Novelty Assessment Report

Paper: MENLO: From Preferences to Proficiency – Evaluating and Modeling Native-like Quality Across 47 Languages

PDF URL: https://openreview.net/pdf?id=QOWYX3Q2XS

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

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

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Evaluating and Improving Native-Like Response Quality in Multilingual Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multilingual Evaluation Frameworks and Benchmarks**
- **Evaluation Metrics and Methodologies**
- **Model Architectures and Training Approaches**
- **Task-Specific Applications and Specialized Domains**
- **Machine-Generated Text Detection and Analysis**
- **Comparative and Empirical Studies**

Complete Taxonomy Tree

- Evaluating and Improving Native-Like Response Quality in Multilingual Language Models Survey Taxonomy
- Multilingual Evaluation Frameworks and Benchmarks
  - General Multilingual Generation and Understanding Benchmarks (6 papers)
  - [4] Mega: Multilingual evaluation of generative ai (Kabir Ahuja, 2023) View paper
  - [5] IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages (Harman Singh, 2024) View paper
  - [9] The gem benchmark: Natural language generation, its evaluation and metrics (Sebastian Gehrmann, 2021) View paper
  - [20] Mmlu-prox: A multilingual benchmark for advanced large language model evaluation (Xuan, 2025) View paper
  - [21] Evaluating the elementary multilingual capabilities of large language models with multiq (Carolin Holtermann, 2024) View paper
  - [35] MTG: A benchmark suite for multilingual text generation (Lei, 2022) View paper
  - Native-Like Quality and Naturalness Evaluation ★ (3 papers)
  - [0] MENLO: From Preferences to Proficiency – Evaluating and Modeling Native-like Quality Across 47 Languages (Anon et al., 2026) View paper
  - [13] Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs (Yanzhu Guo, 2025) View paper
  - [43] Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI (Wang Yu-xia, 2025) View paper
  - Machine-Generated Text Detection Benchmarks (2 papers)
  - [12] Overview of ELOQUENT 2025: shared tasks for evaluating generative language model quality (Jussi Karlgren, 2025) View paper
  - [31] MULTITuDE: Large-scale multilingual machine-generated text detection benchmark (Macko Dominik, 2023) View paper
  - Domain-Specific and Task-Specific Multilingual Benchmarks (4 papers)
  - [25] Mdia: A benchmark for multilingual dialogue generation in 46 languages (Zhang Qingyu, 2022) View paper
  - [47] Your Next Token Prediction: A Multilingual Benchmark for Personalized Response Generation (Ding, 2025) View paper
  - [49] Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning (Antreas Ioannou, 2025) View paper
  - [50] Unlocking Markets: A Multilingual Benchmark to Cross-Market Question Answering (Yifei Yuan, 2024) View paper
- Evaluation Metrics and Methodologies
  - LLM-Based Automated Evaluation (4 papers)

  - [11] MTQ-Eval: Multilingual Text Quality Evaluation for Language Models (Rhitabrat Pokharel, 2025) View paper
  - [17] Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? (Hada, 2023) View paper
  - [38] MUG-Eval: A Proxy Evaluation Framework for Multilingual Generation Capabilities in Any Language (Seyoung Song, 2025) View paper
  - [42] RDF-Based Structured Quality Assessment Representation of Multilingual LLM Evaluations (Jonas Gwozdz, 2025) View paper
  - Translation Quality Evaluation (5 papers)
  - [3] Error analysis prompting enables human-like translation evaluation in large language models (Qingyu Lu, 2024) View paper
  - [6] Cross-lingual evaluation of multilingual text generation (S Chollampatt, 2025) View paper
  - [7] Cross-lingual auto evaluation for assessing multilingual llms (Sumanth Doddapaneni, 2025) View paper
  - [10] Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): Proposing a unifying framework for research (Han, 2025) View paper
  - [32] A corpus-based multilingual comparison of AI-based machine translations (Cuilin Liu, 2024) View paper
  - Specialized Quality Metrics (3 papers)
  - [16] Comparing Hallucination Detection Metrics for Multilingual Generation (Kang, 2024) View paper
  - [29] INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback (Xu, 2023) View paper
  - [39] Measuring the Quality of AI-Generated Clinical Notes: A Systematic Review and Experimental Benchmark of Evaluation Methods (A Dahlberg, 2025) View paper
- Model Architectures and Training Approaches
  - Multilingual Pre-Training and Foundation Models (3 papers)
  - [26] Advancements in Natural Language Processing: Leveraging Transformer Models for Multilingual Text Generation (Goyal, 2024) View paper
  - [27] Multilingual Large Language Models: A Systematic Survey (Zhu Shaolin, 2024) View paper
  - [30] MedMT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain (GarcÃa-Ferrero, 2024) View paper
  - Fine-Tuning and Adaptation Methods (2 papers)
  - [14] Enhancing text generation in joint Nlg/Nlu learning through curriculum learning, semi-supervised training, and advanced optimization techniques (Rahimanuddin Shaik, 2025) View paper
  - [33] KInIT at SemEval-2024 Task 8: Fine-tuned LLMs for Multilingual Machine-Generated Text Detection (Michal Spiegel, 2024) View paper
  - Cross-Lingual Transfer and Alignment (4 papers)
  - [15] On the Consistency of Multilingual Context Utilization in Retrieval-Augmented Generation (Jirui Qi, 2025) View paper
  - [19] Align, Generate, Learn: A Novel Closed-Loop Framework for Cross-Lingual In-Context Learning (Mateo Alejandro Rojas, 2024) View paper
  - [36] Cross-Lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-RoBERTa Alignment (Li Bing, 2021) View paper
  - [48] Cross-Lingual Transfer Learning for Neural Machine Translation: A Novel Approach to Improved Fluency and Accuracy (Alsheikhidris, 2024) View paper
  - Translation Model Improvement (3 papers)
  - [1] Exploring human-like translation strategy with large language models (Zhiwei He, 2024) View paper
  - [2] Is translation all you need? a study on solving multilingual tasks with large language models (Chaoqun Liu, 2025) View paper
  - [34] Research on Intelligent English to Chinese Translation Based on Transformer Model (Tian Tian, 2023) View paper
- Task-Specific Applications and Specialized Domains
  - Multimodal and Visual Text Generation (3 papers)
  - [8] AnyText: Multilingual Visual Text Generation And Editing (Xiang, 2023) View paper
  - [45] Unpaired cross-lingual image caption generation with self-supervised rewards (Yuqing Song, 2019) View paper
  - [46] Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment (Wu, 2023) View paper
  - Conversational AI and Dialogue Systems (1 papers)
  - [28] Studies in conversational AI: multilingual capabilities, world knowledge, and evaluation strategies (Bruyn, 2024) View paper
  - Specialized Domain Applications (3 papers)
  - [18] Overview of the multilingual text detoxification task at pan 2024 (D Dementieva, 2024) View paper
  - [22] Language Learning Meets Generative AI: Utilizing Large Language Models for Metalinguistic Explanations (Behzad, 2024) View paper
  - [24] ECGText: Human-Centric Text Generation with Enhanced Emotional Intelligence (Banamali Roy, 2024) View paper
  - Cross-Lingual Speech and Multimodal Speech Synthesis (1 papers)
  - [44] Cross-lingual multispeaker text-to-speech under limited-data scenario (Zexin Cai, 2020) View paper
- Machine-Generated Text Detection and Analysis (2 papers)
  - [23] BadRock at SemEval-2024 Task 8: DistilBERT to Detect Multigenerator, Multidomain and Multilingual Black-Box Machine-Generated Text (Marco Siino, 2024) View paper
  - [41] Unveiling the Source: Differentiating Human and Machine-Generated Texts in a Multilingual Setting (Gurunameh Singh Chhatwal, 2024) View paper
- Comparative and Empirical Studies (2 papers)
  - [37] BATAYAN: A Filipino NLP benchmark for evaluating Large Language Models (Jann Railey Montalan, 2025) View paper
  - [40] Benchmarking Human-Like Quality in Neural Machine Translation Systems (Basharat, 2024) View paper

Narrative

Core task: Evaluating and improving native-like response quality in multilingual language models. The field has organized itself around several complementary branches. Multilingual Evaluation Frameworks and Benchmarks establish standardized testbeds spanning diverse languages and tasks, often emphasizing coverage of low-resource settings and culturally grounded phenomena. Evaluation Metrics and Methodologies develop both automatic and human-centered measures to capture fluency, adequacy, and naturalness, with recent work exploring LLM-based evaluators and cross-lingual consistency checks. Model Architectures and Training Approaches investigate how pretraining strategies, alignment techniques, and curriculum learning can enhance multilingual generation quality, while Task-Specific Applications and Specialized Domains adapt these methods to translation, summarization, dialogue, and domain-specific contexts such as medical or legal text. Machine-Generated Text Detection and Analysis examines whether outputs can be distinguished from human writing, and Comparative and Empirical Studies benchmark systems across languages to reveal performance gaps and guide future improvements.

A particularly active line of work focuses on native-like quality and naturalness evaluation, where the challenge is to move beyond surface-level metrics toward assessments that capture idiomatic expression, cultural appropriateness, and human-like fluency. MENLO[0] situates itself squarely in this space, proposing methods to evaluate whether multilingual outputs feel genuinely native rather than merely correct. This emphasis contrasts with broader benchmarks like IndicGenBench[5], which prioritizes task coverage across many Indic languages, and with works such as Human Translation Strategy[1] or Translation All You Need[2], which concentrate on translation fidelity and adequacy. Nearby efforts like English Accent LLMs[13] and Human-Like Text Preference[43] explore related dimensions of naturalness and human-likeness, highlighting ongoing debates about what constitutes truly native quality and how best to measure it across diverse linguistic and cultural contexts.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs

Authors: Yanzhu Guo, Simone Conia, Ze-Lin Zhou, Zelin Zhou, Min Li, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ mon space across languages, we can perform crosslingual â¦ fluent and relevant to the given prompt, making them suitable for analyzing the general linguistic patterns of language modelsâ¦

Relationship Analysis

Both papers belong to the Native-Like Quality and Naturalness Evaluation category, focusing on assessing how well multilingual LLMs produce natural, native-like outputs. The original paper (MENLO) evaluates native-like quality across 47 languages using human preference annotations on four dimensions (fluency, tone, localized tone, localized factuality) and trains RL-based judges, while the candidate paper evaluates naturalness in French and Chinese using automatic corpus-level metrics (lexical and syntactic divergence) and proposes a DPO-based alignment method. The key difference is that MENLO relies on human-annotated preferences and pairwise RL training for evaluation, whereas the candidate paper introduces automatic distributional metrics and focuses on improving naturalness through preference tuning without extensive human annotation.

2. Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI

Authors: Wang Yu-xia, Xing Rui, Yuxia Wang, Mansurov, Jonibek, et al. (59 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Prior studies have shown that distinguishing text generated by large language models (LLMs) from human-written one is highly challenging, and often no better than random guessing. To verify the generalizability of this finding across languages and domains, we perform an extensive case study to identify the upper bound of human detection accuracy. Across 16 datasets covering 9 languages and 9 domains, 19 annotators achieved an average detection accuracy of 87.6\%, thus challenging previous conclu...

Relationship Analysis

Both papers belong to the Native-Like Quality and Naturalness Evaluation category, focusing on assessing human-like characteristics of multilingual LLM outputs. While MENLO evaluates native-like quality across 47 languages using four dimensions (fluency, tone, localized tone, localized factuality) with structured rubrics and trains judges via RL, the candidate paper examines human detection of AI-generated text across 9 languages and 9 domains, investigating whether humans can distinguish and prefer human-written versus machine-generated content. The key difference is that MENLO develops a framework for automated evaluation and improvement of native-like quality, whereas the candidate paper focuses on human perception, detection capabilities, and preferences regarding human-like text.

Contributions Analysis

Overall novelty summary. The paper introduces MENLO, a framework for evaluating native-like response quality in multilingual LLMs through audience design-inspired mechanisms, accompanied by a dataset of 6,423 human-annotated preference pairs across 47 language varieties. Within the taxonomy, this work resides in the 'Native-Like Quality and Naturalness Evaluation' leaf, which contains only three papers total. This represents a relatively sparse research direction compared to neighboring leaves like 'General Multilingual Generation and Understanding Benchmarks' (six papers) or 'LLM-Based Automated Evaluation' (four papers), suggesting the specific focus on native-like quality assessment remains an emerging area.

The taxonomy structure reveals that MENLO's leaf sits within the broader 'Multilingual Evaluation Frameworks and Benchmarks' branch, which also encompasses general benchmarks emphasizing task coverage, machine-generated text detection, and domain-specific evaluation. The sibling papers in this leaf explore related naturalness dimensions—one examining English accent variations in LLMs and another investigating human-like text preferences—but the scope note explicitly distinguishes native-like quality frameworks from general benchmarks lacking explicit naturalness metrics. Neighboring branches address LLM-based automated evaluation and translation quality metrics, indicating that MENLO bridges evaluation methodology with multilingual framework development.

Among the 30 candidates examined through semantic search, none were identified as clearly refuting any of MENLO's three core contributions: the evaluation framework itself, the annotated preference dataset, and the RL-trained judges as reward models. Each contribution was assessed against 10 candidate papers, with zero refutable overlaps found. The framework contribution appears most distinctive given the sparse population of its taxonomy leaf, while the dataset and RL-based reward modeling contributions show no substantial prior work within the limited search scope. This suggests relative novelty across all three dimensions, though the analysis acknowledges its bounded coverage.

Based on the limited literature search of 30 semantically similar papers, MENLO appears to occupy a relatively underexplored niche at the intersection of native-like quality evaluation and multilingual preference alignment. The sparse taxonomy leaf and absence of refuting prior work within the examined candidates suggest meaningful novelty, though the analysis cannot claim exhaustiveness. The framework's emphasis on audience design mechanisms and structured rubrics distinguishes it from broader multilingual benchmarks, while the RL-based reward modeling approach extends beyond existing naturalness evaluation methods within the surveyed scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: MENLO framework for evaluating native-like response quality

Description: The authors develop a framework that breaks down native-like response quality into four key dimensions (language quality and coherence, alignment with cultural and linguistic nuances, factual correctness and grounding in local context, and overall writing style and helpfulness) using principles from audience design with tailored prompts and structured annotation rubrics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. The dimensions and adaptation of partner models in human-machine dialogue

URL: View paper

Brief Assessment

Partner Models Dialogue[56] focuses on human-machine dialogue partner models and their adaptation dimensions, not on evaluating native-like response quality across languages using audience design mechanisms. The candidate addresses dialogue interaction modeling rather than multilingual response evaluation frameworks.

2. Human-Like Embodied AI Interviewer: Employing Android ERICA in Real International Conference

URL: View paper

Brief Assessment

Android ERICA Interviewer[54] focuses on embodied AI systems for conducting human-like interviews at conferences, not on evaluating native-like response quality across languages using audience design mechanisms.

3. Chatbots in healthcare curricula: the case of a conversational virtual patient

URL: View paper

Brief Assessment

Virtual Patient Chatbots[57] focuses on conversational virtual patients in healthcare education, not on evaluating native-like response quality across languages using audience design mechanisms. The domains and evaluation objectives are fundamentally different.

4. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality

URL: View paper

Brief Assessment

Survey Chatbot Humanization[51] focuses on applying humanization techniques to survey chatbots and measuring their impact on interaction experience and data quality. This is a different application domain (survey research) compared to MENLO's focus on evaluating native-like response quality across 47 language varieties using audience design principles.

5. English as a second language writing and automated essay evaluation

URL: View paper

Brief Assessment

Automated Essay Evaluation[59] focuses on assessing English writing quality for second language learners, not evaluating native-like conversational response quality across 47 language varieties using audience design mechanisms as MENLO does.

6. Communication style adaptation in human-computer interaction: An empirical study on the effects of a voice assistant's politeness and machine-likeness on people's â¦

URL: View paper

Brief Assessment

Voice Assistant Politeness[52] focuses on communication style adaptation (politeness and machine-likeness) in human-computer interaction, specifically measuring reactions to recipe recommendations. This is fundamentally different from MENLO's framework for evaluating native-like response quality across multiple linguistic dimensions and 47 language varieties.

7. Artificial intelligence platforms enabling conversational chatbots: the case of tiledesk. com

URL: View paper

Brief Assessment

Tiledesk Platform[53] focuses on conversational chatbot infrastructure and customer service automation, not on evaluating native-like response quality across languages using audience design mechanisms or structured annotation rubrics.

8. Audience-centric natural language generation via style infusion

URL: View paper

Brief Assessment

Style Infusion Generation[55] focuses on infusing audience-centric stylistic preferences (persuasiveness, memorability) into text generation models using pairwise comparisons, not on evaluating native-like response quality across languages using audience design mechanisms.

9. A socio-onomastic study of Spanish receptive bilinguals: Attitudes, ascription, and audience design

URL: View paper

Brief Assessment

Spanish Bilingual Attitudes[58] is a socio-onomastic study examining attitudes and identity among Spanish receptive bilinguals. It does not address LLM evaluation frameworks, audience design mechanisms for AI systems, or structured annotation rubrics for assessing model-generated responses.

10. Styling the other to define the self: A study in New Zealand identity making

URL: View paper

Brief Assessment

New Zealand Identity[60] examines language performance and identity construction in television advertisements using audience design theory. It does not address LLM evaluation frameworks, multilingual response quality assessment, or computational methods for measuring native-like language proficiency.

Contribution 2: MENLO dataset with human-annotated preference pairs

Description: The authors construct a multilingual dataset consisting of 6,423 annotated prompt-response preference pairs across 47 language varieties, achieving high inter-annotator agreement (Krippendorff's alpha = 0.84) through carefully designed annotation guidelines and rubrics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A Multilingual Similarity Dataset for News Article Frame

URL: View paper

Brief Assessment

News Frame Similarity[80] focuses on news article frame similarity across languages, not LLM response quality evaluation. The datasets serve fundamentally different purposes in distinct research domains.

2. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large â¦

URL: View paper

Brief Assessment

PRISM Alignment[73] focuses on participatory human feedback for LLM alignment across diverse demographics and geographies, not on multilingual native-like quality evaluation. The datasets target different objectives and domains.

3. AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging

URL: View paper

Brief Assessment

AraTraditions10k[76] focuses on cross-lingual image annotation with Arabic-English captions for visual content, not multilingual text preference pairs for evaluating LLM response quality across language varieties.

4. Nativqa: Multilingual culturally-aligned natural query for llms

URL: View paper

Brief Assessment

NativQA[75] focuses on natural question-answering datasets for cultural and regional alignment across languages, not on preference-based evaluation of LLM response quality with pairwise comparisons and Likert ratings as in MENLO.

5. Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation

URL: View paper

Brief Assessment

User Preference Metrics[78] focuses on machine translation preference data using automatic metrics to induce preferences, while MENLO addresses native-like conversational quality across 47 language varieties with human annotations on four distinct dimensions (fluency, tone, localized tone, localized factuality). The datasets serve fundamentally different purposes and evaluation contexts.

6. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization

URL: View paper

Brief Assessment

MAPO[79] focuses on multilingual reasoning alignment using machine translation models for preference estimation, not human annotation. The candidate constructs preference pairs automatically via translation probability scores rather than through human annotators rating response quality across cultural and linguistic dimensions.

7. Dialectal Toxicity Detection: Evaluating LLM-as-a-Judge Consistency Across Language Varieties

URL: View paper

Brief Assessment

Dialectal Toxicity Detection[77] focuses on toxicity detection across dialectal variations with synthetic transformations and human-assisted translations, not on native-like quality preference pairs for LLM evaluation.

8. SeamlessM4T: massively multilingual & multimodal machine translation

URL: View paper

Brief Assessment

SeamlessM4T[74] focuses on multilingual speech-to-speech and speech-to-text translation with automatically aligned data (SeamlessAlign), not human-annotated preference pairs for evaluating native-like quality across language varieties.

9. HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages

URL: View paper

Brief Assessment

HelpSteer3-Preference[72] focuses on general-domain instruction-following tasks (STEM, coding, multilingual) for training reward models, while MENLO specifically targets native-like quality evaluation across four dimensions (fluency, tone, localized tone, localized factuality) with audience design principles. The datasets serve different purposes and evaluation frameworks.

10. Aya model: An instruction finetuned open-access multilingual language model

URL: View paper

Brief Assessment

Aya Model[71] focuses on instruction-finetuned multilingual language models with human-annotated instruction-response pairs, not preference pairs for native-like quality evaluation across specific dimensions like fluency, tone, and localized factuality as in MENLO.

Contribution 3: RL-trained judges as generative reward models

Description: The authors demonstrate that judges trained with reinforcement learning, reward shaping, and multi-task learning can be used as generative reward models to directly improve policy model language proficiency, though they note discrepancies between LLM and human evaluations remain.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MPO: Multilingual Safety Alignment via Reward Gap Optimization

URL: View paper

Brief Assessment

MPO[70] focuses on multilingual safety alignment using reward gap optimization between languages, not on training judges with RL for multilingual proficiency evaluation. The technical approaches and objectives differ fundamentally.

2. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback

URL: View paper

Brief Assessment

Okapi[68] focuses on using RLHF to train multilingual LLMs themselves, not on training judges/reward models to evaluate native-like quality. The candidate uses reward models as part of the RLHF pipeline for policy optimization, whereas the original paper trains judges that can themselves serve as generative reward models for improving language proficiency.

3. Implicit Cross-Lingual Rewarding for Efficient Multilingual Preference Alignment

URL: View paper

Brief Assessment

Implicit Cross-Lingual Rewarding[67] focuses on multilingual preference alignment using implicit rewards derived from DPO-aligned models for cross-lingual transfer, not on training judges with RL for multilingual proficiency evaluation as generative reward models.

4. The Role of Generative AI in the Evolution of Digital Advertising Products

URL: View paper

Brief Assessment

Digital Advertising AI[66] focuses on reinforcement learning from human feedback (RLHF) for optimizing advertising content based on user behavior metrics (CTR, CVR), not on training judges for multilingual proficiency evaluation or using them as generative reward models for language quality improvement.

5. Evaluating and improving cultural awareness of reward models for llm alignment

URL: View paper

Brief Assessment

Cultural Reward Models[63] focuses on evaluating and improving cultural awareness of reward models for LLM alignment across diverse cultures, not on using RL-trained judges as generative reward models for multilingual proficiency improvement as in the original paper.

6. Cross-lingual transfer of reward models in multilingual alignment

URL: View paper

Brief Assessment

Cross-lingual Reward Transfer[61] focuses on cross-lingual transfer of reward models for multilingual alignment, not on training judges with RL for generative reward modeling. The candidate examines classifier reward models trained on preference data, while the original develops RL-trained judges that serve dual purposes as evaluators and generative reward models for improving language proficiency.

7. mR3: Multilingual Rubric-Agnostic Reward Reasoning Models

URL: View paper

Brief Assessment

mR3[69] focuses on multilingual reward models trained via supervised fine-tuning (SFT) on diverse multilingual data, not on RL-trained judges. The candidate does not address RL training of judges or their use as generative reward models for improving policy model proficiency.

8. Query in your tongue: Reinforce large language models with retrievers for cross-lingual search generative experience

URL: View paper

Brief Assessment

Cross-lingual Search Generation[65] focuses on cross-lingual information retrieval using LLMs with a retriever-responder architecture. It does not address RL-trained judges, reward models for multilingual proficiency evaluation, or preference alignment—distinct technical domains from the original paper's contribution.

9. Language imbalance driven rewarding for multilingual self-improving

URL: View paper

Brief Assessment

Language Imbalance Rewarding[64] focuses on leveraging language imbalance as a reward signal for multilingual self-improvement through iterative DPO training, not on training judges with RL to serve as generative reward models for evaluating native-like quality across languages.

10. M-rewardbench: Evaluating reward models in multilingual settings

URL: View paper

Brief Assessment

M-RewardBench[62] focuses on evaluating existing reward models in multilingual settings through benchmark construction and analysis, not on training judges with RL or using them as generative reward models for policy improvement.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Styling the other to define the self: A study in New Zealand identity making

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] MENLO: From Preferences to Proficiency – Evaluating and Modeling Native-like Quality Across 47 Languages View paper
- [1] Exploring human-like translation strategy with large language models View paper

- [2] Is translation all you need? a study on solving multilingual tasks with large language models View paper
- [3] Error analysis prompting enables human-like translation evaluation in large language models View paper
- [4] Mega: Multilingual evaluation of generative ai View paper
- [5] IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages View paper
- [6] Cross-lingual evaluation of multilingual text generation View paper
- [7] Cross-lingual auto evaluation for assessing multilingual llms View paper
- [8] AnyText: Multilingual Visual Text Generation And Editing View paper
- [9] The gem benchmark: Natural language generation, its evaluation and metrics View paper
- [10] Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): Proposing a unifying framework for research View paper
- [11] MTQ-Eval: Multilingual Text Quality Evaluation for Language Models View paper
- [12] Overview of ELOQUENT 2025: shared tasks for evaluating generative language model quality View paper
- [13] Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs View paper
- [14] Enhancing text generation in joint Nlg/Nlu learning through curriculum learning, semi-supervised training, and advanced optimization techniques View paper
- [15] On the Consistency of Multilingual Context Utilization in Retrieval-Augmented Generation View paper
- [16] Comparing Hallucination Detection Metrics for Multilingual Generation View paper
- [17] Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? View paper
- [18] Overview of the multilingual text detoxification task at pan 2024 View paper
- [19] Align, Generate, Learn: A Novel Closed-Loop Framework for Cross-Lingual In-Context Learning View paper
- [20] Mmlu-prox: A multilingual benchmark for advanced large language model evaluation View paper
- [21] Evaluating the elementary multilingual capabilities of large language models with multiq View paper
- [22] Language Learning Meets Generative AI: Utilizing Large Language Models for Metalinguistic Explanations View paper
- [23] BadRock at SemEval-2024 Task 8: DistilBERT to Detect Multigenerator, Multidomain and Multilingual Black-Box Machine-Generated Text View paper
- [24] ECGText: Human-Centric Text Generation with Enhanced Emotional Intelligence View paper
- [25] Mdia: A benchmark for multilingual dialogue generation in 46 languages View paper
- [26] Advancements in Natural Language Processing: Leveraging Transformer Models for Multilingual Text Generation View paper
- [27] Multilingual Large Language Models: A Systematic Survey View paper
- [28] Studies in conversational AI: multilingual capabilities, world knowledge, and evaluation strategies View paper
- [29] INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback View paper
- [30] MedMT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain View paper
- [31] MULTITuDE: Large-scale multilingual machine-generated text detection benchmark View paper
- [32] A corpus-based multilingual comparison of AI-based machine translations View paper
- [33] KInIT at SemEval-2024 Task 8: Fine-tuned LLMs for Multilingual Machine-Generated Text Detection View paper
- [34] Research on Intelligent English to Chinese Translation Based on Transformer Model View paper
- [35] MTG: A benchmark suite for multilingual text generation View paper
- [36] Cross-Lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-RoBERTa Alignment View paper
- [37] BATAYAN: A Filipino NLP benchmark for evaluating Large Language Models View paper
- [38] MUG-Eval: A Proxy Evaluation Framework for Multilingual Generation Capabilities in Any Language View paper
- [39] Measuring the Quality of AI-Generated Clinical Notes: A Systematic Review and Experimental Benchmark of Evaluation Methods View paper
- [40] Benchmarking Human-Like Quality in Neural Machine Translation Systems View paper
- [41] Unveiling the Source: Differentiating Human and Machine-Generated Texts in a Multilingual Setting View paper
- [42] RDF-Based Structured Quality Assessment Representation of Multilingual LLM Evaluations View paper
- [43] Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI View paper
- [44] Cross-lingual multispeaker text-to-speech under limited-data scenario View paper
- [45] Unpaired cross-lingual image caption generation with self-supervised rewards View paper
- [46] Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment View paper
- [47] Your Next Token Prediction: A Multilingual Benchmark for Personalized Response Generation View paper
- [48] Cross-Lingual Transfer Learning for Neural Machine Translation: A Novel Approach to Improved Fluency and Accuracy View paper
- [49] Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning View paper
- [50] Unlocking Markets: A Multilingual Benchmark to Cross-Market Question Answering View paper
- [51] Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality View paper
- [52] Communication style adaptation in human-computer interaction: An empirical study on the effects of a voice assistant's politeness and machine-likeness on people's â¦ View paper
- [53] Artificial intelligence platforms enabling conversational chatbots: the case of tiledesk. com View paper
- [54] Human-Like Embodied AI Interviewer: Employing Android ERICA in Real International Conference View paper
- [55] Audience-centric natural language generation via style infusion View paper
- [56] The dimensions and adaptation of partner models in human-machine dialogue View paper
- [57] Chatbots in healthcare curricula: the case of a conversational virtual patient View paper
- [58] A socio-onomastic study of Spanish receptive bilinguals: Attitudes, ascription, and audience design View paper
- [59] English as a second language writing and automated essay evaluation View paper
- [60] Styling the other to define the self: A study in New Zealand identity making View paper
- [61] Cross-lingual transfer of reward models in multilingual alignment View paper
- [62] M-rewardbench: Evaluating reward models in multilingual settings View paper
- [63] Evaluating and improving cultural awareness of reward models for llm alignment View paper
- [64] Language imbalance driven rewarding for multilingual self-improving View paper
- [65] Query in your tongue: Reinforce large language models with retrievers for cross-lingual search generative experience View paper
- [66] The Role of Generative AI in the Evolution of Digital Advertising Products View paper
- [67] Implicit Cross-Lingual Rewarding for Efficient Multilingual Preference Alignment View paper

- [68] Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback View paper
- [69] mR3: Multilingual Rubric-Agnostic Reward Reasoning Models View paper
- [70] MPO: Multilingual Safety Alignment via Reward Gap Optimization View paper
- [71] Aya model: An instruction finetuned open-access multilingual language model View paper
- [72] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages View paper
- [73] The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large â¦ View paper
- [74] SeamlessM4T: massively multilingual & multimodal machine translation View paper
- [75] Nativqa: Multilingual culturally-aligned natural query for llms View paper
- [76] AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging View paper
- [77] Dialectal Toxicity Detection: Evaluating LLM-as-a-Judge Consistency Across Language Varieties View paper
- [78] Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation View paper
- [79] Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization View paper
- [80] A Multilingual Similarity Dataset for News Article Frame View paper