# Challenges and Recommendations for LLMs-as-a-Judge in Multilingual Settings and Low-Resource Languages

**Anonymous ACL submission**

## Abstract

LLM-as-a-Judge has become the dominant evaluation paradigm for many natural language generation tasks, due to shortcomings of conventional metrics and high correlations with human judgment, albeit mostly in English. There are now attempts to extend LLM-as-a-Judge to multilingual settings including low-resource languages. However, LLMs have limited proficiency in low-resource languages, and there is often no adequate human validation in these settings. To highlight the scope of the problem and current practices, we explore the use of LLM-as-a-Judge evaluators in ACL Anthology papers focusing on multilingual settings and low-resource languages across a diverse set of tasks. Out of 650 papers mentioning LLM-as-a-judge, only 33 of them focus on low-resource or multilingual settings. Our in-depth analysis of these papers indicates inconsistent evaluation outcomes, a tendency to overtrust LLM judgments in multilingual settings, and the widespread reliance on a single judge model per study. To help the NLP community further, we conclude with recommendations about how to use LLM-as-a-Judge in multilingual and low-resource settings.

## 1 Introduction

Until recently, evaluation in NLP systems has mainly been performed by humans, with advantages (e.g., reliability, faithfulness, interpretability, quality control) and disadvantages like high time investments and financial costs (Muhammad et al., 2025; Hellwig et al., 2025; Wiechetek et al., 2025; Rouzegar and Makrehchi, 2024, *inter alia*).

Large language models (LLMs) have become central to the development of Natural Language Processing (NLP) systems across a diverse range of tasks (OpenAI et al., 2024; DeepSeek-AI et al., 2025; Gemini Team et al., 2024). In addition to directly performing natural language generation (NLG) and understanding tasks, LLMs are increasingly used for evaluating the outputs of other language models (Zheng et al., 2023; Adlakha et al., 2024a) as well. This paradigm is known as **LLM-as-a-Judge** and it profits from many features of state-of-the-art generative language modeling. These features include instruction following, multi-step reasoning, ease of use through chat conversation capabilities, high linguistic proficiency in high-resource languages, and the ability to generate explanations alongside outputs and judgments.

While human experts are still the upper bound for overall evaluation quality and trust, LLM-as-a-Judge is considered to be easier, cheaper and faster to scale. However, there is limited research about whether LLM judges deliver reliable and trustworthy evaluations that correlate strongly with human judges. These questions have previously been raised by Zheng et al. (2023), who also coined the term LLM-as-a-Judge, and subsequently by Chen et al. (2024b) and Bavaresco et al. (2025), *inter alia*. While some of these studies observe high correlations between LLM and human judges, the focus remains only on English.

Nevertheless, the current trend of using LLM-as-a-Judge has been extended and scaled to languages other than English, including low-resource languages (LRLs). Although some studies have highlighted low reliability in such settings (Hada et al., 2024a,b), LLM-as-a-Judge is becoming a mainstream evaluation method due to (1) its perceived high correlation with human judgments (which is often only validated for a small number of languages); (2) fast and cheap evaluation; and (3) limitations and biases of traditional metrics, especially ones that are reference-based. As a result, LLM-as-a-Judge has become a predominant evaluation paradigm for NLP tasks (e.g., question answering and instruction following), and is now being adopted across additional modalities, including visual question answering (Chen et al., 2024a) and AudioQA (Ivry and Watanabe, 2026).

However, several critical issues arise when LLM-as-a-Judge systems are applied without sufficient caution across languages, including: (1) inconsistent evaluation outcomes depending on the prompt language, with performance often overestimated for low-resource languages, (2) inaccurate performance estimation due to the limited proficiency or inadequate understanding of outputs expressed in low-resource languages, and (3) the widespread reliance on a single judge for evaluation, without considering multi-judge or ensemble-based assessment. At the same time, (4) the tendency to overtrust LLM judgments across settings;[1] clashes with (5) the extremely high requirements for reliability when evaluating high-risk outputs involving safety, fairness, or cultural biases.

Previous work on multilingual LLM-as-a-Judge shows that judgment consistency is particularly weak for low-resource languages, raising concerns that LLM judges may be less reliable in the lower-resource settings where automatic evaluation is especially tempting due to the scarcity of expert annotators (Fu and Liu, 2025). Yet, we lack a systematic analyses about to what extent current multilingual and low-resource uses of LLM-as-a-Judge are reliable (i.e., whether judges are validated in the target languages, whether low-resource languages receive direct human or gold-label checks, and whether studies avoid over-reliance on single general-purpose judges). Our paper addresses this gap in the literature through a systematic survey of 33 in-scope papers retrieved from multilingual and low-resource LLM-as-a-Judge search criteria, analyzing how judges are deployed and validated across languages, tasks, and model families, and offering recommendations for more reliable cross-lingual evaluation.

## 2 Related Work

The LLM-as-a-Judge paradigm was introduced by Zheng et al. (2023), who proposed MT-Bench and Chatbot Arena to evaluate chatbot alignment with human judgments. Their findings (based on evaluation in English) show that frontier LLMs (e.g., GPT-4) match human judgment at over 80% agreement. They also identified key failure modes such as *position bias* (sensitivity to the order in which candidate responses are presented), *verbosity bias* (preference for longer responses regardless of quality), and *self-enhancement bias* (tendency of a model to favor its own outputs).

Building on this foundation, several **surveys** have since synthesized the growing body of work on LLM-based evaluation across NLP tasks. For example, Gu et al. (2024) provide a comprehensive overview about how to build reliable LLM-as-a-Judge systems, covering bias mitigation, consistency improvement, and prompt design strategies and Li et al. (2024) analyze the LLMs-as-judges paradigm from five perspectives (e.g., functionality, methodology, application, meta-evaluation, and limitations) Similarly, Li et al. (2025a) organize the literature around LLM-as-a-Judge in three dimensions: *what* to judge (quality attributes such as helpfulness, safety, and reliability), *how* to judge (tuning and prompting strategies), and *how to benchmark* LLM judges (categorizing benchmarks for LLM-as-a-Judge across general performance, bias quantification, challenging tasks, and domain-specific settings). However, these surveys focus mostly on English without systematically addressing LLM-as-a-Judge in multilingual or low-resource settings.

Alongside these surveys, a number of **empirical studies** and position papers have raised important caveats about the paradigm's reliability. Bavaresco et al. (2025) conduct a large-scale empirical study across 20 NLP tasks and find that LLM judges show substantial variability across tasks and datasets, cautioning that they should be carefully validated before deployment. Chehbouni et al. (2025) argue that LLM-as-a-Judge has been widely adopted before its validity and reliability as an evaluation method have been thoroughly examined.

Although recent benchmarks such as MM-Eval (Son et al., 2024) quantify cross-lingual judge reliability, no prior study has surveyed the research landscape of LLM-as-a-Judge specifically in multilingual settings and for low-resource languages.

## 3 Literature Search and Annotation Methodology

**Definition** **LLM-as-a-Judge** is used broadly to cover both evaluator-oriented and annotator-oriented uses of LLMs in the literature. *Evaluator-*

[1] Overtrusting LLM judgments becomes particularly insidious when what is judged are human-generated outputs and annotations. For example, Belay et al. (2026) use LLM-as-a-Judge to audit the quality of human translation for LRLs even though related work shows this is unreliable without in-context learning examples (Li et al., 2025c). Similarly, Cheng et al. (2026) find that, in difficult social scenarios, humans trust LLMs that evaluate their own stances favourably, even when they are clearly in the wrong.

oriented settings use LLM judgments to measure, compare, or validate items (e.g., model responses, system outputs, retrieved evidence, or benchmark examples). Given task-specific context (e.g., an instruction, source text, candidate response, reference answer, rubric, or label definitions), the judge produces an evaluative output (e.g., a score, label, ranking, preference judgment, or textual assessment). *Annotator*-oriented settings use LLMs to produce labels, metadata, explanations, or error annotations for downstream analysis or training. In this paper, we focus on the first meaning and only analyze LLM-as-evaluator settings.

**Literature Search Methodology** We conducted a keyword-based search over metadata from the ACL Anthology (Bollmann et al., 2023), which provides comprehensive and structured coverage of research in NLP, including work on multilinguality and low-resource languages. Our goal is not to exhaustively enumerate all relevant publications across venues, but to obtain a representative overview of LLM-based evaluation research within the NLP community.

We implemented the search as a lightweight, fully reproducible pipeline in our project repository,[2] which contains the code for parsing Anthology XML files and performing keyword matching.

**Data Source and Reproducibility** Our search operates on the official ACL Anthology repository (Apache 2.0 license), which provides structured XML metadata (e.g., titles, abstracts, and venue information) for papers published at major ACL venues. To ensure reproducibility, we fixed the Anthology snapshot to one commit (370911e, 2025-11-14). It includes the major ACL conferences (e.g., ACL, EMNLP, NAACL, EACL, AACL), two peer-reviewed journals (*Computational Linguistics* and *TACL*), other recurring NLP conferences (e.g., LREC, COLING), and several hundred specialized workshops.

**Keyword Design and Matching** The search is applied to the concatenation of each paper's title and abstract, and is driven by three manually curated keyword groups targeting (i) LLMs, (ii) evaluation or judging functionality, and (iii) low-resource or multilingual contexts. Matching is case-insensitive, and a paper is counted as a hit if it contains at least one keyword from each of the three groups:

- **LLM**: `["LLM", "large language model"]`
- **Judge**: `["judge", "evaluator", "LLM-based evaluation", "LLM-as-a-judge", "LLM-based assessment"]`
- **Low-resource**: `["low-resource", "low resource", "underresourced", "under-resourced", "underresearched", "under-researched", "multilingual"]`

We include *multilingual* in the low-resource group to increase recall, as other low-resource-specific terms alone produce relatively few matches. We also found that some early candidate keywords (e.g., *annotator*) frequently triggered false positives, as they often refer to human annotators rather than automated or LLM-based evaluators.

Enforcing low-resource related keywords substantially reduces the number of retrieved papers. Specifically, across the full ACL Anthology directory, the number of hits decreases from 650 to 49 under the same constraint.[3] This indicates that comparatively few papers explicitly position LLM-based evaluation methods in low-resource or under-represented language settings.

**Annotation and Exclusion Criteria** We manually reviewed each of the 49 candidate papers and annotated the role of the LLM, the task being judged, the languages covered, and the validation protocol against human judgments or gold-standard benchmark labels. We identified 33 papers in which an LLM is used to assess model-generated or human-produced outputs. The remaining 16 papers were excluded for the following reasons:

We excluded annotation-oriented papers in which LLMs are used to directly label text data rather than evaluate outputs of a previous processing step. These include work using LLMs for toxicity ratings (Bell et al., 2025; Faisal et al., 2025), multiclass classification (Upadhayay and Behzadan, 2025), and binary decisions (Tran and Nam, 2025).

Some papers contain both the concepts of "LLMs" and "judgment," but they do not necessarily investigate "LLM-as-a-Judge". For exam-

[2] [URL removed for anonymous review]

[3] The raw search returned 50 results, but one paper (Song et al. 2025) appeared twice (once in the Findings of ACL and once in the Proceedings of the Workshop on Multilingual Representation Learning). Therefore, we counted it only once, yielding 49 unique papers.

ple, several papers examine human judgments of LLM output quality (Anh et al., 2024; Thakur et al., 2024), moral judgments made by LLMs (Agarwal et al., 2024), LLMs' judgments of semantic similarity (Brglez et al., 2024; Wang et al., 2024), or grammaticality and acceptability (Zhang et al., 2024). However, they do not use LLMs as evaluators of model outputs, and therefore fall outside our definition of LLM-as-a-Judge.

We also excluded papers that do not focus on natural language, but instead evaluate programming languages (Chen et al., 2024c) or domain-specific query languages (Li et al., 2025b), as well as two papers[4] that did not use LLMs.

Finally, two papers explicitly avoid LLM-as-a-Judge evaluation in favor of objective evaluation, citing self-enhancement bias (Dussolle et al., 2025) and low reliability in low-resource languages (Song et al., 2025) as reasons.

## 4 LLM-as-a-Judge in Multilingual Research

We analyze the 33 in-scope papers along five dimensions (i.e., judging paradigm, downstream task, judge model, language coverage, and judge evaluation). This section organizes our findings around three cross-cutting claims that draw on these dimensions and motivate the recommendations in Section 5. Figure 1 provides an extensive distribution of the papers across these five dimensions, with a detailed per-paper breakdown of language coverage in Appendix A.

We focus on five judging paradigms: **direct scoring** (the LLM judge produces a scalar score, label, or short verdict from input and candidate output, sometimes with a reference or rubric as in Dinh et al. 2024; Gupta et al. 2024; Qian et al. 2024), **pairwise comparison and ranking** (two or more candidate outputs are compared, sometimes with ties as in Raju et al. 2024; Watts et al. 2024; Thakur et al. 2025), **multi-step or reasoning-based judging** (the judge generates explicit reasoning steps before issuing a verdict as in Lu et al. 2025), **multi-dimensional judging** (the judge rates several predefined quality aspects separately rather than evaluating quality holistically in one step as in Boughorbel et al. 2024; Mukherjee et al. 2025; Okewunmi et al. 2025), and **multi-agent or ensemble judging** (verdicts from multiple distinct judges are aggregated into a single final label as in Niklaus et al. 2025).

The downstream tasks evaluated by these judges cover tasks such as machine translation (MT), summarization, instruction following, question answering, retrieval-augmented generation, safety and moderation as well as more specialized tasks (e.g., text style transfer, referential ambiguity resolution, tutoring, and benchmark quality assessment). The distribution of papers and tasks is provided in Figure 1.

The models used as judges include both generalist instruction-tuned LLMs and more specialized judge models. Generalist judges include closed-source models such as GPT, Gemini, and Claude, as well as open-source models from families such as Llama and Qwen. Specialized judges, such as Llama Guard for safety evaluation and domain-adapted evaluators for specialized tasks, are used in narrower settings where the output space is more constrained.

Our review shows that LLM-as-a-Judge is already broadly used across NLP tasks, different judge model families, and diverse evaluation settings. However, a closer reading of the surveyed papers reveals that this breadth also masks several structural weaknesses which we report below as the results of our analyses.

**Broad Multilingual Coverage but Shallow Depth for Low-Resource Languages** Broad multilingual coverage in the surveyed literature does not necessarily indicate a focus on low-resource languages. We classify languages using Joshi et al. (2020)'s taxonomy, treating its Class 0–3 as low-resource.[5] Among the 19/33 papers (58%) that include at least one low-resource language, coverage is still often skewed toward higher-resource ones: Sitaram et al. (2025) covers 37 languages, Kocmi et al. (2025) 30, and Fu and Liu (2025) 25. However, only 36–47% of these language lists can be considered as low-resource. Marchisio et al. (2024) cover 22 languages but only 4 (18%) of them can be considered as low-resource. Only 8 of the 33 papers (24%) include low-resource languages as at least half of their evaluation languages.

[4]The string "LLM" appeared only as part of the cited name "Fi**llm**ore" (Pimentel, 2012; Minnema et al., 2022).

[5]We use a binary classification of language resource level, assigning each language to either the low-resource or high-resource category. While this pre-LLM classification, based on labeled-data availability, may not fully capture resourcedness of languages at the time of writing. We still adopt it because it remains the most widely cited reference in the literature and provides a transparent, reproducible criterion for our purposes.

Figure 1: Taxonomy of the 33 papers that use LLM-as-a-Judge in an **evaluator** role (i.e., assessing model-generated or human-produced outputs) in multilingual settings. Papers are organized along five dimensions: **Judging Paradigm** (how the judge scores), **Downstream NLP Tasks**, **Judge LLMs** (model families used), **Language Coverage**, and **Evaluation of the Judge** (how judge reliability is assessed). A paper may appear in multiple categories.

The distribution gets even more skewed under a stricter cutoff that excludes Joshi Class 3, the so-called "Rising Stars" of low-resource NLP (e.g., Indonesian, Romanian, and Thai), which already have substantial LLM support. Only 13/33 papers (39%) include at least one low-resource language under this stricter definition.

The low-resource languages that do receive attention are themselves unevenly distributed. Among the 8 papers in which low-resource languages constitute at least half of the covered languages, 6 focus primarily on South Asian languages (Watts et al., 2024; Sato et al., 2024; Qian et al., 2024; Lu et al., 2025; Doddapaneni et al., 2025; Duwal et al., 2025), while the remaining cases represent smaller pockets of research on Yorùbá (Okewunmi et al., 2025) and Romanian (Niculae et al., 2025). Many low-resource languages are entirely or nearly entirely absent. Indigenous American and Pacific Island languages do not appear in any of the papers in the survey. Sub-Saharan African languages are represented only by Maasai, Swahili, and Yorùbá, which are precisely the settings where dominant judge models are the weakest and where the human reference data needed to validate them is scarcest (Adelani et al., 2024; Smart et al., 2024; Ojo et al., 2025).

**Narrow and Closed LLMs in the Judge Ecosystem** The current judge ecosystem is dominated by closed and proprietary models. More precisely, GPT-family models are used in 26/33 papers (79%), and 14/33 (42%) use *only* closed-source judges, compared with just 4/33 (12%) that use *only* open-source ones. Cross-judge validation is not a common practice. 16/33 papers (48%) rely on a single judge model family, and 11/33 (33%) use GPT as their sole judge. Among open-source generalist alternatives, Llama (12/33) and Qwen (7/33) are the most common. However, in most papers they appear alongside GPT rather than on their own (i.e., 10 of the 12 Llama papers and 5 of the 7 Qwen papers also use a GPT model as judge). The majority of judges are generalist instruction-tuned LLMs, repurposed via rubrics or comparison prompts. Only 7/33 papers (21%) use a model fine-tuned or specialized for the evaluation task itself, including safety classifiers (Gupta et al., 2024), judge-tuned models like JudgeLM (Bennie et al., 2025; Farhan, 2025; Preciado Márquez et al., 2025), a domain-fine-tuned variant (Niculae et al., 2025), and evaluator models fine-tuned for specialized evaluation tasks such as multilingual coherence (Mendonca et al., 2024) or cross-lingual evaluation (Doddapaneni et al., 2025).

**Problems with Validation** The third pattern concerns how judge reliability is checked. On the surface, validation appears widespread. That is to say, 24/33 papers (73%) report some form of human comparison or expert evaluation of the judge in at least one deployment language. However, this aggregate figure could be obscuring a subtler problem. Sometimes the more revealing question is not *whether* papers validate, but ***which* languages they validate on**. Some papers that deploy LLM-as-a-Judge in multilingual settings check reliability only on a subset of languages. For example, Mendonca et al. (2024) build a dialogue coherence judge for five languages (i.e., English, French, German, Italian, and Chinese). However, the check against human ratings is performed on an existing English-only benchmark of human-annotated dialogue responses. In other words, the reliability of LLM-as-a-Judge in the other languages (i.e., French, German, Italian, and Chinese) is never compared to a human judgment. This choice is convenient since a human-annotated English benchmark is readily available. However, English is also the language where the reliability of LLM-as-a-Judge is perhaps the least controversial (considering how extensively frontier models have been evaluated on it). As a result, the validation only takes place where it is least needed. On the other hand, the languages for which an independent check of validity would be more informative are left unchecked.

The same pattern is also observed when LLM-as-a-Judge is deployed directly in low-resource languages. Three papers in our corpus use LLM-as-a-Judge without any human or gold-label check. Duwal et al. (2025) use GPT-4o as the only judge for Nepali generation. Instead of comparing its Nepali outputs to human or gold labels, this evaluation choice is justified by citing prior work on GPT-4o's multilingual coverage. Thakur et al. (2025) deploy an LLM judge across 18 languages, including low-resource ones (e.g., Telugu, Swahili, Yorùbá, Bengali, Indonesian, and Thai). They treat the judge's pairwise preferences as gold labels by default, so the judge's reliability is assumed rather than tested. Okewunmi et al. (2025) use GPT-4o, Gemini 2.0 Flash, and Claude 3.7 Sonnet to score Yorùbá question answering without human evaluation of the outputs by the LLM judges.

These decisions are difficult to justify, because the existing evidence (both within our corpus and outside it) indicates thatjudge reliability is language-conditional rather than uniform. Among the papers that test judges directly in low-resource languages, this finding is consistent. Watts et al. (2024) report that human-LLM agreement drops for direct assessment, particularly for Bengali and Odia. Fu and Liu (2025) report an average Fleiss' Kappa of approximately 0.3 across 25 languages, with consistency being particularly poor in low-resource languages, and find that neither multilingual training nor model scale directly improves this result. Studying GPT-4 as an evaluator across eight languages, Hada et al. (2024b) document a bias toward higher scores when human opinions differ and it is most pronounced in lower-resource and non-Latin-script languages.

Therefore, the problem is not that validation is rare. It is rather the case that validation is often skipped for low-resource languages. In these cases, it is not clear that the LLM-as-a-Judge evaluation can be trusted without human validation, given the underrepresentation of low-resource languages in LLM training to begin with.

## 5 Discussion and Recommendations

Evaluation of an NLP system is inherently difficult, no matter if it is carried out by humans or LLMs. To be able to perform the evaluation properly with humans takes a lot of time, effort and resources (e.g., recruiting and training qualified and representative/suitable annotators, collecting multiple annotations per items for a task in a language, identifying and understanding disagreements between annotators, making decisions about how the annotations are aggregated and used for evaluation). Therefore, it could be tempting to ignore many of these difficulties with human judgements and opt for LLMs which seem faster and cheaper. However, by reviewing the relevant literature (section 3), we have identified several recurring issues with this practice for languages other than English, particularly for low-resource ones. (section 4). Based on our survey, we provide the following recommendations for future research using LLM-as-a-Judge.

**Recommendation 1:** *Validate LLMs in multilingual contexts and for low-resource languages.* While using LLM-as-a-Judge, it is crucial to investigate whether a judge is reliable for a given language. For instance, Zheng et al. (2023) evaluate LLM-as-a-Judge on an English multi-turn question answering dataset. Boughorbel and Hawasly (2023) translate this dataset into Arabic and use the same LLM for judging model outputs. However, their claim that the judge is reliable is only based on the English dataset and is not re-investigated for Arabic. Given the fact that LLM judges tend to overestimate the quality of text generated by the same LLM (Liu et al., 2024a; Panickssery et al., 2024), it is especially important to properly validate a judge's performance on languages that are only covered by few language models, where the set of potential generator and judge models is extremely small. In multilingual contexts, it is particularly important to compare the validity of LLM-as-a-Judge across the target languages (Mendonca et al., 2024; Cruz Blandón et al., 2025).

**Recommendation 2:** *Keep humans in the loop.* Combining metrics and including human evaluations is a standard procedure in fields like speech synthesis, where automatic metrics capture different properties of the output and an ideal output depends on the context in which a system is deployed (Wagner et al., 2019). Similarly, shared tasks for MT systems often complement automatic evaluation metrics with human evaluation on a subset of the data to rank participants' systems (e.g., Kocmi et al., 2025). When LLM-as-a-Judge is used for low-resource languages, human evaluation (at least on a small subset of the data) and statistical confidence estimates should also be conducted (cf. Zheng et al., 2023; Niklaus et al., 2025).

It is also important to acknowledge that humans might disagree on how to judge a given text (Plank, 2022) or may introduce their own biases (Chen et al., 2024b). This can depend on many factors, including lack of familiarity with the specific dialect a text is written in (Keleg et al., 2024). When using LLM-as-a-Judge, there is a need to document *which* humans the judge LLM is compared to (e.g., which human population the human evaluators represent Doğruöz et al., 2023), and which populations are modelled (or not modelled) by the LLM judges. Researchers should document their selection criteria, and both the language competence and domain expertise of the human evaluators, as well as any other relevant criteria beyond their academic status.

**Recommendation 3:** *Check if non-LLM-based metrics are equally or more feasible.* While LLM-as-a-Judge is a new paradigm for evaluating LLM-generated outputs or human annotations, there is a

need to carefully consider which existing metrics can actually be replaced by LLM judges, and which should still be usedalongside them. This should highly depend on the competence of the LLM on the target language and their agreement with human judgment. For instance, traditional metrics (e.g, Exact Match and F1-score) in Question Answering may prefer short answers while LLM generations are generally more verbose. LLM-as-a-Judge has been shown to provide better and more flexible extraction of answers than these reference-based metrics (Adlakha et al., 2024b). However, this is only validated for English, and even in this setting models often do not stay faithful to provided references. If references are available for a dataset, conventional if flawed metrics (e.g., chrF) should still be used alongside LLM judges to catch cases where the LLM that is being judged generated text in the wrong language variety.

The weaker the LLM's competence in a language, the more errors and biases are likely to be introduced by LLM-as-a-Judge evaluation (e.g., rewarding longer answers (Zheng et al., 2023; Chen et al., 2024b)), preferring their own generations in direct comparison (Liu et al., 2024a; Panickssery et al., 2024), or relying on their own internal knowledge and ignoring explicit references (Lee et al., 2026). Therefore, LLM-as-a-Judge should only be used when there is absolutely no better or equally good existing metric is available. In addition, it should be used in conjunction and comparison with validated metrics (if possible) but it should be avoided when there is a high risk of conflicting information between task-specific references and the LLM's learned representation.

**Recommendation 4:** *Consider real-world relevance and representativeness beyond language.* The amount of available language data is not the only relevant dimension with respect to using LLM-as-a-Judge in low-resource settings, and failures on different dimensions should be distinguished. For instance, linguistic competence and cultural competence can be two independent dimensions. Watts et al. (2024) show that LLM evaluators agree less with humans on evaluating responses with culturally specific nuances. LLMs reproduce and amplify social and cultural stereotypes in their outputs (Mitchell et al., 2025) and even introduce these (without being asked) in text-to-image generation, purely based on the language the prompt is written in (Holtermann et al., 2026). Besides investigating cultural competence, future work could also focus on low-resource domains (e.g., how well LLM-as-a-Judge works in medical or legal contexts across languages, cf. Diekmann et al., 2025; Spiegel et al., 2025; Rolshoven et al., 2025).

## 6 Conclusion

Evaluation is an important and necessary part of NLP pipelines, and it can be challenging in multilingual and low-resource settings. Currently, LLMs are widely used in the evaluation process across NLP tasks. Our goal in this paper was to explore the use of LLM-as-a-Judge paradigm for multilingual and low-resource settings in the literature. While application of LLM-as-a-Judge increasingly includes low-resource languages, our survey shows that this does not automatically lead to inclusive application, as validation of LLM judges is still mostly done for English only. Although LLM judges may reduce the annotation costs and may deliver faster results, this benefit is limited when their judgments require additional verification or when the target language is weakly supported by the LLM judge. Therefore, human validation remains essential, especially for low-resource languages, where judge reliability cannot be assumed by default. Ideally, studies should validate LLM judges on a representative subset of target-language examples, report agreement with human or gold-label references, and clearly document the limitations of automatic judgments. If this is not done, the ease and ubiquity of LLM-as-a-Judge evaluation may become its own downfall. Although there is an increase in multilingual NLP evaluations, many of the included languages are still low-resource. Therefore, LLM performance may not be properly validated on them before the LLMs are used as judges on these same languages. This cycle will degrade the evaluation quality while superficially pretending to increase the representation breadth.

We hope that the results of our study will raise awareness about the limitation of LLMs-as-a-Judge paradigm in critical evaluation contexts and our recommendations will serve as guidelines for NLP researchers who work on low-resource languages and multilingual settings.

## Limitations

Although we performed a thorough search of the ACL Anthology, we might have missed papers by not searching for specific language names or includ-

ing other search terms (e.g., “dialects”, “varieties”). Furthermore, our keyword search only matches papers whose titles and/or abstracts are in English. We only cover papers that appeared in the ACL Anthology (in order to point out current research trends in the *ACL community), but other papers (or paper drafts) using LLM-as-a-Judge in multilingual settings and/or for LRLs might be in other research venues or on preprint servers.

Our paper consists of an analysis of already published papers rather than new experiments. This is in part precisely due to the challenges (e.g., lack of data, lack of good existing metrics for NLG tasks in LRL languages, difficulties with finding human experts for evaluating the LLM judges) of working with LRL data and in part because our aim is to show shortcomings of current research trends in *ACL research.

## Ethical Considerations

The annotators of the ACL Anthology papers are the authors of this paper, who carried out the annotation work as part of their jobs.

We use data from the ACL Anthology in accordance with one of its intended uses: academic research about NLP research (cf. Bollmann et al., 2023).

In parts of this paper, LLMs were used for helping with phrasing and/or correcting English (e.g., grammar, mistakes), and for assisting with the preparation of figures and tables.

## References


David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024a. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024b. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Tadesse Destaw Belay, Henok Biadglign Ademtew, Idris Abdulmumin, Sukairaj Hafiz Imam, Abubakar Juma Chilala, Godfred Agyapong, Chinedu Emmanuel Mbonu, Basil Friday Ovu, Catherine Nana Nyaah Essuman, Alfred Malengo Kondoro, Sonia Adhiambo, Daud Abolade, Ponts’o Mpholle, Nicholaus Dismas Ladislaus, Saminu Mohammad Aliyu, Gali Ahmad Samuel, Fabrice Hakuzimana, Mike Nzirainengwe, Temitayo Olatoye, and 8 others. 2026. Trust but check: LLM-assisted review of human translations in african languages. In *7th Workshop on African Natural Language Processing*.

Samuel Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-Jussà. 2025. Translate, then detect: Leveraging machine translation for cross-lingual toxicity classification. In *Proceedings of the Tenth Conference on Machine Translation*, pages 253–268, Suzhou, China. Association for Computational Linguistics.

Michael Bennie, Bushi Xiao, Chryseis Xinyi Liu, Demi Zhang, Jian Meng, and Alayo Tripp. 2025. CODEOFCONDUCT at multilingual counterspeech generation: A context-aware model for robust counterspeech generation in low-resource languages. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 37–46, Abu Dhabi, UAE. Association for Computational Linguistics.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the ACL Anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore. Association for Computational Linguistics.

Sabri Boughorbel and Majd Hawasly. 2023. Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. In *Proceedings of ArabicNLP 2023*, pages 128–139, Singapore (Hybrid). Association for Computational Linguistics.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

Mojca Brglez, Špela Vintar, and Aleš Žagar. 2024. How human-like are word associations in generative models? An experiment in Slovene. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 42–48, Torino, Italia. ELRA and ICCL.

Khaoula Chehbouni, Mohammed Haddou, Jackie CK Cheung, and Golnoosh Farnadi. 2025. Neither valid nor reliable? Investigating the use of LLMs as judges. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Mouxiang Chen, Hao Tian, Zhongxin Liu, Xiaoxue Ren, and Jianling Sun. 2024c. JumpCoder: Go beyond autoregressive coder via online modification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11500–11520, Bangkok, Thailand. Association for Computational Linguistics.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352.

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. MEMERAG: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22577–22595, Vienna, Austria. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 technical report. *arXiv*, 2412.19437 (v2).

Peter Devine. 2024. Are you sure? Rank them again: Repeated ranking for better preference datasets. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 93–105, Miami, Florida, USA. Association for Computational Linguistics.

Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah, and Joyce Ho. 2025. LLMs as medical safety judges: Evaluating alignment with human annotation in patient-facing QA. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 217–224, Viena, Austria. Association for Computational Linguistics.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA. Association for Computational Linguistics.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, and Mitesh M Khapra. 2025. Cross-lingual auto evaluation for assessing multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29297–29329, Vienna, Austria. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Antoine Dussolle, Andrea Cardeña Díaz, Shota Sato, and Peter Devine. 2025. M-IFEval: Multilingual

instruction-following evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6176–6191, Albuquerque, New Mexico. Association for Computational Linguistics.

Sharad Duwal, Suraj Prasai, and Suresh Manandhar. 2025. Domain-adaptative continual learning for low-resource tasks: Evaluation on Nepali. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 144–153, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Lukas Ellinger and Georg Groh. 2025. It depends: Resolving referential ambiguity in minimal contexts with commonsense knowledge. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 229–246, Suzhou, China. Association for Computational Linguistics.

Fahim Faisal, Md Mushfiqur Rahman, and Antonios Anastasopoulos. 2025. Dialectal toxicity detection: Evaluating LLM-as-a-judge consistency across language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12429–12452, Suzhou, China. Association for Computational Linguistics.

Md Shariq Farhan. 2025. Hyderabadi pearls at multilingual counterspeech generation : HALT : Hate speech alleviation using large language models and transformers. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 65–76, Abu Dhabi, UAE. Association for Computational Linguistics.

Jessica Zosa Forde, Ruochen Zhang, Lintang Sutawika, Alham Fikri Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19476–19493, Miami, Florida, USA. Association for Computational Linguistics.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, 2403.05530.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv*, 2411.15594.

Prannaya Gupta, Le Qi Yau, Hao Han Low, I-Shiang Lee, Hugo Maximus Lim, Yu Xin Teoh, Koh Jia Hng, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2024. WalledEval: A comprehensive safety evaluation toolkit for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 397–407, Miami, Florida, USA. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025. Do we still need human annotators? Prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 153–172, Vienna, Austria. Association for Computational Linguistics.

Carolin Holtermann, Florian Schneider, and Anne Lauscher. 2026. SoS: Analysis of surface over semantics in multilingual text-to-image generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3955–3995, Rabat, Morocco. Association for Computational Linguistics.

Amir Ivry and Shinji Watanabe. 2026. LALM-as-a-judge: Benchmarking large audio-language models for safety evaluation in multi-turn spoken dialogues. *arXiv*, 2602.04796.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the level of dialectness predicts inter-annotator agreement in multi-dialect Arabic datasets.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–789, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.

Dongryeol Lee, Yerin Hwang, Taegwan Kang, Minwoo Lee, Younhyung Chae, and Kyomin Jung. 2026. Judging against the reference: Uncovering knowledge-driven failures in LLM-judges on QA evaluation. *arXiv*, 2601.07506.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods. *arXiv*, 2412.05579.

Miaoran Li, Jiangning Chen, Minghua Xu, and Xiaolong Wang. 2025b. Hallucination detection in structured query generation via LLM self-debating. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16102–16113, Suzhou, China. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025c. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12979–12998, Suzhou, China. Association for Computational Linguistics.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024a. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, and Nancy Chen. 2024b. Optimizing code-switching in conversational tutoring systems: A pedagogical framework and evaluation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Kyoto, Japan. Association for Computational Linguistics.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

John Mendonca, Isabel Trancoso, and Alon Lavie. 2024. ECoh: Turn-level coherence evaluation for multilingual dialogues. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 516–532, Kyoto, Japan. Association for Computational Linguistics.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

*Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Sourabrata Mukherjee, Atul Kr. Ojha, John P. McCrae, and Ondřej Dušek. 2025. Evaluating text style transfer evaluation: Are there any reliable metrics? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 418–434, Albuquerque, USA. Association for Computational Linguistics.

Andrei Niculae, Adrian Cosma, Cosmin Dumitrache, and Emilian Radoi. 2025. Dr. copilot: A multi-agent prompt optimized assistant for improving patient-doctor communication in Romanian. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1780–1792, Suzhou (China). Association for Computational Linguistics.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. SwiLTra-bench: The Swiss legal translation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Paul Okewunmi, Favour James, and Oluwadunsin Fajemila. 2025. Evaluating robustness of LLMs to typographical noise in Yorùbá QA. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 195–202, Vienna, Austria. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *arXiv*, 2303.08774.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Janine Pimentel. 2012. Identifying equivalents of specialized verbs in a bilingual comparable corpus of judgments: A frame-based methodology. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1791–1798, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Salvador Preciado Márquez, Helena Gómez Adorno, Ilia Markov, and Selene Baez Santamaria. 2025. NLP@IIMAS-CLTL at multilingual counterspeech generation: Combating hate speech using contextualized knowledge graph representations and LLMs. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 29–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

Ravi Shanker Raju, Swayambhoo Jain, Bo Li, Jonathan Lingjie Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for LLM-as-a-judge. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 167–181, Miami, Florida, USA. Association for Computational Linguistics.

Luca Rolshoven, Vishvaksenan Rasiah, Srinanda Brügger Bose, Sarah Hostettler, Lara Burkhalter, Matthias Stürmer, and Joel Niklaus. 2025. Unlocking legal knowledge: A multilingual dataset for judicial summarization in Switzerland. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15382–15411, Suzhou, China. Association for Computational Linguistics.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through LLM-driven active learning and human annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 98–111, St. Julians, Malta. Association for Computational Linguistics.

Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2024. TMU-HIT's submission for the WMT24 quality estimation shared task: Is GPT-4 a good evaluator for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pages 529–534, Miami, Florida, USA. Association for Computational Linguistics.

Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. 2025. A multilingual, culture-first approach to addressing misgendering in LLM applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31159–31183, Suzhou, China. Association for Computational Linguistics.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. Socially responsible data for large multilingual language models. *arXiv*, 2409.05247.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: a multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models. *arXiv*, 2410.17578.

Seyoung Song, Seogyeong Jeong, Eunsu Kim, Jiho Jin, Dongkwan Kim, Jay Shin, and Alice Oh. 2025. MUG-eval: A proxy evaluation framework for multilingual generation capabilities in any language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19488–19514, Suzhou, China. Association for Computational Linguistics.

Sören Spiegel, Seid Muhie Yimam, Philipp Breitfeld, and Frank Ückert. 2025. Adaption and evaluation of generative large language models for German medical information extraction. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 33–47, Hannover, Germany. HsH Applied Academics.

Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "Knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad. 2025. MIRAGE-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 274–298, Albuquerque, New Mexico. Association for Computational Linguistics.

Hanh Thi Hong Tran and Nguyen Tien Nam. 2025. L3i++ at GenAI detection task 1: Can label-supervised LLaMA detect machine-generated text? In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 155–160, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Bibek Upadhayay and Vahid Behzadan. 2025. X-guard: Multilingual guard agent for content moderation. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pages 54–86, Vienna, Austria. Association for Computational Linguistics.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and Jana Voße. 2019. Speech synthesis evaluation – state-of-the-art assessment and suggestion for a novel research program. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 105–110.

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking STS and NLI in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982, St. Julian's, Malta. Association for Computational Linguistics.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.

Linda Wiechetek, Flammie A Pirinen, and Maja Lisa Kappfjell. 2025. How to create treebanks without human annotators – an indigenous language grammar checker for treebank construction. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 119–128, Ljubljana, Slovenia. Association for Computational Linguistics.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena.

In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

## A Per-paper language coverage

This appendix provides a per-paper breakdown of language coverage across the 33 in-scope papers (Table 1). For each paper, we list the languages covered, sorted by Joshi et al.'s (2020) language resource taxonomy. We treat languages of class 0–3 as low-resource. Note that this taxonomy is from 2020 and might thus underestimate the current availability of language resources. Also note that the framing of a language's resourcedness may be different in the papers surveyed. E.g., the surveyed papers that include Basque (class 4) frame it as a low-resource language (Bennie et al., 2025; Farhan, 2025; Preciado Márquez et al., 2025).

| Paper | #LR | Language resource classes (Joshi et al., 2020) and languages |
|---|---|---|
| Sitaram et al. (2025) | 15 | *0: Slovene / 1: Norwegian, Welsh / 3: Bulgarian, Danish, Estonian, Greek, Hebrew, Indonesian, Latvian, Lithuanian, Romanian, Slovak, Thai, Ukrainian /* 4: Catalan, Croatian, Czech, Dutch, Finnish, Hungarian, Italian, Korean, Polish, Portuguese, Russian, Serbian, Swedish, Turkish, Vietnamese / 5: Arabic, Chinese, English, French, German, Japanese, Spanish |
| Kocmi et al. (2025) | 14 | *0: Maasai / 1: Bhojpuri, Kannada / 2: Icelandic, Marathi / 3: Bengali, Egyptian Arabic, Estonian, Greek, Indonesian, Lithuanian, Romanian, Thai, Ukrainian /* 4: Czech, Hindi, Italian, Korean, Persian, Russian, Serbian, Swedish, Turkish, Vietnamese / 5: Chinese, English, French, German, Japanese, Spanish |
| Watts et al. (2024) | 9 | *1: Gujarati, Kannada, Malayalam, Odia, Telugu / 2: Marathi, Punjabi / 3: Bengali, Tamil /* 4: Hindi |
| Fu and Liu (2025) | 9 | *1: Telugu / 2: Swahili / 3: Bengali, Greek, Hebrew, Indonesian, Romanian, Thai, Ukrainian /* 4: Czech, Dutch, Hindi, Italian, Korean, Portuguese, Russian, Turkish, Vietnamese / 5: Arabic, Chinese, English, French, German, Japanese, Spanish |
| Thakur et al. (2025) | 6 | *1: Telugu / 2: Swahili, Yoruba / 3: Bengali, Indonesian, Thai /* 4: Finnish, Hindi, Korean, Persian, Russian / 5: Arabic, Chinese, English, French, German, Japanese, Spanish |
| Qian et al. (2024) | 5 | *0: Sinhala / 1: Nepali / 2: Marathi / 3: Estonian, Romanian /* 4: Russian / 5: Chinese, English, German |
| Lu et al. (2025) | 4 | *1: Assamese, Kannada, Maithili / 2: Punjabi /* 4: Russian / 5: Chinese, English, German |
| Marchisio et al. (2024) | 4 | *3: Greek, Indonesian, Romanian, Ukrainian /* 4: Czech, Dutch, Hindi, Italian, Korean, Persian, Polish, Portuguese, Russian, Turkish, Vietnamese / 5: Arabic, Chinese, English, French, German, Japanese, Spanish |
| Sato et al. (2024) | 3 | *1: Gujarati, Telugu / 3: Tamil /* 4: Hindi / 5: English |
| Doddapaneni et al. (2025) | 3 | *1: Telugu / 3: Bengali, Urdu /* 4: Hindi / 5: French, German |
| Hada et al. (2024a) | 2 | *2: Swahili / 3: Bengali /* 4: Hindi, Russian, Turkish / 5: Arabic, Chinese, English, French, Japanese |
| Raju et al. (2024) | 2 | *3: Slovenian, Thai /* 4: Hungarian, Russian, Serbian, Turkish / 5: Arabic, Japanese |
| Duwal et al. (2025) | 1 | *1: Nepali* |
| Niklaus et al. (2025) | 1 | *1: Romansh /* 4: Italian / 5: English, French, German |
| Okewunmi et al. (2025) | 1 | *2: Yorùbá /* 5: English |
| Forde et al. (2024) | 1 | *3: Indonesian /* 5: Chinese, English |
| Gupta et al. (2024) | 1 | *3: Filipino /* 4: Hindi, Russian, Serbian / 5: Arabic, English, French, Spanish |
| Mukherjee et al. (2025) | 1 | *3: Bengali /* 4: Hindi / 5: English |
| Niculae et al. (2025) | 1 | *3: Romanian* |
| Devine (2024) | 0 | 4: Russian / 5: Chinese, English, French, German, Japanese; also 62 non-specified languages |
| Hada et al. (2024b) | 0 | 4: Czech, Italian, Portuguese / 5: Chinese, English, French, German, Japanese, Spanish |
| Liu et al. (2024b) | 0 | 4: Korean / 5: Chinese, English |
| Mendonca et al. (2024) | 0 | 4: Italian / 5: Chinese, English, French, German |
| Bennie et al. (2025) | 0 | 4: Basque, Italian / 5: English, Spanish / |
| Cruz Blandón et al. (2025) | 0 | 4: Hindi / 5: English, French, German, Spanish |
| Ellinger and Groh (2025) | 0 | 4: Russian / 5: Arabic, Chinese, English, French |
| Farhan (2025) | 0 | 4: Basque / 5: English |
| Preciado Márquez et al. (2025) | 0 | 4: Basque, Italian / 5: English, Spanish |
| Rolshoven et al. (2025) | 0 | 4: Italian / 5: French, German |
| Umutlu et al. (2025) | 0 | 4: Turkish |
| Boughorbel and Hawasly (2023) | 0 | 5: Arabic, English |
| Boughorbel et al. (2024) | 0 | 5: Arabic |
| Dinh et al. (2024) | 0 | 5: English, German |

Table 1: Language coverage per surveyed paper (n=33). Languages are sorted using Joshi et al.'s (2020) taxonomy (0 = least resources, 5 = most resources). "LR" = low-resource, counting languages from classes 0–3 (language names in italics). Papers are sorted by number of LR languages covered (descending), then by lowest resource class included, then publication year (ascending), then alphabetically by author. For MT, both source and target languages are counted.