When Abundance Conceals Weakness: Knowledge Conflict in Multilingual Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) encode vast world knowledge across multiple languages, yet their internal beliefs are often unevenly distributed across linguistic spaces. When external evidence contradicts these language-dependent memories, models encounter *cross-lingual knowledge conflict*, a phenomenon largely unexplored beyond English-centric settings. We introduce **CLEAR**, a **C**ross-**L**ingual knowl**E**dge conflict ev**A**luation f**R**amework that systematically examines how multilingual LLMs reconcile conflicting internal beliefs and multilingual external evidence. CLEAR decomposes conflict resolution into four progressive scenarios, from multilingual parametric elicitation to competitive multi-source cross-lingual induction, and systematically evaluates model behavior across two complementary QA benchmarks with distinct task characteristics. We construct multilingual versions of ConflictQA and ConflictingQA covering 10 typologically diverse languages and evaluate six representative LLMs. Our experiments reveal a task-dependent decision dichotomy. In reasoning-intensive tasks, conflict resolution is dominated by language resource abundance, with high-resource languages exerting stronger persuasive power. In contrast, for entity-centric factual conflicts, linguistic affinity, not resource scale, becomes decisive, allowing low-resource but linguistically aligned languages to outperform distant high-resource ones.

1 Introduction

Large Language Models (LLMs) are trained on vast corpora and encode substantial world knowledge in their parameters (Hurst et al., 2024; Team et al., 2023; Yang et al., 2025; Grattafiori et al., 2024). In practice, however, modern LLM systems rarely rely on parametric memory alone. To mitigate errors and hallucinations, external information is commonly injected at inference time–most prominently via Retrieval-Augmented Generation (RAG) (Chen et al., 2022; Cattan et al., 2025; Park and Lee, 2025), where retrieved documents provide additional evidence for answering user queries.

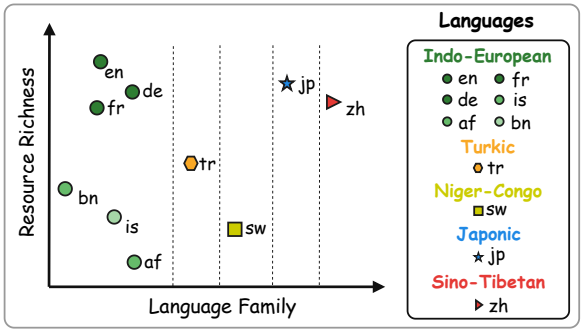


Figure 1: language distribution in the CLEAR framework: languages are mapped based on their resource richness and taxonomic family, enabling a systematic study of how linguistic affinity and data scale influence cross-lingual knowledge conflict resolution.

Yet, introducing external evidence gives rise to a critical failure mode: **knowledge conflict** (Chen et al., 2022), in which a model's internally stored belief contradicts information presented in context. How LLMs react to and resolve such conflicts is central to system robustness and faithfulness, particularly in high-stakes or evidence-grounded applications. Crucially, these systems are increasingly **multilingual**. In real-world RAG pipelines, queries and retrieved sources often span multiple languages, requiring models to reconcile evidence across linguistic boundaries while producing coherent outputs. Despite this reality, existing studies remain largely English-centric, leaving the mechanisms underlying **Cross-Lingual Knowledge Conflict (CLKC)** largely unexplored. The challenge of CLKC is threefold.

- **Multilingual LLMs exhibit language-conditioned memories**: knowledge correct in one language may be incomplete or incorrect in another (Kassner et al., 2021). Prior multilingual studies focus on source-language

preference (Park and Lee, 2025), implicitly assuming fixed internal beliefs, leaving unclear how query language activates parametric memory and when multilingual evidence overrides or reinforces it.

- **Existing alignment studies often conflate distinct knowledge types.** While cross-lingual consistency is typically measured at the output level (Wang et al., 2025a), *entity-centric factual knowledge* differs fundamentally from that requiring *multi-step logical reasoning*, and their interaction with multilingual evidence remains underexplored.
- **Research on knowledge conflict has been overwhelmingly monolingual** (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2024; Jin et al., 2024), particularly centering on English-centric scenarios. Consequently, it remains unclear 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 navigate knowledge conflicts. As illustrated in Figure 1, CLEAR spans 10 languages chosen to vary in both *resource abundance* and *linguistic affinity*, enabling controlled analysis of whether cross-lingual decision-making is driven primarily by training data scale or structural proximity between languages. Unlike prior work that treats knowledge conflict as binary, CLEAR decomposes conflict resolution into four progressive tasks, ranging from multilingual parametric elicitation to competitive multi-source cross-lingual induction, and evaluates model behavior on two QA benchmarks with different conflict patterns.

We conduct extensive experiments on two newly curated multilingual benchmarks: ConflictQA-PopQA and ConflictQA-StrategyQA, covering 10 typologically diverse languages. Our results reveal a task-dependent decision dichotomy: in reasoning-intensive tasks, conflict resolution is dominated by language resource abundance, whereas in entity-centric factual tasks, linguistic affinity, rather than data scale, emerges as the primary driver of persuasion. Notably, low-resource but linguistically aligned languages can exert stronger influence than distant high-resource languages, exposing an abundance-weakness paradox in multilingual LLMs. The contributions of this work are summarized as follows:

- **Cross-lingual Knowledge Conflict (CLKC).** We introduce CLKC as a new evaluation paradigm that reframes knowledge conflict as a tension between *language-dependent parametric beliefs* and *multilingual external evidence*, enabling systematic analysis of belief activation and revision across languages.
- **Task-dependent cross-lingual conflict resolution.** We uncover a novel behavioral pattern in CLKC: conflict resolution in reasoning-intensive tasks is driven primarily by *language resource abundance*, whereas entity-centric factual conflicts are governed by *linguistic affinity* rather than data scale.
- **Multilingual conflict benchmarks.** We construct multilingual versions of *ConflictQA-PopQA* and *ConflictQA-StrategyQA* across 10 typologically diverse languages, enabling controlled comparison of cross-lingual knowledge conflict in entity-centric and reasoning-intensive QA settings.

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 evi-

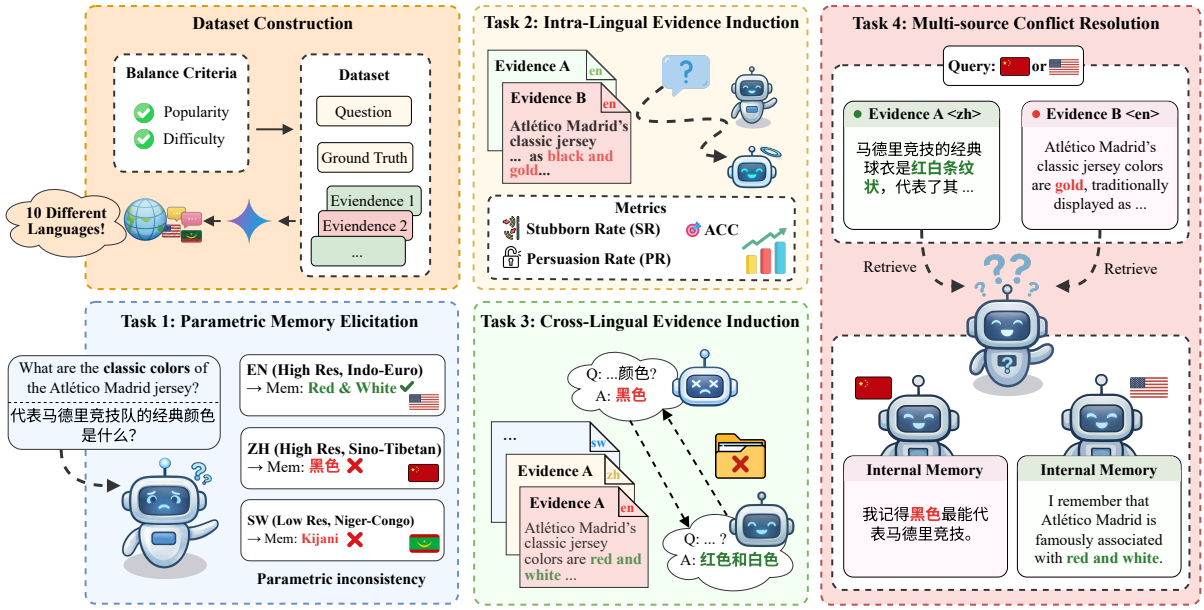


Figure 2: Overview of the **CLEAR** framework.

dence 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 eliciting multilingual parametric memory from LLMs, constructing controlled counter-memory across languages, and evaluating how models resolve cross-lingual knowledge conflicts. Our design ensures that all conflicts are grounded in the model's own internal beliefs rather than annotation artifacts.

3.1 Multilingual Dataset Construction

Question Answering (QA) Datasets. Following prior work (Longpre et al., 2021; Chen et al., 2022), we adopt the QA task as the primary testbed. We employ a QA benchmark suite based on **ConflictQA** (Xie et al., 2024), which is derived from PopQA (Mallen et al., 2023) (entity-centric factual knowledge) and StrategyQA (Geva et al., 2021) (commonsense questions with higher reasoning demands).

From ConflictQA, we curate high-quality subsets by filtering for knowledge popularity, question type, and difficulty:

- **ConflictQA-PopQA** includes 898 entity-centric queries with Wikipedia-based evidence, refined via human annotation to ensure strong entity-level contradictions.
- **ConflictQA-StrategyQA** contains 1,000 reasoning-intensive samples, where external evidence is synthesized by LLMs to preserve fluency while minimizing interference with implicit reasoning.

Multilingual Dataset Construction. To study cross-lingual dynamics, we translate all benchmarks into 10 diverse languages:

$$\mathcal{L} = \{\text{af, bn, de, fr, is, ja, sw, zh, tr, en}\},$$

covering five language families and a wide range of resource levels (Joshi et al., 2020). Translation is performed using Gemini-2.5-Pro (Team et al., 2023), followed by human verification to ensure semantic fidelity.

For entity-centric tasks, we apply an *entity-presence constraint*: the ground-truth entity must appear explicitly in supportive evidence and be absent or replaced in conflicting evidence. This guarantees that model decisions reflect a genuine

choice between *parametric memory* and *external cues*, rather than surface ambiguity.

3.2 Tasks

CLEAR decomposes cross-lingual knowledge conflict into four progressively complex tasks, each isolating interactions between language-dependent parametric memory and external evidence.

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 within a single (typically English) language, 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 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.

Task 4: Multi-Source Conflict Resolution. To simulate realistic, high-pressure settings, we present the model with **two explicitly contradictory evidence sources** expressed in different languages. Using a symmetric 2×2 permutation design for each language pair, we control for query-language bias and isolate source competition.

This task investigates the dynamic interplay between parametric memory and multilingual evidence competition, focusing on two states:

(1) **Memory-Supportive Competition**, where one source aligns with internal belief,

(2) **Memory-Conflicting Competition**, where internal belief is incorrect or absent.

By analyzing outcomes across language pairs, we test whether certain languages exhibit inherent dominance that can override both memory and competing sources.

3.3 Evaluation Metrics

We define three metrics to quantify the frequency of LLMs adhering to their parametric memory or being influenced by 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, 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 $\mathcal{C}$.

4 Experiments

4.1 Experimental Setup

To investigate cross-lingual knowledge conflict mechanisms, we evaluate our framework across a diverse set of Large Language Models (LLMs), encompassing both proprietary and open-source systems. For proprietary models, we include **GPT-4o mini** (Hurst et al., 2024) and **Gemini 2.5 Flash**(Team et al., 2023), both accessed through the **OpenRouter API** to ensure consistent response generation. For open-source models, we select representative systems spanning various parameter scales and multilingual optimizations: **Qwen3-8B**, **Qwen3-80B** (Yang et al., 2025), **Llama-3.1-8B** (Grattafiori et al., 2024), and **Aya Expanse 8B** (Üstün et al., 2024).

All open-source models are deployed and executed on a single **NVIDIA A100 GPU (80GB)**. Notably, for models equipped with advanced reasoning or "thinking" capabilities, such as Qwen3-8B, we explicitly **disable the thinking process** during evaluation. This ensures that the model's responses reflect its direct parametric knowledge and its immediate reaction to linguistic conflicts, rather than an iterative self-correction process that could mask the underlying cross-lingual misalignment.

To robustly evaluate multilingual outputs under minor surface variations (e.g., spelling and name formatting), we adopt an **LLM-as-a-judge** protocol. We use **Gemini-2.5-Flash** as the judge to determine whether the model output a_{out} matches the language-specific ground truth y^L for each query. Table 1 summarizes closed-book answer accuracy for each model across languages. Additional implementation details can be found in Appendix B.

4.2 Dataset Disparities: Facts vs. Reasoning

Our results in Table 2 reveal a significant disparity in how LLMs reconcile conflicts across different knowledge types. In the entity-centric **PopQA**, models exhibit a relatively low average Stubborn Rate (SR) of 13.4%, coupled with a high Persuasion Rate (PR) of 81.0%. This suggests that for simple factual entities, LLMs are highly susceptible to external evidence, even when it directly contradicts their internal parametric beliefs.

Models	af	bn	de	en	fr	is	ja	sw	tr	zh
PopQA										
GPT-4o-mini	47.3	26.1	51.4	52.7	49.7	41.0	34.4	40.5	43.8	27.6
Gemini-2.5-Flash	54.1	42.2	58.7	59.9	56.3	53.9	45.2	49.8	51.4	40.1
LLaMA-3.1-8B	34.0	15.4	39.4	43.5	36.6	23.9	18.7	26.5	29.4	19.2
Qwen3-8B	23.2	10.6	29.1	32.6	30.2	14.0	17.8	14.8	21.7	23.9
Qwen3-80B	45.7	22.0	47.8	53.8	47.6	35.0	27.4	36.0	37.8	34.6
Aya-Expanse-8B	30.0	10.1	35.3	38.4	35.6	16.3	22.3	18.6	32.2	23.9
StrategyQA										
GPT-4o-mini	68.1	68.2	70.8	74.1	71.6	65.8	70.4	68.7	68.1	70.0
Gemini-2.5-Flash	64.7	62.6	67.7	71.1	65.2	62.3	64.9	64.4	64.3	64.2
LLaMA-3.1-8B	51.4	54.4	54.9	65.6	58.3	47.0	55.5	50.8	54.3	61.7
Qwen3-8B	53.2	52.4	57.7	65.5	60.4	47.7	54.0	48.0	52.6	57.5
Qwen3-80B	55.7	59.5	65.0	70.2	67.3	53.4	62.6	52.9	59.8	63.5
Aya-Expanse-8B	49.6	48.0	55.0	62.8	57.9	50.0	57.4	49.6	52.7	54.0

Table 1: The accuracy of LLMs' responses in ConflictQA-PopQA fashion (Step 1 in Figure 1). We examine eight LLMs, including two closed-source LLMs and four open-source LLMs.

In contrast, the reasoning-heavy **StrategyQA** dataset elicits a much stronger internal resistance, with an average SR of 28.4%—nearly double that of PopQA. This indicates that when knowledge is embedded within a multi-step logical chain, models are significantly more prone to ignore external evidence in favor of their internal reasoning pathways. Interestingly, PR remains high across both datasets (averaging over 80%), demonstrating that models are generally capable of utilizing correct external context to rectify incorrect memories, regardless of the task complexity.

4.3 Linguistic Variances and Cross-lingual Stability

Linguistic factors play a crucial role in conflict resolution dynamics. As Table 2 shows: In PopQA, we observe that languages such as **Chinese (zh)** and **Japanese (ja)** exhibit the highest SR (16.4% and 15.3% respectively), potentially indicating stronger memory alignment or higher confidence in entity representations within these scripts. Conversely, low-resource languages like **Afrikaans (af)** show the lowest SR (10.7%), suggesting a higher reliance on external evidence due to weaker parametric representations.

In StrategyQA, resource-rich languages such as **English (en)** and **German (de)** show higher stubbornness (averaging over 30%), whereas **Turkish (tr)** and **Bengali (bn)** are more easily persuaded. This trend suggests that model "confidence" in logical reasoning is closely tied to the resource

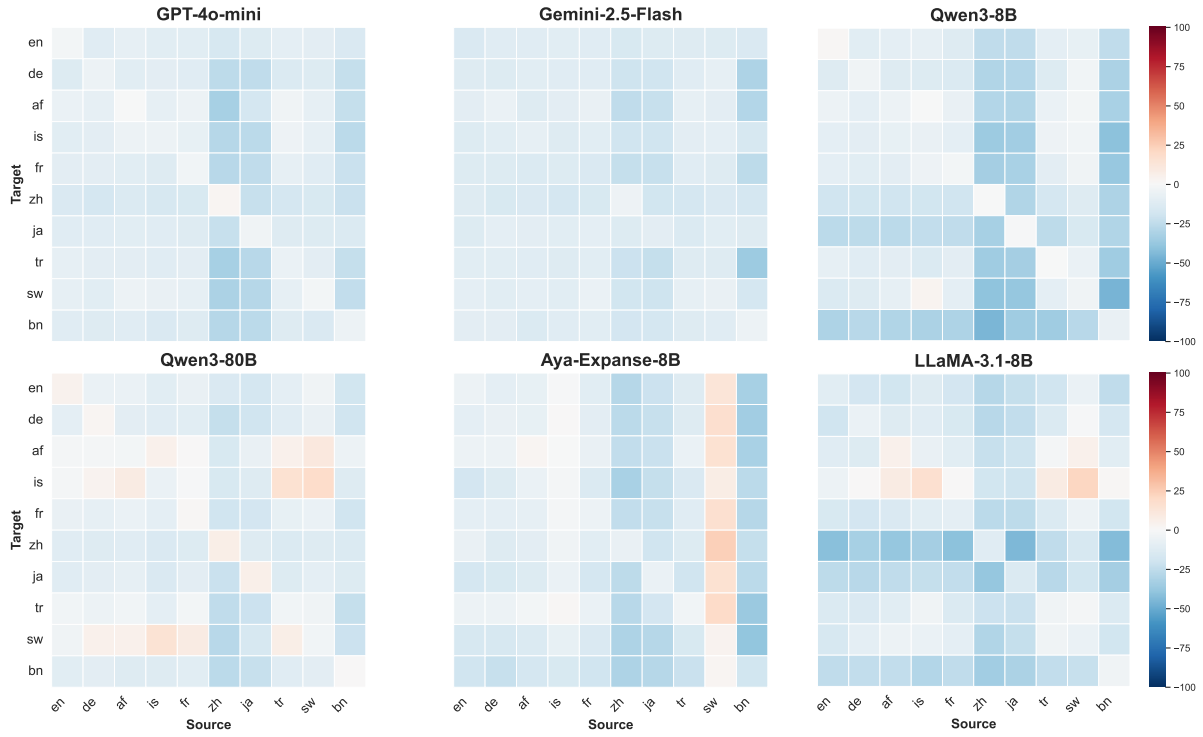


Figure 3: The Persuasion Rate difference between PopQA and StrategyQA, Δ Persuasion is defined as PopQA−StrategyQA

abundance of the language, with high-resource languages fostering more resilient (or stubborn) internal reasoning chains.

4.4 The Dual-Pathway of Cross-lingual Authority

This discovery stems from **Experiment 3**, where we examine the task-dependent authority of various languages. Figure 3 reports the gap in Persuasion Rate (PR) between PopQA and StrategyQA across six models. Our findings reveal a stark contrast: for all models, high-resource non-Latin languages such as **Chinese (zh), Japanese (ja), and Bengali (bn)** exhibit a PR in PopQA that is significantly lower than their PR in StrategyQA (indicated by the deep blue regions). This gap is particularly pronounced in smaller open-source models like **LLaMA-3.1-8B** and **Aya-Expanse-8B**, where these languages possess strong logical authority but fail to rectify factual entities.

Conversely, low-resource languages such as **Icelandic (is)** and **Swahili (sw)** display the opposite behavior. Their PR in PopQA is nearly equal to, or in some instances even higher than, their performance in StrategyQA. For example, in **Aya-Expanse-8B**, these low-resource languages show a neutral or slightly positive delta (warm colors), suggesting they are more effective at resolving entity-centric conflicts than their resource-rich counterparts. This anomalous inversion demonstrates that "knowledge authority" in LLMs follows two distinct pathways:

Path 1: The Logic-Resource Path. In reasoning tasks, the persuasive power of a language is almost strictly linear to its pre-training volume. High-resource languages like **Chinese (zh), Japanese (ja), and Bengali (bn)** emerge as dominant anchors. Their abundance of training data provides a dense and robust logical scaffold. When these languages serve as the source of evidence, they can effectively "rescue" models from reasoning fallacies in any target language. In this domain, the model's strength is a direct artifact of scale—more data yields a more resilient logic engine.

Path 2: The Representation-Affinity Path. However, this resource-driven authority crumbles when the task shifts to entity-centric factual conflicts (PopQA). Despite their logical prowess, **zh, ja, and bn** show a surprising inability to correct factual errors involving Latin-based entities. Here, the **Script Barrier** acts as a profound isolator; the symbolic distance between non-Latin scripts and the original entity names (often stored in a Latin-centric global latent space) hinders effective knowledge retrieval and alignment.

In a remarkable reversal, low-resource languages

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	9.2	13.3	9.7	10.2	9.4	10.9	11.3	11.0	11.7	13.3	85.6	81.2	78.9	79.5	79.2	81.3	80.5	82.2	80.0	84.5
Gemini-2.5-Flash	10.1	10.6	8.7	10.2	8.7	10.3	10.8	10.7	10.2	12.8	79.1	83.8	79.0	76.1	77.0	80.2	81.1	82.9	78.7	84.0
LLaMA-3.1-8B	15.4	13.0	16.7	13.3	14.3	26.1	19.6	12.2	17.4	21.5	81.6	78.6	80.2	79.1	77.9	77.8	78.4	75.2	79.7	77.6
Qwen3-8B	9.6	15.8	16.5	17.1	15.9	13.5	20.6	10.5	12.3	18.6	84.2	78.8	83.4	87.6	83.3	78.6	84.2	75.8	83.6	83.9
Qwen3-80B	11.7	13.1	16.1	17.8	14.8	12.1	14.2	9.6	15.0	15.4	82.0	80.6	78.0	78.3	78.6	80.3	84.2	79.7	80.3	82.3
Aya-Expanse-8B	8.2	8.8	12.6	13.3	13.8	13.0	15.0	12.6	12.8	16.7	86.8	67.9	86.2	87.3	87.0	77.3	86.1	70.3	86.4	85.5
Average	10.7	12.4	13.4	13.7	12.8	14.3	15.3	11.1	13.2	16.4	83.2	78.5	81.0	81.3	80.5	79.3	82.4	77.7	81.5	83.0
StrategyQA																				
GPT-4o-mini	34.7	32.8	37.0	39.3	37.6	34.8	32.8	34.4	32.2	34.3	86.2	86.8	84.6	82.2	82.8	87.4	84.5	84.7	86.8	82.3
Gemini-2.5-Flash	17.8	19.3	20.8	21.0	19.5	19.1	19.3	17.6	17.6	21.0	92.1	89.8	92.6	91.7	92.5	93.4	91.7	90.7	91.3	89.4
LLaMA-3.1-8B	32.7	19.5	24.6	26.1	24.4	47.2	13.5	24.6	23.4	17.3	76.5	83.1	87.1	90.1	88.0	61.5	93.0	81.7	84.3	89.6
Qwen3-8B	25.2	23.7	32.9	34.4	34.3	22.0	30.2	26.9	24.3	34.8	89.7	86.3	87.7	86.1	85.9	86.4	85.2	80.2	84.2	84.0
Qwen3-80B	50.6	44.4	53.2	51.4	51.0	37.8	48.2	33.7	40.5	49.6	84.9	80.5	75.7	73.8	77.1	86.9	78.6	82.8	83.6	75.6
Aya-Expanse-8B	17.1	15.4	13.5	11.8	12.8	22.0	11.9	31.9	12.7	12.8	84.9	86.7	93.3	93.6	92.6	79.2	92.5	66.5	90.3	93.3
Average	29.7	25.9	30.3	30.7	29.9	30.5	26.0	28.2	25.1	28.3	85.7	85.5	86.8	86.3	86.5	82.5	87.6	81.1	86.8	85.7

Table 2: Stubborn Rate (SR) and Persuasion Rate (PR) across 10 languages for PopQA and StrategyQA. Higher SR indicates stronger adherence to internal memory, while higher PR indicates better error correction via external evidence.

that share the Latin script—such as **Swahili (sw), Icelandic (is), and Afrikaans (af)**—become the most effective rescuers in PopQA. Their strength lies not in the "volume" of what they know, but in the **"proximity" of how they represent it**. Because they share a common alphabet and morphological roots with the target entities, they provide a more direct and reliable "address" in the model's memory, bypassing the alignment rot that plagues even the most data-rich non-Latin languages.

4.5 Unveiling the "Abundance-Weakness" Paradox

These findings lead us to a fundamental conclusion: **Data abundance in one cognitive dimension often masks a structural fragility in another.** The apparent multi-lingual competence of high-resource non-Latin models is partially an illusion sustained by their superior reasoning capabilities. While they excel at "thinking" through a problem, they are structurally hampered at the level of "seeing" and "linking" entities across linguistic boundaries.

This paradox suggests that the cross-lingual latent space of modern LLMs is deeply fragmented. A model can be a "master logician" in Chinese but a "clumsy librarian" when it comes to reconciling those same logical truths with factual evidence presented in a different script. This misalignment is precisely where cross-lingual knowledge conflicts take root and flourish.

4.6 Validating Knowledge Pathways via Multi-source Cross-lingual Conflict

Figure 4 reports accuracy under two-source conflicts, where the *truth* source and *fake* (interfering) source are written in different languages. Comparing $Query$ = Truth Lang against $Query$ = Fake Lang allows us to isolate how query-language priming modulates cross-lingual source competition.

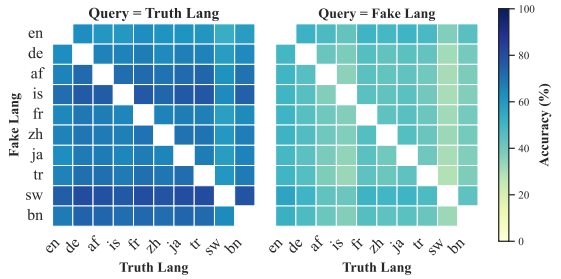
Across models, we observe a consistent **task-dependent split**. On **StrategyQA**, conflict resolution follows a **resource-driven hierarchy**: evidence from high-resource languages is both harder to override (higher robustness against interference) and more effective at rescuing errors under competition. On **PopQA**, this hierarchy weakens substantially. High-resource languages remain locally robust, but their cross-lingual rescuing power degrades when the query is in a different script or representation space, suggesting a **representation barrier** in entity-centric conflicts. Detailed figures and analyses are provided in Appendix C.1.

4.7 Interplay Between Parametric Memory and Cross-lingual Conflicts

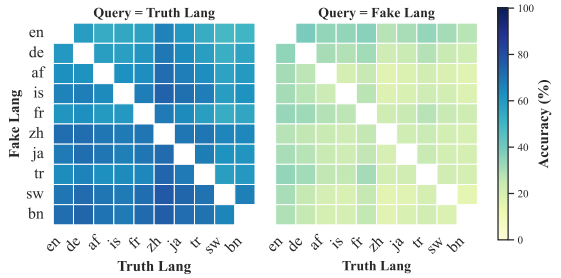
To understand how internal knowledge shapes multi-source arbitration, we analyze results by the four parametric-memory quadrants (TT/TF/FT/FF; Table 3) and relate them to aggregate trends in Figure 3.

Dataset	Model	Q = Pos					Q = 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 3: Accuracy (%) on StrategyQA and PopQA with two ordered evidence snippets (Evidence 1, Evidence 2). Q = Pos/Neg indicates whether the query language matches Evidence 1/Evidence 2, respectively. TT/TF/FT/FF denote the truthfulness of (Evidence 1, Evidence 2); e.g., TF means (true, false) and FT means (false, true). TF–FT compares the two mixed-evidence conditions under the same query-language setting. All values are multiplied by 100 and rounded to one decimal place.



(a) StrategyQA



(b) PopQA

Figure 4: Aggregate accuracy across six evaluated models in Task 4 (Multi-source Conflict Resolution), partitioned by query language and evaluation benchmark.

First, we find strong **alignment-driven robustness**. Accuracy is highest in **TT** (internally correct and aligned across languages), indicating that consistent parametric memory provides a stable anchor under conflict. Accuracy is lowest in **FF**, where the model lacks a reliable internal reference and becomes most vulnerable to misleading evidence.

Second, we identify a **query-memory congruence effect**. When the query language activates a *correct* internal belief (e.g., TF under Q = Pos), the model is more likely to select the correct source under competition. When the query activates an *incorrect* belief (e.g., FT under Q = Pos), that belief can bias evidence selection and suppress correct sources written in other languages. We report full quadrant-wise breakdowns and language-pair analyses in Appendix C.2.

5 Conclusions

In this work, we study **Cross-Lingual Knowledge Conflict (CLKC)**, where an LLM's language-conditioned parametric beliefs contradict multilingual external evidence, a setting largely overlooked by English-centric conflict evaluations. We introduce **CLEAR**, the first evaluation framework that systematically probes and stresses CLKC through four progressive scenarios, from multilingual parametric elicitation to competitive multi-source cross-lingual conflict resolution, enabling a structured analysis across QA tasks with different conflict patterns. Extensive experiments on 10 diverse languages and six representative LLMs reveal a **task-dependent decision dichotomy**: reasoning conflicts are dominated by language resource abundance, whereas entity-centric factual conflicts are governed more by linguistic affinity and representational compatibility. Our findings show that multilingual robustness cannot be inferred from English-only benchmarks or resource scale alone. CLEAR provides a principled testbed for diagnosing cross-lingual misalignment and guiding the development of more faithful, evidence-grounded multilingual LLMs and RAG systems.

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 task 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.

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*.

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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

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.

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*.

A Prompts

A.1 Prompts for Task 1: Parametric Memory Elicitation

Task 1 evaluates the model's internal parametric memory without external context. The prompts are tailored to the output format required by each dataset.

StrategyQA Prompt (Task 1)

```
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 Prompt (Task 1)

```
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 Prompts for Tasks 2–4

For Tasks 2, 3, and 4, the prompts instruct the model to answer based on the provided external context. The prompt templates are consistent across these tasks, while the input components vary by task definition.

StrategyQA

```
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

```
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}
```

Input Configurations for Tasks 2–4. The {Context} and {Question} slots in the above templates are populated as follows:

- **Task 2 (Intra-lingual):** Context(L_n) + Question(L_n)
- **Task 3 (Cross-lingual):** Context(L_n) + Question(L_m)
- **Task 4 (Multi-source):** Context(L_n) + Context(L_m) + Question(L_{norm})

Recommended formatting for Task 4 multi-source context. To reduce the risk that models ignore one of the sources, we concatenate the two contexts with explicit source headers before inserting into {Context}:

Task 4 Context Formatting (Multi-source)

```
Context A (Language L_n): {Context_Ln}
Context B (Language L_m): {Context_Lm}
```

A.3 Prompt for Dataset Translation

The following prompt is used to translate the experimental dataset into target languages while preserving deliberate misinformation and script-specific constraints.

Dataset Translation Prompt

```
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
```

A.4 Prompt for 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

```
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. For all experiments, we disable any explicit chain-of-thought or reasoning mode provided by the models. Closed-source models are accessed via the OpenRouter API. For open-weight models with relatively small parameter sizes, inference is performed on an NVIDIA A100 (40GB) GPU using FP16 precision. For larger open-weight models, we rely on the Alibaba Cloud API. Unless otherwise specified, the decoding temperature is set to 0.01 to ensure stable and 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.

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

Table 4: 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.

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.

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.

Table 4 reports the gap between AI Judge and exact match accura- cies for the parametric memory task. We observe that AI Judge scores are generally higher than exact match scores across models and languages. The difference is more noticeable in languages with richer morphology or less standardized orthography (e.g., BN, SW, IS, JA).

This discrepancy is primarily due to the strict nature 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-level variations and can therefore provide a smoother estimate of model performance under these conditions..

C Additional Experimental Results

This appendix reports complete results for Task 3 and Task 4 and provides complementary analyses. Figure 5 reports full Persuasion Rate (PR) and Stubborn Rate (SR) results for Task 3 on PopQA and StrategyQA for four representative models drawn from the six evaluated systems. Figures 6 and 7 report Task 4 accuracy heatmaps under two-source cross-lingual competition for three representative models. Appendix C.1 interprets the Task 4 heatmaps, and Appendix C.2 analyzes outcomes by parametric-memory quadrants.

Across the four representative models in Figure 5, both PR and SR are consistently higher on StrategyQA than on PopQA. This pattern suggests stronger state dependence in reasoning-intensive conflicts: models are more resistant to misleading evidence when their initial belief is correct (higher SR), yet remain capable of updating when their initial belief is wrong (higher PR).

C.1 Validating Knowledge Pathways via Multi-source Cross-lingual Knowledge Conflict

Figures 7 and 6 visualizes Task 4 accuracy when the *truth* source (Truth Lang) and the *interfering* source (Fake Lang) are written in different languages. We compare two conditions: $Query =$ Truth Lang (left), where the query aligns with the truthful source, and $Query =$ Fake Lang (right), where the query aligns with the interference. This contrast helps separate query-language priming from source-language authority, and provides additional evidence for the dual-pathway behavior identified in Task 3.

StrategyQA: Resource-driven authority in reasoning conflicts. On StrategyQA, performance follows a clear resource-driven hierarchy. High-resource languages such as English (*en*), Chinese (*zh*), and Japanese (*ja*) tend to be (i) harder to override when they serve as the interfering language, and (ii) more effective at rescuing the model when they serve as the truth language. Concretely,

- **Interference resistance.** When high-resource languages appear on the vertical axis (Fake Lang), the model more often withstands interference, yielding higher accuracy.
- **Cross-lingual rescuing.** When high-resource languages appear on the horizontal axis (Truth Lang), accuracy remains high even under competition, indicating stronger corrective influ-

ence.

- **Boundary crossing.** Importantly, this advantage persists even in the harder condition $Query$ = Fake Lang, suggesting that in reasoning tasks, high-resource evidence can override query priming and transfer across languages.

PopQA: Representation barriers in entity-centric conflicts. PopQA exhibits a qualitatively different pattern. While high-resource languages can be robust within their own setting, their corrective influence is less reliable when the query is written in a different script or representation space. In particular,

- **Local robustness.** Under $Query$ = Truth Lang, languages such as *zh* and *ja* can still show strong robustness, consistent with stable in-language representations.
- **Degraded rescuing under mismatch.** Under $Query$ = Fake Lang, the same languages often lose corrective power as Truth Lang, indicating that truthful evidence does not consistently transfer to counteract interference when entity forms are less directly alignable.
- **Isolation under competition.** Overall, PopQA suggests that entity-centric conflict resolution is more sensitive to cross-lingual representational compatibility than to resource scale alone.

Takeaway. Together, these heatmaps support a functional fragmentation of cross-lingual behavior: reasoning conflicts (StrategyQA) exhibit resource-dominated authority that generalizes across languages, whereas entity-centric conflicts (PopQA) are constrained by representational barriers that limit cross-lingual correction.

C.2 Interplay between Parametric Memory and Cross-lingual Conflicts

To further isolate the role of parametric memory, we analyze Task 4 performance under four parametric-memory quadrants (TT, TF, FT, FF), reported in Table 3, 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 = 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 = Neg.** The pattern flips: FT outperforms 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.

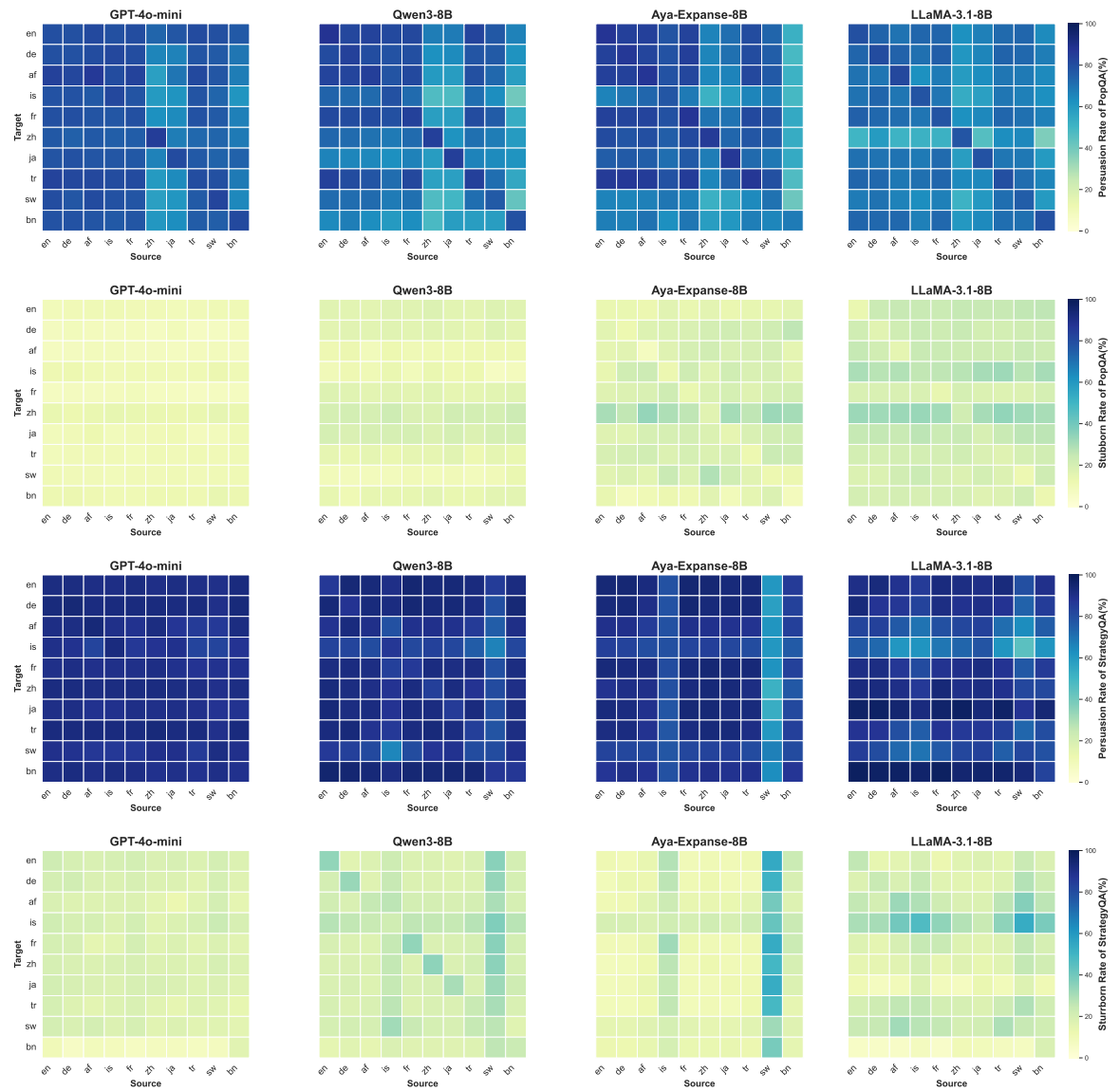
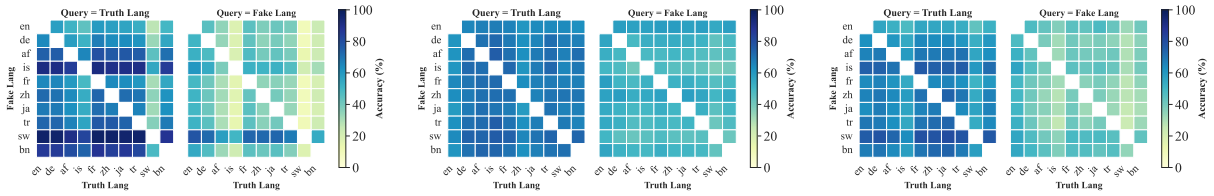
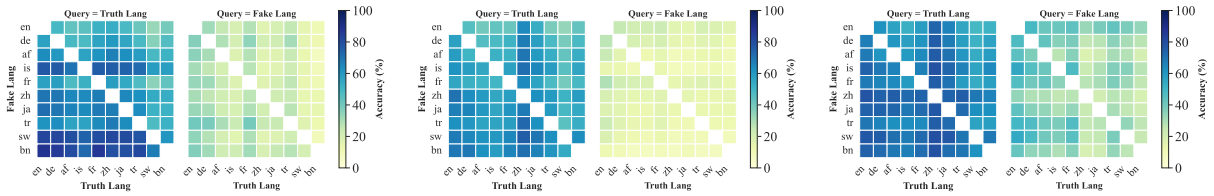


Figure 5: Task 3 results: Persuasion Rate and Stubborn Rate on PopQA and StrategyQA across four representative models.



(a) Aya-Expanse-8B (b) Qwen3-80B (c) LLaMA-3.1-8B

Figure 6: Task 4 results: Language-dependent ACC on StrategyQA across three representative models.



(a) Aya-Expanse-8B (b) Qwen3-80B (c) LLaMA-3.1-8B

Figure 7: Task 4 results: Language-dependent ACC on PopQA across three representative models.