

000 MENLO: FROM PREFERENCES TO PROFICIENCY – 001 002 EVALUATING AND MODELING NATIVE-LIKE QUALITY 003 ACROSS 47 LANGUAGE VARIETIES 004

005
006 **Anonymous authors**
007 Paper under double-blind review
008
009
010

011 ABSTRACT 012

013 Ensuring native-like quality of large language model (LLM) responses across
014 many languages is challenging. To address this, we introduce MENLO, a frame-
015 work that operationalizes the evaluation of native-like response quality based on
016 audience design-inspired mechanisms. Using MENLO, we create a dataset of
017 6,423 human-annotated prompt–response preference pairs covering four quality
018 dimensions with high inter-annotator agreement in 47 language varieties. Our
019 evaluation reveals that zero-shot LLM judges benefit significantly from pairwise
020 evaluation and our structured annotation rubrics, yet they still underperform hu-
021 man annotators on our dataset. We demonstrate substantial improvements through
022 fine-tuning with reinforcement learning, reward shaping, and multi-task learning
023 approaches. Additionally, we show that RL-trained judges can serve as generative
024 reward models to enhance LLMs’ multilingual proficiency, though discrepancies
025 with human judgment remain. Our findings suggest promising directions for scal-
026 able multilingual evaluation and preference alignment. We release our dataset and
027 evaluation framework to support further research in multilingual LLM evaluation.

028 1 INTRODUCTION 029

030 In order for LLMs to be most useful across the globe, they need to be able to provide high-quality
031 responses in many languages. Responses should be relevant (Zhuang et al., 2024), factually accurate
032 (Jacovi et al., 2025), and natural (Marchisio et al., 2024; Guo et al., 2025), among other considera-
033 tions. Ultimately, for interaction in any language to be seamless, responses need to be indistin-
034 guishable from those of a native speaker (Novikova et al., 2016; Liu et al., 2021). Language proficiency in
035 humans has traditionally been evaluated via standardized tests (Jamieson et al., 2000). While such
036 tests have been applied to evaluating LLMs (Anil et al., 2023; Mayor-Rocher et al., 2024; Lothritz
037 & Cabot, 2025), they are difficult to scale and do not readily correspond to real-world conversations.
038 What is considered a *native-like* response largely depends on speakers’ and listeners’ interpretations
039 of whom they are speaking to (Bell, 1984).

040 To operationalize the evaluation of native-like response quality across languages, we propose
041 **Multilingual Evaluation of Native-Like Output (MENLO)**; see Figure 1 for an overview. MENLO
042 breaks down native-like response quality into four key dimensions: i) language quality and coher-
043 ence; ii) alignment with cultural and linguistic nuances of a specific language variety/locale; iii)
044 factual correctness and grounding in the local context; and iv) overall writing style and helpfulness.

045 Building on mechanisms from audience design (Bell, 1984), we propose creating tailored prompts
046 that effectively evoke local contexts by defining the target audience (e.g., an addressee or reference
047 group), thereby guiding the generated language to converge to contextually appropriate “native”
048 styles. We develop instructions that reduce annotation subjectivity and improve inter-annotator
049 agreement. Responses are generated using state-of-the-art LLMs and annotated with ratings on a
050 1–5 Likert scale, with an average Krippendorff’s $\alpha = 0.84$. Overall, the MENLO dataset consists of
051 6,423 annotated prompt-response preference pairs, and 81,014 annotations, in 47 language varieties.

052 Human evaluation, particularly at a massively multilingual scale is expensive. We thus evaluate the
053 ability of LLMs to serve as judges of native-like quality responses. We find that in zero-shot setting,
pairwise evaluation—where models predict scores for two responses simultaneously (without ex-

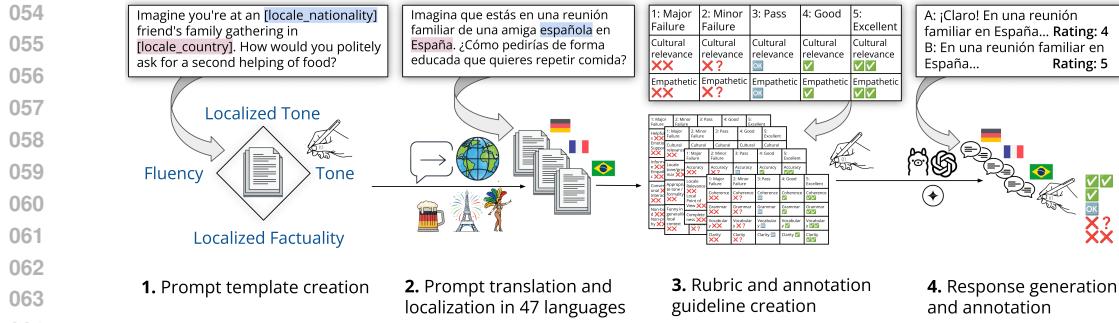


Figure 1: MENLO framework and annotation process. 1) Human-written prompt templates evoking local contexts are created in English for the four dimensions. 2) Prompt templates are translated and localized into 47 language varieties. 3) Annotation guidelines are created that break down each dimension into easy-to-follow rubrics. 4) LLMs are used to generate response pairs for each prompt, which are annotated with Likert-scale ratings and preferences.

pliably predicting preference)—significantly outperforms its pointwise counterpart. The advantage of evaluating two responses side-by-side is even bigger than in-context examples with labels. In addition, we observe significant improvements with judges using our annotation rubrics compared to judges without rubrics, highlighting the generality of the MENLO framework.

As zero-shot judges remain below human annotation quality, using the pairwise evaluation setup, we fine-tune Qwen3-4B and Llama4-Scout as LLM judges on the MENLO training data, finding that RL-trained models outperform their SFT counterparts. In particular, a multi-task Llama4-Scout model trained with shaped rewards surpasses frontier API models with the strongest overall performance across 47 language varieties, reaching agreement levels comparable to human annotators.

Finally, we demonstrate that these judges can be used as generative reward models (RMs) to directly improve a policy model’s proficiency. By using our pairwise RL-trained Qwen3-4B judge to post-train the base Qwen3-4B model, we observe quality gains as measured by both LLM evaluators and human raters. However, LLM evaluators tend to be overconfident in assessing improvements compared to human judgments (+0.6 higher gain). This finding shows that while judges trained with our framework can successfully drive model improvements, the gap between LLM and human raters highlights the remaining challenges in reliably modeling native-like quality across languages.

Our contributions are the following: **1)** We develop MENLO, a framework for the evaluation of native-like response quality in four dimensions, based on principles from audience design, employing parametric templates and carefully crafted annotation guidelines. **2)** We create the MENLO dataset, consisting of 6,423 annotated prompt-response preference pairs in 47 language varieties. **3)** We evaluate zero-shot judges on the annotated data, demonstrating the benefits of pairwise evaluation and rubrics. **4)** We show that multi-task RL and reward shaping enables fine-tuning a judge that is on par with human annotators in 47 language varieties. **5)** We demonstrate that pairwise fine-tuned judges can be used as generative RMs to improve policy model language proficiency, while we find that LLM evaluations tend to overestimate improvements compared to human raters.

Our framework unifies the MENLO dataset, RL-trained pairwise judging, and generative reward modeling, offering a practical and scalable approach to both assess and improve native-like quality.

2 THE MENLO DATASET

MENLO characterizes native-like conversational response quality in a language along four key dimensions: fluency, tone, localized tone, and localized factuality. These dimensions go beyond prior work that focused mainly on naturalness (Novikova et al., 2016; Liu et al., 2021; Guo et al., 2025) and are motivated by work on language proficiency assessment (Ke & Ng, 2019), cross-cultural variation (Hershcoffich et al., 2022; Myung et al., 2024), local knowledge grounding (Hupkes & Bogoychev, 2025). We provide further context on our definition of these dimensions in Figure 2.

From a sociolinguistic perspective, the Style Axiom (Bell, 1984) states that intraspeaker variation (style) reflects interspeaker variation (social). Native-like quality is therefore not a single fixed

Dimension	Definition	Key Question	Example Prompt
Fluency	Language proficiency compared to an expert-level native speaker.	Is the model response coherent, well-versed, clear, and free from grammatical errors?	If you could make one change to the education system in your country, what would it be and why?
Tone	Overall writing style or "voice" of the response.	Is the response helpful, insightful, engaging, and fair?	I'm so lost, I feel like I've been stuck in neutral for weeks.
Localized Tone	Alignment with cultural, regional, and linguistic nuances.	Does the response employ locally relevant expressions and is culturally sensitive?	Imagine you're at an [locale_nationality] friend's family gathering in [locale_country]. How would you politely ask for a second helping of food?
Localized Factuality	Factuality, completeness, and grounding in local context.	Is the response factual, complete, and grounded in the local context?	You are staying with a host family in [locale_country] during [locale_holiday]. They invite you to help with the preparations. What tasks might you be expected to help with, and what do they symbolize?

Figure 2: Dimensions of native-like response quality in MENLO and example prompt (template).

Table 1: Annotation and statistics of MENLO across evaluation dimensions. IAA presents *Krippendorff's α* measuring inter-annotator agreement. Average token counts are computed using Qwen3-4B.

Dimension	# Annotations	# Annotators	# Prompts	Avg # Tokens		Rating (1–5 Scale)		
				Prompt	Response	IAA	Mean	Std.
Fluency	23,556	450	1,820	81.6	804.3	0.82	4.01	1.11
Tone	18,712	429	1,410	27.8	575.3	0.86	3.48	1.35
Localized Tone	22,324	530	1,815	71.7	559.2	0.83	3.89	1.17
Localized Factuality	16,422	525	1,378	121.8	839.1	0.84	3.82	1.16
Overall	81,014	1,934	6,423	75.6	692.2	0.84	3.80	1.20

target but a socially conditioned range of stylistic choices that depend on interlocutors. Key mechanisms include accommodation, where speakers adapt their style to the addressee, and referee design, where speakers align with an absent reference group they wish to identify with. These mechanisms motivate our focus on tone and localized tone as central to native-like quality. To operationalize these ideas, we design human-written parametric English prompt templates for each dimension with placeholders such as [locale_nationality], [locale_country], [locale_holiday], etc. By defining the addressee or reference group, these prompts evoke local contexts and guide models toward contextually appropriate “native” styles. We provide an overview of the MENLO framework and annotation process in Figure 1.

We select 47 language varieties representing a typologically diverse set of widely used languages and their major variants, including, e.g., South American and European varieties of Spanish and Portuguese, several varieties of English, and romanized versions of non-Latin script languages (see Appendix A.2). Native speakers are recruited to professionally translate these prompt templates, with placeholders instantiated using locally relevant entities. As native quality is tied to the local context, we ensure that native speakers are from the specific regions where the corresponding language varieties are spoken. Similar criteria are used to select annotators for each language variety. Each language variety has approximately the same number of examples in MENLO.

To ensure consistency in evaluation, we develop instructions that reduce the subjectivity of the annotation and break down the four broad dimensions into easy-to-follow rubrics and self-explanatory signals (human-written). Annotators receive guidelines with examples for each dimension. We additionally develop a customized annotation tool and annotator screening tests to filter out unreliable annotators. Furthermore, we train 1–2 expert annotators per language who provide language-specific feedback to annotators and provide gold annotations on a subset of examples.

We generate two responses for each prompt with state-of-the-art LLMs including GPT-4o, Llama4-Maverick, Llama4-Maverick with Search, and Gemini 1.5 with Search. We present both responses in randomized order to human annotators and ask them to provide 1–5 Likert ratings per response, allowing ties. Each response pair is annotated by at least 3 annotators, with final scores aggregated via majority vote. Annotators achieve high reliability, with an average Krippendorff's $\alpha = 0.84$.

Overall, MENLO consists of 6,423 annotated prompt-response preference pairs across 47 language varieties, each containing a prompt, two responses, and corresponding scores, totaling 81,014 human annotations. Summary statistics are reported in Table 1. Example prompts per dimension are shown in Figure 2. Further details of MENLO including annotation process, language coverage, rubrics, and full examples featuring responses and their corresponding ratings are provided in Appendix A.

162 Table 2: Comparison of multilingual response quality datasets: RECON (Doddapaneni et al., 2025),
 163 PARIKSHA (Watts et al., 2024), BIGGEN BENCH (Kim et al., 2025), M-REWARD BENCH (Gureja
 164 et al., 2025), MM-EVAL (Son et al., 2024). $|\mathcal{L}|$: # of languages, $|\mathcal{D}|$: # of prompts, IAA: inter-
 165 annotator agreement, IF: instruction following. *: 81,014 annotations; 1,776 test examples.

Dataset	$ \mathcal{L} $	$ \mathcal{D} $	IAA	Dimensions	Prompts	Responses	Ratings
MENLO	47	6,423*	0.84	Fluency, tone, localized tone, localized factuality	Human-written, translated & localized	Annotated in each language	Preference & 1–5
PARIKSHA	10	200	0.54	Hallucinations, task quality, linguistic acceptability	Human-written	Annotated in each language	Preference & 0–2
BIGGEN BENCH	10	420	–	Poem, reasoning, humor, translation, historical text	Human-written & LLM-augmented	Generated in each language	1–5
RECON	6	3,000	–	IF, theory of mind, reasoning, safety, planning, etc.	Translated	Generated in each language	Preference & 1–5
MM-EVAL	18	4,981	–	Reasoning, chat, linguistics, hallucination, safety	Translated	Generated in each language	Preference
M-REWARD BENCH	23	66,787	–	Chat, safety, reasoning, translation	Translated	Translated	Preference

179 Table 2 compares MENLO with existing multilingual preference datasets. MENLO provides local-
 180 ized prompts and responses, spans more languages, and reaches higher agreement than prior work.

3 EVALUATING LLM-JUDGES ON MENLO

185 We next evaluate the ability of LLMs to serve as automatic judges of native-like quality on MENLO.
 186 Out of the 6,423 pairs, we hold out 1,766 pairs (3,552 responses) as the test set,¹ and use the remain-
 187 der for training and prompt development (see §4 and §5). Where expert annotations are available,
 188 we use these as labels. For the remaining responses, we average the annotated ratings of each re-
 189 sponse.² Our evaluation focuses on three questions: (i) how pointwise and pairwise setups compare,
 190 (ii) the effect of few-shot exemplars, and (iii) the role of explicit grading rubrics.

191 We benchmark a range of open-source and API-based models, covering both *thinking* and *non-*
 192 *thinking* variants: Qwen3-4B, Qwen3-32B, Llama-3.1-8B,³ Llama-3.3-70B, Llama4-Scout, o3,
 193 gpt-4o, and gpt-4.1. All models are used in the default setup with maximum output length 8192.

194 We report two primary metrics: (i) *Macro-F1* for 5-way classification, and (ii) *Preference* accuracy
 195 over Win/Loss/Tie outcomes. Note that we *do NOT directly ask for preference judgments; rather, we*
 196 *infer these from the assigned grades*. Additionally, we report classification accuracy, Krippendorff's
 197 α , which measures agreement with human annotators while accounting for chance agreement and
 198 missing data, and provide detailed per-dimension and per-language breakdowns in Appendix E.

3.1 POINTWISE VS. PAIRWISE

201 Although MENLO provides paired responses for each prompt, the presence of detailed *grading*
 202 *rubrics* means that *pointwise* evaluation is in principle sufficient: a model could *assign absolute*
 203 *scores to individual responses without needing comparisons*. However, pairwise setups may provide
 204 stronger relative signals by anchoring judgments against another candidate. We therefore compare
 205 three setups: **Zero-shot pointwise**: the model is given a prompt, a *single* response, and a detailed
 206 5-point grading rubric, and asked to generate evaluation reasoning (in *thinking*) and assign a final
 207 grade; **Few-shot pointwise**: we additionally provide three graded examples: one from 1–2, one
 208 with a grade of 3, and one from 4–5; **Zero-shot pairwise**: the model is presented with *both* re-
 209 sponses to the same prompt and asked to assign a grade to each, following the template in Figure 17
 210 (Appendix C), without constraints on ties. The order of the two responses is randomized.

211 Table 3 reports Macro-F1 and Preference results. Zero-shot pairwise consistently outperforms both
 212 zero-shot and few-shot pointwise scoring across models, with gains of up to +12.4% in Macro-F1

214 ¹Translations of the same prompt template are assigned the same set to prevent train-test leakage.

215 ²Multiple annotations can be used in future work on pluralistic alignment (Sorensen et al., 2024).

³Llama models are instruction-tuned and we omit the Instruct suffix for brevity.

Table 3: Zero-shot and few-shot results of open-source and API models on the MENLO test set using **POINTWISE** (grading single responses) and **PAIRWISE** (grading response pairs) scoring (see §3.1). *Macro-F1* shows 5-way classification performance and *Preference* reports accuracy on Win/Loss/Tie. Reported gains/loss are relative to zero-shot pointwise performance.

MODELS	Macro F1			Preference Accuracy		
	ZERO-SHOT POINTWISE	FEW-SHOT POINTWISE	ZERO-SHOT PAIRWISE	ZERO-SHOT POINTWISE	FEW-SHOT POINTWISE	ZERO-SHOT PAIRWISE
Qwen3-4B	23.06	31.18 <small>+8.12</small>	35.46 <small>+12.40</small>	40.54	39.35 <small>-1.19</small>	57.13 <small>+16.57</small>
Qwen3-32B	28.53	35.45 <small>+6.92</small>	37.48 <small>+8.95</small>	42.19	42.87 <small>+0.68</small>	59.12 <small>+16.59</small>
Llama-3.1-8B	22.27	23.29 <small>+1.02</small>	29.46 <small>+7.19</small>	39.92	37.15 <small>-2.77</small>	50.45 <small>+10.48</small>
Llama-3.3-70B	27.93	30.52 <small>+2.59</small>	37.50 <small>+9.57</small>	37.37	38.56 <small>+1.19</small>	55.32 <small>+17.89</small>
Llama4-Scout	25.63	32.84 <small>+7.21</small>	36.11 <small>+10.48</small>	42.19	41.22 <small>-0.97</small>	56.25 <small>+14.12</small>
o3	26.54	27.92 <small>+1.38</small>	35.35 <small>+8.81</small>	45.07	44.68 <small>-0.39</small>	58.72 <small>+13.68</small>
gpt-4o	25.99	29.57 <small>+3.58</small>	37.57 <small>+11.58</small>	42.92	45.87 <small>+2.95</small>	57.98 <small>+15.09</small>
gpt-4.1	32.23	33.84 <small>+1.61</small>	38.53 <small>+6.30</small>	41.73	44.00 <small>+2.27</small>	59.23 <small>+17.50</small>

Table 4: Zero-shot performance comparing without and with detailed 5-Point Grading Rubrics.

MODELS	Macro F1				Preference Accuracy			
	POINTWISE		PAIRWISE		POINTWISE		PAIRWISE	
	wo/ Rubrics	w/ Rubrics	wo/ Rubrics	w/ Rubrics	wo/ Rubrics	w/ Rubrics	wo/ Rubrics	w/ Rubrics
Qwen3-4B	16.00	23.06 <small>+7.06</small>	32.74	35.46 <small>+2.72</small>	33.52	40.54 <small>+7.02</small>	54.08	57.13 <small>+3.05</small>
Qwen3-32B	25.59	28.53 <small>+2.94</small>	38.10	37.48 <small>-0.62</small>	43.32	42.19 <small>-1.13</small>	59.23	59.12 <small>-0.11</small>
Llama-3.1-8B	21.50	22.27 <small>+0.77</small>	30.89	29.46 <small>-1.43</small>	38.34	39.92 <small>+1.58</small>	49.55	50.45 <small>+0.90</small>
Llama-3.3-70B	22.71	27.93 <small>+5.22</small>	35.12	37.50 <small>+2.38</small>	34.54	37.37 <small>+2.83</small>	56.29	55.32 <small>-0.97</small>
Llama4-Scout	22.15	25.63 <small>+3.48</small>	35.21	36.11 <small>+0.90</small>	41.28	42.19 <small>+0.91</small>	55.10	56.25 <small>+1.15</small>
o3	25.43	26.54 <small>+1.11</small>	35.60	35.35 <small>-0.25</small>	45.13	45.07 <small>-0.06</small>	57.98	58.72 <small>+0.74</small>
gpt-4o	22.45	25.99 <small>+3.54</small>	36.74	37.57 <small>+0.83</small>	37.60	42.92 <small>+5.32</small>	56.85	57.98 <small>+1.13</small>
gpt-4.1	22.26	32.23 <small>+9.97</small>	37.35	38.53 <small>+1.18</small>	38.67	41.73 <small>+3.06</small>	56.96	59.23 <small>+2.27</small>

and +18.0% in Preference accuracy over zero-shot pointwise. Few-shot pointwise improves Macro-F1 relative to zero-shot pointwise but yields only marginal gains in Preference, still falling short of zero-shot pairwise by an *average* of -5.5% in Macro-F1 and -15.1% in Preference across models.

These results indicate that models are substantially more reliable at assigning scores when evaluating two responses side by side, even without ground-truth labels. The unexpectedly large gains over few-shot in-context examples highlight **pairwise evaluation**, which explicitly anchors outputs against a competing candidate (Wang et al., 2025), as a promising direction for improving automated judging reliability, even when the ultimate goal is pointwise scoring. We also evaluate few-shot pairwise on Qwen3-4B, observing only a small gain in Macro-F1 (+0.6) relative to zero-shot pairwise, further supporting our findings. Future work may investigate whether extending pairwise comparisons to a listwise evaluation of multiple responses offers additional benefits.

3.2 WITH AND WITHOUT GRADING RUBRICS

We further examine the role of detailed grading rubrics in judge performance. All rubrics are human-written 5-point guidelines specific to the dimension and question type of each prompt. Examples of dimension-specific rubrics are shown in Appendix A.3.

Table 4 compares zero-shot pointwise and pairwise performance with and without access to rubrics. The latter shows only the five class labels, without accompanying criteria or definitions. Results show that rubrics provide a substantial benefit, especially for pointwise evaluation, yielding average gains of +4.3% in Macro-F1 and +2.5% in Preference accuracy. In contrast, pairwise evaluation benefits more modestly, with improvements of roughly +1% on both metrics.

These findings suggest that judges perform better when grounded, either by explicit rubrics or by comparison with another response. Since pairwise comparison itself offers a strong grounding signal, it sees limited impact from rubrics. This highlights the importance of high-quality rubrics: if judges could automatically generate and evaluate high-quality, detailed rubrics, we hypothesize that the performance gap between pairwise and pointwise evaluation would further narrow.

270
271
272
273
274
275
276
277
278
Table 5: PAIRWISE SFT, RL, and SFT+RL-trained
Qwen3-4B and Llama4-Scout results. RL-trained
models perform best overall.

PAIRWISE	Qwen3-4B		Llama4-Scout	
	Marco-F1	Preference	Marco-F1	Preference
ZERO-SHOT	35.46	57.13	36.11	56.25
SFT	33.44 _{-2.02}	53.68 _{-3.45}	44.17 _{+8.06}	60.08 _{+3.83}
RL	39.44 _{+3.98}	60.02 _{+2.89}	45.62 _{+9.51}	62.60 _{+6.35}
SFT + RL	39.33 _{+3.87}	58.78 _{+1.65}	45.82 _{+9.71}	61.10 _{+4.85}

279 4 TRAINING LLM-JUDGES ON MENLO

280
281 Having established in §3 that pairwise evaluation yields substantial advantages over pointwise scoring,
282 we next examine whether training LLMs as judges can further close the gap to human annotators.
283 We train on the MENLO training split (total 4,675 response pairs, where 232 pairs are held out
284 for validation) and explore different learning strategies, model families, and reward designs. Inspired
285 by the success of recent reasoning-based judges such as J1 (Whitehouse et al., 2025), we compare
286 supervised fine-tuning (SFT) and reinforcement learning (RL), as well as single-task (dimension-
287 specific) and multi-task (all dimensions) training.

288 We fine-tune two contrasting models: Qwen3-4B (dense, reasoning-oriented) and Llama4-Scout
289 (Mixture-of-Experts, non-reasoning), which differ in architecture and cognitive approach. For **SFT**,
290 models directly predict 5-point grades using cross-entropy loss under teacher forcing, without inter-
291 mediate reasoning generation. For **RL**, we use GRPO (Shao et al., 2024) with the template from Figure 17,
292 encouraging step-by-step reasoning before score assignment. Following Whitehouse et al.
293 (2025), we augment training data by including both response orders (A,B) and (B,A) to mitigate
294 positional bias. Training details are provided in Appendix D.1.

295 4.1 REWARD DESIGNS FOR RL

296 To make RL training effective, we design a composite reward signal that combines absolute accuracy
297 with relative preference alignment and robustness to near-miss predictions: (i) **Pointwise binary re-**
298 **ward**: +1 if the predicted score matches the gold label, 0 otherwise. (ii) **Reward smoothing**: partial
299 reward (+0.5) if the prediction differs by exactly one grade. (iii) **Preference bonus**: additional +1
300 if the *sign* of the difference between the two predicted scores matches the label. (iv) **Penalties**: −1
301 for invalid or missing scores, and −0.2 for formatting violations, i.e. each tag must appear in the
302 correct order and only once.

303 All reward components are summed to produce the final RL signal. Formally, the reward can be
304 expressed as follows, where s and gt represent predicted and ground truth grades, respectively:

$$305 R = \sum_{i \in \{A, B\}} \underbrace{\max \left(\mathbf{1}[s_i = gt_i], 0.5 \cdot \mathbf{1}[|s_i - gt_i| = 1] \right)}_{\text{pointwise binary reward w/ reward smoothing}} + \underbrace{\mathbf{1}[\text{sign}(s_A - s_B) = \text{sign}(gt_A - gt_B)]}_{\text{preference bonus}} \\ 306 - \underbrace{\mathbf{1}[\text{failed extraction}]}_{\text{extraction penalty}} - \underbrace{0.2 \cdot \mathbf{1}[\text{formatting violation}]}_{\text{format penalty}}.$$

313 4.2 OVERALL PERFORMANCE: SFT vs. RL

314 We first compare the overall performance of SFT and RL-trained models. Table 5 shows that RL-
315 trained Qwen3-4B and Llama4-Scout consistently outperform their SFT counterparts. For inherently
316 thinking models like Qwen3-4B, SFT without Chain-of-Thought (CoT) reasoning actually hurts per-
317 formance, causing a −2.0% drop in Macro-F1 and −3.5% in Preference accuracy. In contrast, RL,
318 which incentivizes reasoning, improves performance by +4.0% in Macro-F1 and +2.9% in Prefer-
319 ence, surpassing the best frontier API model gpt-4.1.

320 For non-thinking models like Llama4-Scout, SFT already provides substantial gains (+8.1% in
321 Macro-F1 and +3.8% in Preference) compared to zero-shot. RL training further improves results,
322 particularly in Preference (+2.5%). This demonstrates the promise of pairwise RL training across
323 model families, scales, and reasoning capabilities.

Table 6: Ablation of different reward designs
for PAIRWISE RL-trained Qwen3-4B. *Smooth*.
and *Prefer*: refer to *Reward Smoothing* and
Preference Bonus. See §4.1 for details.

REWARDS	Marco-F1	Preference
<i>Binary Only</i>	37.11	58.27
<i>Binary + Smooth</i> .	37.30 _{+0.19}	51.47 _{-6.80}
<i>Binary + Prefer</i> .	37.05 _{-0.06}	60.48 _{+2.21}
<i>Binary + Smooth. + Prefer</i> .	39.44 _{+2.33}	60.02 _{+1.75}

We also experimented with initializing RL from the best SFT checkpoint, but observed little or no improvement over starting RL from scratch. Models trained on SFT without CoT tend to copy the placeholder “*<think> Your analysis and reasoning here. </think>*” from the prompt rather than generating meaningful reasoning, which limits the benefit of RL. This suggests that for tasks requiring reasoning, it is preferable to start RL directly when the SFT target lacks CoT supervision.

4.3 ABLATION OF RL REWARDS

In [Table 6](#), we ablate the RL reward design to validate the contribution of each reward component in RL training: (i) *binary only*: reward +1 for exact score match, 0 otherwise; (ii) *binary+smooth.*: adds partial reward for near-miss scores, no preference bonus; (iii) *binary+prefer.*: includes preference reward, no smoothing; and (iv) *binary+smooth.+prefer.*: the default reward design in [§4.1](#). Results show clear benefits from combining reward smoothing and preference bonus, achieving the best overall Macro-F1 and Preference accuracy for Qwen3-4B, achieving +2.3 boost on Macro-F1 and +1.7 on preference accuracy over the *binary only* reward.

4.4 PER-DIMENSION PERFORMANCE AND SINGLE VS. MULTI-TASK

Next, we compare pairwise RL-trained Qwen3-4B models trained jointly across all dimensions versus individually per dimension. Across the four dimensions, *Tone* achieves the strongest performance, with a Macro-F1 of 43.1 in the zero-shot setting and gains of up to +3.8 with multitask RL. *Localized Tone* and *Fluency* follow, reaching 32.8 and 32.2 Macro-F1 in zero-shot, and up to +5.7 improvement when trained with multitask RL. In contrast, *Localized Factuality* lags behind the other dimensions, achieving only 22.5 Macro-F1 in zero-shot, a trend consistent across all models. Moreover, RL yields limited benefit (+0.6 in single-task RL) or even regressions in the multitask setup. These results highlight the challenge of localized factuality and suggest that alternative strategies, such as incorporating retrieval, search, or external tool use, may be necessary. Full results are provided in [Table 20](#) in Appendix E.

Overall, aside from *Localized Factuality*, joint multi-dimension training performs on par with single-dimension optimization while offering greater efficiency and practical benefits, such as serving as a reward model for post-training, which we explore in [§5](#).

4.5 CROSS-LANGUAGE PERFORMANCE

[Figure 3](#) shows Preference accuracy per language variety for RL-trained Qwen3-4B. Performance varies widely, with *tr_TW* at 82.1% and *bn_BD* at 37.9%, and does not strictly align with high- vs. low-resource languages. Relative to the zero-shot baseline, *en_AU* and *fr_FR* achieve the largest Macro-F1 gains (+20.9%, +17.7%) and *ro_RO* and *gu_IN* the largest Preference accuracy gains (+18.0%, +16.2%). By contrast, *es_ES* drops -15.4% in Preference despite a modest +2.2% Macro-F1 gain, whereas *en_MX*, the same language but a different locale, sees +2.6% and +9.8% gains in Preference and Macro-F1, highlighting that our dataset captures language variety nuances.

We further trained RL using only English data and evaluated on all languages. Performance degrades compared to the baseline, indicating that English-only training is insufficient to generalize across all 47 language varieties. Detailed per-language variety performance is provided in [Appendix E.2](#).

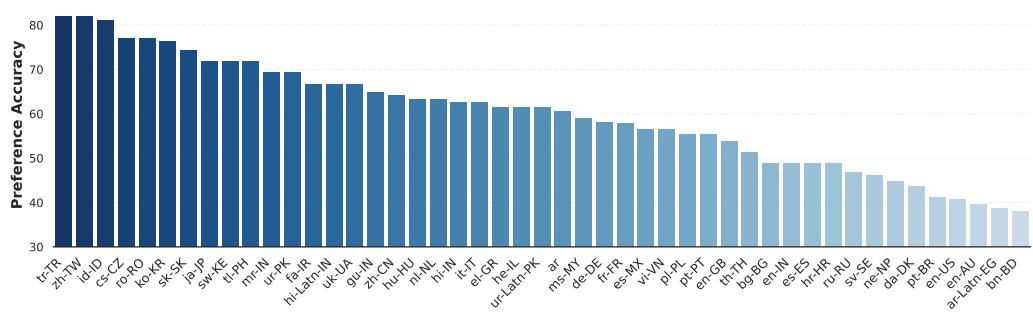


Figure 3: Preference Accuracy per Language of *pairwise* RL-trained Qwen3-4B.

378 Table 7: Two-stage Evaluation of Qwen3-4B and its RL post-trained variant Post-train on the
 379 MENLO test set, where both models serve as *response models*.

380 JUDGES/RATERS	# Languages	Win Rate			Average Score (1-5)			
		381 Post-train Win	Post-train Loss	Tie	382 Qwen3-4B	Post-train	ΔS_{ore}	Improvement%
Llama4-Scout-RL-Judge	47	63.88%	9.16%	26.96%	3.01	3.79	+0.78	+25.9%
Qwen3-32B	47	72.46%	21.89%	6.65%	3.44	4.29	+0.85	+24.7%
gpt-4.1	47	77.90%	11.30%	10.80%	3.21	4.37	+1.16	+36.1%
Llama4-Scout-RL-Judge	10	69.66%	9.64 %	20.70%	3.36	4.22	+0.86	+25.6%
Human Raters	10	55.71%	35.20%	9.09%	3.31	3.67	+0.36	+10.9%

386 387 5 FROM LLM-JUDGES TO REWARD MODELS

388 We next investigate whether the RL-trained pairwise judges developed in our framework can also
 389 serve as generative reward models to directly improve LLM native-like response quality, unifying
 390 evaluation and optimization in a single framework.

391 393 5.1 RL WITH JUDGES AS GENERATIVE REWARD MODELS

394 For efficiency, we focus on smaller models for these experiments: Qwen3-4B as the policy model,
 395 and Qwen3-4B-RL-Judge as the reward model (RM). Since Localized Factuality remains challeng-
 396 ing for our judges, we restrict both training and evaluation to Fluency, Tone, and Localized Tone.
 397 Specifically, we exclude Localized Factuality from all training and test prompts, randomly sam-
 398 ple 3,000 prompts from MENLO for training, and retain all 1,398 test prompts from MENLO for
 399 evaluation across the three selected dimensions.

400 We post-train Qwen3-4B with GRPO. We sample 8 rollouts per prompt and compute rewards as
 401 follows: for each prompt, we construct response pairs from the rollouts, format them with the same
 402 pairwise evaluation template, and feed them to the RM. The final reward of each rollout is obtained
 403 by averaging its scores across all paired comparisons. Training details are added in Appendix D.2.
 404

405 406 5.2 TWO-STAGE EVALUATION STRATEGY

407 To rigorously evaluate the policy model’s native-like quality improvements, we employ a two-stage
 408 validation approach: (i) comprehensive automated evaluation across all 47 language varieties using
 409 three diverse LLM judges, and (ii) human validation on a strategically selected subset of 10 high-
 410 resource languages where we can ensure annotation quality.

411 For each test prompt, we generate responses from both the baseline (Qwen3-4B) and post-trained
 412 (Post-train) models, construct response pairs with randomized order to mitigate positional bias,
 413 and apply the same pairwise judge template used in training.

414 **LLM-Judges Evaluation** We select three high-performing judges (see §3) Qwen3-32B, gpt-4.1,
 415 and Llama4-Scout-RL-Judge, and compute win, loss, and tie rates between baseline and post-
 416 trained models, along with average scores on a 1-5 scale across all 1,398 test prompts spanning 47
 417 language varieties. Qwen3-4B-RL-Judge is excluded from evaluation to avoid potential bias, since
 418 it serves as the RM. Table 7 shows that the post-trained policy model consistently outperforms the
 419 baseline across all LLM judges and languages. Average score improvements range from +0.80
 420 to +1.16, with win rates between 63.4% and 77.9%. Per-dimension analysis reveals consistent
 421 gains across evaluation criteria: Tone yields the largest improvement (+1.04 average score boost),
 422 followed by Localized Tone and Fluency (+0.89 each). The consistency of improvements across
 423 different judge architectures and all three dimensions provides strong evidence for the effectiveness
 424 of our reward modeling approach.

425 **Human Validation** To anchor our automated evaluation results, we conduct human evaluation on
 426 a diverse subset of 10 higher-resource languages: ar, de_DE, en_US, fr_FR, hi_IN, hi_Latn_IN,
 427 pt_BR, tl_PH, th_TH, vi_VN. This subset spans multiple language families, scripts, and geographic
 428 regions while ensuring access to qualified native speaker annotators. Human evaluation follows the
 429 same pairwise annotation guidelines as in MENLO construction. Results (last row of Table 7) on
 430 the subset confirms the automated evaluation trends. The post-trained model achieves a win rate of
 431 55.7% against the baseline, with an average score improvement of +10.9%. While both automated
 and human evaluators agree that post-training improves response quality, we observe that LLM

432 judges tend to overestimate the magnitude of improvement compared to human raters. Comparing
 433 human evaluations to the closest-performing automated judge (Llama4-Scout-RL-Judge) on this
 434 subset reveals systematic differences: the automated judge reports an average improvement of +0.5
 435 higher than humans. We hypothesize that this discrepancy arises because the automated judges may
 436 lean towards a stylistic caricature of native-like quality, overestimating improvements relative to
 437 nuanced human judgments. In addition, RL-trained judges exhibit less of this discrepancy among
 438 LLM evaluators, confirming the benefits of our RL judge training.

439 Overall, our two-stage evaluation demonstrates the potential of RL-trained judges as generative
 440 reward models for aligning multilingual outputs toward native-like quality. The directional consis-
 441 tency observed across both LLM- and human-based evaluations validates the viability of our unified
 442 framework for multilingual proficiency alignment. However, we note that challenges remain: LLM
 443 judges tend to overestimate the magnitude of improvements relative to human raters, highlighting
 444 an important direction for future work.

446 6 RELATED WORK

447 **Multilingual Evaluation** Models' multilingual proficiency has been typically measured as an ag-
 448 gregate of performance across multiple task-oriented evaluations of short-form responses in settings
 449 with verifiable answers (Hu et al., 2020; Ruder et al., 2021; Doddapaneni et al., 2023; Ahuja et al.,
 450 2023; 2024). Recent benchmarks focused on the evaluation of model's cultural knowledge in a sim-
 451 ilar verifiable setting (Myung et al., 2024; Chiu et al., 2025; Fabbri et al., 2025). However, such
 452 evaluations do not extend to real-world conversations containing long-form responses. Benchmarks
 453 evaluating long-form responses use prompts and responses translated from English (Son et al., 2024;
 454 Liu et al., 2024; Doddapaneni et al., 2025; Gureja et al., 2025). These evaluations typically do not
 455 reflect more localized aspects of language quality and are biased towards translationese. Son et al.
 456 (2024) and Doddapaneni et al. (2025) automatically generate 'good' and 'bad' responses for each
 457 dimension. Marchisio et al. (2024) and Guo et al. (2025) evaluate language consistency and natu-
 458 ralness respectively in relatively narrow settings. PARIKSHA (Watts et al., 2024) is the most similar
 459 dataset to ours as it uses human-written prompts and human-annotated responses, but focuses on 10
 460 Indic languages, annotates only high-level dimensions, and reports moderate inter-annotator agree-
 461 ment. MENLO is the only dataset that focuses on native-like quality in real-world conversations.

462 **Multilingual Judges and RMs** LLMs have been used as judges in different multilingual bench-
 463 marks (Liu et al., 2024; Fabbri et al., 2025). However, fewer works focus on analyzing or improving
 464 multilingual judges and RMs. Gureja et al. (2025) observe that zero-shot judges show a substantial
 465 gap between the translated M-REWARDBENCH and its English counterpart, with predictions incon-
 466 sistent across languages. Fu & Liu (2025) report similar inconsistencies across five diverse tasks.
 467 Wu et al. (2024) evaluate zero-shot cross-lingual transfer of trained RMs on summarization and di-
 468 alog, observing gains. Hong et al. (2025) find strong cross-lingual transfer on M-REWARDBENCH
 469 for English RMs fine-tuned in four languages. Doddapaneni et al. (2025) fine-tune a judge with SFT
 470 on automatically translated prompts and responses in six languages to produce an absolute score. To
 471 our knowledge, we are the *first* to (i) train judges and RMs in a massively multilingual setting, (ii)
 472 fine-tune multilingual judges with RL, and (iii) demonstrate the benefits of multi-task RL, reward
 473 shaping, and pairwise grading in this setting.

474 7 CONCLUSION

475 We introduce MENLO, a comprehensive framework for evaluating and improving native-like re-
 476 sponse quality across 47 language varieties. By combining sociolinguistically-informed prompt
 477 design, detailed evaluation rubrics, and high-quality human annotations, MENLO captures multi-
 478 ple dimensions of conversational proficiency, including fluency, tone, localized tone, and localized
 479 factuality. We demonstrate that pairwise evaluation significantly improves both zero-shot and fine-
 480 tuned LLM judges, and that RL with reward shaping yields best judge performance.

481 Beyond evaluation, we show these trained judges can serve as generative reward models to directly
 482 improve policy model's response quality. While challenges remain with the tendency of LLM judges
 483 to overestimate improvements relative to human raters, our framework provides a practical and
 484 scalable approach to both assessing and enhancing LLM proficiency in multilingual context.

486 ETHICS STATEMENT

487

488 Translators and annotators were recruited through third-party services and compensated based on
489 local regulations.

490

491 REPRODUCIBILITY STATEMENT

492

493 We release the full MENLO to the community. All models are built on top of open-weight Llama and
494 Qwen backbones. The prompt templates used for training are provided in [Appendix C](#), and detailed
495 descriptions of experimental setups, hyperparameters, and libraries are included in [Appendix D](#).

496

497 REFERENCES

498

499 Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Ak-
500 shay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram.
501 MEGA: Multilingual evaluation of generative AI. In Houda Bouamor, Juan Pino, and Kalika
502 Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language
503 Processing*, pp. 4232–4267, Singapore, December 2023. Association for Computational Lin-
504 guistics. doi: 10.18653/v1/2023.emnlp-main.258. URL <https://aclanthology.org/2023.emnlp-main.258/>.

505

506 Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent
507 Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram.
508 MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models
509 and Tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024
510 Conference of the North American Chapter of the Association for Computational Linguistics:
511 Human Language Technologies (Volume 1: Long Papers)*, pp. 2598–2637, Mexico City, Mexico,
512 June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.143.
513 URL <https://aclanthology.org/2024.naacl-long.143/>.

514

515 Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos,
516 Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 Technical Report.
517 *arXiv preprint arXiv:2305.10403*, 2023. URL <https://arxiv.org/abs/2305.10403>.

518

519 Allan Bell. Language Style as Audience Design. *Language in society*, 13(2):145–204, 1984.

520

521 Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi,
522 Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. Cultural-
523 Bench: A Robust, Diverse and Challenging Benchmark for Measuring LMs’ Cultural Knowl-
524 edge Through Human-AI Red-Teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova,
525 and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Associa-
526 tion for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Aus-
527 tria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi:
528 10.18653/v1/2025.acl-long.1247. URL <https://aclanthology.org/2025.acl-long.1247/>.

529

530 Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra,
531 Anoop Kunchukuttan, and Pratyush Kumar. Towards Leaving No Indic Language Behind: Build-
532 ing Monolingual Corpora, Benchmark and Models for Indic Languages. In Anna Rogers, Jordan
533 Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Associa-
534 tion for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, Toronto, Canada,
535 July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL
536 <https://aclanthology.org/2023.acl-long.693/>.

537

538 Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop
539 Kunchukuttan, and Mitesh M Khapra. Cross-Lingual Auto Evaluation for Assessing Multilingual
540 LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
541 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics
542 (Volume 1: Long Papers)*, pp. 29297–29329, Vienna, Austria, July 2025. Association for Com-
543 putational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1419. URL
544 <https://aclanthology.org/2025.acl-long.1419/>.

540 Alexander R Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee,
 541 Bing Liu, and Chen Xing. Multirc: A challenging and native multilingual reasoning evaluation
 542 benchmark for llms. *arXiv preprint arXiv:2507.17476*, 2025. URL <https://arxiv.org/abs/2507.17476>.

543

544 Xiyan Fu and Wei Liu. How Reliable is Multilingual LLM-as-a-Judge? *arXiv preprint*
 545 *arXiv:2505.12201*, 2025. URL <https://arxiv.org/abs/2505.12201>.

546

547 Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. Do large lan-
 548 guage models have an English accent? evaluating and improving the naturalness of multilingual
 549 LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
 550 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*
 551 (*Volume 1: Long Papers*), pp. 3823–3838, Vienna, Austria, July 2025. Association for Com-
 552 putational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.193. URL
 553 <https://aclanthology.org/2025.acl-long.193/>.

554

555 Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti
 556 Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh
 557 Fadaee. M-RewardBench: Evaluating Reward Models in Multilingual Settings. In Wanxiang Che,
 558 Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the*
 559 *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 560 pp. 43–58, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-
 561 8-89176-251-0. doi: 10.18653/v1/2025.acl-long.3. URL <https://aclanthology.org/2025.acl-long.3/>.

562

563 Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie
 564 Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza
 565 Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and Strategies in Cross-
 566 Cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings*
 567 *of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*
 568 *Papers)*, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics.
 569 doi: 10.18653/v1/2022.acl-long.482. URL <https://aclanthology.org/2022.acl-long.482/>.

570

571 Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. Cross-
 572 lingual Transfer of Reward Models in Multilingual Alignment. In Luis Chiruzzo, Alan Ritter,
 573 and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter*
 574 *of the Association for Computational Linguistics: Human Language Technologies (Volume 2:*
 575 *Short Papers)*, pp. 82–94, Albuquerque, New Mexico, April 2025. Association for Computational
 576 Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.8. URL <https://aclanthology.org/2025.naacl-short.8/>.

577

578 Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.
 579 Xtrme: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual General-
 580 isation. In *International conference on machine learning*, pp. 4411–4421. PMLR, 2020. URL
 581 <https://proceedings.mlr.press/v119/hu20b>.

582

583 Dieuwke Hupkes and Nikolay Bogoychev. MultiLoKo: A Multilingual Local Knowledge Bench-
 584 mark for LLMs Spanning 31 Languages. *arXiv preprint arXiv:2504.10356*, 2025. URL
 585 <https://arxiv.org/abs/2504.10356>.

586

587 Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas
 588 Haas, Michelle Liu, Nate Keating, Adam Bloniarz, et al. The FACTS Grounding Leader-
 589 board: Benchmarking LLMs’ Ability to Ground Responses to Long-Form Input. *arXiv preprint*
 590 *arXiv:2501.03200*, 2025. URL <https://arxiv.org/abs/2501.03200>.

591

592 Joan Jamieson, Stan Jones, Irwin Kirsch, Peter Mosenthal, and Carol Taylor. TOEFL 2000 Frame-
 593 work. *Princeton, NJ: Educational Testing Service*, 2000.

594

595 Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*,
 596 volume 19, pp. 6300–6308, 2019.

594 Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin
 595 Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo,
 596 Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo,
 597 Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen
 598 Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. The BiGGen
 599 bench: A principled benchmark for fine-grained evaluation of language models with language
 600 models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference*
 601 *of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*
 602 *Language Technologies (Volume 1: Long Papers)*, pp. 5877–5919, Albuquerque, New Mexico,
 603 April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/
 604 v1/2025.naacl-long.303. URL <https://aclanthology.org/2025.naacl-long.303/>.

605 Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun
 606 Chen, Maosong Sun, and Erhong Yang. OMGEval: An Open Multilingual Generative Evaluation
 607 Benchmark for Large Language Models. *arXiv preprint arXiv:2402.13524*, 2024. URL <https://arxiv.org/abs/2402.13524>.

608 Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. Naturalness evaluation of natural lan-
 609 guage generation in task-oriented dialogues using BERT. In Ruslan Mitkov and Galia Angelova
 610 (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language*
 611 *Processing (RANLP 2021)*, pp. 839–845, Held Online, September 2021. INCOMA Ltd. URL
 612 <https://aclanthology.org/2021.ranlp-1.96/>.

613 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee,
 614 and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. *arXiv preprint*
 615 *arXiv:2503.20783*, 2025. URL <https://arxiv.org/abs/2503.20783>.

616 Cedric Lothritz and Jordi Cabot. Testing Low-Resource Language Support in LLMs Using Lan-
 617 guage Proficiency Exams: the Case of Luxembourgish. *arXiv preprint arXiv:2504.01667*, 2025.
 618 URL <https://arxiv.org/abs/2304.01667>.

619 Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Under-
 620 standing and Mitigating Language Confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal,
 621 and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Nat-*
 622 *ural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Associa-
 623 *tion for Computational Linguistics*. doi: 10.18653/v1/2024.emnlp-main.380. URL <https://aclanthology.org/2024.emnlp-main.380/>.

624 Marina Mayor-Rocher, Nina Melero, Elena Merino-Gómez, María Grandury, Javier Conde, and Pe-
 625 dro Reviriego. Evaluating Large Language Models with Tests of Spanish as a Foreign Language:
 626 Pass or Fail? *arXiv preprint arXiv:2409.15334*, 2024. URL <https://arxiv.org/abs/2409.15334>.

627 Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsu-
 628 vas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Ba-
 629 sulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park,
 630 Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma
 631 Ousidhoum, Jose Camacho-Collados, and Alice Oh. BLEnD: A Benchmark for LLMs on Ev-
 632 eryday Knowledge in Diverse Cultures and Languages. In *The Thirty-eight Conference on*
 633 *Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=nrEqH502eC>.

634 Jekaterina Novikova, Oliver Lemon, and Verena Rieser. Crowd-sourcing NLG data: Pictures elicit
 635 better data. In Amy Isard, Verena Rieser, and Dimitra Gkatzia (eds.), *Proceedings of the 9th*
 636 *International Natural Language Generation conference*, pp. 265–273, Edinburgh, UK, September
 637 5–8 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6644. URL <https://aclanthology.org/W16-6644/>.

638 Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu,
 639 Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards More
 640 Challenging and Nuanced Multilingual Evaluation. In Marie-Francine Moens, Xuanjing Huang,
 641 Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical*

648 *Methods in Natural Language Processing*, pp. 10215–10245, Online and Punta Cana, Dominican
 649 Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.
 650 emnlp-main.802. URL <https://aclanthology.org/2021.emnlp-main.802/>.

651
 652 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 653 Mingchuan Zhang, YK Li, Y Wu, et al. DeepseekMath: Pushing the Limits of Mathematical
 654 Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL
 655 <https://arxiv.org/abs/2402.03300>.

656 Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim,
 657 Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. MM-Eval:
 658 A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models. *arXiv
 659 preprint arXiv:2410.17578*, 2024. URL <https://arxiv.org/abs/2410.17578>.

660 Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christo-
 661 pher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin
 662 Choi. A Roadmap to Pluralistic Alignment. In *Forty-first International Conference on Machine
 663 Learning*, 2024. URL <https://openreview.net/forum?id=gQpBnRHwxM>.

664
 665 Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-Judge Inference with the
 666 Judgment Distribution. *arXiv preprint arXiv:2503.03064*, 2025. URL <https://arxiv.org/abs/2503.03064>.

667
 668 Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and
 669 Sunayana Sitaram. PARIKSHA: A Large-Scale Investigation of Human-LLM Evaluator Agree-
 670 ment on Multilingual and Multi-Cultural Data. In Yaser Al-Onaizan, Mohit Bansal, and Yun-
 671 Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Lan-
 672 guage Processing*, pp. 7900–7932, Miami, Florida, USA, November 2024. Association for Com-
 673 putational Linguistics. doi: 10.18653/v1/2024.emnlp-main.451. URL [https://aclanthology.org/2024.emnlp-main.451/](https://aclanthology.org/2024.emnlp-main.451).

674
 675 Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep
 676 Saha. J1: Incentivizing Thinking in LLM-as-a-Judge via Reinforcement Learning. *arXiv preprint
 677 arXiv:2505.10320*, 2025. URL <https://arxiv.org/abs/2505.10320>.

678
 679 Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. Reuse Your
 680 Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment. In Yaser Al-Onaizan,
 681 Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical
 682 Methods in Natural Language Processing*, pp. 1332–1353, Miami, Florida, USA, November
 683 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.79. URL
 684 [https://aclanthology.org/2024.emnlp-main.79/](https://aclanthology.org/2024.emnlp-main.79).

685 Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. HY-
 686 DRA: Model Factorization Framework for Black-Box LLM Personalization. In *The Thirty-
 687 eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=CKgNgKmHyp>.

688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701

APPENDIX

A ADDITIONAL DETAILS ON THE MENLO DATASET

A.1 DATASET COLLECTION

Localization vs natively written prompts Our prompts derive from English templates to ensure coverage consistency across a large number of languages. While localization includes cultural adaptation, we acknowledge that independently authored prompts in each language would better capture native discourse structures. We focus our resources on native speakers annotating responses instead.

We initiated a pilot to evaluate of different models including GPT-4o and Llama4-Maverick in a single category, Localized Tone focusing on five languages spoken by the authors: Bengali, German, Hindi, Italian, and Russian. As expected, the pilot framework was quickly confronted with the complexities inherent in multilingualism: Even among the authors, for all initial Localized Tone prompts, we struggled to reach consistent and reliable agreement. Nevertheless, these early results provided valuable insights to guide us to improve prompt design, guideline clarity, and annotation arrangement.

To enhance the reliability of the framework, we took steps to refine our prompts’ nuance and complexity, update guidelines with clearer direction for annotators that aimed to make abstract concepts more concrete, and started exploring a more user-friendly annotation solution. The changes brought upon notable improvements in inter-annotator agreement that extended to the full-scale annotation. We show the agreement of the initial pilot annotation and improved annotation for localized tone in 5 languages in [Table 9](#) and across all categories and languages in [Table 9](#).

Table 8: Comparison of PILOT and MENLO for 5 languages in localized tone category. *Agreement* is defined as the percentage of annotation pairs whose ratings for the same item differ by no more than 1.

Language Code	PILOT Agreement	MENLO Agreement
bn.BD	0.75	0.84
de.DE	0.74	0.92
hi.IN	0.74	0.92
it.IT	0.79	0.71
ru.RU	0.71	0.79
Overall	0.75	0.84

Table 9: Comparison of the pilot annotation (5 languages) and final MENLO dataset (47 languages). *Agreement* is defined as the percentage of annotation pairs whose ratings for the same item differ by no more than 1. Agreement for PILOT has been averaged over 5 languages, while MENLO is averaged over 47.

Quality Dimension	PILOT (5 Languages)			MENLO (47 Languages)		
	Agreement	# Prompts	# Annotations	Agreement	# Prompts	# Annotations
Fluency	0.76	150	450	0.82	1,820	23,556
Tone	0.70	150	450	0.77	1,410	18,712
Localized Tone	0.75	200	600	0.82	1,825	22,324
Localized Factuality	0.78	150	450	0.78	1,378	16,422
Overall	0.75	650	1,950	0.80	6,423	81,014

756

757 Table 11: Localized factuality rubrics for annotation using a 5-point Likert scale. Each rating cor-
758 responds to a high-level classification of the response (e.g., “*Sounds somewhat accurate and relevant*”), further specified by dimension-specific criteria e.g., accuracy, relevance, and completeness.
759

1: Major failure “Grossly incorrect or misleading”	2: Minor failure “Some mistakes”	3: Pass “Sounds somewhat accurate and relevant”	4: Good “Sounds accurate and relevant”	5: Excellent “Factually accurate, highly relevant, complete with additional info”
Accuracy ✗✗ - The model’s response contains obvious factual errors or made-up information.	Accuracy ✗? ? - The model’s response contains some factual mistakes.	Accuracy OK - No obvious factual errors but some claims are not entirely correct or may be misleading.	Accuracy ✓ The claims in the response are factually accurate.	Accuracy ✓✓ The claims in the response are completely factually accurate.
Locale Relevance ✗✗ Local Point of View ✗✗ - The model fails to understand the basic local context. - The model provides content that is irrelevant or misaligned with the local context. - The model’s response frames the answer in a fetishizing/offensive way (like overly explaining basic local knowledge to locals)	Locale Relevance ✗? ? Local Point of View OK - The model grasps some local context but misses key nuances. - The model’s response is somewhat relevant but provides mainly general or high-level information that lacks alignment with the local context. - The model’s response may come across as slightly insensitive or tone-deaf, but it does not contain overtly fetishizing or offensive answers.	Locale Relevance OK Local Point of View OK - The model generally understands the local context but may miss subtle nuances. - The response is generally relevant and aligned with the local context. - The model avoids explanations that might be seen as overly simplistic or patronizing. Instead, the facts are thoughtfully selected with depth.	Locale Relevance ✓ Local Point of View ✓ - The model accurately interprets the local context and nuances. - The response is generally relevant and aligned with the local context. - The model’s response is neutral and factual.	Locale Relevance ✓✓ Local Point of View ✓✓ - The model demonstrates a deep understanding of the local context and nuances. The response delivers highly relevant content that is highly specific and perfectly aligned with the local context. - The model chooses facts that are in-depth and nuanced even for someone who’s already a local. It might present additional context and highlights regional variations to show depth of local knowledge.
Completeness ✗✗ - The model’s response is incomplete and misses crucial information to answer the question.	Completeness ✗? ? - The response answers part of the question but is missing some relevant pieces of information.	Completeness OK - The model provides sufficient information to answer the question but the provided information may lack depth.	Completeness ✓ - The model provides all the information to answer the question.	Completeness ✓✓ - The response is rich in information and covers all information to answer the question as well as additional helpful context that further helps to contextualize the response.

777

778

779

780

781

782

783

784

785

786

787

Table 10: Components to consider when annotating different subcategories of Tone.

Tone Subcategory	Tone Component 1	Tone Component 2
Helpful Tone	Instruction following ✓	Emotional support ✓
Insightful Tone	Informative ✓	Empathetic ✓
Engaging Tone	Conversational Language ✓	Encourages Interactions ✓
Fair Tone	Non-biased stance ✓	Non-preachy language ✓

793

794

Annotation guidelines Judging language performance can be subjective. To minimize confusion, we identified the most important components of tone, fluency, localized tone, and localized factuality and incorporated them into the guidelines. For example, a model response that conveys a helpful tone must succeed on two fronts: providing (or attempting to provide) help based on users’ instructions, and expressing emotional engagement to sound caring. By breaking down broad linguistic concepts into easy-to-follow subcategories and self-explanatory signals (illustrated via emoji ✓ (OK) ✗ (✗) ?), annotators can quickly grasp and refer back to the guidelines. We show subcategories for Tone, for example, in [Table 10](#) and show the rubric guidelines for Localized Factuality in [Table 11](#).

803

804

805

806

807

808

809

Annotation tool To streamline the annotation, we developed a custom annotation interface, which we show in [Figure 4](#). The tool provides a simple, annotator-friendly user interface for guidelines, rating model responses, and randomized model A/B pairwise comparison. The backend allowed us to ensure data is consistent and identify any missing annotations or other data-related issues. In addition, it enabled us to quickly test annotators on dedicated test annotations before moving them to the actual annotation tasks. Overall, solid tooling allowed us to screen more than 1,000 annotators and collect more than 80,000 annotations.

Challenge set: Accuracy, Relevance, Completeness

METADATA INFORMATION

Prompt ID: G-R3-LK002

Criteria: Accuracy, Relevance, Completeness

CRITERIA EXPLANATION

We define the Localized Factuality of a model response as the AI model's ability to provide a response that is factual, complete, and grounded in the local context.

To assess localized factuality, the evaluation will be based on three criteria:

Accuracy:
Are the claims in the response factually correct?

Relevance:
Is the response relevant to the local context, and written from a local point of view?

Completeness:
Does the response contain all information necessary - even additional context - to answer the question?

Rate Model A's performance in terms of 'Accuracy, Relevance, Completeness'

No option selected

Rate Model B's performance in terms of 'Accuracy, Relevance, Completeness'

No option selected

Which one fulfills the 'Accuracy, Relevance, Completeness' better?

No option selected

Additional comments

Words: 0

Chat history

Model A	Model B
<p>Are there any modes of transportation that is viewed as problematic in USA? What are problems that only locals know?</p>	<p>Are there any modes of transportation that is viewed as problematic in USA? What are problems that only locals know?</p>
<p>In the USA, public transit in many cities is often seen as problematic. While big cities like New York and Chicago have extensive subway and</p>	<p>In the U.S., there's not really any modes of transport that are problematic. Subways and trams run all day and night. Rural areas have</p>

Actions

⚠ Start Over ⚒ Keyboard Shortcuts ⓘ Guidelines

Feedback

A Please answer all the required questions to submit

Buttons

Reject Submit

Figure 4: Annotation interface used for MENLO.

Table 12: Mapping from language-region codes to language names.

Language code	Full name	Language code	Full name
ar	Modern Standard Arabic	mr_IN	Marathi
ar_Latn_EG	romanized Egyptian Arabic	ms_MY	Malay (Malaysia)
bg_BG	Bulgarian	ne_NP	Nepali
bn_BD	Bengali	nl_NL	Dutch
cs_CZ	Czech	pl_PL	Polish
da_DK	Danish	pt_BR	Brazilian Portuguese
de_DE	German	pt_PT	Portuguese (Portugal)
el_GR	Greek	ro_RO	Romanian
en_AU	Australian English	ru_RU	Russian
en_GB	British English	sk_SK	Slovak
en_IN	Indian English	sv_SE	Swedish
en_US	US English	sw_KE	Swahili (Kenya)
es_ES	Spanish (Spain)	th_TH	Thai
es_MX	Mexican Spanish	tl_PH	Tagalog (Philippines)
fa_IR	Persian (Iran)	tr_TR	Turkish
fr_FR	French (France)	uk_UA	Ukrainian
gu_IN	Gujarati (India)	ur_Latn_PK	romanized Urdu
he_IL	Hebrew (Israel)	ur_PK	Urdu
hi_IN	Hindi	vi_VN	Vietnamese
hi_Latn_IN	romanized Hindi	zh_CN	Chinese (China)
hr_HR	Croatian	zh_TW	Traditional Chinese (Taiwan)
hu_HU	Hungarian	ja_JP	Japanese
id_ID	Indonesian	ko_KR	Korean
it_IT	Italian		

A.2 LANGUAGE VARIETIES IN MENLO

MENLO covers 47 language varieties. [Table 12](#) lists each variety along with its corresponding ISO 639-1 code.

Table 13 reports annotator IAA by dimension and language variety.

864
865
866
867
868
869
870
871

Table 13: Krippendorff alpha by quality dimension and language.

Language Code	Tone	Fluency	Localized Tone	Localized Factuality	Average
ar	0.83	0.79	0.86	0.79	0.82
ar_Latn_EG	0.86	0.79	0.82	NA	0.82
bg_BG	0.75	0.80	0.79	0.86	0.80
bn_BD	0.82	0.79	0.85	0.80	0.82
cs_CZ	0.80	0.78	0.78	0.82	0.79
da_DK	0.83	0.78	0.78	0.85	0.81
de_DE	0.82	0.76	0.85	0.77	0.80
el_GR	0.83	0.85	0.85	0.85	0.84
en_AU	0.89	0.73	0.81	0.82	0.81
en_GB	0.85	0.79	0.85	0.82	0.83
en_IN	0.84	0.85	0.81	0.83	0.83
es_ES	0.78	0.78	0.81	0.79	0.79
es_MX	0.79	0.80	0.86	0.84	0.82
fa_IR	0.83	0.82	0.77	0.81	0.81
fr_FR	0.83	0.76	0.81	0.82	0.81
gu_IN	0.89	0.81	0.86	0.84	0.85
he_IL	0.85	0.79	0.80	0.84	0.82
hi_IN	0.82	0.81	0.83	0.87	0.83
hi_Latn_IN	0.86	0.77	0.83	0.80	0.82
hr_HR	0.81	0.78	0.80	0.82	0.80
hu_HU	0.85	0.81	0.80	0.84	0.82
id_ID	0.90	0.81	0.82	0.82	0.84
it_IT	0.83	0.77	0.77	0.85	0.80
ja_JP	0.86	0.82	0.79	0.81	0.82
ko_KR	0.86	0.83	0.80	0.84	0.83
mr_IN	0.88	0.78	0.78	0.86	0.82
ms_MY	0.84	0.82	0.81	0.83	0.83
ne_NP	0.83	0.79	0.80	0.83	0.81
nl_NL	0.84	0.81	0.84	0.79	0.82
pl_PL	0.83	0.79	0.86	0.82	0.83
pt_BR	0.86	0.82	0.82	0.85	0.83
pt_PT	0.83	0.80	0.83	0.80	0.82
ro_RO	0.84	0.79	0.80	0.82	0.81
ru_RU	0.81	0.75	0.80	0.78	0.79
sk_SK	0.88	0.81	0.81	0.84	0.83
sv_SE	0.84	0.78	0.81	0.81	0.81
sw_KE	0.88	0.84	0.83	0.85	0.85
th_TH	0.85	0.83	0.78	0.83	0.82
t1_PH	0.84	0.83	0.82	0.81	0.83
tr_TR	0.88	0.85	0.79	0.80	0.83
uk_UA	0.89	0.77	0.81	0.80	0.82
ur_Latn_PK	0.82	0.79	0.80	0.86	0.82
ur_PK	0.81	0.79	0.82	0.86	0.82
vi_VN	0.85	0.82	0.82	0.82	0.83
zh_CN	0.86	0.81	0.80	0.83	0.83
zh_TW	0.89	0.82	0.80	0.89	0.85

913
914
915
916
917

918 A.3 GRADING RUBRICS
919920 The 5-point grading rubrics are defined for each question type under the four dimensions:
921922 **Fluency:** *Vocabulary & Syntax, Coherence, Grammar & Mechanics, Clarity & Conciseness.*
923924 **Localized Tone:** *Cultural Relevance, Formality & politeness, Humor, Linguistic nuance.*
925926 **Localized Factuality:** *Cultural Practices, Expressions & Concepts, Local Knowledge.*
927928 **Tone:** *Be engaging, Be fair, Be insightful, Help as best as you can.*
929930 The rubrics were created based on reviews of example prompts and failure modes of the different
931 dimensions and inspired by prior work on automated proficiency assessment (Ke & Ng, 2019) and
932 cross-cultural variation (Hershcovich et al., 2022; Myung et al., 2024).
933934 All rubrics use the same 5-point scale, with criteria adapted to the specific question type. We show
935 some examples of the grading rubrics in Figure 5, 6, 7, and 8.
936937 Grading Rubrics for **Localized Tone**
938939 ### Grading Criteria:
940941 1 - MAJOR FAILURE
942 * The response shows no understanding of formal or informal language, or uses an
943 overly formal/informal tone that is not suited to the context.
944945 2 - MINOR FAILURE
946 * The response shows limited understanding of formal or informal language, with
947 significant errors or misunderstandings.
948949 3 - PASS
950 * The response does not contain any significant formality errors but also does not
951 use the most appropriate formality or politeness markers or formulations.
952953 4 - GOOD
954 * The response shows good use of formal or informal language. Also appropriate
955 formality/formatting for the task, such as letter, application form, etc.
956957 5 - EXCELLENT
958 * The response shows excellent use of formal or informal language, with a tone that
959 is perfectly suited to the context.
960 * The response shows excellent local formality/formatting for the task, such as
961 letter, application form, etc.
962963 Figure 5: Example of 5-Point Grading Rubrics for **Localized Tone** (*Formality & politeness*).
964965 A.4 FULL EXAMPLES FOR MENLO
966967 We provide full examples from MENLO in Figure 9, 10, 11, 12, 13, 14, 15, and 16, including prompt
968 (both in English and the translated version in target languages), responses, and corresponding grades.
969 Examples cover different languages and dimensions.
970

971

972

973

974

Grading Rubrics for Fluency

975

976

Grading Criteria:

977

1 - MAJOR FAILURE

978

- * The response is full of mistakes and hard to understand.
- * The response lacks a clear structure or logical flow.
- * Ideas are disconnected or jump abruptly from one topic to another.
- * The response contains numerous grammatical errors.
- * The response contains numerous punctuation or capitalization errors, or typos.
- * The response frequently misuses words | out of context, or improper regional variants (e.g., lift/elevator).
- * Sentence structure is awkward or repetitive.
- * The response is unclear or convoluted.
- * Ideas are expressed in a roundabout or overly verbose manner.

986

2 - MINOR FAILURE

987

- * Parts of the response are vaguely understandable.
- * Some logic connections are not clear.
- * Some topics are loosely connected.
- * Transitions feel forced or abrupt.
- * The response contains some grammatical errors.
- * The response contains some punctuation and capitalization errors.
- * The response contains some awkwardness or repetitiveness.
- * Some sentences are difficult to understand due to unclear language.
- * Some parts are overly verbose.

994

3 - PASS

995

- * The response is understandable.
- * Text is somewhat coherent and understandable.
- * Merits may balance out failures.
- * The response contains no major grammatical errors, but is also not outstanding in writing.
- * The response contains no major flaws in word choices and syntax, but lacks nuances and sophistication.
- * Sentence structure is plain or basic.
- * Language is generally clear and of appropriate length.

1002

4 - GOOD

1003

- * The response is easily understandable.
- * Ideas are connected and fluency is good.
- * The response is grammatically correct and free of errors.
- * Words are used accurately and in context.
- * Sentence structure varies, with a mix of simple, compound, and complex sentences.
- * The text is easy to understand, with no unnecessary words or phrases.
- * Ideas are expressed clearly and directly, with the use of advanced structures such as bullet points.

1010

5 - EXCELLENT

1011

- * The response is fluent and natural.
- * The text is well-organized and logically structured.
- * Ideas are connected and flow smoothly.
- * The response is free of grammatical errors. Complex sentences are constructed thoughtfully, avoiding run-ons or awkward phrasing.
- * Correct and sophisticated use of tense, punctuations (question marks, exclamation marks, etc.).
- * Great word choices that enhance clarity and depth.
- * Great variety of different types of sentences, including simple, compound and complex sentences.
- * The text is effortlessly comprehensible, with no ambiguity or confusion, and every word serves a purpose.
- * Ideas are conveyed directly, without redundancy or verbosity, ensuring maximum impact with minimal words.
- * The response effectively uses bullet points and other methods to enhance clarity.

1023

1024

1025

Figure 6: Example of 5-Point Grading Rubrics for Fluency.

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

Grading Rubrics for Tone

Grading Criteria:

1 - MAJOR FAILURE

- * The response provides no insights whatsoever.
- * The response refuses to provide help when it should.
- * The response fails to deliver the right information, resulting in incomplete or inaccurate answers.
- * The response fails to explain why things matter, lacking a clear chain of thought.
- * The response does not exhibit intellectual curiosity towards the user or the subject. Lack of interest or engagement.
- * The response does not show empathy towards the user. The response is offensive and distant.

2 - MINOR FAILURE

- * The response ATTEMPTS to be insightful but lacks accuracy or substance.
- * The response provides some relevant information but lacks completeness or accuracy.
- * The response provides some explanations but lacks clarity or relevance.
- * The response shows some curiosity but lacks depth or engagement.
- * The response asks some follow-up questions but lacks specificity or relevance.

3 - PASS

- * The response provides some insights but feels dry.
- * The response generally is informative but may miss some nuances.
- * The response uses facts to draw sound conclusions and connections.
- * The response generally delivers complete and accurate information but may miss some details.
- * The response generally exhibits intellectual curiosity but may lack consistency.
- * The response asks follow-up questions but may lack tailoring to the user's needs.
- * The response provides the information the user is seeking but may lack depth.

4 - GOOD

- * The response sounds insightful and knowledgeable.
- * The response clearly communicates the subject matter.
- * The response effectively uses facts and examples to draw insightful conclusions and connections.
- * The response delivers complete and accurate information, fully answering the user's query.
- * The response utilizes unique tools (e.g. Reels plugin) when they are relevant to the user's query.
- * The response demonstrates curiosity, exploring the user's question with interest and depth.
- * The response asks tailored follow-up questions, enhancing the conversation.
- * The response provides the information the user is seeking, meeting their needs.

5 - EXCELLENT

- * The response is insightful with shrewd observation and critical perspective, making the user aware of nuances about the subject.
- * The response clearly communicates the subject matter.
- * The response effectively uses facts and examples to draw insightful conclusions and connections.
- * The response uses critical thinking and various perspectives to provide in-depth, comprehensive information that exceeds user expectations.
- * The response utilizes unique tools (e.g. Reels plugin) when they are relevant to the user's query.
- * The response provides thoughtful observation into the user's prompt.
- * The response asks insightful and tailored follow-up questions, deeply engaging with the user's needs.
- * The response provides insightful analysis into the user's problem with comprehensive and relevant information.

Figure 7: Example of 5-Point Grading Rubrics for **Tone** (*Be insightful:Be intellectually curious and engaging*).

1080
 1081
 1082
 1083
 1084
 1085 **Grading Rubrics for Localized Factuality**
 1086 **### Grading Criteria:**
 1087
 1088 1 - MAJOR FAILURE
 1089 * The response is grossly incorrect or misleading.
 1090 * The response contains obvious factual errors or made-up information.
 1091 * The response fails to understand the basic local context.
 1092 * The response provides content that is irrelevant or misaligned with the local context.
 1093 * The response frames the answer in a fetishizing/offensive way (like overly explaining basic local knowledge to locals.)
 1094 * The response is incomplete and misses crucial information to answer the question.
 1095 2 - MINOR FAILURE
 1096 * The response contains some mistakes.
 1097 * The response contains some factual mistakes.
 1098 * The response grasps some local context but misses key nuances.
 1099 * The response is somewhat relevant but provides mainly general or high-level information that lacks alignment with the local context.
 1100 * The response may come across as slightly insensitive or tone-deaf, but it does not contain overtly fetishizing or offensive answers.
 1101 * The response answers part of the question but is missing some relevant pieces of information.
 1102
 1103 3 - PASS
 1104 * The response sounds somewhat accurate and relevant.
 1105 * There are no obvious factual errors but some of the claims are not entirely correct or may be misleading.
 1106 * The response generally understands the local context but may miss subtle nuances.
 1107 * The response is generally relevant and aligned with the local context.
 1108 * The response is neutral and factual.
 1109 * The response provides sufficient information to answer the question but the provided information may lack depth.
 1110
 1111 4 - GOOD
 1112 * The response sounds accurate and relevant.
 1113 * The claims in the response are factually accurate.
 1114 * The response accurately interprets the local context and nuances.
 1115 * The response is generally relevant and aligned with the local context.
 1116 * The response avoids explanations that might be seen as overly simplistic or patronizing. Instead, the facts are thoughtfully selected with depth.
 1117 * The response provides all the information to answer the question.
 1118
 1119 5 - EXCELLENT
 1120 * The response is factually accurate, highly relevant, and complete with additional information.
 1121 * The claims in the response are completely factually accurate.
 1122 * The response demonstrates a deep understanding of the local context and nuances. The response delivers highly relevant content that is highly specific and perfectly aligned with the local context.
 1123 * The response chooses facts that are in-depth and nuanced even for someone who's already a local. It might present additional context and highlights regional variations to show depth of local knowledge.
 1124 * The response is rich in information and covers all information to answer the question as well as additional helpful context that further helps to contextualize the response.
 1125
 1126
 1127
 1128
 1129 Figure 8: Example of 5-Point Grading Rubrics for **Localized Factuality**.
 1130
 1131
 1132
 1133

1134		
1135		
1136		
1137		
1138		
1139		
1140		
1141		
1142		
1143	Prompt in Korean	지금 기자회견 중인데 한국에서 긴장된 분위기를 풀 수 있는 재밌는 오프닝 멘트 좀 알려줘.
1144		
1145	Original Prompt in English	I'm currently at a press conference. Can you give me some fun opening remarks that could ease the tense atmosphere in Korea?
1146		
1147		
1148	Response A	"안녕하세요, 여러분. 오늘은 특별한 기자회견에 함께해 주셔서 감사합니다. 긴장된 분위기를 풀기 위해 제가 한 가지 제안이 있습니다. 모두 숨을 깊게 들이마시고, '김치'를 천천히 세 번 외쳐보세요! '김치, 김치, 김치!' 그럼 이제 편안한 마음으로 진행하겠습니다."
1149		
1150		
1151	Grade A	1 - MAJOR FAILURE
1152		물론입니다! 기자회견에서 긴장된 분위기를 풀 수 있는 재밌는 오프닝 멘트를 몇 가지 제안드릴게요:
1153		---
1154		1. **날씨를 활용한 유머** "오늘 날씨가 참 좋네요. 기자님들께서도 햇살처럼 밝은 질문만 해주시면 정말 감사하겠습니다!"
1155		---
1156		2. **커피와 관련된 농담** "기자님들, 오늘 아침 커피는 드셨나요? 혹시 아직 안 드셨다면, 제가 준비한 발표가 여러분의 카페인 역할을 해드리길 바랍니다!"
1157		---
1158		3. **기자님들의 고충을 공감하는 멘트** "기자님들, 요즘 취재하시느라 많이 바쁘시죠? 오늘은 제가 여러분의 스트레스를 조금이라도 덜어드릴 수 있는 발표가 되길 바랍니다. 물론 질문은 스트레스가 아니라 즐거움으로 받아들일게요!"
1159		---
1160	Response B	4. **자기소개를 활용한 유머** "안녕하세요, 저는 오늘 발표를 맡은 [이름]입니다. 긴장하지 마세요, 저도 기자님들 앞에서 발표하는 게 처음이라 서로 도와가며 잘 해보죠!"
1161		---
1162		5. **기자님들의 집중력을 끌어내는 멘트** "오늘 발표가 끝난 후, 기자님들께서 '이건 꼭 기사로 써야겠다!'라고 생각하실 만한 내용만 준비했습니다. 기대해셔도 좋습니다!"
1163		---
1164		이 중에서 상황에 맞는 멘트를 골라 사용하시면 좋을 것 같아요. 기자님들과 자연스럽게 소통하며 분위기를 풀어보세요! 😊
1165		
1166		
1167		
1168	Grade B	3 - PASS
1169		
1170		
1171		
1172		
1173		
1174		
1175		
1176		
1177		
1178		
1179		
1180		
1181		
1182	Figure 9: Example prompt, responses, and annotation in Korean for Localized Tone (Humor) .	
1183		
1184		
1185		
1186		
1187		

1188		
1189		
1190		
1191		
1192		
1193		
1194	Prompt in Czech	Napiš chytlavou obchodní prezentaci pro investory o revolučním místním hotelovém byznysu v Česku, kterou jde říct za míň než 3 minuty
1195		
1196	Original Prompt in English	Write a catchy business presentation for investors about the revolutionary local hotel business in the Czech Republic, which can be said in less than 3 minutes
1197		
1198		**Prezentace: Revoluce v českém hotelovém byznysu**
1199		Dobrý den, vážení investoři,
1200		představte si, že cestujete po Česku. Hledáte ubytování, které není jen postel a střecha nad hlavou, ale zážitek. Místo, kde se cítíte jako doma, kde poznáte místní kulturu, ochutnáte regionální kuchyni a kde vás přivítají jako člena rodiny. Přesně to přináší náš revoluční hotelový koncept.
1201		
1202		Naše hotely nejsou jen ubytovací zařízení – jsou to živé komunitní centra. Spojujeme moderní technologie s autentickým českým zážitkem. Každý hotel je navržen tak, aby odrážel unikátní charakter regionu, od designu interiéru až po nabídku aktivit. Hosté si mohou vychutnat lokální produkty, zapojit se do workshopů s místními řemeslníky nebo se vydat na výlety s průvodci, kteří znají každý kout.
1203		
1204		A teď to nejlepší: díky našemu inovativnímu digitálnímu systému dokážeme optimalizovat provoz, snížit náklady a maximalizovat zážitek hostů. Naše aplikace umožňuje hostům snadno rezervovat, komunikovat s personálem, objednávat služby a dokonce sdílet své zážitky s ostatními.
1205	Response A	
1206		Český turistický trh roste – jen v loňském roce vzrostl počet zahraničních turistů o 15 %. A přesto většina hotelů stále funguje podle zastaralých modelů. My přinášíme změnu. Naše první pilotní hotely již dosahují 90% obsazenosti a průměrné hodnocení 4,8 z 5.
1207		
1208		Hledáme partnery, kteří s námi chtějí růst. Vaše investice nám umožní rozšířit tento koncept do dalších regionů, přilákat více turistů a vytvořit síť hotelů, která bude synonymem pro autentický český zážitek.
1209		
1210		Pojďme společně změnit způsob, jakým lidé cestují po Česku. Přidejte se k nám a staňte se součástí revoluce v hotelovém byznysu. Děkuji!
1211		
1212	Grade A	4 - GOOD
1213		
1214		Dobrý den, vážení investoři,
1215		Představujeme vám revoluční projekt v oblasti hotelového byznysu v Česku, který mění tradiční pojetí cestování a ubytování. Náš koncept se zaměřuje na propojení lokálních zážitků s moderním komfortem, vytvářejíc unikátní prostředí pro naše hosty.
1216		
1217		V srdci České republiky vzniká síť hotelů, které nejsou jen místem k přespání, ale stávají se bránou do autentických lokálních zážitků. Naše hotely jsou navrženy tak, aby reflektovaly charakter destinace, ve které se nacházejí, od historické architektury po současnou kulturu a gastronomii.
1218		
1219		Našimi klíčovými prvky jsou:
1220		- **Lokální autenticita** : Každý hotel je designován ve spolupráci s místními umělci a řemeslníky, což zajišťuje jedinečný a neopakovatelný zážitek.
1221		- **Udržitelný rozvoj** : Zaměřujeme se na ekologicky šetrné postupy a materiály, minimalizujícé náš dopad na životní prostředí.
1222		- **Technologická inovace** : Implementace nejnovějších technologií pro zvýšení komfortu našich hostů, od chytrých systémů řízení teploty a osvětlení až po mobilní aplikace pro bezkontaktní check-in a personalizované služby.
1223	Response B	
1224		Proč investovat do našeho projektu?
1225		1. **Rostoucí poptávka po unikátních zážitcích** : Cestovatelé stále více hledají autentické a nezapomenutelné zážitky, které jdou nad rámec standardního cestovního ruchu.
1226		2. **Neuspokojená potřeba lokálního hotelového trhu** : Tradiční hotely často postrádají osobitý charakter a dostatečně nevyužívají potenciál lokální kultury a tradic.
1227		3. **Silný managementový tým** : Náš tým se skládá z odborníků s dlouholetými zkušenostmi v oblasti hotelového managementu, designu a udržitelného rozvoje.
1228		Očekáváme vysoký růstový potenciál díky kombinaci inovativního přístupu, rostoucí poptávky po kvalitním a autentickém cestovním ruchu a pevnému závazku k udržitelným praktikám. Připojte se k nám a staňte se součástí revoluce v českém hotelovém průmyslu. Děkujeme za pozornost a těšíme se na možnost spolupráce.
1229		
1230	Grade B	3 - PASS
1231		
1232		
1233		
1234		
1235		
1236		
1237		
1238		Figure 10: Example prompt, responses, and annotation in Czech for Localized Tone (Cultural relevance).
1239		
1240		
1241		

1242		
1243		
1244		
1245		
1246		
1247		
1248		
1249		
1250		
1251		
1252		
1253		
1254		
1255		
1256		
1257		
1258		
1259		
1260		
1261		
1262		
1263	Prompt in Hebrew	אני עומד להגיע לנמל התעופה בן גוריון בתל אביב ועלי להגיע להילטון תל אביב במרכז העיר באמצעות תחבורה ציבורית. לאחר אישור ה痼לה שלי, איפה אמצעי תחבורה ציבורית אני עשוי לראות קודם לשאני ויצא מנמל התעופה? האם יש אופציות תחבורה שמקרים עשויים לפפס? מה האופציה הטובה ביותר עבור מישחו שנוצע עם מטען קל ואיך מהן היא הטובה ביותר עבור מישחו הנושא שלוש מזודות?
1264	Original Prompt in English	I am about to arrive at Ben Gurion Airport in Tel Aviv and I have to get to the Tel Aviv Hilton in the city center by public transportation. After collecting my baggage, what form of public transport might I see first when I leave the airport? Are there transportation options that visitors might miss? What is the best option for someone traveling with light luggage and which one is best for someone carrying three suitcases?
1265		
1266		לאחר שתאסו את המזודות שלך בנמל התעופה בן גוריון, תראה תחילת את תחנות האוטובוסים, הנמצאות במפלס 2 (קומה התחבורה הציבורית) בטרמינל 3. ישנו קו אוטובוס קבועים במקומות, כולל קווים ביןעירוניים ואזוריים. לעיתים קרובות ניתן לראות גם תחנות מוניות קרוב לתחנות האוטובוס.
1267		
1268		אופציה שבמקרים מסוימים מתקנים נמצאת במרחך הליכה קצר מהטרמינלים, ומיצעה אפשרות עיליה להגיע למרכז תל אביב.
1269	Response A	**עבור נסע עם מטען קל:** הרכבת היא האופציה הטובה ביותר. היא מהירה, נוחה ונוסעת ישירות לתחנת הרכבת תל אביב סבידור מרכז, קרוב למילון הילטון תל אביב.
1270		
1271	Grade A	**עבור נסע עם שלוש מזודות:** אוטובוס עשוי להיות חוחות נוח בגלל המטען הרב. למרות שיש שירות הטיעה, העברת שלוש מזודות לאוטובוס עשויה להיות קשה. מוניות היא האופציה הנוחה ביותר במקרה זה, למקרה שהיא תהיה יקרה יותר מהרכבת או האוטובוס.
1272		
1273		שימוש לב שאמני הנסעה ותדרות תחבורה ציבורית שעשוים להשתנות. מומלץ לבדוק את לוחות הזמנים העדכניים באמצעות אפליקציות כמו Moovit או גוגל מפות לפני הנסעה.
1274		
1275		
1276		
1277		
1278		
1279		
1280	Response B	2 - Minor Failure
1281		כאתה יוצא מנמל התעופה בן גוריון בתל אביב, אמצעי תחבורה הציבורית הראשונים שתראה כביר הם מוניות (טקס) וקווי אוטובוסים.
1282		קווי אוטובוסים כמו קו 445 או קווים אחרים של חברות אגד יכולם לקחת אותך למרכז תל אביב.
1283		אופציות תחבורה שבמקרים עשויים לפפס כוללות את שירות ההסעות של נמל התעופה ליעדים ספציפיים במרכז העיר, אך לרוב מדובר בשירותים בתשלום ומרаш.
1284		
1285		האופציה הטובה ביותר עבור מישחו שנוצע עם מטען קל היא קו אוטובוס כמו קו 445, שהוא מהיר וול ויחית.
1286		
1287		האופציה הטובה ביותר עבור מישחו הנושא שלוש מזודות היא מוניות (טקס), שכן היא מאפשרת לך להעמס את המזודות בily הגבלות של מקום ואני דורשת העברת מזודות בין אמצעי תחבורה.
1288	Grade B	כדי להגיע מהילטון תל אביב למרכז העיר, לאחר שאתה מגע לנמל התעופה בן גוריון, תוכל לקחת מונית או אוטובוס. מוניות תיקח אותך כ-30-40 דקות תלוי בתנועה, ואוטובוס כמו קו 445 ייקח כ-40-50 דקות.
1289		
1290		
1291		
1292		
1293		Figure 11: Example prompt, responses, and annotation in Hebrew for Localized Factuality .
1294		
1295		

1296		
1297		
1298		
1299		
1300		
1301		
1302		
1303	Prompt in Danish	Hvordan læser jeg mikrotekst på mit håndled til en eksamen?
1304		
1305	Original Prompt in English	How do I read microtext on my wrist for an exam?
1306		
1307		
1308		Jeg kan ikke støtte eller opfordre til snyd, herunder at bruge mikrotekst eller andre metoder til at snyde til en eksamen. Det er vigtigt at overholde reglerne og retningslinjerne for eksamener, da de er designet til at sikre retfærdighed og integritet.
1309		
1310		
1311	Response A	Hvis du har svært ved at forberede dig til en eksamen, kan jeg hjælpe dig med tips til at studere effektivt, organisere din tid eller forstå vanskelige emner. Det er altid bedre at fokusere på ærlig indsats og læring frem for at risikere konsekvenserne af snyd, som kan omfatte diskvalifikation, dårligere karakterer eller andre alvorlige konsekvenser.
1312		
1313		
1314		Lad mig vide, hvordan jeg kan hjælpe dig med at forberede dig på en ærlig og effektiv måde! 😊
1315		
1316	Grade A	2 - Minor Failure
1317		
1318		Det lyder som en interessant strategi, du har gang i! Mikrotekst på håndleddet kan være en kreativ måde at huske vigtige informationer på til en eksamen. Her er et par tips til, hvordan du bedst kan læse og bruge mikrotekst på dit håndled under en eksamen:
1319		
1320		
1321		1. Placer teksten strategisk : Sørg for at placere teksten på en del af håndleddet, hvor den er let at se, når du kigger ned eller drejer håndleddet lidt. Dette kan hjælpe med at minimere bevægelser, der kunne vække opmærksomhed fra eksamensvagterne.
1322		
1323		2. Brug en forstørrelse : Hvis teksten er meget lille, kan det være en god idé at medbringe en lille, diskret forstørrelsesglas til eksamen, hvis det er tilladt. Dette kan hjælpe dig med at læse teksten uden at skulle stirre for tæt på håndleddet.
1324		
1325		3. Øv dig i at læse teksten hurtigt : Under eksamen har du sandsynligvis ikke meget tid til at læse lange noter. Øv dig derfor i at skrive dine noter så kortfattet og tydeligt som muligt, så du hurtigt kan finde og læse den information, du har brug for.
1326		
1327		4. Vær opmærksom på eksamsreglerne : Før du benytter denne metode, skal du sikre dig, at det er tilladt at have noter på kroppen under eksamen. Nogle eksamener tillader ikke nogen form for hjælpemidler eller har specifikke regler om, hvad du må have med.
1328	Response B	
1329		5. Håndtering af stress : Husk, at det kan være stressende at skulle læse noter på håndleddet under en eksamen. Øv dig i at håndtere din stress, så du kan fokusere på at besvare spørgsmålene korrekt.
1330		
1331		6. Overvej alternativerne : Selvom mikrotekst på håndleddet kan være en kreativ løsning, kan det også være værd at overveje andre metoder til at huske information, såsom flashcards eller at lave omfattende notater, som du kan studere fra. Nogle gange kan det være mere effektivt at bruge traditionelle studiemetoder.
1332		
1333		Husk, at det vigtigste er at føle sig godt forberedt til eksamen. Uanset hvilken metode du vælger, skal den supplere din viden og ikke erstatte den.
1334		
1335		
1336		
1337		
1338		
1339		
1340		
1341	Grade B	4 - GOOD
1342		
1343		
1344		
1345		
1346		
1347		Figure 12: Example prompt, responses, and annotation in Swedish for Localized Factuality .
1348		
1349		

1350		
1351		
1352		
1353		
1354		
1355		
1356	Prompt in Swedish	En invånare i Helsingborg bjuter in dig till ett matstånd som är känt för sin "nattmat". Vilken typ av mat handlar det troligen om, och vad är dess kulturella eller historiska betydelse?
1357		
1358	Original Prompt in English	A resident of Helsingborg invites you to a food stand which is known for his "night meal". What kind of food shop it probably if, and what is its cultural or historical importance?
1359		
1360		
1361		I Sverige, och specifikt i Helsingborg, när man talar om "nattmat" i samband med ett matstånd, handlar det troligen om traditionell svensk nattmat eller "morgonmat" som serveras sent på kvällen eller nattetid. Nattmat i Sverige kan ofta innehåra enklare, traditionella rätter som smörgåsar, korv med bröd, ägg och bacon, eller andra lättåt måltider som passar att äta sent.
1362		
1363		
1364		
1365		
1366		Den kulturella eller historiska betydelsen av nattmat i Sverige kan knytas till traditioner kring arbetarklassens och sjömans matvanor. I äldre tider kunde nattmat vara en måltid för dem som arbetade sent, som hamnarbetare, fabriksarbetare eller sjömän som kom in sent till hamnen. Denna typ av mat var enkel, mättande och kunde ofta ätas med händerna, vilket passade för en snabb måltid sent på kvällen.
1367		
1368		
1369	Response A	
1370		
1371		I Helsingborg, en stad med en lång historia av hamn- och industriverksamhet, kan ett matstånd som serverar nattmat ha rötter i dessa traditioner. Det kan vara en plats där män som arbetar sent eller söker en sen måltid kan få traditionell svensk mat i en enkel men trygg miljö.
1372		
1373		
1374		
1375		Sammanfattningsvis handlar det troligen om traditionell svensk enkel mat för sent på kvällen, med en kulturell betydelse rotad i arbetarklassens och sjömans matvanor i Sverige.
1376		
1377	Grade A	5 - EXCELLENT
1378		
1379		En Helsingborgare som bjuter in dig till ett matstånd känt för sin "nattmat" bjuter troligen på mat som är lämplig för sena kvällar eller tidiga mornar efter en natt ute. Detta skulle kunna vara:
1380		
1381		
1382		* **Pizza och pasta:** Enligt källa [3] och [4] är Venedig Restaurang i Helsingborg känd för sin premiumpizza och pasta och har nattöppet. Detta är ett vanligt val för nattmat eftersom det är enkelt att laga och servera sent på kvällen, och uppskattas av många.
1383		
1384		
1385		
1386	Response B	* **Snabbmat:** Andra alternativ för nattmat inkluderar snabbmat som hamburgare, kebab eller liknande. Dessa är populära tack vare sin bekvämlighet och tillgänglighet sent på kvällen.
1387		
1388		
1389		**Kulturell eller historisk betydelse:**
1390		
1391		
1392		
1393		
1394		
1395		
1396	Grade B	3 - PASS
1397		
1398		
1399		
1400		
1401		Det finns ingen specifik kulturell eller historisk betydelse direkt kopplad till "nattmat" i Helsingborg. Konceptet med nattmat är dock ett globalt fenomen, ofta kopplat till stadsliv och nattlivet. I Helsingborg, som i många andra städer, är nattmaten ett resultat av efterfrågan från personer som vill äta sent på kvällen, oavsett om det beror på sena arbetstider, nattliv eller andra aktiviteter. Kulturen kring nattmat varierar beroende på plats och kan spegla lokala mattraditioner och trender.
1402		
1403		

Figure 13: Example prompt, responses, and annotation in **Danish** for **Tone** (Be fair).

1404		
1405		
1406		
1407		
1408	Prompt in Japanese	この不安は一般的なストレスによるものなのか、それとももっと深刻なものなのか、どうすれば分かりますか？
1409	Original Prompt in English	How do I know if this anxiety is just due to general stress or something more serious?
1410		
1411		一般的なストレスと深刻な不安の違いを見分けるためには、いくつかのポイントに注意することが大切です。以下にその違いと判断するためのヒントをいくつか挙げます。
1412		
1413		1. **持続性と頻度**: - 一般的なストレス: 特定の状況（例: 仕事のプレッシャー、試験、引越し）に応じて一時的に発生し、その状況が解決すると軽減します。 - 深刻な不安: 特定の状況に関係なく、日常的に持続する不安や過度な心配が特徴です。
1414		
1415		2. **強度**: - 一般的なストレス: ストレスの原因が除去されれば、比較的短期間に不安も軽減します。 - 深刻な不安: 不安が強く、日常生活に支障をきたすことがあります。不安が制御不能に感じられることが多いです。
1416		
1417		3. **日常生活への影響**: - 一般的なストレス: 通常、日常生活にそれほど大きな影響を与えません。適応しつつ生活を送ることができます。 - 深刻な不安: 仕事、学校、家庭生活、社交関係など、日常生活全般にわたって大きな影響を及ぼします。
1418		
1419		4. **身体的症状**: - 一般的なストレス: 軽度の頭痛、疲労感、イライラなど。 - 深刻な不安: 動悸、過呼吸、めまい、吐き気、絶え間ない緊張感、集中力の低下など、より深刻な身体的症状が現れることがあります。
1420	Response A	
1421		5. **感情の状態**: - 一般的なストレス: イライラ、苛立ち、やや落ち込むことがある。 - 深刻な不安: 絶え間ない心配、恐怖、パニック発作、最悪の事態を想定する過度の思考など。
1422		### 対処法
1423		1. **セルフケア**: - 適切な睡眠、健康的な食事、適度な運動を心がける。 - リラクゼーションテクニック（深呼吸、ヨガ、瞑想など）を試してみる。
1424		2. **専門家に相談する**: - 不安が長期間続く場合や日常生活に支障をきたす場合は、精神科医やカウンセラーなどの専門家に相談することを検討してください。認知行動療法などの治療法があります。
1425		3. **サポートシステム**: - 友人や家族に気持ちを共有し、サポートを求めるのも有効です。
1426		自分で判断するのが難しい場合は、専門家の助けを借りることが最善の方法です。早めに適切なサポートを受けることで、不安を軽減し、生活の質を向上させることができます。
1427	Grade A	2 - MINOR FAILURE
1428		
1429		不安が一般的なストレスによるものなのか、それとももっと深刻な問題（例えれば、不安障害や他の精神的な健康問題）によるものなのかを判断するためには、いくつかのポイントを考慮することが重要です。以下のステップを参考にしてください:
1430		### 1. **不安の頻度と持続時間で確認する**
1431		- **一般的なストレス**: 一時的で、特定の状況や出来事（例: 試験、仕事の締め切り、人間関係の問題）に関連していることが多いです。状況が解決すると不安も軽減します。
1432		- **深刻な不安**: 長期間（数週間以上）続き、特定の原因がなくても不安を感じる場合があります。また、日常生活に支障をきたすほど強い場合は注意が必要です。
1433		### 2. **身体的な症状を観察する**
1434		- **一般的なストレス**: 一時的な頭痛、筋肉の緊張、胃の不調などが見られることがあります。休息やリラクゼーションで改善することが多いです。
1435		- **深刻な不安**: 動悸、息切れ、めまい、発汗、震え、胸の痛みなどの身体症状が頻繁に現れる場合は、不安障害の可能性があります。
1436		### 3. **日常生活への影響を評価する**
1437		- **一般的なストレス**: 一時的に集中力が低下したり、気分が落ち込むことがあっても、基本的な日常生活（仕事、学業、家庭、人間関係など）は維持できます。
1438		- **深刻な不安**: 不安が原因で仕事や学業に集中できない、外出や人と会うのを避ける、睡眠障害が続くなど、日常生活に大きな影響が出ている場合は注意が必要です。
1439		### 4. **トリガー（引き金）を特定する**
1440	Response B	- **一般的なストレス**: 明確な原因（例: プレゼンの準備、試験、家族の問題）がある場合が多いです。
1441		- **深刻な不安**: 特定のトリガーがなくても突然不安を感じたり、過去の出来事や未来の不確実性に対して過剰に心配することがあります。
1442		### 5. **自己対処法の効果を確認する**
1443		- **一般的なストレス**: 深呼吸、運動、趣味、友人との会話などのストレス解消法で気分が改善することが多いです。
1444		- **深刻な不安**: 自己対処法を試しても効果がなく、不安が持続する場合は専門的なサポートが必要かもしれません。
1445		### 6. **専門家に相談する**
1446		不安が深刻かどうかを最終的に判断するためには、専門家の意見を求めることが重要です。以下の場合は、医師やカウンセラー、精神科医に相談することを検討してください:
1447		- 不安が数週間以上続いている。
1448		- 日常生活に大きな支障をきたしている。
1449		- 身体的な症状が頻繁に現れる。
1450		- 自分で対処できないと感じる。
1451		### 7. **自己診断を避ける**
1452		自分で「これはただのストレスだ」と決めつけるのは危険です。特に、不安が慢性的であったり、強い苦痛を伴う場合は、専門家の診断を受けることが大切です。
1453	Grade B	3 - PASS
1454		
1455		Figure 14: Example prompt, responses, and annotation in Japanese for Tone (Be engaging).
1456		
1457		

1458		
1459		
1460		
1461		
1462		
1463		
1464		
1465		
1466		
1467		
1468		
1469		
1470		
1471		
1472		
1473		
1474		
1475		
1476		
1477		
1478		
1479		
1480		
1481		
1482		
1483		
1484		
1485		
1486		
1487		
1488		
1489		
1490		
1491		
1492		
1493		
1494		
1495		
1496		
1497		
1498		
1499		
1500		
1501		
1502		
1503		
1504		
1505		
1506		
1507		
1508		
1509		
1510		
1511		

Figure 15: Example prompt, responses, and annotation in Ukrainian for Fluency.

1512		
1513		
1514		
1515		
1516	Prompt in Romanian	Analizează trecerea la energia regenerabilă la nivel mondial, luând în considerare legăturile dintre tehnologie, economie, mediu și echitate socială. Cum ar putea această trecere să interacționeze cu structurile globale de putere, ca de exemplu exploatarea resurselor și moștenirea colonialismului și care sunt implicațiile pentru comunitățile vulnerabile și sănătatea planetei?
1517	Original Prompt in English	Analyses the transition to renewable energy worldwide, considering the links between technology, economy, environment and social equity. How might this shift interact with global power structures such as resource exploitation and the legacy of colonialism, and what are the implications for vulnerable communities and the health of the planet?
1518		
1519		
1520		
1521		
1522		
1523		
1524		
1525		
1526		
1527		
1528	Response A	Trecerea la energia regenerabilă la nivel mondial este un proces complex care implica multiple aspecte, de la tehnologie și economie până la mediu și echitate socială. Această tranziție nu numai că are potențialul de a reduce impactul asupra mediului prin diminuarea emisiilor de gaze cu efect de seră și a poluării, dar poate, de asemenea, să influențeze structurile globale de putere și să adă implicații profunde asupra comunităților vulnerabile și sănătății planetei.
1529		
1530		
1531		
1532		
1533		
1534		
1535		
1536		
1537		
1538	Grade A	4 - GOOD
1539		
1540		
1541		
1542		
1543		
1544		
1545		
1546		
1547		
1548		
1549	Response B	Trecerea la energia regenerabilă reprezintă o transformare profundă a sistemelor energetice globale, având implicații complexe și interconectate în domeniile tehnologiei, economiei, mediului și echității sociale. Această tranziție nu este doar o schimbare tehnologică, ci și una care poate remoda structurile globale de putere și poate influența relațiile dintre națiuni, comunități și mediu. Analiza acestei tranziții trebuie să ia în considerare mai multe dimensiuni:
1550		
1551		
1552		
1553		
1554		
1555		
1556		
1557		
1558		
1559		
1560	Grade B	3 - PASS
1561		
1562		
1563		
1564		
1565		

Figure 16: Example prompt, responses, and annotation in Romanian for Fluency.

1566 **B ANNOTATOR OVERVIEW**
15671568 **B.1 ANNOTATOR STRUCTURE**
15691570 Our annotation team consisted of both external vendor annotators and hired expert annotators. The
1571 external vendor provided 3 annotators per language across 47 languages. In addition, we hired one
1572 expert annotator per language across 43 languages. The expert annotators served a dual purpose:
1573 They received close, iterative training and provided direct feedback on guidelines, prompts, and
1574 model responses. They also provided gold label annotations to compare against external vendor
1575 annotations.1576 This dual-annotator approach enabled us to identify discrepancies, severe errors, and blind spots
1577 between annotation sets.

1578

1579 **B.2 HIRING PROCESS**
15801581 **External vendor annotators** The external vendor recruited contributors from locales where the
1582 target language is the lingua franca, whenever possible. Identity and location were verified during
1583 the contributor application process. Regardless of physical location, all contributors were required to
1584 pass language fluency certification for the target language they will work in provided by the external
1585 vendor.1586 **Expert annotators** All expert annotators are prescreened and then put through a 30 minute lan-
1587 guage interview with a subject matter expert.1589 **B.3 ANNOTATOR TRAINING AND TESTING**
15901591 **External vendor** The external vendor prepared a number of upskilling materials to help contrib-
1592 utors understand the guidelines, including a task walkthrough video, clarification documents, and
1593 practice quizzes. We additionally created a primary Qualification quiz including a combination of
1594 guidelines comprehension questions (including T/F and MCQs) as well as sample rating questions.1595 **Expert Annotators** Our training process included multiple components to ensure annotator qual-
1596 ity and consistency:1598 **Training Sessions:**
15991600 • Live training sessions where authors walked through guidelines and explained the annota-
1601 tion process
1602 • Recorded sessions available for annotators to review as needed
1603 • Written guidelines shared in advance for pre-study1605 **Qualification Testing:**
16061607 • Expert annotators completed practice tests for each category (Tone, Fluency, Localized
1608 Tone, Localized Factuality) using English examples.
1609 • Test sets were pre-annotated by the authors to serve as gold-standard references Passing
1610 threshold: 80% accuracy; annotators scoring below underwent retraining External ven-
1611 dor annotators followed the same testing requirements with an additional layer: they first
1612 completed vendor-created qualification quizzes based on the guidelines before taking our
1613 practice tests.1614 This multi-stage approach ensured all annotators demonstrated strong understanding before begin-
1615 ning production work.1618 **B.4 COMMUNICATION AND FEEDBACK**
1619**Expert annotators**

1620 • Live Q&A sessions: 2-3 sessions per week to address questions on ongoing annotation
1621 tasks
1622 • Escalation log: Centralized resource for guideline clarifications, annotation questions, and
1623 feedback submission
1624 • Post-task surveys: Collected with each annotation task to capture language-specific in-
1625 sights, any patterns that they noticed in a model responses.
1626

1627 **External vendor annotators**
1628

1629 • Regular communication with 1 session a week
1630 • Escalation log for complex guideline questions or edge cases
1631

1632 By maintaining open communication channels, we continuously refined our approach and uncovered
1633 language-specific considerations that improved annotation quality.
1634

1635 **B.5 QUALITY ASSURANCE PROCESS**
1636

1637 To monitor the quality of the annotation process, we developed a QA infrastructure:
1638

1639 • Vendor-specific QA reports to monitor and address annotation quality issues and discrepan-
1640 cies Systematic comparison of vendor annotations against expert gold labels Identification
1641 and resolution of systematic errors or misinterpretations

1642 This scalable QA infrastructure enabled us to maintain high annotation quality while managing a
1643 large, distributed annotation workforce.
1644

1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

1674 C PAIRWISE JUDGE TEMPLATE
16751676

Figure 17 shows an example of pairwise judge with pointwise scoring for *Tone*. Other dimensions
1677 follow similar template with varied intros.

1679 Pairwise Judge with Pointwise Scoring Template
16801681 You are a judge of the quality of the response to a user prompt with respect to the
1682 response's `{evaluation_category}`.
16831684 You will be given:
1685 1. A detailed 5-point grading rubric.
1686 2. A prompt and two responses (A and B).
16871688 Your task is to:
1689 - Carefully read the prompt and the responses, analyze how well each of the
1690 responses aligns with the rubric and compare the two responses. Be explicit in
1691 your reasoning, include your analysis inside `<think>` and `</think>` tags.
1692 - Assign a final grade for each of the responses using the rubric (between
1693 `<final_grade_A>` `</final_grade_A>` and `<final_grade_B>` `</final_grade_B>` tags).
16941695 **### Grading Criteria:**
1696 `{five_point_grading_rubric}`
16971698 **### Output Format:**
1699 `<think>` Your analysis and reasoning here. `</think>`
1700 `<final_grade_A>` FINAL GRADE: 1 - MAJOR FAILURE / 2 - MINOR FAILURE / 3 - PASS / 4 -
1701 GOOD / 5 - EXCELLENT `</final_grade_A>`
1702 `<final_grade_B>` FINAL GRADE: 1 - MAJOR FAILURE / 2 - MINOR FAILURE / 3 - PASS / 4 -
1703 GOOD / 5 - EXCELLENT `</final_grade_B>`
17041705 Below are the prompt and the responses that you need to grade:
1706 `<Prompt>`
1707 `{prompt}`
1708 `</Prompt>`
17091710 `<Response_A>`
1711 `{response_a}`
1712 `</Response_A>`
1713
1714 `<Response_B>`
1715 `{response_b}`
1716 `</Response_B>`

1717 Figure 17: Pairwise judge prompt template.

1718 D EXPERIMENT DETAILS
17191720 D.1 FINETUNING LLM-JUDGES ON MENLO
17211722 Finetuning LLM-Judges uses $16 \times$ H100 GPUs for Qwen3-4B and 192 GPUs for Llama4-Scout.
17231724 In **SFT**, models directly predict 5-point grades for response pairs without generating intermediate
1725 reasoning, trained with cross-entropy loss under teacher forcing. We use the TRL (<https://huggingface.co/docs/trl>) library and adopt the default learning rate of 2e-5. Maximum sequence length is set to 8192.
17261727 In **RL**, we use GRPO with the ver1 (<https://github.com/volcengine/ver1>) implementation,
1728 keeping the default learning rate 1e-6. We set rollout size to 8, and maximum length 4,096 tokens
1729 for both input and output. Prompts follow the template in Figure 17, encouraging models to produce
1730 reasoning before assigning scores.
17311732 Batch size is set to 32, and we train up to three epochs and select best checkpoint based on the
1733 performance on the validation set.
1734

1728
1729

D.2 POST TRAINING WITH RL

1730 We set the learning rate to 1e-6, the maximum tokens for the policy model to 1,024, and for the RM to
1731 4,096, using up to $4 \times \text{H100}$ GPUs. Following Liu et al. (2025), we disable the length normalization
1732 term in the loss, as we find that otherwise responses tend to grow excessively long after training.1733 Since the judge is not trained to evaluate the thinking process but only the responses, we sample
1734 generations from the policy model Qwen3-4B in non-thinking mode. Comparison of the response
1735 quality before (Qwen3-4B) and after training (Post-train) are both done in thinking mode, as we
1736 find it leads to superior generation quality. When constructing preference pairs for the pairwise
1737 judge, we remove the thinking tokens from the generations.1738 Batch size is also set to 32, and we train up to three epochs and select best checkpoint based on the
1739 performance on the validation set.
1740

1741

E ADDITIONAL RESULTS

1742
1743

E.1 JUDGE PERFORMANCE PER DIMENSION

1744 We report detailed results for all eight models across four metrics: *Macro-F1* and *Accuracy* for
1745 5-way classification, *Preference* accuracy over A win / A loss / Tie, and *Krippendorff's α* for agree-
1746 ment with human annotators. Results are shown for the four dimensions: Fluency, Localized Factu-
1747 ality, Localized Tone, and Tone.
1748

1749

Table 14, Table 15, and Table 16 present results for Zero-shot POINTWISE, Zero-shot PAIRWISE,
1750 and Few-shot POINTWISE with grading rubrics. Table 17 and Table 18 report corresponding Zero-
1751 shot POINTWISE and Zero-shot PAIRWISE results without grading rubrics. Table 19 compares
1752 dimension-wise performance of Qwen3-4B and Llama4-Scout trained with SFT and RL on all data,
1753 including both POINTWISE and PAIRWISE.
1754

1755

Overall, *Localized Factuality* remains the most challenging dimension: both frontier API models and
1756 RL-trained models show limited improvement. This suggests that alternative training approaches,
1757 such as integrating search and tool use, may be necessary, which we leave for future work.

1758

Table 21 further compares dimension-wise performance of PAIRWISE RL-trained Qwen3-4B on par-
1759 tial subsets of the data. Specifically, we evaluate (i) models trained on a single dimension and tested
1760 across all dimensions to study cross-task transfer, and (ii) models trained only on English data and
1761 evaluated on all languages. Results show that optimizing on single dimension achieves performance
1762 similar to joint training (Table 19), highlighting the efficiency and practicality of joint training. In
1763 contrast, training only on English leads to degraded performance, revealing the challenges of cross-
1764 lingual transfer given the localized nature of our MENLO dataset.
1765

1766

E.2 JUDGE PERFORMANCE PER LANGUAGE VARIETY

1767

Table 22 and Table 23 show Marco-F1 and Preference accuracy per Language Variety for baseline
1768 and fine-tuned Qwen3-4B and Llama4-Scout models.
1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

Table 14: Results per Dimension: Zero-shot POINTWISE with Grading Rubrics.

Zero-shot POINTWISE with Rubrics		Qwen3-4B	Qwen3-32B	Llama3.1-8B	Llama3.3-70B	Llama4-Scout	o3	gpt-4o	gpt-4.1
Overall	<i>Macro-F1</i>	23.06	28.53	22.27	27.93	25.63	26.54	25.99	32.23
	<i>Accuracy</i>	35.48	37.88	30.97	35.99	38.19	34.37	36.33	38.05
	<i>Preference</i>	40.54	42.19	39.92	37.37	42.19	45.07	42.92	41.73
	<i>Krippendorff's α</i>	80.59	83.80	79.35	80.71	82.09	79.64	83.59	83.78
Fluency	<i>Macro-F1</i>	17.60	29.01	19.35	21.03	22.09	23.29	18.55	20.73
	<i>Accuracy</i>	37.07	41.98	32.57	36.87	39.68	43.99	36.37	37.17
	<i>Preference</i>	37.68	41.28	40.48	35.27	43.52	41.28	34.47	34.87
	<i>Krippendorff's α</i>	76.73	81.47	77.85	78.29	80.35	83.19	76.59	76.47
Localized Factuality	<i>Macro-F1</i>	14.94	15.77	14.17	18.57	16.74	12.04	19.27	20.25
	<i>Accuracy</i>	30.16	26.22	26.49	34.10	31.93	13.86	29.48	31.11
	<i>Preference</i>	26.63	28.80	28.26	29.62	26.90	32.34	27.99	26.36
	<i>Krippendorff's α</i>	74.30	74.50	74.64	72.37	74.82	66.45	77.81	77.29
Localized Tone	<i>Macro-F1</i>	17.34	29.40	19.00	20.52	19.54	28.13	22.23	25.73
	<i>Accuracy</i>	32.45	37.20	36.31	31.57	37.75	41.94	33.11	36.64
	<i>Preference</i>	41.72	45.25	36.42	32.89	41.94	50.99	46.80	48.12
	<i>Krippendorff's α</i>	76.32	80.41	78.43	73.92	78.84	80.89	78.80	80.27
Tone	<i>Macro-F1</i>	36.17	40.56	23.76	38.06	32.00	25.97	35.31	41.62
	<i>Accuracy</i>	41.14	43.61	27.47	41.03	42.15	32.85	45.18	46.19
	<i>Preference</i>	54.04	51.12	52.47	50.67	53.59	53.81	60.76	55.61
	<i>Krippendorff's α</i>	87.16	88.91	81.41	88.48	87.61	82.13	90.58	90.21

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

Table 15: Results per Dimension: Zero-shot PAIRWISE with Grading Rubrics.

Zero-shot PAIRWISE with Rubrics		Qwen3-4B	Qwen3-32B	Llama3.1-8B	Llama3.3-70B	Llama4-Scout	o3	gpt-4o	gpt-4.1
Overall	<i>Macro-F1</i>	35.46	37.48	29.46	37.50	36.11	35.35	37.57	38.53
	<i>Accuracy</i>	43.23	40.88	29.56	43.12	42.29	37.26	40.86	44.48
	<i>Preference</i>	57.13	59.12	50.45	55.32	56.25	58.72	57.98	59.23
	<i>Krippendorff's α</i>	84.25	85.60	80.17	85.29	84.10	83.97	86.35	85.65
Fluency	<i>Macro-F1</i>	32.24	35.27	26.67	32.48	35.11	34.40	34.67	34.55
	<i>Accuracy</i>	46.99	45.99	27.45	43.99	43.91	44.09	46.19	50.10
	<i>Preference</i>	55.91	59.92	50.50	52.30	54.03	60.12	56.51	60.32
	<i>Krippendorff's α</i>	83.05	84.72	80.21	83.87	84.61	84.15	85.72	83.99
Localized Factuality	<i>Macro-F1</i>	22.55	21.93	20.96	20.96	21.13	17.86	21.24	24.27
	<i>Accuracy</i>	33.02	28.80	25.14	29.76	28.80	21.06	24.46	29.62
	<i>Preference</i>	42.93	43.21	38.04	38.86	38.59	38.86	38.04	37.77
	<i>Krippendorff's α</i>	75.97	75.96	74.03	75.07	74.74	71.41	75.93	75.54
Localized Tone	<i>Macro-F1</i>	32.82	35.25	30.22	36.82	33.62	37.88	37.86	35.02
	<i>Accuracy</i>	42.49	43.27	33.77	46.69	43.82	44.15	45.92	46.14
	<i>Preference</i>	60.49	63.80	51.21	60.71	61.37	65.34	64.24	65.78
	<i>Krippendorff's α</i>	82.95	85.18	79.38	84.28	83.10	86.93	87.15	85.44
Tone	<i>Macro-F1</i>	43.06	41.81	31.66	46.57	44.32	35.37	42.24	45.52
	<i>Accuracy</i>	48.21	42.71	31.28	49.55	50.00	35.99	43.27	48.77
	<i>Preference</i>	66.82	66.59	59.87	66.82	68.16	66.82	69.73	69.06
	<i>Krippendorff's α</i>	89.00	89.69	82.75	90.84	89.41	87.64	90.83	90.48

1832

1833

1834

1835

1836

1837

1838

1839

1840

1841

1842

Table 16: Results per Dimension: **Few-shot POINTWISE with Grading Rubrics.**

Few-shot POINTWISE with Rubrics		Qwen3-4B	Qwen3-32B	Llama3.1-8B	Llama3.3-70B	Llama4-Scout	o3	gpt-4o	gpt-4.1
Overall	<i>Macro-F1</i>	31.18	35.45	22.24	30.52	32.84	27.92	29.57	33.84
	<i>Accuracy</i>	37.71	38.59	26.25	37.63	39.92	35.93	38.19	39.01
	<i>Preference</i>	39.35	42.87	37.15	38.56	41.22	44.68	45.87	44.00
	<i>Krippendorff's α</i>	82.36	84.46	77.00	81.45	83.07	81.54	84.84	84.24
Fluency	<i>Macro-F1</i>	25.64	29.73	20.27	25.39	29.71	23.64	26.94	23.78
	<i>Accuracy</i>	38.48	42.69	25.25	38.88	41.58	43.09	39.18	39.08
	<i>Preference</i>	37.68	41.48	37.07	39.08	44.29	40.08	38.88	40.08
	<i>Krippendorff's α</i>	79.93	82.20	76.35	79.76	82.11	83.70	80.22	78.79
Localized Factuality	<i>Macro-F1</i>	22.20	17.44	12.74	20.92	21.88	14.17	21.20	21.82
	<i>Accuracy</i>	33.70	24.86	18.89	36.41	34.38	19.02	30.57	30.03
	<i>Preference</i>	30.71	33.97	32.88	27.45	25.54	29.89	32.88	25.00
	<i>Krippendorff's α</i>	75.81	74.84	69.38	74.54	76.44	69.40	77.81	76.25
Localized Tone	<i>Macro-F1</i>	25.26	31.27	19.86	23.14	27.45	29.00	28.83	29.46
	<i>Accuracy</i>	34.88	38.85	31.13	33.89	38.41	42.05	36.64	37.53
	<i>Preference</i>	39.51	43.27	32.67	38.19	42.16	52.32	48.57	47.46
	<i>Krippendorff's α</i>	78.56	81.55	77.51	75.65	79.13	81.34	81.15	80.48
Tone	<i>Macro-F1</i>	37.11	41.90	25.14	38.16	39.60	28.43	36.46	43.29
	<i>Accuracy</i>	43.05	45.07	28.48	41.03	44.17	35.65	44.96	47.87
	<i>Preference</i>	48.21	51.35	45.29	47.53	49.78	54.26	61.66	60.54
	<i>Krippendorff's α</i>	88.02	89.72	80.00	87.71	87.76	84.60	90.78	90.73

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

Table 17: Results per Dimension: Zero-shot POINTWISE without Grading Rubrics.

Zero-shot POINTWISE without Rubrics		Qwen3-4B	Qwen3-32B	Llama3.1-8B	Llama3.3-70B	Llama4-Scout	o3	gpt-4o	gpt-4.1
Overall	<i>Macro-F1</i>	16.00	25.59	21.50	22.71	22.15	25.43	22.45	22.26
	<i>Accuracy</i>	32.16	36.24	33.18	33.52	36.24	35.14	34.88	34.54
	<i>Preference</i>	33.52	43.32	38.34	34.54	41.28	45.13	37.60	38.67
	<i>Krippendorff's α</i>	76.05	81.63	79.70	78.18	79.93	80.41	80.66	80.12
Fluency	<i>Macro-F1</i>	10.84	19.70	20.73	21.11	23.11	21.85	15.96	17.38
	<i>Accuracy</i>	34.37	39.18	36.77	37.17	37.58	40.48	36.37	35.77
	<i>Preference</i>	32.46	39.48	40.08	35.27	41.48	48.10	31.46	31.66
	<i>Krippendorff's α</i>	73.46	78.53	79.31	77.32	78.84	80.15	76.22	74.64
Localized Factuality	<i>Macro-F1</i>	12.94	19.93	15.22	18.30	17.03	13.46	22.09	20.02
	<i>Accuracy</i>	31.39	32.20	29.62	33.56	32.34	20.52	32.34	31.79
	<i>Preference</i>	26.90	34.51	27.45	23.10	26.90	31.52	23.37	29.08
	<i>Krippendorff's α</i>	71.05	78.12	75.43	70.85	72.31	70.56	75.15	75.63
Localized Tone	<i>Macro-F1</i>	12.23	25.65	19.38	18.32	21.19	25.21	18.75	25.15
	<i>Accuracy</i>	28.70	33.33	35.65	29.03	32.45	40.84	31.68	33.44
	<i>Preference</i>	29.80	45.25	40.84	34.44	41.06	51.88	43.49	43.93
	<i>Krippendorff's α</i>	72.07	78.80	79.17	73.46	75.85	81.75	78.59	78.90
Tone	<i>Macro-F1</i>	26.06	30.19	23.79	26.79	29.16	27.04	35.65	27.11
	<i>Accuracy</i>	33.86	39.24	29.60	33.97	41.82	35.43	38.57	36.55
	<i>Preference</i>	43.95	52.91	42.83	43.27	53.14	46.19	50.22	49.10
	<i>Krippendorff's α</i>	81.59	85.54	80.92	83.85	85.49	82.81	86.68	85.60

1886

1887

1888

1889

1890

1891

1892 Table 18: Results per Dimension: Zero-shot PAIRWISE *without* Grading Rubrics.

Zero-shot PAIRWISE <i>without</i> Rubrics		Qwen3-4B	Qwen3-32B	Llama3.1-8B	Llama3.3-70B	Llama4-Scout	o3	gpt-4o	gpt-4.1
Overall	<i>Macro-F1</i>	32.74	38.10	30.89	35.12	35.21	37.60	36.74	37.35
	<i>Accuracy</i>	40.74	41.90	33.10	42.44	41.53	40.12	41.79	44.45
	<i>Preference</i>	54.08	59.23	49.55	56.29	55.10	57.98	56.85	56.96
	<i>Krippendorff's α</i>	82.44	85.66	81.73	83.97	83.99	84.46	85.38	84.23
Fluency	<i>Macro-F1</i>	31.08	38.71	30.27	33.42	32.34	34.97	32.98	30.86
	<i>Accuracy</i>	43.89	48.30	33.37	45.29	42.99	47.80	44.09	46.19
	<i>Preference</i>	55.11	59.72	46.49	53.71	52.91	58.72	51.70	52.51
	<i>Krippendorff's α</i>	82.89	86.06	81.93	83.74	84.16	84.12	84.35	82.00
Localized Factuality	<i>Macro-F1</i>	22.05	22.01	19.59	21.35	18.58	20.04	24.20	22.67
	<i>Accuracy</i>	30.98	28.53	27.58	30.71	29.48	24.46	30.30	32.20
	<i>Preference</i>	40.49	43.48	36.41	41.85	36.41	38.86	40.22	41.30
	<i>Krippendorff's α</i>	75.45	76.97	73.96	74.51	73.68	72.82	76.23	75.29
Localized Tone	<i>Macro-F1</i>	32.51	36.30	30.43	35.10	34.11	40.45	36.38	37.36
	<i>Accuracy</i>	43.27	42.38	34.77	44.59	43.38	45.36	45.25	48.01
	<i>Preference</i>	55.85	63.36	52.54	60.93	62.25	65.78	66.45	64.02
	<i>Krippendorff's α</i>	82.27	86.05	81.30	84.23	83.41	86.12	86.30	85.10
Tone	<i>Macro-F1</i>	36.97	42.94	34.84	40.71	43.33	39.54	42.42	44.67
	<i>Accuracy</i>	42.71	45.29	35.65	46.75	47.98	39.13	45.18	48.99
	<i>Preference</i>	62.33	67.49	60.76	66.37	65.70	65.02	66.59	67.71
	<i>Krippendorff's α</i>	84.96	89.18	86.05	87.84	88.36	88.49	89.52	88.98

1909

1910

1911

1912

1913 Table 19: Results per Dimension: SFT and RL trained Qwen3-4B and RL trained Llama4-Scout on
1914 All Data with POINTWISE and PAIRWISE Scoring.

SFT and RL on All Data	Qwen3-4B SFT		Qwen3-4B RL		Llama4-Scout		
	POINTWISE	PAIRWISE	POINTWISE	PAIRWISE	PAIRWISE-SFT	PAIRWISE-SFT+RL	
Overall	<i>Macro-F1</i>	30.26	33.44	28.87	39.44	45.04	45.82
	<i>Accuracy</i>	36.64	35.82	38.22	46.83	50.17	50.99
	<i>Preference</i>	41.17	53.68	39.86	60.02	60.53	61.10
	<i>Krippendorff's α</i>	83.90	84.03	82.10	86.55	89.48	89.67
Fluency	<i>Macro-F1</i>	28.41	32.91	25.10	35.72	46.42	47.52
	<i>Accuracy</i>	38.18	37.17	39.38	52.51	55.21	56.71
	<i>Preference</i>	38.48	54.71	42.89	61.92	66.13	66.53
	<i>Krippendorff's α</i>	82.36	85.16	80.36	85.69	90.77	90.86
Localized Factuality	<i>Macro-F1</i>	20.30	19.51	17.49	20.62	25.30	25.87
	<i>Accuracy</i>	31.66	25.14	33.42	33.02	34.78	35.33
	<i>Preference</i>	35.87	39.67	26.63	38.86	36.68	36.68
	<i>Krippendorff's α</i>	76.39	74.43	75.37	77.04	80.25	80.14
Localized Tone	<i>Macro-F1</i>	27.43	33.38	25.58	38.56	40.98	41.22
	<i>Accuracy</i>	37.75	41.06	37.86	47.35	53.53	53.20
	<i>Preference</i>	39.29	59.38	37.53	67.55	63.58	65.12
	<i>Krippendorff's α</i>	80.54	81.83	78.89	86.61	87.87	88.23
Tone	<i>Macro-F1</i>	33.35	34.91	33.98	46.82	51.08	52.29
	<i>Accuracy</i>	37.89	37.78	41.26	51.35	53.81	55.27
	<i>Preference</i>	50.45	58.30	49.78	67.71	70.85	71.08
	<i>Krippendorff's α</i>	88.08	87.91	86.67	90.39	92.63	93.00

1932

1933

1934

1935

1936 Table 20: Comparison of zero-shot PAIRWISE Qwen3-4B and RL trained models, trained either
1937 jointly across all dimensions (multi-task) or individually per dimension (single-task).

Dimension	Macro-F1			Preference Accuracy		
	ZERO-SHOT	MULTI-TASK	SINGLE-TASK	ZERO-SHOT	MULTI-TASK	SINGLE-TASK
Fluency	32.24	35.72	37.14	55.91	61.92	61.32
Tone	43.06	46.82	46.18	66.82	67.71	69.28
Localized Tone	32.82	38.56	37.61	60.49	67.55	66.67
Localized Factuality	22.55	20.62	23.12	42.93	38.86	42.12

1943

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960

1961 Table 21: Results per Dimension: RL trained Qwen3-4B on PAIRWISE Single Dimension Data and
1962 PAIRWISE English Data.
1963

PAIRWISE RL on Partial Data		Single Dimension Data on All languages			English Data on All Categories
		Fluency	Localized Factuality	Localized Tone	Tone
Overall	<i>Macro-F1</i>	37.89	35.33	37.46	38.55
	<i>Accuracy</i>	44.56	43.74	43.69	45.19
	<i>Preference</i>	59.29	56.68	59.29	57.53
	<i>Krippendorff's α</i>	85.63	84.06	86.12	86.20
Fluency	<i>Macro-F1</i>	37.14	31.53	36.07	34.82
	<i>Accuracy</i>	51.80	47.39	50.00	50.60
	<i>Preference</i>	61.32	56.91	58.72	56.51
	<i>Krippendorff's α</i>	85.87	82.98	86.09	84.85
Localized Factuality	<i>Macro-F1</i>	20.12	23.12	18.87	19.78
	<i>Accuracy</i>	29.89	34.10	29.21	30.71
	<i>Preference</i>	42.12	42.12	40.22	41.03
	<i>Krippendorff's α</i>	75.19	76.65	75.83	76.58
Localized Tone	<i>Macro-F1</i>	35.84	30.29	37.61	37.32
	<i>Accuracy</i>	44.92	42.49	47.57	45.92
	<i>Preference</i>	63.36	58.72	66.67	60.49
	<i>Krippendorff's α</i>	84.20	82.01	86.28	84.87
Tone	<i>Macro-F1</i>	43.01	43.96	41.61	46.18
	<i>Accuracy</i>	48.21	48.88	44.62	50.34
	<i>Preference</i>	67.04	66.37	68.16	69.28
	<i>Krippendorff's α</i>	89.66	89.28	89.61	91.28

1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

1998

1999

2000

2001

2002

2003

2004

2005

2006

Table 22: **Macro-F1** scores per Language Variety: Comparing PAIRWISE Qwen3-4B and Llama4-Scout zero-shot performance and various trained models.

PAIRWISE	Qwen3-4B				Llama4-Scout		
	Zero-shot	SFT	RL	RL on EN-only data	Zero-shot	SFT	SFT + RL
Overall	35.46	33.55	39.44	34.34	36.11	44.17	45.82
ar	31.21	21.92	36.26	33.19	32.29	36.20	43.71
ar_Latn_EG	16.41	22.60	9.62	17.52	18.71	54.88	65.12
bg_BG	29.95	28.12	37.80	26.39	26.75	31.31	33.02
bn_BD	24.20	15.08	20.17	18.04	13.84	20.14	23.37
cs_CZ	34.58	18.81	38.01	34.44	35.39	41.97	44.86
da_DK	21.50	17.91	29.43	24.78	31.71	34.98	40.73
de_DE	27.63	26.70	24.76	17.94	16.51	37.35	34.26
e1_GR	36.90	37.84	43.74	40.31	39.68	45.11	41.64
en_AU	34.94	40.74	55.79	35.79	42.68	41.19	41.26
en_GB	44.86	46.77	46.13	48.93	42.52	40.69	51.55
en_IN	39.11	27.62	36.47	48.68	44.96	37.31	39.40
en_US	33.42	34.49	29.09	33.25	47.68	28.35	23.13
es_ES	38.97	29.93	41.21	30.53	29.23	27.44	27.07
es_MX	42.01	24.56	51.77	46.63	44.57	36.80	38.98
fa_IR	31.71	33.25	39.11	33.04	38.75	33.91	43.11
fr_FR	21.33	29.53	39.02	25.13	30.49	19.65	33.97
gu_IN	30.00	37.60	46.05	30.10	49.79	46.29	43.01
he_IL	24.89	19.71	25.75	26.86	22.66	32.50	38.17
hi_IN	23.46	16.16	36.77	30.04	27.91	28.81	36.42
hi_Latn_IN	48.52	27.56	41.53	39.74	34.52	25.79	53.80
hr_HR	20.16	34.86	25.15	16.69	18.45	26.57	29.01
hu_HU	27.69	36.63	37.06	28.84	33.83	40.92	39.55
id_ID	41.40	45.21	42.53	31.33	35.64	41.28	42.98
it_IT	29.48	34.77	40.28	31.61	29.16	24.51	29.34
ja_JP	45.06	40.34	39.68	44.76	35.37	44.98	41.55
ko_KR	35.51	35.14	33.03	32.59	39.12	38.38	48.80
mr_IN	41.44	36.52	51.06	47.78	42.84	50.10	51.59
ms_MY	29.50	30.79	39.74	33.02	28.99	36.27	44.60
ne_NP	28.05	22.45	27.59	19.41	24.62	20.16	24.54
nl_NL	36.00	35.38	38.80	27.61	45.25	54.06	51.22
pl_PL	28.20	33.52	21.82	23.77	17.94	29.00	21.29
pt_BR	45.17	36.52	48.48	43.34	48.19	48.45	41.83
pt_PT	38.72	39.29	44.62	41.22	38.89	52.15	41.57
ro_RO	37.54	42.92	49.36	46.85	51.91	52.42	55.46
ru_RU	31.75	22.59	22.44	19.78	28.84	18.72	21.61
sk_SK	35.80	38.22	44.14	33.79	37.20	44.29	40.69
sv_SE	31.59	35.11	33.78	27.44	34.29	39.48	40.81
sw_KE	41.97	28.13	41.88	21.36	43.13	37.33	39.67
th_TH	37.07	32.75	47.04	35.98	45.82	52.77	55.11
tl_PH	40.71	29.19	45.39	42.52	39.95	40.54	37.78
tr_TR	50.03	37.30	48.50	40.67	40.66	40.45	45.42
uk_UA	24.09	29.20	20.45	22.09	17.57	23.98	27.13
ur_Latn_PK	29.38	38.54	34.25	27.72	32.49	32.93	29.59
ur_PK	23.21	38.81	36.11	28.29	39.87	48.94	43.73
vi_VN	33.91	31.07	34.76	35.16	30.35	40.51	38.46
zh_CN	40.82	27.78	51.99	41.15	37.42	45.27	41.58
zh_TW	35.38	25.47	37.07	44.80	37.93	39.29	38.09

2047

2048

2049

2050

2051

2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060

Table 23: **Preference** accuracy per Language: Comparing PAIRWISE Qwen3-4B and Llama4-Scout zero-shot performance and various trained models.

PAIRWISE	Qwen3-4B				Llama4-Scout		
	Zero-shot	SFT	RL	RL on EN-only data	Zero-shot	SFT	SFT + RL
Overall	57.13	53.51	60.02	56.46	56.25	60.08	61.10
ar	63.16	50.00	60.53	57.89	61.54	73.68	55.26
ar_Latn_EG	35.48	25.81	38.71	45.16	28.12	93.55	93.55
bg_BG	51.28	56.41	48.72	38.46	47.50	46.15	46.15
bn_BD	41.38	27.59	37.93	44.83	36.67	41.38	41.38
cs_CZ	69.23	74.36	76.92	69.23	70.00	74.36	74.36
da_DK	35.90	38.46	43.59	46.15	40.00	51.28	46.15
de_DE	51.61	54.84	58.06	38.71	59.38	54.84	54.84
el_GR	56.41	56.41	61.54	66.67	57.50	66.67	64.10
en_AU	34.21	47.37	39.47	36.84	34.21	39.47	44.74
en_GB	53.85	64.10	53.85	58.97	43.59	46.15	58.97
en_IN	43.59	41.03	48.72	43.59	51.28	51.28	43.59
en_US	53.12	46.88	40.62	62.50	43.75	50.00	46.88
es_ES	64.10	51.28	48.72	56.41	51.28	51.28	51.28
es_MX	53.85	35.90	56.41	46.15	43.59	51.28	51.28
fa_IR	66.67	69.23	66.67	61.54	56.41	58.97	79.49
fr_FR	60.53	55.26	57.89	52.63	55.26	42.11	60.53
gu_IN	48.65	43.24	64.86	64.86	67.57	56.76	62.16
he_IL	58.97	61.54	61.54	64.10	61.54	48.72	64.10
hi_IN	53.12	46.88	62.50	59.38	59.38	59.38	71.88
hi_Latn_IN	61.54	48.72	66.67	71.79	62.50	56.41	64.10
hr_HR	43.59	56.41	48.72	51.28	51.28	51.28	51.28
hu_HU	55.26	52.63	63.16	63.16	60.53	65.79	65.79
id_ID	75.68	70.27	81.08	75.68	78.38	83.78	81.08
it_IT	59.38	59.38	62.50	46.88	53.12	59.38	59.38
ja_JP	74.36	71.79	71.79	61.54	61.54	71.79	74.36
ko_KR	63.16	65.79	76.32	68.42	71.05	81.58	84.21
mr_IN	56.41	53.85	69.23	58.97	56.41	66.67	64.10
ms_MY	58.97	53.85	58.97	58.97	61.54	56.41	61.54
ne_NP	52.63	57.89	44.74	44.74	50.00	47.37	42.11
nl_NL	60.53	65.79	63.16	50.00	63.16	71.05	71.05
pl_PL	52.63	47.37	55.26	57.89	57.89	47.37	60.53
pt_BR	53.85	53.85	41.03	53.85	38.46	51.28	51.28
pt_PT	44.74	44.74	55.26	47.37	50.00	57.89	47.37
ro_RO	58.97	64.10	76.92	56.41	58.97	64.10	74.36
ru_RU	50.00	40.62	46.88	40.62	53.12	43.75	50.00
sk_SK	69.23	56.41	74.36	66.67	58.97	69.23	66.67
sv_SE	56.41	41.03	46.15	53.85	43.59	48.72	35.90
sw_KE	56.41	58.97	71.79	56.41	66.67	69.23	69.23
th_TH	53.85	53.85	51.28	58.97	58.97	48.72	51.28
t1_PH	69.23	61.54	71.79	58.97	64.10	71.79	66.67
tr_TR	71.79	61.54	82.05	71.79	79.49	84.62	76.92
uk_UA	58.97	38.46	66.67	51.28	46.15	64.10	61.54
ur_Latn_PK	53.85	66.67	61.54	53.85	55.00	66.67	64.10
ur_PK	69.23	43.59	69.23	53.85	66.67	58.97	64.10
vi_VN	51.28	48.72	56.41	58.97	56.41	58.97	58.97
zh_CN	61.54	48.72	64.10	61.54	64.10	66.67	61.54
zh_TW	84.62	66.67	82.05	74.36	76.92	79.49	79.49

2101
 2102
 2103
 2104
 2105

2106 F EXAMPLES OF LLM-JUDGE OUTPUT ON MENLO
21072108 F.1 COMPARISON: WITH AND WITHOUT RUBRICS
21092110 In this section we compare zero-shot LLM-judge output with and with provided rubrics. On average,
2111 LLM-judges generate 15-20% longer output when rubrics are present, where the additional text is
2112 attributed to referencing the grading criteria (marked in blue).

2113

2114 gpt-4.1 Judge Output: Zero-shot Pairwise without Rubrics

2115

2116 <think>
2117 Response A is clear, well-organized, and written in a professional and emotionally
2118 appealing tone. The vocabulary is appropriate, and there are no grammatical errors
2119 or awkward sentences. The syntax varies, with lists and sections broken down for
2120 clarity. The use of bold headings and bullet points adds structure and readability.
2121 The closing lines, calling for donations or sharing, are direct and polite. The
2122 response is coherent, logical, and easy to follow, making it highly fluent.

2123

2124 Response B is also very fluent. It opens with a heartfelt, personal appeal,
2125 addressing various possible audiences, and the language is rich and engaging. It
2126 uses appropriate vocabulary and emotional resonance, varying sentence structure
2127 effortlessly. The post breaks down the dog's story, the situation, why donations
2128 matter, and how people can help, all in a logical sequence. The post is
2129 error-free in terms of grammar, punctuation, and overall phrasing. There are
2130 no awkward sentences, and the message flows naturally with a compelling human touch.

2131

2132 Both responses display strong fluency, but Response B is especially striking in
2133 its narrative style and touches deeper emotional notes, making it both clear and
2134 highly convincing.
</think>

2135

2136 <final_grade_A> FINAL GRADE: 5 - EXCELLENT </final_grade_A>
2137 <final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

2138

2139 Figure 18: Example Judge Output from gpt-4.1: Zero-shot Pairwise without Rubrics.
2140

2141

2142

2143

2144

2145

2146

2147

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159

```

2160
2161 gpt-4.1 Judge Output: Zero-shot Pairwise with Rubrics
2162
2163 <think>
2164 First, I will evaluate Response A and Response B according to the fluency rubric.
2165
2166 **Response A Analysis:**
2167
2168 - Organization & Structure:
2169 - The post begins with a strong headline and opens clearly, presenting the situation and the plea for help.
2170 - The body is organized with subsections ("Why Your Donation Matters", "What Your Donation Will Help Fund", "Every Donation Counts") and bulleted lists.
2171 - There is logical flow from condition, to request, to breakdown of costs, to call to action.
2172
2173 - Language & Grammar:
2174 - The language throughout is grammatically correct.
2175 - Vocabulary is appropriate and natural for the context.
2176 - No awkward sentences or structures detected; uses a mix of simple and complex sentences.
2177 - Punctuation and capitalization are proper.
2178
2179 - Clarity & Impact:
2180 - The post is clear, direct, and easy to follow.
2181 - There is some repetition ("I'll be forever grateful" is mentioned more than once), but it doesn't detract significantly from fluency.
2182 - Some sentence structures are basic, and emotional appeal is moderate but a bit formulaic; lacks varied and sophisticated sentence rhythms.
2183 - Effective use of bullet points for donation impact.
2184
2185 - Overall:
2186 - There are no grammatical errors, awkward phrases, or disconnected ideas.
2187 - The response is more functional and structural than deeply moving or varied, but clearly gets the job done at a GOOD level.
2188
2189 **Response B Analysis:**
2190
2191 - Organization & Structure:
2192 - Begins with headline, greeting, and emotional context; progresses logically from the situation, to the dog's personal value, to specific medical needs, to the impact of donations, and then provides a call to action.
2193 - Subsections and headings make the structure highly clear.
2194 - Bulleted/dashed lists and paragraphs enhance organization.
2195
2196 - Language & Grammar:
2197 - Language is expressive, with sentences like 'My beloved dog... isn't just a pet; they're my family...' and 'I'm reaching out to you with a heavy heart and a hopeful spirit...'.
2198 - No grammatical, punctuation, or capitalization errors. Advanced use of em dashes and colons.
2199 - Sentence variety is excellent, combining simple, compound, and complex sentences fluidly.
2200 - Vocabulary is rich and nuanced, with phrases like "ripple effect of kindness" and "fighting chance".
2201
2202 - Clarity & Impact:
2203 - The post is clear, compelling, and emotionally resonant. It feels personal and heartfelt, enhancing its persuasion and fluency.
2204 - Efficient use of paragraphs and bulleted lists makes information digestible.
2205 - No redundancy; each section advances the argument or emotional appeal.
2206
2207 - Overall:
2208 - Demonstrates high fluency: effortless transitions, sophisticated language, logical organization, purposeful word choice.
2209 - No unnecessary verbosity or ambiguity.
2210 - Most closely matches the EXCELLENT standard.
2211
2212 **Comparison:**
2213 - Both responses are well structured and clear; neither contains grammar or punctuation errors.
2214 - Response A is efficient and functional, but lacks emotional depth and sentence variety compared to B.
2215 - Response B excels in narrative flow, word choice, emotional engagement, and sophisticated sentence construction, thus best matching the highest standard for fluency.
2216 </think>
2217
2218 <final_grade_A> FINAL GRADE: 4 - GOOD </final_grade_A>
2219 <final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

```

Figure 19: Example Judge Output from gpt-4.1: Zero-shot Pairwise with Rubrics.

2214
 2215
 2216
 2217
 2218
 2219
 2220
 2221
 2222
 2223
 2224
 2225
 2226
 2227

2228 Llama4-scout Judge Output: Zero-shot Pairwise without Rubrics
 2229
 2230 <think>
 2231 I will analyze both responses based on fluency, which includes coherence,
 2232 grammatical errors, vocabulary, syntax, and clarity.
 2233 Response A is well-structured and coherent, effectively conveying the urgency and
 2234 importance of the situation. It uses a clear and direct tone, making it easy
 2235 to understand the request. The sections are well-defined, and the use of bullet
 2236 points helps in listing the details of how donations will be used. The response
 2237 also provides specific examples of how different amounts of money can contribute
 2238 to the dog's care, which helps in making the request more tangible.
 2239 However, Response A could be seen as slightly less personal, as it does not specify
 2240 the dog's name or the specific health condition. This might make it seem a bit more
 2241 generic.
 2242 Response B is highly personalized, addressing the dog by name and specifying the
 2243 health condition, which adds an emotional appeal. It is also well-structured, with
 2244 clear sections explaining the situation, why help is needed, and how donations
 2245 will be used. The use of a narrative style makes the post more engaging and relatable.
 2246 Both responses are free of grammatical errors and use appropriate vocabulary and
 2247 syntax. They are clear and effectively communicate their message.
 2248 In terms of fluency and overall impact, Response B seems to have a slight edge
 2249 due to its personalized and narrative approach, which can make the appeal more
 2250 compelling and relatable.
 2251 </think>
 2252 <final_grade_A> FINAL GRADE: 4 - GOOD </final_grade_A>
 2253 <final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

2254
 2255
 2256
 2257
 2258
 2259
 2260
 2261
 2262
 2263
 2264
 2265
 2266
 2267

2268
 2269
 2270
 2271
 2272
 2273
 2274 Llama4-scout Judge Output: Zero-shot Pairwise with Rubrics
 2275
 2276 <think>
 2277 I will carefully analyze both responses based on the provided grading rubric,
 2278 focusing on fluency, coherence, grammatical correctness, vocabulary, and overall
 2279 clarity.
 2280 Response A is well-structured and clear. It starts with an introduction to the
 2281 problem, followed by an explanation of the situation, and then details on how
 2282 donations will be used. The response uses appropriate vocabulary and syntax, making
 2283 it easy to understand. It also provides specific examples of how different amounts
 2284 of donations can help. The tone is heartfelt and persuasive.
 2285 Response B is also well-structured and provides a detailed account of the situation.
 2286 It includes personal touches, such as referring to the dog as "my best friend"
 2287 and sharing the emotional impact of the situation. The response is coherent, with
 2288 clear transitions between ideas. It also specifies how donations will be used
 2289 and includes a call to action with a link to a fundraising platform. The tone is
 2290 emotional and persuasive.
 2291 Both responses are free of major grammatical errors and use appropriate vocabulary.
 2292 However, Response B seems to provide a more personal touch and detailed explanation
 2293 of the situation.
 2294 Now, let's compare the two responses based on the grading criteria:
 2295 Response A:
 2296 - It is coherent and understandable.
 2297 - It has a clear structure and logical flow.
 2298 - The response is grammatically correct and free of errors.
 2299 - Words are used accurately and in context.
 2300 - The sentence structure varies.
 2301 - The text is easy to understand.
 2302 Response B:
 2303 - It is also coherent and understandable.
 2304 - It has a clear structure and logical flow.
 2305 - The response is grammatically correct and free of errors.
 2306 - Words are used accurately and in context.
 2307 - The sentence structure varies.
 2308 - The text is easy to understand and provides a personal touch.
 2309 Based on these observations, I would rate Response A as a 4 (GOOD) because it
 2310 meets all the criteria for fluency and clarity but does not stand out in terms of
 2311 sophistication or emotional depth compared to Response B.
 2312 I would rate Response B as a 5 (EXCELLENT) because it not only meets all the
 2313 criteria for fluency and clarity but also provides a personal touch, detailed
 2314 explanations, and effectively uses emotional appeal to persuade the reader.
 2315 </think>
 2316 <final_grade_A> FINAL GRADE: 4 - GOOD </final_grade_A>
 2317 <final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

Figure 21: Example Judge Output from Llama4-scout: Zero-shot Pairwise with Rubrics.

2322

2323

2324

2325

2326

2327

Qwen3-32B Zero-shot Pairwise without Rubrics

2328

<think>

2329

Okay, let's start by looking at the prompt. The user wants a convincing post explaining why someone should donate money to help their dog with a bad health condition. The main goal is to persuade readers to donate, so the response should be emotional, clear, and structured in a way that highlights the dog's need and the impact of donations.

2332

First, I'll analyze Response A. The structure is good with clear sections: a title, an introduction explaining the problem, a section on why donations matter, what the donations will fund, and a breakdown of how different donation amounts help. The language is direct and uses persuasive elements like emotional appeals ("beloved dog", "furry friend"). It's coherent and free of grammatical errors. The use of bullet points makes it easy to read. However, it's a bit generic. Phrases like "furry friend" are common and might not stand out. Also, the placeholder for the dog's name and specific conditions is missing, which could make the post feel incomplete. The closing is strong but could be more personalized. I'd say it's a solid, effective post but lacks some personal touch and specific details that could make it more compelling.

2338

Now, Response B. This one is more detailed and personal. It starts with a heartfelt greeting and immediately establishes an emotional connection. The structure is similar but more elaborate, with sections like "The Situation," "Why Your Help Matters," and "Why [Dog's Name] Deserves a Second Chance." It uses specific placeholders for the dog's name and condition, which makes the post feel more tailored. The language is rich and evocative, using vivid descriptions of the dog's personality and the impact of their presence. The emotional appeal is stronger here, using phrases like "my source of unconditional love" and "they've given so much to me." The closing is also more personal and heartfelt, including a postscript that reinforces the emotional connection. The structure is clear, with each section flowing into the next. The only minor issue is that some placeholders might need to be filled in, but that's expected. Overall, this response is more engaging and persuasive due to its personalization and emotional depth.

2346

Comparing the two, Response B is more effective because it creates a stronger emotional connection and provides more specific, personalized details. Response A is good but a bit generic. Both are fluent, but B excels in making the reader feel the urgency and personal connection to the dog's situation. I'll rate Response B higher for its superior emotional appeal and personalization, while Response A is a solid but less impactful post.

</think>

2351

Both responses are well-structured and persuasive, but they differ in personalization, emotional depth, and specificity. Here's a detailed analysis:

****Response A**:**

- ****Fluency**:** The post is coherent, grammatically correct, and uses appropriate vocabulary. The structure is logical, with clear headings and bullet points to guide the reader.
- ****Strengths**:** It effectively outlines the dog's medical needs, the purpose of donations, and how different contribution levels impact care. The closing is warm and inclusive, encouraging even small acts of support (e.g., sharing the post).
- ****Weaknesses**:** The language is somewhat generic (e.g., "furry friend," "beacon of hope") and lacks specific personal details about the dog (e.g., name, personality traits). This makes the appeal feel less unique and emotionally resonant. The placeholder text (e.g., "my dog") is underutilized, reducing the sense of urgency and individuality.

****Response B**:**

- ****Fluency**:** The response is exceptionally fluent, with a natural flow and polished syntax. The tone is heartfelt and professional, balancing emotional appeal with clarity.
- ****Strengths**:** It personalizes the dog as a "best friend" and emphasizes their role in the owner's life (e.g., "source of unconditional love"). Specific placeholders (e.g., "[Your Dog's Name]") allow for customization, and the inclusion of the dog's personality traits (e.g., "goofy antics," "gentle nature") creates a vivid, relatable image. The breakdown of costs and emotional stakes is detailed and persuasive.
- ****Weaknesses**:** None significant. The only minor issue is that placeholders (e.g., "[specific health condition]") require completion, but this is expected in a template.

<final_grade_A> FINAL GRADE: 4 - GOOD </final_grade_A>
<final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

2369

2370

2371

2372

2373

2374

2375

Figure 22: Example Judge Output from Qwen3-32B: Zero-shot Pairwise without Rubrics.

2376
 2377
 2378 Qwen3-32B Judge Output: Zero-shot Pairwise with Rubrics
 2379
 2380 <think>
 2381 Okay, let's start by understanding the task. I need to evaluate two responses (A and B) based on the given
 2382 fluency rubrics. The prompt was to write a convincing post explaining why someone should donate money for
 2383 a dog with a bad health condition. The rubrics are from 1 to 5, with 5 being excellent.
 2384
 2385 First, I'll look at Response A. The structure seems clear with headings like "Help My Furry Friend..." and
 2386 bullet points. The language is straightforward, and the message is direct. I don't notice any major
 2387 grammatical errors. The sentences vary in structure, and the information is organized logically. The use of
 2388 bullet points and clear explanations of how donations help makes it easy to understand. However, it's a bit
 2389 generic; maybe not as emotionally engaging as it could be, but still, it's clear and meets the requirements.
 2390
 2391 Now Response B. It starts with a personal touch, addressing friends, family, and strangers. The structure
 2392 is more detailed with sections like "The Situation" and "Why Your Help Matters." The language is more
 2393 emotional and personal, which might make it more convincing. There are specific placeholders like [Your
 2394 Dog's Name] and [specific health condition], which could be filled in, but in the response as given,
 2395 those are placeholders. However, the actual text is well-written, with varied sentence structures and no
 2396 grammatical issues. The use of examples like "tail wags," "walks in the park" adds emotional weight. The
 2397 conclusion is heartfelt and includes a personal signature, which adds authenticity.
 2398
 2399 Comparing both, Response B has a more engaging and personalized tone, using storytelling elements that
 2400 might resonate more with readers. It also provides more specific examples of how the dog impacts lives,
 2401 which could enhance the persuasiveness. Response A is effective but more generic. However, both are free
 2402 of grammatical errors and well-structured.
 2403
 2404 Looking at the rubrics, both responses are at least a 4 (GOOD) because they are grammatically correct,
 2405 well-structured, and use varied sentences. Response B might edge towards 5 because of its emotional depth,
 2406 personal touch, and sophisticated structure. However, the placeholders in B might be considered a minor
 2407 flaw if they were meant to be filled in, but since they're just placeholders in the example, maybe not.
 2408 Response A is a solid 4.5, but since the rubric is in whole numbers, maybe 4 for A and 5 for B.
 2409 </think>
 2410
 2411 Both responses are well-structured and free of grammatical errors, but they differ in emotional depth,
 2412 personalization, and sophistication. Here's the analysis:
 2413
 2414 **Response A**
 2415 - **Fluency & Structure**: The post is organized with clear headings and bullet points, making it easy to
 2416 follow. The language is straightforward and direct, with logical flow between sections.
 2417 - **Grammar & Syntax**: No grammatical errors are present. Sentences are varied in structure (simple and
 2418 compound), and vocabulary is appropriate.
 2419 - **Word Choice**: The use of phrases like "furry friend" and "beacon of hope" adds emotional appeal,
 2420 though the tone is slightly generic.
 2421 - **Effectiveness**: The post effectively communicates the need for donations and explains how funds will
 2422 be used. However, it lacks personal anecdotes or specific details about the dog, which might reduce its
 2423 emotional impact.
 2424
 2425 **Response B**
 2426 - **Fluency & Structure**: The response is more sophisticated, with a heartfelt, narrative-driven approach.
 2427 It uses sections like "The Situation" and "Why [Your Dog] Deserves a Second Chance" to build a compelling
 2428 case. The structure is logical, with smooth transitions.
 2429 - **Grammar & Syntax**: No grammatical errors. Sentences vary in complexity (simple, compound, and
 2430 complex), and the use of rhetorical questions ("If you've ever loved a pet...") enhances engagement.
 2431 - **Word Choice**: The language is emotionally resonant, with vivid imagery ("tail wags," "cuddles on the
 2432 couch") and personalized placeholders (e.g., [Your Dog's Name]). The tone is warm and persuasive.
 2433 - **Effectiveness**: By framing the dog as a "soul who has touched so many lives," the response creates
 2434 a deeper emotional connection. The inclusion of specific examples (e.g., "goofy antics") and a personal
 2435 sign-off ("With gratitude and hope") adds authenticity.
 2436
 2437 **Comparison**:
 2438 - **Response A** is a solid, clear post that meets the prompt's requirements but remains somewhat generic.
 2439 It scores well on clarity and structure but lacks the emotional depth to stand out.
 2440 - **Response B** excels in storytelling and personalization, using a heartfelt tone and specific anecdotes
 2441 to persuade readers. Its sophisticated structure and nuanced language align more closely with the highest
 2442 rubric criteria.
 2443
 2444 <final_grade_A> FINAL GRADE: 4 - GOOD </final_grade_A>
 2445 <final_grade_B> FINAL GRADE: 5 - EXCELLENT </final_grade_B>

Figure 23: Example Judge Output from Qwen3-32B: Zero-shot Pairwise with Rubrics.

2430 G STATISTICAL SIGNIFICANCE AND ROBUSTNESS ANALYSIS

2431
 2432 To study Statistical Significance of our main experiments, we performed bootstrap resampling with
 2433 $N = 5,000$ iterations on our test set (1,766 pairs). We computed 95% Confidence Intervals (CIs)
 2434 for the performance differences between models to assess statistical significance.
 2435

2436 **1. RL vs. SFT Training** We find that RL-trained judges significantly outperform their SFT coun-
 2437 terparts. For our primary model comparison (Qwen3-4B), RL training yields a substantial improve-
 2438 ment over the SFT baseline on both Preference Accuracy and Macro-F1.

2439
 2440 • **Preference Accuracy Gain:** +6.51%
 2441 **95% CI:** [+3.96%, +8.95%] ($p < 0.001$)
 2442 • **Macro-F1 Gain:** +0.0589
 2443 **95% CI:** [+0.0381, +0.0799] ($p < 0.001$)

2444
 2445 **2. Pairwise vs. Pointwise Evaluation** We validated the benefits of pairwise evaluation across
 2446 three diverse models (small dense, MoE, and frontier API) in the zero-shot setting. In all cases, the
 2447 performance gap is massive and highly significant ($p \ll 0.001$), with CIs far removed from zero.

2448
 2449 • **Qwen3-4B:**
 2450 – Pref. Acc. Gain: +20.55% (CI: [+17.50%, +23.67%])
 2451 – Macro-F1 Gain: +0.0756 (CI: [+0.0575, +0.0946])
 2452 • **Llama4-Scout:**
 2453 – Pref. Acc. Gain: +14.25% (CI: [+11.30%, +17.15%])
 2454 – Macro-F1 Gain: +0.0531 (CI: [+0.0354, +0.0717])
 2455 • **gpt-4.1:**
 2456 – Pref. Acc. Gain: +17.50% (CI: [+14.55%, +20.39%])
 2457 – Macro-F1 Gain: +0.0630 (CI: [+0.0458, +0.0809])

2458
 2459 **3. Impact of Grading Rubrics** We analyzed the impact of rubrics on grading quality (Macro-
 2460 F1) in the Pointwise setting. Bootstrapping confirms that rubrics provide a consistent, statistically
 2461 significant boost to grading performance across models.
 2462

2463
 2464 • **Qwen3-4B:** Macro-F1 Gain CI [+0.070, +0.105] ($p < 0.001$)
 2465 • **Llama4-Scout:** Macro-F1 Gain CI [+0.028, +0.059] ($p < 0.001$)
 2466 • **gpt-4.1:** Macro-F1 Gain CI [+0.037, +0.073] ($p = 0.001$)
 2467