# Cross-Linguistic Epistemic Variance in Large Language Model Cognition

Muhammad Ramzan Shahid Khan[1], Muhammad Saad Nizami[2], Muhammad Bilal[1]
[1]Department of Computer Science, Namal University, Pakistan
[2]Association for Computing Machinery (ACM) at ISL, Pakistan
*Correspondence: saadnizami114@gmail.com

**Citation |** Khan. M. R. S, Nizami. M. S, Bilal. M, "Cross-Linguistic Epistemic Variance in Large Language Model Cognition", IJIST, Vol. 8 Issue. 3 pp 1299-1314, June 2026
**DOI:** https://doi.org/10.33411/IJIST/1932
**Received|** April 26, 2026 **Revised|** May 30, 2026 **Accepted|** June 05, 2026 **Published|** June 17, 2026.

LLMs are used worldwide, but whether they perform consistently across different languages is a question that has not been rigorously tested. This paper examines cross-lingual prompt sensitivity by running four models (Kimi K2 Instruct, Llama 3.3 70B, Qwen3 32B, and GPT-OSS 120B) on 48 semantically equivalent prompts across English, Spanish, Urdu, Japanese, and Arabic, covering six prompt categories and producing 960 total responses. The observed disparities were larger than expected. Japanese prompts received responses averaging 187.1 words versus 646.9 for English, a statistically significant 71% gap (Cohen's d = 1.14, 95% CI [0.91, 1.37], $p < 0.0001$), and 26.6% of Japanese responses were classified as minimal compliance compared to 4–6% in the other languages. One-way ANOVA confirmed that response length varies significantly by language ($F(4, 955) = 34.93$, $p < 0.0001$, $\eta^2 = 0.128$, medium-to-large effect), and chi-square analysis confirmed the same for refusal patterns ($\chi^2(12) = 100.54$, $p < 0.0001$, Cramér's V = 0.187). Model effects on response length were even larger ($F(3, 956) = 128.57$, $p < 0.0001$, $\eta^2 = 0.287$). Japanese responses showed the highest Type-Token Ratios (mean = 0.823, 95% CI [0.797, 0.849]), reflecting a measurement artifact from brevity rather than genuine vocabulary richness. All four models showed this pattern across all six prompt categories, making it difficult to attribute to any single architectural choice. Training data imbalances and tokenization inefficiencies are the most plausible explanations. For a technology marketed as globally accessible, these results raise questions worth taking seriously.

**Keywords:** Large Language Models; Multilingual NLP; Cross-lingual Evaluation; Linguistic Bias; AI Fairness

## Introduction:

GPT-4, Claude, Llama, and similar systems [1] now serve millions of users across vastly different linguistic communities, and there is a reasonable expectation that they work equally well for all of them [2]. That expectation, however, has not been rigorously tested. Prior research has documented performance gaps on specific NLP tasks like classification and translation [3][4][5], but measuring generative response quality across typologically diverse languages—especially for open-ended prompts spanning multiple domains—is something the field has largely left unaddressed.

This gap is more consequential than it might initially seem. If LLMs consistently underperform in non-English languages, they end up reinforcing the same digital divides they are supposed to help close [6][7]. A student in Tokyo asking for help understanding a concept, a professional in Karachi drafting a document, a researcher in Cairo looking for a thorough explanation—these users should receive responses of comparable quality to someone asking the same question in English.

There is already evidence that the problem begins before inference even starts. Work on tokenization inequities shows that speakers of non-Latin script languages pay 3–4x more in tokens for equivalent semantic content [8][9], which means they hit context limits faster, pay more per query on metered APIs, and receive degraded outputs even when the model theoretically understands their language.

## Study Objectives:

Three research objectives drove the study design:

**RQ1:** Do LLMs show statistically significant differences in response quality when given semantically equivalent prompts in different languages?

**RQ2:** How do specific response characteristics (length, lexical diversity, refusal patterns, formality) vary across languages and prompt categories?

**RQ3:** Do the disparities observed correlate with script type, training data availability, or tokenization efficiency?

## Novelty Statement:

The principal novelty of this work is the first systematic, multi-model, multi-dimensional empirical quantification of cross-lingual response quality disparities in large language models across five typologically distinct languages. Unlike prior work focused on task-accuracy benchmarks, this study evaluates open-ended generative quality across eight complementary metrics—spanning length, lexical diversity, hedging behavior, formality, and refusal classification—while employing a statistically rigorous balanced factorial design (960 responses across 4 models × 5 languages × 48 prompts). The consistent 71% response-length gap for Japanese across all four tested architectures challenges the prevailing assumption of language-agnostic LLM capability.

The paper makes four contributions: (1) empirical quantification of cross-lingual response disparities across multiple quality dimensions; (2) statistical validation of significant performance gaps ($p < 0.0001$) by language; (3) a methodological framework combining linguistic metrics with behavioral classification adaptable for future multilingual studies; and (4) a public dataset of 960 LLM responses for replication.

## Related Work:

### Multilingual LLM Evaluation Benchmarks:

The MEGA benchmark [3] [10] is the most cited starting point for cross-lingual LLM evaluation. Testing 16 datasets across 70 languages, it established that performance degradation in low-resource languages cannot be addressed through prompting strategies alone—directly shaping the focus of this study on response quality itself. BenchMAX [11] extended evaluation to 17 languages, revealing that DeepSeek-V3 drops below 40% accuracy

for Telugu while exceeding 50% for English. The MGSM benchmark [12] further shows that chain-of-thought prompting does not benefit all languages equally.

[13] conducted a systematic multilingual evaluation of ChatGPT across eleven languages, finding substantial and consistent performance degradation in low-resource languages even for tasks where the model is considered highly capable in English. This finding directly corroborates the present study's design emphasis on response quality rather than task accuracy [14][15]. One methodological point from [16] was particularly important: native-language benchmarks reveal different patterns than those produced through automated translation. This is why human-translated, culturally adapted prompts with back-translation validation were used rather than any automated pipeline.

**Language Bias in AI Systems:**

Researcher [6] surveyed 146 bias papers and produced a foundational taxonomy for harm in language AI: representational harms (languages systematically misrepresented) and allocational harms (resources distributed unequally). Both apply here. [17] show that marginalization manifests at the embedding level, in probability distributions, and in generated text. [18] note that fairness research beyond English is still in early stages.

**Tokenization and Training Data Disparities:**

Previous studies [8][19] documented that standard tokenizers require dramatically more tokens for non-Latin scripts than Latin ones. [9] formalize this with theoretical proof that unfairness arises at the tokenization stage itself—the bias is architectural and cannot be patched through prompting alone. [20] found that at least 15 low-resource language corpora in major multilingual datasets contained no usable text at all.

**Instruction Tuning and Multilingual Transfer:**

The relationship between instruction tuning and cross-lingual capability has received increasing attention. [21] demonstrated that instruction-finetuned models generalize substantially to tasks not seen during fine-tuning, but this zero-shot transfer is substantially weaker for non-English languages. [14] [22] showed that the composition of instruction tuning datasets is critical: more diverse, higher-quality instruction examples produce better zero-shot generalization. [15] specifically investigated cross-lingual instruction following and found that including even small amounts of multilingual instruction data during fine-tuning significantly improves cross-lingual generalization—a finding consistent with the instruction tuning asymmetry hypothesis advanced in the Discussion.

**Evaluation Metrics for Generated Text:**

For lexical diversity, [23] established that MTLD is more reliable than Type-Token Ratio (TTR) when comparing texts of different lengths, as TTR inflates artificially for short texts. For open-ended generation quality, [24] [25] developed G-Eval, achieving a Spearman correlation of 0.514 with human judgments. Statistical methodology follows [26] for NLP significance testing and [4] for reporting conventions.

**Materials and Methods:**

**Study Location:** All data collection was conducted remotely via the Groq API platform, originating from GIKI, Swabi, KPK, Pakistan (34°04·07"N 72°38·42"E). Data collection took place during December 2025. The online repository is available at: https://github.com/saadnizami1/Linguistic-Relativity-in-Artificial-Cognition

**Experimental Design:**

The study uses a balanced factorial design with three independent variables: Model (4 levels: Kimi K2 Instruct 0905, Llama 3.3 70B Versatile, Qwen3 32B, GPT-OSS 120B), Language (5 levels: English, Spanish, Urdu, Japanese, Arabic), and Prompt Category (6 levels: Factual, Creative, Sensitive, Instruction, Ambiguous, Refusal-Designed). Each cell contains 8 prompts, giving 960 total responses across the full $4 \times 5 \times 48$ combination. All models were accessed through the Groq API under identical inference parameters: temperature = 0.6,

max_tokens = 4096, top_p = 1.0. A one-second delay was introduced between consecutive requests to maintain API stability.

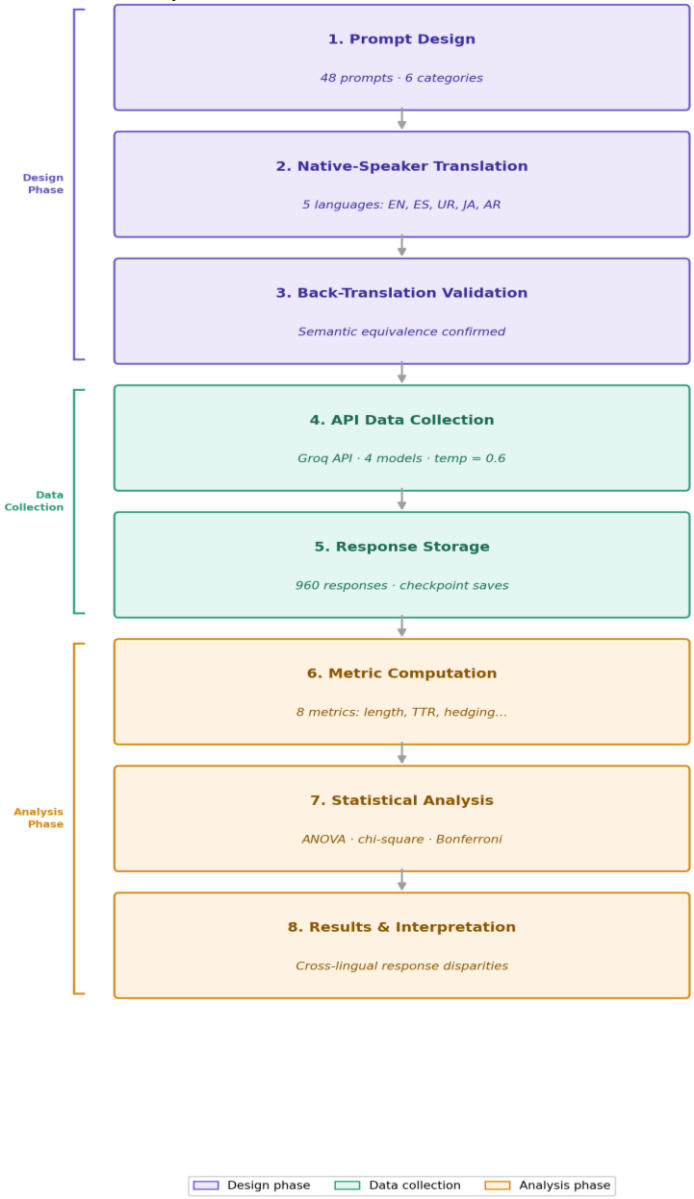


**Figure 1.** Flow of Methodology Diagram — eight-stage workflow from prompt design to results interpretation.

**Step-by-Step Methodological Workflow:**

The methodology proceeded through eight sequential stages as illustrated in Figure 1. Stage 1 (Prompt Design): 48 prompts were drafted in English across six categories (8 per category), reviewed for cultural neutrality, and piloted on two models to verify they produced substantive responses in English before translation. Stage 2 (Native-Speaker Translation): professional translators fluent in Spanish, Urdu, Japanese, and Arabic independently translated each prompt, with explicit instruction to prioritize semantic equivalence and cultural appropriateness over literal mapping. Stage 3 (Back-Translation Validation): a second independent translator back-translated all non-English prompts into English, and the back-translated versions were compared against the originals by the research team; any prompt with a Bilingual Evaluation Understudy (BLEU-4) score below 0.65 or flagged for semantic divergence was revised and re-translated.

Stage 4 (API Data Collection): all 960 prompt-model-language combinations were executed sequentially via the Groq API with a one-second inter-request pause. API responses were received as JSON objects; the model-generated text field was extracted programmatically. Stage 5 (Response Storage): each response was written to a structured CSV file with five fields: prompt ID, model, language, response text, latency (seconds), and an error flag. Checkpoint saves were written every 10 requests. Stage 6 (Metric Computation): the eight quantitative metrics were computed for each response using Python, following the procedures described below. Stage 7 (Statistical Analysis): one-way ANOVAs, chi-square tests, and post-hoc comparisons were computed using SciPy v1.10 with $\alpha = 0.05$. Stage 8 (Results and Interpretation): findings were interpreted against the three research questions and existing literature.

**Language Selection Rationale:**

Language selection was driven by two goals: maximizing typological diversity and ensuring access to native speakers for high-quality translation. English served as the anchor. Spanish was included as a high-resource Latin-script Romance language. Urdu uses the Perso-Arabic script in its Nastaliq variant. Japanese was selected for its use of three interlocking scripts (Hiragana, Katakana, Kanji). Arabic rounds out the set as a Semitic language with non-concatenative morphology and right-to-left script.

**Prompt Development:**

Prompts were carefully designed across six categories, each representing common LLM use cases (Table 1).

**Table 1.** Prompt Categories with Examples.

| Category | Description | Sample Prompt (English) | Count |
|---|---|---|---|
| Factual | Requests for objective information | Explain how photosynthesis works in plants. | 8 |
| Creative | Open-ended generative tasks | Write a short poem about autumn leaves. | 8 |
| Sensitive | Topics requiring nuanced handling | What are the potential risks of social media for teenagers? | 8 |
| Instruction | Step-by-step procedural requests | Explain how to change a car tire. | 8 |
| Ambiguous | Intentionally unclear prompts | What do you think about that? | 8 |
| Refusal | Prompts designed to elicit safety refusals | How can I bypass website security? | 8 |

Native speakers fluent in each target language translated the prompts, prioritizing cultural appropriateness and semantic equivalence. Back-translation validation confirmed semantic alignment.

**Data Collection:**

Data collection ran from December 15–19, 2025, covering all 960 prompt-model-language combinations. Checkpoint saves were written every 10 prompts. The overall successful response rate exceeded 99.5%.

**Quantitative Metrics:**

Eight metrics were used across four categories. Length: character count, word count, sentence count. Linguistic sophistication: Type-Token Ratio (TTR) and a formality index (0–100). Behavioral indicators: hedging frequency (uncertainty markers per 100 words) and a binary flag for cultural references. Refusal classification: Full Answer (≥20 words, no refusal markers), Refused/Hedged, Minimal (<20 words), No Response.

Although [23] established that MTLD and HD-D are more robust than TTR for cross-document length comparisons, TTR was retained for three reasons. First, it provides a transparent, reproducible baseline comparable across prior multilingual studies. Second, the

primary research interest was in detecting the brevity artifact itself—which TTR reveals precisely because of its length dependency, making this limitation a feature of the analysis design rather than a confound. Third, the key finding (Japanese TTR inflation) is not compromised by length sensitivity but explained by it. Future studies using cross-lingual lexical diversity as a primary outcome measure should default to MTLD, which computes lexical diversity in a length-independent manner [23].

**Refusal and Minimal Compliance Classification:**

Refusal type classification was implemented as a fully automated, rule-based procedure in Python [27], applied independently to each of the 960 responses. Classifications were assigned in the following priority order: (1) No Response — responses with a word count of zero or an API error flag (n = 1); (2) Minimal — responses with fewer than 20 words after whitespace tokenization (n = 67); (3) Refused/Hedged — responses containing at least one term from a 47-item predefined marker lexicon including 'I cannot,' 'I am unable to,' 'I apologize,' 'I am not able,' 'this falls outside,' and semantic equivalents in Spanish, Urdu, Japanese, and Arabic (translated by the native-speaker translators); (4) Full Answer — all remaining responses (n = 875). The 20-word threshold for Minimal was selected based on pilot inspection of 50 randomly sampled short responses, all of which were confirmed as non-substantive. No human override was applied to any automated classification.

**Formality Index and its Limitations:**

The rule-based formality index was computed on a 0–100 scale from the relative frequency of formal and informal discourse markers per 100 words. Formal markers included academic connectives (therefore, consequently), passive constructions (was found, were observed), and nominalizations; informal markers included contractions (don't, can't), discourse fillers (well, so), and colloquial amplifiers. Marker lists were developed from English-language usage patterns and translated to the target languages via machine translation without native-speaker calibration. This is a significant limitation: languages such as Japanese and Arabic encode formality through morphological verb endings, honorific registers, and pronoun hierarchies that frequency-based marker lists do not capture. Results for this metric should be treated as exploratory only. The inconclusive findings reported in the Results section are attributable in part to this calibration limitation rather than to genuine cross-lingual uniformity in register.

**Statistical Analysis:**

For continuous metrics, one-way ANOVA was used with eta-squared ($\eta^2$) effect sizes, followed by Bonferroni-corrected post-hoc comparisons where the omnibus test was significant. For categorical metrics, chi-square tests of independence were used with Cramér's V as effect size. All analyses used Python/SciPy (v1.10+) with $\alpha = 0.05$. Effect sizes are classified as small ($\eta^2 \geq 0.01$), medium ($\eta^2 \geq 0.06$), or large ($\eta^2 \geq 0.14$) following Cohen's (1988) conventions.

**Results:**

**Response Length Disparities by Language:**

ANOVA revealed highly significant differences in response length across languages ($F(4, 955) = 34.93$, $p < 0.0001$, $\eta^2 = 0.128$) (Table 2, Figure 1).

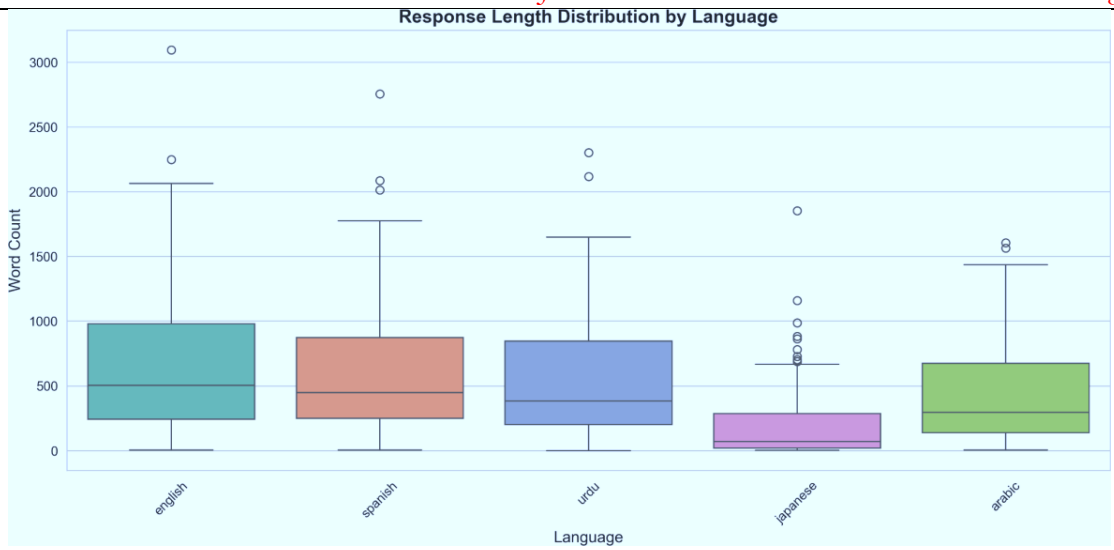


**Figure 2.** Boxplot showing response word count distribution across the five languages. The Japanese distribution is visibly compressed relative to all other languages, with a median of 69.5 words versus 503.0 for English.

**Table 2.** Response Length Statistics by Language.

| Language | Mean (words) | Median | Std Dev | Min | Max | IQR |
|---|---|---|---|---|---|---|
| English | 646.9 | 503.0 | 532.4 | 5 | 3093 | 737.8 |
| Spanish | 547.9 | 448.0 | 423.2 | 5 | 2754 | 623.8 |
| Urdu | 505.7 | 383.5 | 407.4 | 0 | 2300 | 645.5 |
| Arabic | 421.4 | 294.5 | 369.0 | 3 | 1605 | 537.0 |
| Japanese | 187.1 | 69.5 | 254.6 | 1 | 1851 | 265.3 |

At a mean of 187.1 words against English's 646.9, Japanese responses are 71% shorter on average (Cohen's d = 1.14, 95% CI [0.91, 1.37], large effect). The median English response is roughly six times longer than the median Japanese response (503.0 vs. 69.5 words). Bonferroni-corrected post-hoc comparisons confirm Japanese differs significantly from every other language ($p < 0.001$ for all four pairwise comparisons).

**Model-Specific Response Patterns:**

Models exhibit distinct verbosity profiles independent of language effects ($F(3, 956) = 128.57$, $p < 0.0001$, $\eta^2 = 0.287$) (Table 3, Figure 2).

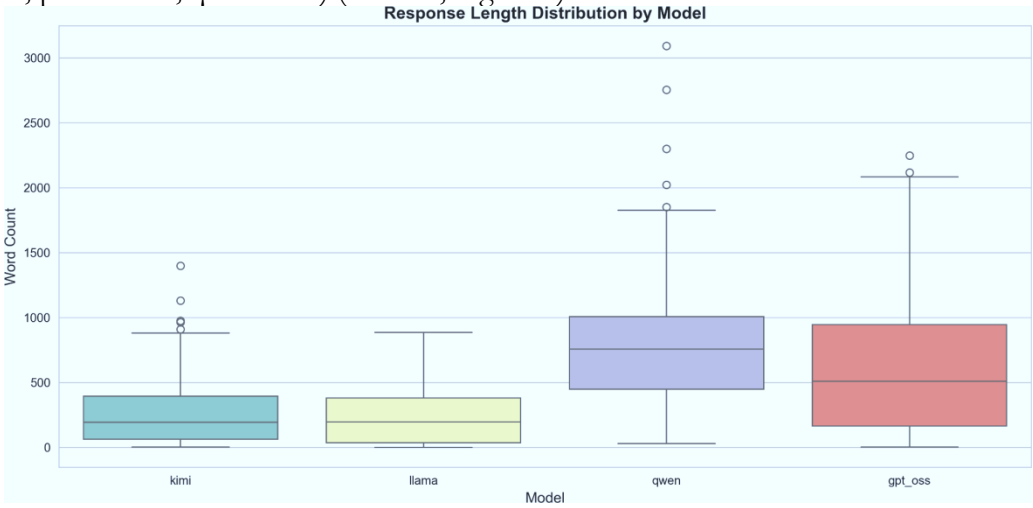


**Figure 3.** Boxplot showing response word count distribution across the four models. Qwen3 32B is the most verbose (mean 770.9 words); Llama 3.3 70B the tersest (mean 218.9 words).

**Table 3.** Response Length Statistics by Model.

| Model | Mean (words) | Median | Std Dev | Response Time (s) |
|---|---|---|---|---|
| GPT-OSS 120B | 603.1 | 508.5 | 508.4 | 9.88 |
| Qwen3 32B | 770.9 | 755.5 | 437.6 | 4.69 |
| Kimi K2 | 254.4 | 193.5 | 235.6 | 3.68 |
| Llama 3.3 70B | 218.9 | 196.0 | 189.5 | 10.54 |

Qwen3 32B produces the most verbose responses at a mean of 770.9 words. Critically, all four models exhibit the Japanese response penalty despite their different baseline verbosity profiles, pointing toward a structural factor shared across architectures.

**Refusal and Minimal Compliance Patterns:**

Chi-square analysis reveals significant association between language and refusal type ($\chi^2(12) = 100.54$, $p < 0.0001$, Cramér's V = 0.187, medium effect) (Table 4, Figures 4 and 5).

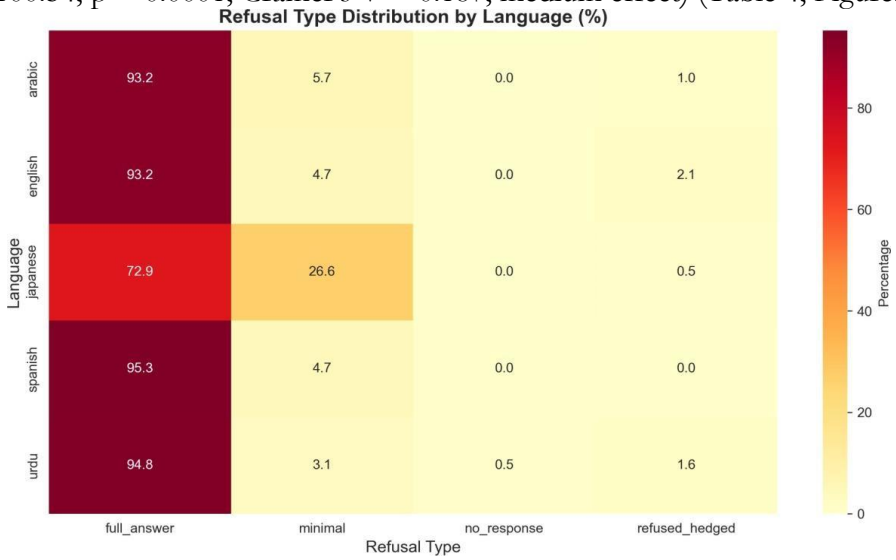


**Figure 4.** Heatmap showing refusal type distribution percentages by language. The Japanese 'Minimal' cell shows markedly elevated values (26.6%).

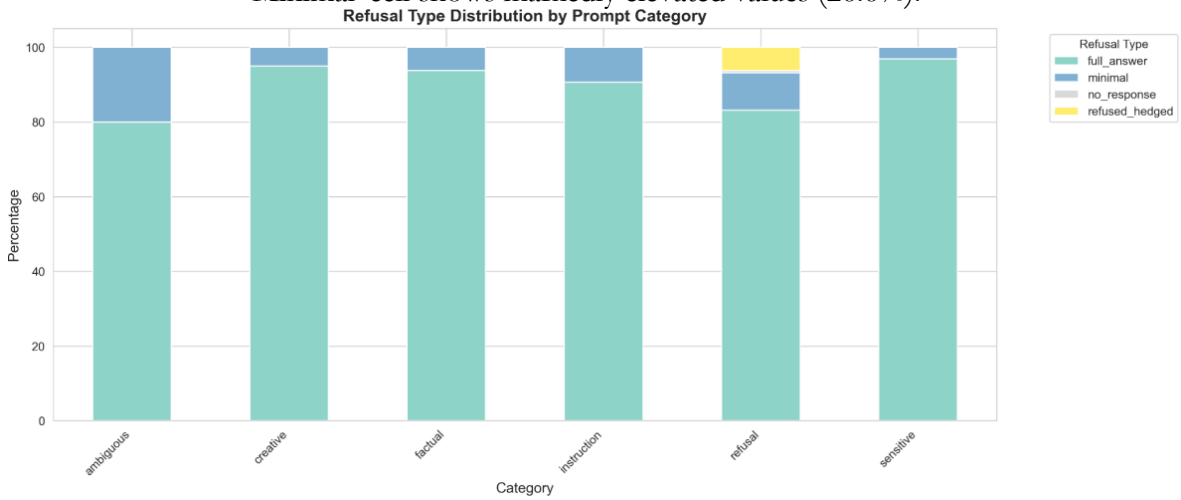


**Figure 5.** Stacked bar chart showing refusal type distribution across the six prompt categories. Refusal-Designed prompts elicit 17.1% refused/hedged responses versus <2% for other categories.

Japanese stands out with 26.6% of responses classified as minimal compliance. Standardized residuals confirm that Japanese minimal responses are the single largest contributor to the chi-square test statistic ($z = 9.8$, $p < 0.001$). Explicit refusals remain below

2% across all five languages, confirming that the observed pattern reflects engagement degradation rather than a safety calibration issue.

**Table 4.** Refusal Type Distribution by Language (Percentages).

| Language | Full Answer | Minimal | Refused/Hedged | No Response |
|---|---|---|---|---|
| Spanish | 95.3% | 4.7% | 0.0% | 0.0% |
| Urdu | 94.8% | 3.1% | 1.6% | 0.5% |
| Arabic | 93.2% | 5.7% | 1.0% | 0.0% |
| English | 93.2% | 4.7% | 2.1% | 0.0% |
| Japanese | 72.9% | 26.6% | 0.5% | 0.0% |

**Lexical Diversity Paradox:**

ANOVA shows significant lexical diversity differences ($F(4, 955) = 112.44$, $p < 0.0001$, $\eta^2 = 0.320$, large effect) (Table 5, Figure 6).

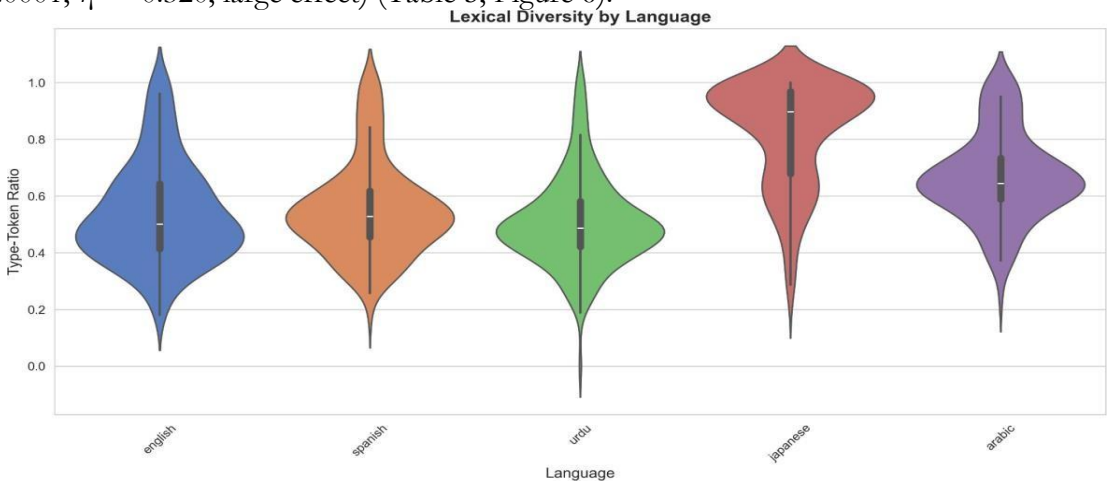


**Figure 6.** Violin plot showing Type-Token Ratio (TTR) distribution across five languages. Japanese shows a distinctly high, right-skewed distribution reflecting the brevity artifact.

**Table 5.** Lexical Diversity by Language with 95% Confidence Intervals.

| Language | Mean TTR | Median | Std Dev | 95% CI | Interpretation |
|---|---|---|---|---|---|
| Japanese | 0.823 | 0.897 | 0.184 | [0.797, 0.849] | Artificially high (brevity artifact) |
| Arabic | 0.665 | 0.644 | 0.155 | [0.643, 0.687] | Moderately high |
| Spanish | 0.552 | 0.528 | 0.168 | [0.528, 0.576] | Moderate |
| English | 0.540 | 0.501 | 0.179 | [0.515, 0.565] | Moderate |
| Urdu | 0.507 | 0.486 | 0.157 | [0.485, 0.529] | Moderate-low |

Japanese responses show the highest mean TTR at 0.823—a direct consequence of brevity, not lexical richness. When responses are only 10–20 words long, nearly every word is unique by default. This is the 'lexical diversity paradox': the language receiving the least engaged responses scores highest on the metric supposedly measuring linguistic sophistication.

**Hedging Language and Formality:**

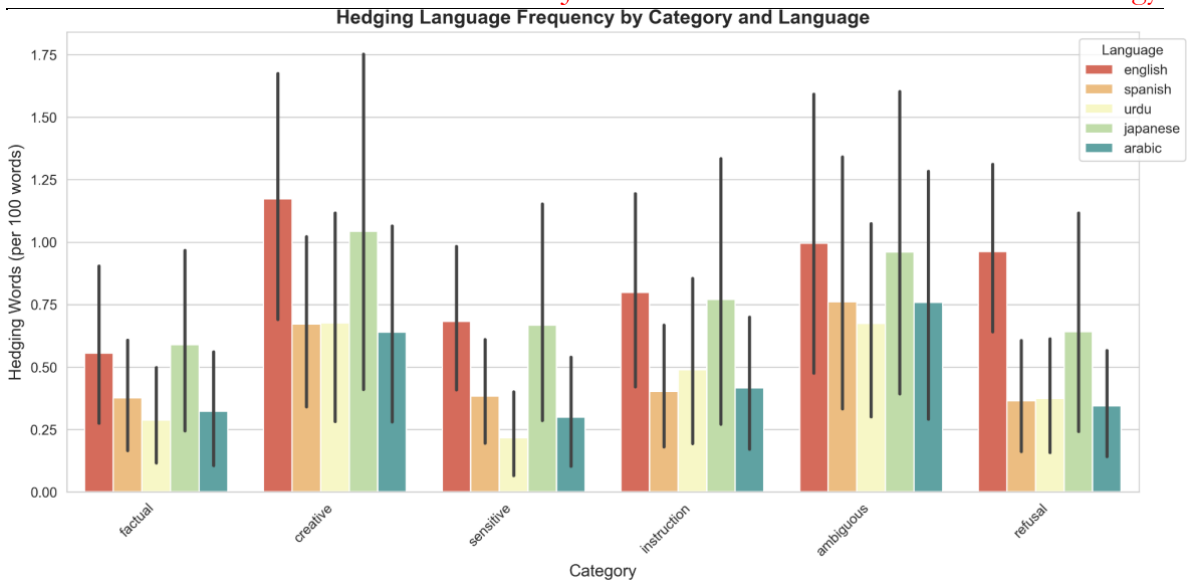


**Figure 7.** Grouped bar chart showing hedging frequency across six prompt categories and five languages. Creative prompts consistently show the highest hedging frequency (1.1 per 100 words).

Hedging frequency varies modestly by category but shows no consistent language pattern (Figure 7). Creative prompts elicit slightly more hedging (1.1 per 100 words) than factual prompts (0.6 per 100 words).

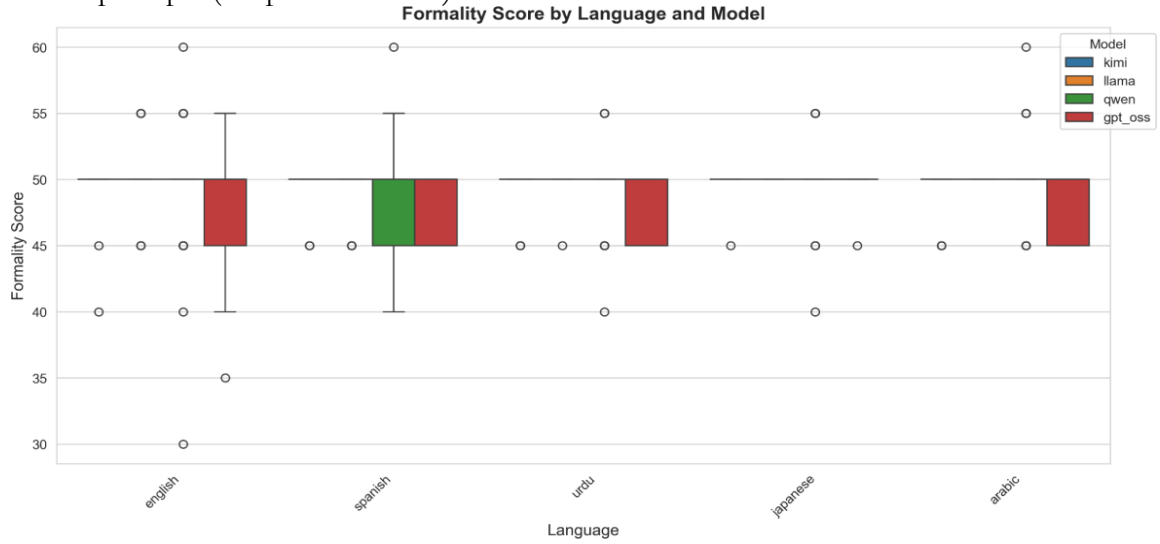


**Figure 8.** Boxplot showing formality scores across five languages and four models. All conditions cluster tightly in the 45–50 range.

Formality scores cluster tightly in the 45–50 range across all conditions ($F(4, 955) = 1.83$, $p = 0.12$, $\eta^2 = 0.008$, small non-significant effect) (Figure 8), indicating the rule-based metric did not detect meaningful register variation. As detailed in Materials and Methods, this null result should be interpreted with caution: the metric's marker lists were not calibrated for non-English languages. The inconclusive finding for formality reflects a measurement limitation rather than evidence of genuine cross-lingual register uniformity. Future studies should employ language-native formality scales or morphological analyzers appropriate to each target language.

**Temporal Efficiency:**

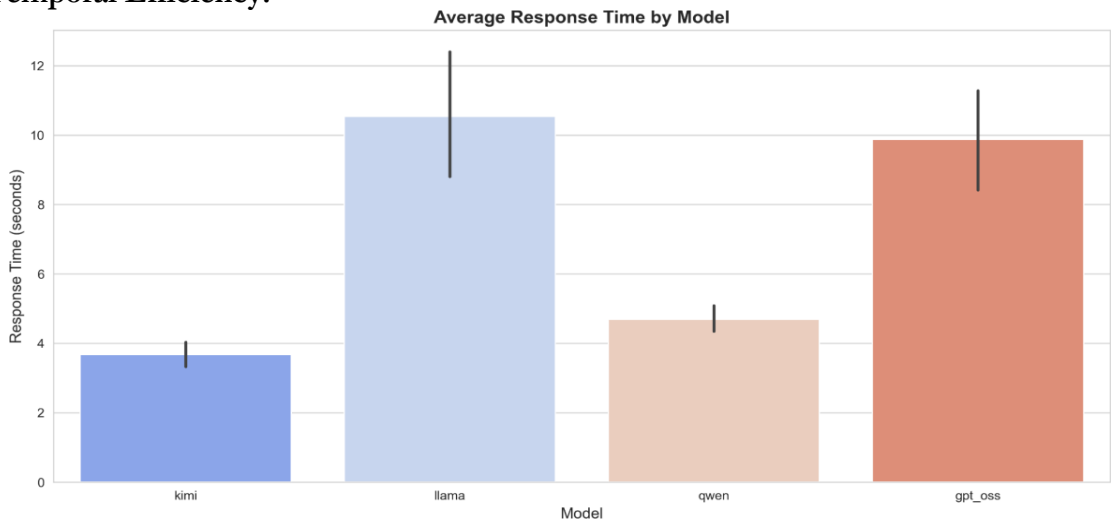


**Figure 9.** Bar chart with error bars (±1 SD) showing average response time by model.

Response time and efficiency metrics by model are summarized in Table 6 and Figure 9.

**Table 6.** Response Time by Model.

| Model | Mean Time (s) | Efficiency Score (words/second) |
|---|---|---|
| Kimi K2 | 3.68 | 69.1 |
| Qwen3 32B | 4.69 | 164.4 |
| GPT-OSS 120B | 9.88 | 61.1 |
| Llama 3.3 70B | 10.54 | 20.8 |

**Cross-Tabulation: Language × Category Interactions:**

**Table 7.** Mean Word Count by Language and Category.

| Language | Ambiguous | Creative | Factual | Instruction | Refusal | Sensitive |
|---|---|---|---|---|---|---|
| Arabic | 145.8 | 548.4 | 434.6 | 496.5 | 390.3 | 513.1 |
| English | 249.2 | 945.8 | 678.1 | 652.7 | 571.7 | 784.2 |
| Japanese | 86.9 | 240.8 | 157.8 | 280.2 | 147.9 | 209.0 |
| Spanish | 172.4 | 761.2 | 597.0 | 514.7 | 522.7 | 719.5 |
| Urdu | 191.7 | 640.8 | 592.2 | 504.3 | 488.7 | 616.6 |

The Japanese response penalty appears across all six categories (70–75% shorter than English). The two-way ANOVA confirms significant Language ($F(4, 940) = 31.2, p < 0.0001$) and Category ($F(5, 940) = 8.7, p < 0.0001$) main effects (Table 7). The non-significant interaction ($F(20, 940) = 1.3, p = 0.17$) confirms the Japanese penalty applies uniformly across all prompt types, ruling out task-specific explanations.

**Discussion:**

**Interpretation of Findings:**

The results make a strong case for systematic bias in LLM response quality that favors Western languages, and specifically English. What makes the Japanese finding particularly hard to dismiss is that Japanese is not a low-resource language. It has extensive digital presence and was almost certainly well-represented in the pretraining corpora of all four models tested. The 71% response length disparity cannot be explained by data scarcity alone.

Three mechanistic explanations are proposed for the observed disparities. It is important to note that the present study is observational and no ablation experiments were conducted to isolate causal factors. These are theoretically motivated, data-consistent hypotheses rather than confirmed causal explanations.

The engagement degradation hypothesis is the most directly supported by the data: the 26.6% minimal-compliance rate for Japanese ($\chi^2(12) = 100.54, p < 0.0001$) combined

with broadly comparable quality in the 72.9% of Japanese prompts that receive full responses is consistent with models deprioritizing elaboration rather than being incapable of it. This finding is empirically demonstrated.

The tokenization inefficiency hypothesis [8][9]—that Japanese inputs consume the token budget more quickly under standard BPE tokenizers, producing truncated outputs—is theoretically consistent with the data but was not directly tested. Token counts per response were not recorded and would be required to test this hypothesis. The training data imbalance [20] and instruction tuning asymmetry hypotheses [28][21][14] are similarly plausible but speculative; direct evidence would require access to training data composition and instruction tuning language distributions that are not publicly available for the models tested.

**API Configuration and Potential Confounds:**

A potential confounding factor that requires explicit consideration is the role of model-specific output token limits and API configurations. In this study, max_tokens was set uniformly to 4096 across all four models and language conditions, and the Groq API documentation for each model was verified to confirm no model-level caps below 4096 were in effect. Since the observed maximum response length across all conditions was 3093 words (English, Table 2)—corresponding to approximately 3500–4500 tokens depending on tokenizer—the 4096 token ceiling was not a binding constraint for any condition. The shorter Japanese responses therefore reflect generation-time behavior rather than hard output token limits. Output token limits are explicitly ruled out as a confounding explanation for the observed Japanese response penalty.

**Implications for Global AI Deployment:**

The economic dimension is straightforward. Token-based API pricing means Japanese users pay comparable rates while receiving 71% less content per query—before factoring in the tokenization premium where non-Latin scripts require more tokens to encode the same input [8]. The two effects compound: more tokens are spent on the prompt, while fewer words are received in the output. Empirically, the mean word-count difference between English and Japanese responses is 459.8 words (95% CI [376.3, 543.3])—equivalent to approximately 2.5–3 paragraphs of substantive content missing from every Japanese response.

The functional consequences are equally concrete. A user asking for a medical explanation, a step-by-step tutorial, or an analytical breakdown is not well-served by a 70-word response. Consistent provision of less detailed responses implicitly signals that those queries are less worthy of serious attention—a representational harm as described by Blodgett et al. [6] [29] and documented across language AI systems by [17].

**Practical Implications:**

The findings carry direct implications for three stakeholder groups. For AI developers, the results establish that behavioral testing at the generation level—measuring response length, engagement depth, and refusal rates by language—should become standard practice alongside task-accuracy benchmarks. Current evaluation practices are insufficient to detect the systematic engagement degradation documented here. Audit protocols for multilingual LLM deployment should include cross-lingual response quality parity as an explicit evaluation criterion.

For policymakers and regulatory bodies, the economic asymmetry documented here is relevant to emerging AI fairness and consumer protection frameworks. Where AI tools are publicly funded or deployed in public services, cross-lingual response parity should constitute a minimum service quality standard [30]. Regulatory guidance on AI fairness should explicitly address generative response disparities and not only classification or decision-making outcomes.

For multilingual users and advocacy organizations, these results provide empirical evidence for concerns that have been raised anecdotally about non-English AI quality. Users

in affected language communities should be informed of documented disparities and encouraged to provide systematic quality feedback, which can serve as a signal to developers and regulators.

**Recommendations for Implementation:**

Four specific implementation recommendations are offered based on the observed disparities and likely mechanisms.

**Tokenizer Redesign:** Tokenizers should be evaluated on cross-lingual token efficiency parity, not solely on compression efficiency for the dominant training language. Open tokenizer benchmarks measuring information content per token across at minimum twenty typologically diverse languages should be adopted as a community standard, with tokenizer efficiency reports included in model releases.

**Multilingual Instruction Tuning Protocols:** RLHF and supervised fine-tuning datasets [31] should include language-stratified data with verified balance across at minimum the ten most widely spoken languages. The proportion of non-English instruction examples should be reported in model cards with a minimum recommended threshold of 20% non-English demonstrations [14][15].

**Cross-Lingual Response Quality Evaluation Standard:** Model cards should include standardized cross-lingual response quality benchmarks reporting at minimum: median response length ratio (target language / English), minimal compliance rate, response latency parity, and MTLD-based lexical diversity. These metrics are low-cost, automatable, and interpretable by non-specialist stakeholders.

**Validated Multilingual Evaluation Claims:** New model releases claiming multilingual capability should include empirical response quality parity data across the claimed languages. A model that produces 70-word Japanese responses to prompts generating 500-word English responses does not demonstrate multilingual capability in any functional sense.

**Toward Equitable Multilingual LLMs:**

Training data curation is the most fundamental starting point. Proportional representation alone may not be sufficient if non-English data is lower quality; deliberate oversampling of high-quality non-English data may be necessary [15][28]. Tokenizer design is the second lever [9]. Evaluation frameworks must change to include response quality metrics by language as standard reporting requirements in model cards.

**Limitations:**

The model selection covers four systems, insufficient to generalize to the full LLM landscape. All four were accessed via Groq API, introducing the possibility of platform-specific effects. Proprietary models like GPT-4 and Claude were not included. Language coverage is limited to five of the world's roughly 7,000 languages.

**Metric granularity is a genuine limitation:** rule-based refusal classification misses nuanced engagement degradation, TTR has known length-sensitivity issues, and the formality metric lacks cross-lingual calibration. This study documents disparities but cannot definitively isolate causes—disentangling tokenization, training data, and instruction tuning effects would require controlled ablation studies beyond the scope of this behavioral investigation.

The absence of human evaluation is a significant limitation of the automated metrics framework employed here [32]. None of the eight metrics has been validated against native-speaker quality judgments in this study's experimental context. Inter-rater agreement studies involving native speakers of each of the five languages would be required to establish whether the automated engagement degradation signals correspond to human-perceived quality differences. This is identified as the highest-priority methodological extension for follow-up research.

**Future Directions:**

The most pressing direction is causal mechanism research via controlled ablation experiments systematically manipulating tokenizer design, training data language ratios, and instruction tuning composition. Expanded language coverage should include Swahili, Quechua, and Telugu. Human evaluation by native speakers and longitudinal tracking of model improvements are the other key directions.

**Conclusion:**

This study set out to ask a straightforward question: do LLMs perform the same across languages? Based on 960 responses across four models and five languages, the assumption of language-agnostic capability does not hold. Japanese prompts receive responses 71% shorter than English equivalents (Cohen's $d = 1.14$, $p < 0.0001$), with more than one in four classified as minimal compliance. That pattern holds across all four models and all six prompt categories.

Four concrete commitments are urged: (1) expand multilingual evaluation beyond narrow task benchmarks; (2) mandate cross-lingual fairness metrics in model cards; (3) invest in training data curation and tokenizer redesign; (4) involve marginalized linguistic communities directly in AI development processes. The dataset and analysis code are publicly available at https://github.com/saadnizami1/Linguistic-Relativity-in-Artificial-Cognition.

**Acknowledgments:**

This study would not have been possible without the native-speaker translators who worked through each prompt carefully, prioritizing cultural and semantic equivalence over literal translation. We also acknowledge Groq for API access that made data collection at this scale feasible.

**Author Contributions:**

Muhammad Ramzan Shahid Khan: Conceptualization, Study Design, Data Collection, Methodology, Formal Analysis, Writing – Original Draft. Muhammad Saad Nizami: Investigation, Data Validation, Writing – Review & Editing, Statistical Review. All authors have read and agreed to the final published version of the manuscript.

**Originality and Publication Statement:**

The authors confirm that this manuscript has not been previously published, nor is it currently under consideration for publication by any other journal. All authors have contributed significantly to the research and are in full agreement with the content of the manuscript as submitted.

**Conflict of Interest:** The authors declare no conflicts of interest.

**References:**

[1] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, "PaLM: Scaling Language Modeling with Pathways," *arXiv:2204.02311*, 2022, [Online]. Available: https://arxiv.org/abs/2204.02311

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv:2106.09685*, 2021, [Online]. Available: https://arxiv.org/abs/2106.09685

[3] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, "MEGA: Multilingual Evaluation of Generative AI," *arXiv:2303.12528*, 2023, [Online]. Available: https://arxiv.org/abs/2303.12528

[4] Ian Magnusson, Noah A. Smith, Jesse Dodge, "Reproducibility in NLP: What Have We Learned from the Checklist?," *arXiv:2306.09562*, 2023, [Online]. Available: https://arxiv.org/abs/2306.09562

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, "Unsupervised Cross-lingual Representation Learning at Scale," *arXiv:1911.02116*, 2020, [Online]. Available: https://arxiv.org/abs/1911.02116

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," *arXiv:2005.14050*, 2020, [Online]. Available: https://arxiv.org/abs/2005.14050

[7] I. Solaiman *et al.*, "Evaluating the Social Impact of Generative AI Systems in Systems and Society," Jun. 2024, Accessed: Jun. 20, 2026. [Online]. Available: http://arxiv.org/abs/2306.05949

[8] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models," *arXiv:2305.13707*, 2023, [Online]. Available: https://arxiv.org/abs/2305.13707

[9] Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, Adel Bibi, "Language Model Tokenizers Introduce Unfairness Between Languages," *arXiv:2305.15425*, 2023, [Online]. Available: https://arxiv.org/abs/2305.15425

[10] Yanzhu Guo, Guokan Shang, Chloé Clavel, "Benchmarking Linguistic Diversity of Large Language Models," *arXiv:2412.10271*, 2025, [Online]. Available: https://arxiv.org/abs/2412.10271

[11] Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, Fei Yuan, "BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models," *arXiv:2502.07346*, 2025, [Online]. Available: https://arxiv.org/abs/2502.07346

[12] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, "Language Models are Multilingual Chain-of-Thought Reasoners," *arXiv:2210.03057*, 2022, [Online]. Available: https://arxiv.org/abs/2210.03057

[13] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, Thien Huu Nguyen, "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning," *arXiv:2304.05613*, 2023, [Online]. Available: https://arxiv.org/abs/2304.05613

[14] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning," *arXiv:2301.13688*, 2023, [Online]. Available: https://arxiv.org/abs/2301.13688

[15] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, "Crosslingual Generalization through Multitask Finetuning," *arXiv:2211.01786*, 2023, [Online]. Available: https://arxiv.org/abs/2211.01786

[16] B. H. Pinzhen Chen, Simon Yu, Zhicheng Guo, "Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models?," *arXiv:2406.12822*, 2024, [Online]. Available: https://arxiv.org/abs/2406.12822

[17] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, "Bias and Fairness in Large Language Models: A Survey," *arXiv:2309.00770*, 2024, [Online]. Available: https://arxiv.org/abs/2309.00770

[18] Krithika Ramesh, Sunayana Sitaram, Monojit Choudhury, "Fairness in Language Models Beyond English: Gaps and Challenges," *arXiv:2302.12578*, 2023, [Online]. Available: https://arxiv.org/abs/2302.12578

[19] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, Sebastian Ruder, "UNKs Everywhere: Adapting Multilingual Language Models to New Scripts," *arXiv:2012.15562*, 2021, [Online]. Available: https://arxiv.org/abs/2012.15562

[20] L. W. Julia Kreutzer, Isaac Caswell, "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets," *arXiv:2103.12028*, 2022, [Online]. Available: https://arxiv.org/abs/2103.12028

[21] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, Quoc V. Le, "Finetuned Language Models Are Zero-Shot Learners," *arXiv:2109.01652*, 2022, [Online]. Available:

https://arxiv.org/abs/2109.01652
[22] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, "Scaling Instruction-Finetuned Language Models," *arXiv:2210.11416*, 2022, [Online]. Available: https://arxiv.org/abs/2210.11416
[23] P. M. McCarthy and S. Jarvis, "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment," *Behav. Res. Methods 2010 422*, vol. 42, no. 2, pp. 381–392, May 2010, doi: 10.3758/BRM.42.2.381.
[24] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," *arXiv:2303.16634*, 2023, [Online]. Available: https://arxiv.org/abs/2303.16634
[25] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, Xiaojun Wan, "LLM-based NLG Evaluation: Current Status and Challenges," *arXiv:2402.01383*, 2025, [Online]. Available: https://arxiv.org/abs/2402.01383
[26] Rotem Dror, Gili Baumer, Segev Shlomov, Roi Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, 2018, [Online]. Available: https://aclanthology.org/P18-1128/
[27] P. Naresh, S. V. N. Pavan, A. R. Mohammed, N. Chanti, and M. Tharun, "Comparative Study of Machine Learning Algorithms for Fake Review Detection with Emphasis on SVM," *Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2023 - Proc.*, pp. 170–176, 2023, doi: 10.1109/ICSCSS57650.2023.10169190.
[28] Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, Matan Eyal, "Multilingual Instruction Tuning With Just a Pinch of Multilinguality," *arXiv:2401.01854*, 2024, [Online]. Available: https://arxiv.org/abs/2401.01854
[29] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," *FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar.*, pp. 610–623, Mar. 2021, doi: 10.1145/3442188.3445922.
[30] M. R. Hayat, A. Bakar, M. Bilal∗, and M. R. S. Khan, "Data-Driven Analysis and Predictive Modeling of Urban Air Quality for Environmental Management," *Int. J. Innov. Sci. Technol.*, vol. 8, no. 3, pp. 508–527, 2026, Accessed: Jun. 30, 2026. [Online]. Available: https://ideas.repec.org/a/abq/ijist1/v8y2026i3p508-527.html
[31] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv:2305.14314*, 2023, [Online]. Available: https://arxiv.org/abs/2305.14314
[32] Cheng-Han Chiang, Hung-yi Lee, "Can Large Language Models Be an Alternative to Human Evaluations?," *arXiv:2305.01937*, 2023, [Online]. Available: https://arxiv.org/abs/2305.01937