3 \pm 0.8$
$Q_1$	$3.4 \pm 0.7$	$3.2 \pm 0.9$	$3.2 \pm 0.9$	$3.1 \pm 0.8$	$3.1 \pm 0.9$
$Q_2$	NA	$3.5 \pm 0.8$	$3.3 \pm 0.9$	$3.4 \pm 0.9$	$3.3 \pm 1.0$
$Q_3$	NA	$3.3 \pm 0.9$	$3.0 \pm 1.0$	$3.2 \pm 1.1$	$3.0 \pm 1.3$
$Q_4$	NA	$3.3 \pm 0.8$	$3.1 \pm 1.0$	$3.1 \pm 1.1$	$3.0 \pm 1.3$
$Q_5$	NA	$3.3 \pm 0.9$	$3.1 \pm 1.0$	$2.9 \pm 1.1$	$2.6 \pm 1.3$
$Q_6$	$3.6 \pm 0.7$	$3.3 \pm 0.8$	$3.2 \pm 0.9$	$3.7 \pm 0.6$	$3.6 \pm 0.7$
$Q_7$	$3.6 \pm 0.7$	$3.1 \pm 0.9$	$3.1 \pm 1.0$	$3.5 \pm 0.8$	$3.6 \pm 0.7$
$Q_8$	$2.7 \pm 0.7$	$2.5 \pm 0.9$	$2.6 \pm 0.8$	$2.5 \pm 0.6$	$2.5 \pm 0.6$
$Q_0$	$2.3 \pm 0.7$	$3.1 \pm 0.8$	$3.0 \pm 0.8$	$3.0 \pm 0.9$	$2.9 \pm 0.9$

Table 4: Mean and standard deviation of human annotations for different sets of synthetic dialogues. As each column has  $n \approx 148$  examples, the standard error of the mean is about  $\frac{1}{12}$  of the standard deviation shown. Thus a 95% confidence interval on the mean is  $\pm \frac{1}{6}$  of the standard deviation, ranging here from  $\pm 0.1$  to  $\pm 0.2$ .

	<b>DS1</b>	<b>DS2</b>	<b>DS3</b>
# conversation	76	71	76
$Q_1$	$3.329 \pm 0.817$	$3.197 \pm 0.973$	$3.066 \pm 0.964$
$Q_2$	NA	$3.155 \pm 0.816$	$2.763 \pm 0.930$
$Q_3$	NA	$2.971 \pm 0.903$	$2.631 \pm 0.900$
$Q_4$	NA	$3.112 \pm 0.943$	$2.618 \pm 0.959$
$Q_5$	NA	$3.014 \pm 1.013$	$2.631 \pm 1.049$
$Q_6$	$3.473 \pm 0.734$	$3.436 \pm 0.686$	$3.473 \pm 0.678$
$Q_7$	$3.552 \pm 0.768$	$3.295 \pm 1.012$	$3.500 \pm 0.716$
$Q_8$	$2.803 \pm 0.487$	$2.788 \pm 0.501$	$2.739 \pm 0.440$
$Q_0$	$2.668 \pm 0.817$	$2.915 \pm 0.835$	$2.697 \pm 0.707$

Table 5: Mean and standard deviation of different sets of real human-agent dialogues. In all cases,  $\pm 0.24$  gives a 95% confidence interval on the mean.

Table 4 shows the mean and standard deviation of the human judgments for the questionnaires that passed the DQQ0 quality check. The results on DQQ1 and DQQ2 suggest that all systems often simulated the user turns competently, and the results on  $Q_1$ – $Q_8$  and  $Q_0$  suggest that all systems often produced reasonably good assistant responses to these simulated users. In fact, the DS2 and DS3 systems obtained an average  $\geq 3.0$  over their dialogues for *all* questions (except for  $Q_8$ , where the response scale is 1–3).

Of course, the point of our LLM-RUBRIC experiments is not to generate good dialogues but to determine which dialogues show more satisfactory behavior by the assistant. These generated dialogues simply provide synthetic training and development data for that task.

**Questions on the naturalness of dialogues (DQQ1, DQQ2,  $Q_1$ ).** Table 4 indicates that system DS1 produces the most natural conversations. This is a non-RAG system that simply asks the LLM to write a plausible dialogue on the given topic. The other four systems perform comparably in terms of generating natural dialogues. DS2 performs slightly better than the rest; this it may be due to the high quality of its references, which can be less noisy and confusing than the other variants.

**Questions on citations ( $Q_2$ ,  $Q_3$ ,  $Q_4$ ,  $Q_5$ ).** On citation quality and usage, DS2 achieves the highest average rating, thanks to its “oracle” RAG. Among the methods that perform RAG with various BM25 queries, DS3 and DS4 perform slightly better than DS5, which prompts an LLM to generate the BM25 query.

**Questions on conciseness ( $Q_6$ ,  $Q_7$ ,  $Q_8$ ).** All systems are similar at generating an appropriate number of turns ( $Q_8$ ). DS2 and DS3 seem to have less concise dialogues ( $Q_6$ ,  $Q_7$ ), perhaps because the simulated

user has access to the retrieved documents.

**Question on overall satisfaction ( $Q_0$ ).** The results suggest that the quality of retrieved documents is the most important factor for our judges, with DS1 clearly doing worst and DS2 doing slightly better than the others.

## Azure Enterprise Chat

Imagine that you are a user of an enterprise chat bot related to Azure. Have an interactive conversation with the bot to learn about the following topic.

**Topic: azure cdn ip range**

If you cannot have a meaningful conversation with the bot about this topic, feel free to refresh the page to receive another topic.

Please feel free to ask follow up questions if you need more information to fully understand the topic

Once you are done having the conversation, click on the **End of Conversation** button to complete a survey about the conversation you just had

The screenshot shows a web-based chat interface for Azure Enterprise Chat. On the left, there is a sidebar with two sections: one for claims citation and one for dialogue citation. The main area shows a conversation between a user and an AI agent. The user asks about Azure CDN IP ranges, and the agent provides a detailed response. Below the conversation, there is an 'End of Conversation!' button. The right side of the interface contains a sidebar for evaluation, which includes a 'References:' section with several links and a survey question asking for user feedback on their experience with the agent.

**(a) First, we ask the user to have a conversation with the agent about a given topic.**

Independent of what references are cited in the conversation, to what degree the claims made by the assistant are followed by a citation. If no references are provided at the end of conversation, please select NA.

- NA
- None of the claims are followed by a citation.
- Less than half of the claims are followed by a citation.
- Half, or more than half of the claims are followed by a citation.
- All claims are followed by a citation.

What percentage of citations in the dialogue accurately support the claims made in the conversation? If no references are provided at the end of conversations, please select NA.

- NA
- None of the citations accurately support the provided claims.

Message... ➤

Hello! How can I assist you today?

What is azure cdn ip range?

The IP ranges for Azure CDN are the same as the data center IP ranges. You can find the data center IP ranges easily from the network. [1]

To find the IP ranges for Azure CDN, you can go to the Azure portal and select your Azure Media Services account. Then, in the Settings blade, select Streaming endpoints. From there, you can view and configure the properties of the streaming endpoint, including the IP addresses associated with it. Additionally, you can try to trace the CDN IP address from your location or ping the CDN URL to confirm the IP address ranges.

End of Conversation!

Thanks for having a conversation with our bot! This bot had access to the following sources related to this topic.

**References:**

- [1]: <https://social.msdn.microsoft.com/forums/azure/en-us/082a16af-065f-4183-96ab-be7f17e84c/azure-cdn-ip-ranges>
- [2]: <https://docs.microsoft.com/en-us/azure/media-services/previous/media-services-portal-manage-streaming-endpoints>
- [3]: <https://azure.microsoft.com/en-gb/resources/templates/front-door-standard-premium-container-instances-application-gateway-public>
- [4]: <https://docs.microsoft.com/en-us/azure/cdn/cdn-billing>
- [5]: <https://docs.microsoft.com/en-us/azure/architecture/aws-professional/networking>

If references are provided, you need to click on the above links and read them in order to answer the survey questions in the sidebar.

Message... ➤

**(b) Once the user clicks on 'End of Conversation', we present all the search engine results available to the agent and ask the user to read them. Finally, we ask the user to evaluate their experience with the agent by answering the evaluation rubric questions (Appendix C).**

Figure 3: User interface for real dialogue collection and evaluation.

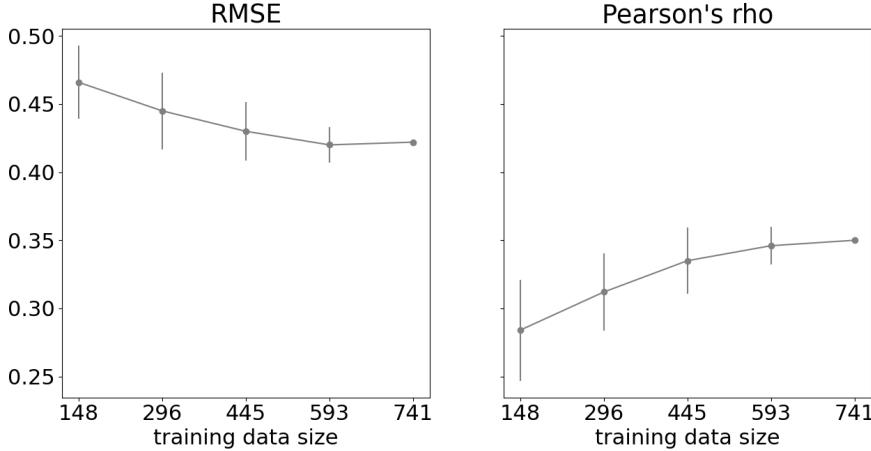


Figure 4: Learning curve for training the personalized calibration network in LLM-RUBRIC on synthetic conversations and testing on the real conversation data. The model’s performance becomes relatively stable after observing 80% of the training data. Note that the LLM itself is not fine-tuned to predict any judge’s responses.

## H The User Interface for Human-Agent Dialogue Collection and Evaluation

We designed a web interface (Figure 3) to enable humans to converse with a dialogue system as users and then evaluate their interactions as judges (§3.3). In each session, the website shows the human a random Azure-related topic from the set described in §3.1, and randomly selects one of the dialogue systems DS1–DS3 for the human to converse with. The human does not know which system was selected (although DS1 might be distinguishable as it does not give citations).

This is a standard guided data collection procedure that has been previously used in prior work (Zamani et al., 2023). If the user does not understand the topic, they may refresh the website to get a different topic. Once the user is finished with their conversation, they click on the ‘End of Conversation’ button and judge the conversation (see Appendix C).

## I Evaluating the Collected Human-Agent Dialogues

We asked 13 trained judges to use the website for dialogue collection and evaluation. The judge set for the synthetic dialogues presented above includes these 13 judges. We collected a total of 223 conversations, ranging from 14–27 conversations per judge. The judge scores for the three dialogue systems evaluated are summarized in Table 5.

## J How much human judge data is needed to train calibration?

We plot learning curves in Figure 4. To make these plots, we train the model on the synthetic data and test on the real conversation data, but reduce the training portion of the data to a random  $x\%$  sample. To reduce the impact of random selection, we repeat this process 50 times and plot the average performance,  $\pm 1$  standard deviation. As expected, the average performance improves and drops in variance as we increase the amount of training data per judge.<sup>32</sup>

Performance is reasonably good with even 20% of our training set, and appears to have essentially converged by the time we reach 80–100% of our training set. (Here 100% represents 741 dialogues, where each judge has evaluated only  $\sim 30$  dialogues on average.) Further improvements would, therefore, require more dimensions or more accurate modeling of each dimension, or perhaps training data targeted at fixing the residual errors.

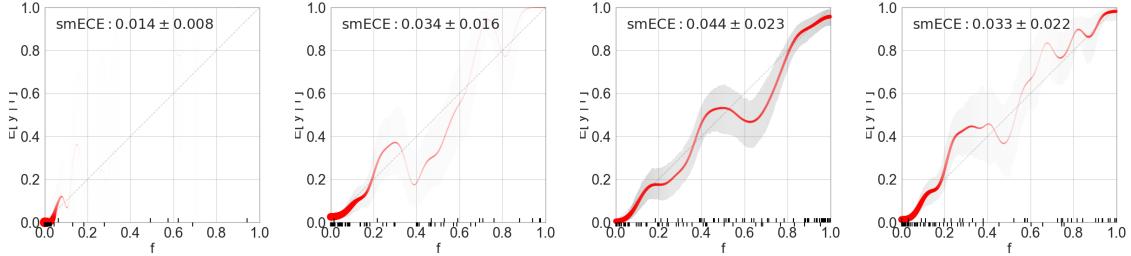


Figure 5: Calibration plots for  $Q_0$  on held-out synthetic dialogues, as explained in §6 and Appendix K. These are plots for  $y_0 \in \{1, 2, 3, 4\}$  respectively. They show low calibration error.

## K Calibration Plots (Reliability Diagrams)

Well-trained neural networks tend to produce well-calibrated probabilities (Niculescu-Mizil and Caruana, 2005), an outcome that is incentivized by the log-likelihood training and validation objectives. Figure 5 shows smoothed calibration plots when training and evaluating our system on synthetic dialogues, as explained in §6. The system is indeed well-calibrated, meaning that the red curves stay close to the diagonal, as measured by smoothed expected calibration error (smECE).

The plots are produced by the `replot` package<sup>33</sup> of Błasiok and Nakkiran (2023), using 5-fold cross-validation. All graphs plot the same examples, namely the tuples  $(T, i, a, y_i^a) \in \mathcal{D}$ , where  $T$  is a synthetic dialogue. However, each graph considers the calibration for a different possible score  $y_0$ . Tick marks just above (or below) the horizontal axis are a sample of held-out examples for which the true judgment  $y_i^a$  is (or is not)  $y_0$ . Their position along the horizontal axis shows their predicted probabilities  $\hat{p}_a(y_0 | T, Q_0)$  under cross-validation. Thus, the tick marks above the axis (true score is  $y_0$ ) tend to appear farther to the right ( $y_0$  is predicted with high probability).

For each predicted probability  $p$ , the height of the red curve estimates the actual probability that  $y_i^a = y_0$  among held-out examples where  $\hat{p}_a(y_0 | T, Q_0) \approx p$ . One may think of this visually as the approximate fraction of tick marks that are plotted near  $p$  on the horizontal axis that are just above the axis rather than just below. The thickness of the red curve at  $p$  corresponds to the density of tick marks near  $p$ . The gray band around the red curve is a 95% bootstrap confidence interval.

The smoothed expected calibration error (smECE) is the average absolute distance between the height of the red curve and the diagonal, weighted by the thickness of the curve. In other words, it estimates the average difference between the predicted and actual probability for a random held-out example. The smECE number for each graph is printed on the graph with a 95% confidence interval.

Calibration of  $\hat{p}_a$  is important because it means that the predicted probabilities are meaningful. The system can use them to assess its own uncertainty and make decisions accordingly. Some applications in our case:

- **Text evaluation.** The expected score (2) will be approximately unbiased: that is, on average over held-out examples, it will match the judge’s actual score. Table 1 assesses this match directly. Beyond expected score, various other interesting quantities that we might derive from  $\hat{p}_a$  are also approximately unbiased: for example, the probabilities that the score is  $\leq 1$ ,  $\leq 2$ , and  $\leq 3$ .
- **Text selection.** At runtime, a dialogue system might generate several candidate responses, and then choose the one with maximum expected reward. If the reward has the form  $\sum_{i,a} f_{i,a}(y_i^a)$ , for any set of reward functions  $f_{i,a}$ , then its expectation under  $\hat{p}$  will be unbiased.
- **Dynamic feature selection.** LLM-RUBRIC can be sped up at test time by asking fewer questions of the LLM. As briefly mentioned in §8 and Appendix B,  $\hat{p}_a$  can be used to greedily choose the

<sup>32</sup>The reduction in variance is partly because the larger training sets are more similar to the population and thus to each other, but also because they overlap more and thus are less independent.

<sup>33</sup><https://github.com/apple/ml-calibration>

question with the greatest information gain—that is, whose answer is predicted to most reduce the variance of the evaluation  $\hat{y}_0$  or most reduce the entropy of a text selection decision.

- **Distillation.** LLM-RUBRIC can be used to stochastically label a large dataset of naturally occurring texts according to  $\hat{p}_a$  (“multiple imputation”). A faster scoring function can then be trained to have low loss on this dataset.
- **Rubric improvement.**  $\hat{p}_a$  can be used to identify difficult texts where LLM-RUBRIC is unsure what judge  $a$  would say, or controversial texts where LLM-RUBRIC predicts that two judges will disagree more than usual. This can be used to improve the LLM questions or the human questions, respectively.

As a caveat, the plots in Figure 5 measure calibration only for the dataset as a whole. One could create calibration plots for subsets of the data to verify that the model remains calibrated within each user category, judge, or dialogue topic that is sufficiently represented in training data—as one would expect with maximum-likelihood training—rather than overestimating probabilities in some categories and underestimating them in others. The correlation coefficients in Figure 5 do show that the predicted scores provide a reasonable ranking of examples.

<b>Prompt</b>	
<b>DS1</b>	A user wants to know about “{topic}”. Write a conversation between a user and a helpful assistant about user’s need.
<b>DS2 &amp; DS3</b>	A user wants to know about “{topic}”. Write a conversation between the user and a helpful assistant about user’s need. The assistant should provide factual responses using the following Sources. Cite Sources as needed, like [3] or [2]. Sources: [ 1 ] {Reference 1} [ 2 ] {Reference 2} ...
<b>DS4 &amp; DS5 (Init)</b>	Imagine a user wants to know about “{topic}” by talking to an intelligent assistant. What would be the first question that the user asks? Generate the output in this format: “User: utterance”.
<b>DS4 &amp; DS5 (Assistant)</b>	Imagine a user is interacting with an intelligent assistant to solve their problem. The assistant should provide factual responses using the following sources, or if that is not possible, asks useful questions to get a better understanding of the user need. Cite Sources as needed, like [3] or [2]. Sources: [ 1 ] {Reference 1} [ 2 ] {Reference 2} ... Complete the following dialogue with only one utterance. If there is no need for a response, only generate “END OF CONVERSATION!” Assistant: How can I help you? ... User: {Last Generated User Utterance} Assistant:
<b>DS4 &amp; DS5 (User)</b>	Imagine a user is interacting with an intelligent assistant to solve their problem about “{topic}”. Complete the following dialogue with only one utterance. If there is no need for a response, only generate "END OF CONVERSATION!" Assistant: How can I help you? ... Assistant: {Last Generated Assistant Utterance} User:
<b>DS5 (QGen)</b>	Assume that you plan to answer the user’s question in the following conversation: Assistant: How can I help you? ... User: {Last Generated User Utterance} What query will you submit to a search engine to find the answer? Only generate the query.

Table 6: Prompts used for synthetic data generation using gpt-3.5-turbo-16k.