

Reasoning over Non-Parametric Memory in Multilingual LLMs: A Cognitive-Based Analysis of Relational Knowledge

Zohreh Abbasi Liasi

Qatar Computing Research Institute
zohreabbasiliasi@gmail.com

Batoul Hammoud

Qatar Computing Research Institute
batoul.hammoud@ru.nl

Anas Madkoor

Qatar Computing Research Institute
amadkoor@hbku.edu.qa

Ehsaneddin Asgari

Qatar Computing Research Institute
easgari@hbku.edu.qa

Abstract

We introduce WIKI-BLOOM, a multilingual diagnostic benchmark for evaluating how large language models reason over supplied relational knowledge, with the goal of separating evidence-conditioned reasoning from memorized entity associations. WIKI-BLOOM contains 450 multiple-choice questions grounded in 30 curated Wikidata relations, instantiated under factual, counterfactual, and neutralized conditions and spanning five levels adapted from Bloom’s revised taxonomy: Remember, Understand, Apply, Analyze, and Evaluate. We evaluate 43 models, 41 open-weight models across eight series and two closed-weight Gemini models, in six languages (Arabic, Chinese, English, Hindi, Persian, and Spanish) and two multilingual settings, yielding 226,800 predictions. Matched accuracy contrasts use relation-clustered inference over the 30 source relations, with dedicated correlation-based tests for cross-language stability.

Several concrete patterns emerge. Open-weight families consistently exhibit an Apply-to-Analyze accuracy drop, but its magnitude varies substantially with family and scale and disappears entirely for both Gemini models. Evaluate, however, remains a persistent bottleneck even when the earlier cliff has closed. Gemini 3.1 Pro, for example, averages 99.5% across Remember, Understand, Apply, and Analyze, but only 60.7% on Evaluate. Item-difficulty rankings are significantly stable across languages in every evaluated family and size bucket, and relation-difficulty hierarchies show similarly strong cross-language stability.

The interaction between parametric knowledge and supplied context also changes with cognitive demand. The Neutralized-over-Factual advantage is often larger at Analyze and Evaluate than at Remember and Understand, with gains reaching more than 20 percentage points in some settings, suggesting that entity-specific associations become increasingly disruptive as reasoning demands increase. Counterfactual contexts are generally the weakest condition, and the factual-counterfactual gap widens with cognitive complexity in several model families, including both Gemini models, showing that context conflict can remain damaging even near ceiling performance.

Together, these results show that increasing model capability alone does not eliminate failures in cognitively demanding use of supplied knowledge, and that the structure of these failures is remarkably stable across languages.

1 Introduction

Large language models are increasingly deployed in knowledge-intensive settings, most notably retrieval-augmented generation (RAG), where a model must reason over facts supplied in its context rather than rely solely on information acquired during pretraining (Lewis et al., 2020). A fundamental challenge in evaluating such systems is disentangling genuine evidence-conditioned reasoning from the retrieval of memorized associations: a model may produce a correct answer by activating parametric knowledge about a named entity, even when the provided context specifies a different or contradictory relation (Longpre et al., 2021). Existing benchmarks rarely operationalize this distinction directly, making it difficult to determine whether strong performance reflects faithful reasoning over supplied knowledge or sophisticated pattern matching against associations learned during pretraining.

Bloom’s revised taxonomy (Krathwohl, 2002) offers a principled framework for characterizing the cognitive complexity of tasks, organizing cognitive processes into six hierarchical levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. Originally designed to classify human learning objectives, the taxonomy has been adopted in NLP and LLM research to evaluate models across varying depths of reasoning (Fei et al., 2024; Cao et al., 2024). A critical assumption underlying such work remains insufficiently tested, however: whether Bloom’s cognitive hierarchy, derived from observations of human learning, accurately reflects the computational difficulty landscape of language models. In particular, stronger models may not exhibit the same monotonic progression of difficulty implied by the human-oriented taxonomy, and different cognitive transitions may expose qualitatively different model weaknesses.

Multilingual evaluation adds a further dimension to this challenge. Benchmarks such as XTREME (Hu et al., 2020) and MEGA (Ahuja et al., 2023) have established cross-lingual performance baselines, yet multilingual evaluations have largely focused on whether models can perform tasks across languages. Less attention has been paid to whether the underlying structure of model behavior itself is stable across languages: whether the same items and relations remain difficult, whether models exhibit similar sensitivity to knowledge format, and whether cognitive difficulty profiles and interference from memorized associations reflect shared model behavior or language-specific artifacts.

We address these gaps simultaneously with WIKI-BLOOM, a multilingual diagnostic benchmark for evaluating how language models reason over supplied relational knowledge. WIKI-BLOOM contains 450 multiple-choice questions grounded in 30 curated Wikidata relations (Wang et al., 2021). Each question is instantiated under three veracity conditions: Factual, in which real entities are paired with the correct relation; Counterfactual, in which the original entities are replaced with incorrect real entities while preserving the relation; and Neutralized, in which entity identities are replaced with typed abstract placeholders. This design allows us to vary the degree to which parametric entity associations can assist or interfere with reasoning over the supplied context. Questions span five levels adapted from Bloom’s revised taxonomy: Remember, Understand, Apply, Analyze, and Evaluate, with Create omitted because it is not naturally compatible with controlled multiple-choice evaluation.

We evaluate 43 models: 41 open-weight models drawn from eight model series, Qwen2.5 Base, Qwen2.5 Instruct, Qwen3, Qwen3.5, Gemma, LLaMA, Granite, and DeepSeek, together with two closed-weight Gemini models. Evaluation covers six languages, Arabic, Chinese, English, Hindi, Persian, and Spanish, and two multilingual settings, yielding 226,800 predictions. Our analysis considers model scale, Bloom cognitive level, knowledge veracity, relation usage frequency, and language. Matched accuracy contrasts are evaluated using relation-clustered inference over the 30 source relations, while cross-language stability is assessed using dedicated correlation-based procedures.

Our study yields four main findings. **First**, improvements with model size are widespread but do not follow a universal trajectory. Most open-weight families become more accurate as model size increases, particularly when moving beyond the smallest models. However, the relative gain obtained at successive size regimes differs substantially across families,

and only a subset exhibits clear diminishing returns. This variation shows that performance gains associated with increasing model size do not follow a single trajectory across model families.

Second, the empirical difficulty structure of Bloom levels differs markedly from a simple human-inspired hierarchy. A recurring transition appears between Apply and Analyze for open-weight models, but its severity changes considerably across families and model sizes and is absent in the Gemini models. More importantly, strong performance at Analyze does not imply that the subsequent level has been mastered: Evaluate remains difficult across the evaluated families, including highly capable models whose performance on the preceding levels is close to ceiling. The two transitions therefore appear to diagnose different limitations rather than successive points on a single smooth difficulty curve.

Third, whether parametric knowledge helps or interferes with the supplied context depends on the type of cognitive operation required. Removing entity identities has a comparatively larger benefit at the higher Bloom levels in many model families, consistent with entity-specific associations becoming more disruptive when the task requires deeper manipulation of the presented relation. Counterfactual examples provide a complementary view of this interaction. They are generally harder than factual examples, and several families show stronger factual-counterfactual separation at higher cognitive levels. These effects indicate that conflicts between contextual evidence and prior associations are shaped jointly by task complexity and model family rather than by knowledge veracity alone.

Fourth, the structure of benchmark difficulty transfers strongly across languages. Items that are relatively difficult in one language tend to retain similar difficulty rankings in the others, and the same pattern holds when difficulty is aggregated at the relation level. This cross-lingual agreement goes beyond similarity in average accuracy: it indicates that the benchmark exposes recurring sources of difficulty across linguistic realizations. Consequently, the diagnostic patterns observed in WIKI-BLOOM are unlikely to be explained solely by language-specific proficiency differences.

These findings carry practical implications for knowledge-intensive and retrieval-augmented systems. They suggest that stronger models can still exhibit sharply localized weaknesses when retrieved information must be manipulated at higher cognitive levels, even when their aggregate performance is high. They also highlight the importance of how retrieved knowledge interacts with parametric associations: representations that reduce irrelevant entity-specific interference or make the required relation easier to identify may help models follow supplied evidence more faithfully. More broadly, WIKI-BLOOM provides a controlled setting for examining how capability, cognitive demand, knowledge condition, and language jointly shape reasoning over information introduced at inference time.

2 Related Work

Our work sits at the intersection of three active research threads: cognitive-complexity benchmarking of LLMs, context faithfulness under knowledge conflicts, and the evaluation of reasoning in retrieval-augmented and knowledge-graph settings.

2.1 Bloom’s Taxonomy as a Cognitive Complexity Framework for LLMs

Bloom’s revised taxonomy (Krathwohl, 2002) has become a widely adopted scaffold for stratifying the cognitive demands placed on language models, moving evaluation beyond aggregate accuracy toward a diagnosis of where competence breaks down. Early adoption appeared in domain-specific benchmarks such as LawBench (Fei et al., 2024) and scientific evaluation suites such as SciEval (Sun et al., 2024) and SciAssess (Cai et al., 2025). Subsequent work extended this to law (Dong et al., 2025), medicine (Zhou et al., 2025), software engineering (Ma et al., 2025), and biographical reasoning (Cekinmez et al., 2025), consistently revealing that performance degrades at higher cognitive levels. StructEval (Cao et al., 2024) generates test instances automatically for all six levels, while Khoramfar et al. (2026) report accuracy drops of up to 70% as tasks ascend the hierarchy, echo-

ing findings from Todorovikj et al. (2025), who apply Bloom’s framework to characterize tasks in knowledge-graph benchmarks. BloomWise (Zoumpoulidi et al., 2025) reveals a non-monotonic pattern in mathematical reasoning, models score higher on Apply than on Understand; we do not reproduce this ordering in the relational setting. Raimondi & Gabbrielli (2026) further argue, via mechanistic probing, that Bloom’s levels correspond to distinguishable internal representations.

Despite this growing body of work, relatively little attention has been paid to controlled multilingual comparisons of cognitive difficulty over matched items. Prior studies also typically examine cognitive complexity independently of how supplied knowledge interacts with parametric associations. WIKI-BLOOM brings these dimensions together by evaluating matched relational knowledge across Bloom level, factual, counterfactual, and entity-neutralized conditions, and six languages.

2.2 Context Faithfulness Under Knowledge Conflicts

A parallel line of work studies whether models ground responses in supplied context or default to memorized knowledge. Longpre et al. (2021) formalize this as an entity-level knowledge conflict, measuring models’ tendency to favor parametric answers over counterfactual context. Li et al. (2023) demonstrate that large models exhibit poor controllability in conflict scenarios and introduce Knowledge-Aware Fine-Tuning (KAFT), incorporating counterfactual and irrelevant contexts to improve robustness. Basmov et al. (2024) show that faithfulness further degrades for hypothetically or modally framed contexts, while Meng et al. (2022) provide mechanistic evidence linking this behavior to localized factual associations stored in mid-layer feed-forward modules. Liu et al. (2025) introduce a generative framework for relational reasoning that recreates classic cognitive-science paradigms, such as transitive inference and the symbolic distance effect, and report that several human reasoning biases reappear in LLMs. Their manipulation targets relational structure and difficulty; ours holds the relation fixed and varies whether the supplied fact is real, contradictory, or entity-neutralized, isolating the role of parametric interference. Our work extends this literature by introducing a Neutralized condition that removes direct entity-identity cues without introducing entity-specific conflict, enabling a cleaner comparison between relational and associative processing, and by characterizing how the influence of entity-specific associations varies with cognitive complexity across six languages.

2.3 RAG Reasoning and Knowledge-Graph QA

The original RAG framework (Lewis et al., 2020) demonstrated that combining parametric memory with non-parametric retrieval improves factuality on knowledge-intensive tasks. Pan et al. (2024) survey the landscape of LLM-KG integration, and GraphRAG variants (Thompson et al., 2025) leverage entity-relation graphs for multi-hop traversal and associative retrieval. Sarthi et al. (2024) organize documents into hierarchical summaries for improved retrieval. Zhang et al. (2025) report that the average factual correctness rate across 16 popular KGQA datasets is only 57%, and Lin et al. (2026) propose HybridRAG-Bench to control for parametric memorization. Todorovikj et al. (2025) specifically examine how Bloom’s taxonomy can characterize KG tasks, identifying gaps in coverage of higher-order reasoning. Our benchmark derives questions from Wikidata relations (Wang et al., 2021) and instantiates them under factual, counterfactual, and neutralized conditions, asking not merely whether a model uses retrieved context, but how the influence of entity-specific associations changes with cognitive demand.

3 Methods

3.1 Benchmark Overview

We introduce WIKI-BLOOM, a compact diagnostic benchmark for sentence-level understanding of Wikidata-style relations. Each item consists of a fact sentence, a multiple-choice question with four answer options, and metadata describing the Bloom level, veracity set-

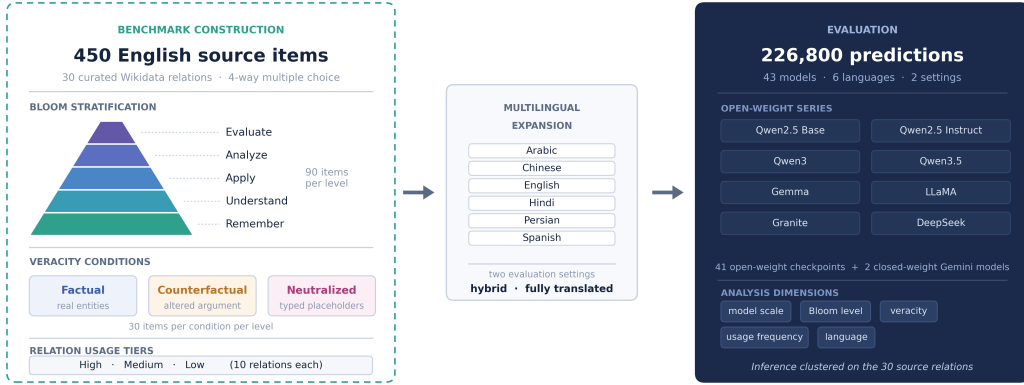


Figure 1: Overview of WIKI-BLOOM. A pool of 450 English items is built from 30 curated Wikidata relations and stratified by Bloom level, veracity condition, and relation usage tier. The pool is then expanded across six languages under two multilingual settings and evaluated on 43 models, yielding 226,800 predictions. Matched accuracy contrasts use relation-clustered inference over the 30 source relations, with dedicated correlation-based tests for cross-language stability.

ting, and relation usage tier. WIKI-BLOOM spans five Bloom-inspired cognitive levels, Remember, Understand, Apply, Analyze, and Evaluate, with 90 questions per level, totalling 450 questions. We omit the sixth level (Create) because it requires open-ended generation that is incompatible with a controlled multiple-choice format. Within each Bloom level, questions are distributed equally across three veracity settings (Factual, Counterfactual, and Neutralized, 30 questions each), and each set of 30 is further balanced across three relation usage tiers (High, Medium, and Low, 10 questions each).

Figure 1 summarizes the benchmark’s structure and evaluation pipeline.

3.2 Relation Selection and Usage Tiers

We draw candidate relations from Wikidata5M (Wang et al., 2021) and stratify them into Low, Medium, and High usage tiers using MediaWiki property-frequency counts as a proxy for relation prevalence. We manually curate 10 relations per tier, yielding 30 source relations. The complete sampling, frequency transformation, and stratification procedure is provided in Appendix A.3.

3.3 Triple Extraction and Manual Curation

For each candidate relation, we extract one representative factual triple (h, p, t) and manually curate a final set of 30 relations, with 10 from each usage tier. Curation prioritizes semantic clarity, domain diversity, and non-overlapping entities, while excluding sensitive or excessively technical relations. Each selected relation contributes one source triple that is reused across Bloom levels and veracity conditions, prioritizing controlled comparison over broad relation coverage. Further extraction and curation details are provided in Appendix A.3.

3.4 Fact Verbalization and Question Generation

For each extracted triple (h, p, t) , GPT-5¹ verbalizes the fact and assists in generating level-specific multiple-choice questions. Table 1 summarizes the operationalization of the five Bloom levels.

¹We accessed GPT-5 in early 2026. All verbalized facts were subsequently verified manually by the authors.

Table 1: Bloom-level operationalization in WIKI-BLOOM.

Bloom level	Operationalization in WIKI-BLOOM
Remember	Extract the value explicitly stated in the fact.
Understand	Select a paraphrase that preserves the relation semantics while competing options express different relations or attributes.
Apply	Apply an abstract relation schema to identify a new statement that follows the same subject–relation–object structure and meaning.
Analyze	Discriminate the target relation from near-miss alternatives, where distractors are semantically related or structurally similar but logically incorrect for the relation expressed in the fact.
Evaluate	Select the strongest justification grounded in the defining semantic criterion of the relation, while distractors provide plausible but non-definitional explanations.

Table 2: Veracity variants in WIKI-BLOOM, generated with GPT-5 and manually verified.

Variant	Construction rule	What it tests
Factual	Verbalize the extracted triple (h, p, t) as a sentence-level fact.	Baseline relation understanding with real entity surface forms.
Counterfactual	Edit exactly one argument of (h, p, t) , preserving coarse entity type and relation context.	Evidence-following under conflict with parametric priors.
Neutralized	Replace entity mentions with typed placeholders (e.g., [PERSON 1], [SPORT 1]) while preserving the question template.	Relation abstraction with reduced entity-prior shortcuts.

All generated questions, answer options, and answer keys are manually reviewed for factual correctness, preservation of the intended relation, answer uniqueness, and distractor plausibility. As an additional validation pass, Gemini and Grok independently flag disagreements with the provided answers; all flagged cases are manually inspected, and ambiguous items are revised and re-evaluated. Additional quality-control details are provided in Appendix A.4.

3.5 Veracity Variants

To disentangle memorization, evidence-following, and relational abstraction, we construct three matched veracity variants for each Bloom level, summarized in Table 2. The variants preserve the underlying relation and Bloom-level question structure, enabling controlled comparisons across factual, counterfactual, and neutralized conditions.

3.6 Multilingual Evaluation Settings

WIKI-BLOOM is constructed in English and evaluated in Arabic, Chinese, English, Hindi, Persian, and Spanish under two settings. In the *hybrid cross-lingual setting*, the fact and question remain in English while the answer options are translated into the target language, isolating cross-lingual answer selection from a shared English stimulus. In the *fully translated setting*, the fact, question, and options are all presented in the target language

while preserving the relation, Bloom-level operation, veracity condition, answer key, and option order. We analyze the two settings separately.

3.7 Hypothesis Testing Framework

We evaluate ten hypotheses spanning model scale, Bloom-level difficulty, knowledge veracity, relation usage frequency, and cross-language stability. For scale-dependent analyses, open-weight models are grouped into Small ($< 7\text{B}$), Medium ($\geq 7\text{B}$ and $< 14\text{B}$), and Large ($\geq 14\text{B}$) buckets; Gemini models are reported individually because their parameter counts are undisclosed.

H1 tests whether accuracy gains diminish with scale, comparing the Small-to-Medium improvement with the Medium-to-Large improvement. H2 tests the broader monotonic scaling prediction that Medium outperforms Small and Large outperforms Medium. H3 tests whether Apply accuracy exceeds Analyze accuracy; we additionally mark a pronounced “cliff” when the statistically supported drop has a lower 95% confidence bound above 10 percentage points.

H4 tests the veracity ordering Neutralized $>$ Factual $>$ Counterfactual. H5 tests whether the Factual–Counterfactual gap is larger at the higher Bloom levels (Analyze and Evaluate) than at the lower levels (Remember and Understand). H6 tests whether relation usage frequency has a practically negligible effect, using a five-percentage-point equivalence margin over Low-, Medium-, and High-usage relations. H7 tests whether the Neutralized-over-Factual advantage is larger at Analyze and Evaluate than at Remember and Understand. H8 tests whether Evaluate underperforms the mean of the other four Bloom levels.

Finally, H9 and H10 assess cross-language stability. H9 measures the degree to which item-level difficulty rankings are preserved across the six languages, summarized by the mean of the 15 pairwise Spearman correlations, while H10 applies the same idea to relation-level difficulty rankings.

Because items derived from the same Wikidata relation share a common source triple, matched accuracy contrasts are evaluated using relation-clustered inference over the 30 source relations rather than treating the 450 questions as independent. Directional accuracy hypotheses use one-sided relation-cluster sign-flip tests with 95% relation-cluster bootstrap confidence intervals. H6 uses an equivalence criterion, while H9 and H10 use dedicated correlation-based permutation procedures for cross-language stability. Full mathematical definitions and statistical procedures are provided in Appendix A.2.

4 Results

Tables 3 and 4 summarize hypothesis support across the fully translated and hybrid settings. Gemini models are reported individually in the fully translated setting. The most consistent findings are cross-language item and relation stability (H9–H10), which hold in every series and bucket; the recurrent Apply-to-Analyze drop in open-weight models (H3); the Evaluate bottleneck (H8), including both Gemini models; and the Small-to-Medium scaling improvement (H2). H1 and H4–H7 show greater dependence on model family and size regime.

4.1 Scaling Behavior

Accuracy generally increases with model size (H2). In the fully translated setting, Medium significantly exceeds Small in all eight open-weight series, while Large significantly exceeds Medium in seven; DeepSeek is the only exception. The same broad pattern holds in the hybrid setting. For example, Qwen2.5 Instruct gains 23.8 percentage points from Small to Medium and 6.1 points from Medium to Large.

Diminishing returns (H1) are less universal, holding in five of eight series when fully translated and six of eight in the hybrid setting. LLaMA and Granite instead show larger gains at the upper transition, while DeepSeek supports H1 only in the hybrid setting and slightly

declines from Medium to Large when fully translated. Qwen2.5 Base and Instruct otherwise exhibit broadly similar scaling trends, with instruction tuning providing a modest accuracy advantage.

4.2 Bloom-Level Patterns

Apply exceeds Analyze across all eight open-weight series in both settings (H3), but the magnitude of the drop varies sharply with family and scale. A “cliff” denotes a statistically supported drop whose lower 95% relation-cluster bootstrap bound exceeds 10 percentage points. At Large, Granite exhibits a 47.2-point drop, whereas Qwen3.5 drops only 7.2 points. The transition disappears entirely for Gemini: neither Gemini 3 Flash nor Gemini 3.1 Pro shows the expected Apply-to-Analyze decrease.

The cliff is therefore not scale-invariant. Its magnitude can shrink substantially with capability, although this pattern is not monotonic across families.

Evaluate presents a different pattern (H8). It remains below the mean of the other four Bloom levels across all eight open-weight series and both Gemini models. The effect persists even when the Apply-to-Analyze transition has disappeared: Gemini 3.1 Pro averages approximately 99.5% across the other four levels but 60.7% on Evaluate. Analyze and Evaluate therefore expose distinct limitations rather than successive manifestations of the same difficulty.

4.3 Veracity Effects

The full Neutralized > Factual > Counterfactual ordering (H4) is common directionally but rarely supported in full. Counterfactual is generally the weakest condition, whereas Neutralized does not consistently exceed Factual and reverses in several high-capability models, including Qwen3.5, Gemma, and both Gemini models. Removing entity identity is therefore not uniformly beneficial.

The Factual-Counterfactual gap (H5) also depends on cognitive demand. Its expected widening from lower to higher Bloom levels appears in several open-weight families but is statistically established only selectively. The effect is strongest for Gemini, where both models show significant widening despite near-ceiling performance on the lower-complexity levels. Context conflict can therefore remain consequential even in highly capable models.

4.4 Neutralization and Cognitive Demand

H7 compares the Neutralized-over-Factual advantage at Analyze and Evaluate with that at Remember and Understand. In the fully translated setting, the high-complexity advantage is significantly larger in the Small bucket for seven of eight open-weight series and remains present in several Medium and Large regimes.

The effect weakens or reverses in some stronger families: it remains substantial for Qwen2.5 Base and Instruct, becomes smaller for Qwen3, and is absent or reversed for several of the strongest models, including Gemini 3.1 Pro. We therefore interpret H7 as a recurring, but not universal, interaction between entity-specific associations and cognitive demand.

4.5 Relation Usage Frequency

H6 receives limited support. Although several Medium and Large buckets exhibit observed Low/Medium/High spreads below five percentage points, no open-weight series satisfies the stronger interval-based equivalence criterion in the fully translated setting. Gemini 3.1 Pro also shows a small observed spread, whereas Gemini 3 Flash lies at the five-point boundary under the rounded result. We therefore regard the influence of relation usage frequency as inconclusive rather than absent.

Table 3: Hypothesis support by model series in the **fully translated setting**. S, M, and L indicate that the required direction is observed in the Small, Medium, and Large buckets, respectively; starred labels indicate statistical support under the hypothesis-specific test after Holm correction. † marks an H3 cliff whose lower 95% relation-cluster bootstrap bound exceeds 10 percentage points. x indicates lack of support and — indicates that the comparison is not applicable. H1 is a single series-level test rather than a per-bucket test. Gemini models are reported individually because their parameter counts are undisclosed and are therefore not evaluated under H1 or H2. For Gemini, * denotes statistical support and ✓ denotes the required direction without statistical support. ‡ marks a value at the H6 five-percentage-point boundary under the rounded output.

Series	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
Qwen2.5 Base	*	S,M*,L*	S*,M*†,L*†	S,M,L	S,M	M	S*,M*,L*	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen2.5 Instruct	*	S,M*,L*	S*,M*†,L*†	S,M,L*	S	x	S*,M*,L*	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen3	*	S,M*,L*	S*,M*†,L*†	S,M,L	S,M*,L	M,L	S*,M*,L	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen3.5	*	S,M*,L*	S*†,M*†,L*	S,M	S,L	x	S*,M	S*,M*,L*	S*,M*,L*	S*,M*,L*
Gemma	*	S,M*,L*	S*†,M*†,L*	S,M	S,M,L	x	S*,M	S*,M*,L*	S*,M*,L*	S*,M*,L*
LLaMA	x	S,M*,L*	S*,M*†,L*†	L	x	S	S*,M,L*	S*,M,L*	S*,M*,L*	S*,M*,L*
Granite	x	S,M*,L*	S*,M*†,L*†	S,M,L	S,M	x	S*,M*,L*	S*,M*,L	S*,M*,L*	S*,M*,L*
DeepSeek	x	S,M*	S,M*,L*	M,L	S,L	x	S,M*,L	S*,M*,L*	S*,M*,L*	S*,M*,L*
Gemini 3 Flash	—	—	x	x	*	✓†	✓	*	*	*
Gemini 3.1 Pro	—	—	x	x	*	✓	x	*	*	*

Table 4: Hypothesis support by model series in the **hybrid cross-lingual setting**. Notation follows Table 3. Gemini models were not evaluated in this setting.

Series	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
Qwen2.5 Base	*	S,M*,L*	S*†,M*†,L*†	S,M,L	x	M,L	S*,M*,L*	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen2.5 Instruct	*	S,M*,L*	S*,M*†,L*†	S,M,L*	x	M	S*,M*,L*	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen3	*	S,M*,L*	S*,M*†,L*†	S,M,L	S,M,L	M,L	S*,M*,L*	S*,M*,L*	S*,M*,L*	S*,M*,L*
Qwen3.5	*	S,M*,L*	S*†,M*†,L*	S,M*,L	S,L*	x	S*,M,L	S*,M*,L*	S*,M*,L*	S*,M*,L*
Gemma	*	S,M*,L*	S*†,M*†,L*	S*,M,L	S,M,L*	L	S*,M,L	S*,M*,L*	S*,M*,L*	S*,M*,L*
LLaMA	x	S,M*,L*	S,M*†,L*†	L*	L	S,M	S*,M*,L*	S*,M,L*	S*,M*,L*	S*,M*,L*
Granite	x	S,M*,L*	S*†,M*†,L*†	S,M*,L	M	M	S*,M*,L*	S*,M*,L	S*,M*,L*	S*,M*,L*
DeepSeek	*	S,M*,L	S,L*†	M,L	L	S	S*,M*,L*	M*,L*	S*,M*,L*	S*,M*,L*

4.6 Cross-Language Stability

Item-level difficulty rankings (H9) show consistently positive cross-language stability across every model series and size bucket in both settings and for both Gemini models. In the fully translated setting, the mean pairwise correlations reach approximately $\rho = 0.54$ for Gemini 3 Flash and $\rho = 0.74$ for Gemini 3.1 Pro. Because the 15 language-pair correlations are statistically dependent, we interpret H9 primarily through the magnitude and consistency of these observed correlations rather than as a confirmatory significance test.

Relation-level difficulty rankings show an even stronger inferential pattern (H10): all series and size buckets, together with both Gemini models, exhibit significantly positive cross-language stability under the relation-identity permutation test. Thus, multilingual differences in aggregate accuracy coexist with a shared structure in which many of the same items and relations remain comparatively easy or difficult across languages.

5 Analysis

5.1 Entity Cues Change Role with Cognitive Demand

Averaged over Remember and Understand, Neutralized contexts usually underperform Factual contexts across model-series and size-bucket combinations, suggesting that entity names can provide useful cues for simpler tasks. At Analyze and Evaluate, the pattern often reverses, consistent with entity-specific associations becoming more disruptive as relational demands increase.

Table 5: Representative errors consistent with memory competition. In the Factual condition, the model selects a salient entity-associated answer rather than the relation required by the question.

Entity	Wrong answer	Correct relational answer
Apollo 11	Returned safely to Earth	Recovered by USS Hornet
Bohemian Rhapsody	A six-minute song	Tonality is B-flat major
Amanita muscaria	Red cap with white spots	Spore print color is white

This crossover underlies H7. The concentration of the Neutralized advantage becomes smaller in several higher-performing model series. Because this comparison is exploratory, we do not interpret it as an established scaling law.

5.2 Errors Consistent with Memory Competition

We inspect matched Neutralized–Factual pairs for which Qwen2.5-32B-Instruct succeeds under Neutralized input but fails under Factual input. The Factual errors often select a salient attribute or related fact associated with the named entity rather than the relation requested by the question. Table 5 shows representative examples. Although these cases do not establish an internal causal mechanism, they are consistent with entity-specific parametric associations competing with the relational evidence supplied in context.

6 Conclusions

We presented WIKI-BLOOM, a multilingual diagnostic benchmark for studying reasoning over supplied relational knowledge across model scale, cognitive complexity, knowledge conditions, and language. Across 43 models, scaling generally improves performance but follows family-dependent trajectories, while Bloom’s hierarchy does not correspond to a smooth model-difficulty progression. The Apply-to-Analyze transition is widespread among open-weight models but varies substantially in magnitude and disappears for Gemini, whereas Evaluate remains difficult even for models that perform near ceiling on preceding levels.

The veracity experiments show an equally conditional role for parametric knowledge. Neutralization is not uniformly advantageous: entity identities are associated with better performance on simpler tasks but with larger interference-related effects at higher cognitive demand in several model families. Counterfactual evidence is generally harder to use than factual evidence, with the resulting gap increasing at higher Bloom levels in several families. Contextual and parametric knowledge can therefore cooperate or compete depending on task complexity and model capability.

Cross-language behavior is more consistent. Both item-level and relation-level difficulty rankings are strongly preserved across languages and model families, indicating a shared underlying organization of benchmark difficulty despite differences in aggregate multilingual performance. Together, these results show that stronger models can overcome some reasoning weaknesses without resolving others and motivate evaluation at the level of specific cognitive operations rather than aggregate accuracy alone.

Limitations and future work. WIKI-BLOOM deliberately prioritizes controlled comparison over breadth, using 30 source relations and one source triple per relation. This limits relation coverage and statistical power for small effects, particularly the relation-usage equivalence analysis. Our qualitative memory-competition analysis is also restricted to one checkpoint and should be replicated across model families. Future work can expand relation and language coverage, investigate the observed effects at the representation level, and test whether explicit reasoning strategies alter the persistent difficulty of Evaluate.

Ethics Statement

In accordance with COLM 2026 policy, we disclose that **Claude**, **GPT-5**, and **Gemini** were used to assist in rewriting and improving manuscript drafts across revision rounds. All scientific content, experimental design, and conclusions are the authors' own responsibility.

References

- Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Krithika Ramesh, S. Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. MEGA: Multilingual evaluation of generative AI. In *Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.258.
- Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. LLMs' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *arXiv preprint arXiv:2404.06283*, 2024. doi: 10.48550/arXiv.2404.06283.
- Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, et al. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2335–2357. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.125.
- Boxi Cao, Mengjie Ren, Hongyu Lin, Xianpei Han, Feng Zhang, Junfeng Zhan, and Le Sun. StructEval: Deepen and broaden large language model assessment via structured evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5300–5318. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.314.
- Jasin Cekinmez, Omid Ghahroodi, Saad Fowad Chandle, D. Gupta, and Ehsaneddin Asgari. ADAM: A diverse archive of mankind for evaluating and enhancing LLMs in biographical reasoning. *arXiv preprint arXiv:2509.22991*, 2025. doi: 10.48550/arXiv.2509.22991.
- N. Dong, Minh-Anh Nguyen, T. D. Hoang, N. Ngoc, Dao Xuan Quang Minh, P. Hai, Nguyen Thi Ngoc Anh, Dan Tu, and Binh Vu. VLegal-Bench: Cognitively grounded benchmark for vietnamese legal reasoning of large language models. *arXiv preprint arXiv:2512.14554*, 2025. doi: 10.48550/arXiv.2512.14554.
- Zhiwei Fei, Xiaoyu Shen, D. Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.452.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *International Conference on Machine Learning*, pp. 4411–4421, 2020.
- Ali Khoramfar, Ali Ramezani, Mohammad Mahdi Mohajeri, Mohammad Javad Dousti, Majid Nili Ahmadabadi, and Hesham Faili. DEEPQUESTION: Systematic generation of real-world challenges for evaluating LLMs performance. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 11451–11460. European Language Resources Association (ELRA), 2026.
- David R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218, 2002. doi: 10.1207/s15430421tip4104_2.

- Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.
- Daliang Li, A. Rawat, M. Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Surinder Kumar. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.112.
- Junhong Lin, Bing Zhang, Song Wang, Ziyang Liu, D. Gutfreund, Julian Shun, and Yada Zhu. How much reasoning do retrieval-augmented models add beyond LLMs? a benchmarking framework for multi-hop inference over hybrid knowledge. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9403–9414. ACM, 2026. doi: 10.1145/3770855.3817481.
- Andrew Liu, Henry Prior, Gargi Balasubramaniam, Rivka Moroshko, Amir Zait, Ilia Labzovsky, Danny Karmon, Ishita Dasgupta, Kim Stachenfeld, and Kenneth Marino. ReCogLab: A framework testing relational reasoning and cognitive hypotheses on LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.
- S. Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565.
- Ying-jeou Ma, Jiho Shin, L. D. Silva, Z. Jiang, Song Wang, F. Khomh, and Shin Hwei Tan. BloomAPR: A Bloom’s taxonomy-based framework for assessing the capabilities of LLM-powered APR solutions. *arXiv preprint arXiv:2509.25465*, 2025. doi: 10.48550/arXiv.2509.25465.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.
- Bianca Raimondi and Maurizio Gabbriellini. Mechanistic interpretability of cognitive complexity in LLMs via linear probing using Bloom’s taxonomy. *SSRN preprint*, 2026. doi: 10.2139/ssrn.6997673.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.
- Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhe-Wei Shen, Baocai Chen, Lu Chen, and Kai Yu. SciEval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19053–19061. Association for the Advancement of Artificial Intelligence (AAAI), 2024. doi: 10.1609/aaai.v38i17.29872.
- Travis Thompson, Seung-Hwan Lim, Paul Liu, Ruoying He, and Dongkuan Xu. Inference scaled GraphRAG: Improving multi-hop question answering on knowledge graphs. *arXiv preprint arXiv:2506.19967*, 2025. doi: 10.48550/arXiv.2506.19967.
- Sara Todorovikj, Lars-Peter Meyer, and Michael Martin. Characterizing knowledge graph tasks in LLM benchmarks using cognitive complexity frameworks. In *International Conference on Semantic Systems*, 2025. doi: 10.48550/arXiv.2509.19347.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9: 176–194, 2021. doi: 10.1162/tacl.a_00360.

Liangliang Zhang, Zhuorui Jiang, H. Chi, Haoyang Chen, Mohammed Elkoumy, Fali Wang, Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, et al. Diagnosing and addressing pitfalls in KG-RAG datasets: Toward more reliable benchmarking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75490–75527, 2025.

Yuxuan Zhou, Xien Liu, Chenwei Yan, Chen Ning, Xiao Zhang, Boxun Li, Xiangling Fu, Shijin Wang, Guoping Hu, Yu Wang, et al. Evaluating LLMs across multi-cognitive levels: From medical knowledge mastery to scenario-based problem solving. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2506.08349.

Maria-Eleni Zoumpoulidi, Georgios Paraskevopoulos, and A. Potamianos. Bloom-Wise: Enhancing problem-solving capabilities of large language models using Bloom’s-taxonomy-inspired prompts. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pp. 34–49. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mathnlp-main.3.

A Appendix

A.1 Model Checkpoint Inventory

We evaluate 43 model checkpoints in total.

Qwen2.5 Base. Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, and Qwen2.5-32B.

Qwen2.5 Instruct. Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Qwen2.5-32B-Instruct.

Qwen3. Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, and Qwen3-14B.

Qwen3.5. Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, and Qwen3.5-27B.

Gemma. Gemma 3 270M IT, Gemma 3 4B IT, Gemma 3 12B IT, Gemma 3 27B IT, Gemma 4 E4B IT, Gemma 4 26B-A4B IT, and Gemma 4 31B IT.

LLaMA. Llama 3.2 1B Instruct, Llama 3.2 3B Instruct, Llama 3.1 8B Instruct, and Llama 3.1 70B Instruct.

Granite. Granite 3.3 2B Instruct, Granite 4.0 Micro, Granite 4.0 H-Tiny, Granite 3.3 8B Instruct, and Granite 4.0 H-Small.

DeepSeek. DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Qwen-14B.

Gemini. Gemini 3 Flash and Gemini 3.1 Pro.

A.2 Full Hypothesis Definitions and Statistical Procedures

This appendix provides the complete mathematical definitions corresponding to the compact hypothesis descriptions in the main paper. For hypotheses involving model scale, A_S , A_M , and A_L denote mean accuracy within the Small ($< 7B$), Medium ($\geq 7B$ and $< 14B$), and Large ($\geq 14B$) buckets of a coherent model series.

H1: Diminishing Gains with Model Scale Let

$$G_{SM} = A_M - A_S$$

denote the gain from Small to Medium and

$$G_{ML} = A_L - A_M$$

the gain from Medium to Large. We define

$$D_{\text{scale}} = G_{SM} - G_{ML}.$$

H1 is directionally supported when

$$G_{SM} > 0, \quad G_{ML} > 0, \quad D_{\text{scale}} > 0.$$

Thus, both scaling transitions must improve accuracy, and the Small-to-Medium gain must exceed the Medium-to-Large gain.

H2: Accuracy Increases with Model Size H2 tests the two adjacent scaling contrasts

$$A_M > A_S$$

and

$$A_L > A_M.$$

Support is reported separately by transition. A Medium result indicates that $A_M > A_S$, while a Large result indicates that $A_L > A_M$.

H3: Apply-to-Analyze Performance Drop We define

$$D_{AA} = A_{\text{Apply}} - A_{\text{Analyze}}.$$

H3 tests

$$D_{AA} > 0.$$

In addition to directional and statistical support, we mark an Apply-to-Analyze “cliff” when the drop is statistically significant and the lower bound of its 95% relation-cluster bootstrap interval exceeds 10 percentage points.

H4: Veracity Ordering H4 tests the ordering

$$A_{\text{Neutralized}} > A_{\text{Factual}} > A_{\text{Counterfactual}}.$$

A size bucket supports H4 only when both component inequalities

$$A_{\text{Neutralized}} > A_{\text{Factual}}$$

and

$$A_{\text{Factual}} > A_{\text{Counterfactual}}$$

hold in the expected direction.

H5: Factual-Counterfactual Gap Widens with Cognitive Complexity For Bloom level b , define

$$G_b = A_{\text{Factual},b} - A_{\text{Counterfactual},b}.$$

We summarize lower-complexity levels as

$$G_{\text{low}} = \frac{G_{\text{Remember}} + G_{\text{Understand}}}{2},$$

and higher-complexity levels as

$$G_{\text{high}} = \frac{G_{\text{Analyze}} + G_{\text{Evaluate}}}{2}.$$

Apply is omitted because it occupies the middle position in the five-level sequence.

We define

$$D_{\text{FC}} = G_{\text{high}} - G_{\text{low}},$$

and H5 tests

$$D_{\text{FC}} > 0.$$

H6: Relation Usage Frequency Has a Negligible Effect For each size bucket, let

$$A_{\text{Low}}, \quad A_{\text{Medium}}, \quad A_{\text{High}}$$

denote mean accuracy on relations in the three usage-frequency tiers.

We define the observed spread as

$$R_{\text{usage}} = \max\{A_{\text{Low}}, A_{\text{Medium}}, A_{\text{High}}\} - \min\{A_{\text{Low}}, A_{\text{Medium}}, A_{\text{High}}\}.$$

The effect is considered practically negligible when

$$R_{\text{usage}} < 0.05,$$

corresponding to a maximum spread below five percentage points.

For statistical equivalence, we consider the three pairwise contrasts:

$$A_{\text{Low}} - A_{\text{Medium}},$$

$$A_{\text{Low}} - A_{\text{High}},$$

and

$$A_{\text{Medium}} - A_{\text{High}}.$$

A bucket receives statistical support only when the corresponding 95% relation-cluster bootstrap intervals lie entirely inside the ± 5 percentage-point equivalence region.

The confirmatory H6 claim concerns models at or above 7B parameters, so Medium and Large are the confirmatory buckets. Small is reported only as an exploratory diagnostic.

H7: Neutralized Advantage Is Concentrated at Higher Bloom Levels For each Bloom level b , define the Neutralized-over-Factual advantage

$$N_b = A_{\text{Neutralized},b} - A_{\text{Factual},b}.$$

We summarize the lower levels by

$$N_{\text{low}} = \frac{N_{\text{Remember}} + N_{\text{Understand}}}{2},$$

and the higher levels by

$$N_{\text{high}} = \frac{N_{\text{Analyze}} + N_{\text{Evaluate}}}{2}.$$

Apply is omitted as the middle level.

We define

$$D_{\text{NF}} = N_{\text{high}} - N_{\text{low}},$$

and H7 tests

$$D_{\text{NF}} > 0.$$

Thus, H7 asks whether removing entity identity provides a larger relative benefit under higher cognitive demand.

H8: Evaluate Is a Recurring Bloom Bottleneck Let

$$A_{\text{other}} = \frac{A_{\text{Remember}} + A_{\text{Understand}} + A_{\text{Apply}} + A_{\text{Analyze}}}{4}.$$

We define the Evaluate bottleneck gap as

$$D_{\text{Eval}} = A_{\text{other}} - A_{\text{Evaluate}}.$$

H8 tests

$$D_{\text{Eval}} > 0.$$

This definition asks whether Evaluate is harder than the average of the other four Bloom levels rather than requiring it to be worse than each level individually.

H9: Item-Level Difficulty Is Stable Across Languages For item i and language ℓ , let $a_{i,\ell}$ denote its mean accuracy within a model series and size bucket. For every pair of languages (ℓ_1, ℓ_2) , we compute

$$\rho_{\ell_1, \ell_2}^{\text{item}} = \text{Spearman}(\{a_{i, \ell_1}\}_i, \{a_{i, \ell_2}\}_i).$$

With six languages, this yields 15 language-pair correlations. We summarize item-level cross-language stability by

$$R_{\text{item}} = \frac{1}{15} \sum_{\ell_1 < \ell_2} \rho_{\ell_1, \ell_2}^{\text{item}},$$

and H9 tests

$$R_{\text{item}} > 0.$$

Significance is evaluated using an item-identity permutation test. One reference language is held fixed, while item identities are independently permuted in each of the remaining languages. The complete set of 15 pairwise Spearman correlations and their mean are recomputed after each permutation, thereby preserving the dependence structure among language-pair correlations. The one-sided permutation p -value is the proportion of permuted mean correlations at least as large as the observed value.

Confidence intervals are obtained using a relation-cluster bootstrap: the 30 source relations are resampled with replacement, all items associated with each sampled relation are carried together, and the complete mean-correlation statistic is recomputed for each bootstrap sample.

H10: Relation-Level Difficulty Hierarchy Is Stable Across Languages For relation r and language ℓ , let $A_{r,\ell}$ denote mean accuracy on that relation. For every pair of languages, compute

$$\rho_{\ell_1, \ell_2}^{\text{rel}} = \text{Spearman}(\{A_{r, \ell_1}\}_r, \{A_{r, \ell_2}\}_r).$$

We summarize relation-level stability by

$$R_{\text{rel}} = \frac{1}{15} \sum_{\ell_1 < \ell_2} \rho_{\ell_1, \ell_2}^{\text{rel}}.$$

H10 tests

$$R_{\text{rel}} > 0.$$

A positive result means that relations that are comparatively difficult in one language tend to retain their relative difficulty in the others.

Statistical Inference Questions derived from the same Wikidata relation share a source triple and therefore are not treated as independent observations.

For H1–H5, H7, and H8, we first construct the hypothesis-specific matched contrast and then average the contrast within each of the 30 source relations. Directional hypotheses are evaluated using one-sided relation-cluster sign-flip permutation tests with 10^5 Monte Carlo permutations. We report 95% relation-cluster bootstrap confidence intervals based on 2×10^4 resamples.

Within each model series, p -values are Holm-corrected over the relevant size-bucket comparisons, with

$$\alpha = 0.05.$$

H6 is an equivalence hypothesis rather than a directional hypothesis. We therefore use the predefined five-percentage-point equivalence region and assess whether the relation-cluster bootstrap intervals for the relevant pairwise usage-frequency contrasts lie entirely inside that region.

For H9, the statistic is the mean of the 15 pairwise item-level Spearman correlations. Significance is evaluated using an item-identity permutation test: one reference language is held fixed, while item identities are independently shuffled in each of the remaining languages, and the complete mean-correlation statistic is recomputed after each permutation. Confidence intervals are obtained using a relation-cluster bootstrap that resamples source relations and carries all associated items together.

For H10, significance is evaluated using a relation-identity permutation test. Relation correspondences are independently shuffled across languages to break any stable cross-language relation ranking, after which the mean of the 15 pairwise Spearman correlations is recomputed. Confidence intervals are obtained by resampling relations and recomputing the full correlation statistic.

A.3 Benchmark Construction Details

Relation candidate pool and usage tiers. We construct a candidate pool of Wikidata properties starting from the relations in Wikidata5M (Wang et al., 2021). For each relation, we take its primary English alias as a human-readable handle, deduplicate, and randomly shuffle the resulting list using a fixed seed, retaining up to $N=1000$ candidates for downstream sampling. We estimate relation usage frequency by querying the MediaWiki search API with `haswbstatement:PID` and using the returned `totalhits` as a proxy. This measures the prevalence of the property in Wikidata-indexed content and should not be interpreted as a direct estimate of model pretraining exposure. We apply a $\log(1 + \text{hits})$ transformation to reduce skew and split values into tertiles, defining Low, Medium, and High usage tiers. We then sample up to 30 properties per tier using a fixed random seed, yielding a candidate pool of up to 90 properties for manual curation.

Triple extraction and manual curation. For each sampled property p , we extract exactly one representative factual triple (h, p, t) from popular English Wikipedia pages, ranked by sitelink count as a proxy for popularity. From the automatically sampled candidates, we manually curate a final set of 30 relations, 10 per usage tier, excluding relations involving sensitive topics, overly technical semantics, or cases where multiple relations feature the same entity. We prioritize semantic clarity and domain diversity across geography, science, engineering, sports, entertainment, and linguistics.

A.4 Item Quality Control

All generated questions, answer options, and answer keys were manually reviewed by the authors for factual correctness, preservation of the intended relation, answer uniqueness, and distractor plausibility. As an additional quality-control step, every item was independently evaluated using Gemini and Grok. Each model received the fact sentence, question, answer options, and intended answer key and was asked to identify cases in which it disagreed with the provided answer. All flagged cases were manually inspected. When disagreement occurred, we examined whether it reflected model error or plausible ambiguity in the question or answer options. Valid objections led to revisions of wording or distractors, followed by re-evaluation before inclusion.