MENLO: From Preferences to Proficiency

Evaluating and Modeling Native-like Quality Across 47 Languages

AUTHORS

Chenxi Whitehouse*, Sebastian Ruder*, Tony Zhiyang Lin, Oksana Kurylo, Haruka Takagi, Janice Lam, Nicolò Busetto, Denise Diaz, Francisco Guzmán

* Equal Contribution

The Core Problem — LLMs Still Have an "Accent"

- **LLMs struggle to produce truly native-like responses across languages**
- LLMs exhibit systematic gaps when evaluated by native speakers:
  - Responses may be **grammatically correct** but **culturally tone-deaf**
  - Local factual knowledge is often missing or generic
  - Writing style may feel translated rather than native
  - Existing datasets focus on **task accuracy,** not conversational naturalness

Key Question:

How do we rigorously define, measure, and improve *native-like* quality at scale across many languages?

The MENLO Framework

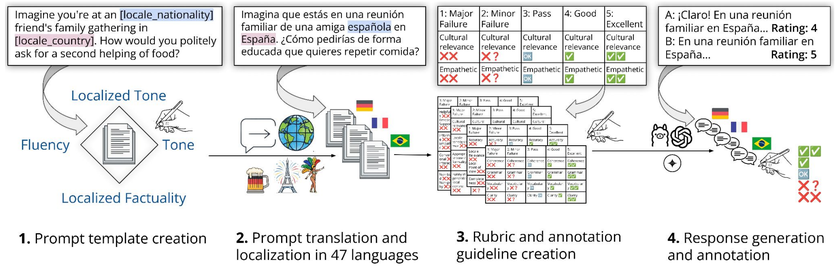
MENLO: Multilingual Evaluation of Native-Like Output

The Four Dimensions of Native-Like Quality

Fluency	Language proficiency compared to an expert-level native speaker
Tone	Overall writing style or "voice" of the response
Localized Tone	Alignment with cultural, regional, and linguistic nuances
Localized Factuality	Factuality, completeness, and grounding in the local context

Dataset Construction — Parametric Prompt Design

Prompts are designed to evoke local contexts through audience design mechanisms



1. Prompt template creation **2.** Prompt translation and localization in 47 languages **3.** Rubric and annotation guideline creation **4.** Response generation and annotation

Dataset Statistics — Scale and Quality

📊 6,423 preference pairs · 47 language varieties
👥 450+ native-speaker annotators · 81,014 total annotations
✓ Krippendorff's α = 0.84 — high inter-annotator agreement
📝 Avg. response: 804 tokens (Fluency) · 575 tokens (Tone)

Evaluating Zero-Shot LLM Judges

Can LLMs automatically evaluate native-like quality?

- **Zero-shot Pointwise** | **Few-shot Pointwise** | **Zero-shot Pairwise**

Pairwise Eval Dramatically Outperforms Pointwise

Model	Macro-F1 (Pointwise ZS)	Macro-F1 (Pairwise ZS)	Pref. Acc. (Pointwise ZS)	Pref. Acc. (Pairwise ZS)
Qwen3-4B	23.06	**35.46** (+12.4)	40.54	**57.13** (+16.6)
Qwen3-32B	28.53	**37.48** (+8.9)	42.19	**59.12** (+16.6)
Llama4-Scout	25.63	**36.11** (+10.5)	42.19	**56.25** (+14.1)
gpt-4.1	32.23	**38.53** (+6.3)	41.73	**59.23** (+17.5)

Key finding: Pairwise evaluation gives a clear **relative grounding signal**. Models are more reliable when judging two responses side-by-side. Gains vs few-shot pointwise: **+5.5% Macro-F1** and **+15.1% Preference**.

Grading Rubrics Provide Substantial Benefit

Model	Pointwise w/o Rubrics	Pointwise w/ Rubrics	Gain
Qwen3-4B	16.00	23.06	**+7.06**
gpt-4.1	22.26	32.23	**+9.97**
Llama-3.3-70B	22.71	27.93	**+5.22**

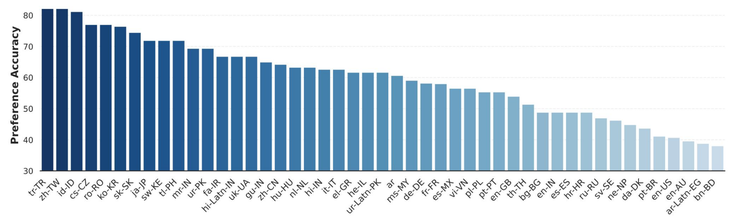
Takeaway: Rubrics can bridge the gap between pointwise and pairwise evaluation. High-quality model-generated rubrics can fill the same gap.

Training LLM Judges — SFT vs. RL

Model	Zero-shot	SFT	RL	SFT+RL
Qwen3-4B (Macro-F1)	35.46	33.44 (−2.0)	**39.44 (+4.0)**	39.33 (+3.9)
Qwen3-4B (Pref. Acc.)	57.13	53.68 (−3.5)	**60.02 (+2.9)**	58.78 (+1.7)
Llama4-Scout (Macro-F1)	36.11	44.17 (+8.1)	**45.62 (+9.5)**	**45.82 (+9.7)**
Llama4-Scout (Pref. Acc.)	56.25	60.08 (+3.8)	**62.60 (+6.4)**	61.10 (+4.9)

Critical insight: For thinking models (Qwen3-4B), SFT (without CoT) hurts performance. RL incentivizes reasoning and achieves best results, also for non-reasoning models.

Cross-Language Judge Performance Varies Widely



From Judges to Generative Reward Models

MENLO's high-quality judges can be deployed as **Generative Reward Models (GenRMs)** to align policy models with native-like preferences.

1. **Train the Judge**
   Use MENLO data to train a judge model via multi-task RL.
2. **Generate & Score**
   Policy model generates responses; the GenRM judge scores them using the pairwise setup.
3. **Policy Optimization**
   Use the judge's scores as reward signals to optimize the policy model (e.g., via GRPO).

Post-Training Results — Consistent Quality Gains

Two-stage evaluation of RL post-trained Qwen3-4B vs Qwen3-4B on the MENLO test set.

Evaluator	# Langs	Win Rate	Average Score (1–5 Scale)		
			Baseline	Post-train	Δ Score
Llama4-Scout-RL-Judge	47	63.88%	3.01	3.79	**+0.78**
Qwen3-32B	47	72.46%	3.44	4.29	**+0.85**
gpt-4.1	47	77.90%	3.21	4.37	**+1.16**
Human Raters	**10**	**55.71%**	**3.31**	**3.67**	**+0.36**

Key Takeaway: Both automated and human evaluators agree that post-training with the RL-judge significantly improves response quality across languages. However, LLM judges tend to overestimate the magnitude of improvement compared to nuanced human judgments.

Conclusion

MENLO provides a scalable, sociolinguistically-grounded framework for evaluating and improving native-like quality in LLMs across 47 language varieties.

Grounded Evaluation

Moves beyond generic quality metrics by operationalizing audience design theory into four distinct, measurable dimensions.

Actionable Signals

Demonstrates that multi-task RL with pairwise evaluation creates judge models that achieve human-level agreement.

Policy Improvement

Proves that these judge models can serve as effective Generative Reward Models to drive consistent gains via RLHF.

🤗 huggingface.co/datasets/facebook/menlo