

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

Anonymous ACL submission

Abstract

A search agent answers a Hindi or Telugu question less accurately than the same question in English. Two causes can produce this drop, and each needs a different fix. The agent may write poor queries or search in the wrong language (the *query-side gap*). Or the evidence may not exist in the user’s language (the *evidence-side gap*). We separate the two. We use human-parallel Hindi/English and Telugu/English questions whose answer appears in both Wikipedias (287 Hindi, 127 Telugu). The agent searches a frozen pool of 1.9 million Wikipedia passages served as English-only, Indic-only, or both. We log the language of every query and then force it. Four models (three Gemini, one open-weight Qwen3) run the full 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 points (not significant). Telugu gaps are 1–5 points and not significant. Changing the query language or the served pool recovers almost nothing; with the gold passage in the prompt, Hindi is still 4–6 points worse, so the remaining gap is in reading. Under BM25 the Hindi gap grows to 19–20 points, and dual-language queries recover 9–14 of them. The Hindi and Telugu Wikipedias contain the answer for only 29% and 24% of questions. Measured 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).

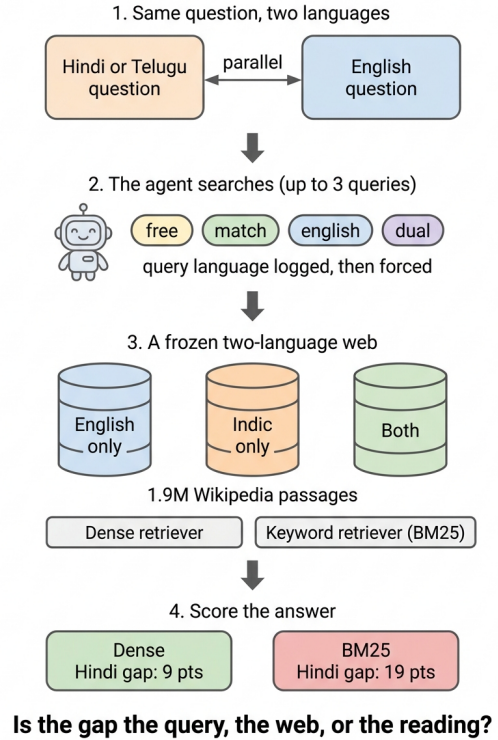


Figure 1: Study design. The same question is asked in an Indic language and in English. The agent searches a frozen Wikipedia pool served as English-only, Indic-only or both, under a free or forced query-language policy, with a dense or keyword retriever. The gaps shown are for Gemini 2.5 Flash-Lite,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

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 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., 2026a; 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 et al., 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 et al., 2025; Moon et al., 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 et al., 2026b). Dense multilingual retrievers match across languages where lexical matching cannot (Goworek et al., 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 et al., 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). Appendix A 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 (Appendix E). 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 et al., 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 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 spelling variants such as ओरेगन/ऑरेगॉन defeat exact token match. 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 Appendix D.

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 (Appendix F). Every contrast is a paired within-item difference with a sign-flip permutation test and

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.

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% (Appendix Figure 3). 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. 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 et al., 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; Appendix Figure 4 plots it.

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%,

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 Appendix Table 5.

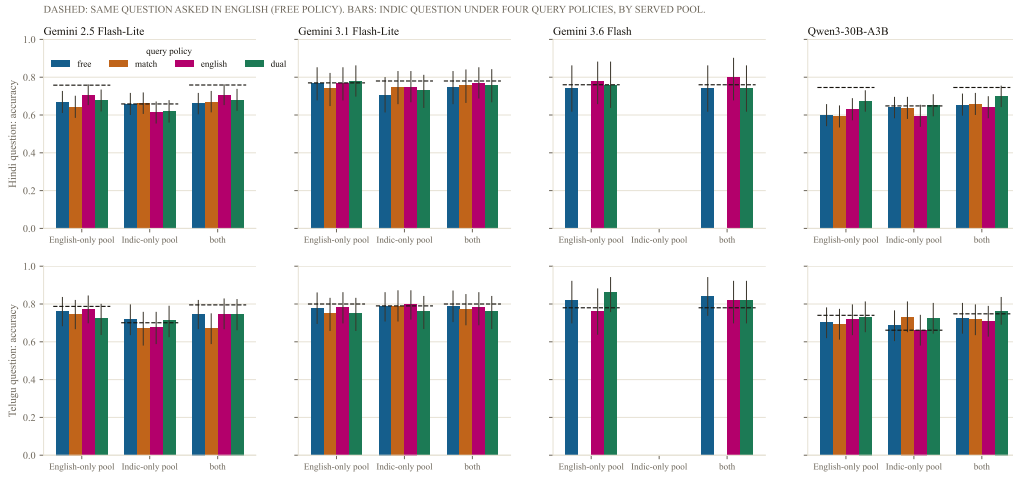


Figure 2: Accuracy of the Indic question under four query policies and three served pools (dense retriever), with 95% intervals. The dashed line is the English question under the free policy. Top row Hindi, bottom row Telugu.

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 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 (Appendix Figure 5).

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 re-

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.

triever, 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

- Arkadeep Acharya, Rudra Murthy, Vishwajeet Kumar, et al. 2024. [Hindi-BEIR : A Large Scale Retrieval Benchmark in Hindi](#). *arXiv preprint arXiv:2408.09437*.
- Alham Fikri Aji, Faiz Rizki Ramadhan, Zayd M. K. Zuhri, et al. 2026. [HyperBrowseComp: A Multilingual and Multimodal Stress Test for Web-Browsing Agents](#). *arXiv preprint arXiv:2610.03574*.
- Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, et al. 2025. [The Cross-Lingual Cost: Retrieval Biases in RAG over Arabic-English Corpora](#). *arXiv preprint arXiv:2507.07543*.
- Akari Asai, Jungo Kasai, Jonathan H. Clark, et al. 2020. [XOR QA: Cross-lingual Open-Retrieval Question Answering](#). *arXiv preprint arXiv:2010.11856*.
- Yoav Benjamini and Yosef Hochberg. 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.
- Andrea Caciolai, Pere-Lluís Huguet Cabot, Chierh Cheng, et al. 2026. [OmnilingualGAIA2: Evaluating the Multilingual Gap in Frontier AI Agents](#). *arXiv preprint arXiv:2608.08775*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, et al. 2024. [M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation](#). *arXiv preprint arXiv:2402.03216*.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, et al. 2025. [BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent](#). *arXiv preprint arXiv:2508.06600*.
- Lynn Chua, Badih Ghazi, Yangsibo Huang, et al. 2024. [Crosslingual Capabilities and Knowledge Barriers in Multilingual Large Language Models](#). *arXiv preprint arXiv:2406.16135*.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- Omer Goldman, Uri Shaham, Dan Malkin, et al. 2025. [ECLeKTic: a Novel Challenge Set for Evaluation of Cross-Lingual Knowledge Transfer](#). *arXiv preprint arXiv:2502.21228*.
- Roksana Goworek, Olivia Macmillan-Scott, and Eda B. Özyigit. 2025. [What Drives Cross-lingual Ranking? Retrieval Approaches with Multilingual Language Models](#). *arXiv preprint arXiv:2511.19324*.
- Aman Gupta, Yingying Zhuang, Zhou Yu, et al. 2025. [How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting](#). *arXiv preprint arXiv:2507.22923*.
- Saiful Haq, Ashutosh Sharma, and Pushpak Bhat-tacharyya. 2023. [IndicIRSuite: Multilingual Dataset and Neural Information Models for Indian Languages](#). *arXiv preprint arXiv:2312.09508*.
- 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*, pages 291–300.
- Omer Hofman, Jonathan Brokman, Oren Rachmil, et al. 2025. [MAPS: A Multilingual Benchmark for Agent Performance and Security](#). *arXiv preprint arXiv:2505.15935*.
- Dieuwke Hupkes and Nikolay Bogoychev. 2025. [MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages](#). *arXiv preprint arXiv:2504.10356*.
- Rinit Jain, Tirthraj Mahajan, Advait Joshi, 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*.
- Youngjoon Jang, Seongtae Hong, Hyeonseok Moon, et al. 2026. [MLAIRE: Multilingual Language-Aware Information Retrieval Evaluation Protocol](#). *arXiv preprint arXiv:2605.07249*.
- Minbyul Jeong. 2026. [Ko-WideSearch: A Korean Breadth-Search Benchmark for Exhaustive Set Enumeration by Web Agents](#). *arXiv preprint arXiv:2606.27595*.

560	Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, et al. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models . <i>arXiv preprint arXiv:2010.06189</i> .	614
561		615
562		616
563		617
564	Sachin Kale. 2026. MarathiWeb: Evaluating Web-Search Language Models on Native Marathi Information Needs . In <i>Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> . Title transcribed from the ACM DOI record; author list as indexed by researchr.	618
565		
566		
567		
568		
569		
570		
571	Omar Kamali. 2026. wikipedia-monthly: monthly Wikipedia snapshots for 348 languages (snapshot 20261001) . https://huggingface.co/datasets/omarkamali/wikipedia-monthly .	619
572		620
573		621
574		622
575	Dayeon Ki, Marine Carpuat, Paul McNamee, et al. 2025. Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG . <i>arXiv preprint arXiv:2509.13930</i> .	623
576		624
577		625
578		626
579	Dayeon Ki, Zhiqiang Zhang, Silviu Cucerzan, Ryen W. White, and Jianfeng Gao. 2026. Where’s Waldo? Query-language Preference under Cross-lingual Knowledge Disparities . <i>arXiv preprint arXiv:2610.00606</i> .	627
580		
581		
582		
583		
584	Yunsu Kim, Kaden Uhlig, and Joern Wuebker. 2026. GAIA-v2-LILT: Multilingual Adaptation of Agent Benchmark beyond Translation . <i>arXiv preprint arXiv:2604.24929</i> .	628
585		629
586		630
587		
588	Nahyun Lee, Dongkeun Yoon, Guijin Son, et al. 2026a. K-BrowseComp: A Web Browsing Agent Benchmark Grounded in Korean Contexts . <i>arXiv preprint arXiv:2606.02404</i> .	631
589		632
590		633
591		634
592	Nayeon Lee, Jiwoo Song, and Byeongcheol Kang. 2026b. CORAL: Adaptive Retrieval Loop for Culturally-Aligned Multilingual RAG . <i>arXiv preprint arXiv:2604.25676</i> .	635
593		636
594		637
595		
596	Patrick Lewis, Barlas Oğuz, Ruty Rinott, et al. 2019. MLQA: Evaluating Cross-lingual Extractive Question Answering . <i>arXiv preprint arXiv:1910.07475</i> .	638
597		639
598		640
599	Zhixiang Liang, Yifei Liu, Yidan Huang, et al. 2026. SearchAuditor: Auditing and Attributing Failures in Long-Horizon Search Agents . <i>arXiv preprint arXiv:2608.05212</i> .	641
600		642
601		643
602		644
603	Dingying Liu, Yunshun Zhong, Wentao Zhang, et al. 2026a. CIVI: A Framework for Diagnosing Search Agent Failures in Civic Information . <i>arXiv preprint arXiv:2609.08094</i> .	645
604		646
605		647
606		648
607	Qi Liu, Jiaxin Mao, Fengbin Zhu, et al. 2026b. Diagnosing Search Behavior and Failure Modes in Long-Horizon Search Agents . <i>arXiv preprint arXiv:2608.01913</i> .	649
608		650
609		651
610		652
611	Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, et al. 2025. XRAG: Cross-lingual Retrieval-Augmented Generation . <i>arXiv preprint arXiv:2505.10089</i> .	653
612		654
613		655
	Yuheng Lu, Qingcheng Zeng, Heli Qi, et al. 2026. Beyond Monolingual Deep Research: Evaluating Agents and Retrievers with Cross-Lingual BrowseComp-Plus . <i>arXiv preprint arXiv:2606.15345</i> .	656
		657
		658
	Jiaxin Luo, Chandra Kutralingam, Nnamdi Okoani, et al. 2026. Lost in Execution: On the Multilingual Robustness of Tool Calling in Large Language Models . <i>arXiv preprint arXiv:2601.05366</i> .	659
		660
	Debajyoti Mazumder, Divyansh Pathak, Prashant Kodali, et al. 2026. Evaluating Cross-lingual Knowledge Consistency in Code-Mixed vis-a-vis Indian Languages using IndicKLAR . <i>arXiv preprint arXiv:2605.29637</i> .	661
		662
		663
		664
		665
	Hoyeon Moon, Byeolhee Kim, and Nikhil Verma. 2025. Quality-Aware Translation Tagging in Multilingual RAG system . <i>arXiv preprint arXiv:2510.23070</i> .	
	Jeonghyun Park, Byeongjeong Kim, Seojin Hwang, et al. 2026. Enhancing Multilingual RAG Systems with Debaised Language Preference-Guided Query Fusion . <i>arXiv preprint arXiv:2601.02956</i> .	
	Jeonghyun Park and Hwanhee Lee. 2025. Investigating Language Preference of Multilingual RAG Systems . <i>arXiv preprint arXiv:2502.11175</i> .	
	Vihari Piratla, Purvam Jain, Darshan Singh, et al. 2025. Rethinking Cross-lingual Gaps from a Statistical Viewpoint . <i>arXiv preprint arXiv:2510.15551</i> .	
	Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task . <i>arXiv preprint arXiv:2504.03616</i> .	
	Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. <i>Foundations and Trends in Information Retrieval</i> , 3(4):333–389.	
	Pritika Rohera, Chaitrali Ginimav, Gayatri Sawant, et al. 2025. Better To Ask in English? Evaluating Factual Accuracy of Multilingual LLMs in English and Low-Resource Languages . <i>arXiv preprint arXiv:2504.20022</i> .	
	Sheldon Ross. 2014. <i>A First Course in Probability</i> , 9 edition. Pearson.	
	Harman Singh, Nitish Gupta, Shikhar Bharadwaj, et al. 2024. IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages . <i>arXiv preprint arXiv:2404.16816</i> .	
	Nandan Thakur, Xinyu Zhang, Sahel Sharifmoghaddam, Xueguang Ma, Zijian Chen, and Jimmy Lin. 2026. MAST: Multilingual Agentic Search Track at FIRE 2026 . https://mast-benchmark.github.io . Shared task; organiser baselines only as of 2026-10-10.	

666	Sshubam Verma, Mohammed Safi Ur Rahman Khan,	- If the evidence does not contain	717
667	Vishwajeet Kumar, et al. 2024. MILU: A Multi-task	the answer, you may still answer from	718
668	Indic Language Understanding Benchmark . <i>arXiv</i>	your own knowledge, but prefer the	719
669	<i>preprint arXiv:2411.02538</i> .	evidence.{policy}	720
670	Dan Wang, Guozhao Mo, Yafei Shi, et al. 2026. All	Policy sentence: free	721
671	Languages Matter: Understanding and Mitigating	(none)	722
672	Language Bias in Multilingual RAG . <i>arXiv preprint</i>	Policy sentence: match	723
673	<i>arXiv:2604.20199</i> .	- Write every search query in	724
674	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak	{qlang}, the language of the	725
675	Shafran, Karthik Narasimhan, and Yuan Cao. 2022.	question.	726
676	ReAct: Synergizing Reasoning and Acting in Lan-	Policy sentence: english	727
677	guage Models . <i>arXiv preprint arXiv:2210.03629</i> .	- Write every search query in	728
678	Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo,	English, even though the question is	729
679	et al. 2022. Making a MIRACL: Multilingual Infor-	in {qlang}. Translate the	730
680	mation Retrieval Across a Continuum of Languages .	information need into an English	731
681	<i>arXiv preprint arXiv:2210.09984</i> .	query.	732
682	Peilin Zhou, Bruce Leon, Xiang Ying, et al. 2025.	Policy sentence: dual	733
683	BrowseComp-ZH: Benchmarking Web Browsing	- At every step call `search` twice:	734
684	Ability of Large Language Models in Chinese . <i>arXiv</i>	once with a query in {qlang} (the	735
685	<i>preprint arXiv:2504.19314</i> .	language of the question) and once	736
686	A Comparison with the closest prior	with the same information need as an	737
687	work	English query. Both calls count as	738
688	Table 4 lists what each close neighbour established	one step.	739
689	and what our design adds.	Tool schema (JSON)	740
690	B All paired contrasts	{	741
691	Table 5 lists every paired contrast for both lan-	"type": "function",	742
692	guages and all four models.	"function": {	743
693	C Additional figures	"name": "search",	744
694	D Verbatim prompts and tool schema	"description": "Search the document	745
695	All prompts are reproduced exactly as sent. The	collection. Returns the most	746
696	policy sentence is appended to the system prompt;	relevant passages with their	747
697	{qlang} is the name of the question's language	titles.",	748
698	and {max_searches}=3, {k}=5.	"parameters": {	749
699	Agent system prompt	"type": "object",	750
700	You are a search agent. You answer	"properties": {	751
701	factual questions by searching a	"query": {	752
702	document collection with the `search`	"type": "string",	753
703	tool, reading the returned passages,	"description": "The search query."	754
704	and then giving a short final answer.	}	755
705	Rules:	},	756
706	- You may call `search` at most	"required": [	757
707	{max_searches} times. Each call	"query"	758
708	returns the top {k} passages.	]	759
709	- When you have enough evidence, or	}	760
710	when you have used all your searches,	}	761
711	stop calling tools and reply with the	}	762
712	final answer only: a short phrase (a	Empty-result message	763
713	name, date, number, or term), no	No passages found for this query.	764
714	explanation.		
715	- Answer in the same language as the		
716	question.		

Closest neighbour	What it established / what it does not cover
XBCP (Lu et al., 2026)	Deep-research agents on BrowseComp-Plus with the corpus translated into 12 languages; oracle decomposition into retrieval vs integration. Question always English; query language never logged or forced; no Indic language; translated rather than native evidence.
MarathiWeb (Kale, 2026)	Two web-search LLMs on native Marathi queries; failures attributed to integration by inspection. One language; evidence availability observed, not manipulated; no parallel English arm.
MAST Indic track (Thakur et al., 2026)	Nine Indic query languages over the English BrowseComp-Plus corpus. English-only corpus, so the evidence-side gap is zero by construction; a systems shared task.
DELTA (Park et al., 2026)	Names the gold-availability prior and regresses it out of single-shot RAG. No agent, no Indic, statistical rather than experimental control.
Where’s Waldo (Ki et al., 2026)	Query-language preference with fixed paired context when a fact exists in one edition only; Bengali worst. No retrieval, no agent, no search-query language.
Diagnosing search failures (Liu et al., 2026b)	Retrieval gap vs utilisation gap with oracle evidence. English only; the retrieval gap is not split into query formulation vs corpus coverage.
Cross-Lingual Cost (Amiraz et al., 2025); XRAG (Liu et al., 2025); tRAG/CrossRAG (Ranaldi et al., 2025); CORAL (Lee et al., 2026b)	Query-language vs document-language effects in RAG; one controlled Google ablation; an agent that picks the corpus language. Non-agentic or end-accuracy only; no Indic; no decomposition.

Table 4: Novelty delineation. Comparison with the closest prior work.

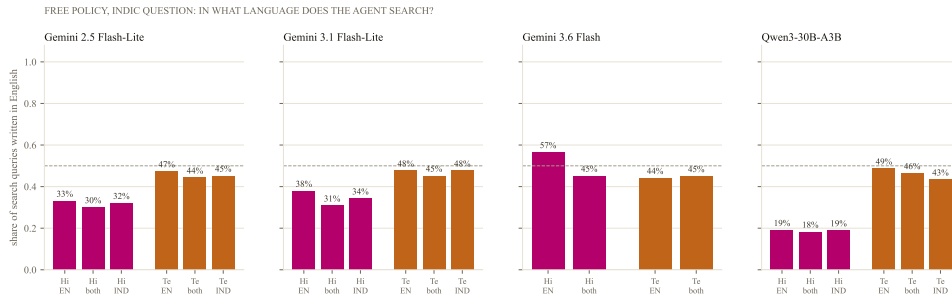


Figure 3: Share of search queries written in Latin script when the question is in Hindi (magenta) or Telugu (orange), free policy, by served pool (EN = English-only, IND = Indic-only). Dashed line: one half.

Closed-book system prompt

You answer factual questions from your own knowledge. Reply with the final answer only: a short phrase (a name, date, number, or term), no explanation. Answer in the same language as the question.

Oracle system prompt

You answer a factual question using the passage provided. Reply with the final answer only: a short phrase (a name, date, number, or term), no explanation. Answer in the same language as the question.

Oracle user message

Passage:

"""

{passage}

"""

Question: {question}

Judge system prompt

You are a strict grader of short answers to factual questions. Decide whether the candidate answer expresses the same fact as any of the gold answers. Accept equivalent forms: a transliteration or translation of the gold answer into another language or script, a date written differently, a number with or without units, a name with or without honorifics, a more specific answer that contains the gold answer. Reject answers that name a different entity, date, or number, that hedge between several candidates. If the candidate declines to answer, says the information is unavailable, or asks a question back, that is ABSTAIN.

Judge user message

Question (English): {q_en}

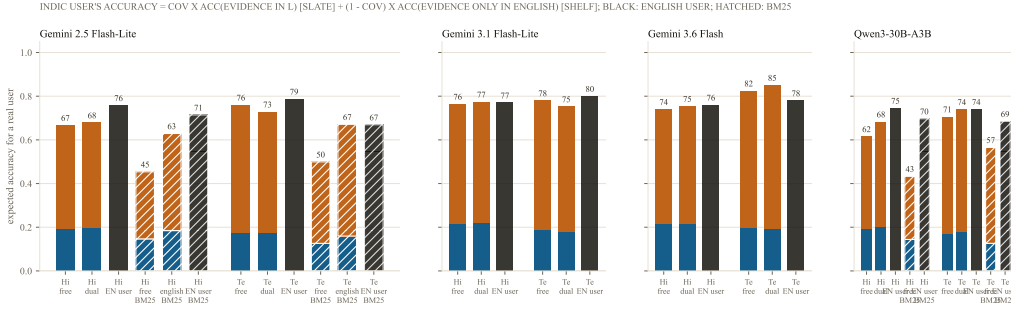


Figure 4: Expected accuracy of a real Indic user from Eq. 1: the slate part is coverage times accuracy with evidence in both languages; the orange part is the remaining share times accuracy on the English-only pool; black is the English user. Hatched bars use BM25.

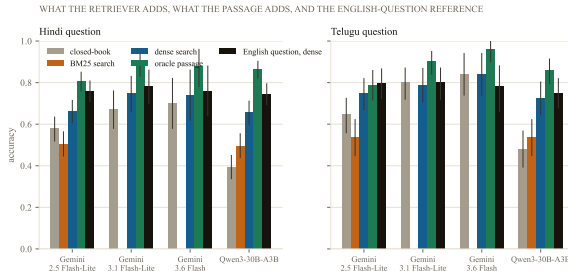


Figure 5: Indic question, free policy, both pools: closed-book, BM25 search, dense search, gold passage in the prompt, and the English question with dense search, per model.

Question (original): {q_L}
 Gold answers: {gold}
 Candidate answer: {cand}
 Reply with exactly one word: CORRECT, WRONG, or ABSTAIN (the candidate says it does not know, cannot find the answer, or asks for clarification instead of answering).

E Coverage measurement and audit

For each of the 539 XorQA-In test questions per language we resolved the English source article to its Hindi or Telugu counterpart through Wikidata sitelinks (916 of 918 distinct titles resolved; two had no Wikidata item) and looked the counterpart up in the 20261001 snapshot. An article exists for 364 Hindi and 283 Telugu items. Coverage was then decided in three steps. (1) *String match*: the Indic gold answer, or the English gold answer, occurs in the article after NFC normalisation and punctuation stripping; 119 Hindi and 94 Telugu items (the 237 Telugu items whose released translated answer is the function word “అని” were matched on the English answer only). (2) *Two independent LLM reads*: gemini-2.5-flash-lite and

then gemini-3.6-flash were each given the question in both languages, the gold answers and the full article (up to 14,000 characters) and asked to quote verbatim the sentence that states the answer or reply NONE; a read counts only if the quoted sentence is found in the article. The first read accepted 205 Hindi and 179 Telugu items; the second accepted 158 and 129; both agree on 157 and 127, which are the verified-coverage counts (29.1% and 23.6%). (3) *Manual audit*: before the second read, we drew 25 items per language stratified over first-read outcomes and labelled them by reading the quoted sentence against the question. Of the 30 first-read acceptances, 21 were correct, 7 were uncertain (the quote did not visibly contain the answer) and 2 were wrong (the Hindi article gives 13 South American countries where the gold says twelve; the Telugu Gandhi article never names Subhas Chandra Bose). The second read rejected 1 of the 2 wrong, 4 of the 7 uncertain and 8 of the 21 correct acceptances, so the final coverage figure is conservative: it under-counts rather than over-counts. The 20 first-read rejections in the audit were not labelled by hand; the second read accepted 3 of them with a verbatim quote, but an item needs both reads to count, so they stay outside the controlled grid and the coverage figure. The audit file with every label is shipped as `results/coverage_audit.json`.

F Judge audit

Of the judged agent answers, the string rule decided 41% (exact or contained match to a gold alias in either script after digit normalisation); the rest went to the LLM judge (gemini-2.5-flash, not in the grid), which returned CORRECT, WRONG or ABSTAIN. We drew 40 judged answers (20 per language: 10 string-correct, 5 LLM-

correct, 5 LLM-wrong or abstain) and labelled them by hand against the question and gold aliases. All 40 verdicts agreed with the manual label. Cases worth noting: a Korean-script answer accepted for the romanised gold “mumu”; “labour movement” accepted for “trade union movement”; an answer naming one of three states in the gold marked WRONG, which we consider defensible; and seven abstentions (“the search results do not contain...”) marked ABSTAIN and scored as wrong. The sample is shipped as results/audit_sample.json.

G Per-cell tables

H Cost

I Per-cell tables

J Cost

Contrast (percentage points)	Gemini 2.5 Flash-Lite		Gemini 3.1 Flash-Lite		Gemini 3.6 Flash		Qwen3-30B-A3B	
	Hindi	Telugu	Hindi	Telugu	Hindi	Telugu	Hindi	Telugu
Total gap: English minus Indic question (free, both pools)	+9.4 [+4.5, +14.3]*	+4.7 [-1.6, +11.0]	+3.0 [-2.0, +8.0]	+1.0 [-5.1, +7.1]	+2.0 [-4.0, +8.0]	-6.0 [-14.0, +0.0]	+9.1 [+3.8, +14.6]*	+2.4 [-4.7, +9.4]
Recovered by dual-language queries	+1.8 [-3.2, +6.8]	-2.5 [-9.2, +4.2]	+1.0 [-4.0, +6.0]	-3.0 [-10.1, +3.0]	+0.0 [-6.0, +6.0]	-2.0 [-8.0, +4.0]	+4.5 [+0.7, +8.4]*	+3.9 [-0.8, +9.4]
Recovered by English-only queries	+4.5 [+0.0, +9.1]	-0.0 [-6.3, +6.3]	+2.0 [-3.0, +7.0]	-1.0 [-6.1, +4.0]	+6.0 [+0.0, +14.0]	-2.0 [-10.0, +6.0]	-1.4 [-6.3, +3.5]	-1.6 [-7.1, +3.9]
Recovered by matched-language queries	+0.7 [-3.1, +4.9]	-7.9 [-14.2, -1.6]*	+1.0 [-3.0, +6.1]	-2.0 [-6.1, +2.0]	-	-	+0.3 [-2.8, +3.5]	-0.8 [-6.3, +4.7]
Evidence-side: both pools minus English-only pool	-0.7 [-5.6, +3.8]	-1.6 [-7.9, +4.7]	-2.0 [-7.0, +3.0]	+1.0 [-3.0, +5.1]	+0.0 [-6.0, +6.0]	+2.0 [-6.0, +10.0]	+5.6 [+1.0, +10.5]*	+2.4 [-3.1, +7.9]
Evidence-side: both pools minus Indic-only pool	+0.3 [-3.5, +4.2]	+3.1 [-2.4, +8.7]	+4.0 [-2.0, +10.1]	+0.0 [-6.1, +6.1]	-	-	+1.4 [-1.7, +4.9]	+3.9 [-2.4, +11.0]
English queries minus matched queries, English-only pool	+6.3 [+1.7, +11.2]*	+2.4 [-3.1, +7.9]	+3.0 [-2.0, +8.0]	+3.0 [-2.0, +8.0]	-	-	+3.8 [-0.7, +8.7]	+2.4 [-4.7, +8.7]
English queries minus matched queries, Indic-only pool	-4.9 [-10.5, +0.3]	+0.8 [-7.1, +8.7]	+0.0 [-5.0, +5.0]	+1.0 [+0.0, +3.0]	-	-	-4.2 [-8.4, +0.0]	-7.1 [-14.2, +0.0]
Symmetric monolingual: en/en minus L/L	+10.1 [+4.5, +15.7]*	+12.6 [+3.9, +21.3]*	-	-	-	-	+13.2 [+7.7, +19.2]*	+3.1 [-4.7, +11.0]
Integration gap: oracle passage, English minus Indic question	+6.3 [+2.8, +9.8]*	+6.3 [+0.0, +12.6]	-2.0 [-6.0, +2.0]	-1.0 [-5.0, +2.0]	-2.0 [-10.0, +4.0]	-2.0 [-6.0, +0.0]	+4.2 [+0.7, +7.7]*	+6.3 [+1.6, +11.0]*
Closed-book gap	+3.1 [-1.0, +7.7]	+3.9 [-2.4, +11.0]	+3.0 [-4.0, +10.0]	-1.0 [-7.0, +5.0]	+2.0 [-4.0, +8.0]	+0.0 [-6.0, +6.0]	+12.9 [+7.3, +18.5]*	+7.1 [+0.0, +14.2]
Total gap under BM25	+18.9 [+12.9, +24.8]*	+11.0 [+2.4, +19.7]*	-	-	-	-	+19.5 [+13.6, +25.8]*	+14.2 [+6.3, +22.0]*
Symmetric monolingual under BM25	+35.3 [+29.0, +42.0]*	+31.0 [+20.6, +41.3]*	-	-	-	-	-	-
Recovered by dual queries under BM25	+9.3 [+3.9, +14.6]*	+7.6 [-1.7, +16.8]	-	-	-	-	+13.6 [+8.4, +18.8]*	+9.4 [+2.4, +17.3]*
Indic question: BM25 minus dense (free, both)	-15.7 [-21.3, -10.1]*	-21.3 [-30.7, -12.6]*	-	-	-	-	-16.0 [-21.6, -10.5]*	-18.9 [-28.3, -9.4]*
Policy $\times$ evidence interaction	+11.2 [+4.2, +18.2]*	+1.6 [-7.9, +11.0]	+3.0 [-4.0, +10.0]	+2.0 [-3.0, +8.0]	-	-	+8.0 [+2.1, +14.3]*	+9.4 [+0.0, +18.9]

Table 5: Paired within-item contrasts (percentage points) with item-bootstrap 95% intervals. *: sign-flip permutation $p < 0.05$. Bold: survives Benjamini–Hochberg over the full family of 144 tests (threshold $p \leq 0.0076$). Positive total gap = the English question is answered more often. Dense retriever unless marked BM25.

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	287	57.8	[51.9, 63.4]	0%	0%
Hindi	dual	Indic-only	281	61.9	[56.2, 67.6]	50%	79%
Hindi	dual	Indic-only [bm25]	280	34.3	[28.6, 39.6]	52%	20%
Hindi	dual	both	280	67.9	[62.5, 73.6]	50%	87%
Hindi	dual	both [bm25]	281	58.7	[53.0, 64.4]	51%	67%
Hindi	dual	English-only	280	67.9	[62.1, 73.2]	50%	83%
Hindi	dual	English-only [bm25]	280	58.2	[52.5, 63.9]	51%	61%
Hindi	english	Indic-only	287	61.3	[55.7, 66.9]	100%	68%
Hindi	english	Indic-only [bm25]	285	32.3	[27.0, 37.9]	100%	8%
Hindi	english	both	287	70.7	[65.5, 76.0]	100%	80%
Hindi	english	both [bm25]	287	63.4	[57.8, 69.0]	100%	63%
Hindi	english	English-only	287	70.7	[65.5, 76.0]	100%	80%
Hindi	english	English-only [bm25]	287	62.7	[57.1, 68.3]	100%	59%
Hindi	free	Indic-only	287	65.9	[60.3, 71.4]	32%	74%
Hindi	free	Indic-only [bm25]	287	40.1	[34.5, 45.6]	38%	17%
Hindi	free	both	287	66.2	[60.6, 71.4]	30%	79%
Hindi	free	both [bm25]	286	50.3	[44.7, 56.3]	28%	40%
Hindi	free	English-only	287	66.9	[61.3, 72.5]	33%	72%
Hindi	free	English-only [bm25]	286	43.4	[37.8, 49.3]	35%	38%
Hindi	free	(gold hidden)	49	55.1	[40.8, 69.4]	27%	0%
Hindi	match	Indic-only	286	66.4	[60.8, 71.7]	1%	74%
Hindi	match	Indic-only [bm25]	286	35.3	[29.7, 40.9]	1%	16%
Hindi	match	both	287	66.9	[61.7, 72.5]	1%	78%
Hindi	match	both [bm25]	287	43.6	[38.0, 49.5]	2%	23%
Hindi	match	English-only	286	64.3	[58.7, 69.9]	1%	66%
Hindi	match	English-only [bm25]	286	19.6	[15.0, 24.5]	2%	3%
Hindi	oracle L	(gold passage)	287	72.8	[67.6, 78.0]	0%	0%
Hindi	oracle en	(gold passage)	287	80.5	[76.0, 85.0]	0%	0%
English	closed	(no tool)	287	61.0	[55.4, 66.6]	0%	0%
English	dual	Indic-only	287	64.5	[58.9, 70.0]	97%	72%
English	dual	Indic-only [bm25]	287	25.8	[20.9, 31.0]	97%	8%
English	free	Indic-only	287	65.9	[60.3, 71.1]	100%	70%
English	free	Indic-only [bm25]	286	32.2	[26.6, 37.8]	100%	8%
English	free	both	286	75.9	[71.0, 80.8]	100%	83%
English	free	both [bm25]	287	69.0	[63.4, 74.2]	100%	65%
English	free	English-only	285	75.8	[70.9, 80.7]	100%	82%
English	free	English-only [bm25]	287	71.4	[66.2, 76.7]	100%	65%
English	match	Indic-only	287	64.8	[59.2, 70.4]	100%	70%
English	match	Indic-only [bm25]	287	38.3	[32.8, 43.9]	100%	8%
English	match	both	287	74.9	[70.0, 79.8]	100%	82%
English	match	both [bm25]	287	71.4	[66.2, 76.7]	100%	65%
English	match	English-only	287	76.3	[71.4, 81.2]	100%	82%
English	match	English-only [bm25]	287	70.7	[65.5, 76.0]	100%	64%
English	oracle en	(gold passage)	287	86.8	[82.9, 90.6]	0%	0%

Table 6: All cells, Gemini 2.5 Flash-Lite, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	127	64.6	[55.9, 72.4]	0%	0%
Telugu	dual	Indic-only	118	71.2	[62.7, 78.8]	50%	81%
Telugu	dual	Indic-only [bm25]	118	44.9	[35.6, 54.2]	51%	13%
Telugu	dual	both	119	74.8	[66.4, 82.4]	50%	89%
Telugu	dual	both [bm25]	119	61.3	[52.9, 69.7]	50%	60%
Telugu	dual	English-only	119	72.3	[63.9, 79.8]	50%	89%
Telugu	dual	English-only [bm25]	118	54.2	[44.9, 62.7]	51%	55%
Telugu	english	Indic-only	127	67.7	[59.1, 75.6]	100%	65%
Telugu	english	Indic-only [bm25]	127	44.1	[35.4, 52.8]	100%	5%
Telugu	english	both	127	74.8	[66.9, 82.7]	100%	87%
Telugu	english	both [bm25]	127	67.7	[59.8, 75.6]	100%	62%
Telugu	english	English-only	127	77.2	[70.1, 84.3]	100%	87%
Telugu	english	English-only [bm25]	127	66.9	[58.3, 74.8]	100%	61%
Telugu	free	Indic-only	127	71.7	[63.8, 79.5]	45%	72%
Telugu	free	Indic-only [bm25]	127	43.3	[34.6, 52.0]	51%	11%
Telugu	free	both	127	74.8	[66.9, 81.9]	44%	81%
Telugu	free	both [bm25]	127	53.5	[44.9, 62.2]	43%	40%
Telugu	free	English-only	127	76.4	[68.5, 83.5]	47%	86%
Telugu	free	English-only [bm25]	127	48.8	[40.2, 57.5]	46%	43%
Telugu	free	(gold hidden)	50	72.0	[60.0, 84.0]	40%	0%
Telugu	match	Indic-only	127	66.9	[58.3, 75.6]	0%	74%
Telugu	match	Indic-only [bm25]	127	37.8	[29.1, 46.5]	2%	11%
Telugu	match	both	127	66.9	[59.1, 74.8]	0%	76%
Telugu	match	both [bm25]	127	35.4	[27.6, 44.1]	2%	14%
Telugu	match	English-only	127	74.8	[66.9, 81.9]	3%	80%
Telugu	match	English-only [bm25]	127	15.0	[9.4, 21.3]	2%	4%
Telugu	oracle L	(gold passage)	127	70.9	[63.0, 78.7]	0%	0%
Telugu	oracle en	(gold passage)	127	78.7	[71.7, 85.8]	0%	0%
English	closed	(no tool)	127	68.5	[60.6, 76.4]	0%	0%
English	dual	Indic-only	127	64.6	[55.9, 72.4]	96%	66%
English	dual	Indic-only [bm25]	127	32.3	[24.4, 40.2]	96%	9%
English	free	Indic-only	127	70.1	[62.2, 78.0]	100%	65%
English	free	Indic-only [bm25]	126	43.7	[34.9, 52.4]	100%	7%
English	free	both	127	79.5	[72.4, 86.6]	100%	88%
English	free	both [bm25]	127	64.6	[55.9, 72.4]	100%	58%
English	free	English-only	127	78.7	[71.7, 85.8]	100%	88%
English	free	English-only [bm25]	127	66.9	[58.3, 74.8]	100%	59%
English	match	Indic-only	127	70.1	[62.2, 78.0]	100%	60%
English	match	Indic-only [bm25]	126	46.0	[37.3, 54.8]	100%	6%
English	match	both	127	80.3	[73.2, 87.4]	100%	87%
English	match	both [bm25]	127	68.5	[60.6, 76.4]	100%	61%
English	match	English-only	127	79.5	[72.4, 86.6]	100%	87%
English	match	English-only [bm25]	126	69.0	[61.1, 77.0]	100%	60%
English	oracle en	(gold passage)	127	85.0	[78.7, 91.3]	0%	0%

Table 7: All cells, Gemini 2.5 Flash-Lite, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	100	67.0	[58.0, 76.0]	0%	0%
Hindi	dual	Indic-only	100	73.0	[64.0, 81.0]	51%	86%
Hindi	dual	both	100	76.0	[67.0, 84.0]	50%	90%
Hindi	dual	English-only	100	78.0	[70.0, 86.0]	50%	90%
Hindi	english	Indic-only	100	75.0	[67.0, 83.0]	100%	75%
Hindi	english	both	100	77.0	[69.0, 85.0]	100%	86%
Hindi	english	English-only	100	77.0	[68.0, 85.0]	100%	86%
Hindi	free	Indic-only	99	70.7	[61.6, 79.8]	34%	79%
Hindi	free	both	100	75.0	[66.0, 83.0]	31%	83%
Hindi	free	English-only	100	77.0	[68.0, 85.0]	38%	83%
Hindi	match	Indic-only	100	75.0	[66.0, 83.0]	5%	79%
Hindi	match	both	99	75.8	[66.7, 83.8]	7%	82%
Hindi	match	English-only	100	74.0	[65.0, 82.0]	11%	81%
Hindi	oracle L	(gold passage)	100	84.0	[76.0, 91.0]	0%	0%
Hindi	oracle en	(gold passage)	100	89.0	[83.0, 95.0]	0%	0%
English	closed	(no tool)	100	70.0	[61.0, 79.0]	0%	0%
English	free	Indic-only	100	78.0	[69.0, 86.0]	100%	76%
English	free	both	100	78.0	[70.0, 86.0]	100%	86%
English	free	English-only	100	77.0	[68.0, 85.0]	97%	84%
English	oracle en	(gold passage)	100	87.0	[80.0, 93.0]	0%	0%

Table 8: All cells, Gemini 3.1 Flash-Lite, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	100	80.0	[72.0, 87.0]	0%	0%
Telugu	dual	Indic-only	100	76.0	[67.0, 84.0]	51%	78%
Telugu	dual	both	100	76.0	[67.0, 84.0]	50%	92%
Telugu	dual	English-only	100	75.0	[66.0, 83.0]	50%	90%
Telugu	english	Indic-only	100	80.0	[72.0, 87.0]	100%	67%
Telugu	english	both	100	78.0	[69.0, 86.0]	100%	87%
Telugu	english	English-only	100	78.0	[70.0, 86.0]	100%	87%
Telugu	free	Indic-only	100	79.0	[71.0, 86.0]	48%	76%
Telugu	free	both	99	78.8	[70.7, 86.9]	45%	85%
Telugu	free	English-only	99	77.8	[69.7, 85.9]	48%	84%
Telugu	match	Indic-only	100	79.0	[71.0, 87.0]	4%	74%
Telugu	match	both	100	77.0	[69.0, 85.0]	4%	76%
Telugu	match	English-only	100	75.0	[66.0, 83.0]	6%	83%
Telugu	oracle L	(gold passage)	100	86.0	[79.0, 92.0]	0%	0%
Telugu	oracle en	(gold passage)	100	90.0	[84.0, 95.0]	0%	0%
English	closed	(no tool)	100	79.0	[71.0, 87.0]	0%	0%
English	free	Indic-only	100	79.0	[71.0, 86.0]	100%	68%
English	free	both	100	80.0	[72.0, 87.0]	100%	91%
English	free	English-only	100	80.0	[72.0, 88.0]	100%	91%
English	oracle en	(gold passage)	100	89.0	[83.0, 95.0]	0%	0%

Table 9: All cells, Gemini 3.1 Flash-Lite, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	50	70.0	[58.0, 82.0]	0%	0%
Hindi	dual	both	50	74.0	[62.0, 86.0]	50%	88%
Hindi	dual	English-only	50	76.0	[64.0, 88.0]	53%	94%
Hindi	english	both	50	80.0	[68.0, 90.0]	98%	90%
Hindi	english	English-only	50	78.0	[66.0, 88.0]	100%	90%
Hindi	free	both	50	74.0	[62.0, 86.0]	45%	80%
Hindi	free	English-only	50	74.0	[62.0, 86.0]	57%	84%
Hindi	oracle L	(gold passage)	50	78.0	[66.0, 88.0]	0%	0%
Hindi	oracle en	(gold passage)	50	88.0	[78.0, 96.0]	0%	0%
English	closed	(no tool)	50	72.0	[60.0, 84.0]	0%	0%
English	free	both	50	76.0	[64.0, 88.0]	100%	90%
English	free	English-only	50	76.0	[64.0, 88.0]	100%	84%
English	oracle en	(gold passage)	50	86.0	[76.0, 94.0]	0%	0%

Table 10: All cells, Gemini 3.6 Flash, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	50	84.0	[74.0, 94.0]	0%	0%
Telugu	dual	both	50	82.0	[70.0, 92.0]	50%	94%
Telugu	dual	English-only	50	86.0	[76.0, 94.0]	51%	94%
Telugu	english	both	50	82.0	[70.0, 92.0]	98%	94%
Telugu	english	English-only	50	76.0	[64.0, 88.0]	100%	88%
Telugu	free	both	50	84.0	[74.0, 94.0]	45%	80%
Telugu	free	English-only	50	82.0	[70.0, 92.0]	44%	92%
Telugu	oracle L	(gold passage)	50	86.0	[76.0, 94.0]	0%	0%
Telugu	oracle en	(gold passage)	50	96.0	[90.0, 100.0]	0%	0%
English	closed	(no tool)	50	84.0	[74.0, 94.0]	0%	0%
English	free	both	50	78.0	[66.0, 88.0]	100%	84%
English	free	English-only	50	78.0	[66.0, 88.0]	100%	88%
English	oracle en	(gold passage)	50	94.0	[86.0, 100.0]	0%	0%

Table 11: All cells, Gemini 3.6 Flash, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	287	39.4	[33.8, 44.9]	0%	0%
Hindi	dual	Indic-only	287	65.2	[59.6, 70.7]	50%	79%
Hindi	dual	both	287	70.0	[64.5, 75.3]	50%	87%
Hindi	dual	both [bm25]	287	63.1	[57.5, 68.6]	50%	62%
Hindi	dual	English-only	287	67.6	[62.0, 72.8]	50%	85%
Hindi	dual	English-only [bm25]	287	58.5	[53.0, 64.5]	50%	57%
Hindi	english	Indic-only	287	59.6	[54.0, 65.2]	100%	64%
Hindi	english	both	287	64.1	[58.5, 69.7]	100%	79%
Hindi	english	English-only	287	63.1	[57.5, 68.6]	100%	78%
Hindi	free	Indic-only	287	64.1	[58.5, 69.3]	19%	65%
Hindi	free	both	287	65.5	[59.9, 71.1]	18%	69%
Hindi	free	both [bm25]	287	49.5	[43.9, 55.4]	19%	23%
Hindi	free	English-only	287	59.9	[54.4, 65.5]	19%	59%
Hindi	free	English-only [bm25]	287	40.8	[35.2, 46.3]	19%	11%
Hindi	match	Indic-only	287	63.8	[58.2, 69.3]	1%	72%
Hindi	match	both	287	65.9	[60.3, 71.4]	0%	74%
Hindi	match	both [bm25]	287	43.2	[37.6, 49.1]	0%	15%
Hindi	match	English-only	287	59.2	[53.7, 64.8]	0%	65%
Hindi	match	English-only [bm25]	287	37.3	[31.7, 42.9]	0%	0%
Hindi	oracle L	(gold passage)	287	81.9	[77.4, 86.1]	0%	0%
Hindi	oracle en	(gold passage)	287	86.4	[82.2, 90.2]	0%	0%
English	closed	(no tool)	287	52.3	[46.3, 58.2]	0%	0%
English	dual	Indic-only	287	70.7	[65.2, 76.0]	100%	74%
English	free	Indic-only	287	64.8	[59.2, 70.4]	100%	55%
English	free	both	287	74.6	[69.3, 79.4]	100%	71%
English	free	both [bm25]	287	69.0	[63.4, 74.2]	100%	53%
English	free	English-only	287	74.6	[69.7, 79.4]	100%	71%
English	free	English-only [bm25]	287	69.7	[64.5, 74.9]	100%	52%
English	match	Indic-only	287	67.9	[62.4, 73.2]	100%	64%
English	match	both	287	75.6	[70.7, 80.5]	100%	80%
English	match	English-only	287	77.0	[72.1, 81.9]	100%	80%
English	oracle en	(gold passage)	287	90.6	[87.1, 93.7]	0%	0%

Table 12: All cells, Qwen3-30B-A3B, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	127	48.0	[39.4, 56.7]	0%	0%
Telugu	dual	Indic-only	127	72.4	[64.6, 80.3]	50%	80%
Telugu	dual	both	127	76.4	[69.3, 83.5]	50%	93%
Telugu	dual	both [bm25]	127	63.0	[54.3, 70.9]	50%	62%
Telugu	dual	English-only	127	73.2	[65.4, 81.1]	50%	94%
Telugu	dual	English-only [bm25]	127	65.4	[56.7, 73.2]	50%	61%
Telugu	english	Indic-only	127	66.1	[58.3, 74.0]	100%	61%
Telugu	english	both	127	70.9	[63.0, 78.7]	100%	87%
Telugu	english	English-only	127	71.7	[63.8, 79.5]	100%	87%
Telugu	free	Indic-only	127	68.5	[60.6, 76.4]	43%	61%
Telugu	free	both	127	72.4	[64.6, 80.3]	46%	70%
Telugu	free	both [bm25]	127	53.5	[44.9, 62.2]	47%	28%
Telugu	free	English-only	127	70.1	[62.2, 78.0]	49%	71%
Telugu	free	English-only [bm25]	127	57.5	[48.8, 66.1]	46%	24%
Telugu	match	Indic-only	127	73.2	[65.4, 81.1]	1%	74%
Telugu	match	both	127	71.7	[63.8, 79.5]	2%	78%
Telugu	match	both [bm25]	126	41.3	[32.5, 50.0]	1%	9%
Telugu	match	English-only	127	69.3	[61.4, 77.2]	1%	77%
Telugu	match	English-only [bm25]	127	37.8	[29.1, 46.5]	1%	1%
Telugu	oracle L	(gold passage)	127	85.0	[78.7, 91.3]	0%	0%
Telugu	oracle en	(gold passage)	127	85.8	[79.5, 91.3]	0%	0%
English	closed	(no tool)	127	55.1	[46.5, 63.8]	0%	0%
English	dual	Indic-only	127	71.7	[63.8, 79.5]	100%	71%
English	free	Indic-only	127	66.1	[57.5, 74.0]	100%	47%
English	free	both	127	74.8	[67.7, 81.9]	100%	69%
English	free	both [bm25]	127	67.7	[59.1, 75.6]	100%	42%
English	free	English-only	127	74.0	[66.1, 81.1]	100%	70%
English	free	English-only [bm25]	127	68.5	[60.6, 76.4]	100%	43%
English	match	Indic-only	127	69.3	[61.4, 77.2]	100%	57%
English	match	both	127	76.4	[68.5, 83.5]	100%	79%
English	match	English-only	127	76.4	[69.3, 83.5]	100%	83%
English	oracle en	(gold passage)	127	92.1	[87.4, 96.1]	0%	0%

Table 13: All cells, Qwen3-30B-A3B, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Stage	USD
core	9.21
uncovered	1.35
judge	1.00
coverage2	0.63
oracle	0.22
native_te	0.22
coverage	0.19
closed	0.08
noise	0.08
none	0.03
pilot	0.03
filter	0.01
Total metered through OpenRouter (100808 calls)	13.04

Table 14: Metered API spend (exact per-call `usage.cost`). The open-weight arm ran on a rented A100 (about 2.5 GPU-hours) and the bge-m3 embeddings took about 0.5 GPU-hour; both outside the metered total.

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	287	57.8	[51.9, 63.4]	0%	0%
Hindi	dual	Indic-only	281	61.9	[56.2, 67.6]	50%	79%
Hindi	dual	Indic-only [bm25]	280	34.3	[28.6, 39.6]	52%	20%
Hindi	dual	both	280	67.9	[62.5, 73.6]	50%	87%
Hindi	dual	both [bm25]	281	58.7	[53.0, 64.4]	51%	67%
Hindi	dual	English-only	280	67.9	[62.1, 73.2]	50%	83%
Hindi	dual	English-only [bm25]	280	58.2	[52.5, 63.9]	51%	61%
Hindi	english	Indic-only	287	61.3	[55.7, 66.9]	100%	68%
Hindi	english	Indic-only [bm25]	285	32.3	[27.0, 37.9]	100%	8%
Hindi	english	both	287	70.7	[65.5, 76.0]	100%	80%
Hindi	english	both [bm25]	287	63.4	[57.8, 69.0]	100%	63%
Hindi	english	English-only	287	70.7	[65.5, 76.0]	100%	80%
Hindi	english	English-only [bm25]	287	62.7	[57.1, 68.3]	100%	59%
Hindi	free	Indic-only	287	65.9	[60.3, 71.4]	32%	74%
Hindi	free	Indic-only [bm25]	287	40.1	[34.5, 45.6]	38%	17%
Hindi	free	both	287	66.2	[60.6, 71.4]	30%	79%
Hindi	free	both [bm25]	286	50.3	[44.7, 56.3]	28%	40%
Hindi	free	English-only	287	66.9	[61.3, 72.5]	33%	72%
Hindi	free	English-only [bm25]	286	43.4	[37.8, 49.3]	35%	38%
Hindi	free	(gold hidden)	49	55.1	[40.8, 69.4]	27%	0%
Hindi	match	Indic-only	286	66.4	[60.8, 71.7]	1%	74%
Hindi	match	Indic-only [bm25]	286	35.3	[29.7, 40.9]	1%	16%
Hindi	match	both	287	66.9	[61.7, 72.5]	1%	78%
Hindi	match	both [bm25]	287	43.6	[38.0, 49.5]	2%	23%
Hindi	match	English-only	286	64.3	[58.7, 69.9]	1%	66%
Hindi	match	English-only [bm25]	286	19.6	[15.0, 24.5]	2%	3%
Hindi	oracle L	(gold passage)	287	72.8	[67.6, 78.0]	0%	0%
Hindi	oracle en	(gold passage)	287	80.5	[76.0, 85.0]	0%	0%
English	closed	(no tool)	287	61.0	[55.4, 66.6]	0%	0%
English	dual	Indic-only	287	64.5	[58.9, 70.0]	97%	72%
English	dual	Indic-only [bm25]	287	25.8	[20.9, 31.0]	97%	8%
English	free	Indic-only	287	65.9	[60.3, 71.1]	100%	70%
English	free	Indic-only [bm25]	286	32.2	[26.6, 37.8]	100%	8%
English	free	both	286	75.9	[71.0, 80.8]	100%	83%
English	free	both [bm25]	287	69.0	[63.4, 74.2]	100%	65%
English	free	English-only	285	75.8	[70.9, 80.7]	100%	82%
English	free	English-only [bm25]	287	71.4	[66.2, 76.7]	100%	65%
English	match	Indic-only	287	64.8	[59.2, 70.4]	100%	70%
English	match	Indic-only [bm25]	287	38.3	[32.8, 43.9]	100%	8%
English	match	both	287	74.9	[70.0, 79.8]	100%	82%
English	match	both [bm25]	287	71.4	[66.2, 76.7]	100%	65%
English	match	English-only	287	76.3	[71.4, 81.2]	100%	82%
English	match	English-only [bm25]	287	70.7	[65.5, 76.0]	100%	64%
English	oracle en	(gold passage)	287	86.8	[82.9, 90.6]	0%	0%

Table 15: All cells, Gemini 2.5 Flash-Lite, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	127	64.6	[55.9, 72.4]	0%	0%
Telugu	dual	Indic-only	118	71.2	[62.7, 78.8]	50%	81%
Telugu	dual	Indic-only [bm25]	118	44.9	[35.6, 54.2]	51%	13%
Telugu	dual	both	119	74.8	[66.4, 82.4]	50%	89%
Telugu	dual	both [bm25]	119	61.3	[52.9, 69.7]	50%	60%
Telugu	dual	English-only	119	72.3	[63.9, 79.8]	50%	89%
Telugu	dual	English-only [bm25]	118	54.2	[44.9, 62.7]	51%	55%
Telugu	english	Indic-only	127	67.7	[59.1, 75.6]	100%	65%
Telugu	english	Indic-only [bm25]	127	44.1	[35.4, 52.8]	100%	5%
Telugu	english	both	127	74.8	[66.9, 82.7]	100%	87%
Telugu	english	both [bm25]	127	67.7	[59.8, 75.6]	100%	62%
Telugu	english	English-only	127	77.2	[70.1, 84.3]	100%	87%
Telugu	english	English-only [bm25]	127	66.9	[58.3, 74.8]	100%	61%
Telugu	free	Indic-only	127	71.7	[63.8, 79.5]	45%	72%
Telugu	free	Indic-only [bm25]	127	43.3	[34.6, 52.0]	51%	11%
Telugu	free	both	127	74.8	[66.9, 81.9]	44%	81%
Telugu	free	both [bm25]	127	53.5	[44.9, 62.2]	43%	40%
Telugu	free	English-only	127	76.4	[68.5, 83.5]	47%	86%
Telugu	free	English-only [bm25]	127	48.8	[40.2, 57.5]	46%	43%
Telugu	free	(gold hidden)	50	72.0	[60.0, 84.0]	40%	0%
Telugu	match	Indic-only	127	66.9	[58.3, 75.6]	0%	74%
Telugu	match	Indic-only [bm25]	127	37.8	[29.1, 46.5]	2%	11%
Telugu	match	both	127	66.9	[59.1, 74.8]	0%	76%
Telugu	match	both [bm25]	127	35.4	[27.6, 44.1]	2%	14%
Telugu	match	English-only	127	74.8	[66.9, 81.9]	3%	80%
Telugu	match	English-only [bm25]	127	15.0	[9.4, 21.3]	2%	4%
Telugu	oracle L	(gold passage)	127	70.9	[63.0, 78.7]	0%	0%
Telugu	oracle en	(gold passage)	127	78.7	[71.7, 85.8]	0%	0%
English	closed	(no tool)	127	68.5	[60.6, 76.4]	0%	0%
English	dual	Indic-only	127	64.6	[55.9, 72.4]	96%	66%
English	dual	Indic-only [bm25]	127	32.3	[24.4, 40.2]	96%	9%
English	free	Indic-only	127	70.1	[62.2, 78.0]	100%	65%
English	free	Indic-only [bm25]	126	43.7	[34.9, 52.4]	100%	7%
English	free	both	127	79.5	[72.4, 86.6]	100%	88%
English	free	both [bm25]	127	64.6	[55.9, 72.4]	100%	58%
English	free	English-only	127	78.7	[71.7, 85.8]	100%	88%
English	free	English-only [bm25]	127	66.9	[58.3, 74.8]	100%	59%
English	match	Indic-only	127	70.1	[62.2, 78.0]	100%	60%
English	match	Indic-only [bm25]	126	46.0	[37.3, 54.8]	100%	6%
English	match	both	127	80.3	[73.2, 87.4]	100%	87%
English	match	both [bm25]	127	68.5	[60.6, 76.4]	100%	61%
English	match	English-only	127	79.5	[72.4, 86.6]	100%	87%
English	match	English-only [bm25]	126	69.0	[61.1, 77.0]	100%	60%
English	oracle en	(gold passage)	127	85.0	[78.7, 91.3]	0%	0%

Table 16: All cells, Gemini 2.5 Flash-Lite, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	100	67.0	[58.0, 76.0]	0%	0%
Hindi	dual	Indic-only	100	73.0	[64.0, 81.0]	51%	86%
Hindi	dual	both	100	76.0	[67.0, 84.0]	50%	90%
Hindi	dual	English-only	100	78.0	[70.0, 86.0]	50%	90%
Hindi	english	Indic-only	100	75.0	[67.0, 83.0]	100%	75%
Hindi	english	both	100	77.0	[69.0, 85.0]	100%	86%
Hindi	english	English-only	100	77.0	[68.0, 85.0]	100%	86%
Hindi	free	Indic-only	99	70.7	[61.6, 79.8]	34%	79%
Hindi	free	both	100	75.0	[66.0, 83.0]	31%	83%
Hindi	free	English-only	100	77.0	[68.0, 85.0]	38%	83%
Hindi	match	Indic-only	100	75.0	[66.0, 83.0]	5%	79%
Hindi	match	both	99	75.8	[66.7, 83.8]	7%	82%
Hindi	match	English-only	100	74.0	[65.0, 82.0]	11%	81%
Hindi	oracle L	(gold passage)	100	84.0	[76.0, 91.0]	0%	0%
Hindi	oracle en	(gold passage)	100	89.0	[83.0, 95.0]	0%	0%
English	closed	(no tool)	100	70.0	[61.0, 79.0]	0%	0%
English	free	Indic-only	100	78.0	[69.0, 86.0]	100%	76%
English	free	both	100	78.0	[70.0, 86.0]	100%	86%
English	free	English-only	100	77.0	[68.0, 85.0]	97%	84%
English	oracle en	(gold passage)	100	87.0	[80.0, 93.0]	0%	0%

Table 17: All cells, Gemini 3.1 Flash-Lite, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	100	80.0	[72.0, 87.0]	0%	0%
Telugu	dual	Indic-only	100	76.0	[67.0, 84.0]	51%	78%
Telugu	dual	both	100	76.0	[67.0, 84.0]	50%	92%
Telugu	dual	English-only	100	75.0	[66.0, 83.0]	50%	90%
Telugu	english	Indic-only	100	80.0	[72.0, 87.0]	100%	67%
Telugu	english	both	100	78.0	[69.0, 86.0]	100%	87%
Telugu	english	English-only	100	78.0	[70.0, 86.0]	100%	87%
Telugu	free	Indic-only	100	79.0	[71.0, 86.0]	48%	76%
Telugu	free	both	99	78.8	[70.7, 86.9]	45%	85%
Telugu	free	English-only	99	77.8	[69.7, 85.9]	48%	84%
Telugu	match	Indic-only	100	79.0	[71.0, 87.0]	4%	74%
Telugu	match	both	100	77.0	[69.0, 85.0]	4%	76%
Telugu	match	English-only	100	75.0	[66.0, 83.0]	6%	83%
Telugu	oracle L	(gold passage)	100	86.0	[79.0, 92.0]	0%	0%
Telugu	oracle en	(gold passage)	100	90.0	[84.0, 95.0]	0%	0%
English	closed	(no tool)	100	79.0	[71.0, 87.0]	0%	0%
English	free	Indic-only	100	79.0	[71.0, 86.0]	100%	68%
English	free	both	100	80.0	[72.0, 87.0]	100%	91%
English	free	English-only	100	80.0	[72.0, 88.0]	100%	91%
English	oracle en	(gold passage)	100	89.0	[83.0, 95.0]	0%	0%

Table 18: All cells, Gemini 3.1 Flash-Lite, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	50	70.0	[58.0, 82.0]	0%	0%
Hindi	dual	both	50	74.0	[62.0, 86.0]	50%	88%
Hindi	dual	English-only	50	76.0	[64.0, 88.0]	53%	94%
Hindi	english	both	50	80.0	[68.0, 90.0]	98%	90%
Hindi	english	English-only	50	78.0	[66.0, 88.0]	100%	90%
Hindi	free	both	50	74.0	[62.0, 86.0]	45%	80%
Hindi	free	English-only	50	74.0	[62.0, 86.0]	57%	84%
Hindi	oracle L	(gold passage)	50	78.0	[66.0, 88.0]	0%	0%
Hindi	oracle en	(gold passage)	50	88.0	[78.0, 96.0]	0%	0%
English	closed	(no tool)	50	72.0	[60.0, 84.0]	0%	0%
English	free	both	50	76.0	[64.0, 88.0]	100%	90%
English	free	English-only	50	76.0	[64.0, 88.0]	100%	84%
English	oracle en	(gold passage)	50	86.0	[76.0, 94.0]	0%	0%

Table 19: All cells, Gemini 3.6 Flash, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	50	84.0	[74.0, 94.0]	0%	0%
Telugu	dual	both	50	82.0	[70.0, 92.0]	50%	94%
Telugu	dual	English-only	50	86.0	[76.0, 94.0]	51%	94%
Telugu	english	both	50	82.0	[70.0, 92.0]	98%	94%
Telugu	english	English-only	50	76.0	[64.0, 88.0]	100%	88%
Telugu	free	both	50	84.0	[74.0, 94.0]	45%	80%
Telugu	free	English-only	50	82.0	[70.0, 92.0]	44%	92%
Telugu	oracle L	(gold passage)	50	86.0	[76.0, 94.0]	0%	0%
Telugu	oracle en	(gold passage)	50	96.0	[90.0, 100.0]	0%	0%
English	closed	(no tool)	50	84.0	[74.0, 94.0]	0%	0%
English	free	both	50	78.0	[66.0, 88.0]	100%	84%
English	free	English-only	50	78.0	[66.0, 88.0]	100%	88%
English	oracle en	(gold passage)	50	94.0	[86.0, 100.0]	0%	0%

Table 20: All cells, Gemini 3.6 Flash, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Hindi	closed	(no tool)	287	39.4	[33.8, 44.9]	0%	0%
Hindi	dual	Indic-only	287	65.2	[59.6, 70.7]	50%	79%
Hindi	dual	both	287	70.0	[64.5, 75.3]	50%	87%
Hindi	dual	both [bm25]	287	63.1	[57.5, 68.6]	50%	62%
Hindi	dual	English-only	287	67.6	[62.0, 72.8]	50%	85%
Hindi	dual	English-only [bm25]	287	58.5	[53.0, 64.5]	50%	57%
Hindi	english	Indic-only	287	59.6	[54.0, 65.2]	100%	64%
Hindi	english	both	287	64.1	[58.5, 69.7]	100%	79%
Hindi	english	English-only	287	63.1	[57.5, 68.6]	100%	78%
Hindi	free	Indic-only	287	64.1	[58.5, 69.3]	19%	65%
Hindi	free	both	287	65.5	[59.9, 71.1]	18%	69%
Hindi	free	both [bm25]	287	49.5	[43.9, 55.4]	19%	23%
Hindi	free	English-only	287	59.9	[54.4, 65.5]	19%	59%
Hindi	free	English-only [bm25]	287	40.8	[35.2, 46.3]	19%	11%
Hindi	match	Indic-only	287	63.8	[58.2, 69.3]	1%	72%
Hindi	match	both	287	65.9	[60.3, 71.4]	0%	74%
Hindi	match	both [bm25]	287	43.2	[37.6, 49.1]	0%	15%
Hindi	match	English-only	287	59.2	[53.7, 64.8]	0%	65%
Hindi	match	English-only [bm25]	287	37.3	[31.7, 42.9]	0%	0%
Hindi	oracle L	(gold passage)	287	81.9	[77.4, 86.1]	0%	0%
Hindi	oracle en	(gold passage)	287	86.4	[82.2, 90.2]	0%	0%
English	closed	(no tool)	287	52.3	[46.3, 58.2]	0%	0%
English	dual	Indic-only	287	70.7	[65.2, 76.0]	100%	74%
English	free	Indic-only	287	64.8	[59.2, 70.4]	100%	55%
English	free	both	287	74.6	[69.3, 79.4]	100%	71%
English	free	both [bm25]	287	69.0	[63.4, 74.2]	100%	53%
English	free	English-only	287	74.6	[69.7, 79.4]	100%	71%
English	free	English-only [bm25]	287	69.7	[64.5, 74.9]	100%	52%
English	match	Indic-only	287	67.9	[62.4, 73.2]	100%	64%
English	match	both	287	75.6	[70.7, 80.5]	100%	80%
English	match	English-only	287	77.0	[72.1, 81.9]	100%	80%
English	oracle en	(gold passage)	287	90.6	[87.1, 93.7]	0%	0%

Table 21: All cells, Qwen3-30B-A3B, Hindi items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Question	Policy	Pool	n	Acc	95% CI	English queries	gold in top-5
Telugu	closed	(no tool)	127	48.0	[39.4, 56.7]	0%	0%
Telugu	dual	Indic-only	127	72.4	[64.6, 80.3]	50%	80%
Telugu	dual	both	127	76.4	[69.3, 83.5]	50%	93%
Telugu	dual	both [bm25]	127	63.0	[54.3, 70.9]	50%	62%
Telugu	dual	English-only	127	73.2	[65.4, 81.1]	50%	94%
Telugu	dual	English-only [bm25]	127	65.4	[56.7, 73.2]	50%	61%
Telugu	english	Indic-only	127	66.1	[58.3, 74.0]	100%	61%
Telugu	english	both	127	70.9	[63.0, 78.7]	100%	87%
Telugu	english	English-only	127	71.7	[63.8, 79.5]	100%	87%
Telugu	free	Indic-only	127	68.5	[60.6, 76.4]	43%	61%
Telugu	free	both	127	72.4	[64.6, 80.3]	46%	70%
Telugu	free	both [bm25]	127	53.5	[44.9, 62.2]	47%	28%
Telugu	free	English-only	127	70.1	[62.2, 78.0]	49%	71%
Telugu	free	English-only [bm25]	127	57.5	[48.8, 66.1]	46%	24%
Telugu	match	Indic-only	127	73.2	[65.4, 81.1]	1%	74%
Telugu	match	both	127	71.7	[63.8, 79.5]	2%	78%
Telugu	match	both [bm25]	126	41.3	[32.5, 50.0]	1%	9%
Telugu	match	English-only	127	69.3	[61.4, 77.2]	1%	77%
Telugu	match	English-only [bm25]	127	37.8	[29.1, 46.5]	1%	1%
Telugu	oracle L	(gold passage)	127	85.0	[78.7, 91.3]	0%	0%
Telugu	oracle en	(gold passage)	127	85.8	[79.5, 91.3]	0%	0%
English	closed	(no tool)	127	55.1	[46.5, 63.8]	0%	0%
English	dual	Indic-only	127	71.7	[63.8, 79.5]	100%	71%
English	free	Indic-only	127	66.1	[57.5, 74.0]	100%	47%
English	free	both	127	74.8	[67.7, 81.9]	100%	69%
English	free	both [bm25]	127	67.7	[59.1, 75.6]	100%	42%
English	free	English-only	127	74.0	[66.1, 81.1]	100%	70%
English	free	English-only [bm25]	127	68.5	[60.6, 76.4]	100%	43%
English	match	Indic-only	127	69.3	[61.4, 77.2]	100%	57%
English	match	both	127	76.4	[68.5, 83.5]	100%	79%
English	match	English-only	127	76.4	[69.3, 83.5]	100%	83%
English	oracle en	(gold passage)	127	92.1	[87.4, 96.1]	0%	0%

Table 22: All cells, Qwen3-30B-A3B, Telugu items (n = items; ‘English queries’ = share of search calls written in Latin script; ‘gold in top-5’ = share of episodes in which a gold chunk was retrieved).

Stage	USD
core	9.21
uncovered	1.35
judge	1.00
coverage2	0.63
oracle	0.22
native_te	0.22
coverage	0.19
closed	0.08
noise	0.08
none	0.03
pilot	0.03
filter	0.01
Total metered through OpenRouter (100808 calls)	13.04

Table 23: Metered API spend (exact per-call `usage.cost`). The open-weight arm ran on a rented A100 (about 2.5 GPU-hours) and the bge-m3 embeddings took about 0.5 GPU-hour; both outside the metered total.