Language-Dependent Miscalibration in Multilingual LLM Evaluators

Ej Zhou Lucas Resck Zheng Hui Anna Korhonen

Language Technology Lab, University of Cambridge

Motivation

- LLMs are increasingly evaluated by other LLMs (e.g., **LLM-as-a-Judge** and **Reward Models**).
- **Implicit Assumption:** Evaluation scores are *language-invariant*. Semantically identical content should receive comparable scores regardless of evaluation language.

★ *Does this assumption actually hold, or are standard metrics hiding critical, language-dependent failure modes?*

Contributions

- We demonstrate evaluators exhibit large, systematic **language bias in pointwise scoring**.
- We show standard **pairwise accuracy is structurally blind** to these absolute score shifts.
- We highlight that fixed-threshold filtering yields **drastically different acceptance rates** across languages.

Summary & Implications

- **Core Conclusion:** Evaluators encode systematic language-conditioned scoring behavior entirely independent of actual content quality.
- **Downstream Impact:** Reward shaping spuriously penalizes languages; uniform thresholds yield highly unfair safety and quality decisions.
- **Future Directions:** Evaluation protocols must move beyond pairwise metrics to explicitly audit pointwise miscalibration.

Language-Dependent Scoring Bias

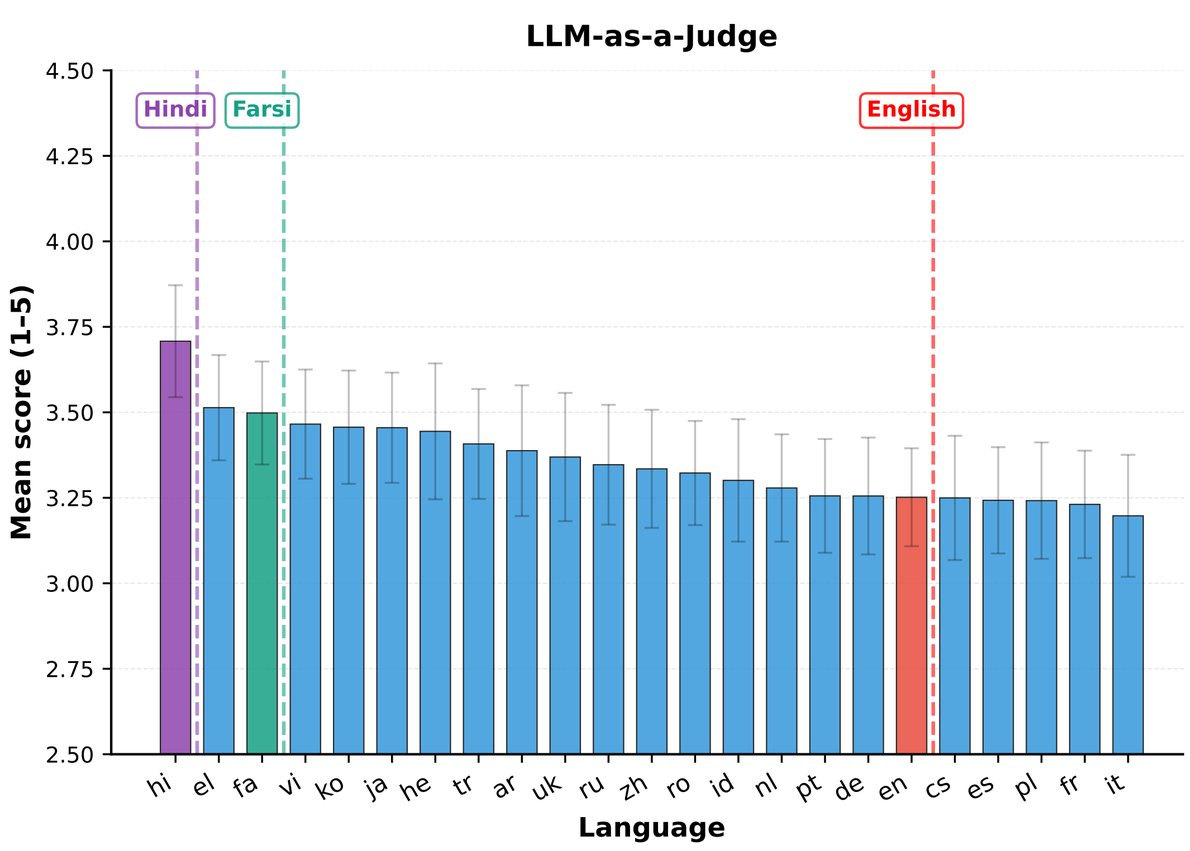

Figure 1. Mean pointwise scores from LLM-as-a-Judge models. Identical responses receive widely different absolute scores based solely on the target language.

Key Findings

- **Absolute Score Shifts:** Evaluators assign scores with rigid shifts of up to **0.4–0.5 points** based on language alone.
- **Stable Hierarchy:** Hindi, Greek, and Hebrew consistently score highest; Western European languages occupy the lower end.
- **Language Resource Level:** High-resource languages are systematically assigned lower scores, while lower-resource languages, more favorably.

Cross-Model Consistency

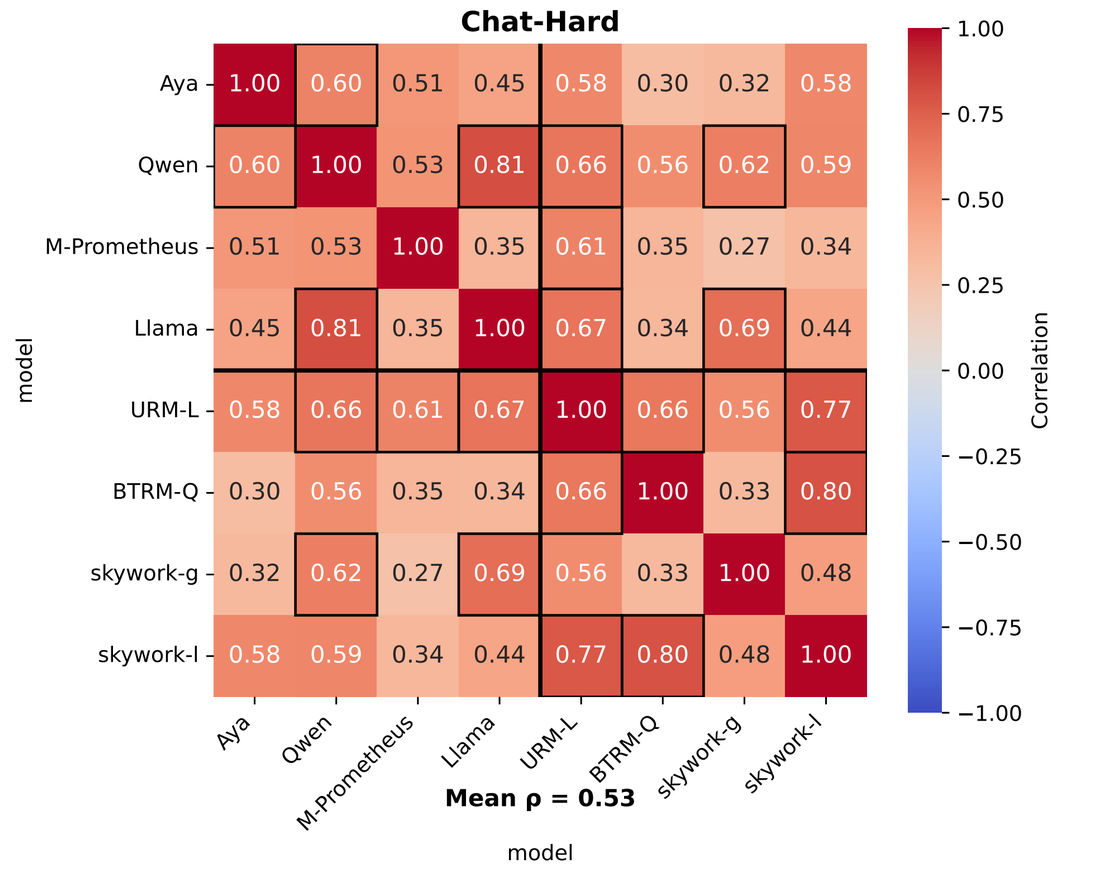

Figure 2. Pearson correlation between per-language mean scores assigned by two evaluators (*Chat-Hard* of M-RewardBench).

Key Findings

- **Strong Agreement:** Cross-model correlations are remarkably high (e.g., mean $\rho = 0.53$ for *Chat-Hard*), showing models largely agree on language rankings.
- **Shared Inductive Bias:** This language-conditioned hierarchy persists despite vast differences in model architecture, training objectives, and score scale.

Pairwise Accuracy Masks Decision Bias

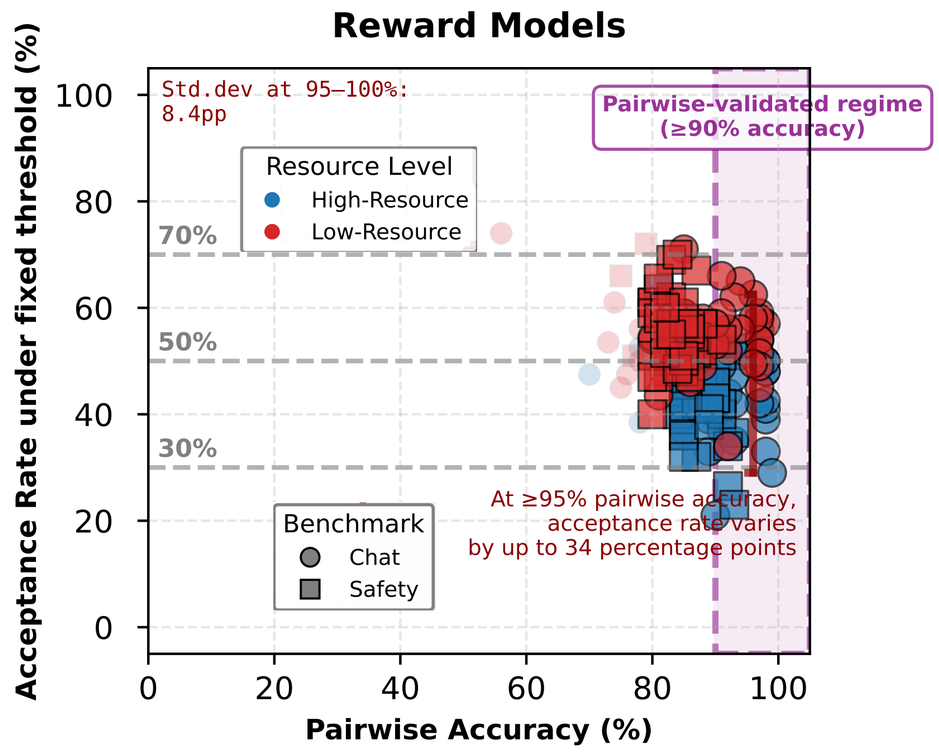

Figure 3. Per-language acceptance rate under a fixed threshold vs. pairwise accuracy for Reward Models.

Key Findings

- **The Pairwise Illusion:** High pairwise accuracy ($> 90\%$) falsely implies language-agnosticism. The metric captures relative order but ignores absolute score distributions.
- **Acceptance Disparities:** Under uniform thresholds, acceptance rates vary by up to **34 percentage points** for models with identical pairwise accuracy.
- **Structural Blindness:** Pairwise validation is fundamentally blind to these severe downstream disparities, leaving biases undetected.

yz926@cam.ac.uk