Cross-Lingual LLM-Judge Transfer via Evaluation Decomposition

Ivaxi Sheth[1*] **Zeno Jonke**[2] **Amin Mantrach**[2] **Saab Mansour**[2]

[1]CISPA Helmholtz Center for Information Security
[2]Amazon

Abstract

As large language models are increasingly deployed across diverse real-world applications, extending automated evaluation beyond English has become a critical challenge. Existing evaluation approaches are predominantly English-focused, and adapting them to other languages is hindered by the scarcity and cost of human-annotated judgments in most languages. We introduce a decomposition-based evaluation framework built around a Universal Criteria Set (UCS). UCS consists of a shared, language-agnostic set of evaluation dimensions, producing an interpretable intermediate representation that supports cross-lingual transfer with minimal supervision. Experiments on multiple faithfulness tasks across languages and model backbones demonstrate consistent improvements over strong baselines without requiring target-language annotations.

1 Introduction

The rapid growth of general-purpose AI systems has led to a dramatic increase in machine-generated text across applications such as summarization (Pu et al., 2023), question answering (Yue, 2025), content moderation (Kolla et al., 2024), and search (Spatharioti et al., 2023). Assessing the quality and correctness of these outputs at scale remains a central challenge (Gu et al., 2025; Ohde et al., 2025). Although human annotation is the gold standard, it is expensive, time-consuming, and difficult to scale to the volume and frequency required by modern development cycles (Ohde et al., 2025; Gu et al., 2025).

Traditional automatic evaluation metrics for text generation, such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), have long served as proxies for output quality. However, they rely primarily on surface-level lexical overlap and fail to capture semantic meaning, reasoning, and contextual appropriateness (Sulem et al., 2018). LLM-based judges can be a more flexible alternative by leveraging broad pretrained knowledge and contextual reasoning. An LLM can be prompted to assess the quality of output using task-specific instructions or rubric-style criteria (Li et al., 2025; Zheng et al., 2023; Chen et al., 2024; Chiang et al., 2024), leading to their growing adoption for summarization, factuality assessment, preference modeling, and model comparison.

The development and benchmarking of LLM judges have mainly focused on English, despite only about 15% of the global population speaking English (Pombal et al., 2025). As AI systems are deployed worldwide, reliable evaluation across diverse languages becomes a critical bottleneck (Wang et al., 2024; Fu and Liu, 2025). Existing multilingual judge approaches face two primary challenges. First, many rely on language-specific fine-tuning, which requires human-annotated data that is scarce for most languages, and often underperform strong English-based baselines (Doddapaneni et al., 2025). Second, English-trained judges do not consistently transfer to typologically distant or low-resource languages (Fu and Liu, 2025). Together, these limitations hinder scalable and reliable multilingual evaluation.

These limitations become particularly apparent in the common scenario of language expansion, where systems developed and evaluated in English are later deployed to additional languages. While model capabilities can often be extended through translation or multilingual prompting, evaluation typically requires collecting new annotation data or adapting evaluation frameworks for each language. As a result, evaluation becomes a bottleneck in scaling LLM systems to new linguistic settings, highlighting the need for approaches that can transfer evaluation behavior from English to other languages with minimal additional supervision.

*Work done as an intern at Amazon. Correspondence to ivaxi.sheth@cispa.de.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 39727–39745
October 24-29, 2026 ©2026 Association for Computational Linguistics

In this paper, we introduce an interpretable decomposition-based framework for LLM judges that enables sample-efficient cross-lingual transfer. Our approach decomposes judgment into a shared set of language-agnostic evaluation criteria, each expressed as a targeted question about a specific dimension of quality. The resulting criterion-level responses define a structured intermediate representation that maps judgments into a predefined, human-interpretable evaluation space, in the spirit of concept-based models that ground intermediate representations in explicit semantic dimensions (Koh et al., 2020; Espinosa Zarlenga et al., 2022; Poeta et al., 2023; Sheth and Ebrahimi Kahou, 2023; Sun et al.), while also capturing reasoning patterns that transfer across languages. A lightweight transfer module trained only on English-labeled data can then be applied directly to new languages, enabling cross-lingual judge transfer without target-language supervision.

We evaluate our framework across multilingual benchmarks and multiple LLM backbones. Our contributions are: (1) we show that LLM-judge reasoning can be factored through a shared, language-agnostic criteria space, enabling consistent judgments across heterogeneous linguistic inputs; (2) we introduce an interpretable intermediate representation derived from criteria-level responses; (3) we demonstrate that this representation supports effective cross-lingual transfer using a lightweight module trained in English and applied to other languages without requiring target-language labels.

2 Related Works

LLM Judges Large language models are increasingly used as automated evaluators for summarization, dialogue, factuality, and preference modeling. Early work (Zheng et al., 2023; Liusie et al., 2024) demonstrated that LLMs can approximate human preferences using carefully designed prompts. Subsequent studies explored rubric-based scoring (Song et al., 2024), pairwise comparison (Liusie et al., 2024), and direct answer classification using models like GPT-4. However, these methods typically rely on single-prompt formulations that are brittle to prompt phrasing and struggle to generalize (Thakur et al., 2025).

Another line of work improves evaluation by incorporating explicit reasoning, such as chain-of-thought prompting (Zheng et al., 2023), justification-first scoring (Trivedi et al., 2024), or predefined criteria prompting (Wei et al.; Lee et al., 2025).

More recent approaches employ multiple LLMs interacting through debate (Feng et al., 2025; **?**), critique (Kim et al., 2024), or competitive assessment. While effective in certain settings, these systems require careful human input for selecting few-shot or seed prompts (Feng et al., 2025; Lee et al., 2025) and substantial role engineering (Alfano et al., 2026), making them costly and difficult to deploy reliably in production environments.

Checklist-style evaluation has recently gained attention as an interpretable alternative to monolithic rubric prompts. Previous work has shown that explicitly decomposing an evaluation task into smaller criteria can improve transparency. Wei et al. generate binary checklist items using a stronger model and apply them to smaller evaluators, effectively distilling high-level judgements into simple, verifiable units. (Lee et al., 2025) prompt models to generate their own criteria and use these checklists for iterative self-improvement, allowing the evaluator to refine its reasoning over multiple rounds.

Multilingual LLM-Based Judges Most LLM-judge design and evaluation has focused on English. Chang et al. (2025) studies the impact of resource availability on multilingual evaluators, while Fu and Liu (2025) analyzes the reliability of multilingual judges. Recent approaches train multilingual evaluators through large-scale pretraining (Pombal et al., 2025) or language-specific fine-tuning (Doddapaneni et al., 2025). To reduce reliance on human annotations, Alfano et al. (2026) generates synthetic multilingual supervision by translating English data and constructing corrupted summaries for training, and further examines how training language affects multilingual evaluation behavior; notably, their strongest results are obtained by fine-tuning on English data only.

However, all of these approaches require either task-specific data construction, full model fine-tuning, or language-specific adaptation. In contrast, our method produces a language-agnostic interpretable representation from a shared criterion set, enabling cross-lingual transfer via a lightweight module trained on only a few labeled English examples with no target-language supervision, no data engineering, and no full-model fine-tuning.

3 Methodology

We propose a criteria-based evaluation framework that decomposes an LLM-based judge's decision into a structured set of sub-criteria. Rather than relying on a single prompt or free-form reasoning, our judge produces answers to a set of targeted evaluation questions, which we refer to as criteria, and these answers are aggregated into an intermediate judge representation used for cross-lingual transfer.

Problem Formulation. Consider an evaluation task in which an input sample x must be assigned a binary label $y \in \{0, 1\}$. Let $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ denote a labeled dataset, where y_i is the ground-truth judgment for sample x_i. Each sample x can consist of any structured input relevant to the task, such as a source–output pair, an instruction–response pair, or any text bundle under evaluation. A standard LLM-based judge is obtained by prompting the model $\mathcal{M}$ with an evaluation instruction to produce a predicted label $\hat{y}_i$:

$$\hat{y}_i = \mathcal{M}(x_i). \tag{1}$$

Standard prompting produces $\hat{y}_i$ directly, either with chain-of-thought reasoning or via multi-agent orchestration.

In this work, we replace this single-prompt-based judgment with a decomposition-based framework in which the LLM first responds to a set of evaluation criteria. These responses form an intermediate judge representation that is used for transfer across languages.

Framework Overview. Our methodology consists of three stages. **Stage 1**: *Criteria Set Generation* constructs a set of language-agnostic evaluation criteria from the task specification. **Stage 2**: *Criteria Set Evaluation* applies these criteria to each input sample, producing a structured intermediate representation from the LLM's criterion-level responses. **Stage 3**: *Cross-Lingual transfer* trains a lightweight transfer module on labeled English data to align the judge representation, enabling transfer to other languages. An overview of the framework is shown in Figure 1.

3.1 Criteria Set Generation

Our approach first constructs a set of evaluation criteria that define the dimensions along which the LLM-based judge evaluates each sample. Formally, we generate a *Universal Criteria Set* (UCS), a collection of evaluation questions

$$Q = \{q_1, \ldots, q_k\} \tag{2}$$

where each q_k specifies a distinct judgment dimension. We generate criteria in English, motivated by evidence that LLMs exhibit more consistent reasoning and intermediate representations in English than in other languages (Schut et al.; Shi et al., 2023; Huang et al., 2024). This choice simplifies the generation of criteria and improves stability for downstream cross-lingual transfer.

The UCS defines a reusable set of evaluation dimensions applicable to all samples for a given task and is shared across all languages. Given a task description t, the LLM generates a set of evaluation concepts $C = \{c_1, \ldots, c_m\}$ representing generic evaluation dimensions such as accuracy, consistency, attribution, or hallucination. The UCS is task-specific, generated from the task description but *language-universal*: the same criteria set is applied to all languages for a given task.

$$C = \mathcal{M}(t) \tag{3}$$

See Appendix E for the prompts used in this stage.

Question Generation For each concept $c_j \in C$, the LLM generates one or more evaluation questions to comprehensively cover its sub-dimensions. Let Q_j denote the subset of criteria generated from the concept c_j, such that :

$$Q_j = \mathcal{M}(c_j) \tag{4}$$

The final Universal Criteria Set is obtained by aggregating the criteria generated for all concepts:

$$Q = \bigcup_{j=1}^{m} Q_j. \tag{5}$$

Thus, $Q = \{q_1, \ldots, q_k\}$ represents the union of all concept-derived questions. The same criteria set Q is applied to every sample, providing stable language-agnostic evaluation features that support cross-lingual transfer.

3.2 Criteria Set Evaluation

In the second stage of our framework, given the UCS Q and an input sample x, the LLM is prompted to assess the extent to which x satisfies each criterion $q_k \in Q$, returning a numerical score:

$$\mathcal{M}(q_k, x) = z_k(x) \in [1, \ldots, 10], \tag{6}$$

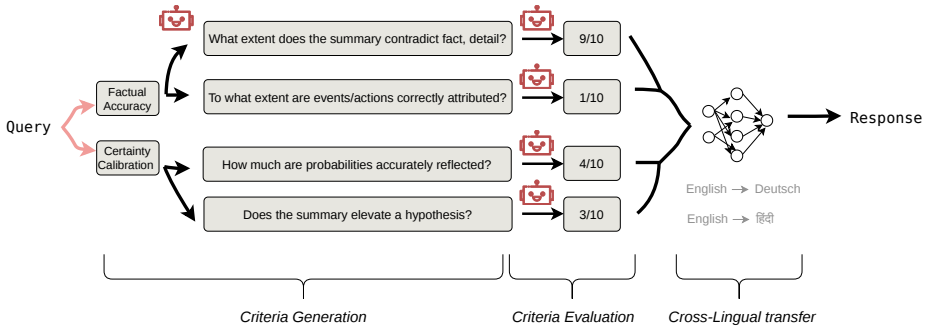


Figure 1: UCS framework. The LLM evaluates an input using criteria questions derived from higher-level concepts. Criterion scores are aggregated into a latent representation, which a transfer module trained on English-labeled data maps to the final judgment and applies across languages.

We consider Likert rating from 1-10 following (Lee et al., 2025). Other rating ranges is left to future explorations. Collecting these outputs yields the *criteria-response vector*:

$$z(x) = [\, z_1(x), \ldots, z_k(x) \,], \tag{7}$$

which serves as a structured and interpretable intermediate judge representation of the sample.

3.3 Cross-Lingual Transfer

The final stage learns a mapping from the criteria-response representation $z(x)$ to a calibrated judgment that generalizes across languages. The key insight is that while raw LLM outputs may vary across languages, the structure of evaluation as expressed through the shared criteria remains stable. The criteria-response vector $z(x)$ therefore defines a language-agnostic intermediate representation space in which a lightweight predictor trained on one language can be applied directly to another.

Concept-Level Aggregation. Rather than operating on the full criteria-response vector $z(x)$, we perform concept-level transfer by aggregating criterion-level scores within each concept. This reduces the dimensionality of the representation, decreases sample complexity, and improves transfer robustness (see Appendix C for details). Concretely, for each concept c_j, we average the scores of its associated criteria:

$$s_j(x) = \frac{1}{|Q_j|} \sum_{q_i \in Q_j} z_i(x). \tag{8}$$

The resulting concept-level representation is

$$s(x) = [\, s_1(x), \ldots, s_m(x) \,]. \tag{9}$$

Transfer. Given labeled data in a source language ℓ, we train a lightweight predictor (e.g. a neural network) on the concept-level representation:

$$\hat{y} = f_\theta(s(x)), \tag{10}$$

where θ is learned on the source-language dataset $\mathcal{D}^\ell = \{(x_i^\ell, y_i^\ell)\}_{i=1}^N$. At inference time, the trained predictor f_{θ^*} is applied directly to samples in a target language ℓ':

$$\hat{y}_j^{\ell'} = f_{\theta^*}(s(x_j^{\ell'})). \tag{11}$$

This enables cross-lingual transfer without any target-language supervision, as the shared criteria space ensures that $s(x)$ remains semantically consistent across languages.

Applying UCS to a new task. Given a new evaluation task, the practitioner specifies a natural-language task definition and, optionally, a small number of high-level evaluation concepts. An LLM first expands this specification into a set of task-relevant concepts and then generates several operational evaluation questions for each concept using the meta-prompts in Appendix E. This generation is performed once per task. The resulting criteria are written in English and fixed thereafter: the identical criteria are used to evaluate every example and every language. Only the source-language labeled examples used to learn the final transfer mapping are required; no target-language labels or translated criteria are needed.

4 Experimental Setup

Models. We evaluate our methods across a range of LLMs to ensure diversity in scale and architecture. We report the results for Qwen3-32B (Yang

Method	MEMERAG					mFACE								
	DE	FR	ES	HI	*Avg*	AM	MY	FR	SW	TH	AR	HI	ES	*Avg*
Zero-shot	77.2	77.5	77.5	74.4	76.7	54.7	**70.2**	74.3	73.0	72.6	69.7	68.0	77.9	70.1
CoT	73.0	76.2	76.2	74.1	74.9	58.7	68.5	72.8	77.5	71.4	69.2	69.4	74.4	70.2
AG	77.2	80.6	76.1	78.0	78.0	59.3	67.4	76.6	77.7	72.6	70.7	69.0	79.5	71.6
ChatEval	71.8	73.0	60.9	62.4	67.0	51.1	65.5	68.4	74.0	63.9	66.1	55.7	70.3	64.4
CheckEval	76.1	80.7	74.8	78.4	77.5	60.8	68.9	74.9	76.8	**75.1**	73.4	65.8	76.4	71.5
RocketEval	76.4	81.0	75.9	79.1	78.1	61.2	69.2	75.4	77.0	74.9	72.9	66.9	78.1	72.0
UCS (EN)	**79.7**	**84.3**	**80.0**	**82.2**	**81.6**	**64.7**	67.7	**82.4**	**79.0**	**75.1**	**74.6**	**70.0**	**85.5**	**74.9**

Table 1: Balanced Accuracy (BA) across methods and languages for the Qwen-235B model. Bold values indicate the best performance per column averaged across 3 runs. Our method UCS is trained on English only data (EN).

et al., 2025), Qwen3-235B-A22B (Yang et al., 2025), OSS-20B (OpenAI et al., 2025), and OSS-120B (OpenAI et al., 2025). All LLM-based judges are queried in a deterministic setting with `temperature` = 0 and `top_p` = 1.0 to reduce randomness in evaluation outputs.

Datasets. Our primary experiments use two multilingual evaluation benchmarks. **MEMERAG** (Blandón et al., 2025) is a multilingual RAG faithfulness benchmark containing evidence–query–answer triples, each annotated with a binary faithfulness label. The dataset spans five languages: English (EN), French (FR), German (DE), Hindi (HI), and Spanish (ES). **mFACE** (Aharoni et al., 2023) is a multilingual summarization evaluation dataset consisting of news articles and human-written summaries collected from BBC regional editions. The task is to judge whether a summary is faithful to the source article. We use a representative subset of languages spanning high-resource, mid-resource, and low-resource languages: English (EN), Amharic (AM), Burmese (MY), French (FR), Swahili (SW), Thai (TH), Arabic (AR), Hindi (HI), and Spanish (ES).

Evaluation Metrics. We report *balanced accuracy* (BA) to account for class imbalance, as it measures judge's sensitivity to both positive and negative classes equally.

Implementation Details. For each dataset and model, all methods are evaluated using consistent prompting templates and fixed criteria-generation procedures. All transfer models are trained exclusively on English-labeled data unless otherwise noted, and are applied to the remaining languages without any target-language supervision. We use shallow neural network for transfer from English to other languages (See B.3 for details).

Baselines. We evaluate our approach against a broad set of strong LLM-based judge baselines spanning single-model prompting, multi-agent methods, and checklist-based evaluators. First, we include zero-shot LLM judges following standard prompting setups from previous work (Blandón et al., 2025; Bavaresco et al., 2025), where the model is directly instructed to assess correctness or faithfulness without additional structure. We also include chain-of-thought (CoT) prompting, which has been shown to improve reasoning in LLM judges (Zheng et al., 2023). Another baseline we consider is prompting LLM with annotation guidelines (AG) that were given to humans, similar to (Blandón et al., 2025), see Appendix B.

We further compare against ChatEval (**?**), a multi-agent debate-style evaluator in which models critique and challenge each other's assessments. Such systems have demonstrated strong performance on complex reasoning and judgment tasks, but require substantial orchestration overhead. Finally, we include two recent checklist-based evaluation frameworks: RocketEval (Wei et al.) and CheckEval (Lee et al., 2025), which generate or refine structured criteria to guide LLM judgments. These methods share our goal of improving structure and interpretability in LLM-based evaluation; however, unlike our approach, they do not produce a unified latent judge representation that supports cross-lingual transfer [1].

5 Results

In this section, we evaluate the proposed criteria-based framework across languages and datasets. We focus on the language expansion setting and report the performance of UCS when trained on the English portion of the multilingual dataset, denoted by UCS (EN). We then present a series of analyses

[1]We reproduce the baseline results using publicly code.

that unpack the sources of these gains: sample-efficiency curves quantifying how much English supervision is required to reach stable cross-lingual performance (§ 5.2), an analysis studying the most predictive criterion dimensions (§ 5.3), a comparison of alternative transfer models (§ 5.4), a comparison of criterion aggregation strategies (§ 5.5) and an analysis of inference cost trade-offs (§ 5.6).

5.1 Cross-Lingual Transfer

Table 1 reports balanced accuracy (BA) for the Qwen-235B judge on MEMERAG and mFACE, comparing the proposed UCS framework trained on English data against prompting-based, debate-style, and checklist-based baselines. Across both datasets, UCS (EN) delivers the strongest and most consistent cross-lingual performance, achieving the best or tied-best results in the majority of evaluated languages (11 out of 12). Appendix Tables 6 and 7 report results averaged over multiple random seeds, showing that UCS maintains consistent improvements across languages while exhibiting low variance across runs.

On MEMERAG, UCS (EN) achieves the highest performance across all four non-English languages, with an average BA of 81.6 compared to 78.1 for RocketEval, the strongest baseline. Improvements are consistent across typologically diverse languages, ranging from 3.1 to 4.1 BA points and spanning Germanic (DE), Romance (FR, ES), and Indo-Aryan (HI) languages. This uniformity supports the claim that the criteria-based representation is genuinely language-agnostic for this task.

On mFACE, UCS (EN) achieves best-or-tied performance on seven of the eight evaluated languages, with an average BA of 74.9 compared to 72.0 for RocketEval. Notably, the largest improvements are observed for French (+7.0) and Spanish (+7.4) – two high-resource languages where strong baseline performance might be expected. This suggests that the structured criteria space captures evaluation dimensions relevant to summarization faithfulness that are not easily elicited through direct prompting, and that this benefit is most pronounced for languages where the LLM can reliably interpret and respond to the criteria. Consistent gains are also observed on lower-resource languages such as Amharic (+3.5) and Hindi (+3.1), further demonstrating the breadth of cross-lingual generalization.

One notable exception is observed on Burmese (MY), where the zero-shot baseline achieves the highest performance across all methods (70.2), outperforming not only UCS (67.7) but also all other baselines. This suggests that Burmese may exhibit language-specific properties, such as script complexity or limited LLM pretraining coverage, that are not well captured by the shared criteria space, and that direct prompting may be more robust in such cases. Despite this exception, UCS maintains the strongest overall performance profile across languages, demonstrating the robustness of the proposed criteria-based representation for cross-lingual judge transfer.Appendix Table 5 reports the full cross-lingual results across all judge backbones evaluated in this work. The results are consistent with the main findings: UCS-based judges achieve the best or near-best performance across most language–model combinations on both MEMERAG and mFACE, indicating that the improvements are not specific to an LLM backbone.

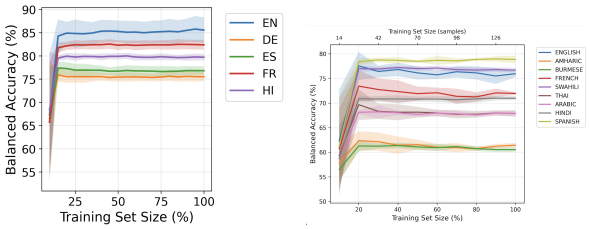
5.2 Sample Efficiency

A practical consideration for language expansion is how much labeled data in a high-resource language is required to enable effective transfer. Although our framework trains only a lightweight transfer module on English-labeled criterion features, understanding the supervision requirements for stable cross-lingual performance is important for assessing the practical viability of the approach.

To study this, we construct a fixed random split of the English data, holding out 30% as a test set. From the remaining 70%, we vary the amount of labeled training data used to fit the transfer module, ranging from 5% to 100% of the available training portion. Each resulting model is evaluated both on the held-out English test set and on all target languages, allowing us to examine how increasing English supervision affects both in-language performance and cross-lingual generalization.

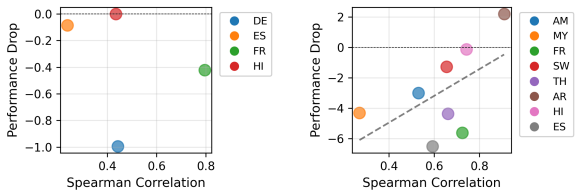
Figure 2 shows that performance improves rapidly with a small fraction of labeled data and stabilizes once approximately 20–30% of the English training data is used. Beyond this point, additional supervision yields only marginal gains across all languages.

Importantly, the same trend is observed for both English and target languages: as the transfer module improves on English, performance increases consistently across languages. This suggests that the criteria-based intermediate representation enables efficient cross-lingual generalization, requiring only a modest amount of labeled data in a single high-resource language.



(a) (b)

Figure 2: Sample efficiency of training the transfer module on English data. Performance is shown as a function of the percentage of English training data used. Results are reported for MEMERAG (a) and mFACE (b).



(a) MEMERAG (b) mFACE

Figure 3: Relationship between English–target criteria importance alignment and performance change when training with only the top-10 English-selected criteria.

5.3 Criteria Importance

Our transfer framework assumes that the criteria representation captures evaluation signals that are meaningful across languages. A natural question is whether the relative importance of these evaluation dimensions is preserved across languages. We hypothesize that languages whose importance profile criteria are more similar to English should exhibit stronger transfer performance.

To examine this, we analyze the importance of individual criteria before concept aggregation. While the transfer model operates on concept-level averages, criterion-level analysis provides a finer diagnostic view of how evaluation signals are prioritized across languages.

For each language ℓ, we train a Random Forest classifier using that language's criterion responses together with the corresponding human labels, and extract feature importance scores using Gini impurity reduction (Nembrini et al., 2018). We then compute the Spearman rank correlation (Spearman, 1961) between the English importance profile and the language-specific importance profile. To evaluate whether this alignment matters in practice, we simulate a constrained setting where only the top-10 criteria selected according to English importance are used to train the classifier for each language. We then measure the performance change relative to using the full criteria set. Figure 3 plots, for each language, the relationship between English–target importance correlation and the resulting performance change.

In mFACE, we observe that languages with lower importance alignment generally experience larger performance degradation. For example, Burmese (MY) shows one of the lowest correlations with English and also corresponds to one of the weaker-performing languages in Table 1. While this observation is only suggestive, it indicates that differences in evaluation priorities across languages can contribute to variation in cross-lingual transfer performance. For MEMERAG, we observe only minor performance differences when restricting the model to English-selected criteria; the drops remain below 1% across languages and do not show a clear relationship with importance alignment.

Figure 4 visualizes full distribution of importance scores across all evaluated languages where some languages exhibit importance profiles similar to English, others show noticeable differences in how evaluation dimensions are prioritized. Overall, these results indicate that the relative importance of evaluation criteria is not universally shared across languages. Selecting a fixed subset of criteria based solely on English importance can harm performance for languages with different importance profiles. This observation supports the design of our full criteria-based representation, which allows the transfer module to learn language-appropriate weightings rather.

5.4 Alternative Training Models

While our primary transfer model uses a shallow neural network to map criteria-response representations to final judgments, we also evaluate alternative transfer modules to assess the sensitivity of our framework to the choice of predictor. These include logistic regression as a linear baseline, SVMs for margin-based classification, KNN as an instance-based method, and tree-based models such as Random Forests and XGBoost that can capture nonlinear feature interactions.

Table 2 shows that several simple models perform competitively when operating on the criteria-based representation. Logistic regression and SVM achieve strong results, indicating that much of the signal captured by the criteria representation is lin-

Method	MEMERAG	mFACE
LogReg	80.9	78.2
SVM	78.5	**79.0**
KNN	77.1	70.9
RF	76.4	69.8
XGBoost	74.3	71.1
NN	**81.5**	78.9

Table 2: BA averaged across languages for different transfer modules on the Qwen-235B model.

Method	MEMERAG	mFACE
UCS (Trained)	**81.5**	**74.9**
UCS (LLM)	77.2	74.3

Table 3: Average BA across languages on MEMERAG and mFACE. UCS (Trained) uses a lightweight transfer module trained on English-labeled data; UCS (LLM) uses the LLM itself as a training-free aggregator over the criteria responses.

early separable.

The shallow neural network achieves the best performance on MEMERAG and remains competitive on mFACE, where SVM slightly outperforms it. Overall, the neural network provides the most consistent performance across datasets and languages.

In contrast, instance-based and tree-based methods perform substantially worse. These models appear to overfit to patterns in the English training data and generalize less effectively to other languages, suggesting that simpler parametric models are better suited for cross-lingual transfer here.

5.5 Transfer Learning vs LLM-based Aggregator

The criteria-response vector $z(x)$ produced by the second stage must be aggregated into a final binary judgment. In our primary framework, this is achieved by the lightweight transfer module f_θ trained on English-labeled data, which learns a calibrated mapping from the concept-level representation $s(x)$ to the predicted label $\hat{y}$.

An alternative we consider an *LLM-based aggregator*, in which the LLM is prompted to produce a final judgment given the input x, the criteria Q, and their corresponding responses $z(x)$. This approach requires no additional training and may leverage the LLM's broader world knowledge and reasoning capabilities to capture interactions between criteria. However, it does not benefit from calibration on human-labeled data and may be sensitive to prompt phrasing and model-specific biases.

Generation	Scoring	MEMERAG	mFACE
Joint	Joint	79.1	77.4
Joint	Per-concept	81.0	78.6
Per-concept	Joint	80.2	78.0
Per-concept	Per-concept	**81.5**	**78.9**

Table 4: Average BA across languages on MEMERAG and mFACE for UCS under different decompositions. Joint denotes evaluating all criteria in one LLM call, while Per-concept denotes one call per concept.

Table 3 demonstrates a trade-off between training-free and trained aggregation. The LLM-based aggregator does not require labeled data beyond the criteria responses and can be applied directly. However, when labeled English data is available, training a lightweight transfer module on the criteria-based representation consistently yields a higher average balanced accuracy on both datasets.

5.6 Inference Cost

Compared to zero-shot prompting, our criteria-based framework incurs additional inference cost due to explicit criteria generation and criterion-level evaluation. While a zero-shot judge requires a single LLM call per sample, our approach introduces structured evaluation steps that increase prompt length and the number of LLM calls.

To analyze the trade-offs between decomposition granularity and inference efficiency, we evaluate four prompt variants differing in how criteria are generated and scored. Generation can be joint i.e. all evaluation dimensions produced in one LLM call or per-concept, i.e. each concept generates its criteria separately). Scoring follows the same options: a single joint judgment over all criteria or separate per-concept scores.

Table 4 compares these prompt variants in terms of both inference cost and performance. Performance improves consistently as the evaluation process is more finely decomposed. In particular, scoring at the per-concept level consistently outperforms joint scoring over all criteria. The best results are obtained when both criteria generation and scoring are performed at the per-concept level, indicating that finer-grained criterion-level signals provide a more discriminative intermediate representation for the downstream transfer module.

6 Conclusion

In this paper, we introduced a decomposition-based framework for LLM judges that enables sample-

efficient cross-lingual transfer through a shared set of language-agnostic evaluation criteria. We proposed Universal Criteria Sets (UCS), which structure evaluation into explicit dimensions and produce a transparent intermediate representation of the judgment process. A lightweight transfer module trained on English-labeled data maps this representation to final judgments and generalizes directly to new languages without requiring target-language supervision. Across experiments on MEMERAG and mFACE, UCS achieves strong and consistent performance while requiring only a small amount of supervision in a single high-resource language. Beyond improved cross-lingual performance, the criteria-based representation provides interpretable insights into how evaluation signals contribute to final judgments, and our criteria-importance analysis reveals that the degree of cross-lingual alignment varies by task. Together, these results highlight the potential of structured, criteria-based representations as a principled foundation for building reliable and interpretable multilingual LLM-based judges. The structured nature of this representation also allows future work to explore criterion-level judgments to construct structured reward signals or rubric-based feedback for reinforcement learning and alignment.

7 Future Work

In this work, we evaluated cross-lingual transfer across multiple languages and datasets. Future work can extend this to other languages not evaluated here. In particular, language-specific analyses could help explain the strong zero-shot performance of the LLM Judge in Burmese. Figure 3 illustrates the relationship between English–target criteria importance alignment, showing that the relative importance of evaluation criteria can vary across languages and datasets. This suggests that not all criteria contribute equally to evaluation quality in cross-lingual settings. In particular, some criteria appear consistently less important across languages, while others show stronger and more stable alignment. These observations indicate the need to identify the most effective evaluation criteria for multilingual assessment. Future work could focus on selecting or learning criteria that provide the strongest signal across languages.

Beyond cross-lingual evaluation, the criteria-based decomposition introduced in this work opens several promising directions. Because the framework represents judgments through explicit evaluation dimensions, it provides a structured interface for analyzing and controlling LLM evaluation behavior.

Another promising direction is leveraging criteria representations to support human-in-the-loop evaluation. Structured criteria could allow human evaluators to provide targeted feedback on specific dimensions, which can then be incorporated to refine the transfer model or adjust evaluation standards. This may enable more transparent and controllable evaluation pipelines.

Finally, the criteria representation provides a natural foundation for studying the reliability of LLM judges. Future work could investigate uncertainty estimation, agreement across multiple judge models, or ensemble approaches that operate at the criterion level.

Future work could also explore how criterion-level representations can support reinforcement learning and alignment of language models. Current alignment methods often rely on scalar reward signals derived from preference comparisons or holistic judgments, which provide limited insight into why a response is preferred. In contrast, criterion-based evaluation decomposes quality into explicit dimensions, such as factual consistency, relevance, or completeness. These structured signals could be used to construct richer reward functions that guide models toward satisfying multiple evaluation dimensions simultaneously. Moreover, criterion-level feedback may enable more interpretable and controllable alignment, allowing training objectives to emphasize specific aspects of behavior or adapt across languages and domains. Investigating how such structured evaluation signals can be integrated into reinforcement learning or preference optimization pipelines is a promising direction for future work.

Limitations

First, the approach inherits the sensitivity from the underlying LLM. Criterion-level responses may vary with prompt phrasing, decoding settings, or model updates. Although decomposition reduces some instability by structuring evaluation through fixed criteria, the system remains dependent on the robustness of the base model. Future work could investigate prompt-invariant formulations or uncertainty-aware calibration. Second, our method assumes that evaluation criteria capture stable se-

mantic dimensions shared across languages. This assumption may not hold for culturally specific judgments, stylistic norms, or tasks where evaluation standards differ substantially across regions. Finally, the framework relies on English supervision to learn the transfer module. Although this reduces the need for multilingual labels, it can introduce biases present in English evaluation data that need to be studied. Finally, we do not systematically explore the impact of variance that cascades through each stage of UCS generation.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591.

Carlo Alfano, Aymen Al Marjani, Zeno Jonke, Amin Mantrach, Saab Mansour, and Marcello Federico. 2026. Multilingual self-taught faithfulness evaluators. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5035–5051.

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.

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. pages 22577–22595.

Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. Exploring the multilingual nlg evaluation abilities of llm-based evaluators. *arXiv preprint arXiv:2503.04360*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

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.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. 35:21400–21413.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual llm-as-a-judge? In *EMNLP (Findings)*, pages 11040–11053.

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. 2025. A survey on llm-as-a-judge.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting llm reasoning in non-english languages. *Advances in neural information processing systems*, 37:34161–34187.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. Debate: Devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1885–1897.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge

framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809.

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. 2025. 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*, Suzhou, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th conference of the European chapter of the Association for Computational Linguistics (volume 1: long papers)*, pages 139–151.

Stefano Nembrini, Inke R König, and Marvin N Wright. 2018. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718.

Joshua W Ohde, Lauren M Rost, and Joshua D Overgaard. 2025. The burden of reviewing llm-generated content.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André FT Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *arXiv preprint arXiv:2504.04953*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Ivaxi Sheth and Samira Ebrahimi Kahou. 2023. Auxiliary losses for learning generalizable concept-based models. *Advances in Neural Information Processing Systems*, 36:26966–26990.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. *arXiv preprint arXiv:2407.00908*.

Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Prapti Trivedi, Aditya Gulati, Oliver Molenschot, Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Keith Stevens, Tanveesh Singh Chaudhery, Jahnavi Jambholkar, James Zou, and Nazneen Rajani. 2024. Self-rationalization improves llm as a fine-grained judge. *arXiv preprint arXiv:2410.05495*.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. Rocketeval: Efficient automated llm evaluation via grading checklist. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Murong Yue. 2025. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. 36:46595–46623.

A Risks

Our framework introduces several potential risks. First, reliance on English-labeled data for training the transfer module may propagate biases present in English evaluation standards, potentially leading to unfair or misaligned judgments in other languages. Second, the assumption of language-agnostic evaluation criteria may overlook culturally specific norms, stylistic preferences, or context-dependent interpretations of quality, resulting in systematic evaluation errors. Third, the approach depends on the stability of underlying LLM outputs; variations due to prompt phrasing or model updates may affect criterion-level responses and downstream predictions. Finally, as with other automated evaluators, there is a risk of over-reliance on LLM-based judgments in high-stakes settings without sufficient human oversight.

B Reproducibility

Our code is available here: https://github.com/amazon-science/cross-lingual-judge-transfer.

B.1 Model Inference Settings.

All LLM-based evaluations were performed with the temperature set to 0 and the top-p set to 1 to ensure deterministic outputs. We report the exact model snapshots used in our experiments. All models were accessed through the Amazon Bedrock API. No additional fine-tuning of the backbone LLMs was performed. We reproduced the results for all of the baselines reported in the paper.

B.2 Cross-Lingual Transfer Setup.

For concept-level transfer, we train calibration models on English criterion-level representations and evaluate zero-shot on other languages without using any target-language labels. All experiments use `random_state=42` for reproducibility.

B.3 Transfer Module Hyperparameters.

We evaluated multiple lightweight predictors. We evaluated several lightweight calibration models for concept-level transfer. The neural network consists of a single hidden layer with 32 units, trained with a learning rate of 0.01 for up to 2000 iterations. Logistic regression uses a regularization strength of C = 0.1, is trained for up to 2000 iterations, and applies balanced class weights to account for label imbalance. The random forest model uses 200 trees with a maximum depth of 10 and balanced class weights. The support vector machine uses C = 0.1, a gamma value set to "scale" enables probability estimates, and applies balanced class weights. Gradient boosting is configured with 100 estimators, a maximum depth of 5, and a learning rate of 0.1. Finally, the k-nearest neighbors classifier uses k = 5 neighbors.

Unless otherwise stated, the reported results correspond to the NN model selected on English validation data.

B.4 Data Splits and Evaluation.

We train the transfer module exclusively on labeled English data (226 samples for MEMERAG and 619 for mFACE) and evaluate on multilingual test splits without retraining or hyperparameter tuning. Balanced accuracy is used as the primary evaluation metric to account for label imbalance.

B.5 Implementation Details.

All experiments were implemented in Python using standard machine learning libraries. Fixed random seeds were used across training runs. During transfer, no target-language supervision, translation, or synthetic augmentation was used .

B.6 Annotation Guideline baseline.

A baseline we consider prompts the LLM with the annotation guidelines (AG) originally provided to human annotators, following the setup of Blandón et al. (2025). For MEMERAG, the annotation guidelines are available in Appendix A of the paper and describe the criteria used by annotators to assess faithfulness between the generated answer and

the supporting evidence. For mFACE, we use the evaluation instructions provided in Figure 2 of the paper, which outline the conditions under which a summary should be considered faithful to the source article. In both cases, these guidelines are directly incorporated into the evaluation prompt to guide the LLM's judgment.

C Concept-level transfer

A central design choice in our framework is to perform transfer at the level of evaluation concepts rather than at the level of raw criterion interactions. Each dimension in $z(x)$ corresponds to a semantically meaningful evaluation question (e.g., faithfulness, completeness, consistency). By learning a linear calibration over these concept-level signals, the transfer module estimates how much each evaluation dimension contributes to the final judgment.

An alternative would be to model interactions between criteria using a more expressive predictor. However, such approaches substantially increase the number of learnable parameters and, consequently, the number of labeled samples required for stable training. In multilingual settings where supervision is typically available only in English, this would lead to overfitting and poor generalization.

By constraining transfer to operate over independent semantic dimensions, we reduce sample complexity and improve stability. This design aligns with the intuition that evaluation structure is largely shared across languages, even when surface realizations differ. As a result, concept-level calibration enables effective cross-lingual transfer using limited source-language supervision.

D Results

Table 5 reports the full cross-lingual evaluation results for all baselines and criteria-based methods across models and languages on MEMERAG and mFACE. Consistent with the main results, UCS-based judges generally achieve the strongest or among the strongest performance across languages and model backbones.

Figure 4 shows the cross-lingual alignment of criterion importance between English and target languages for MEMERAG and mFACE. The heatmaps illustrate how the relative importance of evaluation criteria varies across languages.

Table 6 and Table 7 report MEMERAG and mFACE results for Qwen-235B averaged over three runs with different seeds. The results show that UCS maintains consistent improvements across languages while exhibiting relatively low variance compared to most baselines.

Table 9 compares the transfer-module classifier with LLM-based aggregation of the criteria representations across languages. The trained transfer module generally achieves higher performance, although LLM aggregation remains competitive in some cases.

D.1 Supervised Unstructured Representation Baseline

To isolate the effect of the structured criteria representation from that of source-language supervision, we consider an additional supervised baseline based on an unstructured representation. Specifically, we generate chain-of-thought (CoT) rationales using the same judge LLM, encode these rationales into embeddings, and train the same lightweight classifier used by UCS on English-labeled data. Unlike UCS, this baseline does not use any explicit criteria decomposition.

Lang.	BA	κ	N
DE	56.08	0.1173	272
HI	60.38	0.1967	208
ES	63.79	0.2698	276
Avg.	**60.08**	**0.1946**	–

Table 8: Performance of a supervised baseline trained on unstructured CoT embeddings. The same lightweight classifier and English supervision used by UCS are applied, but without explicit criteria decomposition. BA denotes balanced accuracy and κ denotes Cohen's kappa.

The supervised CoT-embedding baseline performs substantially below UCS and the corresponding zero-shot judge. This indicates that the gains from UCS cannot be attributed solely to the use of English supervision and a lightweight classifier. Instead, the structured criteria representation provides a more transferable intermediate signal for learning the final judgment across languages. In contrast, the unstructured CoT representation appears less suitable for cross-lingual transfer.

E Prompts

Model	Method	MEMERAG (BA)				mFACE (BA)							
		DE	FR	ES	HI	AM	MY	FR	SW	TH	AR	HI	ES
OSS-20B	Zero-shot	70.7	81.6	80.4	81.7	59.4	65.6	73.7	79.2	76.8	69.2	68.5	78.0
	CoT	70.5	81.6	82.0	80.6	60.9	64.9	74.6	76.2	78.3	65.2	67.6	78.0
	AG	71.7	80.8	75.6	80.9	61.4	62.2	75.2	77.8	76.9	72.4	66.7	78.0
	ChatEval	64.2	73.5	71.4	72.0	54.7	60.1	66.9	70.3	68.4	62.9	61.4	71.1
	CheckEval	72.5	81.9	79.9	80.8	60.9	68.3	77.9	76.7	79.7	68.0	65.7	77.0
	RocketEval	67.9	76.9	74.8	75.6	57.9	63.4	70.8	73.5	71.6	66.5	64.9	74.2
	UCS (EN)	74.4	84.2	81.2	85.0	61.2	68.1	79.0	82.7	82.4	73.7	67.8	82.0
Qwen 3 32B	Zero-shot	74.9	73.7	71.4	78.5	56.6	63.5	71.9	72.3	69.2	65.0	67.9	76.0
	CoT	71.7	73.8	72.6	77.2	55.2	62.9	72.0	72.4	70.0	66.7	68.1	76.5
	AG	73.7	73.9	74.8	81.8	58.3	63.9	76.6	73.0	70.9	68.4	66.1	80.0
	ChatEval	66.7	68.5	63.4	66.9	50.9	56.3	61.2	65.1	61.7	59.4	56.9	66.0
	CheckEval	71.1	74.6	72.3	80.9	59.2	63.5	74.5	73.6	71.1	65.8	65.0	79.0
	RocketEval	70.0	72.9	68.7	72.4	54.8	59.7	65.8	69.0	66.9	63.9	61.4	70.4
	UCS (EN)	74.4	82.7	79.5	85.0	57.7	61.7	68.4	73.4	72.1	68.9	66.7	78.9
Llama70b	Zero-shot	69.9	70.3	68.0	73.8	55.6	57.9	65.3	73.2	60.4	63.4	62.8	75.8
	CoT	67.1	72.2	68.4	69.7	56.6	58.2	64.8	75.3	60.7	64.7	62.3	75.8
	AG	75.4	76.4	74.0	80.8	58.2	59.0	68.7	77.6	63.2	66.1	67.3	77.5
	ChatEval	68.4	71.3	66.1	69.8	53.9	57.4	63.9	68.9	64.8	62.1	59.8	69.6
	CheckEval	72.8	76.6	72.1	80.2	59.2	59.2	69.5	77.1	65.4	63.8	67.8	73.9
	RocketEval	71.8	74.9	70.5	73.9	57.1	61.0	67.9	72.3	69.5	66.3	64.0	73.1
	UCS (EN)	77.3	80.8	82.1	84.5	63.5	56.5	72.1	74.0	73.2	67.4	66.1	77.7
OSS-120B	Zero-shot	79.2	83.6	79.3	81.5	64.1	67.6	82.0	78.6	74.9	74.2	69.4	85.1
	CoT	78.3	82.9	78.8	80.9	64.7	67.0	81.4	79.4	74.1	73.9	69.9	83.8
	AG	79.4	84.1	79.0	82.1	65.2	66.7	82.6	79.7	75.0	74.6	69.7	85.3
	ChatEval	73.1	74.2	64.9	66.1	58.3	65.4	69.4	74.0	65.7	66.8	58.0	71.4
	CheckEval	78.6	83.8	78.5	81.9	65.7	67.9	82.1	79.1	75.2	75.0	69.1	84.7
	RocketEval	78.9	84.1	78.9	82.3	66.0	68.2	82.4	79.3	75.4	75.2	69.4	85.0
	UCS (EN)	79.4	83.9	79.7	81.9	64.4	67.8	82.4	79.0	75.1	74.6	69.7	85.6

Table 5: Balanced Accuracy (BA) of all baselines and criteria set judges across languages on MEMERAG and mFACE.

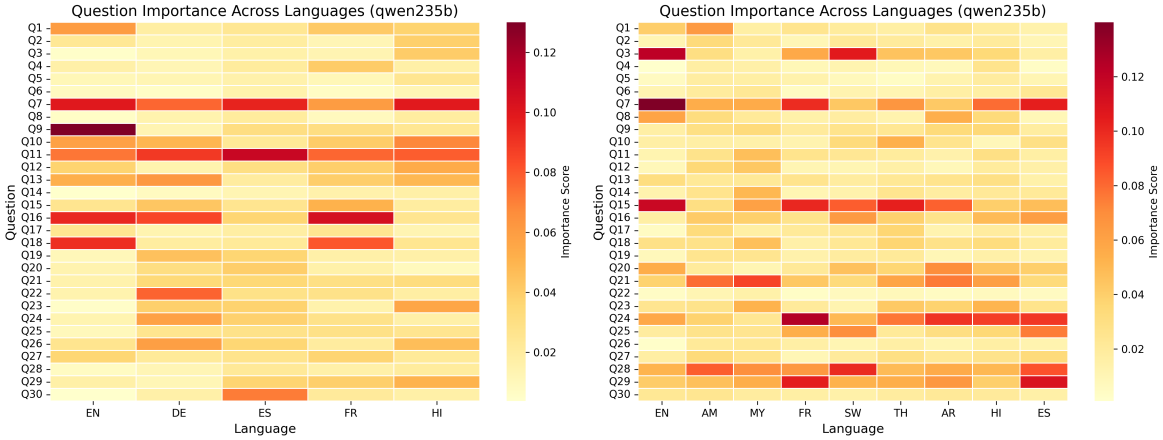


Figure 4: Relationship between English–target criteria importance alignment.

Method	DE	FR	ES	HI
Zero-shot	77.2±0.4	77.5±1.3	77.5±0.9	74.4±1.1
CoT	73.0±1.2	76.2±0.5	76.2±1.4	74.1±0.8
AG	77.2±0.7	80.6±1.1	76.1±0.4	78.0±1.4
ChatEval	71.8±1.5	73.0±0.6	60.9±1.9	62.4±0.7
CheckEval	76.1±0.5	80.7±1.4	74.8±0.7	78.4±1.1
RocketEval	76.4±1.9	81.0±1.2	75.9±1.7	79.1±1.4
UCS (EN)	**79.7**±0.6	**84.3**±1.1	**80.0**±0.4	**82.2**±1.2

Table 6: Balanced Accuracy (BA) on MEMERAG for Qwen-235B, reported as mean ± standard deviation over 3 runs.

Method	AM	MY	FR	SW	TH	AR	HI	ES
Zero-shot	54.7±0.7	**70.2**±1.3	74.3±0.6	73.0±1.1	72.6±0.4	69.7±1.4	68.0±1.0	77.9±0.6
CoT	58.7±1.4	68.5±0.6	72.8±1.2	77.5±0.5	71.4±1.1	69.2±0.5	69.4±1.6	74.4±0.8
AG	59.3±0.8	67.4±1.5	76.6±0.7	77.7±1.1	72.6±0.6	70.7±1.2	69.0±0.7	79.5±1.2
ChatEval	51.1±1.6	65.5±0.9	68.4±1.3	74.0±0.5	63.9±1.8	66.1±0.6	55.7±1.1	70.3±1.5
CheckEval	60.8±1.3	68.9±0.4	74.9±1.6	76.8±0.7	**75.1**±1.1	73.4±0.8	65.8±1.5	76.4±0.6
RocketEval	61.2±2.1	69.2±1.1	75.4±1.7	77.0±1.3	74.9±1.6	72.9±2.0	66.9±1.4	78.1±1.8
UCS (EN)	**64.7**±1.0	67.7±1.4	**82.4**±0.6	**79.0**±1.3	**75.1**±0.4	**74.6**±1.2	**70.0**±0.7	**85.5**±1.0

Table 7: Balanced Accuracy (BA) on mFACE for Qwen-235B, reported as mean ± standard deviation over 3 runs.

Summary Evaluation Questions

Factual Accuracy (6 questions)

1. Does the summary include any claims that are not explicitly supported by information in the source document?
2. Does the summary contradict any specific fact, detail, or relationship stated in the source document?
3. Are events, actions, or attributes in the summary correctly attributed to the individuals, entities, or sources named in the source document?
4. Does the summary present speculative, uncertain, or conditional information from the source as definitive or certain?
5. Does the summary accurately reflect the relative importance or prominence of key points as presented in the source document?
6. Does the summary introduce numerical data, statistics, or quantitative claims that differ from those in the source document?

Contradiction Detection (6 questions)

7. Does the summary attribute a claim, opinion, or action to a source or entity that is not supported or explicitly stated in the source document?
8. Does the summary present a possibility or uncertainty as a definitive fact, thereby increasing the level of certainty beyond what is expressed in the source document?
9. Does the summary state that an event occurred, or a condition exists, when the source document explicitly indicates it did not happen or was not the case?
10. Does the summary include a causal relationship between two events that is not stated, implied, or supported by the source document?
11. Does the summary report a numerical value, statistic, or quantitative detail that contradicts the figures provided in the source document?
12. Does the summary describe an action or outcome as having happened in the past when the source document states it is planned, proposed, or speculative?

Claim Support (6 questions)

13. Does the summary include any claims that are not explicitly supported by evidence or statements in the source document?
14. Are there any statements in the summary that directly contradict information provided in the source document?
15. Does the summary attribute actions, opinions, or statements to individuals or entities that are not assigned to them in the source document?
16. To what extent does the summary present speculative or conditional information from the source as definitive or certain?
17. Does the summary introduce new causal relationships or implications between events that are not stated or logically supported in the source document?
18. Are key details in the summary distorted in a way that alters the meaning or significance of the information presented in the source document?

Misrepresentation Identification (6 questions)

19. Does the summary attribute a claim, opinion, or action to a source or entity that is not supported or explicitly stated in the source document?
20. Does the summary present a speculative or conditional statement from the source as a definitive fact?
21. Does the summary include a key event, outcome, or statistic that is not mentioned or implied in the source document?
22. Does the summary reverse, invert, or otherwise distort the causal or temporal relationship between two events described in the source document?
23. Does the summary exaggerate the strength, scope, or certainty of a finding, trend, or conclusion beyond what is stated in the source document?
24. Does the summary omit a critical qualifying condition, limitation, or exception present in the source that changes the interpretation of the information?

Certainty Calibration (6 questions)

25. Does the summary present a claim as certain or definitive when the source document expresses it as uncertain, tentative, or conditional?
26. Does the summary introduce a level of confidence or precision (e.g., “proves,” “definitely,” “exactly”) in a claim that is not supported by the degree of certainty used in the source document?
27. Are probabilities, frequencies, or likelihoods in the summary accurately reflected in terms of their magnitude and wording compared to the source document (e.g., “likely” vs. “possible” vs. “certain”)?
28. Does the summary omit hedging language (e.g., “may,” “suggests,” “appears to”) present in the source, thereby overstating the strength of a conclusion?
29. To what extent does the summary mirror the source document’s attribution of claims to specific agents, studies, or evidence, without shifting responsibility or generalizing to broader consensus?
30. Does the summary elevate a hypothesis, preliminary finding, or speculative idea from the source to the status of an established fact?

Evidence Support Evaluation Questions

Factual Accuracy (6 questions)

1. To what extent does the evidence directly support the key claims in the answer, with no unsupported assertions?
2. How well do the retrieved passages contradict or fail to support any statements in the answer?
3. To what degree does the answer avoid overgeneralizing or making broad inferences beyond what is stated in the evidence?
4. How precisely does the answer align with the specificity (e.g., timeframes, quantities, conditions) provided in the evidence?
5. To what extent are all named entities, events, and relationships in the answer accurately reflected in the evidence passages?
6. How well does the answer refrain from introducing plausible but unverified details not present in the evidence?

Completeness of Support (6 questions)

7. To what extent does the evidence directly support all key claims in the answer, with no unsupported assertions?
8. How well do the retrieved passages contain specific details or examples that match the level of specificity in the answer?
9. To what degree does the evidence fully cover the scope of the answer, including all sub-claims or components mentioned?
10. How consistently do the passages align with the answer without introducing contradictions or conflicting information?
11. To what extent does the answer avoid overgeneralizing beyond what is reasonably supported by the evidence?
12. How well do the passages provide sufficient context or explanation to justify causal or inferential claims made in the answer?

Specificity Alignment (6 questions)

13. To what extent does the answer reflect the same level of specificity as the evidence, avoiding unwarranted generalizations or oversimplifications?
14. How well do the key claims in the answer map directly to specific details or data points in the evidence, rather than relying on vague or peripheral information?
15. To what degree does the evidence support the precise scope (e.g., time frame, population, location) asserted in the answer without overextension?
16. How closely does the answer avoid introducing concepts or conclusions that are more specific than what is warranted by the evidence?
17. To what extent are named entities, quantities, or relationships in the answer explicitly grounded in corresponding specific mentions within the evidence?
18. How well does the answer maintain alignment with the evidence by neither omitting critical qualifying conditions nor adding unsupported qualifiers?

Consistency with Evidence (6 questions)

19. To what extent does the evidence directly support the key claims in the answer, with no unsupported assertions?
20. How well do the retrieved passages contradict or conflict with any statements in the answer?
21. To what degree does the answer avoid overgeneralizing beyond the scope or specificity of the evidence provided?
22. How closely does the answer align with the factual details and context present in the evidence, avoiding subtle distortions or misrepresentations?
23. To what extent can each component of a multi-part answer be individually justified by at least one evidence passage?
24. How well does the answer reflect the certainty level (e.g., tentative, definitive) expressed in the evidence, without introducing unwarranted confidence or ambiguity?

Source Attribution (6 questions)

25. To what extent does the evidence explicitly attribute the claim to a credible source or original provider of information?
26. How well do the retrieved passages support the specificity of the claim, without introducing unsupported details or omitting critical qualifiers present in the source?
27. To what degree is the claim directly verifiable from the cited evidence, rather than requiring inference beyond what the source states?
28. How consistently does the answer reflect the source's intended meaning, avoiding misrepresentation or overgeneralization of the evidence?
29. To what extent does the evidence rule out contradictions or significant discrepancies with the claim being made?
30. How clearly is the connection between the evidence and the claim articulated, such that the support is traceable and transparent?

Concept Generation - MEMERAG

```
"system_prompt": "You are an evidence-grounding evaluator. Your goal is to identify key verification concepts for checking whether answers are supported by evidence."
"task_prompt": "Generate 3-5 distinct verification concepts that are essential for evaluating whether an answer is supported by evidence passages.\n\nGuidelines:\n- Concepts should cover different aspects of evidence grounding (e.g., factual accuracy, completeness, specificity, consistency, source attribution).\n- Each concept should be general enough to apply across different topics and domains.\n- Concepts should target different failure modes where answers might not be properly supported.\n- Keep concepts concise (2-5 words each).\n\nReturn exactly this format:\n\n<concepts>\n<concept1>[Concept name]</concept1>\n<concept2>[Concept name]</concept2>\n...\n</concepts>"
```

Question Generation - MEMERAG

"system_prompt": "You are an evidence-grounding evaluator. Your goal is to generate reusable verification questions for checking whether an answer is supported by retrieved evidence passages."

"task_prompt": "**Verification Concept:** conceptGenerate 4-6 clear, diverse, and reusable evaluation questions for verifying whether an answer is supported by evidence with respect to the concept 'concept'. These questions will later be scored on a **0-10 scale indicating degree of evidence support**.Guidelines:- Focus on whether an **answer is supported by evidence passages**.- Questions must be reusable across different topics and documents.- Cover different **evidence-grounding failure modes** (e.g., unsupported claims, contradictions, overgeneralization, specificity mismatch).- Each question should be written to support **graded (partial-to-full) scoring**, not just yes/no judgments.- Each question should target a distinct verification angle.- Wording should be concise and evaluative (e.g., "To what extent is the claim supported by the evidence...", "How well do the passages justify...").Return exactly this format:<questions><question1>Question: [Universal evaluation question]</question1>...</questions>"

Question Application - MEMERAG

"system_prompt": "You are an evidence-grounding evaluator. Your task is to determine whether an answer is supported by the provided evidence passages."

"task_prompt": "**Evidence Passages:**\n {context}\n\n **Answer:** {answer_segment}\n **Verification Concept:** {concept}\n\n **Evaluation Questions:**\n {questions}\n\n For each question, assign a score from **0-10** indicating how well the evidence supports the answer with respect to that question.\n\n Base all judgments strictly on the provided evidence passages. Do not assume any external knowledge.\n\n Return exactly this format:\n\n <evaluation>\n {questions_format_example}\n </evaluation>"

Concept Generation - mFACE

concept_generation

"system_prompt": "You are an expert factual faithfulness evaluator. Your goal is to identify key faithfulness concepts for checking whether summaries are factually faithful to their source documents."

"task_prompt": "Generate 3-5 distinct faithfulness concepts that are essential for evaluating whether a summary is factually faithful to its source document.\n\nGuidelines:\n- Concepts should cover different aspects of factual faithfulness (e.g., factual accuracy, contradiction detection, unsupported claims, misrepresentation, distorted certainty).\n- Each concept should be general enough to apply across different topics and domains.\n- Concepts should target different failure modes where summaries might not be properly supported by source documents.\n- Keep concepts concise (2-5 words each).\n\nReturn exactly this format:\n\n<concepts>\n<concept1>[Concept name]</concept1>\n<concept2>[Concept name]</concept2>\n...\n</concepts>"

Question Generation - mFACE

question_generation

"system_prompt": "You are an expert factual faithfulness evaluator. Your role is to construct high-quality, reusable evaluation questions that test whether summaries are factually faithful to their source documents."

"task_prompt": "**Faithfulness Concept:** {concept}\n\nGenerate exactly 6 clear, diverse, and reusable evaluation questions for the faithfulness concept '{concept}'. Each question must require direct comparison between a summary and its source document to assess factual support, contradiction, or distortion under this concept.\n\nAnnotation Guidance to Apply While Writing Questions:\n- The question must be answerable ONLY by checking the source document.\n- The question should detect one of the following: unsupported claims, contradiction, misrepresentation, distorted certainty, or incorrect attribution.\n- The question must not depend on surface form (grammar, fluency, style, or verbosity).\n\nGuidelines:\n- Every question must explicitly or implicitly require verification against the source document.\n- Do NOT ask about readability, grammar, fluency, writing quality, or length.\n- Questions must be context-independent and reusable across domains.\n- All questions must stay strictly within the given faithfulness concept, but vary the factual scenario or common error pattern being tested.\n- Avoid redundancy: each question should target a distinct factual failure mode.\n- Wording must be concise, unambiguous, and evaluative (e.g., \"Does the summary. . . \" or \"To what extent does the summary. . . \").\n\nReturn exactly this format:\n\n<questions>\n<question1>\nQuestion: [Universal evaluation question]\n</question1>\n\n<question2>\nQuestion: [Universal evaluation question]\n</question2>\n\n<question3>\nQuestion: [Universal evaluation question]\n</question3>\n\n<question4>\nQuestion: [Universal evaluation question]\n</question4>\n\n<question5>\nQuestion: [Universal evaluation question]\n</question5>\n\n<question6>\nQuestion: [Universal evaluation question]\n</question6>\n</questions>"

Question Application - mFACE

question_application

"system_prompt": "You are an expert factual faithfulness evaluator. Your task is to score a summary against its source document using predefined faithfulness evaluation questions."

"task_prompt": "**Source Document:**\n{context}\n\n**Summary:** {answer_segment}\n\n**Faithfulness Concept:** {concept}\n\n**Evaluation Questions:**\n{questions}\n\nFor each question, evaluate whether the summary is factually faithful to the source document under that question.\n\nWhile scoring, explicitly consider whether the summary:\n- Contains information not stated in or directly inferable from the source\n- Contradicts the source\n- Introduces unsupported details\n- Misrepresents entities, quantities, relationships, or certainty\n- Overgeneralizes, narrows, or distorts scope\n\nAssign a score from 0-10 for each question:\n- 0 = Completely unfaithful (clear contradiction, fabrication, or unsupported claim)\n- 10 = Completely faithful (fully supported, correctly represented, no distortion)\n\nRules:\n- Base every score strictly on evidence from the source document.\n- Ignore grammar, fluency, style, and summary length.\n- Penalize both direct hallucinations and subtle distortions of meaning or certainty.\n\nProvide a brief, evidence-based justification for each score.\n\nReturn exactly this format:\n\n<evaluation>\n{questions_format_example}\n</evaluation>"

Dataset	Lang.	UCS (transfer)	UCS (LLM)
MEMERAG	DE	**79.7**	76.9
	FR	**84.3**	78.9
	ES	**80.0**	77.2
	HI	**82.2**	75.9
mFACE	AM	**64.7**	62.5
	MY	67.7	**69.6**
	FR	**82.4**	82.3
	SW	**79.0**	78.4
	TH	**75.1**	74.9
	AR	**74.6**	73.1
	HI	**70.0**	68.8
	ES	**85.5**	84.5

Table 9: Comparison of threshold-based training classifier and LLM-based aggregation of intermediate criteria representations for final judgment.