WHERE'S WALDO? QUERY-LANGUAGE PREFERENCE UNDER CROSS-LINGUAL KNOWLEDGE DISPARITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models increasingly serve as interfaces for knowledge-intensive information seeking tasks across languages by synthesizing multilingual evidence. Prior work has shown that they often exhibit query-language preference—the tendency to favor sources written in the language of the query—but has largely examined this behavior in settings where equivalent knowledge is available across languages. However, this bias becomes consequential when sources in different languages provide incomplete or inconsistent accounts of the same fact, since the information users receive then depends on the sources a model selects to use. To characterize query-language preference under such cross-lingual knowledge disparities, we introduce WALDO,[1] a multilingual Question-Answering (QA) benchmark constructed from Wikipedia. WALDO contains 12K QA pairs targeting knowledge *gaps*, where a fact is available in one language but absent in another, and knowledge *conflicts*, where language editions provide conflicting versions of the same fact. Evaluating eight models across five languages, we find that when one language edition merely lacks the relevant fact, models generally use evidence from the other language regardless of the query language. Under conflicting accounts, however, model responses strongly align with the document in the query language, causing semantically equivalent queries to elicit different accounts depending on the user's language. Finally, we explore two different approaches that could mitigate this preference under knowledge conflicts: a mechanistic intervention that ablates attention heads associated with query-language preference, and LoRA-based training, which reduces the preference gap by up to 61.5%.[2]

1 INTRODUCTION

Large Language Models (LLMs) increasingly mediate knowledge-intensive information seeking tasks across diverse languages (Longpre et al., 2021a). When answering a user's query, they may draw on information from sources in multiple languages and select which sources inform their final responses (Wu et al., 2024; Chirkova et al., 2024). As LLMs become interfaces between users and multilingual information ecosystems (Nadalic Sotic & Kamps, 2026), understanding *how* they select among competing knowledge sources becomes increasingly important.

A growing body of evidence suggests that multilingual systems often exhibit query-language preference during generation: a tendency to favor knowledge sources written in the same language as the user query (Ki et al., 2026). Consequently, the language in which a user poses a question can influence which sources inform the response, and therefore what information the user ultimately receives. This is not necessarily problematic when equivalent information is available across languages—an assumption underlying much prior work (Park & Lee, 2025; Qi et al., 2025). In practice, however, multilingual sources are rarely perfectly redundant or mutually consistent (Filatova, 2009).

Query-language preference becomes consequential when knowledge is *heterogeneous* across languages, which can take two forms. First, knowledge may be unevenly available: a fact documented

[1] WALDO stands for "**W**ikipedi**A** **L**ost in **D**isparity of kn**O**wledge," named after the character from the Where's Waldo? series (in Figure 1), where readers search crowded scenes to find him; similarly, WALDO tests whether models can use relevant facts across multilingual documents with missing or conflicting knowledge.

[2] Code, data, and fine-tuned models will be released upon publication.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

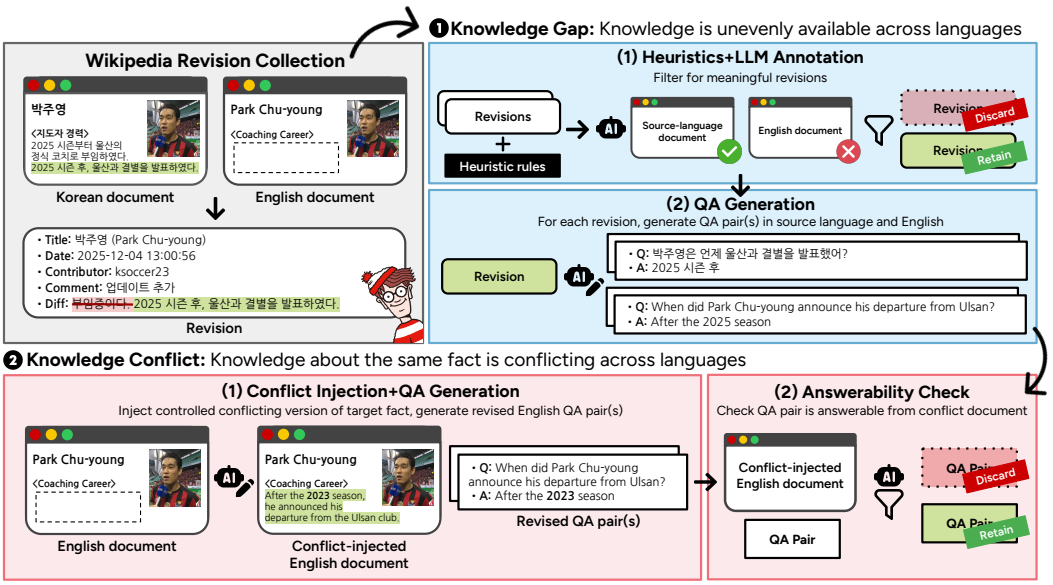


Figure 1: **Overview of WALDO.** For each source language (e.g., Korean), we collect its Wikipedia revisions and pair them with the corresponding English documents. For ❶ **Knowledge Gap**, we identify revisions containing facts present in the source-language document but absent from English, generating aligned source-language and English QA pairs for each. For ❷ **Knowledge Conflict**, we inject a controlled conflicting fact into the English document and generate revised English QA pairs.

in one language may be absent from sources in another, creating a *knowledge gap* (Hecht & Gergle, 2010). Second, sources in different languages may provide incompatible accounts of the same fact, resulting in a *knowledge conflict* (Li et al., 2024a; Cattan et al., 2025). When neither account can be resolved using available signals, semantically equivalent queries in different languages should surface the gap or conflict rather than elicit different responses depending on the query language. At scale, these differences can create asymmetries in answer quality (Jin et al., 2024a), produce uneven downstream performance (Ranaldi et al., 2026), deepen opinion polarization (Sharma et al., 2024), reinforce the marginalization of underrepresented narratives (Phillipson, 1992), and ultimately undermine user trust in multilingual information systems (Arguedas et al., 2023). These risks motivate the study of query-language preference under cross-lingual knowledge disparities.

To this end, we introduce **WALDO**, a multilingual Question-Answering (QA) benchmark for evaluating query-language preference under knowledge gaps and knowledge conflicts (§3). As illustrated in Figure 1, WALDO constructs QA instances around target facts that satisfy two criteria: they appear in Wikipedia documents in languages other than English but are absent from the corresponding English documents (§3.1), and they postdate the reported knowledge cutoffs of the evaluated model. This design helps minimize the possibility that models answer from their parametric knowledge, enabling observed response differences to be more directly attributed to their use of the context (Jin et al., 2024b). We additionally introduce factually inconsistent versions of the target facts by manually perturbing the corresponding English Wikipedia documents to simulate real-world knowledge conflict scenarios. By holding the available multilingual context fixed while varying *only* the query language, WALDO enables a counterfactual test of whether query language alone changes the knowledge source on which a model bases its answer.

Using WALDO, we evaluate eight models across five languages (§4) and diagnose the extent to which they exhibit query-language preference. Under knowledge gaps, models largely overcome this preference, producing consistent responses across query languages. Under knowledge conflicts, however, their response distributions differ substantially by query language (§5). These findings motivate our exploration of potential approaches for mitigating query-language preference in the presence of conflicting evidence. We introduce Query-language Preference Head Ablation (**QPHA**) and adapt Quantized Low-Rank Adapter (**QLoRA**) training (Dettmers et al., 2023) as two targeted methods (§6.1), and show both help reduce the preference gap (§6.2). More broadly, WALDO provides an evaluation testbed for studying how language bias interacts with answer quality in multilingual

systems. Our initial analyses and mitigation approaches open new directions for understanding the underlying sources of these biases and improving the robustness of cross-lingual knowledge transfer.

We make three main contributions in this work:

- **Benchmark:** We present WALDO, a multilingual QA benchmark comprising 12K queries targeting cross-lingually heterogeneous facts drawn from Wikipedia. It supports the study of query-language preference under knowledge gaps, where facts are unevenly available across languages, and knowledge conflicts, where languages provide incompatible accounts of the same fact.
- **Diagnosis:** We introduce two metrics for quantifying query-language preference and systematically evaluate eight models across five languages. We find little preference under knowledge gaps, but show that it becomes pronounced under knowledge conflicts.
- **Mitigation:** We evaluate two approaches targeting query-language preference under conflicting evidence: QPHA and QLORA training with two customized objectives. Both methods effectively reduce preference, with QLORA reducing the gap by up to 61.5%.

2 BACKGROUND: DEFINITION, EVALUATION, AND MITIGATION

Definition. Language preference refers to a model's systematic tendency to favor knowledge sources in some languages over others. Such preference may arise from unequal language representation in pre-training data (Joshi et al., 2020) or from cross-lingual tokenization differences (Petrov et al., 2023). Query-language preference is a specific form of language preference in which a model favors sources written in the same language as the query (Ki et al., 2026). Prior work has studied how query-language preference manifests in parametric knowledge (Goldman et al., 2025), multilingual retrieval (Yang et al., 2024; Park & Lee, 2025), and generation (Sharma et al., 2025; Li et al., 2025a; Ki et al., 2026). We focus on *generation*, holding the multilingual retrieved context fixed to isolate how query language affects the knowledge source on which a model bases its response.

Datasets and Evaluation. Existing work studies query-language preference in two broad settings: settings that assume semantically equivalent knowledge is available across languages, and simulated settings with limited ecological validity (de Vries et al., 2020). One line of work retrieves naturally occurring evidence from real-world sources (Qi et al., 2025; Park et al., 2026; Huo et al., 2026; Wang et al., 2026). However, they cannot isolate query-language preference when language editions differ in the availability or content of relevant facts. Conversely, Sharma et al. (2025) introduces cross-lingual knowledge conflicts, but uses fictional documents. A separate body of work documents cross-lingual knowledge gaps (Massa & Scrinzi, 2012; Bao et al., 2012; Miquel-Ribé & Laniado, 2018; 2020; Hupkes & Bogoychev, 2025) as well as factual and narrative inconsistencies across Wikipedia language editions (Samir et al., 2024; Li et al., 2024a; Zhao et al., 2026a). Yet, this work generally neither provides a reusable benchmark nor connects such disparities to query-language preference. WALDO addresses these gaps by targeting naturally occurring, cross-lingually heterogeneous facts from Wikipedia, covering knowledge gaps and controlled knowledge conflicts.

Mitigation. Prior efforts to mitigate language preference have primarily targeted the knowledge-source retrieval stage, including calibrated relevance estimation (Park et al., 2026; Wang et al., 2026; Jang et al., 2026), query translation (Amiraz et al., 2025), and language-controlled retrieval (Huo et al., 2026; Qi et al., 2026; Lee et al., 2026; Hong et al., 2026). At the generation stage, existing approaches translate retrieved documents into English (Ranaldi et al., 2026) or into the query language (Park & Lee, 2025) to unify the evidence in a single pivot language. To our knowledge, no prior work has specifically studied *query*-language preference at generation time or explored mechanistic interventions or fine-tuning approaches to mitigate it; we take a first step toward doing so.

3 CREATING WALDO

Wikipedia is a valuable source of cross-lingual knowledge disparities: documents about the same entity in different language editions are not merely translations of one another (Hecht, 2013), but contain localized knowledge (Haraway, 1988; Kumar, 2021) shaped by the distinct contributor communities that produce and maintain them (Hale, 2014; Sen et al., 2015; Johnson & Lescak, 2022). For each source language, we collect Wikipedia documents and their contributor-authored revision

histories from publicly available Wikipedia dumps available as of June 1, 2026.[3] We retain revisions made on or after January 1, 2020, representing each revision as a diff to its parent revision.

Source Languages. We collect documents in five source languages spanning diverse resource levels, language families, scripts, and Wikipedia sizes: French (fr), Simplified Chinese (zh), Japanese (ja), Korean (ko), and Bengali (bn). Per-language characteristics are detailed in Appendix A.1.

We refer to the complete collection of revisions per language as the *full* set, which we plan to release. For our experiments, we sample five dump files per language to form a balanced *evaluation* set.

3.1 KNOWLEDGE GAP [❶]

Heuristics & LLM Annotation. To retain *meaningful* revisions for evaluation, we filter candidate revisions using a combination of heuristic rules and LLM-based annotation with GPT-5.6-TERRA (OpenAI, 2026b). After manual auditing, we define a suite of rules that a revision must (1) not be a bot edit or maintenance-related change (e.g., a revert); (2) have a length within a predefined range; (3) introduce at least one factual statement without removing information; and (4) contain an introduced fact that remains in the latest version of the source-language Wikipedia document but is absent from the latest linked English document, identified via the same Wikidata QID.

During LLM annotation, we also record a `timestamp`, formatted as YYYY-MM, indicating the month in which the introduced fact is asserted to have occurred. This procedure retains, on average, 0.3% of candidate revisions across languages. See Appendix A.2 for filtering details, examples of excluded revisions, and per-language retain rates.

QA Generation. For each retained revision, we provide GPT-5.6-TERRA with the revision, its source-language document, and the corresponding English document. The model generates *aligned* QA pairs in the source language ($\mathcal{Q}_{\text{src}}, \mathcal{A}_{\text{src}}$) and English ($\mathcal{Q}_{\text{en}}, \mathcal{A}_{\text{en}}$), which are translations of one another and target the same fact introduced by the revision. We retain a revision only if the source-language and English sets contain the same number of QA pairs and every aligned QA pair achieves an XCOMET translation-quality score of at least 80.0 (Guerreiro et al., 2023). Further details and per-language XCOMET scores are provided in Appendix A.3.

3.2 KNOWLEDGE CONFLICT [❷]

Conflict Injection & QA Generation. For each revision retained in §3.1, we introduce a controlled knowledge conflict by prompting an LLM (GPT-5.6-TERRA) to inject an inconsistent version of the target fact into the corresponding English document and update the answers to the same questions (Chen et al., 2022; Mo et al., 2026; Zhao et al., 2026a;b). To simulate real-world knowledge conflicts, we follow prior conflict-injection practices (Longpre et al., 2021b; Wang et al., 2024). Each injected conflict must satisfy two criteria: (1) it modifies a core attribute of the target fact, such as a name, quantity, date, or location, rather than merely negating it (e.g., "is" → "is not"); and (2) it is applied consistently throughout the document while preserving the original tone. Table 1 reports the distribution of conflict types per language, with numerical and date changes being the most common on average. Appendix A.4 provides qualitative examples of each conflict type.

Answerability Check. To verify that each conflict-injected English document supports its revised answer, we use GPT-5.6-SOL, a different model from the generator, to answer each original English question using the conflict-injected document. We retain only QA pairs that pass this answerability check (Sun et al., 2026); on average, 98.8% of pairs are retained across languages. Our final evaluation set contains approximately 12K QA pairs across five source languages, each of which passes the procedures for both knowledge gaps (§3.1) and conflicts (§3.2). More details are in Appendix A.5.

For each evaluated model, we retain only QA pairs whose (1) `timestamp` postdates the model's reported knowledge cutoff and (2) when given only the question, the model answers incorrectly in both the source language and English, giving an average retention rate of 23.8%. This reduces the likelihood that a model can answer from parametric knowledge alone (Yeh & Li, 2026) or preferentially follows evidence consistent with its prior beliefs (i.e., confirmation bias; Jin et al. (2024a)). It therefore allows observed response differences to be attributed more directly to the provided context.

[3] https://dumps.wikimedia.org/

4 EXPERIMENT SETUP

4.1 STUDY FORMULATION

Each WALDO instance contains an aligned QA pair, a source-language document, its corresponding English document, a conflict-injected English document, and the revised English answers.

Language	Date	Num	Per	Loc	Org	Misc
French	11.3	20.0	15.9	7.6	10.6	34.6
Chinese	17.5	26.3	8.6	9.7	9.6	28.3
Japanese	17.9	18.4	9.9	8.5	12.0	33.3
Korean	27.9	12.6	14.8	5.1	8.2	31.4
Bengali	14.8	25.4	10.1	6.9	8.8	34.0
Average	17.9	20.5	11.9	7.6	9.8	32.3

Table 1: **Distribution of injected conflict types per language (%). Num:** Number; **Per:** Person; **Loc:** Location; **Org:** Organization.

Context. For each instance, the model receives the source-language document together with either the original English document (in knowledge gap) or its conflict-injected version (in knowledge conflict), presented in randomized order to control for position bias (Liu et al., 2024). We hold this multilingual context fixed while varying *only* the query language. Thus, any systematic difference in model responses across the two queries reflects the effect of the query language.

Response. For each knowledge gap and conflict setting s, given the fixed context $\mathcal{C}_s$ and a query $\mathcal{Q}_\ell$ in language $\ell \in \{\text{src}, \text{en}\}$, model $\mathcal{M}$ produces $\hat{\mathcal{A}}_{\ell,s}$. For each model, since we retain only target facts that postdate its knowledge cutoff, the outputs are less likely to be determined by its parametric knowledge than by the provided context. Thus, if a model is invariant to query language, changing the query language should *not* systematically alter which language-specific account it follows.

> We say that a model exhibits **query-language preference** when it follows the source-language account in response to a source-language query but does not provide the equivalent English answer in response to the English query. Under knowledge gaps, this may manifest as abstention or an unsupported answer to the English query; under knowledge conflicts, it may appear as instead selecting the conflict-injected English answer.

Models. We evaluate eight models that satisfy two criteria: they have publicly reported knowledge cutoffs and context windows of at least 128K tokens to accommodate Wikipedia-length documents. To assess the generality of query-language preference, the suite spans model sizes, families, and knowledge cutoffs: LLAMA-3.1 8B (Grattafiori et al., 2024), QWEN-3 8B (Yang et al., 2025), and AYA-EXPANSE 8B (Dang et al., 2024) are open-weight models; GPT-4O (OpenAI et al., 2024), GPT-4.1 (OpenAI, 2025), DEEPSEEK-V4-PRO (DeepSeek-AI et al., 2026), GROK-4.3,[4] and GPT-5.5 (OpenAI, 2026a) are closed-source models. Per-model details are in Appendix Table 9.

4.2 EVALUATION METRICS

We use two complementary metrics for quantifying whether models recover the original target fact as the query language changes. Answer agreement measures whether a model response matches the reference answer expressed in the query language. Answer probability is the length-normalized likelihood assigned to the answer supported by the source-language document.

Answer Agreement. For each QA pair, we use GPT-5.6-TERRA to determine whether the model response matches the reference answer in the query language. We do not directly compare responses across languages to avoid potentially less reliable cross-lingual model judgments (Fu & Lu, 2026).

For each knowledge gap and conflict setting s, we compute the percentage of QA pairs in $\mathcal{P}_s$ whose model response $\hat{\mathcal{A}}_{\ell,s}$ matches the reference answer $\mathcal{A}_\ell$ under query language ℓ:

$$\text{Agreement}(\ell, s) = \frac{1}{|\mathcal{P}_s|} \sum_{i \in \mathcal{P}_s} \text{Agree}\left(\hat{\mathcal{A}}_{\ell,s}^{(i)}, \mathcal{A}_\ell^{(i)}\right) \times 100, \tag{1}$$

where $\text{Agree}(\cdot, \cdot)$ is 1 if they match and 0 otherwise. We define the query-language agreement gap:

$$\Delta_{\text{agree}}(s) = \text{Agreement}(\mathcal{Q}_{\text{src}}, s) - \text{Agreement}(\mathcal{Q}_{\text{en}}, s). \tag{2}$$

A large positive $\Delta_{\text{agree}}(s)$ indicates that the model more often recovers the original target fact when queried in the source language than in English, consistent with query-language preference.

[4] https://docs.x.ai/developers/models/grok-4.3

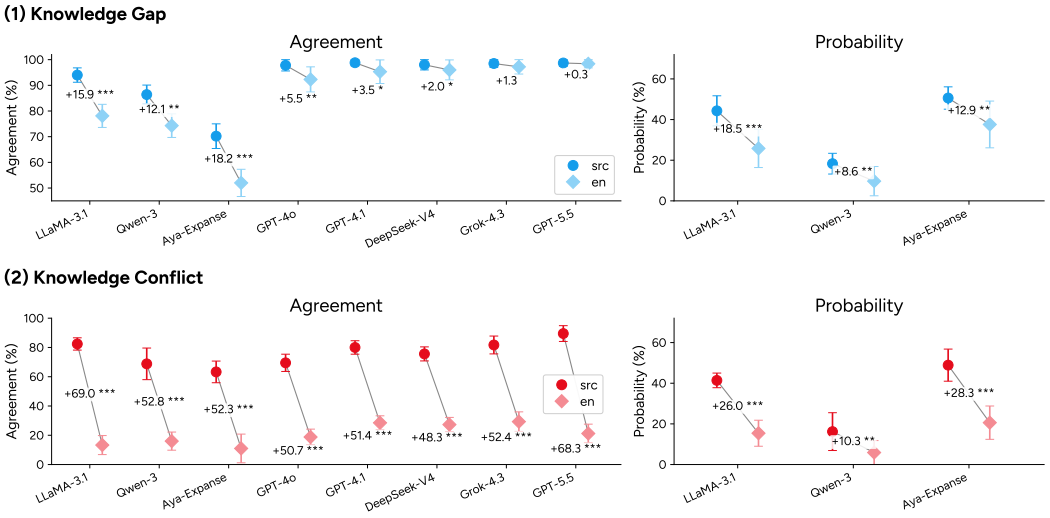


Figure 2: **Answer agreement and probability under knowledge gap and conflict settings.** A larger positive Δ(src-en) denotes stronger query-language preference. Error bars are 95% bootstrap CIs. We use McNemar's test for agreement differences and the Wilcoxon signed-rank test for probability differences, with Bonferroni correction for multiple comparisons. *: significant with $p < .05$; **: $p < .01$; ***: $p < .001$; unmarked: not significant. See Appendix B.2 for per-language results.

Answer Probability. Answer agreement captures whether a decoded response matches the source-supported reference answer, but not how *strongly* the model favors that answer. Thus, we compute the length-normalized conditional likelihood of the source-supported answer $\mathcal{A}_{\text{src}} = (t_1, \ldots, t_k)$ under each query language ℓ:

$$P_\theta(\mathcal{A}_{\text{src}} \mid \mathcal{Q}_\ell, \mathcal{C}_s) = \left[\prod_{m=1}^{k} p_\theta(t_m \mid \mathcal{Q}_\ell, \mathcal{C}_s, t_{<m}) \right]^{1/k}, \tag{3}$$

where p_θ is the model's next-token distribution, and $t_{<m} = (t_1, \ldots, t_{m-1})$. Length normalization makes scores comparable across answers of different lengths. We average this score across evaluated QA pairs (yielding Prob) and report the source-answer likelihood shift:

$$\Delta_{\text{prob}}(s) = \text{Prob}(\mathcal{Q}_{\text{src}}, s) - \text{Prob}(\mathcal{Q}_{\text{en}}, s). \tag{4}$$

A positive $\Delta_{\text{prob}}(s)$ means that the model assigns greater likelihood to the source-supported answer when queried in the source language than in English. Conversely, more comparable likelihoods across query languages could help systems surface cross-lingual disparities.

5 Diagnosis Results

Figure 2 presents our diagnosis for query-language preference.[5] We highlight four findings below.

Closed-source models largely overcome query-language preference under knowledge gaps. As shown in Figure 2 (*top*), closed-source models show little to no statistically significant difference in answer agreement across query languages. In contrast, open-weight models exhibit significant agreement gaps: 15.9% for LLaMA-3.1, 12.1% for Qwen-3, and 18.2% for Aya-Expanse. Notably, among the three, Aya-Expanse shows the largest gap despite explicit training on diverse languages and multilingual tasks (Dang et al., 2024). Answer probability shows a similar pattern: all three open-weight models assign significantly different likelihoods to the source-supported answer across query languages. These findings suggest that, when the required knowledge to correctly answer a query is available only in one language, closed-source models generally recover it regardless of the query language, whereas the smaller open-weight models often fail to.

[5] While our primary experiments use a controlled two-document context (§4.1) to isolate the effect of query language, Appendix B.1 shows that the results generalize when an external retrieval component is added.

Knowledge conflicts amplify query-language preference. Under knowledge conflicts (shown in Figure 2, *bottom*), query-language preference becomes substantially more pronounced. All models exhibit statistically significant answer-agreement gaps between source-language and English queries, with an average gap of 55.7% across models. The answer-probability gap likewise increases relative to the knowledge gap setting, by 7.5 *pp* for LLAMA-3.1, 1.7 *pp* for QWEN-3, and 15.4 *pp* for AYA-EXPANSE. When disaggregated by source language (Appendix B.2), probability gaps are largest for lower-resource languages such as Bengali and generally decline with language-resource availability, reaching their lowest values for French. Overall, even closed-source models that show weak preference under knowledge gaps preferentially follow the query-language-matched document when evidence conflicts across languages.[6]

Query-language preference persists when the document-language asymmetry is reversed. By design, WALDO is asymmetric: the original target fact is in the source-language document, while it is absent (gap) or contains an incompatible version (conflict) in the English document. Is the query-language preference observed in WALDO driven merely by the asymmetric construction of its QA pairs? To test this, we reverse the knowledge conflict condition: the English document now contains the original target fact, whereas the source-language document contains the conflicting account. We use GPT-5.6-TERRA to translate and incorporate each fact into the appropriate document language (see Prompt A.9 of Appendix E). As shown in Table 2, all models continue to exhibit strong query-language preference after the document roles are swapped, with the direction of the effect correspondingly reversed. This result indicates that the strong preference observed under knowledge conflicts is not an artifact of WALDO's particular asymmetry but generalizes across both directions.

Query-language preference is not explained by measured alternative factors. Could the observed preference reflect other properties of the data rather than the query language itself? We investigate this by testing four potential confounders—pretraining effect, embedding similarity, surface cues, and query-topic locality—but find no significant associations with query-language preference metrics (Appendix C). Together, the results suggest that the measured effect of query-language preference is not accounted for by the examined confounders.[7]

Model	Agreement (%)			Probability (%)		
	src	en	Δ_{agree}	src	en	Δ_{prob}
LLAMA-3.1	21.0	86.3	-65.3^{***}	41.4	15.4	26.0^{***}
QWEN-3	28.7	81.6	-52.9^{***}	7.9	4.9	2.9^{*}
AYA-EXPANSE	24.3	69.4	-45.1^{***}	17.9	15.4	2.4^{*}
GPT-4O	15.6	67.7	-52.1^{***}	–	–	–
GPT-4.1	25.3	77.9	-52.6^{***}	–	–	–
DEEPSEEK-V4-PRO	22.5	72.6	-50.1^{***}	–	–	–
GROK-4.3	27.1	80.9	-53.8^{***}	–	–	–
GPT-5.5	20.4	76.5	-56.1^{***}	–	–	–

Table 2: **Answer agreement and probability gaps under knowledge conflict for swapped roles.** Same statistical test performed as in Figure 1. *: significant with $p < .05$; ***: $p < .001$. No probabilities for closed-source models (–).

6 MITIGATION FOR KNOWLEDGE CONFLICTS

Under conflicting evidence across languages, models *strongly* favor the document in the query language (§5). As an initial exploration of how this behavior may be mitigated, we study two targeted approaches (§6.1) and evaluate their effectiveness in reducing query-language preference (§6.2).

6.1 MITIGATION APPROACHES

Query-language Preference Head Ablation (QPHA). Prior work suggests that individual attention heads can mediate distinct information sources (e.g., context versus parametric knowledge (Jin et al., 2024b; Zhu et al., 2026)), and that head ablation can mitigate conflict-driven behavior (Jin et al., 2024c; Li et al., 2025b). Building on this, we investigate whether query-language preference is associated with a localized set of attention heads and whether ablating these heads mitigates the behavior. We identify candidate heads by adapting the comparison underlying our metrics (§4.2).

[6] Consistent with this account, models assign higher likelihood to the conflicting English-supported answer under English queries than source-language queries; see Appendix B.3. Also, as a control, adding the English fact to the English document reduces the preference gap to near zero (Appendix B.5).

[7] Appendix B.4 reports additional evidence from F1 score gap, citation attribution, and response-language analyses, which corroborates the answer agreement and probability results.

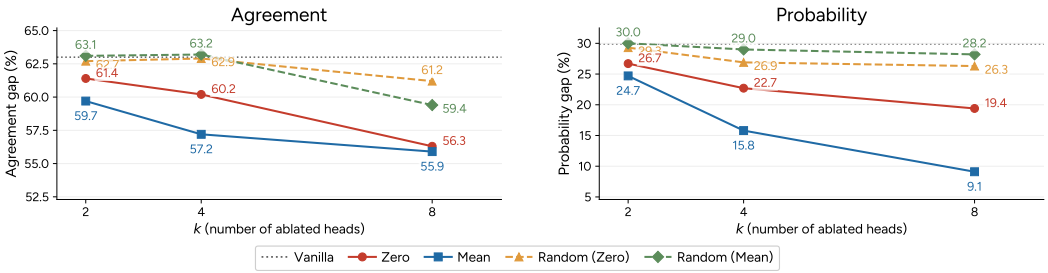


Figure 3: **Answer agreement and probability gaps with varying k.** Both gaps shrink as more attention heads are ablated (i.e., increasing k), and QPHA yields a larger reduction than the random control. Shown for LLAMA-3.1; see Appendix D.4 for other models.

For each conflict instance in WALDO, we score the source-supported answer $\mathcal{A}_{\text{src}}$ and the conflicting English-supported answer $\mathcal{A}'_{\text{en}}$ by their length-normalized log-likelihood under source-language and English queries. We then use gradient–activation attribution (Michel et al., 2019; Jiang et al., 2025; Liu et al., 2026) to measure each attention head's contribution to these likelihoods $\mathcal{S}_\ell(\mathcal{A})$:

$$\mathcal{S}_\ell(\mathcal{A}) = \frac{1}{k}\sum_{i=1}^{k} \log p_\theta(t_i \mid \mathcal{Q}_\ell, \mathcal{C}_{\text{conflict}}, t_{<i}), \qquad \text{Attr}_{l,h}(\mathcal{A}, \ell) = \sum \mathcal{O}^{(l,h)} \odot \frac{\partial \mathcal{S}_\ell(\mathcal{A})}{\partial \mathcal{O}^{(l,h)}}, \tag{5}$$

where $\mathcal{O}^{(l,h)}$ is the output of head h in layer l, and the summation aggregates over its activation dimensions. Here, $\mathcal{A} = (t_1, \ldots, t_k)$ and $t_{<i} = (t_1, \ldots, t_{i-1})$ denotes the preceding answer tokens.

We identify query-language preference heads using the following contrast:

$$\text{QPHA}_{l,h} = [\text{Attr}_{l,h}(\mathcal{A}_{\text{src}}, \text{src}) - \text{Attr}_{l,h}(\mathcal{A}'_{\text{en}}, \text{src})] - [\text{Attr}_{l,h}(\mathcal{A}_{\text{src}}, \text{en}) - \text{Attr}_{l,h}(\mathcal{A}'_{\text{en}}, \text{en})]. \tag{6}$$

A large positive score indicates a head's attribution shifts toward the source-supported answer under source-language queries and the conflicting English-supported answer under English queries. We average this score across conflict instances and select the k heads with the largest positive, statistically significant mean scores (under a two-sided Wilcoxon signed-rank test).

To examine the trade-off between intervention sparsity and strength (Kang et al., 2025), we evaluate $k \in \{2, 4, 8\}$ with two ablation modes: **Zero**, which sets the selected head outputs to zero (Michel et al., 2019), and **Mean**, which replaces each output with its sequence-wise mean activation, reducing potential out-of-distribution effects of zero ablation (Wang et al., 2023). For each, we compare against a **Random** control that ablates k randomly selected heads (Crosbie & Shutova, 2025).

QLORA. Whereas QPHA is a localized inference-time intervention, we ask whether a training-time adaptation can more effectively mitigate query-language preference by encouraging models to recover the original target fact regardless of query language. We fine-tune models using Quantized Low-Rank Adapters (QLORA; Dettmers et al., 2023) with two customized objectives, both based on the length-normalized log-likelihood $\mathcal{S}_\ell(\mathcal{A})$ defined in Eq. 5:

- **Promote:** We promote the likelihood of the original English answers under English queries:
$$\mathcal{L}_{\text{promote}} = -\mathcal{S}_{\text{en}}(\mathcal{A}_{\text{en}}). \tag{7}$$
- **Contrastive:** We additionally set models to forget the conflicting English-supported answer and retain the source-supported answer under source-language queries (Nam et al., 2025):
$$\mathcal{L}_{\text{contrastive}} = \underbrace{\mathcal{S}_{\text{en}}(\mathcal{A}'_{\text{en}})}_{\text{Forget}} - \underbrace{\mathcal{S}_{\text{src}}(\mathcal{A}_{\text{src}})}_{\text{Retain}} - \lambda \underbrace{\mathcal{S}_{\text{en}}(\mathcal{A}_{\text{en}})}_{\text{Promote}}, \tag{8}$$
with the Forget$-$Retain term clipped to $[-20, 20]$ for gradient stability.

As both methods require access to model internals, we evaluate them only on the three open-weight models. We further restrict to knowledge-conflict instances that exhibit query-language preference, yielding 5.1K QA pairs in total across models (1.9K revisions; 52.3% retain rate). For a fair comparison, we allocate 20% of revisions to training and development: QLORA uses it for training and model selection while QPHA uses it for head identification. We use the remaining 80% for evaluating both methods. See Appendix D.1 for implementation details.

6.2 Results & Analyses

QPHA reduces query-language preference. In Table 3, we observe that QPHA ($k = 8$) reduces both agreement and probability gaps across models. Zero and mean ablation each reduce the agreement gap by 4.8 *pp* on average, exceeding their respective random controls (1.3 *pp* for zero and 2.1 *pp* for mean). Similarly for the probability gap, mean ablation of QPHA-selected heads yields the largest reduction (9.2 *pp*), followed by zero ablation (5.3 *pp*), again both exceeding their respective random controls (1.2 *pp* for mean and 2.1 *pp* for zero). These improvements over random ablation suggest that QPHA identifies heads more strongly associated with query-language preference. Across most model–ablation-mode settings, ablating these heads reduces the preference gap, providing an initial proof-of-concept for a mechanistic head-level intervention.

Small-scale QLoRA adaptation yields larger reductions. Fine-tuning QLoRA adapters with targeted objectives further reduces query-language preference more than inference-time QPHA. The promote objective closes 58.0% of the agreement gap and 19.5% of the probability gap, while the contrastive objective pushes this further to 59.0% and 20.7% each. Overall, using only 10% of the data for training, QLoRA substantially attenuates query-language preference under knowledge conflicts, which also further generalizes to the swapped-roles setting (Appendix D.3).

Approach	LLaMA-3.1		Qwen-3		Aya-Expanse	
	Δ_{agree}	Δ_{prob}	Δ_{agree}	Δ_{prob}	Δ_{agree}	Δ_{prob}
Vanilla	63.0	29.9	59.0	12.1	58.8	33.5
QPHA ($k = 8$)						
Random (Zero)	61.2	26.3*	57.0	12.1	58.6	30.8*
Random (Mean)	59.4*	28.2	57.5	11.8	57.5	31.8
Zero	56.3**	19.4***	54.3*	10.5	55.7*	29.7*
Mean	55.9**	9.1***	55.4*	10.1*	55.0*	28.7**
QLoRA						
Promote	2.3***	4.3***	4.3***	7.7**	0.3***	4.9***
Contrastive	1.5***	6.4***	1.4***	5.3***	0.9***	1.7***

Table 3: **Mitigation results on answer agreement and probability gaps.** For both, lower is better (↓). We perform Wilcoxon signed-rank test against Vanilla, with Bonferroni correction. *: significant with $p < .05$; **: $p < .01$; ***: $p < .001$; unmarked: not significant. Appendix D.2 reports full results for $k \in \{2, 4, 8\}$ and each query language.

QPHA's effect grows with the number of ablated heads. As illustrated in Figure 3, zero and mean ablation increasingly reduce agreement and probability gaps as k grows from 2 to 8. This pattern supports that QPHA identifies heads associated with query-language preference: ablating more highly ranked heads yields larger reductions. Notably, ablating $k = 2$ heads already significantly reduces the gap relative to the vanilla baseline ($p < .05$), showing that the preference can be reduced through a small head-level intervention. List of heads is in Appendix D.5.

Which one to use? QPHA provides a lightweight, training-free intervention that directly incorporates the mechanism underlying query-language preference, and is effective with only a few ablated heads. Yet, ablating more heads may affect model behavior beyond query-language preference. In contrast, QLoRA achieves substantially larger reductions, but requires training and knowledge of the desired and conflicting answers across documents to define its objectives. Given these trade-offs, future work could develop more robust approaches that retain the advantages of both methods.

Conclusion & Discussion

We study query-language preference in multilingual systems under two forms of cross-lingual knowledge disparities. Using WALDO, we find that models largely overcome this preference under knowledge gaps, but exhibit it strongly under conflicts, which can be effectively mitigated through a mechanistic intervention or LoRA training. More broadly, addressing query-language preference is important for *equitable* multilingual information access. If the information a system surfaces depends on a user's query language, it may confine users to language-bounded evidence, reinforcing filter bubbles (Flaxman et al., 2016) and information cocoons (Sunstein, 2018). An important next step is to evaluate the real-world impact in deployment, including how source selection affects users' comprehension, trust, and information access. Future benchmarks should also extend beyond Wikipedia to multilingual sources with different authority structures and update dynamics, such as academic papers. We hope that WALDO supports research on robust cross-lingual knowledge transfer and that our mitigation approaches serve as useful tools for studying and mitigating source-selection biases beyond query language, including biases related to formality (Li et al., 2024b), relevance (Wan et al., 2024), recency (Fang et al., 2025), and source origin (Schuster et al., 2026).

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets including translation, implement methods, propose or refine hypotheses, provide feedback on the methodology, and conduct qualitative analyses. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, design experiments, or interpret results. Additionally, we used generative AI tools to formulate questions for surveys, modify figures, propose the name of our benchmark (WALDO), and assist with writing and presentation, including editing the manuscript to improve clarity and readability; making stylistic adjustments; and resolving layout issues such as widows and orphans. The authors independently reviewed all AI-generated hypotheses, feedback, and qualitative analyses and verified the LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Research on query-language preference in multilingual systems raises fairness concerns, particularly for speakers of low-resource and underrepresented languages. When models provide different answers to semantically equivalent queries solely because they are expressed in different languages, they may reinforce existing disparities in access to reliable information. Our work aims to identify, measure, and mitigate such disparities rather than reinforce them. Nevertheless, the proposed mitigation methods should be evaluated carefully across languages and domains before deployment, as reducing one form of language preference may have uneven effects on answer quality, source faithfulness, or linguistic and cultural coverage.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our data construction, measurement, and mitigation experiments. We document the construction of WALDO in Section 3 and provide additional implementation details in Appendix A. We specify our evaluation metrics, statistical tests, and experimental settings in Section 4 and Appendix D.1. We will release both the full and evaluation set of WALDO, code, experiment configurations, and model checkpoints required to reproduce our measurement pipeline, mitigation experiments, and reported analyses.

REFERENCES

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan. The cross-lingual cost: Retrieval biases in RAG over Arabic-English corpora. In Kareem Darwish, Ahmed Ali, Ibrahim Abu Farha, Samia Touileb, Imed Zitouni, Ahmed Abdelali, Sharefah Al-Ghamdi, Sakhar Alkhereyf, Wajdi Zaghouani, Salam Khalifa, Badr AlKhamissi, Rawan Almatham, Injy Hamed, Zaid Alyafeai, Areeb Alowisheq, Go Inoue, Khalil Mrini, and Waad Alshammari (eds.), *Proceedings of The Third Arabic Natural Language Processing Conference*, pp. 69–83, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-352-4. doi: 10.18653/v1/2025.arabicnlp-main.6. URL https://aclanthology.org/2025.arabicnlp-main.6/.

Amy Ross Arguedas, Sayan Banerjee, Camila Mont'Alverne, Benjamin Toff, Richard Fletcher, and Rasmus Kleis Nielsen. News for the powerful and privileged: How misrepresentation and underrepresentation of disadvantaged communities undermine their trust in news. 2023.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1075–1084, 2012.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo,

Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. An expanded massive multilingual dataset for high-performance language technologies (HPLT). 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. 17452–17485, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.854. URL https://aclanthology.org/2025.acl-long.854/.

Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms, 2025. URL https://arxiv.org/abs/2506.08500.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

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/.

Joy Crosbie and Ekaterina Shutova. Induction heads as an essential mechanism for pattern matching in in-context learning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5049–5111, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.283. URL https://aclanthology.org/2025.findings-naacl.283/.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces, 2020. URL https://arxiv.org/abs/2007.14435.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao

Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. Do large language models favor recent content? a study on recency bias in llm-based reranking. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, pp. 85–94, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722189. doi: 10.1145/3767695.3769493. URL https://doi.org/10.1145/3767695.3769493.

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/.

Elena Filatova. Directions for exploiting asymmetries in multilingual wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)*, pp. 30–37, 2009.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

Xiyan Fu and Wei Lu. When languages disagree: Self-evolving multilingual llm judges, 2026. URL https://arxiv.org/abs/2606.08092.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025. URL https://arxiv.org/abs/2502.21228.

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.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. URL https://arxiv.org/abs/2310.10482.

Scott A Hale. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, pp. 99–108, 2014.

Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599, 1988.

Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*, pp. 11–20, 2009.

Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 291–300, 2010.

Brent Jaron Hecht. *The mining and application of diverse cultural perspectives in user-generated content*. PhD thesis, Northwestern University, 2013.

Seongtae Hong, Youngjoon Jang, Jungseob Lee, Seungyoon Lee, and Heuiseok Lim. Lamar: An open language-aware multilingual alignment reranker, 2026. URL https://arxiv.org/abs/2607.22042.

Wenshuai Huo, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Yichong Huang, Hui Wang, and Bing Qin. Breaking language preference in multilingual RAG via language-controllable retrieval and language-agnostic reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7579–7589, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.374. URL https://aclanthology.org/2026.findings-acl.374/.

Dieuwke Hupkes and Nikolay Bogoychev. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages, 2025. URL https://arxiv.org/abs/2504.10356.

Youngjoon Jang, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Mlaire: Multilingual language-aware information retrieval evaluation protocal, 2026. URL https://arxiv.org/abs/2605.07249.

Xinyan Jiang, Hang Ye, Yongxin Zhu, Xiaoying Zheng, Zikang Chen, and Jun Gong. HICD: Hallucination-inducing via attention dispersion for contrastive decoding to mitigate hallucinations in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7764–7786, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.405. URL https://aclanthology.org/2025.findings-acl.405/.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pp. 2627–2638, 2024a.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466/.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Isaac Johnson and Emily Lescak. Considerations for multilingual wikipedia research. *arXiv preprint arXiv:2204.02483*, 2022.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6282–6293, 2020.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding, 2025. URL https://arxiv.org/abs/2503.06287.

Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh. Linguistic nepotism: Trading-off quality for language preference in multilingual RAG. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=MDfCV0UHSb.

Sangeet Kumar. *The digital frontier: Infrastructures of control on the global web*. Indiana University Press, 2021.

Seungyoon Lee, Minhyuk Kim, Seongtae Hong, Youngjoon Jang, Dongsuk Oh, and Heuiseok Lim. CLEAR: Cross-lingual enhancement in retrieval via reverse-training. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 347–362, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.13. URL https://aclanthology.org/2026.acl-long.13/.

Bryan Li, Samar Haider, and Chris Callison-Burch. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. 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. 3855–3871, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.213. URL https://aclanthology.org/2024.naacl-long.213/.

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 *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4215–4241, 2025a.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=0cEZyhHEks.

Jiahuan Li, Yiqing Cao, Shujian Huang, and Jiajun Chen. Formality is favored: Unraveling the learning preferences of large language models on data with conflicting knowledge, 2024b. URL https://arxiv.org/abs/2410.04784.

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/.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. Focusing on language: revealing and exploiting language attention heads in multilingual large language models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i38.40492. URL https://doi.org/10.1609/aaai.v40i38.40492.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021a. doi: 10.1162/tacl_a_00433. URL https://aclanthology.org/2021.tacl-1.82/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Paolo Massa and Federico Scrinzi. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pp. 1–9, 2012.

Paul Michel, Omer Levy, and Graham Neubig. *Are sixteen heads really better than one?* Curran Associates Inc., Red Hook, NY, USA, 2019.

Marc Miquel-Ribé and David Laniado. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in physics*, 6:54, 2018.

Marc Miquel-Ribé and David Laniado. The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, pp. 1–4, 2020.

Kaijie Mo, Siddhartha Venkatayogi, Chantal Shaib, Ramez Kouzy, Wei Xu, Byron C Wallace, and Junyi Jessy Li. Faithfulness vs. safety: Evaluating LLM behavior under counterfactual medical evidence. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37053–37081, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1847. URL https://aclanthology.org/2026.findings-acl.1847/.

Bruno Nadalic Sotic and Jaap Kamps. Information seeking behavior in llm-based rag: Mental models and missing information. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 2684–2695, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808537. URL https://doi.org/10.1145/3805712.3808537.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, April 2025. Accessed: 2026-09-04.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, April 2026a. Accessed: 2026-09-04.

OpenAI. GPT-5.6: Frontier Intelligence That Scales with Your Ambition. https://openai.com/index/gpt-5-6/, 2026b. Accessed: 2026-09-05.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won

Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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/.

Jeonghyun Park, Byeongjeong Kim, Seojin Hwang, and Hwanhee Lee. Enhancing multilingual RAG systems with debiased language preference-guided query fusion. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27116–27136, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1353. URL https://aclanthology.org/2026.findings-acl.1353/.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*,

2023. URL https://arxiv.org/abs/2305.15425.

Robert Phillipson. *Linguistic imperialism*. oxford Un. Press, 1992.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. On the consistency of multilingual context utilization in retrieval-augmented generation. In David Ifeoluwa Adelani, Catherine Arnett, Duygu Ataman, Tyler A. Chang, Hila Gonen, Rahul Raja, Fabian Schmidt, David Stap, and Jiayi Wang (eds.), *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 199–225, Suzhuo, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-345-6. doi: 10.18653/v1/2025.mrl-main.15. URL https://aclanthology.org/2025.mrl-main.15/.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Xu Jinan, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14291–14310, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.700. URL https://aclanthology.org/2026.findings-acl.700/.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive question answering task. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 697–716, 2026.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. Locating information gaps and narrative inconsistencies across languages: A case study of LGBT people portrayals on Wikipedia. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6747–6762, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.384. URL https://aclanthology.org/2024.emnlp-main.384/.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1357. URL https://aclanthology.org/2026.acl-long.1357/.

Shilad W Sen, Heather Ford, David R Musicant, Mark Graham, OS Keyes, and Brent Hecht. Barriers to the localness of volunteered geographic information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 197–206, 2015.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642459. URL https://doi.org/10.1145/3613904.3642459.

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/.

Kaiser Sun, Fan Bai, and Mark Dredze. Task matters: Knowledge requirements shape LLM responses to context–memory conflict. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4154–4176, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.202. URL https://aclanthology.org/2026.findings-acl.202/.

Cass R Sunstein. # republic: Divided democracy in the age of social media. In *# Republic*. Princeton university press, 2018.

John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? 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. 7468–7484, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

Dan Wang, Guozhao Mo, Yafei Shi, Cheng Zhang, Bo Zheng, Boxi Cao, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. All languages matter: Understanding and mitigating language bias in multilingual RAG. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7441–7455, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.338. URL https://aclanthology.org/2026.acl-long.338/.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ptvV5HGTNN.

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.

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*, pp. 2487–2491, 2024.

Samuel Yeh and Sharon Li. How retrieved context shapes internal representations in RAG. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15446–15468, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.758. URL https://aclanthology.org/2026.findings-acl.758/.

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/.

Jiaqi Zhao, Qiang Huang, Haodong Chen, Xiaoxing You, and Jun Yu. When abundance conceals weakness: Knowledge conflict in multilingual models, 2026a. URL https://arxiv.org/abs/2601.07041.

Tianzhe Zhao, Jiaoyan Chen, Shuxiu Zhang, Haiping Zhu, Qika Lin, and Jun Liu. Exploring knowledge conflicts for faithful llm reasoning: Benchmark and method. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 2464–2474, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809650. URL https://doi.org/10.1145/3805712.3809650.

Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, and Xiaohui Liang. Focus directions make your language models pay more attention to relevant contexts, 2026. URL https://openreview.net/forum?id=h7Tm6dSWWp.

APPENDIX

A DETAILS FOR CREATING WALDO

A.1 SOURCE LANGUAGES

We present detailed statistics for each source language in Table 4.

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
	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	80M	437,373
Japonic	Japanese	Japanese	agglutinative	SOV	mid	128M	1,133,444

Table 4: **Characteristics of source 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).

A.2 HEURISTICS & LLM ANNOTATION

Heuristics. We conduct a manual audit of a sample of sourced Wikipedia revisions to develop a filtering procedure for our controlled evaluation. Based on this audit, we define a suite of heuristic rules that retain only revisions meeting our criteria. We apply the following rules sequentially:

- **Required added text:** Drop a row when the added text content is missing, non-string, or empty after trimming whitespace.
- **Remove bot edits:** Drop a row if either the edit comment or contributor username contains a language-specific bot term.
- **Remove maintenance edits:** Drop a row if its comment matches language-specific maintenance keywords such as typo/spelling fixes, links, references, formatting, categories, templates, cleanup, or wikification.
- **Remove reverts:** Drop a row if its comment contains language-specific revert indicators, including terms like `revert`, `undo`, `rv`, `vandal`, or their equivalents in other source languages.
- **Remove high-symbol, low-content edits:** Drop a row if either comment or the added text has at least 12 characters, at least 60% non-content symbols, and no more than 8 alphanumeric or source-script characters.
- **Remove markup-heavy additions:** Drop a row if the added text contains wiki or HTML markup patterns and either at least half of its characters are non-content symbols or it contains four or more pipe characters (|).
- **Remove repetition:** Drop a row if the added text contains a character repeated at least 8 consecutive times.
- **Require punctuation:** Drop a row if the added text does not contain a language-specific sentence terminator.
- **Require source-language text:** Use LINGUA language detector tool[8] on the added text and drop a row if it is undetectable or detected as a language other than the current source language.
- **Remove length outliers:** Compute character-length quartiles of non-empty added text content, then drop rows outside Tukey's $Q1 - 1.5 \times IQR$ to $Q3 + 1.5 \times IQR$ bounds (Tukey, 1977).

LLM Annotation. As shown in Prompts A.1–A.2 (Appendix E), we instruct the model to provide a binary verdict and rationale for whether: (1) the added text introduces at least one factual statement without removing existing information, and (2) the added text is absent from the latest version of the corresponding English Wikipedia document, identified using the same Wikidata QID. To determine

[8] https://pypi.org/project/lingua-language-detector/

whether the added text appears in the latest source-language Wikipedia document, we retrieve the document and perform exact string matching.

We report the retention rate for each source language after (1) heuristic filtering, (2) LLM annotation in A.1, and (3) LLM annotation in A.2 in Table 5. This yields approximately 4.8K retained revisions across the five source languages.

Additional LLM Annotation Variants. We test five additional annotation variants on 100 randomly sampled revisions and manually audited their outputs. These analyses are used to qualitatively refine the LLM Annotation prompt:

- **Adding the first paragraph.** We add the first paragraph of each Wikipedia article as additional context, because it often provides a high-level overview of the topic. This variant increases the retention rate from 8% to 9%, corresponding to one additional retained revision. However, manual inspection found that this additional example did not satisfy our filtering criteria. We therefore decide not to include the first paragraph in the final prompt.
- **Providing Wikipedia context.** We add a brief definition of Wikipedia to provide additional context, which led to no observable change in the retention rate (8%), so we omit this definition.
- **Removing the constraint on deleted text.** We test removing the instruction that the deleted text must not remove meaningful information, to accommodate revisions that replace an incorrect fact with contradictory factual information. However, without this constraint, the LLM fails to retain some revisions that simultaneously add a standalone fact and remove unrelated content. We therefore retain the constraint in the final prompt.
- **Allowing general knowledge.** We instruct the LLM to use both the provided context and general commonsense knowledge during the annotation. In the manual audit, this helped filter revisions containing irrelevant jokes or degenerate edits, thus we retain this instruction.
- **Translating the prompt.** We translate the full prompt into the source language for revisions in that language. This variant did not yield an observable overall improvement and missed several revisions judged meaningful in the manual audit. We use the English prompt for all languages.

Language	Original	(1) Heuristics	(2) LLM Annotation (A.1)	(3) LLM Annotation (A.2)
French	229,438	42,446 (18.5%)	3,917 (1.7%)	504 (0.2%)
Chinese	251,136	107,486 (42.8%)	6,160 (2.5%)	1,227 (0.5%)
Japanese	257,041	112,841 (43.9%)	7,778 (3.0%)	1,125 (0.4%)
Korean	465,500	68,894 (14.8%)	5,293 (1.1%)	943 (0.2%)
Bengali	725,835	172,023 (23.7%)	5,702 (7.9%)	1,043 (0.1%)
Average		(28.7%)	(3.2%)	(0.3%)

Table 5: **Retention rates by each source language after heuristic filtering and LLM annotation.** We compute each retention rate relative to the total number of revisions in the original set.

INFOGAP Framework. As an alternative to LLM annotation for determining whether added text appears in the latest English Wikipedia document, we consider INFOGAP (Samir et al., 2024). INFOGAP identifies information overlaps and gaps between documents in different languages through three stages: (1) fact decomposition, which extracts atomic facts from each document; (2) multilingual alignment, which aligns facts across language versions; and (3) verification, which determines whether aligned facts are semantically equivalent.

Given an English document, we decompose its content into a set of atomic facts using the fact-decomposition prompt of Samir et al. (2024) and the same model used for LLM annotation, GPT-5.6-TERRA. For multilingual alignment, we embed the decomposed English facts and the added text using LABSE (Feng et al., 2022). For each English fact, we compute a hubness score, which is the average cosine similarity to its top-k nearest neighbors, and select the highest-scoring facts as alignment candidates. We then use GPT-5.6-TERRA to verify each candidate. If at least one candidate fact survives verification, we label the added text as "match"; otherwise, as "no match."

Compared with our LLM annotation approach, INFOGAP produces a slightly lower retention rate. This difference primarily arises when the English document contains only part of the information in

the added text. For example, consider the Korean addition "BMW IX는 BMW에서 제조한 크로스오버 SUV이다." (*"BMW iX is a crossover SUV manufactured by BMW."*). The corresponding English document describes the BMW iX as an electric sport utility vehicle but does not state that it is a crossover SUV. Our LLM annotation labels this case as "no match," whereas INFOGAP labels it as "match." After qualitatively examining several such revisions, we decide to use LLM annotation for our final pipeline because it better handles these partial-overlap cases. The two methods agree on 45.5% of "match" labels, 97.1% of "no match" labels, and 87.9% of labels overall.

Examples of Excluded Revisions. Table 6 presents qualitative examples of revisions excluded during the Heuristics & LLM Annotation stage. Most revisions are excluded because they do not meet one or more of the rules. The resulting set is intentionally a highly curated subset rather than a representative sample of Wikipedia revisions, and the low retention rate reflects the rarity of revisions that support our controlled evaluation design.

Stage	Added text	Removed text	Description
Heuristics	* 1895 * 1896 * 1914 * 1919 * 1937 * 1939 * 1945 * 1953 * 1954 * 1963 * 1973 * 1979	*1895 *1896 *1914 *1919 *1937 *1939 *1945 *1953 *1954 *1963 *1973 *1979	**Remove bot edits**; Username with "TedBot" and comment as "Bot: Standardizing template names and styles"
	bgcolor="#ccffcc"|align=left| |bgcolor=#87CEEB|AFC|- bgcolor="#ffffcc"|align=left|		**Remove reverts**
	[...]		**Remove high-symbol, low-content edits**
	bgcolor="#ccffcc"|align=left| |bgcolor=#87CEEB|AFC|- bgcolor="#ffffcc"|align=left|		**Remove markup-heavy additions**
	# # # # # # # # [...]		**Remove repetition**
LLM Annotation	Yanggu County(楊口郡)is	Yanggu County(楊口郡,) is,	Removed two commas without adding any meaningful information
	사용한다. 하지만 러시아가 2030년까지 벨라루스를 합병한다고 하니 오래 가지는 못할 것 같다.	사용한다.	Added text translates to "*However, given reports that Russia plans to annex Belarus by 2030, it seems unlikely to last long,*" which expresses speculation rather than a verifiable fact

Table 6: **Examples of excluded revisions during the Heuristics & LLM Annotation stage.**

A.3 QA GENERATION

We assess LLM-generated QA pairs with XCOMET-XL (Guerreiro et al., 2023) between source-language and English questions $\mathcal{Q}$ and answers $\mathcal{A}$. The average scores are 0.960 for $\mathcal{Q}$ and 0.983 for $\mathcal{A}$, indicating that the content is well-preserved across languages (Table 7).

Language	xCOMET($\mathcal{Q}_{\text{src}}$, $\mathcal{Q}_{\text{en}}$)	xCOMET($\mathcal{A}_{\text{src}}$, $\mathcal{A}_{\text{src}}$)
French	0.971	0.990
Chinese	0.956	0.983
Japanese	0.959	0.982
Korean	0.953	0.979
Bengali	0.970	0.985
Average	0.960	0.983

Table 7: **xCOMET-XL scores by source language.** We evaluate the translation quality of source-language questions ($\mathcal{Q}_{\text{src}}$) and answers ($\mathcal{A}_{\text{src}}$), against their English counterparts, $\mathcal{Q}_{\text{en}}$ and $\mathcal{A}_{\text{en}}$. Higher scores indicate higher translation quality.

A.4 QUALITATIVE EXAMPLES OF KNOWLEDGE CONFLICTS

Table 8 presents representative QA pairs for each conflict type, contrasting the original target fact with the inconsistent version injected into the English document. The injected alternatives are designed to remain plausible, within a reasonable range of possible values (e.g., "*TV Tokyo*" → "*Nippon TV*" in one of the Japanese examples).

Lang.	Type	Title	Target fact	$\mathcal{Q}_{\text{en}}$	$\mathcal{A}_{\text{en}}$	$\mathcal{A}'_{\text{en}}$	Conflicting fact
French	Number	Higashi-ku, Fukuoka	* Kane Tanaka (1903-), supercentenaire japonaise et doyenne de l'humanité depuis 2018 et troisième personne vérifiée la plus âgée de tous les temps.	What was Kane Tanaka's all-time rank among verified oldest people?	Third	Fourth	In 2018, Kane Tanaka was recognized as the fourth verified oldest person of all time.
	Misc	NOAAS Okeanos Explorer	Fichier:Hyocrinida Okeanos 2021.jpglUn crinoïde (Hyocrinida) observé dans l'Atlantique nord en 2021.	What kind of organism was observed in the North Atlantic in 2021 by Okeanos Explorer?	A crinoid (Hyocrinida)	A sea cucumber (Holothuroidea)	A sea cucumber (Holothuroidea) was observed in the North Atlantic in 2021 by Okeanos Explorer.
Chinese	Date	Ruud van Nistelrooy	*2024年10月，曼联主教练坦哈格被解雇，云尼斯达莱以助教身份临危受命成暂代主帅，首次领军曼联于联赛杯大胜李斯特城5:2。	When did Ruud van Nistelrooy take over as Manchester United's interim manager after Ten Hag was fired?	October 2024	September 2024	He was originally tasked to work alongside Erik ten Hag, however following the sacking of the latter on 28 September 2024, he was appointed as interim head coach.
	Location	Lists of rail accidents	*8月12日，阿贝里奥苏格兰铁路一辆InterCity 125型列车于苏格兰阿伯丁郡斯冬希文出轨，并导致机车起火。事故造成至少3人死亡。	Where in Scotland did the Abellio ScotRail InterCity 125 derail on August 12?	Stonehaven, Aberdeenshire	Inverness, Highland	On August 12, an Abellio ScotRail InterCity 125 derailed near Inverness, Highland, Scotland, and its locomotive caught fire, killing at least three people.
Japanese	Date	Cinnamoroll	2023年3月3日、公式YouTubeチャンネルが開設。と同時にYouTube内にてシナモンの誕生日を記念してショートアニメ「シナモンアニメだもん」が開始された。	What date did Cinnamoroll's official YouTube channel launch?	March 3, 2023	April 7, 2023	On April 7, 2023, Sanrio launched an official YouTube channel for Cinnamon to stream a short anime.
	Organization	Shintaro Tsuji	* 2020年8月27日-「普通はテレビ受けないの。今回は絶対に言い残したい事があって最後だから」と約30年ぶりのテレビ取材に応じ、テレビ東京で反戦を語る。	Which TV station aired Shintaro Tsuji talking about anti-war views in a rare TV interview?	TV Tokyo	Nippon TV	In a rare television interview, Tsuji spoke about his anti-war views on Nippon TV.
Korean	Person	Soo Jung Lee	* 이수정·이다혜. 《이수정, 이다혜의 범죄 영화 프로파일》. 민음사. 2020년 3월 31일.	Who co-wrote the book about crime film profiling with Lee Soo-jung?	Lee Da-hye	Kim Da-hye	In 2020, Lee co-wrote Lee Soo-jung and Kim Da-hye's Crime Film Profile, a book about crime film profiling, with Kim Da-hye.
	Number	Economy of Germany	앞으로 동독과 서독의 경제적 상황이 균등해지기 위해 걸리는 시간에 대해서 예측한 결과 할레경제연구소는 2030년까지는 기다려야 한다고 전망했다. 지난 30년간 동독의 노동생산성과 소득은 서독의 85% 수준에 이르렀고, 실업률 또한 양 지역의 격차가 크지 않은 수준에 도달했다.	Over the past 30 years, what percentage of West Germany's level did East Germany's labor productivity and income reach?	85% of West Germany's level	72% of West Germany's level	Over the past 30 years, East Germany's labour productivity and income reached 72% of West Germany's level, and the unemployment-rate gap between the two regions also narrowed substantially.

Table 8: **Examples of original and conflicting facts across languages and conflict types.** $\mathcal{Q}_{\text{en}}$: English question; $\mathcal{A}_{\text{en}}$: Original English answer; $\mathcal{A}'_{\text{en}}$: Conflict-injected English answer.

A.5 Answerability Check

After applying the answerability check in the knowledge conflict setting, we retain 11,888 QA pairs across five source languages, which serve as our final evaluation set. The set contains 957 French, 3,787 Chinese, 3,074 Japanese, 1,937 Korean, and 2,133 Bengali QA pairs.

Type	Model	Knowledge Cutoff	Context Window	Post-cutoff QA pairs (/12K)
🔓 **Open-weight**	LLaMA-3.1 8B	2023-12	128K	3,551 (30%)
	Qwen-3 8B	2023-12	128K	3,551 (30%)
	Aya-Expanse 8B	2024-06	128K	2,667 (22%)
🔒 **Closed-source**	GPT-4o	2023-09	128K	5,129 (43%)
	GPT-4.1	2024-05	1M	3,708 (31%)
	DeepSeek-V4-Pro	2025-05	1M	1,821 (15%)
	Grok-4.3	2025-09	1M	1,283 (11%)
	GPT-5.5	2025-12	1M	968 (8%)

Table 9: **List of evaluated models.** We report each model's knowledge-cutoff date, context-window size, and number of post-cutoff QA pairs. A post-cutoff QA pair is one whose `timestamp` postdates the model's reported knowledge cutoff and the one which the model answers incorrectly when only given the question in both the source language and English. When available, we use non-reasoning mode and the default effort setting for closed-source models.

B Additional Diagnosis with Waldo

B.1 Retrieval-Augmented Robustness Results

Our primary experiments provide paired multilingual Wikipedia documents directly to models in order to isolate the effect of query language on evidence use. To test whether the pattern persists with additional retrieved evidence, we evaluate 100 randomly sampled QA pairs in a retrieval-augmented setting. For each question, we use OpenAI's web-search tool to retrieve up to five documents from external web sources.[9] We then sample two retrieved documents in the source language and two in English, and provide these four documents together with the original source-language and English Wikipedia documents. Each model therefore receives six documents in total.

As shown in Table 10, adding retrieved documents generally lowers absolute answer agreement and probability for both source-language and English queries. Manual inspection suggests that it introduces additional candidate sources, which increases the difficulty of identifying the target fact. Despite this overall reduction, we observe the same pattern as in the controlled two-document setting: answer agreement and probability gaps are small under knowledge gaps but substantially larger under knowledge conflicts. These results suggest that query-language preference is not solely an artifact of the controlled two-document context.

B.2 Per-language Results

We report the per-language results for each knowledge gap and conflict setting in Tables 21 and 22. Across all five source languages, we observe that the general trend holds: query-language preference is minimal under knowledge gaps but pronounced under conflicts. While answer agreement shows no clear pattern by language resource level (Table 9), answer probability gap is consistently larger for lower-resource languages: under knowledge gaps, 18.3% (Bengali), 15.2% (Korean), 14.1% (Japanese), 10.7% (Chinese), and 8.4% (French); under conflicts, 28.0% (Bengali), 25.8% (Japanese), 23.5% (Korean), 17.8% (Chinese), and 12.6% (French).

B.3 Answer Probability for Knowledge Conflicts

For knowledge conflicts, we additionally measure the probability assigned to the *conflicting* English-supported answer under each query language. This complements the main metric (§4.2), which measures the probability assigned to the source-supported answer.

[9] https://developers.openai.com/api/docs/guides/tools-web-search

Type	Model	Agreement (%)			Probability (%)		
		src	en	Δ_{agree}	src	en	Δ_{prob}
Knowledge Gap							
🔓	LLAMA-3.1	86.3	68.7	17.6^{***}	36.8	20.2	16.6^{***}
🔓	QWEN-3	81.2	66.5	14.7^{***}	14.6	7.1	7.5^{**}
🔓	AYA-EXPANSE	59.8	41.3	18.5^{***}	42.1	29.4	12.7^{***}
🔒	GPT-4O	93.1	84.7	8.4^{**}	–	–	–
🔒	GPT-4.1	95.2	90.6	4.6^{*}	–	–	–
🔒	DEEPSEEK-V4-PRO	91.4	88.3	3.1^{*}	–	–	–
🔒	GROK-4.3	95.8	93.7	2.1^{*}	–	–	–
🔒	GPT-5.5	96.4	95.1	1.3	–	–	–
Knowledge Conflict							
🔓	LLAMA-3.1	76.1	10.5	65.6^{***}	34.2	11.8	22.4^{***}
🔓	QWEN-3	59.3	14.2	45.1^{***}	12.8	4.1	8.7^{**}
🔓	AYA-EXPANSE	51.7	7.4	44.3^{***}	38.6	15.3	23.3^{***}
🔒	GPT-4O	61.8	14.3	47.5^{***}	–	–	–
🔒	GPT-4.1	75.5	23.8	51.7^{***}	–	–	–
🔒	DEEPSEEK-V4-PRO	67.4	21.5	45.9^{***}	–	–	–
🔒	GROK-4.3	78.2	26.1	52.1^{***}	–	–	–
🔒	GPT-5.5	84.9	16.8	68.1^{***}	–	–	–

Table 10: **Retrieval-augmented robustness results.** We perform the same statistical tests as in Figure 2. Closed-source models do not expose answer probabilities (–).

Let $\mathcal{A}'_{\text{en}} = (t_1, \ldots, t_k)$ denote the conflicting English-supported answer. For a query in language ℓ, we compute its length-normalized likelihood as:

$$P_\theta(\mathcal{A}'_{\text{en}} \mid \mathcal{Q}_\ell, \mathcal{C}) = \left[\prod_{m=1}^{k} p_\theta(t_m \mid \mathcal{Q}_\ell, \mathcal{C}, t_{<m}) \right]^{1/k}, \tag{9}$$

where p_θ is the model's next-token distribution and $t_{<m}$ denotes the preceding answer tokens. We average this score across evaluated QA pairs for each query language and report the conflicting English-supported answer likelihood shift:

$$\Delta^{\text{en}}_{\text{prob}} = \text{Prob}(\text{en}) - \text{Prob}(\text{src}). \tag{10}$$

A positive $\Delta^{\text{en}}_{\text{prob}}$ indicates that the model assigns greater likelihood to the conflicting English answer when queried in English than when queried in the source language; larger values indicate stronger English-directed query-language preference.

As shown in Figure 4, $\Delta^{\text{en}}_{\text{prob}}$ is positive for all three open-weight models. This result indicates that query-language preference operates in both directions: models favor the source-language document under source-language queries and the English document under English queries.

B.4 ADDITIONAL EVALUATION METRICS

We report three additional metrics: (1) **F1 score gap** complements the LLM-as-judge-based answer agreement metric by measuring token-level overlap with the reference answer through precision and recall. We do not report exact match since model responses are often verbose (e.g., "*It is 43 years.*" compared to the reference answer "*43 years*"). (2) **Citation attribution** measures whether the model cites the source-language or the English document. Models are instructed to append the cited document as `[Document ID]` at the end of their response (see Prompt A.7 in Appendix E) (Zhang et al., 2025; Ki et al., 2026), and we extract this ID . (3) **Response language** is the language of the model response, detected with the LINGUA language detection tool.[10] We assign 1 if the top-1 detected language matches the query language and 0 otherwise.

[10] https://pypi.org/project/lingua-language-detector/

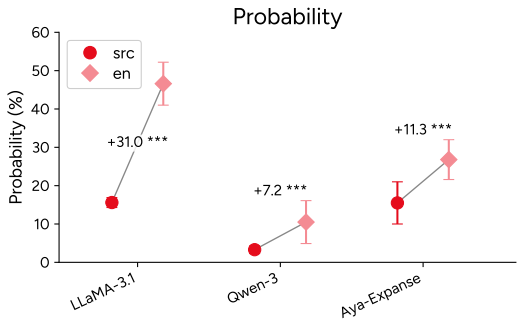


Figure 4: **Answer probability ($\Delta^{\text{en}}_{\text{prob}}$) for conflicting English-supported answer.** A larger positive Δ(en-src) indicates stronger query-language preference (i.e., English preference). We use Wilcoxon signed-rank test for differences, with Bonferroni correction for multiple comparisons. ***: statistically significant with $p < .001$.

	Source-language ✓	**Source-language ✗**
English ✓	Knowledge parity (Appendix B.5)	Swapped roles (Table 2)
English ✗	Main experiments with WALDO (§5)	Not dealt in our paper

Table 11: **A 2×2 view of tested variants.** ✓: contains the original target fact; ✗: does not contain the target fact. For completeness, we test additional variants in Table 2 and Appendix B.5.

As shown in Table 13, F1 score gap mirrors the answer agreement gap. Under knowledge gaps, F1 differences across query languages are small, particularly for closed-source models, whereas under knowledge conflicts, the gap becomes substantially larger. Citation attribution also shows the same pattern: models more frequently cite the source-language document under source-language queries and more English document under English queries, corroborating the preference through their citation behavior.

Response language stays faithful to the query language in both settings (gap: 93.4% source-language, 92.9% English; conflict: 92.1% source-language, 95.1% English). Notably, under knowledge gaps, although the model must retrieve the fact from the source-language document (since it is absent from English), it still mostly responds in the query language. This rules out that low answer agreement under English queries stems from an inability to retrieve information cross-lingually, confirming that our answer agreement metric reflects genuine query-language preference rather than a cross-lingual competence gap.

Type	**Model**	**Agreement (%)**			**Probability (%)**		
		src	en	Δ_{agree}	src	en	Δ_{prob}
🔓	LLAMA-3.1	96.3	98.4	–2.1	34.7	26.5	8.3**
🔓	QWEN-3	95.4	95.9	–0.4	11.7	10.1	1.6
🔓	AYA-EXPANSE	71.6	84.0	–12.4	43.7	34.5	9.2**
🔒	GPT-4O	99.0	99.2	–0.1	–	–	–
🔒	GPT-4.1	99.8	100.0	–0.2	–	–	–
🔒	DEEPSEEK-V4-PRO	98.6	99.2	–0.6	–	–	–
🔒	GROK-4.3	99.2	99.6	–0.4	–	–	–
🔒	GPT-5.5	98.9	99.3	–0.4	–	–	–

Table 12: **Answer agreement (Δ_{agree}) and probability (Δ_{prob}) for knowledge parity experiment. Agreement** (src, en): answer agreement; **Probability** (src, en): probability assigned to the source-supported answer, under source-language and English queries, respectively. We perform the same statistical tests as in Figure 2. Closed-source models do not expose answer probabilities (–).

Type	Model	F1 score (%)			Cite.($\mathcal{Q}_{\text{src}}$) (%)		Cite.($\mathcal{Q}_{\text{en}}$) (%)		Lang. (%)	
		src	en	Δ_{F1}	src	en	src	en	src	en
Knowledge Gap										
🔓	LLAMA-3.1	32.1	25.4	6.7	78.6	18.2	59.0	34.6	93.9	97.1
🔓	QWEN-3	29.2	22.8	6.4	83.7	14.7	66.7	33.0	87.0	97.1
🔓	AYA-EXPANSE	33.2	24.4	8.8	48.2	27.2	38.2	42.3	92.0	69.3
🔒	GPT-4O	33.9	34.2	-0.3	94.3	5.7	89.9	10.1	99.9	94.8
🔒	GPT-4.1	37.6	35.4	2.2	98.6	1.4	94.2	5.8	80.8	98.3
🔒	DEEPSEEK-V4-PRO	37.2	31.0	6.2	97.2	2.7	94.2	5.8	97.7	91.5
🔒	GROK-4.3	39.5	30.5	9.0	97.3	1.9	94.5	5.2	100.0	97.8
🔒	GPT-5.5	44.3	43.0	1.3	99.6	0.4	97.7	2.3	95.9	97.6
Knowledge Conflict										
🔓	LLAMA-3.1	32.7	14.1	18.6	65.6	30.4	14.4	81.3	93.5	98.3
🔓	QWEN-3	29.1	14.3	14.8	64.5	33.4	24.9	72.9	88.7	97.9
🔓	AYA-EXPANSE	33.4	14.1	19.3	46.3	32.2	25.2	60.7	86.1	78.5
🔒	GPT-4O	34.7	16.7	18.0	62.3	37.7	17.1	82.9	99.4	96.6
🔒	GPT-4.1	38.7	18.6	20.1	78.3	21.7	27.3	72.7	78.9	97.8
🔒	DEEPSEEK-V4-PRO	39.4	18.9	20.5	77.4	22.5	28.0	72.0	94.1	95.2
🔒	GROK-4.3	37.9	17.0	21.0	80.2	17.9	27.9	70.7	100.0	98.5
🔒	GPT-5.5	44.6	20.1	24.5	89.9	10.1	19.7	80.3	95.8	98.4

Table 13: **F1 score gap, citation attribution (Cite.($\mathcal{Q}_{\text{src}}$), ($\mathcal{Q}_{\text{en}}$)), and response language for each model and setting.**

B.5 KNOWLEDGE PARITY

To verify that the observed query-language preference is significantly reduced when equivalent knowledge is available across languages, we introduce a control condition with cross-lingual knowledge parity, which corresponds to another cell in Table 11. In this condition, the source-language and English documents support the same target fact and answer: we inject an English translation of the target fact into the English document. As shown in Table 12, the query-language preference gaps measured by answer agreement and answer probability are generally close to zero.

C HYPOTHESES

To identify potential causes of query-language preference and rule out plausible confounders, we formulate and test four hypotheses. We introduce each hypothesis below.

C.1 PRE-TRAINING EFFECT

> For a fixed answer, models assign higher conditional probability when the query and document are in the same language than when they are in different languages.

Rationale. Multilingual language models are pre-trained primarily on large, language-specific web corpora, whereas explicitly aligned cross-lingual text, such as parallel or bilingual documents, is comparatively limited (Burchell et al., 2025). Consequently, semantically aligned same-language query–document pairs may be more common in pre-training data than equivalent cross-language pairs, potentially inducing a higher prior for same-language query–document combinations.

Test. Using three open-weight models, we compare the perplexity of a fixed candidate answer when the query language either matches or differs from the language of the supporting document. For each supported answer $\mathcal{A}_\ell$ and its document $\mathcal{D}_\ell$, where $\ell \in \{\text{src}, \text{en}\}$, we compute:

$$\Delta^{\ell}_{\text{PPL}} = \text{PPL}(\mathcal{A}_\ell \mid \mathcal{Q}_{\neg\ell}, \mathcal{D}_\ell) - \text{PPL}(\mathcal{A}_\ell \mid \mathcal{Q}_\ell, \mathcal{D}_\ell)\,, \tag{11}$$

where $\neg\ell$ denotes the other query language. For $\ell = \text{en}$, $\mathcal{D}_{\text{en}}$ and $\mathcal{A}_{\text{en}}$ refer to the conflict-injected English document and its supported answer, respectively. A positive gap indicates that the fixed answer has higher perplexity, and therefore lower likelihood, when the query and document languages differ. A gap near zero indicates insensitivity to query–document language alignment.

Result. As shown in Table 14, the perplexity gap is generally positive for conflict-supported English answers, whereas results are mixed for source-supported answers. Among the three models, AYA-EXPANSE 8B exhibits the largest effects and the most statistically significant positive or negative gaps. However, most gaps, including those above and below zero, are not statistically significant. Thus, this analysis provides limited evidence that query–document language alignment, as reflected in answer perplexity, is a substantial driver of query-language preference.

Language	Model	$\mathrm{PPL}(\cdot \mid \mathcal{Q}_{\neg\ell}, \mathcal{D}_\ell)$	$\mathrm{PPL}(\cdot \mid \mathcal{Q}_\ell, \mathcal{D}_\ell)$	Δ_{PPL}	p
		$\mathcal{A}'_{\mathrm{en}}$			
French	LLAMA-3.1	4.22	3.42	0.81	
	QWEN-3	8.25	8.09	0.16	
	AYA-EXPANSE	21.76	10.04	11.72	*
Chinese	LLAMA-3.1	3.80	2.93	0.88	
	QWEN-3	8.78	5.43	3.36	
	AYA-EXPANSE	11.57	5.77	5.80	
Japanese	LLAMA-3.1	2.94	2.89	0.05	
	QWEN-3	8.84	6.16	2.68	
	AYA-EXPANSE	15.07	5.83	9.24	
Korean	LLAMA-3.1	2.58	2.46	0.12	
	QWEN-3	7.49	5.82	1.67	
	AYA-EXPANSE	11.21	5.90	5.31	
Bengali	LLAMA-3.1	3.71	2.62	1.09	
	QWEN-3	6.83	4.78	2.05	
	AYA-EXPANSE	12.80	5.24	7.52	**
		$\mathcal{A}_{\mathrm{src}}$			
French	LLAMA-3.1	3.92	2.97	0.95	
	QWEN-3	6.99	5.69	1.31	
	AYA-EXPANSE	18.48	18.74	−0.26	
Chinese	LLAMA-3.1	6.15	4.26	1.88	
	QWEN-3	8.15	8.50	−0.35	
	AYA-EXPANSE	22.41	23.75	−1.34	
Japanese	LLAMA-3.1	4.40	2.88	1.51	
	QWEN-3	5.71	5.18	0.53	
	AYA-EXPANSE	10.76	11.73	−0.96	*
Korean	LLAMA-3.1	4.95	3.81	1.14	
	QWEN-3	6.83	6.19	0.65	*
	AYA-EXPANSE	13.24	14.24	−0.99	
Bengali	LLAMA-3.1	1.40	1.30	0.10	*
	QWEN-3	1.79	1.68	0.11	
	AYA-EXPANSE	1.95	1.85	0.10	*

Table 14: **Results for B.1 Pre-training Effect hypothesis.** Wilcoxon signed-rank test performed against zero; *: statistically significant with $p < .05$; **: $p < .01$; unmarked: not significant.

C.2 EMBEDDING SIMILARITY

> Models are more likely to attribute an answer to the document matching the query language because same-language query–document pairs have higher embedding similarity than cross-language pairs.

Rationale. Retrieval-augmented models may favor documents that are more similar to the query in embedding space. If multilingual representations systematically assign higher similarity to same-language than cross-language query–document pairs, this bias could lead models to preferentially cite the document written in the query language.

Test. For each of the three open-weight models, we compute cosine similarity between each query and the source-language and conflict-injected English documents using the model's own embeddings. For a query in language $\ell \in \{\mathrm{src}, \mathrm{en}\}$, let $\mathcal{D}_\ell$ denote the document in the query language and $\mathcal{D}_{\neg\ell}$ the document in the other language. We compute the same-language similarity advantage:

$$\Delta^{\ell}_{\mathrm{Sim}} = \cos(\mathrm{embed}(\mathcal{Q}_\ell), \mathrm{embed}(\mathcal{D}_\ell)) - \cos(\mathrm{embed}(\mathcal{Q}_\ell), \mathrm{embed}(\mathcal{D}_{\neg\ell}))\,. \tag{12}$$

For English queries, $\mathcal{D}_{\text{en}}$ denotes the conflict-injected English document. A positive value indicates that the query is more similar to the same-language document, while a value near zero indicates little sensitivity to query–document language alignment.

We then compute the point-biserial correlation between $\Delta^{\ell}_{\text{Sim}}$ and citation attribution, a binary indicator of whether the model cites the document in the query language (Appendix B.4). A significant positive correlation indicates that models are more likely to cite the same-language document when it has a larger embedding-similarity advantage, supporting the hypothesis.

Result. As shown in Table 15, correlations between $\Delta^{\ell}_{\text{Sim}}$ and citation attribution are mixed across models and languages, with only a small number of model–language combinations showing a statistically significant positive correlation. Thus, greater query–document embedding similarity does not reliably predict query-language preference from an attribution perspective.

Language	LLAMA-3.1	QWEN-3	AYA-EXPANSE
	$\mathcal{Q}_{\text{en}}$		
French	0.137	0.189	0.171
Chinese	0.055	0.008	0.048
Japanese	0.042	0.274**	0.073
Korean	–0.028	–0.115	0.138
Bengali	0.108	–0.080	0.132
	$\mathcal{Q}_{\text{src}}$		
French	0.066	0.071	0.220*
Chinese	–0.030	0.183	0.237*
Japanese	–0.014	0.026	0.022
Korean	0.201*	0.182	0.053
Bengali	0.231*	–0.319	0.173

Table 15: **Results for B.2 Embedding Similarity hypothesis.** Point-biserial correlation is performed; *: statistically significant with $p < .05$; **: $p < .01$; unmarked: not significant.

C.3 SURFACE CUES

> Query-language preference is driven by surface-level document properties, including relative document length and input order.

Rationale. Language models can be sensitive to superficial properties of evidence that affect their predictions (Wan et al., 2024). Thus, observed query-language preference may instead reflect differences in the source-language and English Wikipedia documents, rather than a preference for the query language itself. We examine two potential confounds: relative document length and document order in the input context. Although we randomize document order, residual position effects could still influence which document a model follows (Liu et al., 2024).

Test. Using the three open-weight models, we measure document length in tokens using each model's tokenizer and compute the length difference:

$$\Delta_{\text{length}} = \text{Length}(\mathcal{D}_{\text{src}}) - \text{Length}(\mathcal{D}'_{\text{en}}). \tag{13}$$

Positive values indicate that the source-language document is longer. We also record whether the source-language document appears first in the input context. For each surface cue, we compute its Pearson correlation with the instance-level answer agreement difference between source-language and English queries. A significant correlation would indicate that the cue is associated with observed query-language preference.

Result. Table 16 shows that correlations between relative document length and the answer agreement difference are generally negative across models and languages: as the source-language document becomes longer relative to the English document, answer agreement gap tends to decrease. However, most correlations are not statistically significant, providing little evidence that document length reliably drives query-language preference.

Correlations between document order and the answer agreement difference are mixed in direction, consistently near zero, and not statistically significant. Thus, document position in the input context also does not appear to confound query-language preference.

Language	LLAMA-3.1	QWEN-3	AYA-EXPANSE
	Δ_{length}		
French	–0.117	–0.107	–0.118
Chinese	0.046	–0.227*	–0.255*
Japanese	–0.098	–0.112	–0.066
Korean	–0.105	–0.142	–0.124
Bengali	–0.082	0.109	–0.032
	Order		
French	0.082	0.033	–0.029
Chinese	0.093	–0.075	0.058
Japanese	–0.062	0.106	–0.042
Korean	0.031	–0.088	0.021
Bengali	–0.131	0.039	0.081

Table 16: **Results for B.3 Surface Cues hypothesis.** Pearson correlation is performed; *: statistically significant with $p < .05$; unmarked: not significant.

C.4 TOPIC LOCALITY

> Query-language preference may reflect a query's cultural specificity rather than the language in which it is asked.

Rationale. Native-language Wikipedia editions often cover regionally relevant topics that are absent from their English counterparts, a pattern associated with self-focus bias (Hecht & Gergle, 2009; Callahan & Herring, 2011) and situated knowledge (Haraway, 1988). Consequently, culturally specific queries may be better supported by documents in the associated language. Models might therefore follow the source-language document because of topic locality rather than query language itself. We test this possibility by comparing query-language preference for culturally specific and culturally agnostic queries.

Test. Following Park et al. (2026), we use GPT-5.6-TERRA to classify each QA pair as culturally *specific* or *agnostic* using Prompt A.10 in Appendix E. Specific queries concern topics culturally associated with the source language or its related region; agnostic queries do not require language- or culture-specific knowledge.

We compare the instance-level answer agreement difference between source-language and English queries across these two groups using Welch's two-sample t-test. A statistically significant group difference would indicate that topic locality contributes to observed query-language preference.

Result. As shown in Table 17, differences in the answer-agreement difference between culturally specific and agnostic queries are not statistically significant. Thus, this analysis provides little evidence that topic locality confounds query-language preference.

Language	Model	Δ_{agree}(Specific)	Δ_{agree}(Agnostic)	Difference
French	LLAMA-3.1	58.3	45.3	0.130
	QWEN-3	48.8	40.0	0.088
	AYA-EXPANSE	16.3	21.8	−0.055
Chinese	LLAMA-3.1	90.0	55.0	0.150
	QWEN-3	57.5	70.0	−0.125
	AYA-EXPANSE	33.8	25.0	0.088
Japanese	LLAMA-3.1	71.6	66.7	0.049
	QWEN-3	72.2	52.4	0.198
	AYA-EXPANSE	38.0	38.1	−0.001
Korean	LLAMA-3.1	50.7	50.0	0.007
	QWEN-3	60.6	55.6	0.050
	AYA-EXPANSE	29.6	55.6	−0.260
Bengali	LLAMA-3.1	61.1	75.0	−0.139
	QWEN-3	55.6	65.0	−0.094
	AYA-EXPANSE	44.4	51.3	−0.068

Table 17: **Results for B.4 Topic Locality hypothesis.** We perform Welch's two-sample t-test, where unmarked: not statistically significant.

D MITIGATION DETAILS

D.1 QLORA IMPLEMENTATION DETAILS

We use 4-bit quantization with `bf16` compute. We set LoRA rank $r = 16$, alpha $a = 32$, dropout 0.05, and use AdamW optimizer (Loshchilov & Hutter, 2017) with learning rate 2e-4 and weight decay 0. We use a linear scheduler with 3% warmup and set the gradient clipping at max norm 1.0. Training uses batch size 1 with gradient accumulation of 8, and runs for 3 epochs. Each model and objective is fine-tuned on a single `L40s` (48GB) GPU. We set the clipping bound c to 20 since the mean raw Forget-Retain score difference ranged from -15.2 to 56.4, with an average of 14.6, allowing us to leave the typical signal unchanged while limiting the effect of extreme likelihood gaps.

D.2 PER-QUERY-LANGUAGE RESULTS

Table 18 reports the full mitigation results for each model, with answer agreement and probability gaps under source-language and English queries, and all k values. We observe that QLORA-based training reduces both gaps the most, while QPHA also reduces them, though to a smaller extent.

D.3 SWAPPED ROLES MITIGATION

To test whether our mitigation approaches generalize, we re-run the mitigation experiments (§6.1) under the swapped-roles setup (Table 2), where the conflicting fact is in the source-language document and the original target fact in the English document.

As shown in Table 19, both QPHA and QLORA remain effective: all models show reduced agreement and probability gaps, with QLORA again yielding the largest reduction. This confirms that our mitigations generalize to the reversed direction, not only the original WALDO configuration.

D.4 INCREASING THE NUMBER OF ABLATED HEADS

We show the effect of increasing the number of ablated attention heads in Figure 5. The trend matches Figure 3: as k grows, both the agreement and probability gap shrink, with QPHA producing a larger reduction than the random ablation control relative to the vanilla baseline.

D.5 ABLATED ATTENTION HEADS

Table 20 lists the attention heads selected for ablation for each model and value of k. Figure 6 visualizes the full attention-head score map for each model, where darker colors indicate higher

Approach	LLAMA-3.1						QWEN-3						AYA-EXPANSE					
	Agreement			Probability			Agreement			Probability			Agreement			Probability		
	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ
Vanilla	78.7	15.7	63.0	47.6	17.7	29.9	76.9	17.9	59.0	17.8	5.7	12.1	72.4	13.6	58.8	58.0	24.5	33.5
QPHA ($k = 2$)																		
Random (Zero)	78.6	15.9	62.7	47.5	18.2	29.3	76.4	19.6	56.8	17.8	6.8	11.0	72.4	13.5	58.9	57.4	24.3	33.1
Random (Mean)	78.4	15.3	63.1	47.8	17.8	30.0	76.5	18.4	58.1	18.1	6.7	11.4	72.7	13.3	59.4	57.1	24.0	33.1
Zero	76.4	15.0	61.4	47.7	21.0	26.7*	76.9	20.7	56.2	18.4	7.4	11.0	72.2	16.1	56.1*	57.3	27.2	30.1*
Mean	78.5	18.8	59.7	46.6	21.9	24.7**	76.8	20.4	56.4	18.2	7.4	10.8	72.9	14.4	58.5	57.8	25.1	32.7
QPHA ($k = 4$)																		
Random (Zero)	77.5	14.6	62.9	48.7	21.8	26.9*	76.0	19.3	56.7	18.2	6.4	11.8	73.5	13.7	59.8	55.7	23.9	31.8
Random (Mean)	78.8	15.6	63.2	50.0	21.0	29.0	76.8	18.5	58.3	18.3	8.2	10.1*	72.7	14.2	58.5	57.1	23.8	33.3
Zero	75.5	15.3	60.2	46.8	24.1	22.7**	76.7	21.5	55.2*	18.4	7.5	10.9	72.7	16.2	56.5	57.4	26.6	30.8*
Mean	77.1	19.9	57.2**	48.8	33.0	15.8***	76.3	20.1	56.2	17.4	6.8	10.6	72.9	16.8	56.1*	58.4	26.0	32.4
QPHA ($k = 8$)																		
Random (Zero)	77.8	16.6	61.2	49.0	22.7	26.3*	75.4	18.4	57.0	18.3	6.2	12.1	72.8	14.2	58.6	56.1	25.3	30.8*
Random (Mean)	77.2	17.8	59.4*	50.0	21.8	28.2	75.5	18.0	57.5	18.2	6.4	11.8	72.6	15.1	57.5	56.7	24.9	31.8
Zero	75.1	18.8	56.3**	46.4	27.0	19.4***	76.0	21.7	54.3*	18.1	7.6	10.5	72.1	16.4	55.7*	57.2	27.5	29.7*
Mean	78.3	22.4	55.9**	47.3	38.2	9.1***	76.7	21.3	55.4*	17.7	7.6	10.1*	72.8	17.8	55.0*	57.1	28.4	28.7**
QLORA																		
Promote	89.3	87.0	2.3***	83.5	79.2	4.3***	86.4	82.1	4.3***	77.8	70.1	7.7**	78.5	78.2	0.3***	74.4	69.5	4.9***
Contrastive	85.5	84.0	1.5***	82.1	75.7	6.4***	82.1	80.7	1.4***	76.3	71.0	5.3***	77.4	76.5	0.9***	72.3	70.6	1.7***

Table 18: **Effect of intervention approaches on agreement (Δ_{agree}) and probability (Δ_{prob}) gaps.** For both metrics, lower is better (↓), which indicate reduced query-language preference. We perform Wilcoxon signed-rank test against Vanilla, with Bonferroni correction. *: significant with $p < .05$; **: $p < .01$; ***: $p < .001$; unmarked: not significant.

Approach	LLAMA-3.1						QWEN-3						AYA-EXPANSE					
	Agreement			Probability			Agreement			Probability			Agreement			Probability		
	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ	src	en	Δ
Vanilla	22.5	86.3	−63.8	30.4	18.5	11.9	23.6	77.9	−54.3	11.9	5.6	6.3	27.8	70.3	−42.5	20.9	17.2	3.7
QPHA ($k = 8$)																		
Zero	24.1	82.1	−58.0**	28.8	21.1	7.7*	27.4	77.0	−49.6*	11.9	7.2	4.7*	30.6	65.0	−34.4***	20.8	19.5	1.3*
Mean	27.7	85.4	−57.7**	30.1	26.5	3.6**	27.0	76.1	−49.1*	12.6	7.3	5.3	32.0	69.4	−37.4**	20.8	18.6	2.2*
Rand-Zero	21.9	85.4	−63.5	31.0	20.5	10.5	24.1	76.4	−52.3	13.1	5.7	7.4	28.4	70.7	−42.3	20.7	17.3	3.4
Mean-Zero	23.1	84.8	−61.7	31.4	20.2	11.2	23.7	76.5	−52.8	14.5	7.2	7.3	29.3	70.5	−41.2	20.8	17.2	3.6
QLORA																		
Promote	92.3	96.9	−4.6***	37.2	37.5	−0.3***	87.8	87.4	0.4***	34.0	34.0	0.0***	70.0	70.3	−0.3***	31.4	31.2	0.2**
Contrastive	89.3	93.1	−3.8***	38.2	35.1	3.1***	86.4	83.1	3.3***	29.9	30.1	−0.2***	69.5	70.4	−0.9***	24.9	26.9	−2.0**

Table 19: **Effect of intervention approaches on agreement (Δ_{agree}) and probability (Δ_{prob}) gaps for swapped roles.** For both metrics, Δ values closer to zero indicate reduced query-language preference. We perform Wilcoxon signed-rank test against Vanilla, with Bonferroni correction. *: significant with $p < .05$; **: $p < .01$; ***: $p < .001$; unmarked: not significant.

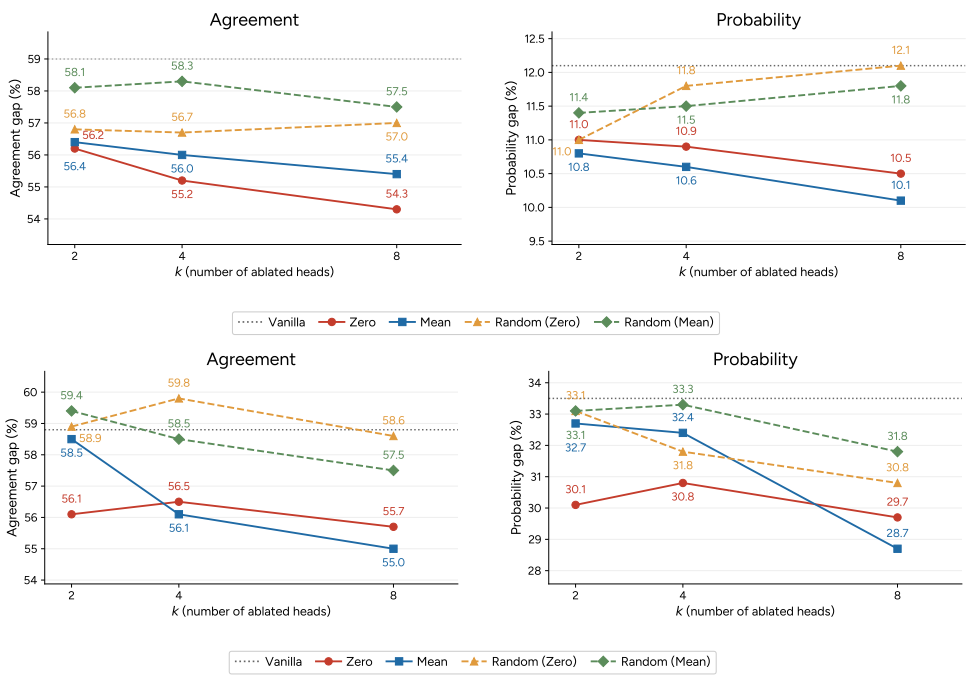


Figure 5: **Answer agreement and probability gap with increasing k.** Both metrics decrease as the number of ablated attention heads (k) increases, with QPHA showing larger gap than random control. Top: QWEN-3; Bottom: AYA-EXPANSE.

QPHA scores. The highest-scoring heads are concentrated in early-to-middle layers: layers 12–15 for LLAMA-3.1, 17–22 for QWEN-3, and 8–12 for AYA-EXPANSE.

Model	k	Ablated pairs (*layer*, *head*)
LLAMA-3.1	2	(12,1), (15,9)
	4	(12,1), (15,9), (31,14), (13,31)
	8	(12,1), (15,9), (31,14), (13,31), (8,6), (16,28), (11,1), (13,19)
QWEN-3	2	(17,18), (22,1)
	4	(17,18), (22,1), (16,19), (20,10)
	8	(17,18), (22,1), (16,19), (20,10), (22,31), (17,30), (10,7), (17,20)
AYA-EXPANSE	2	(8,28), (12,13)
	4	(8,28), (12,13), (11,29), (12,6)
	8	(8,28), (12,13), (11,29), (12,6), (10,27), (29,8), (30,8), (29,11)

Table 20: **Ablated attention heads per model and k.** LLAMA-3.1 and AYA-EXPANSE have a total of 32 layers, and QWEN-3 has 36 layers.

E PROMPTS

We present the prompts used in the data construction pipeline for WALDO (Figure 1). For the knowledge gap setting, we provide the prompts for LLM Annotation (A.1–A.2), and QA Generation (A.3). The in-context examples in A.1 differ by the source language, as illustrated in Figure 7. For the knowledge conflict setting, we provide the prompts for Conflict Injection & QA Generation (A.4) and Answerability Check (A.5). We also include the prompts used to qualitatively analyze the injected conflict types in Table 1 (A.6), the prompt used to query models, adapted from GPTResearcher[11] (A.7), and the prompt used to compare model responses with document-supported answers when computing answer agreement (§4.2) (A.8).

[11] https://github.com/assafelovic/gpt-researcher

Language	Model	Agreement (%)			Probability (%)		
		src	en	Δ_{agree}	src	en	Δ_{prob}
French	LLAMA-3.1	90.4	84.3	6.0	27.9	17.6	10.3
	QWEN-3	80.7	74.7	6.0	10.9	5.1	5.8
	AYA-EXPANSE	67.5	60.2	7.2	42.7	33.5	9.2
	GPT-4O	95.5	93.6	1.8	–	–	–
	GPT-4.1	100.0	94.0	6.0	–	–	–
	DEEPSEEK-V4-PRO	98.9	94.5	4.4	–	–	–
	GROK-4.3	99.0	97.5	1.4	–	–	–
	GPT-5.5	100.0	96.2	3.8	–	–	–
Chinese	LLAMA-3.1	96.4	67.3	29.1	31.5	16.3	15.2
	QWEN-3	92.7	80.0	12.7	9.5	5.2	4.3
	AYA-EXPANSE	69.1	45.5	23.6	43.9	31.2	12.7
	GPT-4O	99.2	88.2	11.0	–	–	–
	GPT-4.1	98.0	95.0	3.0	–	–	–
	DEEPSEEK-V4-PRO	98.4	97.3	1.1	–	–	–
	GROK-4.3	99.4	99.4	0.0	–	–	–
	GPT-5.5	99.3	99.3	0.0	–	–	–
Japanese	LLAMA-3.1	93.8	71.9	21.9	51.1	31.6	19.4
	QWEN-3	79.2	55.7	23.5	13.4	7.1	6.3
	AYA-EXPANSE	75.0	56.2	18.8	51.0	34.5	16.6
	GPT-4O	96.5	86.8	9.6	–	–	–
	GPT-4.1	98.0	93.0	5.0	–	–	–
	DEEPSEEK-V4-PRO	98.6	96.1	2.4	–	–	–
	GROK-4.3	98.8	94.7	4.1	–	–	–
	GPT-5.5	98.7	97.4	1.3	–	–	–
Korean	LLAMA-3.1	98.1	88.7	9.4	40.3	14.9	25.4
	QWEN-3	92.5	84.9	7.4	18.8	6.2	12.6
	AYA-EXPANSE	86.8	66.0	20.8	45.9	38.4	7.5
	GPT-4O	100.0	96.4	3.6	–	–	–
	GPT-4.1	100.0	97.6	2.4	–	–	–
	DEEPSEEK-V4-PRO	100.0	98.2	1.8	–	–	–
	GROK-4.3	97.9	98.1	−0.2	–	–	–
	GPT-5.5	100.0	100.0	0.0	–	–	–
Bengali	LLAMA-3.1	91.4	78.5	12.9	70.7	48.6	22.1
	QWEN-3	87.1	76.3	10.8	39.2	25.1	14.1
	AYA-EXPANSE	52.7	32.3	20.4	69.2	50.5	18.7
	GPT-4O	98.0	96.7	1.3	–	–	–
	GPT-4.1	97.9	96.9	1.0	–	–	–
	DEEPSEEK-V4-PRO	94.3	93.8	0.5	–	–	–
	GROK-4.3	97.5	96.3	1.2	–	–	–
	GPT-5.5	95.6	99.3	−3.7	–	–	–

Table 21: **Per-language answer agreement (Δ_{agree}) and probabilities (Δ_{prob}) under knowledge gap setting. Agreement** (src, en): answer agreement; **Probability** src, en): probability assigned to the source-supported answer, under source-language and English queries, respectively. Closed-source models do not expose answer probabilities (–). A larger positive Δ value indicates stronger query-language preference.

Lang.	Model	Agreement (%)			Probability (%)		
		src	en	Δ_{agree}	src	en	Δ_{prob}
French	LLAMA-3.1	78.3	19.3	59.0	25.9	10.8	15.1
	QWEN-3	53.0	19.3	33.7	9.0	3.3	5.7
	AYA-EXPANSE	54.2	28.9	25.3	37.5	20.4	17.1
	GPT-4O	51.1	25.0	26.1	–	–	–
	GPT-4.1	72.5	20.9	51.5	–	–	–
	DEEPSEEK-V4-PRO	53.1	16.9	36.2	–	–	–
	GROK-4.3	79.7	24.3	55.4	–	–	–
	GPT-5.5	81.5	20.4	61.1	–	–	–
Chinese	LLAMA-3.1	85.5	7.3	78.2	29.7	7.6	22.0
	QWEN-3	78.2	18.2	60.0	7.9	4.0	3.9
	AYA-EXPANSE	63.6	7.2	56.4	44.1	16.7	27.4
	GPT-4O	81.6	10.6	70.9	–	–	–
	GPT-4.1	90.9	21.6	69.2	–	–	–
	DEEPSEEK-V4-PRO	91.8	29.9	61.9	–	–	–
	GROK-4.3	83.0	26.0	57.0	–	–	–
	GPT-5.5	87.5	23.9	63.6	–	–	–
Japanese	LLAMA-3.1	84.4	10.9	73.4	46.5	10.1	36.4
	QWEN-3	71.9	14.1	57.8	10.1	4.0	6.1
	AYA-EXPANSE	59.4	7.8	51.6	47.5	12.6	34.9
	GPT-4O	75.4	12.7	62.7	–	–	–
	GPT-4.1	79.7	35.6	44.1	–	–	–
	DEEPSEEK-V4-PRO	74.6	23.2	51.4	–	–	–
	GROK-4.3	84.8	35.2	49.5	–	–	–
	GPT-5.5	97.0	25.0	72.0	–	–	–
Korean	LLAMA-3.1	83.0	15.1	67.9	35.6	7.7	27.9
	QWEN-3	77.4	20.8	56.6	16.8	4.3	12.5
	AYA-EXPANSE	86.8	5.7	81.1	44.6	14.6	30.0
	GPT-4O	60.9	17.8	43.1	–	–	–
	GPT-4.1	80.3	36.0	44.4	–	–	–
	DEEPSEEK-V4-PRO	71.3	39.1	32.2	–	–	–
	GROK-4.3	73.0	32.0	41.0	–	–	–
	GPT-5.5	93.0	17.0	76.0	–	–	–
Bengali	LLAMA-3.1	80.6	14.0	66.7	69.2	40.6	28.6
	QWEN-3	63.4	7.5	55.9	37.4	14.0	23.4
	AYA-EXPANSE	52.7	5.4	47.3	70.8	38.9	31.9
	GPT-4O	78.6	28.0	50.6	–	–	–
	GPT-4.1	76.5	28.5	48.0	–	–	–
	DEEPSEEK-V4-PRO	87.3	27.6	59.7	–	–	–
	GROK-4.3	88.0	29.1	59.0	–	–	–
	GPT-5.5	88.5	19.8	68.8	–	–	–

Table 22: **Per-language answer agreement (Δ_{agree}) and probabilities (Δ_{prob}) under knowledge conflict setting. Agreement** (src, en): answer agreement; **Probability** src, en): probability assigned to the source-supported answer, under source-language and English queries, respectively. Closed-source models do not expose answer probabilities (–). A larger positive Δ value indicates stronger query-language preference.

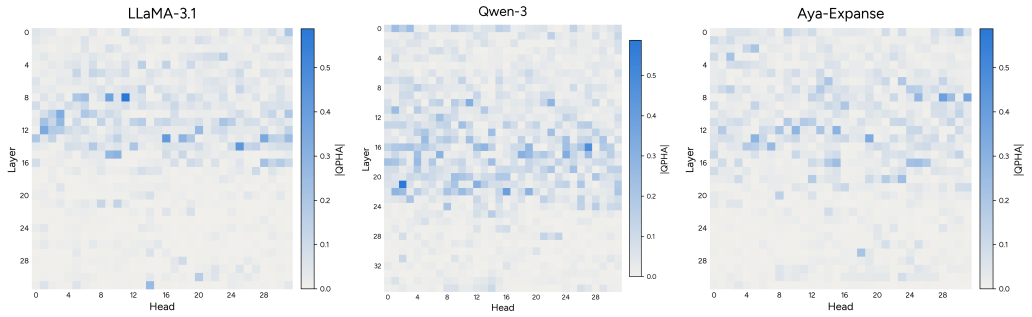


Figure 6: **Full attention-head score map for each model.** Darker-colored cells indicate higher QPHA scores (i.e., heads more strongly associated with query-language preference).

We also provide the prompt used to swap roles of documents (Table 2; A.9) and to classify queries as culturally specific or agnostic for the Topic Locality hypothesis (Appendix C.4; A.10).

Prompt A.1. LLM Annotation 1

Instruction: You are an annotation assistant for Wikipedia revision analysis. Your task is to evaluate a single revision to a French-language Wikipedia article.
You will be given:
- Article title: The title of the Wikipedia article
- Context: The surrounding Wikipedia article text
- Added information: The text newly introduced by the revision
- Removed information: The text deleted by the revision

Your task is to evaluate the revision on two dimensions.
(1) "**is fact**": Is the added information a meaningful, self-contained factual addition?
Mark "true" if the added text:
- Introduces at least one new, standalone fact not present in the removed text (e.g., `{example1}`)
- AND the removed text does not delete meaningful information

Mark "false" if the added text:
- Is only a stylistic, formatting, or phrasing change to existing content (e.g., `{example2}`)
- OR is an incomplete sentence fragment that cannot stand alone as a fact (e.g., `{example3}`)
- If the addition contains multiple facts, or one fact plus a minor fragment, mark "true."

(2) "**date**": When did the event described in the added information occur?
- Provide your best estimate in YYYY-MM format, use not only the provided context but also your general knowledge
- If only the year is determinable, use YYYY-01 as a convention
- If the date is genuinely indeterminate, use None

Return only this JSON object with no other text:
{{
"is fact": true or false,
"date": YYYY-MM or None,
"rationale is fact": One-line explanation of your "is fact" decision,
"rationale date": One-line explanation of your date
}}

Article title: `{title}`
Context: `{context}`
Added information: `{added}`
Removed information: `{removed}`

Language	example1	example2	example3
French	"Guy Chouinard meurt en décembre 2025 à l'âge de 69 ans."	replacing "française." to "française ;"	"obtenu son diplôme en."
Chinese	"2015年，何大一获加州理工学院杰出校友奖。"	replacing "（明朝唯一一个使用两个年号的皇帝）。" to "他是明朝唯一一个使用两个年号的皇帝。"	"个字节构成。这里是维基百科"
Japanese	"かつて存在した熊谷市営球場を移転する形で設置。旧球場跡地は「伊勢町ふれあい公園（南運動場）」として再整備された。"	replacing "KM23系統「西新島」停留所から徒歩（約1.5km）" to "KM23系統「西新島」停留所から徒歩（約1.5 km）"	"を卒業した。"
Korean	"서울 지하철 8호선은 석촌호수 하저터널로 석촌역과 연결된다."	replacing "로스웰 H. 대령은" to "브라운 대령(Colonel Rothwell H. Brown)은"	"받아 졸업하였다."
Bengali	"বাহাওয়ালনগর এমসির জনসংখ্যা, ২০১৭ সালে পাকিস্তানের আদমশুমারি অনুযায়ী, পাকিস্তানের পাঞ্জাবের জনসংখ্যা ২,৯৮১,৯৯৯ জন ছিল।"	replacing "বিশ্বমিত্র। বিশ্বমিত্র," to "বিশ্বামিত্র। বিশ্বামিত্র,"	"থেকে স্নাতক ডিগ্রি লাভ করেন।"

Figure 7: **Per-language examples propagated into Prompt A.1.** Native speakers manually collected these examples from sourced Wikipedia revisions.

Prompt A.2. LLM Annotation 2

Instruction: You are an expert fact-checker. Your task is to determine whether a piece of information from a Wikipedia article is present in the English Wikipedia article on the same topic.
You will be given:
- Source information: A snippet of text from the source-language Wikipedia article
- Source context: The surrounding context of the source information in the article
- English Wikipedia title: The title of the corresponding English article
- English Wikipedia content: The content of the English article

Decision rules:
- Match on meaning, not exact wording. Translations and paraphrases count as matches.
- Return true only if the key claim in the source information is clearly supported by the English content.
- If the claim is only partially present or vague, return false.
- If a number, date, name, or other specific detail in the source information conflicts with the English content, return false.

Return only this JSON object with no other text:
{{
"is present": true or false,
"match": The supporting sentence(s) quoted from the English Wikipedia content, or an empty string if is present is false
"rationale": One-line explanation of your "is present" decision,
}}

Source information: {added}
Source context: {context}

English Wikipedia title: {English title}
English Wikipedia content:
{English article}

Prompt A.3. QA Generation

Instruction: You are a query generator that targets information found in {lang}-language Wikipedia. Your task is to generate realistic user queries that target information present in a {lang} Wikipedia article but absent from (or only partially or vaguely covered by) the corresponding English article. For each query, you also produce the correct answer.

You will be given:
- {lang} article title: The title of the {lang} Wikipedia article
- {lang} context: The surrounding context of the target information within the article
- English article title: The title of the corresponding English Wikipedia article, use this to understand how the topic is rendered in English.
- {lang} information: The text containing the target information the query should ask about
- Rationale: An explanation of why this information is missing or incomplete in the English article, use this only as a guide for choosing which part of the {lang} information to target when the English article already covers part of it

Requirements for each query:
1. Produce the same query in both natural English and natural {lang}. The i-th English query and the i-th {lang} query must express the same question.
2. The query must be answerable using only the {lang} information provided.
3. Keep each query self-contained and focused on one specific fact.
4. Phrase it the way a real person would when typing into a search engine or asking an assistant, using conversational speech.
5. Frame it so it requires a factual answer, not a yes/no response.
6. Do not mention Wikipedia, the article, or the existence of {lang} or English versions.
7. Do not copy long phrases verbatim from the {lang} information; rephrase it in the user's own words.
8. Ensure "en query," "src query," "en answer," and "src answer" are all lists of strings of equal length.

Requirements for each answer:
9. Provide the answer to each query in both natural English ("en answer") and natural {lang} ("src answer"). The i-th answer must correspond to the i-th query.
10. Each answer must be grounded solely in the {lang} information provided; do not add facts that are not present in it.
11. Give only the specific fact that answers the query, such as the value, name, date, or phrase itself. Do not restate or echo the question, and do not wrap the fact in a full sentence.

Return only this JSON object with no other text:
{{
"en query": [List of English-language user queries targeting the {lang} information],
"src query": [List of {lang}-language user queries targeting the {lang} information],
"en answer": [List of English-language answers, aligned by index with en query],
"src answer": [List of {lang}-language answers, aligned by index with src query]
}}

{lang} article title: {title}
{lang} context: {context}
English article title: {English title}

{lang} information: {added}
Rationale: {rationale}

Prompt A.4. Conflict Injection+QA Generation

Instruction: Your task is to insert an altered fact into an English Wikipedia article, based on a fact stated in a source text.
You are given:
- The original English Wikipedia article
- A source text stating a fact
- A question about the article and its current (correct) answer

Produce:
1. "new sent": exactly one English sentence that states a changed version of the key fact from the source text (e.g., a different name, number, date, or place than the correct one).
2. "new article": the original article with "new sent" inserted verbatim, replacing the original statement of that fact so the article still reads naturally.
3. "new answer": the answer to the question if it were answered using "new article" instead of the original article.

Rules:
- "new sent" must change the answer to the given question; i.e., if a reader relied on "new sent" instead of the original fact, they would answer the question differently than the correct answer provided.
- "new sent" must alter the most important fact in the source text (name, quantity, date, location, etc.). If several facts are entangled (e.g., a date tied to a name), alter all of them consistently.
- Do not create the change by simple negation or polarity flip; change the underlying fact.
- "new article" must contain "new sent" verbatim.
- If the original fact appears more than once in the article, update every occurrence so the article remains internally consistent; do not leave the old and new facts contradicting each other.
- Leave all unrelated content unchanged. Do not shorten or skip any unrelated content.
- Maintain a neutral, encyclopedic tone; the insertion must not be stylistically detectable as an edit.
- "new answer" must differ from the correct answer given below; it must be the specific fact, value, name, or date itself, not a full sentence, and must not restate the question.
- Return only a valid JSON object with exactly these three keys: "new sent", "new article", and "new answer"

English Wikipedia title: `{English title}`
English Wikipedia article: `{English article}`
Question: `{question}`
Correct answer: `{answer}`
Source information: `{added}`
Response:

Prompt A.5. Answerability Check

Instruction: You are given a reference article and a question. Answer the question using only the information in the article.

Article: `{context}`
Question: `{question}`
Answer:

Instruction: Does the Model Answer convey the same fact as the Target Answer for the given question? Reply with exactly one word: "match" or "no match."

Question: `{question}`
Target Answer: `{conflict answer}`
Model Response: `{model response}`
Verdict:

Prompt A.6. Conflict Type Analysis

Instruction: You are given a question, its original correct answer, and a conflicting answer that was substituted in its place (an altered version of the same fact). Classify what kind of fact was changed, using exactly one of the following categories:
- date: dates, years, months, or spans/durations of time (e.g. 2021 → 2019, about 30 years → about 25 years)
- number: quantities, counts, statistics, or measurements that are not dates (e.g., vote counts, ages, scores)
- person: names of real people or fictional characters
- location: places, geographic regions, cities, or countries
- organization: companies, media outlets, institutions, teams, or other groups
- misc: anything that doesn't fit the above categories (e.g. roles/occupations, topics/themes, object or species names, reasons)

Respond with exactly one word from: date, number, person, location, organization, misc.
Question: `{question}`
Original answer: `{original answer}`
Conflicting answer: `{conflict answer}`
Altered sentence: `{conflict sentence}`
Category:

Prompt A.7. Querying Models

Information
Document ID: 1
Title: `{title1}`
Article: `{article1}`

Document ID: 2
Title: `{title2}`
Article: `{article2}`

Query: `{query}`

Using the above information, the response is the answer to the 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]).

Response:

Prompt A.8. Answer Agreement Judge

Instruction: Does the Model Answer convey the same fact as the Gold Answer for the given question? Ignore language, phrasing, formatting, and any trailing citation markers like [1] or [2]. Reply with exactly one word: "match" or "no match."

Question: `{question}`
Gold Answer: `{gold answer}`
Model Response: `{model response}`
Verdict:

Prompt A.9. Swapping Roles of Documents

Instruction: Your task is to swap the factual information between two Wikipedia passages while keeping each replacement in the language of its destination article.

Produce exactly two strings:
1. "swapped text": express the complete meaning of conflict fact in the language of the source article. It will replace the source fact exactly.
2. "swapped conflict sent": express the complete meaning of source fact in the language of conflict article. It will replace conflict fact exactly.

Rules:
- Preserve the facts exactly when translating: names, dates, numbers, events, relations, and polarity must not change.
- Match the destination's language, grammar, capitalization, punctuation, and encyclopedic style.
- A target may be a sentence fragment. Use its surrounding context to make the replacement fit the same grammatical position; do not force it into a standalone sentence.
- Do not combine the old and new facts, explain the edit, or add citations.
- Return only one JSON object with exactly the keys "swapped text" and "swapped conflict sent".

Source fact: `{source fact}`
Source article : `{source article}`
Conflict fact: `{conflict fact}`
Conflict article : `{conflict article}`

Prompt A.10. Topic Locality Classification

Instruction: You are given a list of English questions (English query) with their English answers (English answer), and the corresponding original `{lang}` questions (Source-language query) and answers (Source-language answer) they were translated from. The i-th items across all four lists express the same question/answer pair.

Your task is to classify each question/answer pair as exactly one of:
- "**specific**": the fact is tied to `{lang}` language, culture, geography, history, institutions, or audience; the kind of fact that is unlikely to be covered by a source that isn't specifically about that context. - "**agnostic**": the fact is general knowledge that isn't inherently tied to that `{lang}` context, even though it happens to appear in this `{lang}` Wikipedia article (e.g., a fact about a person, place, or event outside that context, or a universal/technical fact).

Requirements:
1. Produce exactly one label per question/answer pair, aligned by index.
2. The label must be exactly "specific" or "agnostic" (lowercase, no other text).
3. Ensure the output list has the same length as English query.

Return only this JSON object with no other text:
{{
"classification": [List of "specific" or "agnostic" labels, aligned by index with the English query]
}}

English query: `{en query}`
English answer: `{en answer}`
Source-language query: `{src query}`
Source-language answer: `{src answer}`