

Search in Whose Language? Separating the Query-Side Gap from the Evidence-Side Gap for Search Agents in Indic Languages

Mohammed Abraar

Vizuara Research
abraar@vizz.vizuara.ai

Abstract

A search agent answers a Hindi or Telugu question less accurately than the same question in English. The cause may be the agent’s queries (the *query-side gap*) or missing evidence in the user’s language (the *evidence-side gap*), and each needs a different fix. We separate them with human-parallel Hindi/English and Telugu/English questions whose answer is in both Wikipedias (287 Hindi, 127 Telugu), a frozen pool of 1.9 million Wikipedia passages served as English-only, Indic-only or both, and a logged, then forced, query language. Four models (three Gemini, one open-weight Qwen3) run the grid under a dense cross-lingual retriever and under BM25. With dense retrieval, the cheapest models lose 9 points on Hindi questions and the stronger Gemini models lose 2–3 (not significant). Changing the query language or pool recovers almost nothing, and with the gold passage in the prompt Hindi is still 4–6 points worse: the residual gap is reading. Under BM25 the Hindi gap is 19–20 points, and dual-language queries recover 9–14. The Hindi and Telugu Wikipedias hold the answer for only 29% and 24% of questions; on the uncovered questions a Hindi user reaches 51% against 60% for an English user.

1 Introduction

Hundreds of millions of people in India now reach the web through chat products backed by search agents: a language model that writes queries, reads the results and answers. When the question is in Hindi, Telugu or Bengali instead of English, the answer is wrong more often. Benchmarks for browsing and deep-research agents report this drop across many languages (Hofman et al. 2025; Caciolai et al. 2026; Lu et al. 2026; Aji et al. 2026). They do not report *why* it happens, and the cause decides the fix.

Two causes produce the same drop. The first is the agent. It may write a worse query in Hindi than in English, or search in a language where the evidence is not, for example an English query for a fact that only Hindi Wikipedia records. We call this the **query-side gap**; prompting and training can fix it. The second is the web. Hindi Wikipedia has 2.4% as many articles as English Wikipedia, Telugu has 1.8%, and Hindi makes up 0.2% of Common Crawl pages. If a fact is not written in the user’s language, no query in that language will find it. We call this the **evidence-side gap**; only cross-lingual

search or new content can close it. A third gap remains when the model has the right passage and still answers worse in Hindi; we call it the **integration gap**.

Prior work has measured parts of this problem but not separated the three gaps for a search agent. XBCP translates an English deep-research corpus into twelve languages and measures how much oracle retrieval recovers. Its questions are always in English, the agent’s query language is never logged or controlled, and no Indic language is included (Lu et al. 2026). MarathiWeb studies two web-search models on native Marathi queries and attributes failures to integration by inspection (Kale 2026). DELTA shows that an apparent preference for English evidence in single-shot RAG partly reflects where the gold evidence lives (Park et al. 2026). Where’s Waldo shows that, with fixed context, models prefer the query’s language when a fact exists in only one Wikipedia edition (Ki et al. 2026). Section 2 covers the wider literature.

We build a controlled setup that measures all three gaps (Figure 1). We fix the question with parallel Indic/English pairs whose evidence exists in both languages. We fix the evidence with a frozen two-language Wikipedia pool that we serve as English-only, Indic-only or both. We log the script of every query under a free policy and then force the query language. We run the grid under a dense cross-lingual retriever and under BM25, because the retriever’s language behaviour is part of the query-side gap. Finally, we measure how often the Hindi and Telugu Wikipedias contain each answer, and use this coverage to estimate the accuracy a real user would see.

Our contributions are:

1. **A controlled setup** that treats the agent’s query language as a measured and then manipulated factor, on parallel questions over a frozen pool where the evidence language is served or withheld.
2. **The retriever as a factor.** With a dense cross-lingual retriever the query-side gap is at most a few points. With BM25 it is 19–20 points for Hindi, and dual-language queries recover 9–14 points.
3. **A measured coverage term** for the Hindi and Telugu Wikipedias (29% and 24%). We also measure the uncovered questions directly instead of extrapolating to them.
4. **Evidence that the residual gap is reading.** With the gold passage in the prompt, the cheaper models still answer

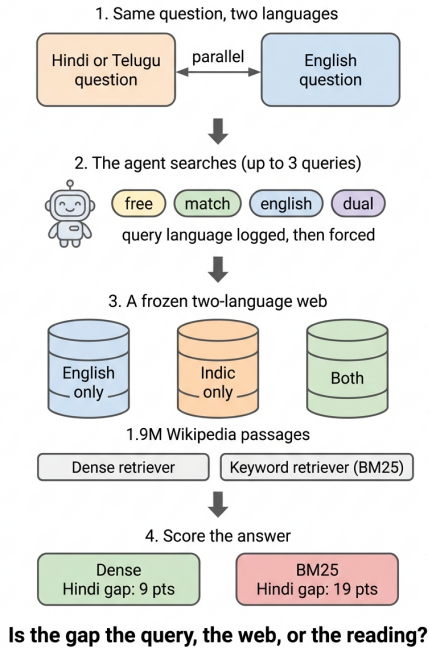


Figure 1: Study design. The same question is asked in an Indic language and in English. The agent searches a frozen Wikipedia pool served as English-only, Indic-only or both, under a free or forced query-language policy, with a dense or keyword retriever. The gaps shown are for Gemini 2.5 Flash-Lite, Hindi question, free policy, both pools.

Hindi questions 4–6 points worse; the stronger Gemini models do not.

2 Related Work

Agentic search beyond English. Browsing benchmarks exist for Chinese (Zhou et al. 2025), Korean (Lee et al. 2026; Jeong 2026), and thirteen natively authored languages without an Indic one (Aji et al. 2026). BrowseComp-Plus fixed the corpus so that retrievers can be swapped (Chen et al. 2025), and XBCP translated it into twelve languages (Lu et al. 2026). Multilingual GAIA variants score 47.4% in English against 35.8% in other languages and place Hindi among the weakest at every model size (Hofman et al. 2025; Caciolai et al. 2026); part of that gap may be translation artefact (Kim, Uhlig, and Wuebker 2026). English-only studies split agent failures into retrieval and use of the retrieved text (Liu et al. 2026b; Liang et al. 2026; Liu et al. 2026a). The MAST shared task runs Indic queries over an English-only corpus (Thakur et al. 2026), so the evidence-side gap is zero by design. None of these logs or forces the agent’s query language for an Indic user, or varies where the evidence lives.

Language preference in multilingual retrieval. Retrievers and generators favour high-resource and query-language passages (Park and Lee 2025; Wang et al. 2026; Jang et al. 2026; Ki et al. 2025), and models write tool arguments in the wrong language, Hindi included (Luo et al. 2026). DELTA

corrects for where gold evidence exists statistically (Park et al. 2026); we control it experimentally. Studies disagree on whether translating the query helps (Ranaldi, Haddow, and Birch 2025; Moon, Kim, and Verma 2025; Gupta et al. 2025); the sign depends on where the documents are (Amiraz et al. 2025). XRAG ran a controlled query-language ablation on a live engine (Liu et al. 2025), and CORAL lets an agent pick the corpus language but reports only end accuracy (Lee, Song, and Kang 2026). Dense multilingual retrievers match across languages where lexical matching cannot (Goworek, Macmillan-Scott, and Özyiğit 2025; Chen et al. 2024), which is why we treat the retriever as a factor.

Indic resources and knowledge gaps. Parallel Indic questions come from IndicGenBench (Singh et al. 2024), built on TyDi XOR (Asai et al. 2020; Clark et al. 2020), and from MLQA (Lewis et al. 2019). India-centric closed-book sets include IndicQuest, MILU and MultiLoKo (Jain et al. 2026; Rohera et al. 2025; Verma et al. 2024; Hupkes and Bogoychev 2025); Indic retrieval resources include MIRACL, IndicIRSuite and Hindi-BEIR (Zhang et al. 2022; Haq, Sharma, and Bhattacharyya 2023; Acharya et al. 2024). Factual recall depends on language and script (Jiang et al. 2020; Chua et al. 2024; Piratla et al. 2025), and India-centric facts are answered better in English (Rohera et al. 2025; Mazumder et al. 2026). ECLeKTic and Where’s Waldo build questions whose evidence exists in one Wikipedia edition only (Goldman et al. 2025; Ki et al. 2026), the closed-book analogue of our evidence manipulation. Wikipedia’s coverage asymmetry has been documented since Hecht and Gergle (2010). The supplementary material compares our design with each close neighbour.

3 Experimental Setup

3.1 Questions

XorQA-In (Singh et al. 2024) provides 539 test questions per language, human-translated from English TyDi XOR questions (Asai et al. 2020), each with the English paragraph that answers it. For each item we found the Hindi or Telugu counterpart of the English source article through Wikidata sitelinks and checked whether it states the answer. We used string match, then two independent LLM reads that had to quote a sentence found verbatim in the article, and a 50-item manual audit (supplementary material). An item enters the controlled grid only when both reads agree: 157 Hindi and 127 Telugu items. **MLQA** (Lewis et al. 2019) adds 130 self-contained Hindi questions with naturally parallel Hindi and English paragraphs, giving 287 Hindi items in total. A further 120 natively written **TyDi XOR** Telugu questions test whether translation drives the results (Kim, Uhlig, and Wuebker 2026).

3.2 Evidence pools and retrievers

All pools come from the October 2026 wikipedia-monthly snapshot (Kamali 2026): the complete Hindi Wikipedia (353,075 chunks), the complete Telugu Wikipedia (518,524 chunks), and a 1.8% sample of English Wikipedia (1,058,175 chunks) plus the English article behind every item. Chunks are about 900 characters with

Model	Access	Hi	Te
Gemini 2.5 Flash-Lite	API	287	127
Gemini 3.1 Flash-Lite	API	100	100
Gemini 3.6 Flash	API	50	50
Qwen3-30B-A3B	open weights	287	127

Table 1: Models and number of items per language. Gemini 2.5 Flash-Lite and Qwen3 run the full grid under both retrievers plus all arms; Gemini 3.1 Flash-Lite runs the dense grid; Gemini 3.6 Flash runs the headline cells.

the title prefixed. The agent’s one tool, `search(query)`, returns the top five chunks from the pool being served: English-only, Indic-only, or both. We use two retrievers: BM25 (Robertson and Zaragoza 2009) over Unicode word tokens, and bge-m3 dense embeddings (Chen et al. 2024) with exact inner-product search. With the question itself as the query, BM25 put the gold chunk in the top five for 50% of English questions but only 8% of Hindi and 7% of Telugu ones, because two Devanagari spellings of the same name (“Oregan” and “Oregon” when transliterated) do not match as tokens. bge-m3 reached 70% for both Indic languages.

3.3 Agent and query policies

The agent is a fixed ReAct loop (Yao et al. 2022) with native tool calling: at most three searches, then a short answer in the question’s language, at temperature 0. The policies differ by one sentence in the system prompt: none (**free**), “write every query in [the question’s language]” (**match**), “write every query in English” (**english**), or “search once in each language” (**dual**). For every call we log the query’s dominant script and whether a gold chunk was retrieved. The full prompts are in the supplementary material.

The design crosses question language (English or Indic), policy (four levels) and served pool (three levels) into 19 cells. Three extra arms isolate specific gaps: **closed-book** (no tool) measures what the model already knows; **oracle** (gold passage in the prompt) measures the integration gap; and a **repeat battery** (30 items $\times$ 5 runs) measures run-to-run noise.

3.4 Models

Table 1 lists the four models. Three are Gemini models accessed through OpenRouter; Qwen3-30B-A3B-Instruct-2507 is open-weight and served with vLLM on a rented GPU. A guardrail on the API account blocked other closed model families during the study.

3.5 Scoring and statistics

An answer is correct if it matches a gold alias in either script after normalisation. Otherwise a judge model outside the grid (Gemini 2.5 Flash) labels it correct, wrong or abstain; a hand audit of 40 judged answers agreed with every label (supplementary material). Every contrast is a paired within-item difference with a sign-flip permutation test and a 95% item-bootstrap interval (Efron and Tibshirani 1994). All p -values form one Benjamini–Hochberg family of 144 tests

(Benjamini and Hochberg 1995). The repeat battery changed the verdict on 1 of 30 Hindi items and 0 of 30 Telugu items, so we do not interpret differences below about two points.

To estimate the accuracy a real user would see, we weight the controlled accuracies by coverage, the share of questions whose Indic Wikipedia article states the answer. This is the law of total probability (Ross 2014):

$$\text{Acc}_L = \text{Cov}_L \cdot \text{Acc}(L \mid \text{in } L) + (1 - \text{Cov}_L) \cdot \text{Acc}(L \mid \text{en only}). \quad (1)$$

4 Results

4.1 Query language under the free policy

Given a Hindi question, the models write 18–57% of their queries in English: Gemini 2.5 Flash-Lite 30–33%, Gemini 3.1 Flash-Lite 31–38%, Gemini 3.6 Flash 45–57% and Qwen 18–19%. For Telugu questions all models use English more often (43–48%). The share barely changes with the served pool, so the agents do not notice when an English-only pool returns no Hindi text.

4.2 The gap under dense retrieval

Table 2 gives the main contrasts. With evidence in both languages, the Hindi question costs Gemini 2.5 Flash-Lite 9.4 points and Qwen 9.1 points; both survive correction. Gemini 3.1 Flash-Lite and 3.6 Flash lose 3.0 and 2.0 points, with intervals that include zero. No Telugu gap is significant. The gap belongs to the cheaper models, not to the languages.

Where the gap exists, changing the query language does not close it. For Gemini 2.5 Flash-Lite, matched, dual and English-only queries change Hindi accuracy by +0.7, +1.8 and +4.5 points; none survives correction. The gold passage is already in the top five for 79% of free-policy episodes, so there is little retrieval left to recover. The served pool also makes no difference: the Hindi question scores within one point on English-only and both pools, because bge-m3 finds the English passage for a Hindi query. Figure 2 shows every policy and pool.

4.3 The gap under keyword retrieval

BM25 gives the opposite result on the same items. The Hindi question now costs 18.9 points for Gemini 2.5 Flash-Lite and 19.5 for Qwen, and scores about 16 points below the same question under dense retrieval (all survive). Here the query language matters. Dual queries recover 9.3 and 13.6 points (both survive), and English-only queries on the English-only pool reach 62.7% against 43.4% for the free policy. The gold passage reaches the top five in 40% of free-policy episodes under BM25 against 79% under dense retrieval, and the agent abstains in 23–39% of BM25 episodes. The query-side gap is large when the search engine matches words in one language and nearly absent when it matches meaning across languages.

4.4 When searching in English helps and hurts

For Gemini 2.5 Flash-Lite on Hindi questions, English queries beat matched-language queries by 6.3 points on the English-only pool and lose 4.9 points on the Hindi-only pool.

Contrast (points), Hindi	Gemini 2.5 FL	Gemini 3.1 FL	Gemini 3.6 F	Qwen3-30B
Total gap, dense (English minus Hindi question)	+9.4 [4.5, 14.3]	+3.0 [−2.0, 8.0]	+2.0 [−4.0, 8.0]	+9.1 [3.8, 14.6]
recovered by dual queries	+1.8 [−3.2, 6.8]	+1.0 [−4.0, 6.0]	+0.0 [−6.0, 6.0]	+4.5 [0.7, 8.4]
recovered by English queries	+4.5 [0.0, 9.1]	+2.0 [−3.0, 7.0]	+6.0 [0.0, 14.0]	−1.4 [−6.3, 3.5]
Evidence-side: both minus English-only pool	−0.7 [−5.6, 3.8]	−2.0 [−7.0, 3.0]	+0.0 [−6.0, 6.0]	+5.6 [1.0, 10.5]
Integration gap (gold passage in prompt)	+6.3 [2.8, 9.8]	−2.0 [−6.0, 2.0]	−2.0 [−10.0, 4.0]	+4.2 [0.7, 7.7]
Closed-book gap	+3.1 [−1.0, 7.7]	+3.0 [−4.0, 10.0]	+2.0 [−4.0, 8.0]	+12.9 [7.3, 18.5]
Total gap, BM25	+18.9 [12.9, 24.8]	—	—	+19.5 [13.6, 25.8]
recovered by dual queries, BM25	+9.3 [3.9, 14.6]	—	—	+13.6 [8.4, 18.8]
Hindi question, BM25 minus dense	−15.7 [−21.3, −10.1]	—	—	−16.0 [−21.6, −10.5]

Table 2: Main paired contrasts for Hindi questions (free policy, both pools unless stated), with 95% item-bootstrap intervals. **Bold**: survives Benjamini–Hochberg correction over 144 tests. FL = Flash-Lite, F = Flash. Telugu and all other contrasts are in the supplementary material.

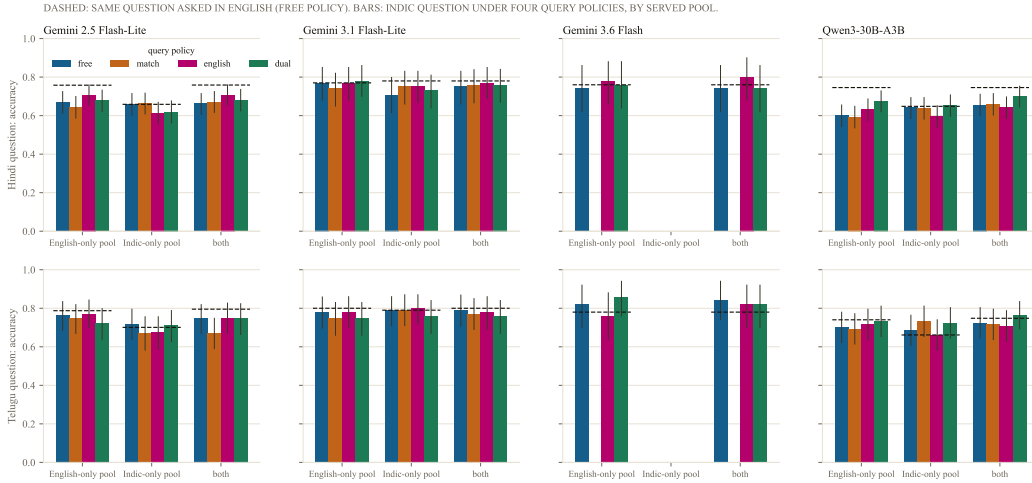


Figure 2: Accuracy of the Indic question under four query policies and three served pools (dense retriever), with 95% intervals. The dashed line is the English question under the free policy. Top row Hindi, bottom row Telugu.

The interaction is 11.2 points [4.2, 18.2] and survives correction. The effect does not appear for Telugu or for Gemini 3.1 Flash-Lite. This reproduces, for one model and language, the sign flip that earlier query-translation studies disagreed about (Amiraz et al. 2025; Ranaldi, Haddow, and Birch 2025): the query language mainly decides which pool the agent can read.

4.5 Coverage and expected user accuracy

The linked Hindi article exists for 67.5% of the 539 XorQA questions and states the answer for 29.1% [25.2, 33.0]. For Telugu the figures are 52.5% and 23.6% [20.0, 27.1]. Table 3 combines coverage with the controlled cells through Eq. 1.

Extrapolating the covered items gives a Hindi user of Gemini 2.5 Flash-Lite 66.7% against 75.8% for an English user under dense retrieval. Under BM25, the Hindi user gets 45.4% against 71.4%, and telling the agent to search in English lifts this to 62.9%. The extrapolation assumes uncovered questions are as easy as covered ones, so we tested it on all 382 uncovered Hindi questions. They are harder for everyone: the English question scores 53.4% on them against 75.8% on covered items. The measured expected accuracy is

51.4% for the Hindi user against 60.5% for the English user. The extrapolation overstates both users by 15–19 points, but the 9-point gap between them holds. On uncovered questions, serving the Hindi pool alongside English costs 5.8 points (survives): Hindi articles on the topic that lack the fact push the English passage that has it out of the top five.

4.6 The residual gap is reading

With the gold English passage in the prompt, Gemini 2.5 Flash-Lite still answers the Hindi question 6.3 points worse than the English one (survives), and Qwen 4.2 points worse. Given the Hindi question, Gemini 2.5 Flash-Lite answers correctly from the English passage 80.5% of the time and from the Hindi passage 72.8%: it reads English evidence better even for a Hindi question. The two stronger Gemini models read as well in Hindi as in English (integration gaps of −2.0 points). Closed-book, the models already know many of these facts (57.8–70.0% for Hindi questions), but 77% of correct agent answers from Gemini 2.5 Flash-Lite had a gold chunk in the trajectory. The natively written Telugu questions are harder (60.0% against 74.8%), but the agent behaves the same way on them: 44% English queries and no gain from

Model, language	Indic user, policy			English
	free	dual	english	user
G2.5 FL, Hi	66.7	67.9	70.7	75.8
measured	51.4	50.4	–	60.5
BM25	45.4	58.4	62.9	71.4
G2.5 FL, Te	76.0	72.9	76.6	78.7
measured	56.8	52.7	–	60.1
BM25	49.9	55.9	67.1	66.9
G3.1 FL, Hi	76.4	77.4	77.0	77.0
G3.1 FL, Te	78.0	75.2	78.0	80.0
G3.6 F, Hi	74.0	75.4	78.6	76.0
G3.6 F, Te	82.5	85.1	77.4	78.0
Qwen3, Hi	61.6	68.3	63.4	74.6
BM25	43.3	59.9	–	69.7
Qwen3, Te	70.6	74.0	71.5	74.0
BM25	56.6	64.8	–	68.5

Table 3: Expected accuracy (%) for a real user from Eq. 1, with coverage 29.1% (Hindi) and 23.6% (Telugu). Dense retriever unless marked BM25. “Measured” rows replace the extrapolated uncovered term with a direct run on the uncovered items. **Bold**: best query policy for the Indic user in each row.

dual queries. Qwen differs from Gemini in one respect: its closed-book Hindi accuracy is 12.9 points below English, so it depends more on retrieval.

5 Discussion

Comparison with prior findings. XBCP found retrieval to be a larger loss than integration for deep-research agents over a translated corpus (Lu et al. 2026). Our BM25 results agree. Our dense-retrieval results do not: with a cross-lingual retriever, retrieval is close to solved and the remaining loss is integration. This is consistent with MarathiWeb’s conclusion from inspection that failures come from integration “rather than missing sources” (Kale 2026), which we now measure under controlled conditions. DELTA’s point that gold-evidence location confounds language preference (Park et al. 2026) holds for agents too: the English-query benefit in Section 4.4 reverses when the evidence is only in Hindi.

What a product team should do. The right fix depends on the search engine. If the engine behind the agent matches words in one language, translate or duplicate the query into English; this recovers about half the Hindi gap. If the engine is cross-lingual, the query language is not the problem, and the remaining gap comes from the model’s reading of Indic text, which a stronger model closes. In both cases agents do not adapt on their own: they keep the same mix of query languages whatever the pool returns.

Pre-registered predictions. We recorded six directional predictions before the first grid call. D3 and D5 held. D1 failed: it predicted that agents would write most queries in English, which held for only one model and pool. D4 held in its first half (coverage below one half) and failed in its second. D2 and D6 split by model.

6 Conclusion

We separated the query-side, evidence-side and integration gaps for search agents answering Hindi and Telugu questions. The answer depends on the retriever. With dense cross-lingual retrieval, neither the query nor the web explains the Indic user’s deficit: the remaining 2–9 points come from the model reading the evidence, and the strongest models show no gap. With keyword retrieval, the query language decides the outcome: the Hindi question costs 19–20 points and dual-language queries recover about half of that. The facts that Indic Wikipedia lacks are also harder for everyone, and measuring them directly lowers both users’ accuracy by 15–19 points while leaving the gap unchanged.

Future work. Three extensions follow directly. First, more languages, starting with Bengali and Tamil, and natively written questions at scale. Second, a live-web arm with a commercial search engine, to test whether its behaviour resembles our dense or our BM25 condition. Third, training or prompting agents to choose the query language from the results they get back, since none of the models tested did so.

Limitations

Wikipedia is not the web. Our English pool is a 1.8% sample plus the articles behind each item, so it has fewer near-topic distractors than a real engine would return. We study two Indic languages. The XorQA questions are human translations of English questions; the natively written Telugu subset checks but does not remove this concern. Coverage relies on string match and two LLM reads with a 50-item audit; the audit found the first read too generous, so the reported figure is conservative. The closed-model arm is Gemini only, because an account guardrail blocked other providers, and the open-weight arm is one model. The judge is a Gemini model; string match decides 41% of items and the hand audit bounds the rest.

Ethical Considerations

We use only public Wikipedia text (CC BY-SA) and public benchmark questions. Models were queried through public APIs or run as open weights on our own GPU. No deployed search product was probed and no human subjects were involved. The pools can be rebuilt from the public snapshot with our scripts.

Use of AI Assistants

We used AI tools at several stages of this work, and the authors checked all outputs. Claude (Anthropic) helped write the experiment code and analysis scripts and helped draft and edit the paper text. Figure 1 was generated with a Gemini image model (Nano Banana) from a text description of our pipeline, and we checked every label against the design; all data figures are plotted from our results. LLMs are also part of the experimental setup: four models are the agents under study, Gemini models perform the two coverage reads, and Gemini 2.5 Flash judges answers that string matching cannot decide (Section 3.5).

References

- Acharya, A.; Murthy, R.; Kumar, V.; et al. 2024. Hindi-BEIR : A Large Scale Retrieval Benchmark in Hindi. *arXiv preprint arXiv:2408.09437*.
- Aji, A. F.; Ramadhan, F. R.; Zuhri, Z. M. K.; et al. 2026. HyperBrowseComp: A Multilingual and Multimodal Stress Test for Web-Browsing Agents. *arXiv preprint arXiv:2610.03574*.
- Amiraz, C.; Fyodorov, Y.; Haramaty, E.; et al. 2025. The Cross-Lingual Cost: Retrieval Biases in RAG over Arabic-English Corpora. *arXiv preprint arXiv:2507.07543*.
- Asai, A.; Kasai, J.; Clark, J. H.; et al. 2020. XOR QA: Cross-lingual Open-Retrieval Question Answering. *arXiv preprint arXiv:2010.11856*.
- Benjamini, Y.; and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1): 289–300.
- Caciolai, A.; Cabot, P.-L. H.; Cheng, C.; et al. 2026. OmnilingualGAIA2: Evaluating the Multilingual Gap in Frontier AI Agents. *arXiv preprint arXiv:2608.08775*.
- Chen, J.; Xiao, S.; Zhang, P.; et al. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv preprint arXiv:2402.03216*.
- Chen, Z.; Ma, X.; Zhuang, S.; et al. 2025. BrowseCompPlus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. *arXiv preprint arXiv:2508.06600*.
- Chua, L.; Ghazi, B.; Huang, Y.; et al. 2024. Crosslingual Capabilities and Knowledge Barriers in Multilingual Large Language Models. *arXiv preprint arXiv:2406.16135*.
- Clark, J. H.; Choi, E.; Collins, M.; Garrette, D.; Kwiatkowski, T.; Nikolaev, V.; and Palomaki, J. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470.
- Efron, B.; and Tibshirani, R. J. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- Goldman, O.; Shaham, U.; Malkin, D.; et al. 2025. ECLeK-Tic: a Novel Challenge Set for Evaluation of Cross-Lingual Knowledge Transfer. *arXiv preprint arXiv:2502.21228*.
- Goworek, R.; Macmillan-Scott, O.; and Özyiğit, E. B. 2025. What Drives Cross-lingual Ranking? Retrieval Approaches with Multilingual Language Models. *arXiv preprint arXiv:2511.19324*.
- Gupta, A.; Zhuang, Y.; Yu, Z.; et al. 2025. How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting. *arXiv preprint arXiv:2507.22923*.
- Haq, S.; Sharma, A.; and Bhattacharyya, P. 2023. IndicIR-Suite: Multilingual Dataset and Neural Information Models for Indian Languages. *arXiv preprint arXiv:2312.09508*.
- Hecht, B.; and Gergle, D. 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*, 291–300.
- Hofman, O.; Brokman, J.; Rachmil, O.; et al. 2025. MAPS: A Multilingual Benchmark for Agent Performance and Security. *arXiv preprint arXiv:2505.15935*.
- Hupkes, D.; and Bogoychev, N. 2025. MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages. *arXiv preprint arXiv:2504.10356*.
- Jain, R.; Mahajan, T.; Joshi, A.; et al. 2026. L3Cube-IndicQuest v2: A Large-Scale Multilingual Benchmark for Evaluating Factual Knowledge of Large Language Models Across Indic Languages. *arXiv preprint arXiv:2608.15535*.
- Jang, Y.; Hong, S.; Moon, H.; et al. 2026. MLAIRE: Multilingual Language-Aware Information Retrieval Evaluation Protocol. *arXiv preprint arXiv:2605.07249*.
- Jeong, M. 2026. Ko-WideSearch: A Korean Breadth-Search Benchmark for Exhaustive Set Enumeration by Web Agents. *arXiv preprint arXiv:2606.27595*.
- Jiang, Z.; Anastasopoulos, A.; Araki, J.; et al. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pre-trained Language Models. *arXiv preprint arXiv:2010.06189*.
- Kale, S. 2026. MarathiWeb: Evaluating Web-Search Language Models on Native Marathi Information Needs. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Kamali, O. 2026. wikipedia-monthly: monthly Wikipedia snapshots for 348 languages (snapshot 20261001). <https://huggingface.co/datasets/omarkamali/wikipedia-monthly>.
- Ki, D.; Carpuat, M.; McNamee, P.; et al. 2025. Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG. *arXiv preprint arXiv:2509.13930*.
- Ki, D.; Zhang, Z.; Cucerzan, S.; White, R. W.; and Gao, J. 2026. Where’s Waldo? Query-language Preference under Cross-lingual Knowledge Disparities. *arXiv preprint arXiv:2610.00606*.
- Kim, Y.; Uhlig, K.; and Wuebker, J. 2026. GAIA-v2-LILT: Multilingual Adaptation of Agent Benchmark beyond Translation. *arXiv preprint arXiv:2604.24929*.
- Lee, N.; Song, J.; and Kang, B. 2026. CORAL: Adaptive Retrieval Loop for Culturally-Aligned Multilingual RAG. *arXiv preprint arXiv:2604.25676*.
- Lee, N.; Yoon, D.; Son, G.; et al. 2026. K-BrowseComp: A Web Browsing Agent Benchmark Grounded in Korean Contexts. *arXiv preprint arXiv:2606.02404*.
- Lewis, P.; Oğuz, B.; Rinott, R.; et al. 2019. MLQA: Evaluating Cross-lingual Extractive Question Answering. *arXiv preprint arXiv:1910.07475*.
- Liang, Z.; Liu, Y.; Huang, Y.; et al. 2026. SearchAuditor: Auditing and Attributing Failures in Long-Horizon Search Agents. *arXiv preprint arXiv:2608.05212*.
- Liu, D.; Zhong, Y.; Zhang, W.; et al. 2026a. CIVI: A Framework for Diagnosing Search Agent Failures in Civic Information. *arXiv preprint arXiv:2609.08094*.
- Liu, Q.; Mao, J.; Zhu, F.; et al. 2026b. Diagnosing Search Behavior and Failure Modes in Long-Horizon Search Agents. *arXiv preprint arXiv:2608.01913*.

- Liu, W.; Trenous, S.; Ribeiro, L. F. R.; et al. 2025. XRAG: Cross-lingual Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.10089*.
- Lu, Y.; Zeng, Q.; Qi, H.; et al. 2026. Beyond Monolingual Deep Research: Evaluating Agents and Retrievers with Cross-Lingual BrowseComp-Plus. *arXiv preprint arXiv:2606.15345*.
- Luo, J.; Kutralingam, C.; Okoani, N.; et al. 2026. Lost in Execution: On the Multilingual Robustness of Tool Calling in Large Language Models. *arXiv preprint arXiv:2601.05366*.
- Mazumder, D.; Pathak, D.; Kodali, P.; et al. 2026. Evaluating Cross-lingual Knowledge Consistency in Code-Mixed vis-a-vis Indian Languages using IndicKLAR. *arXiv preprint arXiv:2605.29637*.
- Moon, H.; Kim, B.; and Verma, N. 2025. Quality-Aware Translation Tagging in Multilingual RAG system. *arXiv preprint arXiv:2510.23070*.
- Park, J.; Kim, B.; Hwang, S.; et al. 2026. Enhancing Multilingual RAG Systems with Debaised Language Preference-Guided Query Fusion. *arXiv preprint arXiv:2601.02956*.
- Park, J.; and Lee, H. 2025. Investigating Language Preference of Multilingual RAG Systems. *arXiv preprint arXiv:2502.11175*.
- Piratla, V.; Jain, P.; Singh, D.; et al. 2025. Rethinking Cross-lingual Gaps from a Statistical Viewpoint. *arXiv preprint arXiv:2510.15551*.
- Ranaldi, L.; Haddow, B.; and Birch, A. 2025. Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task. *arXiv preprint arXiv:2504.03616*.
- Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.
- Rohera, P.; Ginimav, C.; Sawant, G.; et al. 2025. Better To Ask in English? Evaluating Factual Accuracy of Multilingual LLMs in English and Low-Resource Languages. *arXiv preprint arXiv:2504.20022*.
- Ross, S. 2014. *A First Course in Probability*. Pearson, 9 edition.
- Singh, H.; Gupta, N.; Bharadwaj, S.; et al. 2024. IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages. *arXiv preprint arXiv:2404.16816*.
- Thakur, N.; Zhang, X.; Sharifmoghaddam, S.; Ma, X.; Chen, Z.; and Lin, J. 2026. MAST: Multilingual Agentic Search Track at FIRE 2026. <https://mast-benchmark.github.io>. Shared task; organiser baselines only as of 2026-10-10.
- Verma, S.; Khan, M. S. U. R.; Kumar, V.; et al. 2024. MILU: A Multi-task Indic Language Understanding Benchmark. *arXiv preprint arXiv:2411.02538*.
- Wang, D.; Mo, G.; Shi, Y.; et al. 2026. All Languages Matter: Understanding and Mitigating Language Bias in Multilingual RAG. *arXiv preprint arXiv:2604.20199*.
- Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*.
- Zhang, X.; Thakur, N.; Ogundepo, O.; et al. 2022. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. *arXiv preprint arXiv:2210.09984*.
- Zhou, P.; Leon, B.; Ying, X.; et al. 2025. BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese. *arXiv preprint arXiv:2504.19314*.