Think Before You Link: Rarity, Reasoning, and Retrieval in Multilingual Entity Linking

Anonymous ACL submission

Abstract

Entity linking systems degrade on rare entities, but prior work measures rarity primarily through popularity proxies such as Wikipedia pageviews. Using Wikidata structural metrics, we find that unpopularity is not the same as rarity. Structurally sparse entities cause a 53 percentage point accuracy drop, while merely unpopular ones cause only 38pp. Structural sparsity tracks the culturally niche long tail, so popularity-based evaluation hides this failure mode. To investigate it, we build a simple framework in which a reasoning-capable vision-language model iteratively searches and reasons over Wikipedia, gathering evidence dynamically instead of reranking a fixed candidate set. Reasoning alone shows no significant effect on rare entities, and retrieval alone can hurt on common ones, but combined they drive the gains. On MERLIN, a multilingual multimodal entity linking benchmark over five languages (Hindi, Indonesian, Japanese, Tamil, Vietnamese), our best system reaches 87.9% average accuracy, +6.9pp over the state-of-the-art and +37pp on the hardest rare slices. We will release MERLIN-Rare, the rare entity test slices, along with our evaluation framework and code upon acceptance.

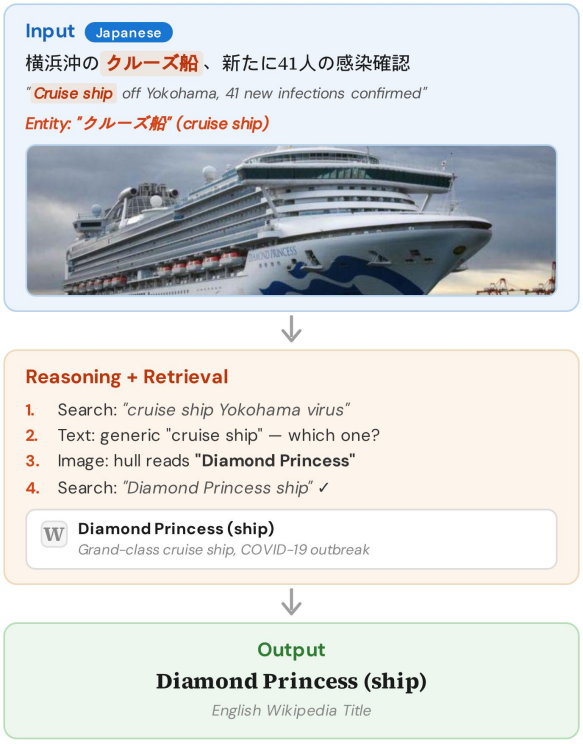

Figure 1: Multimodal entity linking. The Japanese text mentions a generic "cruise ship," which cannot be resolved from text alone. The accompanying image shows the vessel with its name "Diamond Princess" written on the hull, enabling a targeted search to link to the correct entity.

1 Introduction

Entity linking, or the task of grounding textual mentions to knowledge base entries, is foundational for knowledge-intensive NLP applications (Sevgili et al., 2022; Shen et al., 2015). As vision-language models become increasingly capable, there is growing interest in multimodal entity linking, where visual context can help disambiguate mentions that would be ambiguous from text alone (Shi et al., 2024; Moon et al., 2018). This is particularly valuable in multilingual settings, where images provide language-agnostic signal that can bridge gaps in textual coverage (Ramamoorthy et al., 2025).

However, entity linking models exhibit a well-documented failure pattern. They perform well on common, frequently-referenced entities but degrade sharply on the long tail (Boscariol et al., 2025; Hoveyda et al., 2024; Ilievski et al., 2018). This problem is particularly severe for culturally niche entities, or those well-known within specific language communities but sparsely represented in cross-lingual knowledge bases (Veselovsky et al., 2025; Naous et al., 2024), as opposed to entities that are merely unpopular. Prior work, however, has gauged entity rarity through popularity-based proxies such as Wikipedia pageviews and incoming link counts (Graciotti et al., 2025; Mallen et al., 2023; Chen et al., 2021), which capture access frequency but may not reflect cultural specificity.

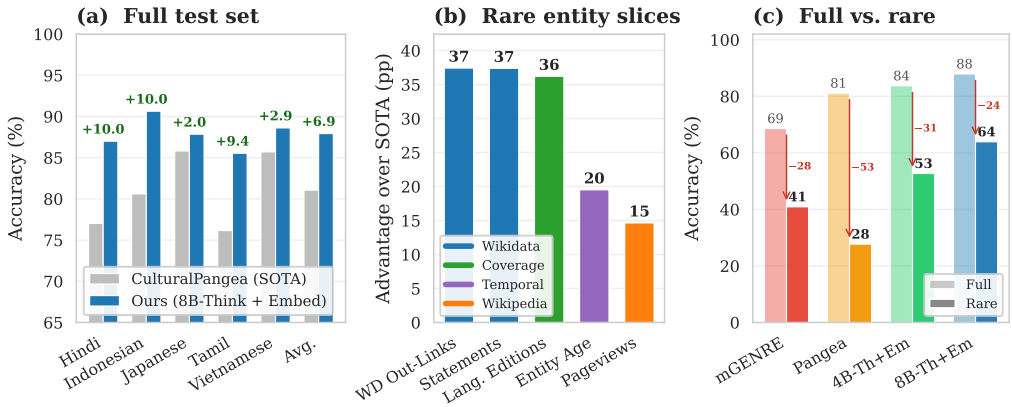

Figure 2: Our best system (8B-Think+Embed) vs. baselines. (a) Full test set accuracy across five languages: +6.9pp average over SOTA. **(b)** Advantage over Pangea on bottom-5% rare entity slices grows to **+37pp** on Wikidata-structural metrics. **(c)** On rare entities (bottom 5% by language editions), Pangea drops to 28% while our system maintains 64%—even our 4B model outperforms Pangea by 25pp.

We study this problem on MERLIN (Ramamoorthy et al., 2025), a multilingual multimodal entity linking benchmark spanning five languages (Hindi, Indonesian, Japanese, Tamil, and Vietnamese), and propose a broader characterization of entity rarity that distinguishes cultural specificity from mere unpopularity. Through systematic analysis of multiple Wikipedia and Wikidata signals, we show that entities with few Wikipedia language editions, sparse Wikidata statements, and limited outgoing links exhibit far more severe model degradation than entities that are simply low in pageviews.

On the current state-of-the-art (de Dieu Nyandwi et al., 2025) model, entities in the bottom 5% by language editions show accuracy drops of 53 percentage points relative to the full test set, compared to 38 percentage points for the bottom 5% by pageviews. This 15pp gap reveals that culturally niche entities pose a distinct and more severe challenge than what prior popularity-based analyses have surfaced.

Rare entities are unlikely to be well-represented in model weights, motivating retrieval-augmented approaches that access external knowledge at inference time (Ding et al., 2025; Pons et al., 2024; Liu et al., 2024). However, retrieval alone may be insufficient as entity disambiguation often involves contextual reasoning. We investigate whether a simple framework combining reasoning-capable vision-language models with retrieval over Wikipedia can address rare entity failures.

As shown in Figure 2, our best system achieves improvements of **+2.1** to **+10.0** percentage points over the SOTA baseline across languages. The gains are substantially larger on rare entities: the full-dataset advantage of **+6.9pp** grows to **+37pp** on the hardest Wikidata-structural slices, reaching **+47pp** on Hindi. This confirms that iterative reasoning with retrieval addresses the long-tail failures that existing models struggle with.

In summary, our contributions are:

- **Rarity characterization:** We propose a multidimensional characterization of entity rarity using Wikidata structural metrics, showing that structurally rare entities cause more severe degradation (up to 53pp) than merely unpopular ones (38pp). We will release MERLIN-RARE, a set of rare entity test slices for targeted evaluation, upon acceptance.
- **Framework:** We present a simple framework combining reasoning-capable VLMs with iterative retrieval over Wikipedia, achieving **+6.9pp** over the state-of-the-art on MERLIN and **+30–47pp** on the rare entity slices.
- **Analysis:** We provide a detailed analysis of model behavior. Notably, retrieval hurts non-reasoning models on common entities but helps on rare ones. Our failure taxonomy and search behavior analysis further reveal that retrieval failure accounts for 72% of residual errors, and that reasoning models make fewer but more targeted retrieval calls.

2 Related Work

Entity Linking with Large Language Models. Entity linking has shifted from classification over candidate sets to direct generation, beginning with GENRE's (Cao et al., 2021a) autoregressive entity retrieval and extended via context enrichment and adaptive routing (Ding et al., 2024, 2025; Xin et al.,

2025; Liu et al., 2024; Li et al., 2025). LELA (Haffoudhi et al., 2026) retrieves once then reasons with self-consistency voting; ELA (Luo et al., 2025) uses a single retrieval call with no query refinement. All operate text-only and predominantly in English; none investigate iterative evidence gathering for culturally niche entities.

Multimodal and Multilingual Entity Linking. Multimodal EL leverages visual context to disambiguate text-ambiguous mentions (Moon et al., 2018; Wang et al., 2022; Shi et al., 2024; Liu et al., 2025), while multilingual EL has advanced via autoregressive and end-to-end methods (Cao et al., 2021b; Limkonchotiwat et al., 2023), but the two have remained largely separate. MERLIN (Ramamoorthy et al., 2025) is one of the first benchmarks at this intersection; Cultural Pangea (de Dieu Nyandwi et al., 2025) establishes the SOTA by fine-tuning a multilingual VLM on culturally grounded data, yet still degrades sharply on structurally rare entities (Section 4).

Entity Rarity and Cultural Representation. EL systems degrade on rare entities (Ilievski et al., 2018; Hoveyda et al., 2024; Boscariol et al., 2025), with rarity typically equated with low popularity (Mallen et al., 2023; Kandpal et al., 2023; Chen et al., 2021). However, unpopularity is not the same as cultural specificity: LLMs exhibit Western-centric entity bias (Naous et al., 2024), VLM performance correlates with per-language Wikipedia size (Bugliarello et al., 2022), and localized cultural knowledge is poorly represented cross-lingually (Veselovsky et al., 2025; Tao et al., 2024; Adilazuarda et al., 2024). We distinguish these through a multidimensional rarity characterization separating structural sparsity from popularity.

Retrieval-Augmented Reasoning. ReAct (Yao et al., 2023) and IRCoT (Trivedi et al., 2023) interleave reasoning with retrieval; reasoning-native models trained via RL (Guo et al., 2025; Team, 2025; Jin et al., 2025; Feng et al., 2025) learn to call search within their thinking, and inference-time compute can substitute for parameters (Snell et al., 2024). This paradigm has been applied to QA but not yet to entity linking, which is the gap we close.

3 Task Definition

Entity linking is the task of mapping textual entity mentions to entries in a knowledge base (Shen et al., 2015). We study a multilingual, multimodal formulation of this task, focusing on entity linking given a marked mention. We utilize the MERLIN dataset (Ramamoorthy et al., 2025) as our evaluation set.

We follow MERLIN's setup. Given a text passage T in a source language, an accompanying image I, and a marked entity mention $m \in T$, the task is to predict the English Wikipedia title of the entity referenced by m. We evaluate using exact-match accuracy against gold annotations.

4 Entity Rarity Analysis

Before presenting our methodology, we characterize the rarity problem that motivates our approach. We define entity rarity along multiple dimensions and show that the current state-of-the-art fails on rare entities.

4.1 Defining Entity Rarity

Prior work commonly operationalizes entity rarity through popularity-based signals such as Wikipedia pageviews or incoming link counts (Graciotti et al., 2025; Xin et al., 2025; Mallen et al., 2023; Ilievski et al., 2018). However, popularity is only one dimension of rarity. An entity may be rare because few languages document it, because it lacks rich semantic structure in knowledge bases, or because it has received limited editorial attention. We propose a broader conceptualization spanning multiple dimensions.

Rarity Metrics. We collect two families of metrics via the Wikipedia and Wikidata APIs. **Wikipedia-based metrics** reflect editorial attention and documentation depth: pageviews (90-day), backlinks, article size, revision count, unique editors, category count, external links, reference count, and image count. **Wikidata-based metrics** reflect structural connectivity and cross-lingual coverage: incoming links, outgoing links, language editions (number of Wikipedia languages with an article), statement count, qualifier count, and entity age.

These dimensions are complementary. An entity could be popular (high pageviews) yet structurally rare (few Wikidata statements), or well-documented in one language yet globally obscure (few language editions). Overlap analysis confirms this. The bottom-5% entity sets for different metrics share only 37% of their entities on average (Appendix A.1), with some pairs overlapping as little as 10%.

We further note that cross-lingual knowledge base coverage tracks cultural representation in models. VLM accuracy on multilingual benchmarks scales with per-language Wikipedia size (Bugliarello et al., 2022), LLMs default to English-centric outputs even when prompted in other languages (Veselovsky et al., 2025; Naous et al., 2024), and digitally underrepresented cultures receive "thin descriptions" that amplify downstream bias (Adilazuarda et al., 2024). We therefore treat structural sparsity in cross-lingual signals as a culturally meaningful rarity signal.

4.2 Baseline Degradation on Rare Entities

We evaluate Cultural Pangea (de Dieu Nyandwi et al., 2025), the current state-of-the-art on MERLIN (81.1% avg), on bottom-5% entity slices for each metric. Crucially, Cultural Pangea was explicitly trained on culturally-grounded data, making it a strong test case for examining whether rarity remains problematic for models designed to handle diverse entities.

Figure 3 shows that entities that are merely unpopular (bottom 5% by pageviews) show a mean accuracy drop of 38pp, while entities that are culturally niche (bottom 5% by language editions) show a far more severe drop of 53pp, suggesting that previous work may have underestimated the performance degradation on rare entities. Table 1 presents per-language breakdowns: Vietnamese and Tamil exhibit the most severe degradation on language editions (18.2% and 23.8%), while Japanese is most robust (42.7%). This degradation is not specific to CulturalPangea, GEMEL and mGENRE show similar drops of 28–30pp on rare slices (Appendix A.2).

Evaluation Slice	Hi	Id	Ja	Ta	Vi
Full Test Set	77.0	80.6	85.8	76.2	85.7
Wikipedia-based metrics (Bottom 5%)					
Pageviews (90d)	42.0	52.9	44.2	29.0	46.0
Backlinks	56.7	55.9	40.7	43.5	50.8
Article Size	34.8	62.9	50.0	38.7	45.2
Revision Count	34.8	55.7	45.3	33.9	50.0
Unique Editors	37.7	55.7	43.5	33.9	48.4
Category Count	**24.1**	55.4	41.1	40.0	42.2
External Links	32.8	60.0	43.0	32.3	47.5
Reference Count	32.8	59.7	48.8	36.8	53.1
Image Count	53.4	62.7	52.6	43.6	56.7
Wikidata-based metrics (Bottom 5%)					
Incoming Links	53.3	43.8	45.9	47.5	37.7
Outgoing Links	26.6	**26.4**	43.5	26.2	21.5
Language Editions	24.2	29.6	**42.7**	**23.8**	18.2
Statement Count	30.0	27.9	44.4	27.6	**13.8**
Qualifier Count	28.0	34.8	45.9	29.6	18.9
Entity Age	50.7	56.9	68.2	28.6	53.7

Table 1: Cultural Pangea accuracy (%) on the full MERLIN test set and bottom 5% slices for each rarity metric. Culturally niche entities (bottom 5% by Wikidata language editions) show the most severe degradation, with accuracy dropping to 18.2% for Vietnamese and 23.8% for Tamil.

To facilitate targeted evaluation on rare entities, we will release **MERLIN-Rare**, bottom-5% entities for each metric across all five languages, upon acceptance.

5 Methodology

Culturally niche entities are those least likely to be well-represented in model parameters, since training data skews toward well-documented entities. Furthermore, disambiguating rare entities may require multi-step reasoning (Trivedi et al., 2023). We propose a simple framework with a reasoning-capable VLM with iterative retrieval over external knowledge sources. The model performs multi-step reasoning, issuing search queries and incorporating retrieved evidence until confident. We evaluate both lexical and semantic retrieval to study how retrieval quality affects entity linking, particularly for rare entities.

5.1 Reasoning with Retrieval

Model Selection. We use the Qwen3-VL model family (Team, 2025), one of the best-performing open-source VLMs, in both Thinking (reasoning-native) and Instruct variants at 2B, 4B, and 8B parameter sizes. This family uniquely provides matched architecture across reasoning and non-reasoning variants at multiple scales, enabling controlled comparisons that isolate the contributions of reasoning, retrieval, and model size.

Retrieval System. While rare entities are unlikely to be encoded in model parameters, they may still be documented in external knowledge sources. Retrieval-augmented approaches can bridge this gap by accessing such sources at inference time. We use English Wikipedia as our retrieval corpus and evaluate two strategies:

BM25 (Lexical). We built a retrieval system using the `wikimedia/structured-wikipedia` dataset from Hugging Face, which provides preprocessed Wikipedia dumps for the English language. We indexed articles using BM25 (Lù, 2024), a lexical retrieval method that matches query terms against document terms. BM25 is fast and effective

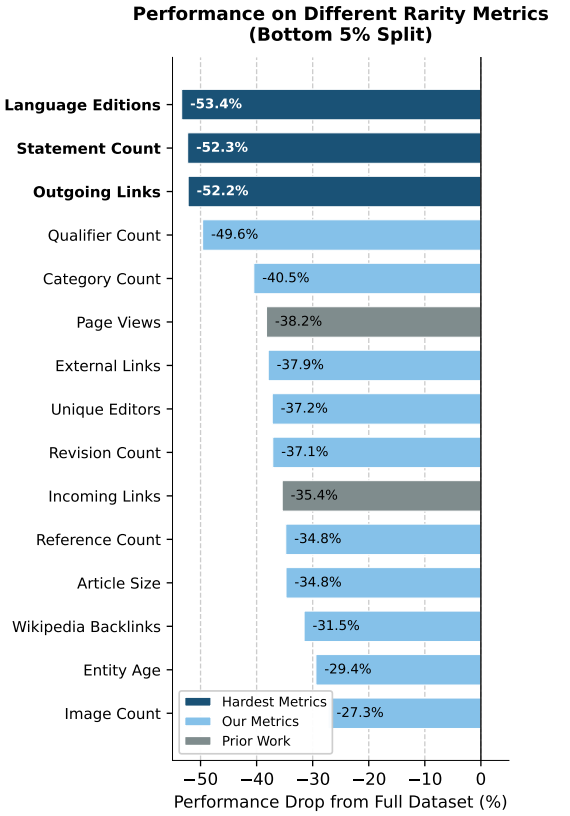

Figure 3: Mean accuracy drop (pp) when evaluating Cultural Pangea on the bottom 5% of entities per rarity metric, relative to full test set performance.

when query terms overlap with target titles, however it has a clear flaw when entity mentions appear in non-Latin scripts and must be transliterated to match English Wikipedia titles for searching.

Embedding (Semantic). To address the cross-lingual limitation of BM25, we also evaluate semantic retrieval using `intfloat/multilingual-e5-large-instruct` (Wang et al., 2024), a multilingual embedding model. We encode all Wikipedia title-description pairs into a FAISS index (Douze et al., 2024), enabling nearest-neighbor search.

Implementation. Given an input image I, text passage T, and entity mention m, the model performs iterative reasoning with retrieval access. At each step, the model will:

1. Analyze the visual and textual context to identify disambiguating signals
2. Issue a search query to the retrieval system
3. Incorporate retrieved Wikipedia snippets into its reasoning
4. Repeat until confident in a final answer

The model is allowed up to 20 retrieval iterations per example. It produces a freeform explanation of its reasoning process, including which candidate entities it considered and why it selected the final answer. A second pass then re-prompts the model with its own complete reasoning trace and instructs it to output only the final Wikipedia title.

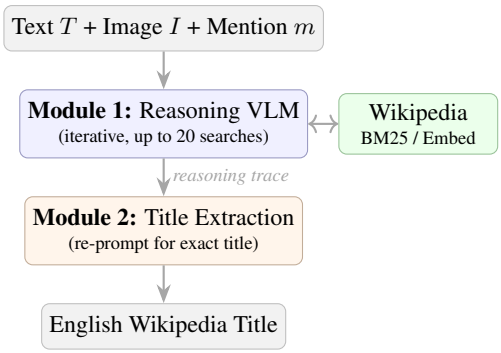

Figure 4: Two-module pipeline. Module 1 performs iterative reasoning with retrieval access over Wikipedia. Module 2 re-prompts the model to extract the final title from the reasoning.

We note that a preliminary single-pass variant where the model was instructed to only emit its final answer at the end of the reasoning trace produced verbose or hedged outputs (restating multiple candidates, appending qualifiers) that failed exact match despite containing the correct entity.

6 Experimental Setup

Model variants. We evaluate Qwen3-VL (Team, 2025) in two variants: **Thinking** (reasoning-native, trained with reinforcement learning to produce extended reasoning traces) and **Instruct** (standard instruction-tuned). Both share the same base architecture and are evaluated at 2B, 4B, and 8B parameter sizes.

Retrieval methods. Each model variant is evaluated under three retrieval conditions: (1) **No RAG**: the model relies solely on parametric knowledge, (2) **BM25**: lexical retrieval over English Wikipedia, and (3) **Embedding**: semantic retrieval using multilingual embeddings with FAISS. This yields 6 configurations per model size (18 total). The resulting $3 \times 2 \times 3$ factorial design (scale $\times$ training mode $\times$ retrieval) isolates the contributions of model scale, reasoning, and retrieval to overall and rare-entity accuracy.

Baselines. We compare against four baselines in total. Three published baselines on MERLIN: **GEMEL** (Shi et al., 2024) (58.7%), a generative multimodal entity linking approach; **mGENRE** (Cao et al., 2021b) (68.6%), which per-

Model	RAG	Hindi	Indo.	Japan.	Tamil	Viet.	Avg.
Prior work							
GEMEL	—	55.5	71.6	73.2	23.9	69.4	58.7
mGENRE	—	55.7	78.9	71.1	63.2	74.1	68.6
CulturalPangea-7B	—	77.0	80.6	85.8	76.2	85.7	81.1
Retrieval-aware baseline							
CulturalPangea-RAG	Embed	67.7	80.0	86.3	55.4	83.1	74.5
Ours							
2B-Instr	—	62.2	74.7	64.4	50.4	65.8	63.5
2B-Instr	BM25	47.7	63.3	54.4	41.7	60.9	53.6
2B-Instr	Embed	54.8	62.1	55.6	43.1	55.7	54.3
2B-Think	—	62.5	74.2	63.6	45.6	67.7	62.7
2B-Think	BM25	62.8	68.2	62.1	45.7	63.3	60.4
2B-Think	Embed	59.1	69.3	59.3	46.0	61.6	59.1
4B-Instr	—	75.4	79.7	75.6	71.2	78.9	76.2
4B-Instr	BM25	74.1	79.0	67.6	70.6	72.3	72.7
4B-Instr	Embed	76.7	83.8	70.7	73.4	76.1	76.1
4B-Think	—	77.9	83.5	79.8	72.7	84.2	79.6
4B-Think	BM25	78.9	85.6	80.5	73.9	84.6	80.7
4B-Think	Embed	83.0	89.6	82.9	76.4	86.7	83.7
8B-Instr	—	81.4	86.6	83.0	80.6	85.7	83.5
8B-Instr	BM25	77.4	80.8	79.4	75.4	79.8	78.6
8B-Instr	Embed	80.0	86.4	86.1	77.3	84.9	82.9
8B-Think	—	81.7	86.8	84.2	82.0	86.1	84.2
8B-Think	BM25	82.8	88.0	86.6	84.4	<u>87.4</u>	85.8
8B-Think	Embed	**87.0**	**90.6**	**87.9**	**85.5**	**88.6**	**87.9**
Δ *best vs. Pangea*		*+10.0*	*+10.0*	*+2.1*	*+9.3*	*+2.9*	*+6.9*

Table 2: Accuracy (%) on the full MERLIN test set. **Bold** = best per column, <u>underline</u> = second best. Our best system (8B-Think+Embed) outperforms Cultural Pangea by +6.9pp on average. Reasoning models consistently outperform instruct models. Embedding retrieval outperforms BM25. BM25 hurts instruct models but helps reasoning models.

forms multilingual autoregressive entity retrieval with constrained beam search over Wikipedia titles; and **Cultural Pangea** (de Dieu Nyandwi et al., 2025) (81.1%), the current SOTA on MERLIN, a multilingual VLM fine-tuned on culturally grounded data.

We also create an additional retrieval-aware baseline using our embedding retrieval. Since CulturalPangea does not support tool calling, **CulturalPangea-RAG** prepends the top-5 retrieved (title, description) pairs to Pangea's input.

7 Results

We evaluate our framework on both the full MERLIN test set and the rare entity slices defined in §4. After presenting overall results, we address three research questions: (RQ1) How does the advantage over existing methods scale with entity rarity? (RQ2) What drives the gains on rare entities: reasoning, retrieval, or their combination? (RQ3) Can smaller reasoning models with retrieval match larger ones?

7.1 Main Results

Table 2 presents accuracy on the full MERLIN test set. Our best system, 8B-Think+Embed, achieves 87.9% average accuracy, outperforming Cultural Pangea by **+6.9** percentage points, with gains of +10.0pp on Hindi and Indonesian. Reasoning models consistently beat their instruct counterparts under retrieval; 4B-Think+Embed (83.7%) matches 8B-Instr (83.5%) with half the parameters, though it uses ∼2.8× more inference tokens than 8B-Instr no-RAG (Appendix A.12, Table 15): the trade is fewer parameters for more compute per example.

RQ1: How Does the Advantage Scale with Entity Rarity? While improvements on the full test set are substantial, the gains become far more dramatic on rare entities. Appendix A.3 shows our advantage over Pangea across all 15 rare entity slices. On the hardest slices (bottom 5% by Wikidata structural metrics), we observe substantially larger gains compared to the full dataset: **+37.5pp** on Wikidata (WD) Out-Links, +37.3pp on Statements, and +37.0pp on Qualifiers, compared to +6.9pp on the full dataset. Even on the less severe popularity-based slices, advantages of +13–15pp

substantially exceed the full-dataset improvement. This pattern is consistent across all rarity metrics.

Table 3 quantifies the key splits. The pattern is clear: Wikidata-structural metrics, which capture how well-connected an entity is in the knowledge graph, identify the entities where our system provides the most value.

Split (Bottom 5%)	Ours	Δ vs Pangea
Full dataset	87.9	+6.9
Wikidata structural		
WD Out-Links	66.3	**+37.5**
Statements	66.1	+37.3
Qualifiers	68.4	+37.0
Lang. Editions	63.9	+36.2
WD In-Links	69.0	+23.4
Wikipedia engagement		
Pageviews	57.5	+14.7
Categories	59.5	+18.9

Table 3: 8B-Think+Embed accuracy (%) and advantage over Pangea on rare entity slices. The +6.9pp full-dataset advantage grows to +37.5pp on the hardest slices (5.4×).

RQ2: What Drives the Gains: Reasoning, Retrieval, or Their Combination? Table 4 reveals the mechanism the benefit of retrieval grows dramatically on rare entities. For 8B-Think+Embed, the RAG delta grows from +3.8pp on the full dataset to +18.2pp on language editions. On common entities, the model may already know the answer parametrically, so retrieval is redundant. On rare entities, parametric knowledge fails and retrieval becomes essential.

Split	Th+Em	Th+BM	In+BM
Full dataset	+3.8	+1.7	−4.9
Lang. Ed.	+18.2	+12.2	+12.6
Statements	+16.2	+12.1	+10.1
WD Out-Links	+15.6	+12.1	+9.0
Qualifiers	+15.7	+12.5	+8.1
Categories	+8.6	+2.0	+0.4

Table 4: RAG delta (pp) vs. no-RAG baseline. Th+Em = 8B-Think+Embed, Th+BM = 8B-Think+BM25, In+BM = 8B-Inst+BM25.

The Instruct Reversal. On the full dataset, BM25 retrieval hurts the instruct model by −4.9pp (Table 4). Yet on rare entity slices, the same BM25 retrieval helps by +9 to +13pp. This reversal occurs because instruct models issue searches 3.2-3.7 searches per example with no deliberation between them, flooding their context with retrieved results they cannot effectively filter. On common entities this noise overwhelms the correct parametric answer. On rare entities any retrieval signal helps because parametric knowledge is absent. Instruct models' queries degrade over successive iterations, with verbatim repetition climbing to 34% by search number 15+ (Appendix A.11). Thinking models avoid all of this: Think+BM25 shows consistent positive deltas on both full (+1.7pp) and rare (+12pp) data.

Reasoning alone does not help on rare entities. The reasoning effect (Think vs. Inst, both without RAG) is not significant on any rare split ($p > 0.5$ across all slices; Table 7). Think and Inst models perform identically at ∼45–53% on rare splits without retrieval. The entire advantage comes from the combination of reasoning and retrieval: reasoning enables effective use of retrieved information, while retrieval provides information that reasoning alone cannot manufacture.

How reasoning models use retrieval differently. Thinking models make fewer but more deliberate searches (1.0–2.2 as opposed to 3.2-5.2 per example for instruct models). They also generate 1,454–1,534 tokens between consecutive searches. This deliberation manifests in query strategy. 56–58% of 8B-Think model's search transitions are refinements that add disambiguation context to the previous query, compared to 35–40% for Instruct variant (Appendix A.13).

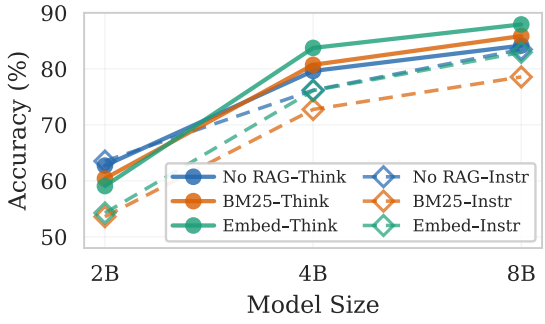

Figure 5: Accuracy vs. model size (2B/4B/8B) across all 6 configurations.

RQ3: Can Smaller Reasoning Models Match Larger Ones? Figure 5 shows accuracy across model sizes for all configurations. Two patterns are notable. First, Embed-Think leads at 4B and 8B. At 4B it already surpasses 8B-Instr, however at 2B the no-RAG baselines outperform all retrieval-augmented configurations. Further analysis reveals that 2B-Think issues only one (forced) search on

97% of examples. It cannot effectively use the tool at all, indicating that effective tool use is an emergent capability requiring sufficient model scale (Appendix A.11). Second, the Think–Instruct gap is substantially wider under retrieval (both BM25 and embedding) than without, suggesting that reasoning and retrieval are complementary.

Table 5 quantifies this on rare entities: 4B-Think+Embed and 8B-Inst are virtually identical on the full dataset (+0.3pp), but on rare entities the smaller reasoning model wins by +5 to +7pp. This demonstrates that reasoning with retrieval compensates for model size, with the advantage growing on harder entities.

Split	4B-Th+Em	8B-Inst	Gap
Full dataset	83.7	83.5	+0.3
Lang. Ed.	52.7	46.8	+5.9
Statements	56.1	49.2	+6.9
WD Out-Links	57.5	50.4	**+7.1**
Qualifiers	59.3	53.3	+6.0

Table 5: 4B-Think+Embed vs. 8B-Inst: nearly identical on the full dataset, but the smaller reasoning model wins by +5–7pp on rare entities.

7.2 Error Analysis

We analyzed 8B-Think+Embed's 841 errors (Table 6) to provide a taxonomy and inform future work. Further error analysis can be found in Appendix A.7 and A.8. Detailed search behavior analysis, computational cost breakdowns, and query content analysis are in Appendices A.11–A.13.

We additionally decompose errors by pipeline stage (Appendix A.8). The dominant bottleneck is retrieval: in 72% of errors, the correct entity was never surfaced by search. Even the deliberate query refinement that Thinking models employ (Appendix A.13) cannot always bridge cross-lingual gaps between non-Latin mentions and English Wikipedia titles. A further 23.5% of errors involve the model engaging with the correct entity in its reasoning but ultimately rejecting it, indicating that disambiguation remains an open challenge even for reasoning-native models. Across configurations (Appendix A.7), Think+Embed has the fewest total errors (841 vs. 1,493 for 8B-Inst+BM25) but the highest proportion of Completely Wrong cases. Retrieval and reasoning resolve the easier categories, thus concentrating residual errors on genuinely hard cases.

Error Category	N	%
Completely Wrong	436	51.8
Name Format	180	21.4
Disambiguation (Specificity)	112	13.3
Wikipedia Variant	55	6.5
Concept Granularity	45	5.4
Empty (Pipeline Error)	13	1.5
Total	841	100.0

Table 6: Error taxonomy for 8B-Think+Embed on the full MERLIN test set across all languages.

Split	Δ (pp)	p-value
RAG effect (Think+Embed vs Think)		
Lang. Ed.	+18.8	<0.001
Statements	+16.6	<0.001
WD Out-Links	+15.8	<0.001
Reasoning effect (Think vs Inst, no RAG)		
Lang. Ed.	-1.4	0.575
Statements	+0.9	0.766
WD Out-Links	+0.3	> 0.999

Table 7: Statistical significance (McNemar's and bootstrap tests). RAG effects are significant on all rare splits. Reasoning alone has no significant effect.

8 Conclusion

We showed that unpopularity is not the same as rarity. Entities with sparse Wikidata structure and few Wikipedia language editions cause accuracy drops of up to 53 percentage points, 15pp worse than what pageview-based analyses would suggest. This gap is consistent across the baselines we tested, indicating that prior work has underestimated the difficulty of the cultural long tail.

Our simple framework, in which a reasoning-capable VLM iteratively searches and reasons over Wikipedia, achieves +6.9pp over the state-of-the-art on MERLIN and up to +37pp on the hardest rare entity slices. The gains are not driven by either component alone: reasoning without retrieval shows no significant effect on rare entities, and retrieval without reasoning can hurt. The two are complementary, with the retrieval benefit growing 4.8× on structurally sparse entities. A 4B reasoning model with retrieval matches an 8B instruct model overall and outperforms it by +5 to +7pp on rare entities, suggesting an alternative to scaling.

Retrieval failure accounts for 72% of errors, making cross-lingual retrieval a useful direction for future work. Evaluation that handle Wikipedia redirects and valid aliases, and extension to languages with sparser coverage, would further advance culturally diverse entity linking.

9 Limitations

Our framework experiments use a single model family (Qwen3-VL (Team, 2025)), chosen because it uniquely supports controlled comparisons across reasoning variants and scales (Section 5). Our rarity characterization is validated across three architecturally distinct baselines (Section 4), but the specific findings about the interaction between reasoning and retrieval may not generalize to other VLM families. As other families develop matched reasoning variants, testing this generalization would strengthen these findings.

We evaluate on five languages from the MERLIN benchmark and target English Wikipedia as the knowledge base. All MERLIN entities have English articles by construction, but this constraint limits applicability to entities without English coverage. Generalization to languages with even sparser Wikipedia representation, such as African languages, remains unknown and would test the limits of retrieval-augmented approaches for the cultural long tail.

Retrieval is the dominant bottleneck in our pipeline, accounting for 72% of errors. Embedding retrieval partially addresses cross-lingual mismatch for non-Latin scripts, but neither retrieval method guarantees recall when transliterations diverge substantially between the source language and English Wikipedia titles. Improving cross-lingual retrieval in the context of LLM tool use remains the most impactful direction for future work.

We follow MERLIN's standard exact-match evaluation protocol, which penalizes predictions that produce valid but non-canonical title strings (e.g., redirects or common abbreviations) identically to genuinely incorrect predictions. This is consistent with all prior work on MERLIN and ensures comparability across systems.

10 Ethical Considerations

This work uses publicly available data: the MERLIN benchmark (Ramamoorthy et al., 2025), Wikipedia, and Wikidata. No personal data or human subjects are involved. Our system inherits biases present in Wikipedia's coverage, which is known to underrepresent non-Western entities and perspectives (Hecht and Gergle, 2010). While our rarity characterization helps surface these gaps, the system itself cannot correct underlying knowledge base biases. We note that entity linking systems, including ours, may perform less reliably on entities from marginalized communities that are systematically underrepresented in Wikipedia.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Marta Boscariol, Luana Bulla, Lia Draetta, Beatrice Fiumanò, Emanuele Lenzi, and Leonardo Piano. 2025. Evaluation of llms on long-tail entity linking in historical documents. *Preprint*, arXiv:2505.03473.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021a. Autoregressive entity retrieval. *Preprint*, arXiv:2010.00904.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021b. Multilingual autoregressive entity linking. *Preprint*, arXiv:2103.12528.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal llms with cultural knowledge. *Preprint*, arXiv:2508.07414.

Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. 2025. Entgpt: Entity linking with generative large language models. *Preprint*, arXiv:2402.06738.

Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024. ChatEL: Entity linking with chatbots. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3086–3097, Torino, Italia. ELRA and ICCL.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025. Retool: Reinforcement learning for strategic tool use in llms. *Preprint*, arXiv:2504.11536.

Arianna Graciotti, Nicolas Lazzari, Valentina Presutti, and Rocco Tripodi. 2025. Musical heritage historical entity linking. *Artificial Intelligence Review*, 58(5).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Samy Haffoudhi, Fabian M. Suchanek, and Nils Holzenberger. 2026. Lela: an llm-based entity linking approach with zero-shot domain adaptation. *Preprint*, arXiv:2601.05192.

Brent Hecht and Darren Gergle. 2010. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 291–300. ACM.

Mohanna Hoveyda, Arjen Vries, Faegheh Hasibi, and Maarten de Rijke. 2024. Real world conversational entity linking requires more than zero-shots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13938–13946, Bangkok, Thailand. Association for Computational Linguistics.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Yajie Li, Albert Galimov, Mitra Datta Ganapaneni, Pujitha Thejaswi, De Meng, Priyanshu Kumar, and Saloni Potdar. 2025. Leveraging the power of large language models in entity linking via adaptive routing and targeted reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 871–882, Suzhou (China). Association for Computational Linguistics.

Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, and Jens Lehmann.

2023. mReFinED: An efficient end-to-end multilingual entity linking system. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15080–15089, Singapore. Association for Computational Linguistics.

Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. Onenet: A fine-tuning free framework for few-shot entity linking via large language model prompting. *Preprint*, arXiv:2410.07549.

Ziyan Liu, Junwen Li, Kaiwen Li, Tong Ruan, Chao Wang, Xinyan He, Zongyu Wang, Xuezhi Cao, and Jingping Liu. 2025. I2cr: Intra- and inter-modal collaborative reflections for multimodal entity linking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 4942–4951. ACM.

Yajie Luo, Yihong Wu, Muzhi Li, Fengran Mo, Jia Ao Sun, Xinyu Wang, Liheng Ma, Yingxue Zhang, and Jian-Yun Nie. 2025. An entity linking agent for question answering. *Preprint*, arXiv:2508.03865.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

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, Toronto, Canada. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Gerard Pons, Besim Bilalli, and Anna Queralt. 2024. *Knowledge Graphs for Enhancing Large Language Models in Entity Disambiguation*, page 162–179. Springer Nature Switzerland.

Sathyanarayanan Ramamoorthy, Vishwa Shah, Simran Khanuja, Zaid Sheikh, Shan Jie, Ann Chia, Shearman Chua, and Graham Neubig. 2025. Merlin: A testbed for multilingual multimodal entity recognition and linking. *Preprint*, arXiv:2510.14307.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7654–7665, Torino, Italia. ELRA and ICCL.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models. *Preprint*, arXiv:2504.10191.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2025. Llmael: Large language models are good context augmenters for entity linking. *Preprint*, arXiv:2407.04020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. Sglang: Efficient execution of structured language model programs. *Preprint*, arXiv:2312.07104.

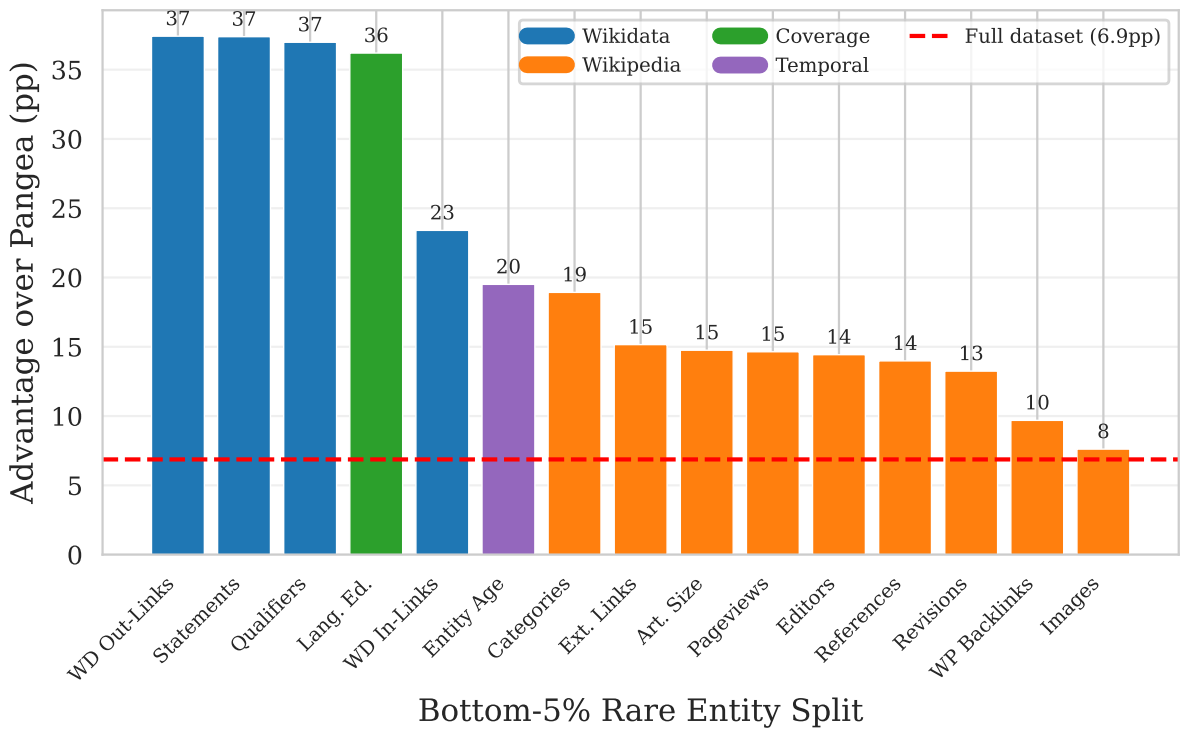

Figure 6: Advantage (pp) of 8B-Think+Embed over Cultural Pangea across all 15 bottom-5% rare entity slices, sorted by magnitude.

A Appendix

A.1 Rarity Metric Independence

To verify that the 15 rarity metrics identify genuinely different entities as "rare" rather than measuring a single underlying construct, we compute pairwise Jaccard overlap between the bottom-5% entity sets for each metric (Figure 7). The mean pairwise overlap is only 37%, confirming that different metrics flag largely non-overlapping entity sets. Within-family overlap is moderate (Wikipedia: 48%, Wikidata: 38%), while between-family overlap is lower still (30%), with some cross-family pairs sharing as few as 10% of entities.

Qualitative analysis reinforces this finding. We identify 50 entities that are Wikidata-rare (bottom-5% on $\geq$3 Wikidata metrics) but Wikipedia-normal (zero Wikipedia metrics flagged), including well-documented news events like the *2016 Indian banknote demonetisation* (36K pageviews, 980 editors, but only 11 Wikidata statements) and the *Pittsburgh synagogue shooting* (88K pageviews but structurally sparse in Wikidata). Conversely, 92 entities are Wikipedia-rare but Wikidata-normal, such as the *Muttahida Qaumi Movement* (minimal English Wikipedia article but 287 Wikidata incoming links). These examples illustrate that editorial attention and structural connectivity are different dimensions of rarity.

A.2 Baseline Degradation Across All Systems

Eval. Slice	GEMEL	mGENRE	Pangea
Full Test Set	59.3	72.9	81.1
Wikidata structural (Bottom 5%)			
Lang. Editions	31.5	40.9	27.7
Statements	30.2	42.9	28.8
WD Out-Links	30.2	42.5	28.8
Qualifiers	30.6	45.0	31.4
WD In-Links	37.0	45.3	45.6
Wikipedia engagement (Bottom 5%)			
Pageviews	28.2	42.7	42.8
Categories	23.8	44.6	40.6

Table 8: Baseline accuracy (%) on the full test set and bottom-5% rare entity slices. All three baselines degrade severely on rare entities, with drops of 28–52 pp on Wikidata-structural slices. The pattern is strongest for Pangea (52 pp structural vs. 39 pp popularity drop) and mGENRE (30 pp vs. 29 pp).

A.3 Full Rare Entity Slice Breakdown

Figure 6 presents the advantage of 8B-Think+Embed over Cultural Pangea across all 15 bottom-5% rare entity slices. Wikidata-structural metrics consistently show the largest gains, while Wikipedia-engagement metrics show

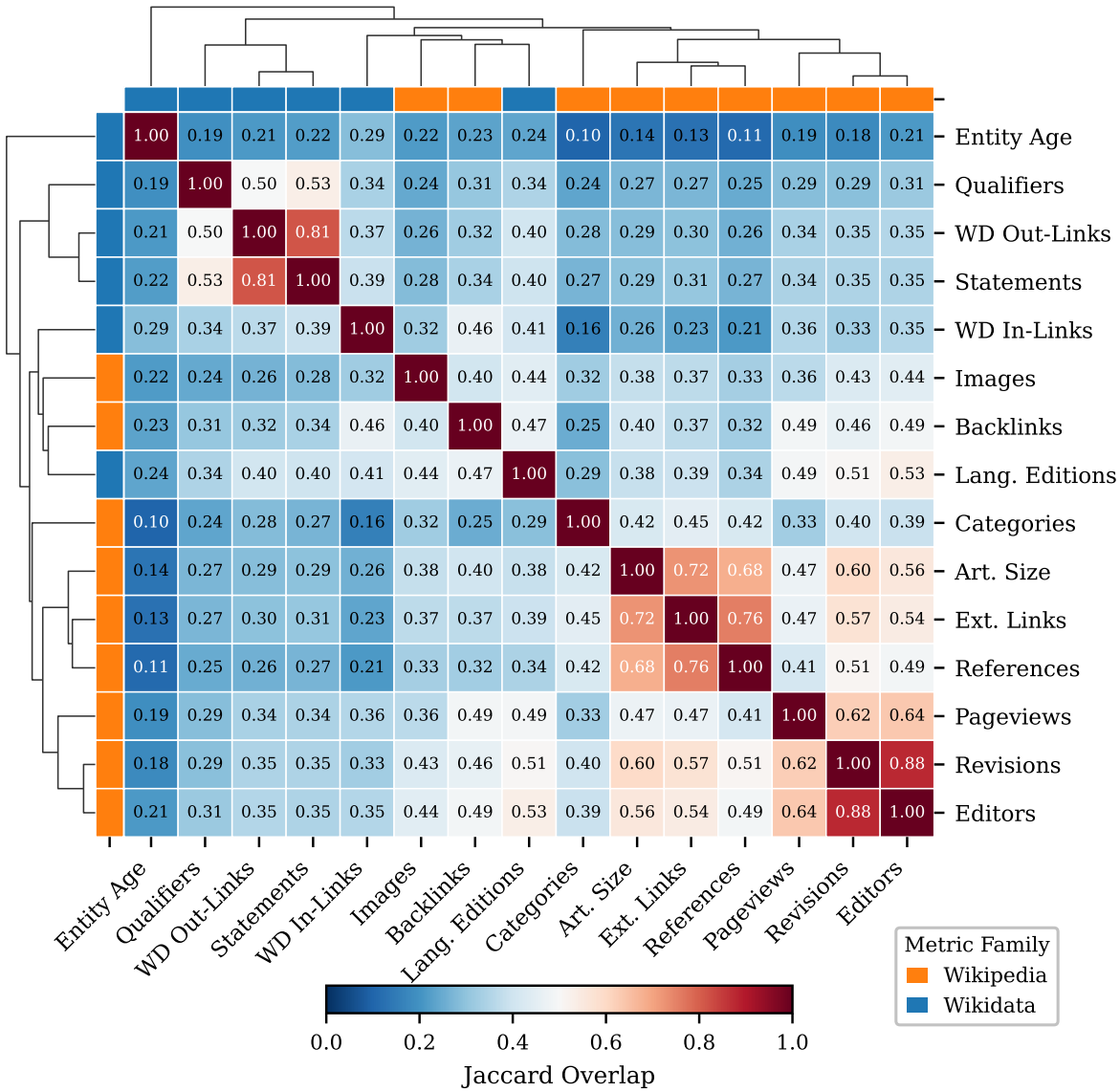

Figure 7: Pairwise Jaccard overlap between bottom-5% rare entity sets for each of the 15 rarity metrics. Mean overlap is 37%, confirming that different metrics identify substantially different entities as rare. Sidebar color indicates metric family (orange = Wikipedia, blue = Wikidata).

more modest improvements.

A.4 RAG Effect by Model Size

Table 9 shows the Think-vs-Instruct gap broken down by model size and retrieval method. The gap is largest with embedding retrieval, consistent across scales.

Size	No RAG	BM25	Embed
2B	−0.8	+6.8	+4.8
4B	+3.4	+8.0	+7.6
8B	+0.7	+7.3	+5.0

Table 9: Think − Instruct gap (pp, average across languages) by model size and retrieval method. Reasoning models benefit more from retrieval at every scale.

A.5 Robustness Analysis

We define the *robustness gap* as the difference between Pangea's accuracy drop and our system's accuracy drop on each rare slice. A positive robustness gap means our system degrades less.

Table 10 presents the full robustness analysis. Our system shows a robustness gap of up to 30.5pp on Wikidata-structural metrics (Statements, WD Out-Links), meaning Pangea's accuracy drops 30pp more than ours on these slices.

A.6 Threshold Robustness

Table 11 shows that our system's advantage over Pangea is stable across 1%, 5%, and 10% rarity

Metric	Ours	Pangea	Gap
Statements	−21.8	−52.3	30.5
WD Out-Links	−21.7	−52.2	30.5
Qualifiers	−19.5	−49.6	30.1
Lang. Editions	−24.0	−53.4	29.3
WD In-Links	−18.9	−35.4	16.5
Entity Age	−16.8	−29.4	12.6
Categories	−28.4	−40.5	12.1
Ext. Links	−29.6	−37.9	8.3
Art. Size	−26.9	−34.8	7.9
Pageviews	−30.5	−38.2	7.8
Editors	−29.7	−37.2	7.6
References	−27.7	−34.8	7.1
Revisions	−30.7	−37.1	6.4
WP Backlinks	−28.7	−31.5	2.8
Images	−26.5	−27.3	0.7

Table 10: Accuracy drop (pp) from full dataset to bottom-5% slices. "Gap" = Pangea drop − our drop. Sorted by robustness gap.

Metric	Thr.	n	Ours	Pangea	Gap
Lang. Ed.	1%	76	60.5	31.6	28.9
	5%	357	65.8	31.1	34.7
	10%	708	69.9	32.5	37.4
Statements	1%	77	63.6	20.8	42.8
	5%	377	67.1	29.4	37.7
	10%	704	70.2	29.8	40.4
WD Out-Links	1%	74	70.3	29.7	40.6
	5%	362	67.1	29.3	37.8
	10%	727	70.7	30.1	40.6
Qualifiers	1%	170	65.3	27.6	37.7
	5%	375	70.7	30.4	40.3
	10%	715	72.3	32.3	40.0
Pageviews	1%	68	55.9	22.1	33.8
	5%	348	58.3	25.0	33.3
	10%	700	65.9	30.0	35.9

Table 11: Accuracy (%) at 1%, 5%, and 10% rarity thresholds for the top 5 metrics. The advantage over Pangea (Gap) is consistent (28–43pp) across all thresholds, confirming robustness to threshold choice.

thresholds. The gap ranges from 28–43pp across all metrics and thresholds, confirming that results are not sensitive to the choice of cutoff.

A.7 Error Taxonomy Shift on Rare Entities

Table 12 presents the full error taxonomy across all configurations.

A.8 Pipeline Decomposition

Table 13 decomposes errors by pipeline stage for RAG-based configurations. Retrieval failures dominate across all configurations, but 8B-Think+Embed shows the lowest retrieval failure rate (72.1%) due to embedding retrieval's better cross-lingual recall. BM25-based configurations (8B-Think+BM25, 8B-Inst+BM25) have retrieval failure rates of 89.5% and 80.9% respectively.

A.9 Reasoning Compensates for Size

Figure 8 compares 4B-Think+Embed against 8B-Inst across the full dataset and rare entity slices. On the full MERLIN test set, the two configurations are nearly identical (83.7% vs. 83.5%, a gap of just +0.3pp), despite the reasoning model having half the parameters. However, the systems diverge sharply on rare entities: 4B-Think+Embed outperforms 8B-Inst by +5.9pp on language editions, +6.9pp on statements, and +7.1pp on Wikidata outlinks. This suggests that reasoning with retrieval is a more cost-effective strategy than scaling model size alone, particularly for the long tail where parametric knowledge is insufficient and effective use of retrieved evidence becomes the dominant factor.

A.10 Prompt Templates

We use the same system prompt across all configurations, varying only the retrieval tool availability. The model receives the article text, image, and marked entity mention, and is instructed to output the English Wikipedia title of the referenced entity. The full prompt will be made available with our released code upon acceptance.

System prompt (with retrieval). The model is instructed that it has access to a Wikipedia search tool. It may issue search queries by performing a tool call, after which the top-k retrieved titles and descriptions are injected into the context. The model may search up to 20 times per example. The prompt guides a six-step methodology: lock in the entity mention, analyze context for disambiguation, evaluate the image, translate or transliterate for search, search strategically with iterative refinement, and verify the answer. After completing its reasoning, the full trace is passed to a second extraction prompt that instructs the model to output only the final answer. This two-pass design decouples open-ended deliberation from structured answer extraction.

The model also receives an OpenAI-compatible tool definition for `search_wikipedia(query, limit)` that instructs it to translate or transliterate entity names to English for best results.

System prompt (without retrieval). Identical to the above, but Step 5 is replaced with "Using your internal knowledge, identify the correct English Wikipedia page title" and the search tool is not provided.

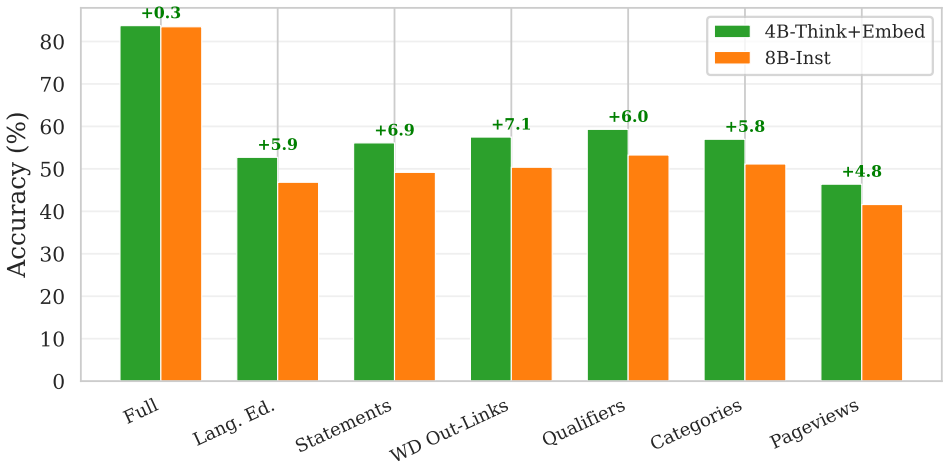

Figure 8: 4B-Think+Embed vs. 8B-Inst on the full dataset and rare entity slices. The two systems are nearly identical on the full dataset but diverge on rare entities, with the smaller reasoning model winning by +5–7pp.

Error Category	8B-Th+Em	8B-Th+BM	8B-Think	8B-Inst	8B-In+BM
Completely Wrong	436 (51.8%)	501 (50.8%)	448 (40.5%)	508 (44.0%)	560 (37.5%)
Name Format	180 (21.4%)	221 (22.4%)	333 (30.1%)	326 (28.2%)	362 (24.2%)
Disambiguation	112 (13.3%)	134 (13.6%)	121 (10.9%)	119 (10.3%)	349 (23.4%)
Wikipedia Variant	55 (6.5%)	73 (7.4%)	130 (11.8%)	121 (10.5%)	61 (4.1%)
Concept Granularity	45 (5.4%)	44 (4.5%)	61 (5.5%)	68 (5.9%)	48 (3.2%)
Empty (Pipeline Error)	13 (1.5%)	13 (1.3%)	13 (1.2%)	13 (1.1%)	113 (7.6%)
Total	841	986	1106	1155	1493

Table 12: Full error taxonomy across 8B configurations. Th+Em has the fewest total errors (841) but the highest *proportion* of "Completely Wrong" (51.8%), because its easier errors (Name Format, Wikipedia Variant) are resolved by retrieval, leaving harder cases.

Stage	Th+Em	Th+BM	In+BM
Retrieval Failure	72.1%	89.5%	80.9%
Selection Failure	2.9%	1.4%	3.7%
Extraction Failure	23.5%	7.8%	7.8%
Empty	1.5%	1.3%	7.6%

Table 13: Pipeline decomposition of errors for RAG-based 8B configurations.

A.11 Search Behavior Analysis

We quantify how Thinking and Instruct models differ in their use of the Wikipedia search tool across all 12 RAG configurations (3 sizes $\times$ 2 variants $\times$ 2 retrieval methods).

Search count and deliberation. Table 14 shows that Instruct models make substantially more searches per example (3.2–5.2 average) than Thinking models (1.0–2.2), yet achieve lower accuracy. Thinking models generate 1,400–3,100 completion tokens between consecutive searches, using this reasoning to analyze retrieved results and decide whether and what to search next. Instruct models generate only 26–30 tokens between searches (bare tool-call overhead) indicating they issue searches with no analysis of previous results. The behavioral contrast is strategic precision vs. brute-force volume. Think models achieve higher accuracy with fewer, more deliberate searches.

Retrieval success. Figure 9 shows per-search and cumulative retrieval success for 8B models. Embedding retrieval finds the target entity in the first search 25–40% of the time, versus 6–15% for BM25. The per-search hit rate peaks at search 2 (the primary refinement step) and drops sharply thereafter, showing diminishing returns from additional searches. The cumulative panel shows that Instruct models end up finding the target in more examples overall, but only because they make more searches, not because any individual search is more effective. Despite this higher cumulative retrieval rate, Instruct models achieve lower accuracy than Think models, indicating that finding the target in results is not the bottleneck, reasoning about the results is.

Model	RAG	Avg	Med	%1	%5+	%20	%Q1
2B-Instr	BM25	5.2	4	0	31	8	91
2B-Instr	Embed	3.3	3	1	8	0	98
2B-Think	BM25	1.0	1	91	0	0	98
2B-Think	Embed	1.0	1	97	0	0	100
4B-Instr	BM25	3.5	3	0	18	1	83
4B-Instr	Embed	3.0	3	0	7	0	88
4B-Think	BM25	1.2	1	80	0	0	92
4B-Think	Embed	1.1	1	93	0	0	97
8B-Instr	BM25	3.6	3	0	12	1	95
8B-Instr	Embed	3.2	3	0	3	0	98
8B-Think	BM25	2.2	2	29	0	0	74
8B-Think	Embed	1.5	1	68	0	0	88

Table 14: Search behavior summary across all 12 RAG configurations. "Avg/Med": searches per example. "%1/%5+/%20": fraction with 1 / 5+ / 20 (max) searches. "%Q1": fraction of searches in first quarter of output.

Query degradation over successive searches. Figure 10 reveals that Instruct models' queries degrade over successive iterations. For all models, the first search is a short entity name (∼10 characters) and the second doubles in length by adding disambiguation context. After this, Thinking models stop (68% of 8B-Think+Embed examples use exactly one search). Instruct models continue with progressively worsening queries: Jaccard diversity drops from 0.63 at transition 1→2 to 0.24 at 19→20, verbatim repetition climbs from 0.3% to 34%, and query length bloats from 9.9 to 45.1 characters as the model appends context words. This indicates Instruct models lack an effective stopping criterion.

A.12 Computational Cost Analysis

Table 15 reports the computational cost per example across all 18 configurations, including the 6 no-RAG baselines. All experiments were run on a single NVIDIA L40S GPU (48 GB VRAM) per configuration, with 10 CPU cores and 25 GB system RAM. Models were served with SGLang (Zheng et al., 2024).

Accuracy–cost tradeoff. Figure 12 shows the Pareto frontier of accuracy vs. token cost. The frontier runs from 4B-Instr (76.2%, 2.8k tokens) through 8B-Think (84.2%, 5.0k tokens) to 8B-Think+Embed (87.9%, 11.1k tokens). Our best configuration costs 3.9× the cheapest (4B-Instr), gaining +11.7pp accuracy. The no-RAG 8B-Think baseline (84.2%, 5.0k tokens) represents a strong low-cost option: 96% of the best system's accuracy at 45% of its token cost.

A.13 Search Query Analysis

Query length. Table 17 shows mean query length across configurations. All models produce queries of similar average length (14–21 characters), suggesting that query length is not a differentiating factor in the Think–Instruct gap.

Query transition patterns. Table 18 categorizes each consecutive query pair by transition type. The dominant patterns are *refinement* (adding disambiguation context to the previous query, 30–55% of transitions) and *variation* (moderate change, 25–50%). Verbatim *repetition* is rare in early searches but increases for Instruct models in long search chains (see Figure 10). *Pivot* transitions (completely different query) occur 5–15% of the time, typically when initial approaches fail.

A.14 Licenses for Artifacts Used and Released

Models. We use Qwen3-VL (Team, 2025) under the Apache License 2.0, and the `multilingual-e5-large-instruct` embedding model (Wang et al., 2024) under the MIT License. The CulturalPangea-7B baseline (de Dieu Nyandwi et al., 2025) is used under the Apache License 2.0. The mGENRE baseline (Cao et al., 2021b) is used under the MIT License, and the GEMEL baseline (Shi et al., 2024) is used under the terms of its original release.

Software. We use FAISS (Douze et al., 2024) under the MIT License, BM25S (Lù, 2024) under the MIT License, and SGLang (Zheng et al., 2024) under the Apache License 2.0.

Data. We use the MERLIN benchmark (Ramamoorthy et al., 2025) for evaluation in accordance with the terms set by its authors. English

Model	RAG	Acc%	M1 In	M1 Out	M2	Total	Time(s)	Ratio
2B-Instr	—	63.5	1485	2856	3056	7397	81.9	2.6×
2B-Instr	BM25	53.6	14961	4867	3699	23527	167.8	8.4×
2B-Instr	Embed	54.3	10622	4048	3210	17880	99.1	6.4×
2B-Think	—	62.7	1487	2984	1286	5757	53.3	2.0×
2B-Think	BM25	60.4	3804	8514	1513	13832	167.9	4.9×
2B-Think	Embed	59.1	3799	7396	1507	12702	113.4	4.5×
4B-Instr	—	76.2	1485	565	765	2815	17.6	1.0×
4B-Instr	BM25	72.7	10145	1458	367	11970	66.2	4.3×
4B-Instr	Embed	76.1	8906	1482	391	10780	51.4	3.8×
4B-Think	—	79.6	1487	2272	1359	5119	62.2	1.8×
4B-Think	BM25	80.7	4322	5150	1327	10799	166.1	3.8×
4B-Think	Embed	83.7	4022	4495	1320	9837	124.6	3.5×
8B-Instr	—	83.5	1485	895	1095	3475	50.7	1.2×
8B-Instr	BM25	78.6	11500	502	600	12602	24.6	4.5×
8B-Instr	Embed	82.9	9851	419	536	10806	19.7	3.8×
8B-Think	—	84.2	1487	2027	1505	5019	81.6	1.8×
8B-Think	BM25	85.8	6658	5566	1486	13709	247.1	4.9×
8B-Think	Embed	87.9	5137	4426	1489	11052	176.9	3.9×

Table 15: Computational cost per example across all 18 configurations. M1 In/Out: Module 1 input/output tokens. M2: Module 2 total tokens. Ratio: number of tokens (cost) relative to cheapest configuration (the most token-efficient).

Wikipedia article content is used under the CC BY-SA 4.0 License, and Wikidata structural metadata is used under the CC0 1.0 Public Domain Dedication.

Released artifacts. MERLIN-Rare, the rare-entity test slices we will release upon acceptance, consists of subsets of the existing MERLIN benchmark identified by structural-rarity metrics computed from Wikidata; we will release MERLIN-Rare under CC BY-SA 4.0. Our code will be released under the MIT License.

All artifacts are used consistently with their intended research purposes.

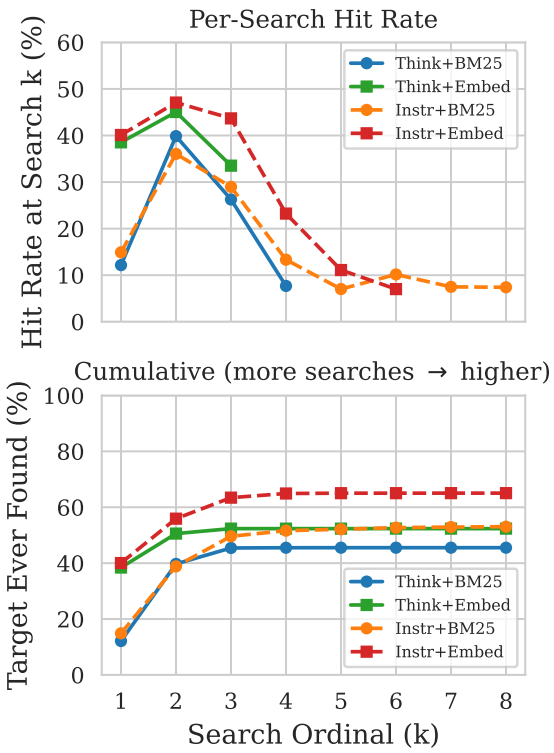

Figure 9: Top: per-search hit rate (fraction of examples where search k returns the target). Search 2 is most productive; later searches have sharply diminishing returns. Bottom: cumulative. Instruct (dashed) ends higher only because it makes more total searches, not because any individual search is better. 8B models shown.

Model	RAG	Tok (F)	Tok (R)	R/F	Dur (F)	Dur (R)
2B-Instr	—	7397	11013	1.49	81.9	133.9
2B-Instr	BM25	23527	27346	1.16	167.8	178.8
2B-Instr	Embed	17880	19523	1.09	99.1	115.8
2B-Think	—	5757	6968	1.21	53.3	71.6
2B-Think	BM25	13832	15728	1.14	167.9	201.9
2B-Think	Embed	12702	14958	1.18	113.4	144.6
4B-Instr	—	2815	3198	1.14	17.6	23.6
4B-Instr	BM25	11970	13451	1.12	66.2	64.8
4B-Instr	Embed	10780	11966	1.11	51.4	79.5
4B-Think	—	5119	6067	1.19	62.2	82.8
4B-Think	BM25	10799	14011	1.30	166.1	250.8
4B-Think	Embed	9837	12700	1.29	124.6	187.5
8B-Instr	—	3475	4704	1.35	50.7	80.3
8B-Instr	BM25	12602	16947	1.34	24.6	37.2
8B-Instr	Embed	10806	11351	1.05	19.7	23.1
8B-Think	—	5019	6234	1.24	81.6	120.3
8B-Think	BM25	13709	15907	1.16	247.1	344.9
8B-Think	Embed	11052	13890	1.26	176.9	260.7

Table 16: Cost on full dataset (F) vs. rare entities (R, bottom-5% popularity). R/F: token ratio.

Model	RAG	Mean Chars	Mean Words
2B-Instr	BM25	18.9	3.0
2B-Instr	Embed	14.6	2.4
2B-Think	BM25	7.8	1.3
2B-Think	Embed	7.8	1.3
4B-Instr	BM25	16.2	2.4
4B-Instr	Embed	15.9	2.4
4B-Think	BM25	10.3	1.6
4B-Think	Embed	9.5	1.5
8B-Instr	BM25	18.5	2.8
8B-Instr	Embed	17.5	2.6
8B-Think	BM25	14.5	2.2
8B-Think	Embed	12.8	2.0

Table 17: Mean search query length across configurations.

Model	RAG	Refine	Variation	Pivot	Repeat	Simplify
2B-Instr	BM25	21	12	18	46	3
2B-Instr	Embed	34	15	19	28	3
2B-Think	BM25	2	1	96	0	0
2B-Think	Embed	6	0	81	12	0
4B-Instr	BM25	38	38	14	7	3
4B-Instr	Embed	46	37	14	2	2
4B-Think	BM25	62	12	21	2	3
4B-Think	Embed	60	9	24	2	6
8B-Instr	BM25	35	41	19	3	3
8B-Instr	Embed	40	38	19	1	2
8B-Think	BM25	56	19	13	9	3
8B-Think	Embed	58	17	16	6	3

Table 18: Query transition patterns (% of all consecutive query pairs). *Refine*: previous query is substring of current (context added). *Variation*: moderate change. *Pivot*: completely different query (Jaccard > 0.8). *Repeat*: verbatim copy. *Simplify*: current is substring of previous.

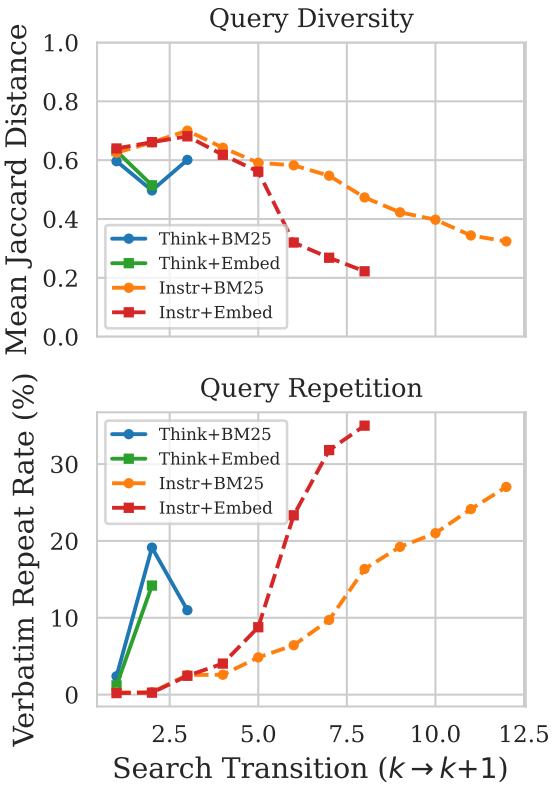

Figure 10: Query diversity (left) and verbatim repetition rate (right) across successive search transitions for 8B models. Instruct+BM25 (orange dashed) shows clear degradation: diversity collapses while repetition climbs. Think models (solid) stop after 2–3 transitions with stable quality.

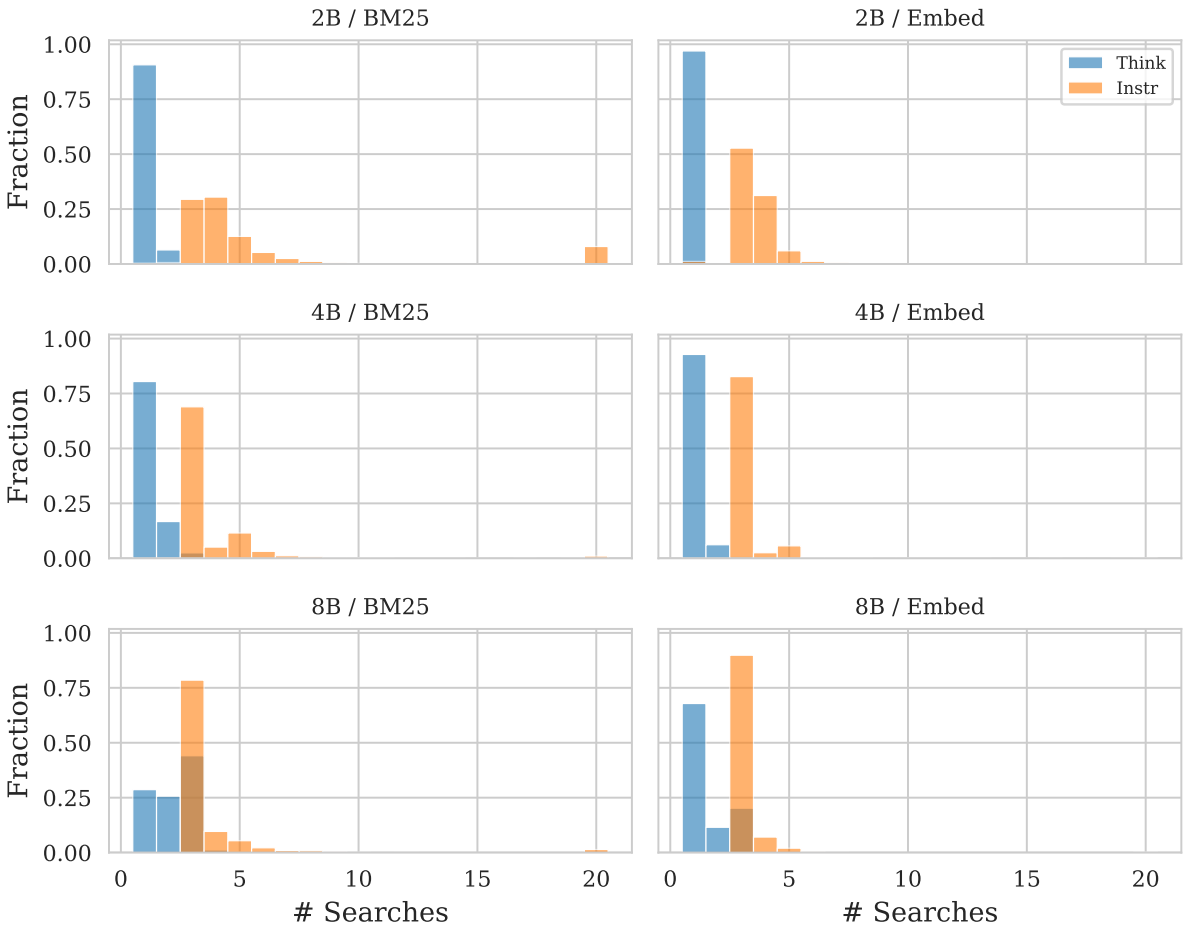

Figure 11: Search count distributions across all 12 RAG configurations. Think (blue) concentrates at 1–2 searches; Instruct (orange) spreads across 3–5+. At 2B, Think barely uses search at all.

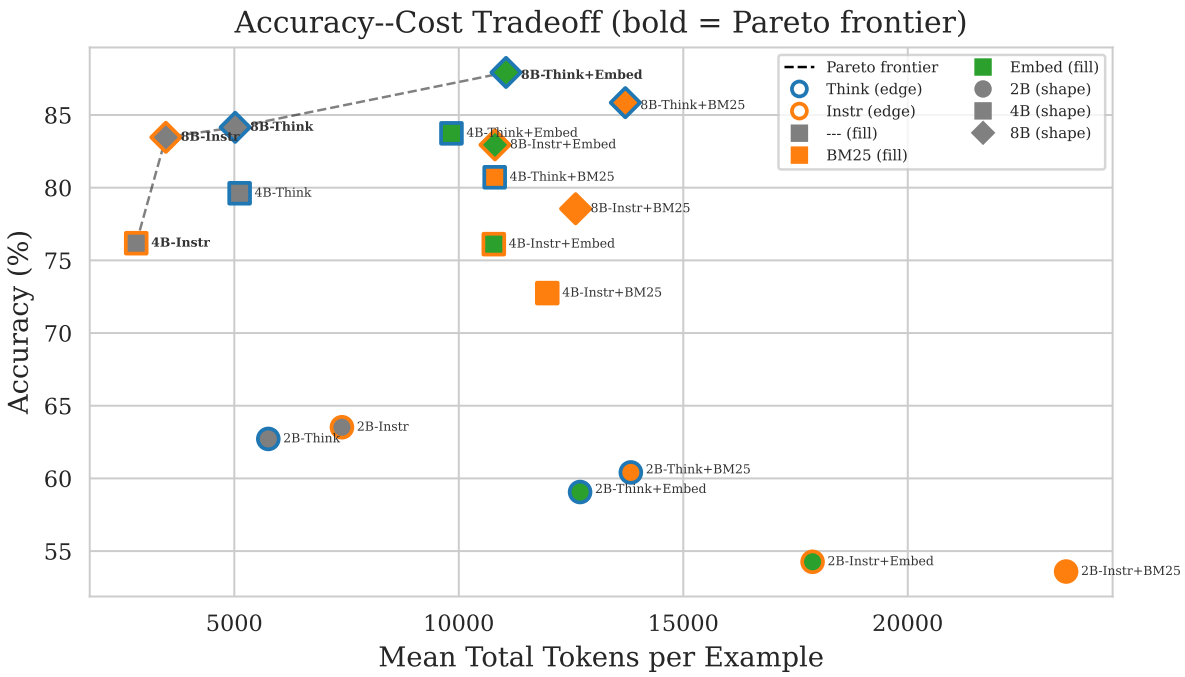

Figure 12: Accuracy vs. average token cost for all 18 configurations. Bold labels mark the Pareto frontier. Shape encodes model size, edge color encodes Think/Instruct, fill color encodes RAG method.