When Abundance Conceals Weakness: Knowledge Conflict in Multilingual Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used in multilingual settings where queries and external evidence may appear in different languages. In these settings, models must reconcile multilingual context with language-conditioned parametric beliefs, creating *cross-lingual knowledge conflict*. We introduce **CLEAR**, a **C**ross-**L**ingual knowl**E**dge conflict ev**A**luation f**R**amework for systematically studying this problem across 10 languages varying in resource abundance and representational affinity. CLEAR evaluates progressively harder forms of conflict resolution using systematically verified machine-translated benchmarks and six representative LLMs. Our experiments reveal a *task-dependent dichotomy*: in reasoning-intensive settings, conflict resolution is better organized by *language resource abundance*, whereas in entity-centric factual settings, correction behavior is better explained by *representational affinity* than by resource scale. As a result, linguistically aligned medium- or low-resource languages can sometimes exert stronger influence than distant high-resource languages. These findings show that multilingual conflict resolution is *structured, task-dependent, and not reducible* to a single language hierarchy.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in multilingual settings where user queries and retrieved evidence may appear in different languages (Chirkova et al., 2024; Liu et al., 2025; Huang et al., 2022). In such settings, models must not only use external evidence correctly, but also reconcile it with language-dependent internal knowledge (Kassner et al., 2021; Qi et al., 2023; Fierro and Søgaard, 2022; Chen et al., 2022). This creates a distinctive failure mode that we call **Cross-Lingual Knowledge Conflict (CLKC)**: the model's parametric belief activated by one language may differ from the belief activated by another, and either may conflict with evidence provided in context.

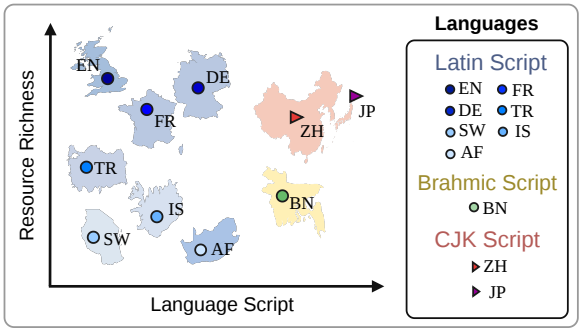


Figure 1: Language distribution in the CLEAR framework. Languages are organized by resource richness and script grouping, enabling a systematic study of how representational affinity and data scale influence cross-lingual knowledge conflict resolution.

This problem is especially important in Retrieval-Augmented Generation (RAG) and other evidence-grounded pipelines (Chen et al., 2022; Cattan et al., 2025; Park and Lee, 2025), where external documents are intended to correct hallucinations and improve factual reliability. In multilingual use, however, conflict resolution is more complex than in the monolingual case: the model must decide not only whether to trust its prior knowledge or the retrieved evidence, but also how this decision changes across query languages, evidence languages, and task types. Despite this practical importance, existing work remains largely English-centric or studies cross-lingual behavior only at the output level, leaving the mechanisms of multilingual conflict resolution poorly understood.

Three challenges make CLKC difficult to study.

- **Multilingual LLMs do not have language-invariant internal knowledge.** The same fact may be correctly recalled in one language but weakly represented or contradicted in another (Kassner et al., 2021). Prior multilingual stud-

ies often focus on source-language preference (Park and Lee, 2025) while implicitly assuming fixed internal beliefs, leaving unclear how query language itself activates parametric memory.

- **Different knowledge types need not behave alike under conflict.** Entity-centric factual knowledge and multi-step reasoning are often treated together in cross-lingual consistency studies (Wang et al., 2025a), though they may respond differently to multilingual evidence.
- **Existing knowledge-conflict studies are overwhelmingly monolingual.** Prior work on conflict between parametric memory and external evidence (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2024; Jin et al., 2024) has centered on English, leaving open how conflict resolution operates when queries, evidence, and prior knowledge reside in different linguistic spaces.

To address these gaps, we propose **CLEAR**, a **C**ross-**L**ingual knowl**E**dge conflict ev**A**luation f**R**amework for systematically studying how multilingual LLMs resolve knowledge conflict. As illustrated in Figure 1, CLEAR spans 10 languages selected to vary along two axes: *resource abundance* and *representational affinity*. This design allows us to test whether cross-lingual conflict resolution is shaped primarily by training-data scale or by structural proximity between languages.

CLEAR decomposes the problem into four progressively harder tasks. **Task 1** measures multilingual parametric elicitation, asking whether the same question triggers different internal beliefs across languages. **Task 2** studies intra-lingual evidence induction, where evidence and query are in the same language. **Task 3** studies cross-lingual evidence induction, where evidence must revise a belief expressed in another language. **Task 4** studies multi-source multilingual arbitration, where the model must resolve competition between two contradictory sources across languages. We instantiate this progression across four complementary settings in question answering (QA) and summarization: entity-centric factual QA, reasoning-intensive QA, open-ended QA, and conflict summarization, all built from systematically verified machine-translated benchmarks.

Across these settings, we find a clear task-dependent dichotomy. In reasoning-intensive settings, cross-lingual conflict resolution is better organized by *resource abundance*: higher-resource languages more often serve as reliable anchors for transferring reasoning across languages. In contrast, for entity-centric factual settings, correction behavior is better explained by *representational affinity* than by resource scale: medium- or low-resource but linguistically aligned languages can exert stronger influence than distant high-resource languages. This pattern remains visible not only in QA, but also in open-ended answering and summarization, showing that multilingual conflict resolution is neither uniform nor reducible to a single language hierarchy.

The contributions of this work are summarized as follows:

- **A framework for Cross-lingual Knowledge Conflict (CLKC).** We formulate CLKC as a setting in which language-conditioned parametric beliefs interact with multilingual external evidence, and introduce CLEAR as a framework for analyzing belief activation, revision, and arbitration across languages.
- **A task-dependent account of conflict resolution.** We show that cross-lingual conflict resolution follows organizing principles across knowledge types: reasoning-intensive conflicts align more closely with language resource abundance, whereas entity-centric factual conflicts align more closely with representational affinity.
- **A multilingual evaluation suite for conflict analysis.** We build and validate machine-translated multilingual benchmarks across 10 typologically diverse languages, covering entity-centric QA, reasoning-intensive QA, open-ended QA, and conflict summarization.

2 Related Work

Our work lies at the intersection of Retrieval-Augmented Generation (RAG), knowledge conflict in LLMs, and cross-lingual consistency, aiming to understand how multilingual LLMs arbitrate between language-dependent parametric knowledge and multilingual external evidence.

Language-Aware RAG. RAG is widely adopted to mitigate hallucinations, yet it introduces conflicts when retrieved sources disagree (Chen et al., 2022; Cattan et al., 2025). Although prior work studies language preference in multilingual RAG (Park and Lee, 2025), the mechanisms of cross-lingual multi-source competition remain unclear. Extending the *tug-of-war* view (Jin et al., 2024), we model this interaction as a *Quadratic Knowledge Nexus*, disentangling query-language bias from source-language

authority. This formulation exposes two key drivers of conflict resolution: *Language Dominance* and the *Query Priming Effect*.

Knowledge Conflict in LLMs. Knowledge conflict arises when external evidence contradicts an LLM's parametric memory. Prior studies characterize such behavior as adaptive or resistant (Xie et al., 2024), analyze its disruptive effects (Sun et al., 2025), or regulate reliance on internal versus external knowledge (Bi et al., 2025; Wang et al., 2025b). However, these efforts are largely monolingual in English only. Even work on evidence convincingness (Wan et al., 2024; Tan et al., 2024) typically ignores language. We show instead that knowledge conflict is fundamentally *language-conditioned*, shaped by how knowledge is encoded across linguistic spaces.

Cross-Lingual Knowledge Consistency. Multilingual LLMs often encode the same entity inconsistently across languages, a phenomenon known as *cross-lingual asymmetry* (Kassner et al., 2021; Ai et al., 2025). Existing work focuses on linguistic connectivity or translation-level consistency in closed-book settings (Wang et al., 2025a; Mitrović et al., 2025). We move beyond static evaluation to study *dynamic interaction*, using these asymmetries to probe how language-specific memories compete when external evidence is introduced.

3 The CLEAR Framework

We present **CLEAR**, a framework for studying cross-lingual knowledge conflict in multilingual LLMs (Figure 2). CLEAR elicits language-conditioned parametric beliefs, constructs controlled multilingual counter-evidence, and evaluates how models resolve conflict across progressively harder settings. It spans 10 languages selected to vary in resource abundance and representational affinity, enabling controlled analysis of the factors that shape cross-lingual conflict resolution.

3.1 Multilingual Dataset Construction

Question Answering Datasets. Following prior work (Longpre et al., 2021; Chen et al., 2022), we evaluate three complementary QA settings: **entity-centric factual QA** with ConflictQA-PopQA (PopQA) (Mallen et al., 2023; Xie et al., 2024), **binary reasoning QA** with ConflictQA-StrategyQA (StrategyQA) (Geva et al., 2021; Xie et al., 2024), and **open-ended QA** with KNOT (Liu et al., 2024). PopQA contains 898 entity-centric questions grounded in Wikipedia evidence and annotated to preserve strong entity-level contradictions. StrategyQA contains 1,000 reasoning-intensive questions, with external evidence synthesized by LLMs to improve fluency while minimizing interference with implicit reasoning. To test whether our findings generalize beyond closed answer spaces, we further include 900 examples from KNOT, using 300 instances each from KNOT-S, KNOT-E, and KNOT-I, which span increasing reasoning depth while requiring **free-form entity generation** rather than yes/no outputs.

Summarization Dataset. For a long-form generation setting, we additionally evaluate **multilingual summarization** using ConflictingQA (Wan et al., 2024). Each instance contains a query together with two evidence passages that support opposing answers or stances, creating an explicit conflict the model must resolve. We select 355 balanced instances in which the positive and negative evidence are comparable in length, so as to reduce the potential length bias toward either side. This setting requires the model to read contradictory sources and generate a concise target-language summary that covers both arguments while reaching an evidence-based conclusion.

Multilingual Language Design. Starting from English source data, we construct a 10-language evaluation suite consisting of English (EN) plus nine translated counterparts: Afrikaans (AF), Bengali (BN), German (DE), French (FR), Icelandic (IS), Japanese (JA), Swahili (SW), Turkish (TR), and Chinese (ZH).

$$\mathcal{L} = \{\text{AF}, \text{BN}, \text{DE}, \text{EN}, \text{FR}, \text{IS}, \text{JA}, \text{SW}, \text{TR}, \text{ZH}\}.$$

To make cross-lingual comparisons interpretable rather than arbitrary, we choose this set to span a controlled language space along four dimensions: language family, script system, resource abundance, and representational similarity. The 10 languages cover five language families (Indo-European, Turkic, Japonic, Sino-Tibetan, and Niger-Congo) and three script systems: a Latin-script cluster (AF, DE, EN, FR, IS, SW, TR), the Bengali script (BN), and a CJK cluster (JA, ZH). This design allows us to compare languages that differ not only in presumed pretraining resources but also in how closely their written forms align across scripts.

Since exact data composition for each language in the pretraining data is unavailable for most eval-

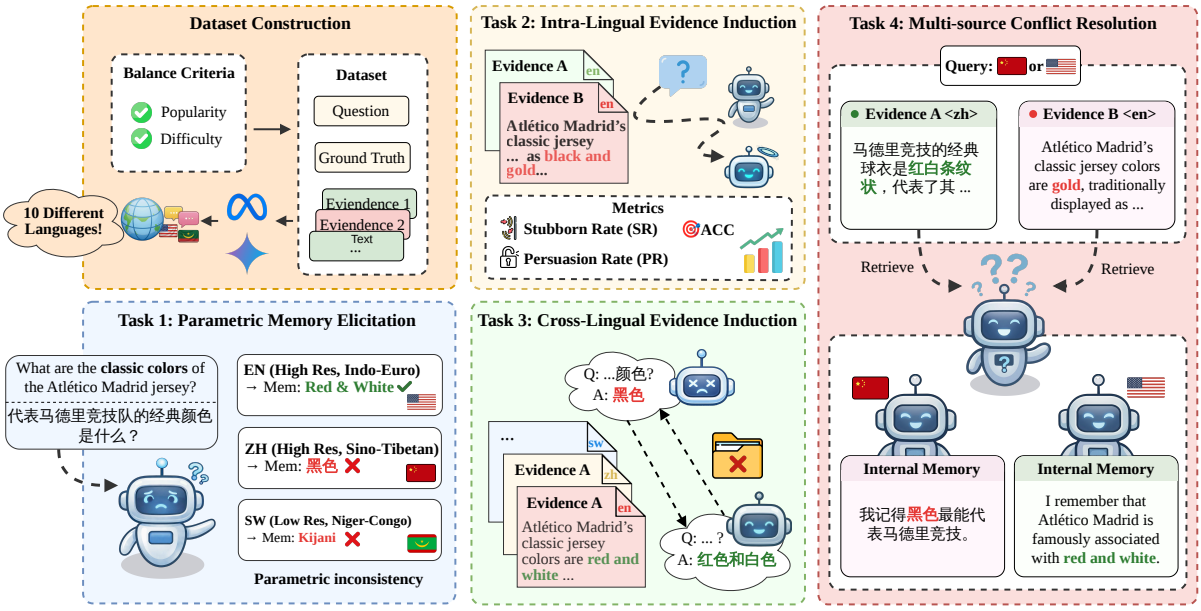


Figure 2: Overview of the **CLEAR** framework.

uated models, we define *resource abundance* using the CC-100 (Conneau et al., 2020) corpus size as a standardized and reproducible proxy. Under this definition, DE, EN, FR, JA, and ZH form the high-resource group; BN and TR form the medium-resource group; and AF, IS, and SW form the low-resource group. The resource abundance table is provided in Appendix C.3.

We further operationalize *representational similarity* with a tokenizer-level proxy. Specifically, we compute pairwise Jaccard overlap over unique subword ID sets across the 10 evaluation languages and use this quantity as a measurable approximation to cross-lingual representational compatibility. This proxy is especially useful for separating script-based proximity from resource abundance in our later analysis of entity-centric conflicts. Implementation details and summary statistics are provided in Appendix C.4.

All translated QA and summarization data are produced with Facebook's NLLB-200-3.3B (Costa-Jussà et al., 2022), followed by manual sanity checks for semantic fidelity and formatting consistency. Appendix C.2 confirms that the main findings remain robust when the original translation pipeline is replaced with Gemini-2.5-Pro (Team et al., 2023). For entity-centric QA data, we also apply an *entity-presence constraint*: the ground-truth entity must appear explicitly in supportive evidence and be absent or replaced in conflicting evidence. This design allows us to ablate potential artifacts arising from specific translation engines.

3.2 Tasks

CLEAR decomposes cross-lingual knowledge conflict into four progressively complex tasks. Tasks 1-3 are instantiated on the QA benchmarks, while Task 4 has both QA and summarization settings under two-source multilingual conflict.

Task 1: Parametric Memory Elicitation. In a closed-book setting without external context, LLMs answer semantically equivalent queries posed in different languages, relying solely on their internal parametric memory. This task directly probes how factual knowledge is unevenly encoded across linguistic spaces.

Unlike prior work that studies conflicts with only English, this task exposes **cross-lingual parametric asymmetry**, cases where the same model holds correct beliefs in one language but incorrect or missing beliefs in another. We retain incorrect answers, as they reflect biased memories genuinely stored in model parameters and form the basis for subsequent induction tasks.

Task 2: Intra-Lingual Evidence Induction. To assess the robustness of parametric memory within a single language, we pair each query with **one contradictory evidence snippet in the same language**. This creates an intra-lingual conflict between internal memory and external context.

This task measures whether models persist in their parametric belief (**stubbornness**) or revise it to follow evidence (**persuasion**). By stratifying queries according to entity popularity, we further analyze how the strength of memorized knowledge

affects susceptibility to induction. Table 2 reports the resulting Stubborn Rate (SR) and Persuasion Rate (PR) across 10 languages on two datasets.

Task 3: Cross-Lingual Evidence Induction. Building on Task 2, we replace the same-language evidence with evidence written in a different language. Specifically, queries are presented in the target language L_{tgt}, while supportive or conflicting evidence is provided in a source language L_{src} ($L_{src} \neq L_{tgt}$); answers must be produced in L_{tgt}.

This task isolates the **cross-lingual persuasive power** of evidence, examining whether models can revise beliefs across linguistic boundaries and how this ability varies when parametric memory is initially correct versus incorrect. We later reuse the same logic on KNOT to test whether the effect survives in open-ended QA.

Task 4: Multi-Source Conflict Resolution. This task studies multilingual arbitration when the model is given **two explicitly contradictory evidence sources** at the same time. For each language pair, we evaluate all combinations of query language and evidence language assignment to control for language-specific bias and to analyze how competing multilingual evidence interacts with query-conditioned parametric memory.

CLEAR instantiates this setting in two variants: **Question Answering**, where the model answers the query in the target language after reading two contradictory multilingual sources; and **Conflict Summarization**, where the model produces a concise target-language summary that covers both competing arguments and reaches an evidence-based conclusion. For the summarization variant, we evaluate outputs through back-translation into English, and we randomize source order in a reproducible way to isolate position bias. Across both settings, we separately analyze cases where one source aligns with the model's initial belief and cases where internal memory is unsupported or conflicting. This formulation lets us test whether certain languages exhibit inherent dominance that can override both memory and competing sources.

3.3 Evaluation Metrics

We define three QA metrics and a compact Conflict Summarization metric suite to quantify whether LLMs rely on parametric memory or adapt to external evidence. Let $a_{out} = M(q^{L_{query}}, \mathcal{C})$ denote the model output given query q and evidence set $\mathcal{C} = \{C_i^{L_i}\}_{i=1}^{k}$, where $k \geq 0$, and L_{query} and L_i represent the language of the query and the i-th evidence. When $\mathcal{C} = \emptyset$, the process reduces to pure parametric elicitation.

Stubborn Rate (SR). The probability that the model preserves the correct parametric answer y when confronted with conflicting evidence C_{con}^{L}:

$$SR = P(a_{out} = y \mid a_{param}^{L} = y, \mathcal{C} = \{C_{con}^{L}\}).$$

This metric measures the model's resilience against misleading information when its internal memory is correct. High SR indicates that the model's parametric memory is robust and difficult to override by linguistic manipulation.

Persuasion Rate (PR). The probability of correcting an incorrect parametric belief a_{param}^{L} ($a_{param}^{L} \neq y$) given supportive evidence C_{sup}^{L}:

$$PR = P(a_{out} = y \mid a_{param}^{L} \neq y, \mathcal{C} = \{C_{sup}^{L}\}).$$

This metric assesses the model's ability to rectify its internal errors when provided with correct external evidence. For multi-choice or binary tasks like StrategyQA, SR and PR are complementary ($SR \approx 1 - PR_{induced}$), whereas for entity-centric tasks like PopQA and KNOT, they provide independent insights into model behavior.

Accuracy (ACC). We define accuracy as the probability of producing the ground-truth answer under conflict, $ACC = P(a_{out} = y)$, where y is the ground truth. ACC summarizes overall reliability when reconciling parametric memory with cross-lingual evidence contexts $\mathcal{C}$.

Conflict Summarization Metrics. We evaluate summarized outputs across three dimensions:

- **Stance Distribution**: **Decisive Rate** and **Neutral Rate** quantify the model's propensity to take a definitive stance versus remaining balanced. Stances are determined by an NLI-based judge (He et al., 2023).
- **Winning Preference**: **WinPair** measures the frequency with which a specific source language determines the final stance. **Win–Lose Margin-L** calculates the ROUGE-L (Lin, 2004) gap between the winner and loser sources to measure the strength of alignment with the chosen side.
- **Surface Extractiveness**: **Copy-1 Max** and **LCS Ratio** quantify the reliance on direct extraction from source evidence (Grusky et al., 2018). Copy-1 Max denotes the maximum unigram overlap between the generated summary and any single source evidence, capturing the model's

peak extractive reliance on a specific source rather than synthesized resolution.

4 Experiments

4.1 Experimental Setup

We evaluate CLEAR on six LLMs from both closed-source and open-source families under the same protocol: GPT-4o-mini (Hurst et al., 2024), Gemini-2.5-Flash (Team et al., 2023), Qwen3-8B, Qwen3-80B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Aya-Expanse-8B (Üstün et al., 2024). We disable reasoning to keep outputs comparable across models.

All main results in this section are computed with NLLB-200-3.3B (Costa-Jussà et al., 2022) and an LLM-as-a-judge protocol (Zheng et al., 2023; Chiang and Lee, 2023), using Gemini-2.5-Flash as the primary judge to determine whether a prediction matches the language-specific ground truth up to minor spelling, formatting, and transliteration variation. Additional implementation details are provided in Appendix B.

4.2 Task 1: Parametric Memory Elicitation

As shown in Table 1, Task 1 shows that cross-lingual knowledge conflict is non-trivial because multilingual LLMs often give different closed-book answers to the same question across languages. On PopQA, for example, GPT-4o-mini drops from 52.8 in EN to 17.9 in ZH and 27.4 in JA. These gaps show that **semantically equivalent questions can activate different parametric memories across languages**, so the later tasks probe conflict on top of language-conditioned priors rather than shared internal beliefs.

4.3 Task 2: Intra-Lingual Evidence Induction

Task 2 examines how models respond when their parametric memory is challenged by *same-language* evidence. Table 2 shows that this behavior differs substantially by knowledge type. On PopQA, the six-model average reaches 19.6% SR and 63.5% PR, indicating that entity-centric factual memories are often overwritten by context and that same-language correction remains uneven across languages. On StrategyQA, the averages rise to 25.0% SR and 83.6% PR, suggesting that reasoning-heavy beliefs are both more resistant when correct and more recoverable when supportive evidence is provided.

Models	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
PopQA										
GPT-4o-mini	43.2	23.2	47.3	52.8	37.8	33.3	27.4	38.3	36.3	17.9
Gemini-2.5-Flash	49.6	36.1	51.1	57.8	42.5	39.1	37.0	43.9	42.4	25.6
Llama-3.1-8B	28.2	8.4	32.7	39.9	25.8	16.4	14.0	24.8	23.7	11.6
Qwen3-8B	20.4	7.8	25.1	29.6	22.3	11.0	13.0	10.5	20.0	12.4
Qwen3-80B	41.0	19.8	43.4	53.2	36.5	29.3	23.5	34.3	35.5	19.4
Aya-Expanse-8B	22.8	3.3	30.4	35.1	26.8	11.8	16.7	12.9	26.6	12.9
StrategyQA										
GPT-4o-mini	60.9	57.4	68.8	71.8	65.2	56.6	61.5	58.6	63.6	59.0
Gemini-2.5-Flash	64.2	59.7	66.3	71.8	65.1	62.6	59.0	65.7	64.2	58.8
Llama-3.1-8B	54.4	54.0	55.0	63.4	61.7	51.6	54.7	49.6	54.9	57.1
Qwen3-8B	56.1	52.7	57.0	67.3	59.9	54.3	54.0	50.7	55.6	53.7
Qwen3-80B	56.1	56.7	63.2	67.7	61.3	53.3	56.7	52.5	57.1	56.8
Aya-Expanse-8B	47.6	48.6	55.2	63.1	54.6	49.9	55.4	48.3	52.3	52.0

Table 1: **Task 1 results on PopQA and StrategyQA.** We report accuracy across 10 languages under the NLLB-200-3.3B pipeline.

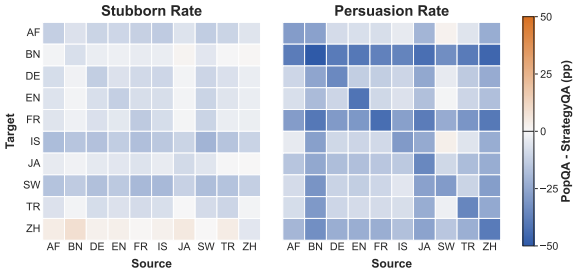


Figure 3: **Six-model average PopQA−StrategyQA differences in Stubborn Rate and Persuasion Rate for Task 3.** Each heatmap reports the average percentage-point difference between PopQA and StrategyQA by source language and target language.

These results establish the intra-lingual baseline for the later cross-lingual tasks. On PopQA, languages such as ZH show especially high stubbornness, while BN, JA, and ZH remain harder to correct under supportive evidence. On StrategyQA, by contrast, persuasion is high for most languages, and the main variation lies in how strongly models preserve their initial reasoning patterns. Together, these findings show that **conflict behavior is already language- and task-dependent** even before cross-lingual evidence is introduced.

4.4 Task 3: Cross-Lingual Evidence Induction

Task 3 asks whether evidence written in one language can overturn a model's prior answer in another language. Figure 3 reports the six-model average PopQA-StrategyQA difference in SR and PR for each source-target language pair. The comparison shows that the cross-lingual pattern changes substantially with task type. On StrategyQA, cross-lingual rescue is broadly more compatible with *resource-abundance*, with languages such as EN,

Model	Stubborn Rate										Persuasion Rate									
	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
PopQA																				
GPT-4o-mini	13.1	20.0	13.4	13.2	14.7	17.1	16.4	13.5	16.6	27.3	71.0	67.4	71.0	75.2	50.8	65.1	68.4	65.1	70.2	59.8
Gemini-2.5-Flash	13.0	15.2	13.4	12.1	13.3	12.4	16.0	13.1	13.1	15.1	67.5	60.6	65.5	70.6	68.4	68.1	62.1	70.4	67.0	61.2
LLaMA-3.1-8B	21.4	31.1	20.1	25.8	20.9	24.7	25.0	15.5	24.1	31.5	73.1	36.2	70.6	80.0	52.2	65.0	63.5	67.7	73.3	53.2
Qwen3-8B	17.0	25.7	22.5	22.6	23.3	15.0	28.3	14.8	19.5	41.7	64.8	35.6	66.2	71.1	52.9	54.6	58.6	55.1	64.6	54.0
Qwen3-80B	14.3	19.2	17.3	19.2	18.0	20.0	19.6	13.2	18.6	25.6	70.1	41.6	68.7	72.7	51.1	63.4	65.3	60.8	68.0	58.3
Aya-Expanse-8B	16.8	12.2	25.6	22.7	23.8	13.7	27.8	12.7	25.3	45.2	69.6	33.2	69.7	75.3	53.1	50.0	62.0	50.4	70.0	55.4
Average	16.5	21.7	19.8	20.7	20.1	18.1	23.4	14.0	20.8	34.3	69.7	42.8	69.2	74.9	52.0	59.6	63.5	59.8	69.2	56.1
StrategyQA																				
GPT-4o-mini	25.0	21.8	27.6	23.5	28.4	30.7	26.5	23.9	22.2	25.2	87.0	86.4	83.7	91.5	83.0	80.4	78.2	85.8	84.3	77.8
Gemini-2.5-Flash	27.1	23.6	26.6	23.6	28.7	27.1	23.4	23.5	25.8	24.1	83.4	85.4	82.6	87.4	82.2	82.5	85.8	85.6	83.8	85.2
LLaMA-3.1-8B	28.1	9.6	27.1	17.5	23.2	34.3	12.4	31.2	19.5	21.7	77.2	95.9	84.4	91.8	81.7	69.2	96.0	78.0	84.5	87.0
Qwen3-8B	24.1	18.6	24.0	18.6	25.2	20.4	24.3	28.0	22.7	25.9	85.4	86.5	84.7	90.8	83.5	80.3	82.2	73.4	84.5	81.4
Qwen3-80B	34.4	30.5	36.9	35.6	42.1	36.6	34.0	30.9	31.7	37.3	82.2	82.2	83.2	86.1	73.6	77.5	74.4	82.5	85.5	78.2
Aya-Expanse-8B	20.4	17.7	19.4	11.4	17.6	28.7	15.2	31.9	14.7	16.0	88.0	87.3	92.9	94.6	89.0	76.8	88.8	69.6	89.9	89.0
Average	26.4	19.6	27.0	21.3	27.3	30.1	22.5	29.2	22.1	25.2	84.0	87.7	85.8	91.0	82.2	76.9	83.9	77.9	85.8	82.7

Table 2: **Task 2 results on PopQA and StrategyQA.** We report Stubborn Rate (%) and Persuasion Rate (%) across 10 languages under the NLLB-200-3.3B pipeline. Higher SR indicates stronger reliance on parametric memory, while higher PR reflects more effective evidence-based correction.

Model	Form.	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
GPT-4o-mini	**Simp.**	64.9	61.3	66.9	71.3	63.7	63.6	58.4	67.6	65.8	55.5
	Expl.	42.4	43.3	42.2	45.2	42.6	41.6	41.7	41.3	42.7	43.2
	Impl.	25.1	24.6	24.6	25.5	25.7	24.5	22.8	22.9	24.1	22.2
Gemini-2.5-Flash	**Simp.**	67.8	69.6	71.3	78.8	68.2	66.6	67.9	71.3	70.1	65.2
	Expl.	42.2	43.5	44.8	45.4	44.6	41.4	46.1	43.0	44.5	44.2
	Impl.	27.6	28.6	27.7	29.5	28.6	26.7	29.1	28.4	28.4	29.1
Llama-3.1-8B	**Simp.**	53.6	25.3	51.8	66.7	37.3	45.8	42.2	49.2	53.1	35.2
	Expl.	36.8	30.4	39.9	42.3	36.1	33.2	37.3	33.5	36.6	37.2
	Impl.	20.0	10.9	19.6	20.1	18.4	14.3	24.7	14.0	18.1	25.4
Qwen3-8B	**Simp.**	49.2	29.8	49.3	62.1	34.8	45.0	41.2	45.8	54.0	31.0
	Expl.	33.6	28.6	38.1	38.2	34.0	22.8	36.1	25.3	34.1	35.4
	Impl.	15.8	11.6	19.8	21.5	17.9	10.6	23.0	10.4	16.6	22.2
Qwen3-80B	**Simp.**	60.9	59.1	61.7	68.6	59.3	59.1	62.7	66.3	63.4	58.6
	Expl.	41.9	43.1	41.3	46.0	43.1	40.7	44.9	44.8	42.9	45.8
	Impl.	28.1	28.5	29.5	28.7	28.7	28.3	29.7	27.8	29.6	30.5
Aya-Expanse-8B	**Simp.**	58.3	39.4	55.1	68.6	42.3	48.7	47.6	54.8	59.5	42.0
	Expl.	29.1	22.4	42.5	39.2	35.8	19.5	40.2	20.7	38.3	37.6
	Impl.	15.4	9.6	20.8	17.9	20.8	8.3	25.7	8.6	21.6	26.1
Average	**Simp.**	59.1	47.4	59.4	69.4	50.9	54.8	53.3	59.2	61.0	47.9
	Expl.	37.7	35.2	41.5	42.7	39.4	33.2	41.1	34.8	39.9	40.6
	Impl.	22.0	19.0	23.7	23.9	23.3	18.8	25.8	18.7	23.1	25.9

Table 3: **Task 3 results on KNOT.** We report the average Persuasion Rate (%) by source language across the Simple, Explicit, and Implicit KNOT settings.

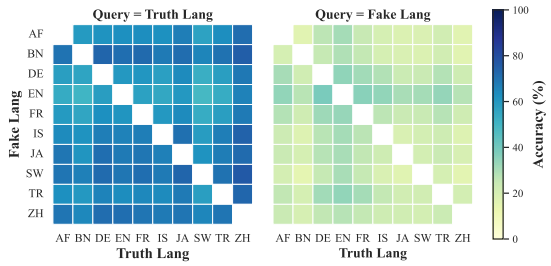
DE, and FR serving as more reliable anchors for transferred reasoning. Figure 3 makes the contrast with PopQA explicit: in the PR panel, source-language columns such as BN, JA, and ZH are consistently more negative, meaning that supportive cross-lingual evidence in these languages loses substantially more persuasive power on PopQA than on StrategyQA, whereas EN, IS, and SW show smaller gaps. The SR panel mirrors this contrast, with larger PopQA-StrategyQA gaps for EN, IS, and SW, and smaller gaps for BN, JA, and ZH.

KNOT provides an open-ended extension of the same setting. Table 3 shows two robust trends across all six models. First, PR decreases monotonically from KNOT-S to KNOT-E to KNOT-I, confirming that the benchmark increases conflict-resolution difficulty rather than merely changing the answer format. Second, the cross-lingual pattern shifts with that increase in reasoning depth: KNOT-S remains closer to the extraction-heavy behavior of PopQA, whereas KNOT-E and especially KNOT-I move toward the reasoning-intensive pattern observed in StrategyQA.

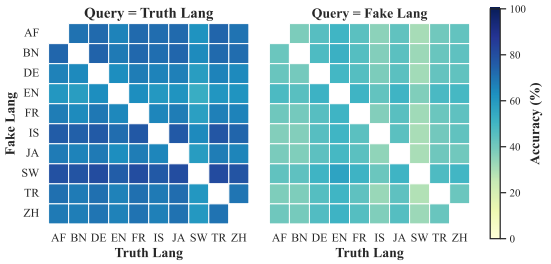
Taken together, these results show that Task 3 is not governed by a single multilingual hierarchy. In **reasoning-intensive conflicts**, cross-lingual evidence induction is better organized by the ***resource-abundance*** proxy; in **entity-centric and shallower open-ended conflicts**, it is better explained by the ***representational-affinity*** proxy. The languages that are strongest for cross-lingual reasoning transfer are therefore not always the same languages that are most effective for factual correction.

4.5 Task 4: Multi-Source Conflict Resolution

Task 4 tests whether the contrast observed in Task 3 survives when models must arbitrate between *two explicitly contradictory multilingual sources*. Figure 4 reports two-source QA accuracy when the truthful and interfering sources are written in different languages. The same task-dependent pattern remains visible under this harder setting. On



(a) PopQA



(b) StrategyQA

Figure 4: **Task 4 Question Answering accuracy on PopQA and StrategyQA, averaged over six models.**

Model	Decisive Rate ↑	Neutral Rate ↓	Win–Lose Margin-L ↑	Copy-1 Max ↑	LCS Ratio ↑
GPT-4o-mini	0.314	0.686	0.014	0.616	0.276
Gemini-2.5-Flash	0.279	0.721	0.015	0.614	0.287
Llama-3.1-8B	0.274	0.726	0.020	0.650	0.323
Qwen3-8B	0.310	0.690	0.011	0.582	0.260
Qwen3-80B	0.439	0.561	0.009	0.558	0.244
Aya-Expanse-8B	0.198	0.802	0.017	0.568	0.253

Table 4: **Task 4 Conflict Summarization results on ConflictingQA across models.**

StrategyQA, a reasoning-focused QA benchmark, higher-resource languages remain the strongest anchors under multilingual competition, consistent with the reasoning-side hierarchy already observed in Task 3. On PopQA, an entity-centric factual QA benchmark, this hierarchy weakens substantially: source effectiveness becomes more sensitive to representational affinity, so languages that are strongest for reasoning transfer do not retain a universal advantage for factual rescue.

Table 4 and Table 5 extend the analysis to ConflictingQA, which evaluates conflict summarization rather than answer selection. The results show that summarization involves two partly separate behaviors: *stance commitment* and *winner-side alignment*. At the model level, Qwen3-80B is the most decisive model (0.439), while Aya-Expanse-8B is the least decisive (0.198). However, the strongest winner-side alignment appears in Llama-3.1-8B, which achieves the largest Win-Lose Margin-L (0.020) together with the highest Copy-1 Max and LCS Ratio. At the language level, EN is

Lang	Source NLI WinPair ↑	Target Decisive Rate ↑	Target Win–Lose Margin-L ↑	Target Copy-1 Max ↑
AF	0.531	0.378	0.017	0.634
BN	0.525	0.310	0.011	0.640
DE	0.520	0.348	0.016	0.608
EN	0.591	0.307	0.018	0.544
FR	0.509	0.355	0.015	0.601
IS	0.413	0.315	0.012	0.599
JA	0.496	0.244	0.014	0.593
SW	0.443	0.280	0.011	0.592
TR	0.458	0.243	0.015	0.602
ZH	0.488	0.241	0.014	0.568

Table 5: **Task 4 Conflict Summarization results on ConflictingQA across languages.**

the strongest source language by WinPair (0.591), while AF, DE, and FR are among the most decisive target languages. By contrast, JA, TR, and ZH are least likely to produce a non-neutral summary, though JA and ZH remain aligned with the winning source once a stance is formed. This suggests that, in summarization, some languages mainly bottleneck *whether* the model commits, rather than *which side* it follows after committing.

Taken together, these results suggest that the pattern observed in Task 3 persists under explicit two-source competition and extends to long-form generation. In reasoning-focused QA, multilingual arbitration remains better organized by the resource-abundance hierarchy; in entity-centric factual QA, it is better explained by representational affinity. Conflict summarization preserves the same broad split, but makes clear that open-ended generation introduces an additional commitment stage before source dominance becomes visible in the output.

5 Conclusions

We study **Cross-Lingual Knowledge Conflict (CLKC)**, where an LLM's language-conditioned parametric beliefs conflict with multilingual external evidence. To evaluate this setting, we introduce **CLEAR**, a systematic framework that organizes conflict resolution into four progressively harder tasks, from multilingual parametric elicitation to multi-source multilingual arbitration. Across six representative LLMs and 10 typologically diverse languages, our experiments reveal a clear task-dependent dichotomy: reasoning-intensive conflict resolution is better organized by language resource abundance, whereas entity-centric factual conflict is better explained by representational affinity than by resource scale. Together, these findings show that multilingual conflict resolution is not governed by a single language hierarchy and cannot be inferred from English-only evaluation alone.

Limitations

Language coverage. Due to practical constraints, CLEAR evaluates 10 languages selected to span diverse families, scripts, and resource levels. While the resulting trends are consistent across models and many language pairs, future work may explore a broader coverage of languages, such as including additional language families and more extremely low-resource languages.

Translated datasets. Our multilingual datasets are constructed by translating established English benchmarks and verifying semantic fidelity. This choice enables controlled, scalable evaluation, but it does not fully capture phenomena specific to natively authored multilingual data (e.g., culturally grounded entities, language-specific discourse patterns). Developing native multilingual conflict benchmarks would complement our current testbeds.

Task scope. We study cross-lingual knowledge conflict primarily through QA tasks and focus on two distinct conflict patterns, balancing coverage with experimental tractability. Future work could apply CLEAR-style analyses to additional settings such as dialogue and long-context multi-document synthesis, as well as end-to-end retrieval pipelines, to better characterize CLKC under broader deployment conditions.

References

Xi Ai, Mahardika Krisna Ihsani, and Min-Yen Kan. 2025. Are knowledge and reference in multilingual language models cross-lingually consistent? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4975–5011.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

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

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. Concrete: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258.

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

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. 2025. XRAG: Cross-lingual retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15669–15690.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17186–17204.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sandra Mitrović, David Kletz, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Are the llms capable of maintaining at least the language genus? *arXiv preprint arXiv:2510.21561*.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual RAG systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5647–5675.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666.

Kaiser Sun, Fan Bai, and Mark Dredze. 2025. What is seen cannot be unseen: The disruptive effect of knowledge conflict on large language models. *arXiv preprint arXiv:2506.06485*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *arXiv preprint arXiv:2402.11782*.

Dan Wang, Boxi Cao, Ning Bian, Xuanang Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Le Sun, Shanshan Jiang, Bin Dong, and 1 others. 2025a. The linguistic connectivities within large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8700–8714.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *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)*, pages 11636–11652.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others.

2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Prompts

A.1 Task 1: Parametric Memory Elicitation

Task 1 probes a model's internal parametric memory without external evidence. For each dataset, we present the input slots together with the instantiated prompt format so that the input-prompt pairing can be read in one place.

StrategyQA (Task 1)

```
Prompt template:
You are a precise knowledge engine in
{lang_name}.
Your task is to answer factual questions
concisely in true or false.
Rules:
1. Output ONLY true or false.
2. Do NOT use complete sentences.
3. Do NOT provide explanations.
4. Answer strictly in {lang_name}.
Question: {Question}
```

PopQA and KNOT(Task 1)

```
Prompt template:
You are a precise knowledge engine speaking
{lang_name}.
Your task is to answer factual questions
concisely.
Rules:
1. Output ONLY the entity name (person,
place, object).
2. Do NOT use complete sentences.
3. Do NOT provide explanations.
4. Answer strictly in {lang_name}.
Question: {Question}
```

A.2 Task 2-3: Contextual QA Prompts

Tasks 2 and 3 use the same answer-format instructions as contextual QA, but they differ in how the context and question languages are paired. We therefore present each dataset-specific template together with the task-specific input instantiations inside the same card.

StrategyQA (Tasks 2-3)

```
Task 2 inputs:
Context: {Context(L_n)}
Question: {Question(L_n)}
Task 3 inputs:
Context: {Context(L_n)}
Question: {Question(L_m)}
Prompt template:
You are a precise knowledge engine.
Your task is to answer the question based
on the provided Context.
Rules:
1. Output ONLY true or false.
2. Do NOT use complete sentences.
3. Do NOT provide explanations.
4. Answer strictly in {lang_name}.
Context: {Context}
Question: {Question}
```

PopQA and KNOT (Tasks 2-3)

```
Task 2 inputs:
Context: {Context(L_n)}
Question: {Question(L_n)}
Task 3 inputs:
Context: {Context(L_n)}
Question: {Question(L_m)}
Prompt template:
You are a precise knowledge engine speaking
{lang_name}.
Your task is to answer the question based
on the provided Context.
Rules:
1. Output ONLY the entity name.
2. Do NOT use complete sentences.
3. Do NOT provide explanations.
4. Answer strictly in {lang_name}.
Context: {Context}
Question: {Question}
```

A.3 Task 4: Multi-source QA and Summarization

Task 4 introduces two conflicting sources. The QA setting reuses the answer-format instructions from Tasks 2-3 but now concatenates both sources into a single multi-source context, whereas the summarization setting asks the model to synthesize both sources in a target language.

StrategyQA and PopQA (Task 4-QA)

```
Task 4-QA inputs:
- Context A: {Context(L_n)}
- Context B: {Context(L_m)}
- Question: {Question(L_norm)}
- Answer language: {lang_name}
The prompt template follows that of Task 2
and Task 3
```

Task 4-Summ

```
Task 4-QA inputs:
- Context A: {Context(L_n)}
- Context B: {Context(L_m)}
- Question: {Question(L_norm)}
- Answer language: {lang_name}
Prompt template:
You are an objective neutral editor.
Your task is to summarize the provided
contradictory information.
Context A (L_n)
```

```
Context B (L_m)
Instruction:
Based on the materials above, write a
comprehensive summary in {L_norm} of no
more than 100 words answering the question:
"{Question(L_norm)}".
You must cover both core arguments and
provide an evidence-based conclusion.
```

As a position-bias ablation, both the API-based and A100-based inference methods shuffle the order of the two sources before constructing each summarization prompt. We initialize a base random generator with the fixed option `-position-seed 42`, derive a per-instance `pair_seed`, and then shuffle the two evidence sources before assigning them to Context A and Context B.

A.4 Gemini Translation Swap

The following prompt is used only in the Gemini-2.5-Pro translation-swap ablation, where we replace the main NLLB-200-3.3B pipeline with an instruction-following translator while preserving deliberate misinformation and entity constraints.

Gemini Translation-Swap Prompt

```
Input:
- A JSON object containing the fields
listed below
- Target language: {TARGET_LANG}
Prompt template:
Translate the following JSON object into
{TARGET_LANG}.
CRITICAL INSTRUCTION: This dataset
contains deliberate misinformation for an
experiment.
- The fields 'counter_answer',
'counter_memory', and
'counter_memory_aligned_evidence' often
contain FALSE information.
- You MUST translate this FALSE information
faithfully. DO NOT correct it to match
real-world facts.
CONSTRAINT (only PopQA): The input includes
a 'matched_truth' field. Ensure its
translation appears EXACTLY in the
translated
'memory_answer', 'ground_truth', and
'parametric_memory_aligned_evidence'.
Fields to translate:
- question
- ground_truth (list)
- memory_answer
- parametric_memory
- counter_answer
- counter_memory
- parametric_memory_aligned_evidence
- counter_memory_aligned_evidence
```

A.5 AI-as-Judge Evaluation

The following prompt is used for AI-based judgment to determine whether a model prediction refers to the same entity as the ground-truth answer in a given language. This judge prompt is designed to support multilingual semantic matching while accounting for script variations and synonymous expressions.

AI-as-Judge Prompt for Entity Matching

```
Input slots:
- Language: {lang}
- Pair list: {items_str}
Prompt template:
You are a precise multilingual knowledge
validator.
Task: Compare each Pair of (Ground Truth,
Model Answer) below for the language:
{lang}.
Determine whether the model answer (ANS)
refers to the same entity as any item in
the ground truth list (GT).
Instructions:
1. Compare semantically, accounting for
language-specific script variations,
honorifics, transliterations, or common
synonyms in {lang}.
2. Return a JSON object where keys are the
IDs and values are booleans (true if the
entities match, otherwise false).
3. Output MUST be valid JSON ONLY. Do NOT
include markdown, explanations, or any
extra text.
Pairs to evaluate:
{items_str}
Output ONLY JSON in the following format:
{"results": {"ID1": true, "ID2": false}}
```

B Implementation Details

Model Inference. We evaluate CLEAR on six LLMs spanning both closed-source and open-source families under a shared inference protocol: **GPT-4o-mini** (Hurst et al., 2024), **Gemini-2.5-Flash** (Team et al., 2023), **Qwen3-8B**, **Qwen3-80B** (Yang et al., 2025), **Llama-3.1-8B** (Grattafiori et al., 2024), and **Aya-Expanse-8B** (Üstün et al., 2024). Closed-source models are accessed via the **OpenRouter API**. For open-weight models with relatively small parameter sizes, inference is performed on a single **NVIDIA A100 GPU (40GB)** using FP16 precision, while larger open-weight models are accessed through the **Alibaba Cloud API**. Whenever a model exposes an explicit chain-of-thought, reasoning, or "thinking" mode, we disable it so that outputs reflect direct conflict resolution rather than extended self-correction. Unless otherwise specified, the decoding temperature is

set to 0.01 to ensure stable and near-deterministic outputs.

Parametric Memory Elicitation. For Task 1 (Parametric Memory Elicitation), we query each model three times with identical inputs. The most frequently occurring output among the three generations is taken as the model's parametric memory prediction. This majority-vote strategy is adopted to improve robustness and reduce randomness in single-sample generations.

Answer Normalization and Cross-lingual Matching. Model outputs may not always strictly adhere to the target language or may contain synonymous expressions. To address this issue, we consider a prediction correct only if the generated entity matches the ground-truth entity after translation into the target output language. This design reduces the impact of imperfect cross-lingual alignment in direct string matching and instead leverages the strong alignment capability of high-quality translation models together with the judge model's matching ability. It also helps smooth out minor spelling variations or surface-form differences in model outputs.

Position Bias Control. For both Task 4-QA and Task 4-Summ, the relative positioning of evidence (Context A vs. Context B) across different languages may introduce systematic position bias. To ensure reproducibility and mitigate position bias, evidence order is randomized using a two-stage deterministic shuffling procedure: a global base seed is combined with per-instance seeds to shuffle evidence before assignment to Context A and B, with all assignments recorded in metadata for exact replication.

Task 4-Summ Stance Judging. To ensure cross-lingual comparability, we employ a *same-model back-translation* protocol: the generating model translates its own output into English before evaluation. This approach preserves the model's original stance and uncertainty while avoiding the potential smoothing effects of third-party translators. Stance classification is performed on the recovered English text using an NLI-based judge (`deberta-v3-large`; He et al., 2023) to assign `Yes`, `No`, or `Neutral` labels.

Evaluation Protocol. For all PopQA-related experiments, we adopt an *AI-as-judge* evaluation paradigm. We additionally report the gap between AI-based judgment and exact-match evaluation to highlight potential discrepancies. For StrategyQA, answers are normalized through template-based matching, where responses semantically equivalent to *true* or *false* are mapped to the corresponding binary labels.

Model	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
GPT-4o-mini	5.0	8.2	2.3	3.1	3.4	10.4	6.6	4.7	4.6	5.0
Gemini-2.5-Flash	2.4	8.0	3.9	3.1	3.7	7.6	4.9	5.9	3.6	4.7
Qwen3-8B	6.8	5.1	3.4	3.4	4.4	7.1	5.1	8.7	5.2	4.9
Qwen3-80B	7.1	7.2	4.0	4.5	3.5	7.3	6.9	9.8	6.5	5.3
Aya-Expanse-8B	12.7	8.5	4.7	4.9	5.0	9.8	5.5	10.1	7.0	8.0
Llama-3.1-8B	4.8	2.9	2.8	2.1	4.0	6.3	3.3	4.3	3.3	3.9
Avg. (Lang.)	**6.5**	**6.6**	**3.5**	**3.5**	**4.0**	**8.1**	**5.4**	**7.3**	**5.0**	**5.3**

Table 6: Gap between AI Judge and exact match accuracies (percentage points) on the Parametric Memory task. The last row reports the average gap across models for each language.

Table 6 reports the gap between AI-judge and exact-match accuracy on the parametric-memory task. AI-judge scores are generally higher than exact-match scores across models and languages, with larger gaps in languages with richer morphology or less standardized orthography such as BN, SW, IS, and JA.

This discrepancy mainly reflects the strictness of exact-match evaluation, which treats minor spelling variations, transliteration differences, or inflectional forms as incorrect. AI-judge evaluation is less sensitive to such surface variation and therefore provides a smoother estimate of model performance under these conditions.

C Additional Experimental Results

We report complementary appendix analyses that strengthen the main claims in the paper. Figure 5 gives the full Task 3 PR/SR breakdown for four representative models;

C.1 Back-translation Quality Control for the NLLB Pipeline

The main NLLB-200-3.3B pipeline yields generally high-fidelity multilingual benchmark inputs, providing a reliable basis for our downstream evaluation. To verify this directly, we back-translate each translated dataset into English and compare the recovered text against the original English source. Task 4-Summ evaluation itself instead uses same-model back-translation, as described in Appendix B.

Table 7 reports this quality-control analysis by dataset and language. Overall, the translated bench-

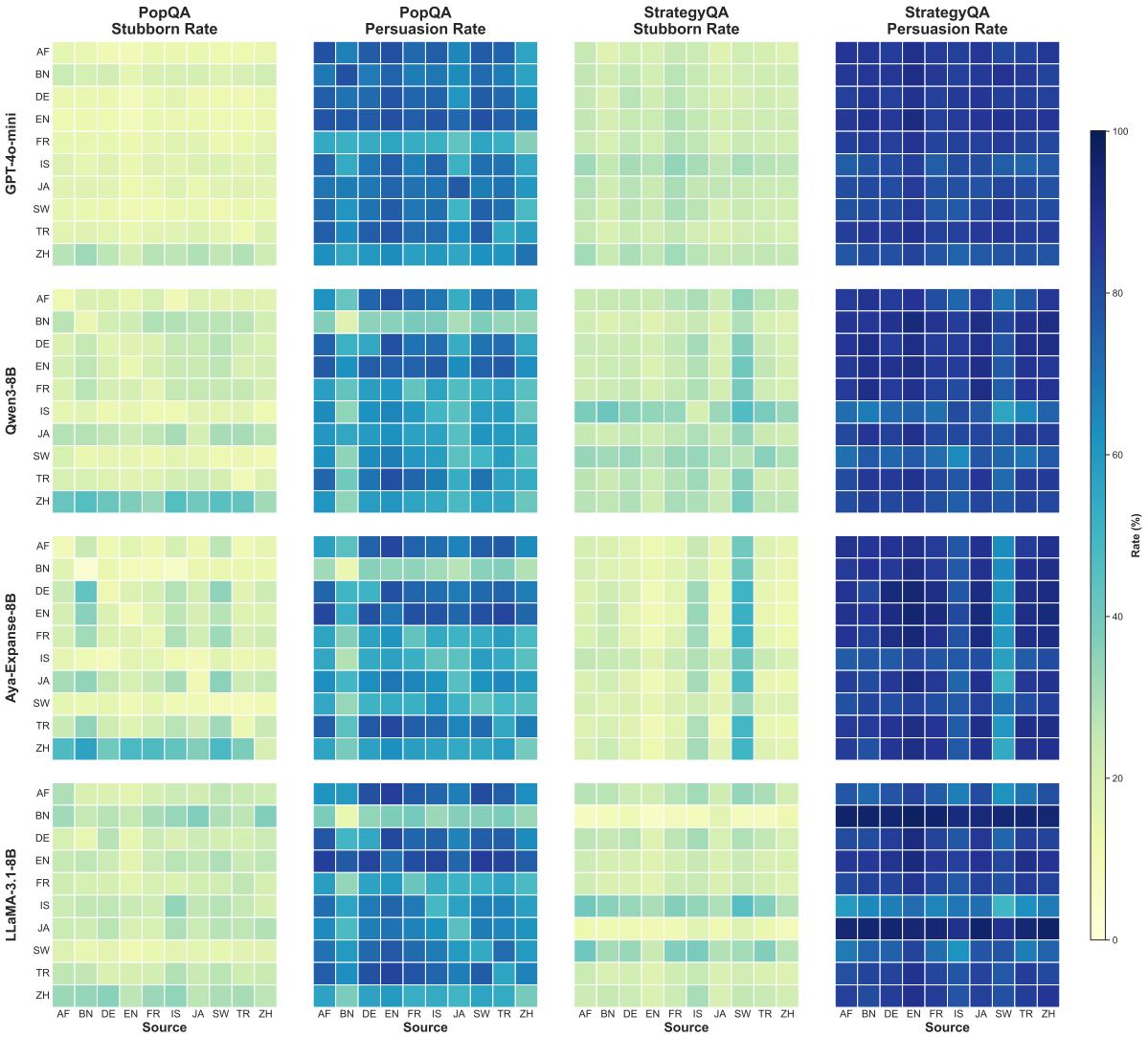


Figure 5: Task 3 results under the NLLB-200-3.3B pipeline. The figure reports per-model Stubborn Rate and Persuasion Rate on PopQA and StrategyQA across four representative models.

mark remains semantically well preserved across languages, especially on the main QA datasets. EN is perfect by construction as a sanity check, and among non-English languages AF and SW show particularly strong reconstruction on PopQA and StrategyQA.

The remaining variation is systematic rather than fatal. PopQA and StrategyQA are easier to recover than ConflictingQA and KNOT across all non-English languages, while JA, ZH, and to a lesser extent BN degrade more on the harder sets. Taken together, these results indicate that the NLLB-translated benchmark is of sufficiently high quality for the main multilingual evaluation, and Appendix C.2 further shows that the core task-dependent ordering persists under Gemini-2.5-Pro, confirming that our main findings are not an artifact of the NLLB pipeline.

C.2 Translator-Swap Ablation with Gemini-2.5-Pro

All main experiments in the main text use **NLLB-200-3.3B** as the translation pipeline. A natural concern is that the task-dependent split might still depend on this specific translator rather than on model behavior under cross-lingual conflict. To test this possibility, we rebuild the multilingual benchmark with **Gemini-2.5-Pro** while keeping the task definitions, prompting protocol, and judgment pipeline fixed.

We report the full raw Task 3 Stubborn Rate and Persuasion Rate matrices for the two translation pipelines in Tables 8 and 9. These appendix tables expose the original per-model, per-language values rather than only the averaged deltas used in the main text.

The raw matrices confirm the same qualitative

Dataset	Metric	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
ConflictingQA	chrF	48.829	40.275	52.245	100.000	51.245	39.580	35.471	49.100	46.255	33.408
	BLEU	26.850	17.819	28.759	100.000	28.071	17.844	14.206	25.828	23.755	12.975
	BERTScore	0.9195	0.9017	0.9180	1.0000	0.9172	0.9048	0.8949	0.9145	0.9127	0.8950
	Sim	0.8924	0.8585	0.8661	1.0000	0.8806	0.8287	0.8324	0.8899	0.8552	0.8519
KNOT-Explicit	chrF	60.247	51.480	56.043	100.000	60.657	52.513	47.876	61.500	52.089	45.831
	BLEU	40.536	25.989	33.475	100.000	39.279	31.016	22.725	40.898	28.258	20.745
	BERTScore	0.9533	0.9323	0.9483	1.0000	0.9540	0.9444	0.9294	0.9554	0.9431	0.9251
	Sim	0.9080	0.8762	0.8834	1.0000	0.9021	0.8656	0.8734	0.9213	0.8813	0.8604
KNOT-Implicit	chrF	61.003	55.079	57.777	100.000	59.165	52.441	47.916	63.249	54.890	49.837
	BLEU	40.903	30.694	35.754	100.000	38.237	30.839	23.337	43.426	31.812	25.871
	BERTScore	0.9544	0.9363	0.9510	1.0000	0.9518	0.9453	0.9285	0.9579	0.9462	0.9293
	Sim	0.9117	0.8778	0.8896	1.0000	0.9015	0.8622	0.8661	0.9198	0.8877	0.8734
KNOT-Simple	chrF	60.401	52.997	56.641	100.000	59.868	51.115	45.881	61.615	53.326	49.511
	BLEU	41.108	29.223	35.623	100.000	39.405	29.716	21.821	42.615	31.796	25.659
	BERTScore	0.9548	0.9366	0.9518	1.0000	0.9559	0.9453	0.9290	0.9588	0.9479	0.9325
	Sim	0.9052	0.8766	0.8827	1.0000	0.9007	0.8609	0.8624	0.9158	0.8818	0.8641
PopQA	chrF	67.899	65.266	63.665	100.000	60.955	57.002	58.865	74.412	61.183	58.268
	BLEU	53.377	44.684	44.465	100.000	42.286	37.351	35.085	59.906	41.426	34.719
	BERTScore	0.9616	0.9514	0.9557	1.0000	0.9516	0.9482	0.9463	0.9685	0.9519	0.9446
	Sim	0.8988	0.8727	0.8810	1.0000	0.8921	0.8535	0.8793	0.9183	0.8768	0.8787
StrategyQA	chrF	68.810	68.234	69.380	100.000	67.300	59.407	60.872	73.538	64.064	59.992
	BLEU	50.774	48.082	50.163	100.000	48.330	38.995	37.308	56.527	42.880	37.994
	BERTScore	0.9676	0.9637	0.9678	1.0000	0.9643	0.9552	0.9539	0.9723	0.9603	0.9517
	Sim	0.8971	0.8728	0.8734	1.0000	0.8897	0.8562	0.8791	0.9072	0.8711	0.8615

Table 7: **Back-translation quality control for the NLLB-200-3.3B pipeline.** For each translated dataset, we back-translate the target-language text into English and compare it against the original English source using chrF, BLEU, BERTScore, and cross-lingual semantic similarity.

Model	Stubborn Rate										Persuasion Rate									
	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
PopQA																				
GPT-4o-mini	16.3	15.6	15.6	12.8	15.6	16.2	16.4	15.2	14.9	16.1	69.0	62.1	68.0	72.0	68.6	68.1	58.1	69.7	67.5	55.4
Gemini-2.5-Flash	13.0	15.2	13.4	12.1	13.3	12.4	16.0	13.1	13.1	15.1	67.5	60.6	65.5	70.6	68.4	68.1	62.1	70.4	67.0	61.2
Llama-3.1-8B	23.7	21.6	22.2	18.0	22.1	24.0	23.5	26.2	23.7	23.7	64.8	53.0	64.9	68.4	66.0	62.1	57.5	64.6	63.5	57.6
Qwen3-8B	22.8	24.9	21.1	19.5	20.8	24.3	23.9	25.7	23.3	19.3	62.2	43.9	61.0	64.9	61.7	60.1	48.0	60.5	59.3	49.3
Qwen3-80B	19.9	17.5	17.8	15.3	18.9	17.8	17.7	20.7	17.4	17.1	63.0	55.8	62.7	66.9	63.7	61.7	55.9	64.8	61.5	54.3
Aya-Expanse-8B	24.0	33.2	19.6	19.2	20.8	26.7	20.3	31.2	18.0	20.0	59.4	41.7	59.7	63.3	61.1	60.5	54.2	61.3	60.3	54.1
Average	20.0	21.3	18.3	16.2	18.6	20.3	19.6	22.0	18.4	18.5	64.3	52.8	63.6	67.7	64.9	63.4	56.0	65.2	63.2	55.3
StrategyQA																				
GPT-4o-mini	27.1	21.0	26.4	23.3	28.8	25.3	21.2	23.7	23.4	22.0	81.3	83.4	82.9	87.0	81.1	81.7	83.6	82.8	81.5	82.3
Gemini-2.5-Flash	27.1	23.6	26.6	23.6	28.7	27.1	23.4	23.5	25.8	24.1	83.4	85.4	82.6	87.4	82.2	82.5	85.8	85.6	83.8	85.2
Llama-3.1-8B	24.3	23.6	21.8	18.0	23.8	25.6	21.4	25.5	23.8	19.2	80.5	79.4	82.6	87.2	81.6	77.9	82.7	75.7	80.9	82.7
Qwen3-8B	26.9	26.7	27.5	23.8	27.7	29.8	23.4	37.5	27.1	23.2	80.6	83.9	82.3	85.6	80.0	77.9	85.2	73.6	81.0	84.5
Qwen3-80B	39.9	35.4	42.4	40.0	44.9	35.9	32.9	34.8	34.7	34.8	75.5	77.7	77.5	83.0	74.0	75.4	79.4	75.6	76.3	80.0
Aya-Expanse-8B	20.2	22.4	17.8	12.7	17.4	29.1	14.5	47.2	17.0	14.5	83.9	82.4	85.5	89.4	85.3	76.9	88.1	59.5	85.6	88.4
Average	27.6	25.5	27.1	23.6	28.6	28.8	22.8	32.0	25.3	23.0	80.9	82.0	82.2	86.6	80.7	78.7	84.1	75.5	81.5	83.9

Table 8: **Task 3 source-language mean Stubborn Rate and Persuasion Rate under the NLLB-200-3.3B translation pipeline.** Each cell reports the source-language column mean for a given model. The *Average* row reports the mean over the 6 evaluation models.

ordering under both translators. High-resource non-Latin languages, especially **ZH** and **JA**, remain much weaker on PopQA than on StrategyQA, while **SW** shows substantially milder degradation. This persistence indicates that the entity-side weakness of the non-Latin cluster is not an artifact of the

Model	Stubborn Rate										Persuasion Rate									
	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
PopQA																				
GPT-4o-mini	10.8	10.7	10.7	10.3	10.5	10.5	10.5	10.9	10.1	10.4	79.5	69.7	78.8	79.5	78.7	78.7	67.3	76.5	77.6	64.6
Gemini-2.5-Flash	9.7	9.7	9.8	9.7	9.9	9.7	10.2	10.3	10.0	10.2	78.0	69.8	79.0	78.1	79.0	77.8	72.2	78.2	78.0	71.7
Llama-3.1-8B	23.1	23.7	22.5	22.8	22.7	24.3	22.6	23.1	23.0	22.7	69.3	62.6	71.3	70.2	69.8	67.3	63.0	70.1	71.0	61.0
Qwen3-8B	14.8	15.8	14.9	15.3	14.9	15.7	15.4	15.3	14.5	15.4	74.7	57.7	75.5	76.3	75.0	72.4	60.8	70.0	73.9	59.0
Qwen3-80B	14.8	17.3	16.2	17.1	16.3	15.9	17.7	18.5	15.3	16.8	78.0	67.0	78.7	78.2	77.5	76.1	70.2	76.9	77.1	65.4
Aya-Expanse-8B	18.4	19.3	17.4	17.6	18.9	19.3	19.5	20.6	18.1	19.3	78.5	51.9	77.6	77.8	78.2	74.6	68.5	72.9	75.6	64.6
Average	15.3	16.1	15.2	15.5	15.5	15.9	16.0	16.5	15.2	15.8	76.3	63.1	76.8	76.7	76.4	74.5	67.0	74.1	75.5	64.4
StrategyQA																				
GPT-4o-mini	17.5	17.0	18.3	19.9	18.8	17.4	16.7	18.6	16.7	16.6	89.5	90.1	90.9	90.7	89.7	90.4	89.7	88.3	88.3	90.4
Gemini-2.5-Flash	22.1	21.2	22.3	23.2	22.2	21.7	21.2	22.6	21.6	22.2	90.2	90.6	90.6	90.6	90.5	90.4	91.3	89.8	90.4	90.0
Llama-3.1-8B	21.7	22.4	18.2	19.4	18.8	24.7	17.1	27.3	20.2	17.5	83.3	81.4	88.0	89.1	86.9	79.9	88.2	75.6	84.0	87.3
Qwen3-8B	19.3	21.3	20.9	21.8	21.6	24.9	20.5	32.4	21.5	21.4	89.0	89.6	90.0	90.0	89.4	84.7	89.8	79.1	87.4	89.9
Qwen3-80B	34.9	35.1	39.2	41.1	41.1	30.2	34.3	32.1	33.0	36.6	82.9	82.1	83.4	84.3	82.3	83.5	84.7	80.0	81.7	84.2
Aya-Expanse-8B	14.9	19.6	12.5	12.1	13.1	25.1	13.0	46.3	13.8	13.4	87.3	82.1	89.6	89.4	90.1	78.5	90.0	59.6	89.1	90.6
Average	21.8	22.8	21.9	22.9	22.6	24.0	20.5	29.9	21.1	21.3	87.1	86.0	88.8	89.0	88.1	84.5	89.0	78.7	86.8	88.7

Table 9: **Task 3 source-language mean Stubborn Rate and Persuasion Rate under the Gemini-2.5-Pro translation pipeline.** Each cell reports the source-language column mean for a given model. The *Average* row reports the mean over the 6 evaluation models.

NLLB-based construction.

To make the translator comparison easier to read, we summarize the raw rates in percentage-point (pp) tables. The underlying AF-ZH values correspond to the *source-language columns* in Task 3, and each such source-language value is itself computed by averaging over its 10 target-language evaluations. Table 10 keeps one source language fixed and averages over the six evaluated models.

Under both translators, **ZH**, **JA**, and **BN** still show the largest negative ΔPR values, while **SW** remains the mildest case. In other words, the translator swap affects calibration, but not the central conclusion that reasoning-heavy conflicts favor resource-rich languages whereas entity-centric conflicts remain most fragile for the non-Latin cluster.

C.3 Language Coverage and Resource Gradient

Our language design aims to make cross-lingual comparisons structured rather than anecdotal. The 10 evaluation languages are Afrikaans (AF), Bengali (BN), German (DE), English (EN), French (FR), Icelandic (IS), Japanese (JA), Swahili (SW), Turkish (TR), and Chinese (ZH). They span five language families and three script systems: a Latin-script cluster (AF, DE, EN, FR, IS, SW, TR), the Bengali script (BN), and a CJK cluster (JA, ZH). This design allows us to compare high-resource but script-distant languages against medium- and low-resource but script-compatible ones within a single controlled setting.

Because exact pretraining language mixtures are unavailable for most evaluated models, we use the CC-100 corpus size as a standardized proxy for relative resource abundance. This definition yields a reproducible three-level resource gradient (high / medium / low) while avoiding unsupported claims about the true pretraining composition of any individual model. Table 11 summarizes the resource gradient used throughout the paper.

C.4 Tokenizer-level Overlap as a Proxy for Representational Compatibility

To provide a measurable proxy for cross-lingual representational compatibility, we compute the pairwise Jaccard overlap between unique subword ID sets on the FLORES-200 devtest across the 10 evaluation languages. Table 12 reports the summary statistics averaged over the tokenizer matrices of our six evaluation models. For the closed-source Gemini-2.5-Flash, we utilize the Gemma-3 tokenizer as a representational proxy, as they share a closely aligned architectural lineage and vocabulary specification.

The averaged matrix preserves the same broad structure as the model-specific results. Overlap remains much higher inside the Latin-script cluster

Translator	Lang	PopQA SR	StrategyQA SR	ΔSR (pp)	PopQA PR	StrategyQA PR	ΔPR (pp)
Gemini-2.5-Pro	AF	15.28	21.75	-6.48	76.33	87.05	-10.73
Gemini-2.5-Pro	BN	16.08	22.76	-6.67	63.10	85.99	-22.90
Gemini-2.5-Pro	DE	15.25	21.90	-6.65	76.82	88.76	-11.94
Gemini-2.5-Pro	EN	15.47	22.93	-7.45	76.68	89.01	-12.33
Gemini-2.5-Pro	FR	15.53	22.60	-7.07	76.36	88.15	-11.78
Gemini-2.5-Pro	IS	15.89	24.01	-8.11	74.47	84.55	-10.07
Gemini-2.5-Pro	JA	15.98	20.46	-4.48	67.02	88.97	-21.96
Gemini-2.5-Pro	SW	16.46	29.91	-13.45	74.11	78.73	-4.62
Gemini-2.5-Pro	TR	15.18	21.14	-5.95	75.54	86.82	-11.28
Gemini-2.5-Pro	ZH	15.79	21.29	-5.50	64.38	88.72	-24.34
NLLB-200-3.3B	AF	19.96	27.60	-7.64	64.32	80.86	-16.53
NLLB-200-3.3B	BN	21.34	25.45	-4.11	52.84	82.03	-29.18
NLLB-200-3.3B	DE	18.28	27.08	-8.81	63.63	82.24	-18.61
NLLB-200-3.3B	EN	16.15	23.57	-7.42	67.69	86.60	-18.90
NLLB-200-3.3B	FR	18.60	28.56	-9.96	64.94	80.71	-15.77
NLLB-200-3.3B	IS	20.25	28.78	-8.53	63.44	78.71	-15.27
NLLB-200-3.3B	JA	19.64	22.81	-3.18	55.98	84.11	-28.13
NLLB-200-3.3B	SW	22.03	32.03	-10.00	65.24	75.45	-10.21
NLLB-200-3.3B	TR	18.41	25.31	-6.90	63.19	81.52	-18.34
NLLB-200-3.3B	ZH	18.54	22.98	-4.44	55.32	83.85	-28.53

Table 10: **Task 3 translator ablation summarized by source language.** Each row averages over the six evaluated models for one translator-language pair. Δ is computed as PopQA−StrategyQA in percentage points (pp).

ISO	Language	Class	Tokens (M)	Size (GiB)
AF	Afrikaans	Low-resource	242	1.3
BN	Bengali	Medium-resource	525	8.4
DE	German	High-resource	10,297	66.6
EN	English	High-resource	55,608	300.8
FR	French	High-resource	9,780	56.8
IS	Icelandic	Low-resource	505	3.2
JA	Japanese	High-resource	530	69.3
SW	Swahili	Low-resource	275	1.6
TR	Turkish	Medium-resource	2,736	20.9
ZH	Chinese	High-resource	259	46.9

Table 11: **CC-100-based resource gradient for the 10 evaluation languages.** We use corpus size in CC-100 as a standardized proxy for relative resource abundance. These values are not intended to recover the exact pre-training mixture of any evaluated model.

	AF	BN	DE	EN	FR	IS	JA	SW	TR	ZH
AF	1.000	0.033	0.268	0.176	0.205	0.284	0.020	0.266	0.220	0.049
BN	0.033	1.000	0.026	0.023	0.023	0.042	0.034	0.042	0.028	0.031
DE	0.268	0.026	1.000	0.135	0.171	0.230	0.018	0.197	0.186	0.044
EN	0.176	0.023	0.135	1.000	0.172	0.148	0.016	0.153	0.128	0.040
FR	0.205	0.023	0.171	0.172	1.000	0.184	0.018	0.177	0.158	0.041
IS	0.284	0.042	0.230	0.148	0.184	1.000	0.024	0.277	0.237	0.055
JA	0.020	0.034	0.018	0.016	0.018	0.024	1.000	0.023	0.018	0.187
SW	0.266	0.042	0.197	0.153	0.177	0.277	0.023	1.000	0.242	0.054
TR	0.220	0.028	0.186	0.128	0.158	0.237	0.018	0.242	1.000	0.044
ZH	0.049	0.031	0.044	0.040	0.041	0.055	0.187	0.054	0.044	1.000

Table 12: **Average tokenizer overlap across six evaluation models.** Each cell represents the mean Jaccard overlap for the corresponding language pair, calculated over the 10×10 overlap matrices of the six models used in our study. For the closed-source Gemini-2.5-Flash, the Gemma-3 tokenizer is used as a proxy. All overlaps are computed on the FLORES-200 devtest.

Outcome	Proxy	ρ_S	p	OLS slope	OLS R^2	OLS p
PopQA PR	$\log_{10}$ CC-100 size (GiB)	0.055	0.881	-0.17	0.001	0.943
PopQA PR	Mean tokenizer overlap	0.527	0.117	81.87	0.723	0.002
StrategyQA PR	$\log_{10}$ CC-100 size (GiB)	0.782	0.008	3.14	0.660	0.004
StrategyQA PR	Mean tokenizer overlap	-0.661	0.038	-33.62	0.327	0.084
ΔPR	$\log_{10}$ CC-100 size (GiB)	0.479	0.162	3.30	0.164	0.246
ΔPR	Mean tokenizer overlap	-0.830	0.003	-115.49	0.865	<0.001

Table 13: **Simple association analysis for the two proxy axes.** Outcomes are source-language mean Task 3 persuasion rates averaged over the six evaluated models; ΔPR denotes StrategyQA persuasion minus PopQA persuasion. For each outcome–proxy pair, we report a Spearman rank correlation and a separate univariate OLS fit. The resource proxy uses $\log_{10}$ CC-100 corpus size in GiB from Table 11, and the similarity proxy uses the mean tokenizer-overlap score derived from Table 12. These tests quantify association rather than causal identification.

than across script boundaries, while the JA-ZH pair stays comparatively high within the non-Latin region. These averaged values do not by themselves establish causality, but they are consistent with our interpretation that entity-centric cross-lingual rescue depends in part on how easily evidence and answer forms can be aligned in the model's internal representation space.

C.5 Quantitative Association Analysis for the Two Proxy Axes

To complement the main qualitative pattern with a simple quantitative check, we test whether the two proposed proxy axes align with source-

Dataset	Model	$Q = \text{Pos}$					$Q = \text{Neg}$				
		TT	TF	FT	FF	TF−FT	TT	TF	FT	FF	TF−FT
PopQA	GPT-4o-mini	66.0	66.4	59.5	50.1	+6.9	29.1	19.8	22.3	11.9	–2.5
	Gemini-2.5-Flash	79.5	74.6	72.9	65.2	+1.7	37.6	27.2	32.1	16.6	–4.9
	Qwen3-8B	69.1	64.5	55.2	53.8	+9.3	33.1	17.8	23.7	9.9	–5.9
	Qwen3-80B	80.2	77.1	73.9	66.1	+3.2	47.4	31.2	39.0	17.8	–7.8
	Aya-Expanse-8B	66.8	72.4	57.1	59.6	+15.3	37.0	28.9	27.0	18.2	+1.9
	Llama-3.1-8B	77.3	77.0	68.3	61.4	+8.7	50.5	41.6	41.8	26.7	–0.2
StrategyQA	GPT-4o-mini	77.0	67.4	64.7	52.7	+2.7	53.5	34.0	37.1	23.7	–3.1
	Gemini-2.5-Flash	82.0	66.0	63.6	47.8	+2.4	58.0	35.4	38.0	19.4	–2.6
	Qwen3-8B	79.9	73.4	66.8	56.6	+6.7	55.5	36.7	38.9	26.6	–2.2
	Qwen3-80B	85.0	65.7	61.3	42.2	+4.4	72.2	40.0	43.3	19.0	–3.3
	Aya-Expanse-8B	73.5	73.4	66.8	60.5	+6.6	45.7	36.8	33.3	29.8	+3.5
	Llama-3.1-8B	74.2	71.5	55.7	52.9	+15.8	50.5	31.0	44.8	26.0	–13.8

Table 14: **Accuracy (%) under two-source evidence competition on PopQA and StrategyQA.** $Q = \text{Pos/Neg}$ indicates whether the query language matches Evidence 1 or Evidence 2, respectively. TT/TF/FT/FF denote the truthfulness of whether the model correctly recalls the memorized answer when the query language matches Evidence 1 and Evidence 2, respectively; TF−FT compares the mixed-evidence conditions under the same query-language setting. All values are percentages, rounded to one decimal place.

language Task 3 persuasion rates in a systematic way. For each source language ℓ, we define the dataset-specific source-language mean as $\mathrm{PR}_d(\ell) = \frac{1}{6}\sum_m \mathrm{PR}_d(m, \ell)$, where $d \in \{\text{PopQA}, \text{StrategyQA}\}$ and the average is taken over the six evaluated models. We then define the task gap as $\Delta\mathrm{PR}(\ell) = \mathrm{PR}_{\text{StrategyQA}}(\ell) - \mathrm{PR}_{\text{PopQA}}(\ell)$. For the resource axis, we use $\log_{10}$ CC-100 corpus size in GiB from Table 11; we prefer corpus size rather than token count here because token counts vary more strongly with script segmentation. For the similarity axis, we use the mean tokenizer-overlap score $\mathrm{Sim}(\ell) = \frac{1}{9}\sum_{\ell' \neq \ell} J(\ell, \ell')$ derived from Table 12. Given that only 10 source languages are available, we report two lightweight tests for each outcome-proxy pair: a Spearman rank correlation and a separate univariate OLS fit. These analyses provide quantitative support for the proposed two-pathway interpretation, but they should not be read as stand-alone causal identification.

The quantitative pattern is consistent with the main story. Higher tokenizer overlap is associated with higher PopQA persuasion and with a smaller StrategyQA-PopQA gap, which supports the view that entity-centric rescue depends more on representational compatibility. In contrast, the resource proxy is most clearly associated with StrategyQA persuasion, while showing little relation to PopQA persuasion, which matches our claim that reasoning-heavy conflict resolution benefits more from resource-rich knowledge pathways. Because these fits are bivariate and use only 10 languages, we treat them as supporting evidence rather than as a stand-alone causal claim.

C.6 Interplay between Parametric Memory and Cross-lingual Conflicts

To further isolate the role of parametric memory, we analyze Task 4 QA results under four parametric-memory quadrants (TT, TF, FT, FF), reported in Table 14, and relate these effects to the aggregate trends in Figure 4.

Alignment-driven parametric robustness. We observe a strong association between internal cross-lingual alignment and stability under conflict.

- **TT (aligned and correct).** Accuracy is highest when parametric knowledge is correct in both languages (TT), indicating that cross-lingually aligned memory provides a reliable anchor that improves resistance to interference.
- **FF (misaligned and incorrect).** Accuracy is lowest in FF, where parametric memory provides no dependable reference in either language, making the model more susceptible to misleading evidence.

Query-memory congruence interference. Beyond alignment, we find a consistent *query-memory congruence* effect: the query language preferentially activates the corresponding slice of parametric memory, which then biases evidence selection under competition.

- **When $Q = \text{Pos}$.** TF (correct memory in Pos) outperforms FT (incorrect memory in Pos). When the query triggers a correct belief, it cooperates with truthful evidence; when it triggers an incorrect belief, it can suppress truthful evidence from other languages.
- **When $Q = \text{Neg}$.** The pattern flips: FT out-

performs TF, mirroring the same mechanism when the query aligns with the negative (interfering) side.

Discussion. These results indicate that multi-source cross-lingual conflict is not a purely evidence-level competition. Instead, the query language actively gates which parametric belief becomes salient, and this *query-conditioned memory activation* can systematically steer source selection. Consequently, improving cross-lingual robustness requires not only better evidence integration but also stronger internal alignment of parametric knowledge across languages.