# LINGUISTIC NEPOTISM: TRADING-OFF QUALITY FOR LANGUAGE PREFERENCE IN MULTILINGUAL RAG

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multilingual Retrieval-Augmented Generation (mRAG) systems enable language models to answer knowledge-intensive queries with citation-supported responses across languages. While such systems have been proposed, an open questions is whether the mixture of different document languages impacts generation and citation in unintended ways. To investigate, we introduce a controlled methodology using model internals to measure language preference while holding other factors such as document relevance constant. Across eight languages and six open-weight models, we find that models preferentially cite English sources when queries are in English, with this bias amplified for lower-resource languages and for documents positioned mid-context. Crucially, we find that models sometimes trade-off document relevance for language preference, indicating that citation choices are not always driven by informativeness alone. Our findings shed light on how language models leverage multilingual context and influence citation behavior.[1]

## 1 INTRODUCTION

Retrieval-Augmented Generation (RAG) systems have become a core component of modern large language model (LLM) pipelines, enabling models to answer knowledge-intensive queries by supplementing their limited parametric knowledge with external information (Lewis et al., 2020; Karpukhin et al., 2020; Gao et al., 2024). Given that over 50% of digital content is produced in languages other than English (Statista, 2025), recent work has extended these systems to multilingual RAG (mRAG) settings, which handle queries and documents in languages beyond English (Chirkova et al., 2024; Wu et al., 2024).

Despite recent advances, prior work highlights a key challenge in mRAG systems: **language preference**—a systematic tendency of models to favor sources written in certain languages during generation (Park & Lee, 2025). Understanding this behavior is crucial, as citation patterns shape both the information users see and the languages prioritized in multilingual knowledge access.

Existing approaches to measuring language preference, however, often fail to capture citation correctness. In short-form mRAG, preference has been estimated via information overlap (Sharma et al., 2025) or embedding similarity (Park & Lee, 2025), which do not directly account for correctness. In long-form mRAG, where outputs contain in-line citations (Zheng et al., 2025; Xu & Peng, 2025), preference has typically been measured by comparing citation frequencies against the language distribution of retrieved documents. This signal is coarse and confounded by the relevance and informativeness of multilingual sources ($C_1$). Moreover, in-line citations are prone to hallucinations (Gao et al., 2023; Zhang et al., 2024), making it unclear whether observed preferences reflect true attribution or spurious citations ($C_2$).

To address these challenges, we propose a controlled methodology for measuring language preference using model internal metrics (illustrated in Figure 1). We first construct a synthetic multi-parallel dataset of relevant documents, which allows us to isolate the effect of language while controlling for other factors such as document content and relevance (Step 1+2; addresses $C_1$). Citation correctness is then verified through a two-step filtering process (Step 3; addresses $C_2$) (§3.1). Next, we compare the accuracy of next token citation predictions (*e.g.,* predicting "2" for document ID 2)

[1] Code and data will be released upon publication.

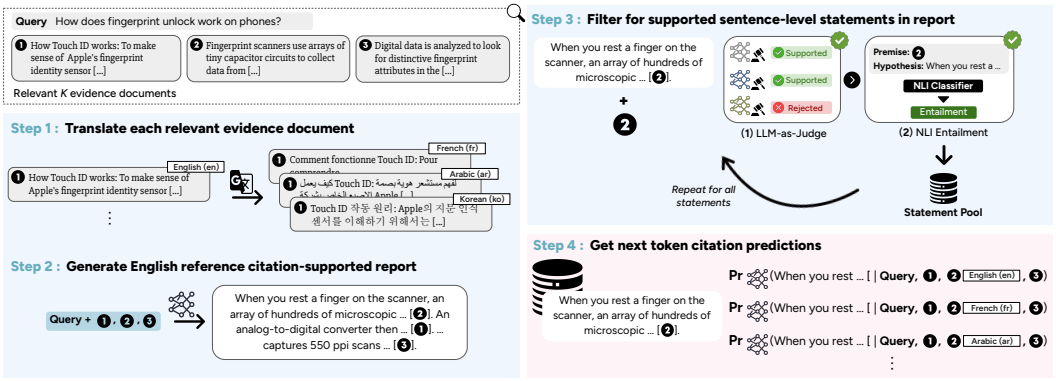



Figure 1: **Overview of our approach for measuring language preference.** We show both synthetic data generation and measurement method. Given an English query $q$ and its $K$ relevant evidence documents $D_{en}$, we first translate the documents into multiple languages $D_{fr}, D_{ar}, D_{ko} \ldots$ (Step 1). We then generate a *reference* citation-supported report $r$ for each query using $q$ and $D_{en}$ (Step 2). The report $r$ consists of sentence-level statements $s_i$, each paired with a single citation ID $c_i$. For each $r$, we retain only statements that are verified (Step 3). Language preference is detected when the next token prediction accuracy for the correct citation ID decreases as the language of the cited document is varied (Step 4).

while varying the language of the same cited document and keeping other variables fixed, including the language of remaining documents, document positions in the input context, and the query language (Step 4). Differences in citation accuracy between languages indicate a preference for the higher-accuracy language (§3.2).

Using this setup across eight languages and six open-weight models, we address the overarching question: Do models preferentially cite documents in certain languages during long-form mRAG? To further inform building more robust mRAG systems, we empirically address three key questions: (a) What factors amplify language preference? (b) What role does the query language play in language preference? and (c) Is citation behavior driven more by document relevance or language?

Our main findings can be summarized as follows:

- **Evidence of strong English preference:** Across all tested models, we find a pronounced tendency to cite English documents when the query is in English. This preference amplifies when: (1) the cited document is in a lower-resource language (*e.g.,* Bengali, Swahili), or (2) the cited document appears in the middle of the input context (§5).
- **Language preference towards query language:** We show that language preference extends beyond English: models favor citing evidence documents written in the query language (§6).
- **Language outweigh relevance:** Last but not least, we show that models frequently cite English documents even when they are irrelevant to the query, suggesting that language itself exerts a stronger influence than document relevance in long-form mRAG (§7).

## 2 Related Work

**Multilingual RAG.** A growing body of work has examined that large language models (LLMs) are prone to hallucinations, especially in knowledge-intensive tasks (Augenstein et al., 2024; Huang et al., 2025a). Retrieval-augmented generation (RAG) mitigates this by retrieving external knowledge sources and incorporating them into generation (Chen et al., 2024; Gao et al., 2024). While early RAG systems largely focused on processing English queries and sources, recent research has extended these methods to multilingual RAG (mRAG), enabling retrieval and generation across a wider range of languages (Asai et al., 2022). Prior mRAG studies primarily examine the effects of query language (Chirkova et al., 2024), the language of relevant or irrelevant evidence documents (Wu et al., 2024; Qi et al., 2025; Liu et al., 2025), document ordering (Ranaldi et al., 2025a), and

prompting strategies (Ranaldi et al., 2025b) on performance. However, due to cost efficiency and scalability (Saad-Falcon et al., 2024; Es et al., 2024), most of this work targets short-form mRAG, where the output is a brief answer to a factoid-style query (*e.g.,* What is the capital of France?). In contrast, we focus on long-form mRAG, where models are asked to generate citation-supported reports in response to open-ended queries (*e.g.,* How does fingerprint unlock work on phones?).

**Long-form (m)RAG.** Long-form RAG systems build upon prior work on long-form question answering (LFQA) datasets (Dasigi et al., 2021; Stelmakh et al., 2023) to generate paragraph level, citation-supported responses for complex, knowledge-intensive queries (Zhao et al., 2024; Wei et al., 2024; Ju et al., 2025; Zhang et al., 2025). Although evaluating models on long-form outputs is notoriously challenging (Qi et al., 2024), it is also increasingly important as it better mirrors how humans naturally interact with search engines (Khashabi et al., 2021), making such systems more easily integrable into search-based workflows like Deep Research platforms (Huang et al., 2025b; Zheng et al., 2025). Similarly, we use a long-form RAG dataset, Explain Like I'm Five (ELI5) (Fan et al., 2019), to measure language preference.

**Language Preference.** Language preference describes a systematic tendency for models to favor sources in certain languages over others. This preference largely arises from differences in training data distribution, tokenization methods, and resource availability (Wu et al., 2024; Sharma et al., 2025; Shen et al., 2024). Such preference manifests at both the retrieval and generation stages. On the retrieval side, prior work shows that multilingual information retrieval (MLIR) systems tend to favor high-resource languages (*e.g.,* English) while under-representing sources in lower-resource languages, which can degrade retrieval quality (Telemala & Suleman, 2022; Yang et al., 2024; Amiraz et al., 2025) and introduce inconsistencies in generation (Chataigner et al., 2024). On the generation side, language models have been found to more effectively utilize sources written in specific languages (Park & Lee, 2025). Existing studies on short-form mRAG typically measure this by querying models in different languages and analyzing information overlap (Sharma et al., 2025) or embedding similarity (Park & Lee, 2025) between outputs and reference answers. In the long-form setting, prior work approximates language preference by comparing citation rates against the distribution of available documents per language, where over-representation in citations signals bias (Li et al., 2025). We build our work on this line of measuring language preference in long-form mRAG, but through a more controlled experimental setup using model internal metrics.[2]

## 3 Measuring Language Preference in Long-form mRAG

Our goal is to measure whether LLMs systematically prefer citing evidence in some languages over others. To do this, we need (a) a multilingual dataset of queries with parallel evidence documents and verifiable citation-supported reports (§3.1), and (b) a measurement method that compares citation accuracy when the same document is presented in different languages (§3.2). Figure 1 shows the pipeline for dataset construction and measurement. All prompts are provided in Appendix A.

### 3.1 Synthetic Data Generation

**Step 1: Evidence Document Translation.** Let $\mathcal{D}_{en} = \{d_1, \ldots, d_K\}$ denote the set of $K$ relevant evidence documents in English associated with a query $q$. Since no parallel long-form mRAG datasets are publicly available, we construct multilingual variants $\mathcal{D}_{\ell_{\text{target}}}$ for each target language $\ell_{\text{target}} \in \mathcal{L}_{\text{target}}$ using machine translation (MT). If $\text{MT}_\ell$ denote a translation function into language $\ell$, we obtain $\mathcal{D}_\ell = \{\text{MT}_\ell(d_1), \ldots, \text{MT}_\ell(d_K)\}$. In our experiments, $\text{MT}_\ell$ is implemented using Google Translate API. Despite the challenges of translating long-context documents (Wang et al., 2023; Cui et al., 2024; Wang et al., 2025b), the translation quality remains reasonable, with average COMET[3] quality estimation scores of 0.541. Per-language scores are reported in Appendix D.1.

**Step 2: Reference Report Generation.** For each query $q$ with associated English evidence document set $\mathcal{D}_{en} = \{d_1, \ldots, d_K\}$, we generate a *reference* citation-supported report using a strong

[2] Our measurement method is complementary to SEPER (Dai et al., 2025)—while SEPER provides a sampling-based measure of retrieval utility in open-ended generation, our task examines the model's token-level probabilities for the citation ID, targeting a more atomic decision.

[3] `Unbabel/wmt22-cometkiwi-da`

LLM $\mathcal{M}_{\text{gen}}$. We select OpenAI o3[4] as $\mathcal{M}_{\text{gen}}$, since its outputs were rated highest by human evaluators in SciArena (Zhao et al., 2025), a benchmark assessing long-form report generation and citation quality. The generated report is: $r = \mathcal{M}_{\text{gen}}(q, \mathcal{D}_{en})$.[5] We segment $r$ into $n$ sentence-level statements: $r = (s_1, [c_1], \ldots, s_n, [c_n])$, where $s_i$ is the $i$-th statement, and $c_i \in \{1, \ldots, K\}$ is the citation ID of the evidence document $d_{c_i} \in \mathcal{D}_{en}$ that $\mathcal{M}_{\text{gen}}$ cites as supporting $s_i$. By construction, $c_i$ denotes the citation token appearing in the report after $s_i$.

**Step 3: Statement Pool Construction.** Long-form generation with citations is prone to hallucination, with LLMs often introducing factual errors (Ji et al., 2023) or misattributing information to incorrect evidence (Gao et al., 2023; Magesh et al., 2024; Zhang et al., 2024). To ensure that only verifiably supported statements are retained for evaluation, we apply a two-stage filtering pipeline to the set of statement-citation pairs $\{(s_i, c_i)\}_{i=1}^{n}$ from Step 2. We perform filtering only if $|c_i| = 1$ (*i.e.,* statements with exactly one citation). First, the LLM-as-Relevance-Judge identifies statements whose cited document is deemed most relevant by the majority of judges—capturing correctness in the statement → cited document direction.[6] Second, the NLI entailment check verifies that the cited document actually entails the information in the statement—capturing in the cited document → statement direction. Using both steps ensure citation correctness from both directions.

**(1) LLM-as-Relevance-Judge:** Let $\mathcal{M}_{\text{judge}} = \{m_1, m_2, m_3\}$ be the set of judge models that rank highest on the SciArena benchmark (OpenAI o4 mini[4], QWEN-3 32B (Yang et al., 2025), and Gemini 2.5 Pro[7]). Each judge $m \in \mathcal{M}_{\text{judge}}$ is prompted with statement $s_i$ and the full evidence document set $\mathcal{D}_{en}$ to return the index of the most relevant document $j_m(s_i, \mathcal{D}_{en})$. Here, $j_m$ implements a relative selection task over all $\mathcal{D}_{en}$ (*i.e.,* "Which document best supports the statement?"), rather than an absolute binary support judgment (*i.e.,* "Does this document support the statement?"), following findings that comparative framing improves LLM evaluation accuracy (Godfrey et al., 2025; Shrivastava et al., 2025). The total number of judges selecting the cited document $d_{c_i}$ is:

$$\textbf{votes}(s_i, c_i) = \sum_{m \in \mathcal{M}_{\text{judge}}} \mathbb{1}(j_m(s_i, \mathcal{D}_{en}) = c_i) \tag{1}$$

We retain $s_i$ if when the majority of judges agree on the correct judgment: $\textbf{votes}(s_i, c_i) \geq 2$.

**(2) NLI Entailment:** We use an off-the-shelf Natural Language Inference (NLI) classifier $\phi$(premise, hypothesis)[8], which outputs 1 if the premise entails the hypothesis, and 0 otherwise. In our setting, $d_{c_i}$ is the premise and $s_i$ the hypothesis. We retain $s_i$ if $\phi(d_{c_i}, s_i) = 1$. This is in accordance with the Attributable to Identified Sources (AIS) framework (Rashkin et al., 2023).

In practice, the LLM-as-Relevance-Judge and NLI Entailment filtering stages achieve retain rates of 90.35% and 96.12%, respectively. The final pool consists of 792 statements that pass both filters[9], ensuring that the correctness of each citation used for evaluation is reliably verified.[10]

### 3.2 MEASUREMENT METHOD

**Step 4: Next Token Prediction Analysis.** Intuitively, if the model predicts the correct citation token when the cited document is in English compared to another language, this indicates a preference for English. To quantify this, for each verified statement-citation pair $(s_i, c_i)$, we measure the accuracy of the model predicting $c_i$ as the top-1 next token.

We first construct a citation prediction prompt ending in the form: $x_i = s_i$ `[`, where the bracket `[` signals the start of the citation. To test for language preference for English, we define the set of evaluation languages as $\mathcal{L}_{\text{eval}} = \{\text{en}\} \cup \mathcal{L}_{\text{target}}$, which includes English and all target languages. For each statement, we construct *contrastive* contexts where only the document to be cited, $d_{c_i}$, is presented in a language $\ell \in \mathcal{L}_{\text{eval}}$, while all other evidence documents remain in English. The full context is denoted as $\text{Context}(d_{c_i} \to \ell, d_{\neg c_i} \to \text{en})$. Given the prompt prefix $x_i$, the model's next

[4] `https://openai.com/index/introducing-o3-and-o4-mini/`

[5] On average, reports contain 148.5 words across 4.90 sentences.

[6] Prior work shows that LLMs provide precise relevance assessments (Ma et al., 2023; Sun et al., 2023).

[7] `https://deepmind.google/models/gemini/pro/`

[8] `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`

[9] On average, each verified statement contains 33.70 words.

[10] Human annotation results in Appendix C show high agreement with the automatic filtering judgments.

token probability of the correct citation ID token $c_i$ corresponding to document $d_{c_i}$ conditioned on this context is: $p_\theta^{(\ell)}(c_i) = \mathcal{P}_\theta(t = c_i | x_i, q, \text{Context}(d_{c_i} \to \ell, d_{\neg c_i} \to \text{en}))$, where $\mathcal{P}$ is the model's next token distribution given a prefix, and $\theta$ denotes model parameters. We define the model's top-predicted citation token as: $\hat{c}_i^{(\ell)} = \text{argmax}_t(p_\theta^{(\ell)})$, and compute citation accuracy in language $\ell$ over $n$ statements as:

$$\mathbf{Acc}^{(\ell)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\hat{c}_i^{(\ell)} = c_i). \tag{2}$$

A model exhibits English preference over a target language $\ell_{\text{target}} \in \mathcal{L}_{\text{target}}$ if it achieves higher citation accuracy when the cited document $d_{c_i}$ is in English than when it is in the target language. We define the citation accuracy gap as:

$$\Delta(\ell_{\text{target}}) = \mathbf{Acc}^{(\ell_{\text{target}})} - \mathbf{Acc}^{(\text{en})}. \tag{3}$$

In other words, $\Delta(\ell_{\text{target}})$ quantifies how much more accurately the model cites English documents compared to the target language, with all other documents fixed to English. To ensure differences in raw scores are statistically meaningful, we perform pairwise two-sided $t$-tests and apply a Bonferroni correction to account for multiple comparisons.

## 4 Experiment Setup

**Dataset.** We use ELI5 dataset (Fan et al., 2019) of long-form questions from the Reddit forum "Explain Like I'm Five". For each query, we adopt the WebGPT test set (Nakano et al., 2022) (270 queries), with relevant evidence documents collected by human annotators using Bing. To successfully answer a query, the generated output must cite all provided relevant documents. To ensure the citation IDs are tokenized as single tokens across all evaluated models, we only use queries with $K < 10$ evidence documents. Detailed dataset statistics are in Appendix Table 2.

**Languages.** For $\mathcal{L}_{\text{target}}$, we study eight languages representing a diverse range of resource levels (measured by number of speakers and Wikipedia articles), language families, scripts, linguistic typologies: Arabic (ar), Bengali (bn), Spanish (es), French (fr), Korean (ko), Russian (ru), Swahili (sw), and Chinese (zh). Detailed characteristics per language are outlined in Appendix Table 3.

**Models.** We use six open-weight LLMs that provide full-access to model weights and support large enough context windows to handle long-context evidence documents and long-form generations. To assess the generality of language preference, we evaluate models varying in size, degree of multilinguality, and architecture family: LLaMA-3.1 8B and LLaMA-3.3 70B (Grattafiori et al., 2024), Qwen-3 8B and 14B (Yang et al., 2025), Gemma-3 27B (Team et al., 2025), and Aya23 8B (Aryabumi et al., 2024). Details for each model can be found in Appendix Table 4.

## 5 Evidence of an English Language Preference

We seek to understand whether models prefer citing evidence documents in English over other languages in long-form mRAG. To do so, we analyze language preference in a controlled setup where all provided evidence documents are relevant to the query. We begin by comparing citation accuracies across languages, then explore factors that may impact language preference (§5.1). Next, we perform a layer-wise analysis of model behavior to unfold how language preference evolves (§5.2).

### 5.1 Do Models Preferentially Cite English Documents?

We define a model exhibits language preference for citing English evidence over the target language if its citation accuracy is higher for English ($\Delta(\ell_{\text{target}}) < 0$ in Eq. 3). Table 1 presents citation accuracies by model and language. Overall, we see a consistent English preference across all tested models and target languages.[11] **Even models explicitly trained on diverse languages and multilingual tasks, such as Aya23 8B, display this preference.** In Appendix D.3, we further show that, for all models, the next token probability of the correct citation ID is the highest—and both the

[11] In Appendix D.2, our embedding-similarity analysis shows that English preference cannot be fully explained by semantic similarity between the query and the cited document *alone*.

| Language | LLAMA-3.1 8B | QWEN-3 8B | AYA23 8B | QWEN-3 14B | GEMMA-3 27B | LLAMA-3.3 70B |
|---|---|---|---|---|---|---|
| **English** | 67.4 | 62.6 | 60.0 | 83.0 | 86.2 | 85.9 |
| **French** | 62.9 (-4.49) | 48.4 (-14.2)*** | 48.5 (-11.5)*** | 76.0 (-7.04)*** | 79.0 (-7.21)** | 77.4 (-8.50)*** |
| **Russian** | 62.1 (-5.30)* | 50.4 (-12.2)*** | 48.1 (-11.9)*** | 74.8 (-8.17)*** | 77.1 (-9.12)*** | 74.5 (-11.4)*** |
| **Spanish** | 62.1 (-5.32)* | 51.9 (-10.7)*** | 49.1 (-10.9)*** | 77.4 (-5.61)* | 80.2 (-6.04)** | 76.0 (-9.90)*** |
| **Korean** | 61.7 (-5.68)* | 49.7 (-12.9)*** | 42.2 (-17.8)*** | 70.3 (-12.7)*** | 77.5 (-8.71)*** | 69.2 (-16.7)*** |
| **Chinese** | 59.9 (-7.51)* | 49.2 (-13.4)*** | 46.3 (-13.7)*** | 73.5 (-9.49)*** | 75.4 (-10.8)*** | 74.1 (-11.8)*** |
| **Arabic** | 59.5 (-7.91)** | 47.6 (-15.0)*** | 43.2 (-16.8)*** | 72.6 (-10.4)*** | 78.4 (-7.82)*** | 67.3 (-18.6)*** |
| **Bengali** | 56.6 (-10.8)*** | 41.3 (-21.3)*** | 27.2 (-32.8)*** | 65.4 (-17.6)*** | 77.9 (-8.33)*** | 68.8 (-17.1)*** |
| **Swahili** | 53.0 (-14.4)*** | 30.4 (-32.2)*** | 22.4 (-37.6)*** | 54.7 (-28.3)*** | 74.0 (-12.2)*** | 67.3 (-18.6)*** |

Table 1: **Citation accuracies (%) by model and language.** We present mean accuracy values $\mathbf{Acc}^{(\ell)}$ with $\Delta(\ell_{\text{target}})$ in subscript. Pairwise two-sided $t$-tests are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. Bonferroni correction is applied for multiple comparisons. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant. Color coding indicates the magnitude of $\Delta(\ell_{\text{target}})$: largest, second largest, others. Columns: increasing model size; rows: decreasing $\Delta(\ell_{\text{target}})$ (of first model). All models consistently show English preference.

Shannon entropy and perplexity of the next token distribution is the lowest—when the cited document is in English, indicating models are not only more accurate but also more confident in their correct predictions for English. We also find that smaller models (8B) have lower English baseline accuracy than larger models (*e.g.,* LLAMA-3.1 70B, GEMMA-3 27B), suggesting that models' general ability to correctly cite English evidence documents tends to improve with model scale.

**Stronger English Preference over Lower-resource Languages.** Having established an overall preference for citing English documents, we next examine which factors amplify this preference. Using the $\Delta(\ell_{\text{target}})$ values from Table 1 (*i.e.,* the drop in citation accuracy relative to English), we find a clear correlation with language resource level: lower-resource languages exhibit largest accuracy decreases. For example, Swahili shows the greatest drop (-23.9% on average, up to -37.6% in AYA23 8B), followed by Bengali (-18.0% on average, up to -32.8% in AYA23 8B), even for models that officially support these languages (QWEN-3 8B, 14B, GEMMA-3 27B; Appendix Table 4). In contrast, higher-resource languages such as Spanish and French show smaller decreases (-8.08% and -8.82% on average, respectively), indicating weaker English preference.

**Position Bias Amplifies Language Preference.** We find that the relative position of an evidence document within the input context impacts citation accuracy. Figure 2 (*left*) shows English citation accuracy binned by the relative position of the cited document: at the beginning (First), the end (Last), or elsewhere (Middle) in the input context. Accuracy is generally lowest when the document appears in the middle (one exception is LLAMA-3 70B, which shows the lowest accuracy for the Last position). This aligns with the "lost in the middle" phenomenon, where LLMs struggle to access and use information in the middle of long contexts (Liu et al., 2024), here demonstrated for citation generation. Figure 2 (*right*) presents the difference in accuracy between English and the average of target languages across these positions.[12] For all models, the largest drop in accuracy occurs when the cited document is positioned in the middle, indicating that document position not only impacts English accuracy but also amplifies models' English preference.

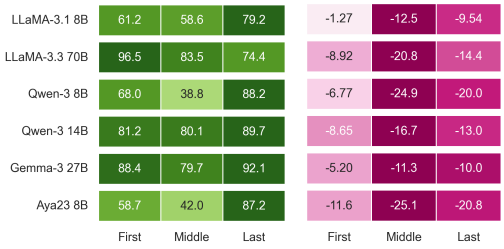



Figure 2: **English accuracy (*left*) and the average of $\Delta(\ell_{\text{target}})$ (*right*) (%) binned by relative position.** Each bin is normalized by sample size. $\Delta(\ell_{\text{target}})$ is largest when the cited document is positioned in the middle, indicating that position bias further amplifies English preference.

In sum, we provide strong evidence that models preferentially cite English evidence documents over target languages. This finding holds not only for *corroborative* attribution, which identifies sources that support a statement, but also for *contributive* attribution, which captures sources that causally

[12]Results for each target language can be found in Appendix D.4.

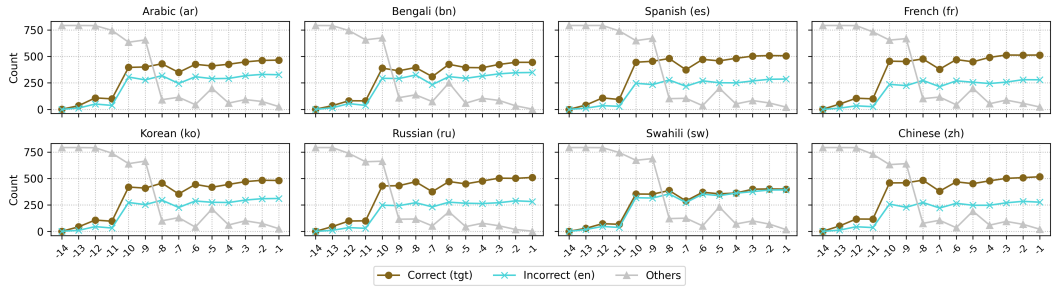



Figure 3: **Logit lens visualization per language for LLAMA-3.1 8B (32 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. ●: Correct citation ID of document in target language; ×: Incorrect citation ID of document in English; ▲: Not in valid citation set. Model makes a specific decision point when selecting which document to cite and largely preserves this choice across later layers. We only show last 14 layers. Results for other models are in Appendix D.5.

influence the model's generation, showing consistent trends (Appendix E). We further identify two key factors that amplify this preference: the resource level of the language and the position of the document within the input context.[13]

### 5.2 MODEL LAYER-WISE ANALYSIS

While our earlier results confirm a strong English preference in citation, we still lack a deeper understanding on *how* this preference unfolds during generation. Does the model settle on its initial choice and persist with it or does it initially favor English documents before shifting toward the correct target language citation? This question extends prior findings from short-form tasks, where multilingual LLMs often align their internal representations with English in early layers, transitioning to target language-specific spaces only in the final layers (Wendler et al., 2024; Zhong et al., 2024; Wang et al., 2025a; Bafna et al., 2025; Schut et al., 2025). We ask whether citation generation in long-form setup follows a similar trajectory: do models initially gravitate toward citing English documents and only later correct themselves, or is the outcome largely decided as soon as the model chooses which document to cite?

To probe this, we employ logit lens (Nostalgebraist, 2020), which maps intermediate state representations of LLMs into the vocabulary space, enabling the ability to track a model's token prediction across layers. Since logit lens is tailored to probe a single token, our citation format is a single digit, and this approach works well for this use case. For each statement, we check whether the top-1 token prediction at a given layer is (1) the correct citation ID $c_i$ (target language document), (2) an incorrect ID $c_j$ ($j \neq i$, English document), or (3) not a valid citation token ($\notin \{1, \ldots, K\}$, Others).

Figure 3 shows results for LLAMA-3.1 8B. Across all languages, layers 1-17 yield no valid predictions, indicating that the model has not yet figured out the expected output format. Around layers 18-20, both correct and incorrect citation IDs begin to appear, with correct IDs slightly more frequent. Layer 22 marks a sharp peak for both correct and incorrect predictions, suggesting this is the stage where the model settles on the output format and decides which document to cite. From layer 23 onward, incorrect IDs remain at a stable rate, showing that once the model commits to an incorrect citation, it rarely changes. Meanwhile, count for correct IDs steadily increases, replacing the earlier invalid predictions (Others). We also find that the gap between correct and incorrect predictions narrows notably for lower-resource languages (*i.e.,* Bengali, Swahili), confirming our earlier findings that these languages exhibit a stronger English preference. Results for other models are provided in Appendix D.5.

Overall, these results indicate that models do not initially favor citing incorrect English documents and then switch to the correct target language. Instead, there is a specific decision point (around

[13] We further show that our findings remain robust to both—(1) stylistic variations in the citation ID (*e.g.,* IDs expressed in different languages or formats; Appendix G) and (2) language variants in the non-cited evidence documents (Appendix F).

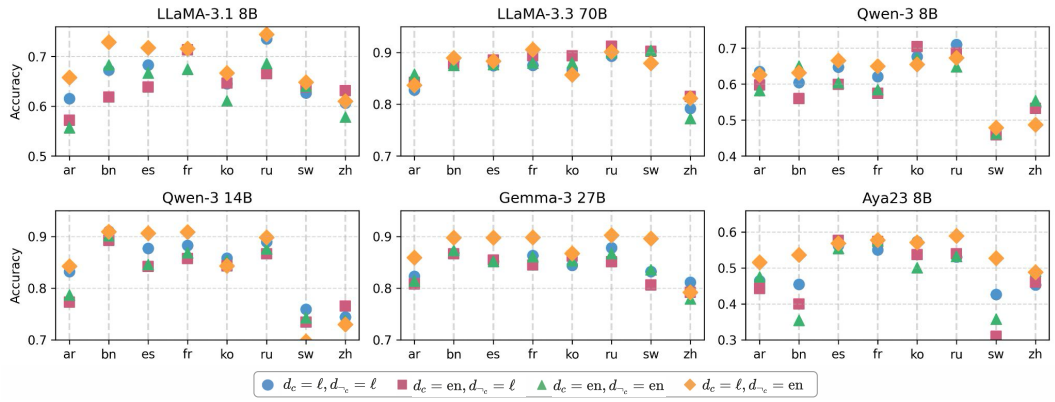



Figure 4: **Accuracy per model for queries in the target language.** ●: $d_c = \ell, d_{\neg c} = \ell$; ■: $d_c = \text{en}, d_{\neg c} = \ell$; ▲: $d_c = \text{en}, d_{\neg c} = \text{en}$; ◆: $d_c = \ell, d_{\neg c} = \text{en}$. Note that $y$-axis scale vary by model. $x$-axis denotes each target language. Models generally show query language preference.

layer 22 for LLAMA-3.1 8B), when they decide which document to cite. From that point on, they largely preserve their initial decision, whether it is correct or not.

## 6 Effect of the Query Language

Our previous analysis demonstrate that models preferentially cite English evidence documents over those in the target language. A natural follow-up question is whether this pattern persists when the query itself is in a language other than English: do models still prefer English documents, or do they prefer documents in the same language as the query?

**Setting.** We follow the same procedure used to measure English preference (Section 3), with one modification in Step 2 (reference report generation). Each user query is translated into the target language $q_{tgt}$, and for each, we generate a reference citation-supported report $r_{tgt}$ using $K$ relevant evidence document translations $\mathcal{D}_{tgt}$.[14] For Step 4 (next token prediction analysis), we consider four context variants differing in the language of the cited document $d_c$ and the remaining evidence documents $d_{\neg c}$: (1) Both $d_c$ and $d_{\neg c}$ in the query language ($\ell$) (●); (2) $d_c$ in English, $d_{\neg c}$ in $\ell$ (■); (3) Both $d_c$ and $d_{\neg c}$ in English (▲); (4) $d_c$ in $\ell$, $d_{\neg c}$ in English (◆). Higher citation accuracy for variants ● and ◆ compared to ■ and ▲ indicates that the model prefers citing documents in the query language. Conversely, higher accuracy for ■ and ▲ suggests a persistent English preference regardless of the query language.

**Results.** We report citation accuracies for the four variants in Figure 4, broken down by target language for each model. Across more than half of the model-language combinations (28 out of 48), we observe the highest citation accuracy when the cited document is in the query language and all other documents are in English (◆). In 17 of these 28 cases, the second-best performance is when all documents are in the query language (●). Since the ◆ configuration generally outperforms the ● variant, this suggest that models benefit from a language contrast between the cited and the remaining documents rather than simply having more documents match the query language. French follows this trend most strongly, with 4 out of 6 models exhibiting it. One possible explanation is that, as a relatively high-resource language, models have strong enough French representations, allowing them to effectively leverage the contrast and identify the most relevant document in context.

We further see that smaller models (8B) generally achieve lower accuracies than larger models (*e.g.,* LLAMA-3.3 70B, GEMMA-3 27B), extending our earlier observation from Section 5.1 that model size improves citation accuracy for English to non-English settings as well. Larger models also exhibit citation accuracies that are more tightly clustered across the four variants, suggesting greater robustness to language variation in the input context.

[14] We use Google Translate API for query translation, with translation quality reported in Appendix Table 5.

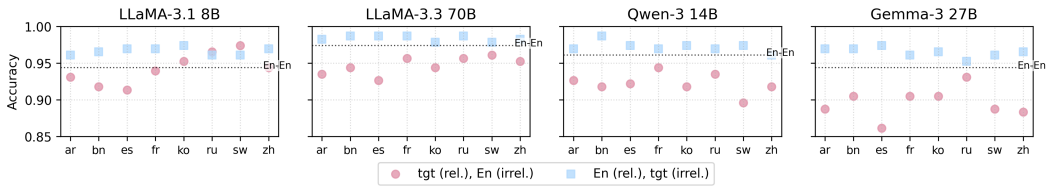



Figure 5: **Accuracy per model with one relevant and one irrelevant evidence document in different languages.** ●: Relevant doc in target language, irrelevant doc in English; ■: Relevant doc in English, irrelevant doc in target language; ·····: Baseline, both docs in English. Models trade off document relevance for language preference. Detailed results are in Appendix Figure 5.

Together, these results suggest that query language plays a key role in models' language preference: models tend to favor citing documents in the same language as the query, even when that language is not English. Interestingly, this mirrors findings in scientometrics literature, where humans also exhibit an "own-language preference", tending to select and cite sources in the language of their writing (Yitzhaki, 1998; EGGHE et al., 2005). Detailed numerical results are in Appendix D.6.

## 7 Relevance vs. Language Preference

Sections 5 and 6 analyzed language preference in a controlled setup where all provided evidence documents were relevant to the query. In reality, however, retrievers are imperfect, and retrieved evidence often contains irrelevant or partially relevant documents (Chen et al., 2023; Jin et al., 2024). To better approximate such conditions, we relax the assumption that all documents are relevant and ask: between relevance and language, which exerts a stronger influence on model citation behavior?

**Setting.** We compare the effects of document relevance and language by varying the language of one relevant and one irrelevant document under three conditions:[15] (1) **En-En** (····): Both relevant and irrelevant documents are in English; (2) **tgt-En** (●): Relevant document in the target language, irrelevant document in English; (3) **En-tgt** (■): Relevant document in English, irrelevant document in the target language. Since ELI5 dataset does not include irrelevant documents, we use MIRACL (Zhang et al., 2023), a multilingual RAG dataset with Wikipedia queries. We use the English subset of the development set, restricting to queries with exactly one relevant document (231 queries). We randomly use one of the irrelevant documents.[16] For each query, we follow the same process described in Section 3. Dataset statistics are in Appendix Table 2.

**Results.** Our hypotheses are: (a) If citation accuracy in **tgt-En** is lower than the **En-En** baseline, it suggests the model is overly influenced by language, preferring to cite an irrelevant English document over a relevant target language one, and (b) If citation accuracy in **En-tgt** exceeds the baseline, it implies the model more easily ignores irrelevant target language distractors, again signaling English preference. As shown in Figure 5, results support both hypotheses. When the relevant document is in the target language, accuracies consistently drop below the baseline, indicating that irrelevant English content more easily mislead the model. Conversely, accuracies for all languages and models rise above the baseline for **En-tgt**, suggesting that target language distractors are easier to dismiss than English distractors. This aligns with recent findings that distractors in the same language as the relevant document degrade performance more severely (Qi et al., 2025). One interesting observation is Swahili. Despite yielding the lowest accuracies in ELI5 experiments (see Table 1), its performance in the **En-tgt** setup is relatively high. A possible explanation is its shared Latin script

[15] We use this setup to ensure that the model's decision can vary only along these two dimensions.

[16] MIRACL simulates a realistic retrieval setting since the irrelevant documents are constructed by (1) using retrievers to gather candidate passages from the query and a Wikipedia dump, and (2) selecting those labeled "irrelevant" by human annotators (Zhang et al., 2023). Therefore, the irrelevant documents are often topically related to the query but not necessary for answering it—simulating realistic retrieval noise.

with English, which may make irrelevant Swahili documents appear more plausible choice. Full numerical results can be found in Appendix D.7.[17]

## 8 Conclusion

We propose a controlled methodology to measure language preference in long-form mRAG by isolating language effects while controlling for document content and relevance. Our analysis shows that models preferentially cite English documents when queries are in English, with this bias stronger for lower-resource languages and mid-context. Importantly, this preference can outweigh relevance, with models often citing irrelevant English documents over relevant non-English ones. Overall, our findings demonstrate how model internals reveal citation behavior in mRAG and offer insights for designing more robust, inclusive systems that balance language and relevance.

**Limitations.** The dataset used in our main experiments, ELI5 (Fan et al., 2019), has known limitations—such as substantial train-validation overlap and answers that are not often grounded in the supporting documents (Krishna et al., 2021). However, ELI5 was the *only* publicly available dataset that met the requirements of our experiment setup: it provides (1) knowledge-extensive queries which need long-form, citation-supported answers to respond, and (2) curated evidence documents that are *all* required to answer the query. To complement our main results, we additionally run experiments on MIRACL (Zhang et al., 2023) in Appendix J, and observe the same English (§5) and query-language preference (§6).

Our analysis uses a controlled setup with several simplifying assumptions—(1) retrieval is complete and all evidence documents are equally relevant; (2) comparisons are only pairwise with English; and (3) multilingual RAG is simulated via machine translations (MT) of English documents since no parallel, long-form mRAG datasets are publicly available. To complement our use of MT, we show that (a) English preference does not meaningfully correlate with MT quality (Appendix I, (b) our findings remain consistent when using naturally occurring queries in the target language (Appendix J.2), and (c) with an alternative translation system (Appendix K). These assumptions may not fully hold in real-world settings, which could limit the generalizability of our results. Even so, our study provides valuable insights into language preference that can guide future work on understanding and improving model citation behavior.

## Ethics Statement

We recognize that research on language preference in multilingual retrieval-augmented generation can raise issues of fairness, particularly for low-resource and underrepresented languages. Our study aims to highlight such disparities rather than reinforce them. We caution that models may amplify English preference in multilingual contexts, with potential implications for inclusivity and equitable knowledge access.

## Reproducibility Statement

All datasets used in our experiments are publicly available (ELI5, MIRACL) and are described in Section 4 and Appendix B, including dataset statistics, splits, and preprocessing steps. Our methodology for constructing multilingual evidence sets, generating English reference reports, and verifying citation correctness is documented in Section 3.1, with prompt templates provided in Appendix A. We report full per-language translation quality scores (Appendix D.1) and include statistical tests to assess the significance of results (Section 3.2, 5.1, and Appendix D). To support reproducibility, we will release our code and processed data upon publication, enabling others to replicate both our measurement pipeline and analyses.

[17] We further show that models also trade-off relevance for language preference when queries are posed in a language other than English in Appendix H.

## REFERENCES


Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks, 2024.

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan. The Cross-Lingual Cost: Retrieval Biases in RAG over Arabic-English Corpora, 2025. URL https://arxiv.org/abs/2507.07543.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open Weight Releases to Further Multilingual Progress, 2024. URL https://arxiv.org/abs/2405.15032.

Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H Clark, and Eunsol Choi. MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages. *arXiv preprint arXiv:2207.00758*, 2022.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, 2024.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. The Translation Barrier Hypothesis: Multilingual Generation with Large Language Models Suffers from Implicit Translation Failure, 2025. URL https://arxiv.org/abs/2506.22724.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. Multilingual Hallucination Gaps in Large Language Models, 2024. URL https://arxiv.org/abs/2410.18270.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. Understanding Retrieval Augmentation for Long-Form Question Answering, 2023. URL https://arxiv.org/abs/2310.12150.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. Retrieval-augmented generation in multilingual settings. In Sha Li, Manling Li, Michael JQ Zhang, Eunsol Choi, Mor Geva, Peter Hase, and Heng Ji (eds.), *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 177–188, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.knowllm-1.15. URL https://aclanthology.org/2024.knowllm-1.15/.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing Model Generation to Context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7CMNSqsZJt.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. Efficiently Exploring Large Language Models for Document-Level Machine Translation with In-context Learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10885–10897, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.646. URL https://aclanthology.org/2024.findings-acl.646/.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *International Conference on Learning Representations (ICLR)*, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*, 2021.

Leo EGGHE, Ronald Rousseau, and M. Yitzhaki. The "own-language preference": Measures of relative language self-citation. *Scientometrics*, 45, 05 2005.

Assaf Elovic. gpt-researcher, July 2023. URL https://github.com/assafelovic/gpt-researcher.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Nikolaos Aletras and Orphee De Clercq (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long Form Question Answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling Large Language Models to Generate Text with Citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2312.10997.

Charles Godfrey, Ping Nie, Natalia Ostapuk, David Ken, Shang Gao, and Souheil Inati. Likert or Not: LLM Absolute Relevance Judgments on Fine-Grained Ordinal Scales, 2025. URL https://arxiv.org/abs/2505.19334.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid

El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,

Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025a.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. Deep Research Agents: A Systematic Examination And Roadmap, 2025b. URL https://arxiv.org/abs/2506.18096.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. *arXiv preprint arXiv:2410.05983*, 2024.

Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. Controlled Retrieval-augmented Context Evaluation for Long-form RAG, 2025. URL https://arxiv.org/abs/2506.20051.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. GooAQ: Open question answering with diverse answer types. *arXiv preprint arXiv:2104.08727*, 2021.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4940–4957, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.393. URL https://aclanthology.org/2021.naacl-main.393/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the*

*34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks: A Benchmark for Cross-lingual Robustness. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4215–4241, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.219. URL https://aclanthology.org/2025.findings-acl.219/.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating Verifiability in Generative Search Engines. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. XRAG: Cross-lingual Retrieval-Augmented Generation, 2025. URL https://arxiv.org/abs/2505.10089.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-Tuning LLaMA for Multi-Stage Text Retrieval, 2023. URL https://arxiv.org/abs/2310.08319.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, 2024. URL https://arxiv.org/abs/2405.20362.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, 2022. URL https://arxiv.org/abs/2203.11147.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

Nostalgebraist. Interpreting GPT: The Logit Lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru, 2020. Accessed: 2025-08-13.

Jeonghyun Park and Hwanhee Lee. Investigating Language Preference of Multilingual RAG Systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5647–5675, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.295. URL https://aclanthology.org/2025.findings-acl.295/.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1314–1324, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1119. URL https://aclanthology.org/N18-1119/.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. On the Consistency of Multilingual Context Utilization in Retrieval-Augmented Generation, 2025. URL https://arxiv.org/abs/2504.00597.

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. $LONG^2RAG$: Evaluating Long-Context & Long-Form Retrieval-Augmented Generation with Key Point Recall. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4852–4872, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.279. URL https://aclanthology.org/2024.findings-emnlp.279/.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task, 2025a. URL https://arxiv.org/abs/2504.03616.

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. Improving Multilingual Retrieval-Augmented Language Models through Dialectic Reasoning Argumentations, 2025b. URL https://arxiv.org/abs/2504.04771.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do Multilingual LLMs Think In English?, 2025. URL https://arxiv.org/abs/2502.15603.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. Faux Polyglot: A Study on Information Disparity in Multilingual Large Language Models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8090–8107, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.411. URL https://aclanthology.org/2025.naacl-long.411/.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of llms in multilingual context. In - *Findings*, 2024. URL https://arxiv.org/abs/2401.13136.

Vaishnavi Shrivastava, Ananya Kumar, and Percy Liang. Language Models Prefer What They Know: Relative Confidence Estimation via Confidence Preferences, 02 2025.

Statista. Most common languages on the internet, 2025. URL https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/. Accessed: 2025-08-05.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid Questions Meet Long-Form Answers, 2023. URL https://arxiv.org/abs/2204.06092.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report, 2025. URL https://arxiv.org/abs/2503.19786.

Joseph P. Telemala and Hussein Suleman. Language-Preference-Based Re-ranking for Multilingual Swahili Information Retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, pp. 144–152, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394123. doi: 10.1145/3539813.3545131. URL https://doi.org/10.1145/3539813.3545131.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-Level Machine Translation with Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16646–16661, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1036. URL https://aclanthology.org/2023.emnlp-main.1036/.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer

Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.253. URL https://aclanthology.org/2025.acl-long.253/.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. DelTA: An Online Document-Level Translation Agent Based on Multi-Level Memory. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=hoYFLRNbhc.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4M9f8VMt2C.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. Not All Languages are Equal: Insights into Multilingual Retrieval-Augmented Generation, 2024. URL https://arxiv.org/abs/2410.21970.

Renjun Xu and Jingwen Peng. A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications, 2025. URL https://arxiv.org/abs/2506.12594.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. Language Fairness in Multilingual Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2487–2491, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657943. URL https://doi.org/10.1145/3626772.3657943.

M. Yitzhaki. The 'Language Preference' in Sociology: Measures of 'Language Self-Citation', 'Relative Own-Language Preference Indicator', and 'Mutual Use of Languages'. *Scientometrics*, 41 (1):243–254, January 1998. ISSN 1588-2861. doi: 10.1007/BF02457981.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA, 2024. URL https://arxiv.org/abs/2409.02897.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. "LongCite: Enabling LLMs to generate fine-grained citations in long-context QA". In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://aclanthology.org/2023.tacl-1.63/.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1259. URL https://aclanthology.org/2024.emnlp-main.1259/.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. SciArena: An Open Evaluation Platform for Foundation Models in Scientific Literature Tasks, 2025. URL https://arxiv.org/abs/2507.01001.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments, 2025. URL https://arxiv.org/abs/2504.03160.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in?, 2024. URL https://arxiv.org/abs/2408.10811.

# APPENDIX

## A PROMPTS

We present the prompts used for generating the gold citation-supported report (Figure 6), obtaining supportedness judgments from LLM-as-judge (Figure 7), and guessing the next token predictions from the evaluated models (Figure 8). We adopt base prompts from GPTResearcher (Elovic, 2023).

**Prompt A.1. Gold Report Generation Prompt**

**Information:**
**Document ID:** {`document ID`}
**Title:** {`title`}
**Content:** {`content`}
—
...
—
Using the above information, respond to the following query or task: {`query`}.
The response should focus on the answer to the query, should be well structured, informative, and concise, with facts and numbers if available.

Please follow all of the following guidelines in your response:
- You MUST write in a single paragraph and at most {`total words`} words.
- You MUST write the response in the following language: {`language`}.
- You MUST cite your sources, especially for relevant sentences that answer the question.
- When using information that comes from the documents, use citation which refer to the Document ID at the end of the sentence (e.g., [1]).
- Do NOT cite multiple documents at the end of the sentence (e.g., [1][2]).
- If multiple documents support the sentence, only cite the most relevant document.
- It is important to ensure that the Document ID is a valid string from the information above and that the information in the sentence is present in the document.

**Response:**

Figure 6: **Prompt for generating gold citation-supported reports.** Information section is populated with the document ID, title, and content of each evidence document. Boldface is only for emphasis.

**Prompt A.2. LLM-as-judge Prompt**

**Instruction:** You are given a query, a document, and a sentence from a generated response that cites the document in answering the query. Determine which document best supports the information in the cited sentence. Respond only with the exact document ID. Do not provide any additional explanation.

**Query:** {`query`}
**Information:**
**Document ID:** {`document ID`}
**Title:** {`title`}
**Content:** {`content`}
—
...
—

**Cited sentence:** {`statement`}
**Response:**

Figure 7: **Prompt for getting supportedness judgments from LLM-as-judge.** Information section is populated with the document ID, title, and content of each evidence document. Boldface is only for emphasis.

Prompt A.3. Next Token Prediction Prompt

**Information:**
**Document ID:** {document ID}
**Title:** {title}
**Content:** {content}
—
...
—
Using the above information, the response is the answer to the query or task: {query} in a single sentence.
You MUST cite the most relevant document by including only its Document ID in brackets at the end of the sentence (e.g., [Document ID]).
Do NOT include any additional words inside or outside the brackets.
Please output ONLY the number of the Document ID that is most relevant to the sentence.

**Response:** {statement} [

Figure 8: **Prompt for guessing the next token prediction.** Information section is populated with the document ID, title, and content of each evidence document. Boldface is only for emphasis.

## B Details of Dataset, Languages, and Models

We provide detailed statistics of the two long-form RAG datasets used in our experiments (ELI5 and MIRACL) in Table 2. The characteristics of the eight tested languages, including their language family, script, linguistic typology, and resource level, are summarized in Table 3. For the models, Table 4 includes their context window size, HuggingFace model identifier, and officially (un)supported languages. Lastly, Table 5 reports COMET-QE (Rei et al., 2020) scores for each target language.

| Dataset | # Queries | Avg. # Words ($q$) | Avg. # Words ($t$) | Avg. # Words ($d$) | Avg. # Sent ($d$) | Avg. # $d$ per $q$ |
|---|---|---|---|---|---|---|
| **ELI5** | 270 | 15.25 | 9.64 | 76.82 | 4.26 | 3.49 |
| **MIRACL** | 231 | 6.87 | 2.63 / 2.83 | 106.59 / 115.80 | 5.41 / 5.88 | 1.00 / 9.31 |

Table 2: **Detailed statistics of long-form RAG datasets used.** We report statistics for ELI5 (Explain Like I'm Five) and MIRACL. For MIRACL, statistics are shown as relevant / irrelevant documents. $q$: query; $t$: title, $d$: evidence document.

| Language Family | Language | Script | Synthesis | Word Order | Resource Level | # Speakers | # Wikipedia Size |
|---|---|---|---|---|---|---|---|
| Indo-European | English | Latin | analytic | SVO | high | 1,130M | 5,758,285 |
| | French | Latin | fusional | SVO | high | 398M | 2,325,608 |
| | Spanish | Latin | fusional | SVO | high | 592M | 1,669,181 |
| | Russian | Cyrillic | fusional | SVO | mid | 260M | 1,476,045 |
| | Bengali | Bengali | fusional | SOV | low | 337M | 63,762 |
| Sino-Tibetan | Chinese | Chinese | analytic | SVO | high | 1,350M | 1,246,389 |
| Koreanic | Korean | Hangul | agglutinative | SOV | mid | 128M | 1,133,444 |
| Afro-Asiatic | Arabic | Arabic | fusional | VSO | mid | 630M | 656,982 |
| Niger-Congo | Swahili | Latin | agglutinative | SVO | low | 83M | 47,793 |

Table 3: **Characteristics of tested languages.** For each language, we show language family, script, linguistic typologies (synthesis and word order), and resource level measured by the number of speakers and Wikipedia articles (Zhang et al., 2023).

| Model | Context Window | HuggingFace Model Identifier | Supported Langs | Unsupported Langs |
|---|---|---|---|---|
| **LLAMA-3 8B** | 128K | `meta-llama/Llama-3.1-8B-Instruct` | en, es, fr | ar, bn, ru, ko, sw, zh |
| **LLAMA-3 70B** | 128K | `meta-llama/Llama-3.3-70B-Instruct` | en, es, fr | ar, bn, ru, ko, sw, zh |
| **QWEN-3 8B** | 33K | `Qwen/Qwen3-8B` | en, ar, bn, es, fr, ru, ko, sw, zh | - |
| **QWEN-3 14B** | 33K | `Qwen/Qwen3-14B` | en, ar, bn, es, fr, ru, ko, sw, zh | - |
| **GEMMA-2 27B** | 128K | `google/gemma-3-27b-it` | en, ar, bn, es, fr, ru, ko, sw, zh | - |
| **AYA23 8B** | 8,192 | `CohereLabs/aya-23-8B` | en, ar, es, fr, ru, ko, zh | bn, sw |

Table 4: **List of evaluated models.** We report the context window size, HuggingFace model identifiers, and the *officially* supported languages during pretraining. Note: Supported language information is extracted from each model's technical report. We use ISO 639-1 codes for languages. We use QWEN-3 series models with `enable thinking=False` mode.

| Language | COMET-QE($q$, $q'$) | COMET-QE($t$, $t'$) | COMET-QE($d$, $d'$) |
|---|---|---|---|
| **Arabic** (ar) | 0.752 | 0.541 | 0.511 |
| **Bengali** (bn) | 0.824 | 0.584 | 0.559 |
| **Spanish** (es) | 0.823 | 0.583 | 0.564 |
| **French** (fr) | 0.822 | 0.582 | 0.566 |
| **Korean** (ko) | 0.816 | 0.584 | 0.555 |
| **Russian** (ru) | 0.780 | 0.557 | 0.528 |
| **Swahili** (sw) | 0.769 | 0.544 | 0.516 |
| **Chinese** (zh) | 0.777 | 0.561 | 0.534 |

Table 5: **COMET-QE scores by language.** We evaluate the machine translation (MT) quality of non-English queries ($q$), titles ($t$), and evidence documents ($d$) in the ELI5 dataset. Apostrophe ($'$) indicates MT. Higher scores indicate better MT quality.

## C HUMAN ANNOTATION

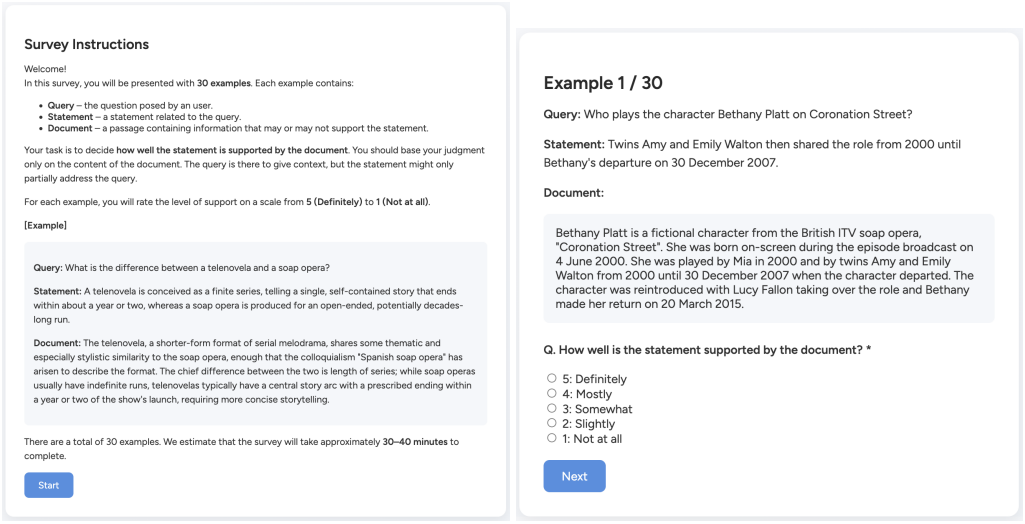
To validate the two-step automatic filtering process described in Section 3 for identifying supported statements, we conduct a small-scale human annotation study on 60 sampled statements. We stratify the sample into 30 "supported" statements (passing both the LLM-as-Judge and NLI entailment filters and included in the final statement pool) and 30 "unsupported" statements (failing one or both filters). We conducted a power analysis to justify our sample size. Using a $t$-test for 2 independent samples[18], we find that 26 statements per label group (supported and unsupported, total 52) are required to detect a minimum effect size of Cohen's $d$ of 0.8 with a significance level of $\alpha$ of 0.05, and desired power of 0.8.

For each query $q$, statement $s_i$, and cited document $d_{c_i}$, we ask annotators: "How well is the statement supported by the provided document?" Responses are given on a five-point Likert scale from 5 (Definitely) to 1 (Not at all), using instructions similar to those provided when prompting the judge LLMs (Figure 7). Figure 9 shows the full instructions and an example provided to annotators.

We recruit six annotators from Prolific[19] who resides in the United States with first, primary, and fluent language as English. We compensate each with USD 8 (equivalent to USD 16/hour), totaling USD 56 including Prolific platform fees. Each annotator evaluates 30 statements (15 supported and 15 unsupported) presented in randomized order. Inter-annotator agreement is moderate, with a Krippendorff's alpha of 0.559. The average rating for supported statements is 4.15 out of 5, while unsupported statements average 2.49 out of 5. These results indicate strong alignment between our automatic filtering process and human judgments of statement supportedness. Figure 10 plots the rating distribution for each label group.

[18] `https://www.statsmodels.org/stable/generated/statsmodels.stats.power.TTestIndPower.html`

[19] `https://www.prolific.com/`




(a) Task Instructions (b) Example Annotation

Figure 9: **Full instructions and example provided to human annotators.** The annotation task was hosted on a custom-built website. Annotators first viewed a brief task instruction (a), then evaluate 30 statements, with an example shown in (b).

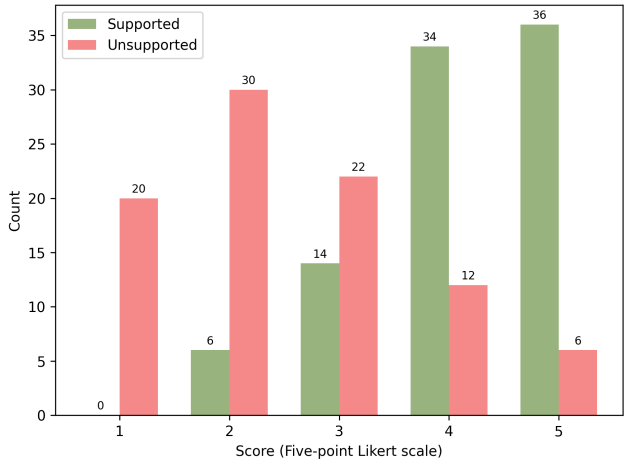
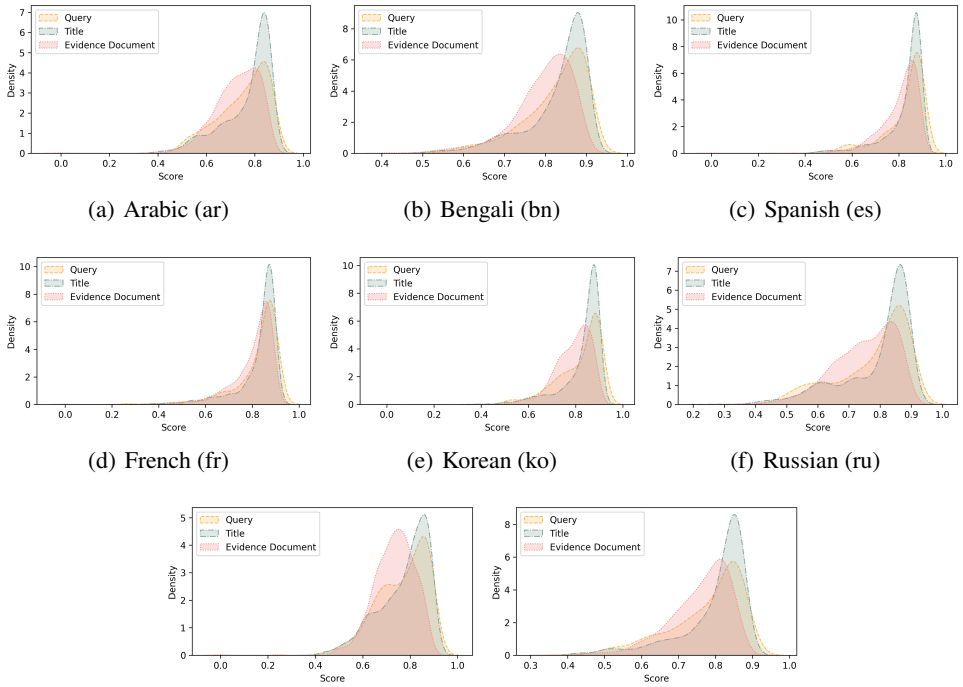



Figure 10: **Rating distribution for each label group.** We plot the distribution of 180 judgments collected during human annotation (90 supported and 90 unsupported statements). Results show that annotators can reliably distinguish supported from unsupported statements based on their ratings.




(a) Arabic (ar) (b) Bengali (bn) (c) Spanish (es)

(d) French (fr) (e) Korean (ko) (f) Russian (ru)

(g) Swahili (sw) (h) Chinese (zh)

Figure 11: **COMET-QE score distributions by language.** Distributions are more skewed for shorter content (*e.g.,* title), while broader distributions for longer content (*e.g.,* evidence document).

# D Detailed Results

## D.1 Machine Translation Quality

We evaluate machine translation (MT) quality for translated queries, titles, and evidence documents using COMET-QE scores. We do not perform any filtering based on these scores. Table 5 reports average scores by language, and Figure 11 shows full score distributions. We find little evidence that MT quality drives English preference. Document COMET-QE scores (last column of Table 5) are lowest for Arabic (0.511) and Swahili (0.516), while Bengali shows a relatively high score (0.559). Yet, citation accuracies (Table 1) show that Arabic's ranking varies widely across models—third lowest for LLaMA-3.1 8B and Qwen-3 8B, lowest for LLaMA-3.3 70B, fourth lowest for Qwen-3 14B and Aya23, but relatively higher for Gemma-3 27B. By contrast, Bengali exhibits the second-strongest English preference after Swahili despite its higher MT quality. This suggests that resource level, rather than MT quality, is a stronger indicator of English preference.

## D.2 Embedding Similarity Analysis

We compute embedding similarity between the query ($q$) and the cited document ($d_c$) using the multilingual encoder LaBSE (Feng et al., 2022). As shown in Table 6, when the query is in English ($q = \text{en}$), the embedding similarities show no statistically significant difference between cases where the cited document is in English vs. non-English languages ($\ell$) (columns 1-2). When the query is in a non-English language ($q = \ell$), we do observe higher similarity scores for cited documents in the same language as the query (columns 3-4). This suggests that English preference observed in Table 1 cannot be fully explained by semantic similarity alone.

| Language | $q = \text{en}, d_c = \text{en}$ | $q = \text{en}, d_c = \ell$ | $q = \ell, d_c = \text{en}$ | $q = \ell, d_c = \ell$ |
|---|---|---|---|---|
| **Arabic** | 0.579 | 0.584 | 0.601 | 0.653 |
| **Bengali** | 0.579 | 0.586 | 0.571 | 0.637 |
| **Spanish** | 0.579 | 0.586 | 0.575 | 0.645 |
| **French** | 0.579 | 0.589 | 0.574 | 0.648 |
| **Korean** | 0.579 | 0.577 | 0.573 | 0.654 |
| **Russian** | 0.579 | 0.589 | 0.571 | 0.654 |
| **Swahili** | 0.579 | 0.588 | 0.574 | 0.648 |
| **Chinese** | 0.579 | 0.586 | 0.573 | 0.651 |

Table 6: **Embedding similarity between query and cited document computed with LABSE.** $q$: query; $d_c$: cited document, $\ell$: target language.

### D.3 Evidence of English Preference

| Language | LLaMA-3.1 8B | LLaMA-3.3 70B | Qwen-3 8B | Qwen-3 14B | Gemma-3 27B | Aya23 8B |
|---|---|---|---|---|---|---|
| **English** | 0.651 | 0.991 | 0.758 | 0.984 | 0.980 | 0.527 |
| **Arabic** | 0.629 (-0.022) | 0.990 (-0.001) | 0.751 (-0.007) | 0.979 (-0.005) | 0.968 (-0.012) | 0.463 (-0.064) |
| **Bengali** | 0.647 (-0.004) | 0.990 (-0.001) | 0.736 (-0.022) | 0.981 (-0.003) | 0.977 (-0.003) | 0.442 (-0.085) |
| **Spanish** | 0.626 (-0.025) | 0.987 (-0.004) | 0.752 (-0.006) | 0.981 (-0.003) | 0.979 (-0.001) | 0.483 (-0.044) |
| **French** | 0.649 (-0.002) | 0.991 (0.000) | 0.728 (-0.030) | 0.983 (-0.001) | 0.973 (-0.007) | 0.499 (-0.028) |
| **Korean** | 0.620 (-0.031) | 0.982 (-0.009) | 0.730 (-0.028) | 0.983 (-0.001) | 0.955 (-0.025) | 0.494 (-0.033) |
| **Russian** | 0.634 (-0.017) | 0.990 (-0.001) | 0.707 (-0.051) | 0.982 (-0.002) | 0.961 (-0.019) | 0.465 (-0.062) |
| **Swahili** | 0.630 (-0.021) | 0.987 (-0.004) | 0.634 (-0.124) | 0.967 (-0.017) | 0.966 (-0.014) | 0.479 (-0.048) |
| **Chinese** | 0.642 (-0.009) | 0.988 (-0.003) | 0.706 (-0.052) | 0.984 (0.000) | 0.976 (-0.004) | 0.488 (-0.039) |

Table 7: **Next token probabilities for the correct citation ID by model and language (↑).** We present mean values along with the difference from English baseline indicated in subscript.

| Language | LLaMA-3.1 8B | LLaMA-3.3 70B | Qwen-3 8B | Qwen-3 14B | Gemma-3 27B | Aya23 8B |
|---|---|---|---|---|---|---|
| **English** | 1.106 | 0.132 | 0.388 | 0.064 | 0.028 | 1.215 |
| **Arabic** | 1.146 (+0.040) | 0.176 (+0.044) | 0.500 (+0.112) | 0.088 (+0.024) | 0.063 (+0.035) | 1.277 (+0.062) |
| **Bengali** | 1.169 (+0.063) | 0.178 (+0.046) | 0.457 (+0.069) | 0.095 (+0.031) | 0.051 (+0.023) | 1.350 (+0.135) |
| **Spanish** | 1.152 (+0.046) | 0.150 (+0.018) | 0.460 (+0.072) | 0.081 (+0.017) | 0.048 (+0.020) | 1.260 (+0.045) |
| **French** | 1.122 (+0.016) | 0.149 (+0.017) | 0.389 (+0.001) | 0.075 (+0.011) | 0.051 (+0.023) | 1.247 (+0.032) |
| **Korean** | 1.150 (+0.044) | 0.166 (+0.034) | 0.394 (+0.006) | 0.087 (+0.023) | 0.059 (+0.031) | 1.269 (+0.054) |
| **Russian** | 1.134 (+0.028) | 0.162 (+0.030) | 0.412 (+0.024) | 0.074 (+0.010) | 0.059 (+0.031) | 1.266 (+0.051) |
| **Swahili** | 1.194 (+0.088) | 0.182 (+0.050) | 0.508 (+0.120) | 0.123 (+0.059) | 0.054 (+0.026) | 1.254 (+0.039) |
| **Chinese** | 1.130 (+0.024) | 0.159 (+0.027) | 0.385 (+-0.003) | 0.084 (+0.020) | 0.067 (+0.039) | 1.255 (+0.040) |

Table 8: **Shannon entropy by model and language (↓).** We present mean values along with the difference from English baseline indicated in subscript.

| Language | LLaMA-3.1 8B | LLaMA-3.3 70B | Qwen-3 8B | Qwen-3 14B | Gemma-3 27B | Aya23 8B |
|---|---|---|---|---|---|---|
| **English** | 3.023 | 1.141 | 1.474 | 1.066 | 1.029 | 3.370 |
| **Arabic** | 3.147** | 1.193*** | 1.649*** | 1.092** | 1.065*** | 3.585*** |
| **Bengali** | 3.219* | 1.194*** | 1.579*** | 1.100*** | 1.052*** | 3.857*** |
| **Spanish** | 3.164*** | 1.162** | 1.584*** | 1.085** | 1.050*** | 3.526*** |
| **French** | 3.072 | 1.161** | 1.476 | 1.078 | 1.052*** | 3.481*** |
| **Korean** | 3.159** | 1.180*** | 1.483*** | 1.091** | 1.061*** | 3.556*** |
| **Russian** | 3.109* | 1.176*** | 1.51 | 1.077 | 1.061*** | 3.548*** |
| **Swahili** | 3.300*** | 1.200*** | 1.662*** | 1.131*** | 1.055*** | 3.506*** |
| **Chinese** | 3.097 | 1.172*** | 1.47 | 1.088** | 1.070*** | 3.509*** |

Table 9: **Perplexity values by model and language (↓).** We present mean values along with the difference from English baseline indicated in subscript. Pairwise two-sided $t$-tests are performed to compare perplexity between English and the target language, with the null hypothesis that the mean perplexity is equal across languages. Bonferroni correction is applied for multiple comparisons. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant.

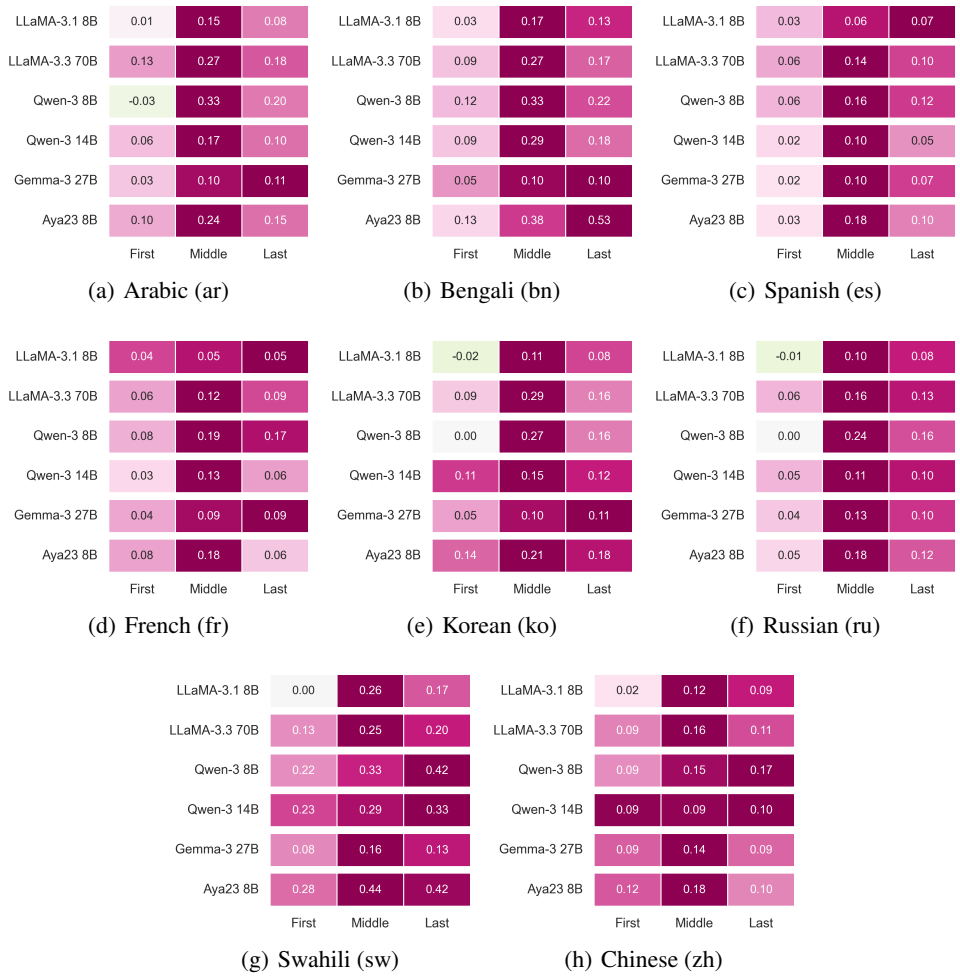



Figure 12: **Accuracy difference between English and each target language binned by relative position.** Each bin is normalized by sample size.

While our main accuracy metric is an intuitive measure, more fine-grained probability changes might not be captured. Therefore, for each model and language, we report the next token probability assigned to the correct citation ID (Table 7) and the Shannon entropy of the next token distribution (Table 8). Across all models, we observe consistently higher probabilities when the cited evidence document is in English, alongside lower entropy values.

Following SEPER (Dai et al., 2025), we further report the perplexity values in Table 9. We show that they are the lowest for English across all models except QWEN-3 8B, where Chinese is slightly lower, but the difference is not statistically significant. Together, this suggests that models are not only more accurate but also more confident when correctly citing English documents.

### D.4 POSITION-WISE ACCURACY PER LANGUAGE

We show accuracy gap between English and each target language in Figure 12. We show that the findings with the aggregated results in Section 5.1 are consistent for all languages: the accuracy drop is generally most pronounced when the cited document appears in the middle of the input context.

### D.5 LOGIT LENS ANALYSIS

Figures 13 to 17 present logit lens visualizations for each model. We observe different trends:

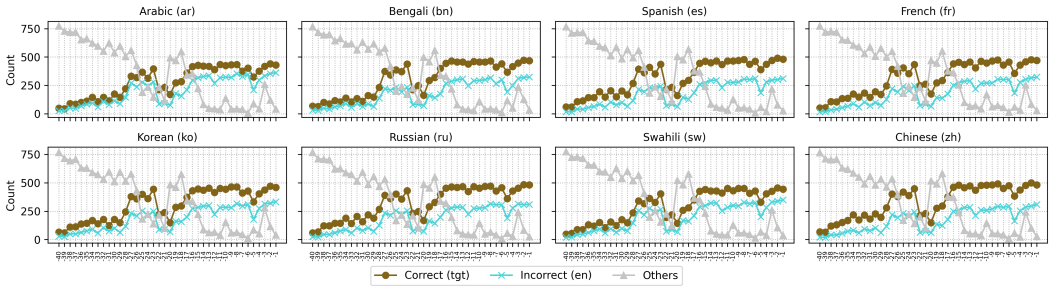



Figure 13: **Logit lens visualization per language for LLAMA-3.3 70B (80 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. We show the last 40 layers. ●: Correct citation ID of document in target language; ×: Wrong citation ID of document in English; ▲: Not in valid citation set.

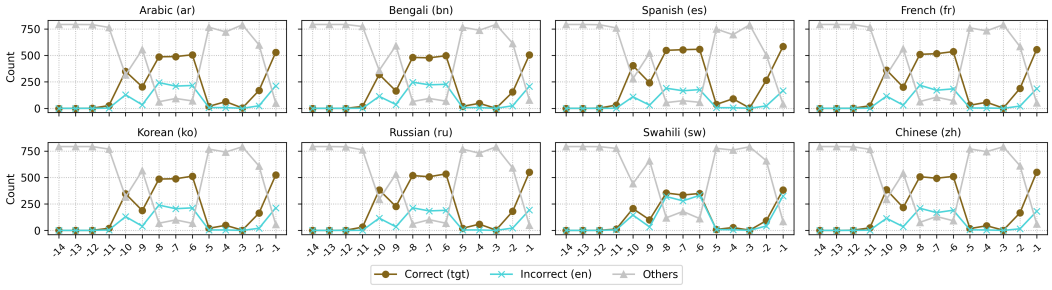



Figure 14: **Logit lens visualization per language for QWEN-3 8B (36 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. We show the last 14 layers.

**LLAMA-3.3 70B.** The model follows a trajectory similar to LLAMA-3.1 8B. Both the correct and wrong citation ID predictions begin to rise around layer 40, peak sharply at layers 52-57, then decline until layer 60 before increasing again and stabilizing toward the final layers. Throughout, correct predictions consistently outnumber incorrect ones. As with LLAMA-3.1 8B, the gap between correct and incorrect predictions narrows for lower-resource languages.

**QWEN-3 8B.** The model exhibits a staggered pattern, where correct citation IDs peak around layer 26, again at layers 28-30, and once more at the final layer, remaining low in between. While the model already predicts the correct IDs in earlier layers (28-30), they are overtaken by invalid predictions just before the final two layers, after which the model uncovers and ends with a final peak in accuracy.

**QWEN-3 14B.** Despite belonging to the same QWEN-3 family, this model exhibits a completely different behavior from QWEN-3 8B. For most of its layers, it fails to predict outputs in the expected citation format. Only in the final layers (38-40), we observe an increase in correct citation predictions, consistently outpacing incorrect ones. This suggests a more conservative prediction strategy, where it delays citation prediction until the very end, or it can only recognize the citation format at the final layers.

**GEMMA-3 27B.** Similar to the QWEN-3 8B, this model shows a staggered pattern, where incorrect predictions remain low, while correct predictions generally increase. There are sharp drops around layers 53-54 and layer 58. However, the model recovers by the final layer, and the count of correct predictions stays high.

**AYA23 8B.** This model stands out from the others, as incorrect predictions generally outnumber correct ones. This aligns with the results in Table 1, where AYA23 8B shows the largest average accuracy drop for target languages. It is also especially pronounced for lower-resource languages like Bengali or Swahili, where the gap between correct and incorrect predictions is even wider.

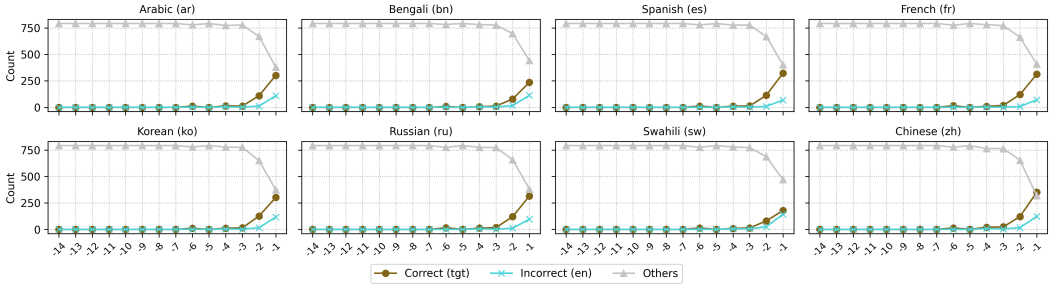



Figure 15: **Logit lens visualization per language for QWEN-3 14B (40 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. We show the last 14 layers.

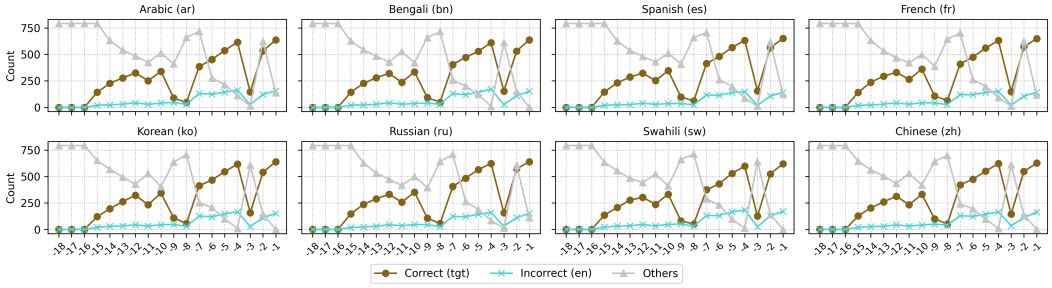



Figure 16: **Logit lens visualization per language for GEMMA-3 27B (62 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. We show the last 18 layers to capture the entire pattern.

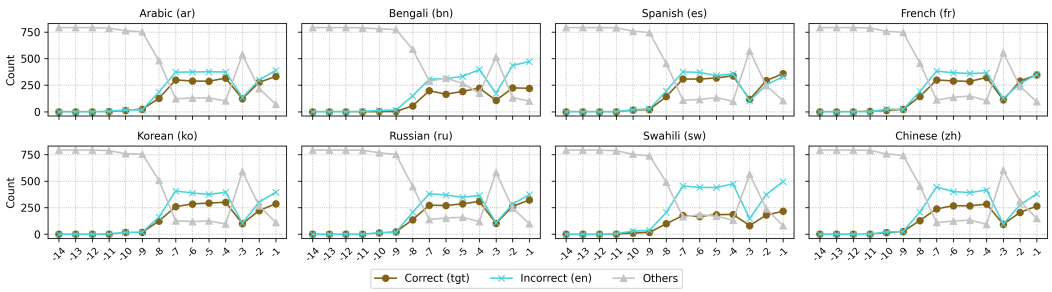



Figure 17: **Logit lens visualization per language for AYA23 8B (32 layers).** $x$-axis: Last layer index; $y$-axis: Statement count. We show the last 14 layers.

| Model | Language | $d_c = \ell, d_{\neg c} = \ell$ (●) | $d_c = \text{en}, d_{\neg c} = \ell$ (■) | $d_c = \text{en}, d_{\neg c} = \text{en}$ (▲) | $d_c = \ell, d_{\neg c} = \text{en}$ (◆) |
|---|---|---|---|---|---|
| LLAMA-3.1 8B | Arabic | 0.616 | 0.572 | 0.557 | **0.658** |
| | Bengali | 0.673 | 0.619 | 0.683 | **0.729** |
| | Spanish | 0.683 | 0.639 | 0.667 | **0.717** |
| | French | **0.716** | 0.713 | 0.674 | **0.716** |
| | Korean | 0.645 | 0.646 | 0.611 | **0.667** |
| | Russian | 0.736 | 0.666 | 0.686 | **0.745** |
| | Swahili | 0.627 | 0.638 | 0.640 | **0.648** |
| | Chinese | 0.607 | **0.631** | 0.579 | 0.610 |
| LLAMA-3.3 70B | Arabic | 0.828 | 0.843 | **0.858** | 0.837 |
| | Bengali | 0.883 | 0.878 | 0.875 | **0.890** |
| | Spanish | 0.875 | **0.886** | 0.877 | 0.883 |
| | French | 0.875 | 0.893 | 0.880 | **0.906** |
| | Korean | 0.866 | **0.893** | 0.880 | 0.857 |
| | Russian | 0.893 | **0.912** | 0.900 | 0.901 |
| | Swahili | 0.902 | 0.902 | **0.904** | 0.879 |
| | Chinese | 0.792 | **0.815** | 0.772 | 0.811 |
| QWEN-3 8B | Arabic | **0.635** | 0.598 | 0.583 | 0.626 |
| | Bengali | 0.605 | 0.560 | **0.650** | 0.632 |
| | Spanish | 0.648 | 0.600 | 0.603 | **0.665** |
| | French | 0.621 | 0.575 | 0.585 | **0.650** |
| | Korean | 0.677 | **0.705** | 0.672 | 0.655 |
| | Russian | **0.710** | 0.686 | 0.648 | 0.673 |
| | Swahili | 0.477 | 0.459 | 0.463 | **0.479** |
| | Chinese | 0.538 | 0.533 | **0.554** | 0.487 |
| QWEN-3 14B | Arabic | 0.832 | 0.773 | 0.787 | **0.843** |
| | Bengali | **0.910** | 0.892 | 0.901 | 0.909 |
| | Spanish | 0.877 | 0.842 | 0.845 | **0.906** |
| | French | 0.883 | 0.857 | 0.868 | **0.908** |
| | Korean | **0.858** | 0.843 | 0.853 | 0.843 |
| | Russian | 0.889 | 0.867 | 0.875 | **0.898** |
| | Swahili | **0.759** | 0.735 | 0.743 | 0.697 |
| | Chinese | 0.744 | **0.765** | 0.737 | 0.730 |
| GEMMA-3 27B | Arabic | 0.823 | 0.808 | 0.814 | **0.859** |
| | Bengali | 0.868 | 0.867 | 0.873 | **0.897** |
| | Spanish | 0.852 | 0.854 | 0.852 | **0.897** |
| | French | 0.863 | 0.845 | 0.861 | **0.898** |
| | Korean | 0.844 | 0.862 | 0.853 | **0.867** |
| | Russian | 0.878 | 0.851 | 0.867 | **0.902** |
| | Swahili | 0.832 | 0.806 | 0.836 | **0.896** |
| | Chinese | **0.811** | 0.792 | 0.779 | 0.792 |
| AYA23 8B | Arabic | 0.464 | 0.443 | 0.475 | **0.516** |
| | Bengali | 0.454 | 0.401 | 0.354 | **0.537** |
| | Spanish | 0.563 | **0.577** | 0.555 | 0.569 |
| | French | 0.551 | 0.574 | 0.575 | **0.578** |
| | Korean | **0.574** | 0.537 | 0.501 | 0.572 |
| | Russian | 0.531 | 0.540 | 0.532 | **0.590** |
| | Swahili | 0.427 | 0.312 | 0.358 | **0.528** |
| | Chinese | 0.453 | 0.460 | **0.492** | 0.488 |

Table 10: **Numerical results when the query is in target language.** We report accuracies for four variants per model and language. We use the same shape notation as in Figure 4. Best scores for each row is **bold**.

### D.6 QUERY LANGUAGE VARIANTS

In Table 10, we report the full numerical results when the query is posed in a target language. We consider four variants, differing in the language of the cited document and the remaining evidence documents, following the same notation introduced in Figure 4: (1) $d_c = \ell, d_{\neg c} = \ell$: all documents in the query language, (2) $d_c = \text{en}, d_{\neg c} = \ell$: cited document in English and all other documents in the query language, (3) $d_c = \text{en}, d_{\neg c} = \text{en}$: all documents in English, and (4) $d_c = \ell, d_{\neg c} = \text{en}$: cited document in the query language and all other documents in English. Overall, we find that models tend to prefer citing evidence in the query language, with $d_c = \ell, d_{\neg c} = \text{en}$ configuration achieving the highest accuracy in more than half of the cases.

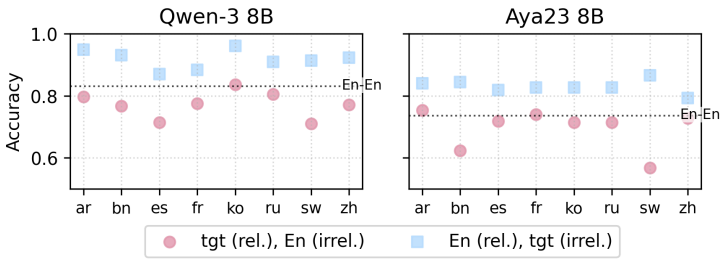



Figure 18: **Accuracy per model with one relevant and one irrelevant evidence document in different languages.** ●: Relevant doc in target language, irrelevant doc in English; ■: Relevant doc in English, irrelevant doc in target language; ····: Baseline, both docs in English.

### D.7 Relevance vs. Language Preference

In Figure 18, we plot citation accuracy for the remaining models (Qwen-3 8B and Aya23 8B) with one relevant and one irrelevant evidence document in different languages, complementing the results in Figure 5. Table 11 reports the full numerical results using the notation from Section 7: (1) **En-En**: both relevant and irrelevant documents are in English, (2) **tgt-En**: relevant document in the target language and irrelevant document in English, and (3) **En-tgt**: relevant document in English and irrelevant document in the target language. Overall, we observe that citation accuracy in **tgt-En** is generally lower than the **En-En** baseline, while **En-tgt** is consistently higher, both indicating a strong English preference that persists regardless of differences in document relevance.

## E Contributive Attribution Patterns

Our analysis of language preference has been based on *corroborative* attribution, measuring the probability of generating in-line citations, which identifies sources that *support* a statement (Menick et al., 2022; Liu et al., 2023). However, if models are citing more English documents, that does not necessarily mean they are actually attributing on their content. If models truly favor English sources, we would expect that preference to also appear when we examine *contributive* attribution, which identifies sources that *cause* a model to generate a specific statement.

To test this, we use an attribution model ContextCite (Cohen-Wang et al., 2024), which estimates the influence of each document on the model's generation. ContextCite is a fitted linear surrogate model that encodes the importance of each source in the context by taking *ablated* contexts $m \in \{0,1\}^K$ as input, where $m_j = 1$ indicates that sentence $j$ is present and $m_j = 0$ indicates that it is masked. The model predicts the ground-truth logit-scaled probability for a given mask $m$ as:

$$f(m) = w^\intercal m + b, \tag{4}$$

where $w \in \mathbb{R}^K$ contains per-sentence attribution weights and $b$ is a bias term.

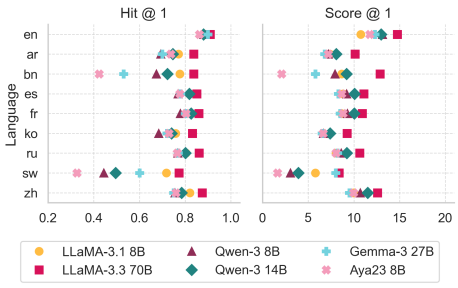



Figure 19: **Hit@1 and Score@1 by model and language.** Higher values indicate more accurate attribution to the cited document.

In our case, given a query $q$, a set of $K$ relevant documents $\mathcal{D} = \{d_1, \ldots, d_k\}$, and a pool of statements $\{s_i\}$, ContextCite returns a ranked list of sentences from $\mathcal{D}$ that most influenced the generation of each $s_i$, along with their attribution scores. Here, $\mathcal{D}$ is composed of the cited document in the target language and all remaining documents in English. We evaluate attribution quality using two metrics: (1) **Hit@1** (↑): whether the top-ranked sentence originates from the cited document, and (2) **Score@1** (↑): the attribution score $w_{j^*}$ of the top-ranked sentence, indicating its estimated relative importance to the model's prediction.

| Model | Language | En-En | tgt-En (↓) | En-tgt (↑) |
|---|---|---|---|---|
| LLAMA-3.1 8B | Arabic | 0.944 | 0.931 | 0.961 |
| | Bengali | 0.944 | 0.918 | 0.965 |
| | Spanish | 0.944 | 0.913 | 0.970 |
| | French | 0.944 | 0.939 | 0.970 |
| | Korean | 0.944 | 0.952 | 0.974 |
| | Russian | 0.944 | 0.965 | 0.961 |
| | Swahili | 0.944 | 0.974 | 0.961 |
| | Chinese | 0.944 | 0.944 | 0.970 |
| LLAMA-3.3 70B | Arabic | 0.974 | 0.935 | 0.983 |
| | Bengali | 0.974 | 0.944 | 0.987 |
| | Spanish | 0.974 | 0.926 | 0.987 |
| | French | 0.974 | 0.957 | 0.987 |
| | Korean | 0.974 | 0.944 | 0.978 |
| | Russian | 0.974 | 0.957 | 0.987 |
| | Swahili | 0.974 | 0.961 | 0.978 |
| | Chinese | 0.974 | 0.952 | 0.983 |
| QWEN-3 8B | Arabic | 0.831 | 0.796 | 0.948 |
| | Bengali | 0.831 | 0.766 | 0.931 |
| | Spanish | 0.831 | 0.714 | 0.870 |
| | French | 0.831 | 0.775 | 0.883 |
| | Korean | 0.831 | 0.836 | 0.961 |
| | Russian | 0.831 | 0.805 | 0.909 |
| | Swahili | 0.831 | 0.710 | 0.913 |
| | Chinese | 0.831 | 0.771 | 0.922 |
| QWEN-3 14B | Arabic | 0.961 | 0.926 | 0.970 |
| | Bengali | 0.961 | 0.918 | 0.987 |
| | Spanish | 0.961 | 0.922 | 0.974 |
| | French | 0.961 | 0.944 | 0.970 |
| | Korean | 0.961 | 0.918 | 0.974 |
| | Russian | 0.961 | 0.935 | 0.970 |
| | Swahili | 0.961 | 0.896 | 0.974 |
| | Chinese | 0.961 | 0.918 | 0.961 |
| GEMMA-3 27B | Arabic | 0.944 | 0.887 | 0.970 |
| | Bengali | 0.944 | 0.905 | 0.970 |
| | Spanish | 0.944 | 0.862 | 0.974 |
| | French | 0.944 | 0.905 | 0.961 |
| | Korean | 0.944 | 0.905 | 0.965 |
| | Russian | 0.944 | 0.931 | 0.952 |
| | Swahili | 0.944 | 0.887 | 0.961 |
| | Chinese | 0.944 | 0.883 | 0.965 |
| AYA23 8B | Arabic | 0.736 | 0.753 | 0.840 |
| | Bengali | 0.736 | 0.623 | 0.844 |
| | Spanish | 0.736 | 0.719 | 0.818 |
| | French | 0.736 | 0.740 | 0.827 |
| | Korean | 0.736 | 0.714 | 0.827 |
| | Russian | 0.736 | 0.714 | 0.827 |
| | Swahili | 0.736 | 0.567 | 0.866 |
| | Chinese | 0.736 | 0.727 | 0.792 |

Table 11: **Numerical results for setup with one relevant and one irrelevant evidence document, in different languages.** We use the same label as in Figure 18. Red denotes **tgt-En** scores that are lower than the **En-En** baseline; Green denotes **En-tgt** scores that are higher than the baseline.

Figure 19 presents both metrics by each model and language. Across all models, both metrics peak when the cited document is in English, outperforming all target language counterparts. This suggests that English preference is not merely a surface-level citation pattern but reflects more reliance on English sources during generation. Full numerical results for Hit@$k$ and Score@$k$ ($k \in \{1, 3\}$) are presented in Table 21.

## F Language Variants of Non-cited Documents

As described in Section 3.2, our main measurement setup constructs contrastive contexts in which only the document to be cited is in English, while all other documents are in the target language. This design choice isolates the effect of the cited document's language by holding other factors constant—changing all other documents would introduce additional confounders that make cross-language comparison less direct.

To ensure that the observed English preference is not merely an artifact of having the non-cited documents in English, we conduct an additional experiment in which the language of all non-cited documents matches the language of the cited document, while keeping the query in English. Specifically, in Step 4 (next token prediction) of Section 3.2, we replace the original configuration $\text{Context}(d_{c_i} \to \ell, d_{\neg c_i} \to \text{en})$ with $\text{Context}(d_{c_i} \to \ell, d_{\neg c_i} \to \ell)$. As shown in Table 12, although citation accuracy increases relative to the original setup (*i.e.,* where non-cited documents are in English), accuracies in this new configuration still remain significantly below the English baseline. This demonstrates that English preference persists even under matched-language contexts.

## G Constrained Decoding Results

While we carefully control our prompt templates (Section A), some valid generations may still fall outside our main accuracy metric—for instance, citation IDs expressed in different languages or stylistic variants. To address this, we conduct an additional experiment using constrained decoding (Post & Vilar, 2018), restricting the model to generate only one of the valid citation ID numbers for each query. This setup effectively removes all stylistic variation. As shown in Table 13, English still achieves the highest citation accuracy, indicating that the English preference persists even when stylistic differences are fully eliminated.

## H Relevance vs. Query Language Preference

We conduct the same set of experiments from Section 7 in a setting where the query is in a language other than English. We use the same dataset, MIRACL (231 queries) with machine translated queries, relevant, and irrelevant documents. While fixing the query in target language, we vary the language of one relevant and one irrelevant document under three conditions: (1) **tgt-tgt** (●): Both relevant and irrelevant documents are in the target (query) language; (2) **tgt-En** (■): Relevant document in the target language, irrelevant document in English; (3) **En-tgt** (▲): Relevant document in English, irrelevant document in the target language.

Our hypotheses are: (a) If citation accuracy in **En-tgt** is lower than **tgt-tgt** or **tgt-En**, it suggests that models trade off relevance for query language preference, citing irrelevant target language documents over relevant English ones, and (b) If citation accuracy of **tgt-En** exceeds **tgt-tgt**, it indicates that models more easily ignore irrelevant English distractors, showing stronger query language preference over English preference.

As shown in Figure 20, results support the first hypothesis: citation accuracies are generally the lowest when the relevant document is in English and the distractor is in the query language, showing that models preferring language over relevance persists for non-English queries. Interestingly, we show that this trend is most evident for (*i*) lower-resource languages such as Bengali (bn) and Swahili (sw) and (*ii*) models that reported to support all tested target languages (Qwen-3 8B and 14B, Gemma-3 27B; Table 4). Conversely, we show mixed results for the second hypothesis. The citation accuracies of **tgt-En** and **tgt-tgt** are largely similar, suggesting that English distractors are not necessarily easier at misleading models than those in the target language. Overall, our results imply that models

| Model | Language | Acc. ($d_{\neg c_i} = \text{en}$) (%, ↑) | Acc. ($d_{\neg c_i} = \ell$) (%, ↑) |
|---|---|---|---|
| **LLAMA-3.1 8B** | English | 67.4 | 67.4 |
| | Arabic | 59.5** | 61.8* |
| | Bengali | 56.6*** | 59.2** |
| | Spanish | 62.1* | 63.5 |
| | French | 62.9 | 64.3 |
| | Korean | 61.7* | 62.7* |
| | Russian | 62.1* | 63.2 |
| | Swahili | 53.0*** | 57.8*** |
| | Chinese | 59.9* | 58.9** |
| **LLAMA-3.3 70B** | English | 85.9 | 85.9 |
| | Arabic | 67.3*** | 77.4*** |
| | Bengali | 68.8*** | 78.6* |
| | Spanish | 76.0*** | 80.3* |
| | French | 77.4*** | 80.6* |
| | Korean | 69.2*** | 75.9*** |
| | Russian | 74.5*** | 78.5** |
| | Swahili | 67.3*** | 76.3*** |
| | Chinese | 74.1*** | 75.7*** |
| **QWEN-3 8B** | English | 62.6 | 62.6 |
| | Arabic | 47.6*** | 51.1*** |
| | Bengali | 41.3*** | 46.8*** |
| | Spanish | 51.9*** | 54.6*** |
| | French | 48.4*** | 50.3*** |
| | Korean | 49.7*** | 55.7** |
| | Russian | 50.4*** | 54.6*** |
| | Swahili | 30.4*** | 39.2*** |
| | Chinese | 49.2*** | 50.4*** |
| **QWEN-3 14B** | English | 83.0 | 83.0 |
| | Arabic | 72.6*** | 73.8*** |
| | Bengali | 65.4*** | 71.5*** |
| | Spanish | 77.4* | 79.3* |
| | French | 76.0*** | 76.9** |
| | Korean | 70.3*** | 73.8*** |
| | Russian | 74.8*** | 78.3* |
| | Swahili | 54.7*** | 62.8*** |
| | Chinese | 73.5*** | 74.4*** |
| **GEMMA-3 27B** | English | 86.2 | 86.2 |
| | Arabic | 78.4*** | 79.9** |
| | Bengali | 77.9*** | 80.3** |
| | Spanish | 80.2** | 82.0 |
| | French | 79.0** | 81.4* |
| | Korean | 77.5*** | 80.7* |
| | Russian | 77.1*** | 79.4** |
| | Swahili | 74.0*** | 78.8*** |
| | Chinese | 75.4*** | 79.4** |
| **AYA23 8B** | English | 60.0 | 60.0 |
| | Arabic | 43.2*** | 49.2*** |
| | Bengali | 27.2*** | 48.3*** |
| | Spanish | 49.1*** | 52.9* |
| | French | 48.5*** | 53.1* |
| | Korean | 42.2*** | 47.4*** |
| | Russian | 48.1*** | 51.7** |
| | Swahili | 22.4*** | 28.9*** |
| | Chinese | 46.3*** | 48.3*** |

Table 12: **Citation accuracies (%) when changing the language of non-cited documents.** Pairwise two-sided $t$-tests are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. Bonferroni correction is applied for multiple comparisons. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant. $d_{\neg c_i}$: non-cited documents; $\ell$: target language.

| Language | LLAMA-3.1 8B | QWEN-3 8B | AYA23 8B | QWEN-3 14B | GEMMA-3 27B | LLAMA-3.3 70B |
|---|---|---|---|---|---|---|
| **English** | 82.5 | 80.8 | 81.7 | 77.5 | 86.7 | 51.0 |
| **Arabic** | 61.2 (-21.3)*** | 60.5 (-20.3)*** | 65.9 (-15.8)*** | 61.2 (-16.3)*** | 80.2 (-6.5)*** | 41.3 (-9.7)*** |
| **Bengali** | 59.3 (-23.2)*** | 63.5 (-17.3)*** | 63.0 (-18.7)*** | 52.7 (-24.8)*** | 81.4 (-5.3)*** | 36.4 (-14.6)*** |
| **Spanish** | 70.5 (-12.0)*** | 67.4 (-13.4)*** | 72.0 (-9.7)*** | 69.1 (-8.4)*** | 83.5 (-3.2)* | 47.5 (-3.5)* |
| **French** | 71.0 (-11.5)*** | 66.7 (-14.1)*** | 70.3 (-11.4)*** | 68.2 (-9.3)*** | 83.0 (-3.7)** | 47.0 (-4.0)** |
| **Korean** | 65.3 (-17.2)*** | 65.5 (-15.3)*** | 65.0 (-16.7)*** | 62.8 (-14.7)*** | 78.9 (-7.8)*** | 39.5 (-11.5)*** |
| **Russian** | 69.2 (-13.3)*** | 65.9 (-14.9)*** | 68.8 (-12.9)*** | 64.8 (-12.7)*** | 82.1 (-4.6)*** | 43.8 (-7.2)*** |
| **Swahili** | 57.1 (-25.4)*** | 61.9 (-18.9)*** | 47.2 (-34.5)*** | 43.6 (-33.9)*** | 78.4 (-8.3)*** | 35.2 (-15.8)*** |
| **Chinese** | 68.8 (-13.7)*** | 68.4 (-12.4)*** | 68.9 (-12.8)*** | 63.8 (-13.7)*** | 79.0 (-7.7)*** | 39.0 (-12.0)*** |

Table 13: **Citation accuracies (%) using constrained decoding.** We present mean accuracy values $\mathbf{Acc}^{(\ell)}$ with $\Delta(\ell_{\text{target}})$ in subscript. Pairwise two-sided $t$-tests are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. Bonferroni correction is applied for multiple comparisons. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant. Color coding indicates the magnitude of $\Delta(\ell_{\text{target}})$: largest, second largest, others.

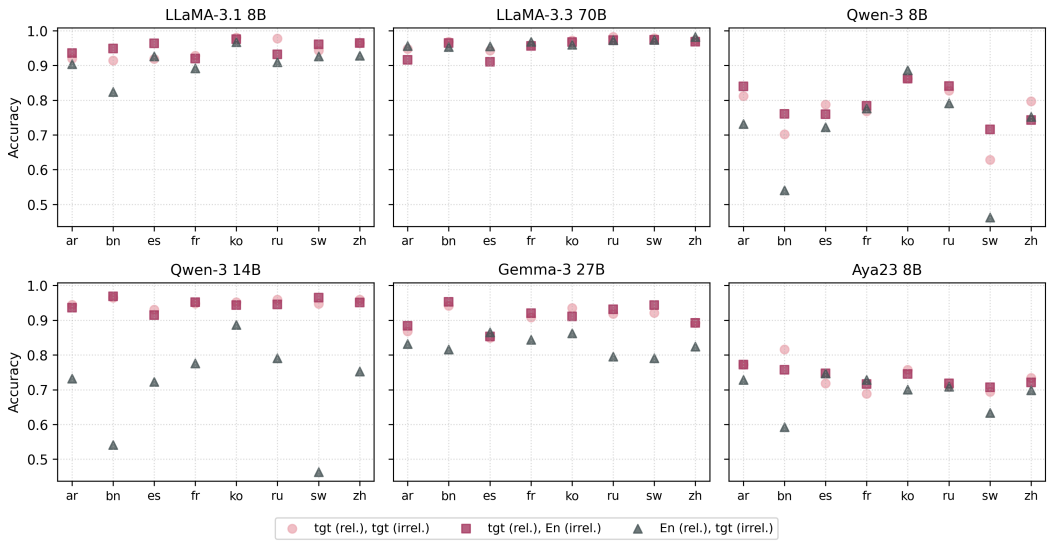



Figure 20: **Accuracy per model with one relevant and one irrelevant evidence document in different languages.** ●: Both docs in target language; ■: Relevant doc in target language, irrelevant doc in English; ▲: Relevant doc in English, irrelevant doc in target language. Models also trade off document relevance for language preference for queries not in English.

show a consistent preference for the query language over relevance, but the distractor's language matters less when the query is not in English.

# I CORRELATION OF MT QUALITY VS. ENGLISH PREFERENCE

We explicitly analyze the relationship between MT quality and English preference. Using COMET-QE scores (Section D.1), we compute segment-level Pearson correlations ($r$) between MT quality and answer accuracy at both (1) the **statement** level: correlating MT quality of the cited document with statement accuracy, and (2) the **query** level: correlating MT quality of the query with its aggregated accuracy. As in Tables 14 and 15, correlations are consistently none to very weak across all models and languages, indicating no meaningful link between MT quality and English preference.

| Language | LLAMA-3.1 8B | LLAMA-3.3 70B | QWEN-3 8B | QWEN-3 14B | GEMMA-3 27B | AYA23 8B |
|---|---|---|---|---|---|---|
| **Arabic** | -0.008 | 0.014 | -0.020 | 0.021 | 0.022 | -0.031 |
| **Bengali** | -0.057 | -0.019 | -0.071 | -0.012 | -0.014 | -0.041 |
| **Spanish** | 0.007 | 0.039 | -0.018 | 0.003 | 0.032 | -0.030 |
| **French** | -0.029 | -0.024 | -0.018 | 0.036 | 0.015 | -0.010 |
| **Korean** | -0.045 | -0.026 | -0.022 | -0.010 | 0.000 | -0.071 |
| **Russian** | 0.001 | -0.057 | -0.040 | -0.002 | -0.007 | -0.034 |
| **Swahili** | -0.028 | 0.014 | -0.002 | 0.025 | 0.029 | -0.083* |
| **Chinese** | -0.030 | -0.040 | 0.014 | -0.039 | 0.005 | -0.026 |

Table 14: **Pearson's correlation ($r$) between MT quality of cited document and statement accuracy.** The reported $p$-values correspond to two-sided significance tests for the null hypothesis that the true correlation is zero. *: significant with $p < 0.05$; non-marked: not statistically significant.

| Language | LLAMA-3.1 8B | LLAMA-3.3 70B | QWEN-3 8B | QWEN-3 14B | GEMMA-3 27B | AYA23 8B |
|---|---|---|---|---|---|---|
| **Arabic** | 0.154* | -0.030 | 0.128* | 0.133* | 0.027 | 0.140* |
| **Bengali** | 0.119 | -0.008 | -0.011 | 0.054 | -0.046 | 0.084 |
| **Spanish** | 0.119 | 0.030 | 0.210 | 0.085 | 0.083 | 0.126 |
| **French** | 0.153* | 0.043 | 0.112 | 0.105 | 0.068 | 0.063 |
| **Korean** | 0.083 | -0.028 | 0.073 | 0.026 | 0.004 | 0.031 |
| **Russian** | 0.097 | -0.030 | 0.142 | 0.033 | 0.034 | 0.035 |
| **Swahili** | 0.183* | 0.144* | 0.120 | 0.155* | 0.076 | 0.052 |
| **Chinese** | 0.195* | -0.001 | 0.058 | 0.028 | -0.068 | 0.083 |

Table 15: **Pearson's correlation ($r$) between MT quality of query and aggregated accuracy.** The reported $p$-values correspond to two-sided significance tests for the null hypothesis that the true correlation is zero. *: significant with $p < 0.05$; non-marked: not statistically significant.

# J ADDITIONAL EXPERIMENTS WITH MIRACL

To complement our main experiment results on ELI5, we conduct additional experiments for measuring English preference (§5) and query-language preference (§6) on an additional dataset, MIRACL (Zhang et al., 2023).

## J.1 ENGLISH PREFERENCE

We follow the same procedure as in ELI5 (§3) using the English portion of the development set. After Step 3 (statement pool construction), we obtain 818 statements. MIRACL is a non-parallel multilingual RAG dataset—where queries are a mix of long- and short-form and not all evidence documents are required for answering the query. Therefore, results should be interpreted with some caution. Despite these differences, Table 16 shows that English preference observed in ELI5 persists in MIRACL, with the English baseline achieving the highest citation accuracy across target languages and models. We further report COMET-QE scores for machine-translated queries, titles, and evidence documents of MIRACL in Table 17, showing reasonable MT quality (average: 0.835 for queries, 0.797 for titles, 0.727 for documents).

## J.2 QUERY LANGUAGE PREFERENCE

MIRACL also provides non-parallel queries and evidence documents *natively* written in eight target languages. For each language, we randomly sample 100 queries and translate their associated evidence documents into English using Google Translate. We then follow the same procedure as in Section 3. For each supported statement, we compute the next token prediction accuracy while varying only the language of the cited document ($d_c$) to English (en) or the query language ($\ell$). All other variables (query and the non-cited documents remain fixed in the query language). This mirrors the setup in Section 6, with the only change being that MIRACL provides naturally occurring data in target languages. As shown in Table 18, we observe the same query-language preference: citation accuracy is consistently higher when $d_c = \ell$ than when $d_c = \text{en}$ across all tested models. This indicates that our findings hold for naturally occurring queries and documents in target languages.

| Language | LLAMA-3.1 8B | QWEN-3 8B | AYA23 8B | QWEN-3 14B | GEMMA-3 27B | LLAMA-3.3 70B |
|---|---|---|---|---|---|---|
| **English** | 75.6 | 83.0 | 66.8 | 87.0 | 85.9 | 65.5 |
| **Arabic** | 54.8 (-20.8)*** | 63.5 (-19.5)*** | 41.7 (-25.1)*** | 68.7 (-18.3)*** | 69.3 (-16.6)*** | 48.4 (-17.1)*** |
| **Bengali** | 54.0 (-21.6)*** | 61.4 (-21.6)*** | 38.8 (-28.0)*** | 60.8 (-26.2)*** | 69.3 (-16.6)*** | 32.6 (-32.9)*** |
| **Spanish** | 60.8 (-14.8)*** | 71.6 (-11.4)*** | 49.4 (-17.4)*** | 77.5 (-9.5)*** | 76.0 (-9.9)*** | 53.6 (-11.9)*** |
| **French** | 60.2 (-15.4)*** | 71.0 (-12.0)*** | 47.3 (-19.5)*** | 76.7 (-10.3)*** | 74.6 (-11.3)*** | 54.5 (-11.0)*** |
| **Korean** | 54.7 (-20.9)*** | 62.7 (-20.3)*** | 45.2 (-21.6)*** | 65.9 (-21.1)*** | 67.9 (-18.0)*** | 45.0 (-20.5)*** |
| **Russian** | 58.2 (-17.4)*** | 68.0 (-15.0)*** | 45.7 (-21.1)*** | 70.7 (-16.3)*** | 71.4 (-14.5)*** | 52.7 (-12.8)*** |
| **Swahili** | 52.0 (-23.6)*** | 64.1 (-18.9)*** | 37.0 (-29.8)*** | 59.5 (-27.5)*** | 69.4 (-16.5)*** | 36.2 (-29.3)*** |
| **Chinese** | 56.0 (-19.6)*** | 64.8 (-18.2)*** | 45.4 (-21.4)*** | 70.3 (-16.7)*** | 66.4 (-19.5)*** | 45.6 (-19.9)*** |

Table 16: **Citation accuracies (%) using MIRACL.** We present mean accuracy values $\mathbf{Acc}^{(\ell)}$ with $\Delta(\ell_{\text{target}})$ in subscript. Pairwise two-sided $t$-tests are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. Bonferroni correction is applied for multiple comparisons. ***: significant with $p < 0.001$. Color coding indicates the magnitude of $\Delta(\ell_{\text{target}})$: largest, second largest, others.

| Language | COMET-QE($q$, $q'$) | COMET-QE($t$, $t'$) | COMET-QE($d$, $d'$) |
|---|---|---|---|
| **Arabic** (ar) | 0.802 | 0.775 | 0.683 |
| **Bengali** (bn) | 0.867 | 0.829 | 0.758 |
| **Spanish** (es) | 0.851 | 0.794 | 0.753 |
| **French** (fr) | 0.847 | 0.798 | 0.756 |
| **Korean** (ko) | 0.862 | 0.814 | 0.741 |
| **Russian** (ru) | 0.821 | 0.808 | 0.703 |
| **Swahili** (sw) | 0.813 | 0.764 | 0.715 |
| **Chinese** (zh) | 0.818 | 0.792 | 0.706 |

Table 17: **COMET-QE scores by language for MIRACL.** We evaluate the machine translation (MT) quality of non-English queries ($q$), titles ($t$), and evidence documents ($d$) in the MIRACL dataset. Apostrophe ($'$) indicates MT. Higher scores indicate better MT quality.

## K ALTERNATIVE MACHINE TRANSLATION SYSTEM

We replicate the main experiments from Section 5 using an alternative machine translation (MT) system to verify whether the English preference trend persists. Specifically, we use TOWER-INSTRUCT 7B[20] (Alves et al., 2024), a model trained for diverse translation-related tasks including general MT, automatic post-editing, and grammatical error correction. Table 19 reports COMET-QE scores by language. Compared to Google Translate (see Table 5), MT quality shows mixed results, with higher scores for all languages except Arabic and Bengali.

Table 20 presents citation accuracies per model and language using TOWER-INSTRUCT translations. We show that the general trend observed with Google Translate persists: models achieve the highest citation accuracy when the cited document is in English. The accuracy gaps between English and target languages are all statistically significant ($p < 0.001$), despite using a stronger MT system compared to Google Translate. This suggests that the English preference cannot be fully attributed to using machine-translated documents. Notably, citing Arabic documents leads to the largest performance drop relative to English across all models, likely reflecting the lower COMET-QE scores for Arabic shown in Table 19.

## L USAGE OF LARGE LANGUAGE MODELS

We used LLMs to support and refine the writing of our work. Importantly, we did not rely on them to generate content or sentences from scratch. Instead, we employed them primarily to polish the clarity and expression of how we presented our results. In addition, we used them for stylistic adjustments, such as improving readability and removing layout issues (*e.g.,* widows and orphans).

[20] `Unbabel/TowerInstruct-7B-v0.2`

| Language | Model | Acc. ($d_c = \text{en}$) (%, ↑) | Acc. ($d_c = \ell$) (%, ↑) |
|---|---|---|---|
| **Arabic** | LLAMA-3.1 8B | 45.3 | 65.7 |
| | LLAMA-3.3 70B | 76.6 | 85.3 |
| | QWEN-3 8B | 43.8 | 64.2 |
| | QWEN-3 14B | 70.9 | 83.8 |
| | GEMMA-3 27B | 64.2 | 82.3 |
| | AYA23 8B | 39.3 | 50.6 |
| **Bengali** | LLAMA-3.1 8B | 50.7 | 72.3 |
| | LLAMA-3.3 70B | 66.2 | 81.1 |
| | QWEN-3 8B | 46.0 | 67.6 |
| | QWEN-3 14B | 71.6 | 89.9 |
| | GEMMA-3 27B | 64.9 | 85.1 |
| | AYA23 8B | 31.1 | 53.4 |
| **Spanish** | LLAMA-3.1 8B | 60.1 | 65.2 |
| | LLAMA-3.3 70B | 75.0 | 78.4 |
| | QWEN-3 8B | 47.9 | 56.1 |
| | QWEN-3 14B | 75.9 | 80.2 |
| | GEMMA-3 27B | 77.4 | 79.0 |
| | AYA23 8B | 46.7 | 52.7 |
| **French** | LLAMA-3.1 8B | 68.9 | 75.4 |
| | LLAMA-3.3 70B | 85.8 | 87.7 |
| | QWEN-3 8B | 69.3 | 73.5 |
| | QWEN-3 14B | 80.9 | 90.0 |
| | GEMMA-3 27B | 84.1 | 91.6 |
| | AYA23 8B | 60.2 | 63.1 |
| **Korean** | LLAMA-3.1 8B | 48.9 | 59.1 |
| | LLAMA-3.3 70B | 69.9 | 76.1 |
| | QWEN-3 8B | 47.7 | 65.9 |
| | QWEN-3 14B | 68.8 | 81.3 |
| | GEMMA-3 27B | 65.3 | 75.6 |
| | AYA23 8B | 44.9 | 64.2 |
| **Russian** | LLAMA-3.1 8B | 54.3 | 58.4 |
| | LLAMA-3.3 70B | 67.0 | 72.6 |
| | QWEN-3 8B | 49.2 | 65.0 |
| | QWEN-3 14B | 67.5 | 80.2 |
| | GEMMA-3 27B | 67.0 | 77.7 |
| | AYA23 8B | 40.1 | 49.2 |
| **Swahili** | LLAMA-3.1 8B | 58.0 | 63.0 |
| | LLAMA-3.3 70B | 74.6 | 79.0 |
| | QWEN-3 8B | 44.9 | 52.2 |
| | QWEN-3 14B | 62.3 | 71.0 |
| | GEMMA-3 27B | 58.7 | 74.6 |
| | AYA23 8B | 49.3 | 57.3 |
| **Chinese** | LLAMA-3.1 8B | 27.8 | 46.3 |
| | LLAMA-3.3 70B | 48.5 | 60.4 |
| | QWEN-3 8B | 35.6 | 38.2 |
| | QWEN-3 14B | 51.9 | 58.2 |
| | GEMMA-3 27B | 45.9 | 63.7 |
| | AYA23 8B | 30.0 | 35.9 |

Table 18: **Results when the query is in target language with MIRACL.** $d_c$: cited document; $\ell$: target language. Note that results are comparable only within each target language, as MIRACL is a non-parallel dataset.

| Language | COMET-QE($q$, $q'$) | COMET-QE($t$, $t'$) | COMET-QE($d$, $d'$) |
|---|---|---|---|
| **Arabic** (ar) | 0.467 | 0.374 | 0.311 |
| **Bengali** (bn) | 0.802 | 0.562 | 0.491 |
| **Spanish** (es) | 0.855 | 0.595 | 0.574 |
| **French** (fr) | 0.859 | 0.598 | 0.548 |
| **Korean** (ko) | 0.857 | 0.597 | 0.554 |
| **Russian** (ru) | 0.839 | 0.588 | 0.549 |
| **Swahili** (sw) | 0.787 | 0.549 | 0.482 |
| **Chinese** (zh) | 0.817 | 0.574 | 0.505 |

Table 19: **COMET-QE scores by language using TOWER-INSTRUCT 7B translations.** We evaluate the machine translation (MT) quality of non-English queries ($q$), titles ($t$), and evidence documents ($d$). Apostrophe ($'$) indicates MT. Higher scores indicate better MT quality.

| Language | LLAMA-3.1 8B | LLAMA-3.3 70B | QWEN-3 8B | QWEN-3 14B | GEMMA-3 27B | AYA23 8B |
|---|---|---|---|---|---|---|
| **English** | 67.4 | 85.9 | 62.6 | 83.0 | 86.2 | 60.0 |
| **Arabic** | 24.6 (-42.8) | 21.1 (-64.8) | 15.9 (-46.7) | 26.3 (-56.7) | 26.4 (-59.8) | 23.2 (-36.8) |
| **Bengali** | 45.5 (-22.0) | 52.9 (-33.0) | 34.1 (-28.5) | 53.4 (-29.6) | 52.4 (-33.8) | 38.6 (-21.4) |
| **Spanish** | 58.5 (-8.91) | 72.5 (-13.4) | 50.2 (-12.4) | 73.4 (-9.64) | 74.2 (-12.0) | 47.5 (-12.5) |
| **French** | 53.0 (-14.4) | 65.7 (-20.2) | 41.8 (-20.8) | 66.4 (-16.6) | 65.0 (-21.2) | 42.5 (-17.4) |
| **Korean** | 55.2 (-12.2) | 62.6 (-23.3) | 45.5 (-17.1) | 65.7 (-17.3) | 69.8 (-16.4) | 41.0 (-19.0) |
| **Russian** | 55.1 (-12.3) | 71.6 (-14.3) | 48.9 (-13.7) | 68.1 (-14.9) | 70.1 (-16.1) | 43.3 (-16.7) |
| **Swahili** | 41.7 (-25.7) | 49.5 (-36.4) | 34.2 (-28.4) | 50.9 (-32.1) | 48.9 (-37.3) | 37.1 (-22.9) |
| **Chinese** | 44.2 (-23.2) | 55.8 (-30.1) | 37.4 (-25.2) | 57.7 (-25.3) | 54.8 (-31.4) | 39.3 (-20.7) |

Table 20: **Citation accuracies (%) by model and language using TOWER-INSTRUCT 7B translations.** We present mean accuracy values $\mathbf{Acc}^{(\ell)}$ along with $\Delta(\ell_{\text{target}})$ in subscript as percent (%). Pairwise two-sided $t$-tests are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. Bonferroni correction is applied for multiple comparisons. All differences are statistically significant ($p < 0.001$). Color coding indicates the magnitude of $\Delta(\ell_{\text{target}})$: largest, second largest, others.

| Model | Language | Hit@1 (↑) | Hit@3 (↑) | Score@1 (↑) | Score@3 (↑) |
|---|---|---|---|---|---|
| LLAMA-3.1 8B | English | 0.880 | 0.971 | 10.737 | 11.114 |
| | Arabic | 0.771 | 0.928 | 7.370 | 7.777 |
| | Bengali | 0.777 | 0.908 | 8.648 | 9.041 |
| | Spanish | 0.821 | 0.934 | 8.797 | 9.195 |
| | French | 0.824 | 0.943 | 8.641 | 9.051 |
| | Korean | 0.758 | 0.924 | 6.649 | 7.064 |
| | Russian | 0.804 | 0.953 | 8.045 | 8.495 |
| | Swahili | 0.718 | 0.903 | 5.768 | 6.197 |
| | Chinese | 0.821 | 0.929 | 9.976 | 10.418 |
| LLAMA-3.3 70B | English | 0.910 | 0.968 | 14.749 | 13.080 |
| | Arabic | 0.837 | 0.955 | 10.138 | 10.594 |
| | Bengali | 0.837 | 0.943 | 12.874 | 13.369 |
| | Spanish | 0.851 | 0.970 | 11.081 | 11.555 |
| | French | 0.860 | 0.970 | 10.918 | 11.401 |
| | Korean | 0.832 | 0.960 | 9.249 | 9.695 |
| | Russian | 0.861 | 0.971 | 10.647 | 11.117 |
| | Swahili | 0.773 | 0.937 | 8.363 | 8.927 |
| | Chinese | 0.875 | 0.984 | 12.576 | 15.075 |
| QWEN-3 8B | English | 0.881 | 0.966 | 13.128 | 13.563 |
| | Arabic | 0.693 | 0.866 | 7.183 | 7.634 |
| | Bengali | 0.674 | 0.826 | 7.912 | 8.376 |
| | Spanish | 0.769 | 0.932 | 9.200 | 9.762 |
| | French | 0.779 | 0.910 | 9.274 | 9.768 |
| | Korean | 0.684 | 0.865 | 6.642 | 7.138 |
| | Russian | 0.768 | 0.929 | 8.602 | 9.154 |
| | Swahili | 0.444 | 0.711 | 3.042 | 3.487 |
| | Chinese | 0.755 | 0.862 | 10.703 | 11.053 |
| QWEN-3 14B | English | 0.880 | 0.973 | 12.986 | 13.399 |
| | Arabic | 0.746 | 0.894 | 8.054 | 8.472 |
| | Bengali | 0.722 | 0.881 | 9.197 | 9.683 |
| | Spanish | 0.818 | 0.957 | 10.075 | 10.562 |
| | French | 0.826 | 0.949 | 10.031 | 10.465 |
| | Korean | 0.740 | 0.899 | 7.393 | 7.877 |
| | Russian | 0.801 | 0.941 | 9.227 | 9.685 |
| | Swahili | 0.495 | 0.745 | 3.918 | 4.396 |
| | Chinese | 0.785 | 0.894 | 11.511 | 12.073 |
| GEMMA-3 27B | English | 0.902 | 0.976 | 12.321 | 12.752 |
| | Arabic | 0.704 | 0.910 | 6.891 | 7.253 |
| | Bengali | 0.535 | 0.722 | 5.784 | 5.781 |
| | Spanish | 0.781 | 0.938 | 8.347 | 8.853 |
| | French | 0.806 | 0.932 | 8.575 | 9.029 |
| | Korean | 0.727 | 0.903 | 6.516 | 7.128 |
| | Russian | 0.776 | 0.911 | 8.223 | 8.771 |
| | Swahili | 0.602 | 0.657 | 8.002 | 2.256 |
| | Chinese | 0.751 | 0.898 | 9.504 | 9.955 |
| AYA23 8B | English | 0.862 | 0.963 | 11.729 | 12.247 |
| | Arabic | 0.737 | 0.912 | 7.180 | 7.671 |
| | Bengali | 0.423 | 0.683 | 2.087 | 2.600 |
| | Spanish | 0.777 | 0.932 | 8.665 | 9.210 |
| | French | 0.804 | 0.927 | 8.811 | 9.280 |
| | Korean | 0.729 | 0.919 | 6.622 | 7.143 |
| | Russian | 0.765 | 0.918 | 8.049 | 8.608 |
| | Swahili | 0.326 | 0.634 | 1.627 | 1.990 |
| | Chinese | 0.760 | 0.883 | 9.955 | 10.443 |

Table 21: **Numerical results for ContextCite.** We report Hit@$k$ and Score@$k$ (where $k \in \{1, 3\}$) for each model across all languages.