Beyond Chunking: Node Integrity and Field Structure in Bengali Retrieval-Augmented Agent Memory

Khan Raiyan Ibne Reza[1], **Sanjana Aktar Maria**[1], **Sumaiya Tabassum Nimi**[1]

[1]North South University, Dhaka, Bangladesh

Correspondence: {raiyan.reza,sanjana.maria}@northsouth.edu

Abstract

Lexical retrieval in low-resource languages can fail when queries use symptoms, colloquial names, or procedural descriptions rather than formal terminology. Chunking can compound this problem by separating query-matching content from the field that identifies a record. We introduce a three-level representation ladder (chunked text, intact full-node text, and field-aware lexical indexing) that separates recovery associated with record integrity from the additional effect of field-aware representation. In the OCR-heavy education corpus, intact records raise Recall@10 from 37.19% to 89.65%. Field-aware indexing adds gains with fixed weights in agriculture and uniform weights in education, but uniform weighting underperforms intact text in law. Ablations indicate that useful fields depend on query formulation; full field-value shuffling reduces Recall@10 to 0.75–5.09%. In agent memory, Recall@5 rises from 27.6% to 61.9% with intact records on 583 agriculture and education queries. Across 872 queries in all three domains, uniform fielded memory has a citation rate about 12 points higher than intact memory, without higher aggregate Recall@5. These results indicate that record integrity and query-aligned content warrant separate evaluation in retrieval and agent memory.

1 Introduction

A farmer in rural Bangladesh may not type "Fusarium wilt of tomato" into a search box; she describes yellowing leaves, a plant wilting despite regular watering, or a smell in the soil. A student preparing for a lab exam may remember the apparatus and steps rather than the formal name of the experiment, while a citizen asking about a law may describe the act as people around her do rather than quote its section number. In each case, retrieval must connect the user's description to a knowledge record expressed in more formal terminology.

Most retrieval-augmented generation (RAG) systems bridge this gap by splitting source documents into fixed-size text chunks and indexing each independently. This becomes problematic when knowledge is naturally organized as a record: a disease name with its symptoms and treatment, an act title with its provisions, or an experiment name with its procedure. A 500-character window can place the symptom a user searches for in one chunk and the disease name that identifies the record in another. The query and the record are then separated before retrieval begins, and downstream reasoning cannot recover what was never retrieved together.

This issue is particularly relevant to Bengali domain retrieval, where query language can differ substantially from formal documents. Regional and colloquial variation is common (Faria et al., 2023; Sami et al., 2026), Bengali remains underresourced relative to its number of speakers (Joshi et al., 2020), and recent work shows that RAG performance can be sensitive to variation between query language and underlying text (Cao et al., 2026b). Bengali therefore provides a useful setting for asking whether the representation of a knowledge record itself affects retrieval.

A natural approach is to preserve record structure and index its fields directly. We treat fielded records as a lightweight, graph-free form of structured memory: each record is a typed unit, but we model no relations between records. GraphRAG and HippoRAG add such relations, which we do not evaluate (Edge et al., 2024; Gutiérrez et al., 2024). However, comparing chunked retrieval with fielded retrieval alone conflates two effects: recovery from splitting a record, and the combined effect of adding field copies and field weights. How much retrieval performance can be recovered simply by keeping a record intact? What additional benefit comes from adding field content and weights?

We answer these questions with a three-level representation ladder applied to the same underlying

Proceedings of the 3rd Workshop on KNOwledge GRaphs & Agentic Systems Interplay (NORA 2026), pages 84–97
November 10, 2026 ©2026 Association for Computational Linguistics

knowledge. At L_1, each record is split into 500-character chunks. At L_2, the full record is kept as one unstructured document, removing chunk boundaries without exposing field information. At L_3, selected field content is added using BM25F-inspired scoring (Robertson et al., 2004). Thus, $L_1 \rightarrow L_2$ measures recovery associated with preserving an intact record, while $L_2 \rightarrow L_3$ measures the combined effect of adding field copies and field weights. We apply the ladder to three Bengali domains with different query regimes: agricultural disease advisory, where queries are predominantly symptom-oriented but may name diseases or pests; statutory law, where users identify a known act; and secondary science education, where users describe experimental procedures. Across 3,670 queries and 1,953 knowledge nodes, the effects are substantial but not uniform. In education, Recall@10 rises from 37.19% at L_1 to 89.65% at L_2, with uniform BM25F adding 1.93 points. In agriculture, both steps help under fixed nonuniform field weights. In law, however, full-node retrieval reaches 79.55%, while uniform BM25F reaches 77.82%.

The remaining question is what makes field structure useful when it does help. Field-group ablations indicate that fields are useful when their surface forms match how users phrase their queries; full field-value shuffling shows that correct content assignment matters. We call such a field a *query-aligned anchor*. In law, `act_title_bn` alone reaches 100.00% Recall@10, whereas `act_no_bn`, which identifies the same acts, reaches only 2.18%. Shuffling all field values across records collapses Recall@10 to 0.75–5.09%, indicating that the observed retrieval signal depends on the content assigned to fields, rather than field boundaries alone.

Finally, we test whether representation matters when the knowledge store is used as agent memory. Poor retrieval can leave an agent without relevant evidence (Gutiérrez et al., 2024; Xu et al., 2025). We hold the agent controller fixed and vary its memory representation. On 583 agriculture and education queries, Recall@5 increases from 27.6% with chunked memory to 61.9% with full-node memory. Legal chunked-agent results have an identifier mismatch and are excluded from this comparison. On the 872-query three-domain subset, uniform fielded memory has a higher citation rate than full-node memory, while benchmark-informed weights provide an exploratory retrieval ceiling.

The results support a narrower conclusion: a substantial portion of the gain attributed to structured retrieval can arise from preserving record integrity, while the additional benefit of field-weighted retrieval depends on the scoring configuration and query-aligned content. Concretely, we contribute:

1. **A representation ladder** that distinguishes recovery from preserving intact records from the combined effect of field copies and field weights, evaluated across three Bengali domains and 3,670 queries.
2. **A field-level account of when structure helps**, using group ablations and counterfactual field permutations to examine how retrieval depends on query-aligned content and varies with the domain's query regime.
3. **A transfer of this representation ladder to agent memory**, examining representation-level recovery and, in an exploratory perturbation, sensitivity to query-aligned fields.

2 Related Work

Chunking, structured RAG, and fielded retrieval. Conventional RAG pipelines commonly split source documents into independently retrievable chunks, but recent work shows that this choice can affect retrieval quality. CausalRAG identifies disrupted contextual integrity as a consequence of chunking and introduces causal graphs to preserve relationships across the resulting fragments (Wang et al., 2025). DyVo studies a related fragmentation problem inside the tokenizer, where entity mentions are split across subword units; our setting is different in that a text splitter separates parts of an already structured record (Nguyen et al., 2024). Other approaches add structure after retrieval: StructRAG reconstructs retrieved passages into task-specific structures, while SRAG augments queries and retrieved chunks with structured information (Li et al., 2025b; Shah et al., 2026). These approaches improve or repair retrieval pipelines, but do not isolate how much of the gain comes from recovering record integrity versus exposing structure itself. Our representation ladder keeps the source knowledge fixed and compares these two representation changes; their remaining scoring and granularity confounds are discussed in the Limitations.

Fielded retrieval provides a complementary way to represent records. BM25F assigns field-specific weights rather than treating a document as a single field (Robertson et al., 2004), while mFAR learns

field importance conditioned on the query (Li et al., 2025a). In our setting, the fields already exist as curated attributes of each domain record, so the question is whether adding field content and weights provides value beyond intact text, and which fields carry that value.

Knowledge graphs and structured agent memory. Graph-based retrieval represents knowledge through explicit entities and relations rather than flat text. GraphRAG builds an entity graph for corpus-level sensemaking, while HippoRAG uses a knowledge graph and personalized PageRank for long-term multi-hop retrieval (Edge et al., 2024; Gutiérrez et al., 2024). Recent work also explores structured linked representations for agent-oriented retrieval (Sade et al., 2026; Volpini et al., 2026). These approaches introduce an additional graph-construction layer, whereas fielded indexing can preserve curated record boundaries and surface forms without constructing a graph. Accordingly, we view graph-based and fielded representations as distinct points on a representation continuum rather than competing solutions.

Agent memory and memory retrieval. Agent-memory research has focused largely on how memories are created, linked, consolidated, and retrieved over time. A-MEM dynamically links and updates memories, MemBench evaluates memory across factual and reflective settings, and other systems explore hierarchical or lightweight long-term memory and structured consolidation (Xu et al., 2025; Tan et al., 2025; Cao et al., 2026a; Zhang et al., 2026; Chhikara et al., 2025). These works vary the memory-management mechanism; we hold the memory controller fixed and vary only the representation of static domain knowledge inside the retrievable memory. This addresses a lower-level question: whether the knowledge entering an agent's memory is represented in a form that retrieval can reliably use.

Bengali and low-resource retrieval. Bengali is a useful setting because the language used in domain queries can differ substantially from formal source text. Dialect and register variation produce measurable performance gaps, and resources such as Vashantor provide broad coverage of regional Bengali variation (Kantharuban et al., 2023; Faria et al., 2023; Sami et al., 2026). Recent Bengali agricultural retrieval evaluations further show that recall depends on query type, language condition, and indexing choices such as passage length (Reza et al., 2026a); KrishokChat provides a complementary citation-grounded agricultural knowledge base derived from authoritative sources (Reza et al., 2026b). These resources establish the difficulty and sensitivity of Bengali domain retrieval, but do not isolate how much of that difficulty is attributable to the representation of the knowledge itself.

Position of this work. Prior work thus shows that chunking can damage retrieval, fields can be weighted, graphs can encode relations, and agent memory can be organized in different ways. What remains unclear is how much of the apparent benefit of structured retrieval comes simply from preserving an intact record, and what additional benefit comes from adding field content and weights. We address this distinction with a representation ladder across three Bengali domains with different query regimes, and then test whether the same representation effects persist when the knowledge is consumed as memory by a retrieval-grounded agent.

3 Benchmark and Knowledge Representation

We evaluate three Bengali knowledge domains that differ in how users refer to information: legal queries identify a known act, agricultural queries are predominantly symptom-oriented, and education queries describe experimental procedures. For each domain, the same knowledge is represented at all three levels, giving 3,670 queries over 1,953 knowledge nodes.

The agricultural corpus contains 1,028 expert-curated disease and pest-management records, evaluated with 2,000 queries covering standard Dhaka, Rajshahi-dialect, and additional regional variants drawn using Vashantor as a reference (Reza et al., 2026a,b; Faria et al., 2023). Many queries describe symptoms, while others name a disease or pest; the disease-name field therefore tests lexical alignment across a mixed advisory query regime. The legal corpus contains 735 statutory acts drawn from the Bangladesh Legal Acts Dataset (Sakhawat, 2024); the 1,100 evaluation queries are derived from open-source Bengali legal QA resources and citizen-oriented statutory inquiries (Millat, 2026), with each query mapped to a validated target act. This known-item regime allows us to distinguish an identifier field from one whose surface form appears in queries. The education corpus contains

Property	Agriculture	Legal	Education
Knowledge Type	Crop disease/pest advisory	Bangladesh statutory law	Secondary science curriculum
Corpus Nodes	1,028	735	190
Unstructured Chunks	4,526	4,202	892
Evaluation Queries	2,000	1,100	570
Dialect/Register Coverage	16 regional + Standard	15 regional + Standard	5 registers + Standard
Record-Identifying Field	`disease_pest_name_bn`	`act_title_bn`	`experiment_name_bn`
Query Regime	Symptom-diagnostic	Known-item (act cited)	Procedural description

Table 1: Summary of the evaluation corpora. The query regime describes how users refer to the target information: legal queries identify a specific act, agricultural queries are predominantly symptom-oriented, and education queries describe experimental procedures.

190 secondary-science experiments from NCTB textbooks, with 570 procedural queries describing apparatus or steps without naming the experiment. Its 892 OCR-derived chunks make it a particularly direct test of fragmentation. The candidate record-identifying fields are `disease_pest_name_bn`, `act_title_bn`, and `experiment_name_bn`, respectively; their query alignment is examined in Section 6.

Query provenance. We evaluate domain-specific benchmark queries with fixed query-to-target mappings. Agricultural queries draw on KrishokChat and related advisory resources, with Vashantor as a reference for regional formulations. Legal queries draw on Bengali legal QA resources and citizen-oriented inquiries, and education queries are based on NCTB science textbooks. Source-derived query formulations may overlap lexically with target records (Appendix C).

We use a three-level representation ladder in which the underlying knowledge is unchanged while its representation and indexing granularity vary. At L_1 (Chunked Unstructured), each record is split into 500-character windows using recursive character splitting and the resulting chunks are indexed independently. At L_2 (Full-Node Unstructured), the intact record text is indexed as one document without exposing individual fields. At L_3, the intact full-text field is retained alongside individual record fields. We write "BM25F" for two BM25F-inspired implementations; neither merges term frequencies across fields as in Robertson et al. (2004). Legal and education standalone retrieval repeat field text in one BM25 document, so no field boundaries reach the scorer. Agriculture standalone retrieval and agent memory instead sum weighted per-field BM25 scores. The main agriculture result uses fixed nonuniform weights (`disease`: 3.5, `aliases`: 2.0, `symptoms`: 1.5, `content`: 1.0; other fields zeroed) set a priori from domain heuristics rather than tuned on benchmark queries; legal and education use unit weights. Implementation details are summarized in Appendix A; benchmark-fitted ceilings are separate (Appendix E).

Representation	Agri	Legal	Edu
L_1: Chunks	62.30%	73.18%	37.19%
L_2: Intact records	67.60%	79.55%	89.65%
L_3: Field-weighted	69.90%[a]	77.82%[b]	91.58%[b]

Table 2: Recall@10 across the three representation levels (ladder defined in Figure 1). [a]Fixed nonuniform sum of field scores set a priori; [b]uniform field-text repetition. All steps are significant under paired bootstrap (5,000 draws, $p \leq 0.0134$); the legal $L_2 \rightarrow L_3$ change is a *significant decrease* (95% CI $[-3.18, -0.27]$ pp). Full statistics in Appendix G.

Thus, L_3 adds individual field content to auxiliary full text, duplicating overlapping record content and changing scoring. A record spans 4.4, 5.7, and 4.7 L_1 retrieval units on average in agriculture, law, and education (Table 1). L_1 is a deliberate stress test of fixed-window retrieval; legal units are annotated passages. Stronger chunking baselines remain untested.

4 Core Retrieval Results

We compare the three representation levels using Recall@10. The primary comparison is L_1 vs. L_2 vs. L_3, distinguishing record-integrity recovery from the combined effect of field copies and weights. Tuned BM25F is treated only as a benchmark-fitted ceiling and is reported in Appendix E.

Table 2 shows that preserving the complete record produces the largest and most consistent recovery. In education, Recall@10 rises from 37.19% with chunked BM25 to 89.65% with full-node BM25, a gain of 52.46 points. Uniform BM25F

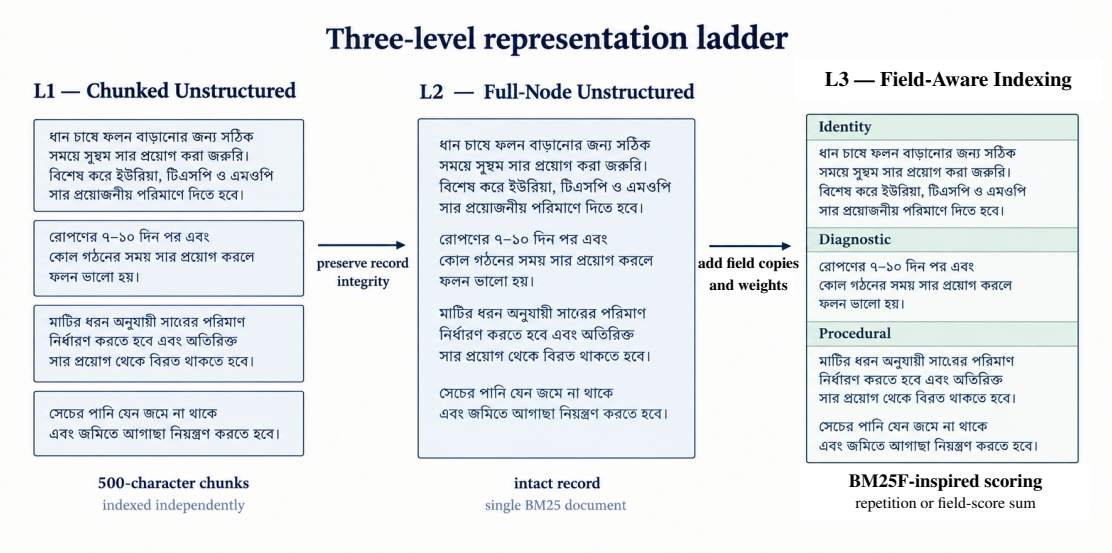


Figure 1: The three knowledge representations used in our experiments. L_1 splits each record into fixed-length chunks, L_2 keeps the record intact as a single unstructured document, and L_3 adds selected field content with BM25F-inspired scoring. The $L_1 \to L_2$ comparison measures recovery associated with preserving record integrity, while $L_2 \to L_3$ measures the combined effect of adding field copies and field weights. The field groups are schematic; L_3 also retains auxiliary full text.

adds a further 1.93 points, reaching 91.58%. The $L_1 \to L_2$ recovery is largest in the OCR-derived education corpus, where information from a single experiment can be split across chunk boundaries.

Agriculture follows the same ordering, but with smaller gains: Recall@10 increases from 62.30% to 67.60% at L_2 (+5.30 points), then to 69.90% at fixed-weight L_3 (+2.30 points). Here, both preserving the record and adding field content with fixed weights improve retrieval.

Legal retrieval is the notable exception. Full-node BM25 improves Recall@10 from 73.18% to 79.55%, whereas uniform BM25F reaches 77.82%. The paired bootstrap interval excludes zero ($\Delta = -1.73$ pp, 95% CI $[-3.18, -0.27]$ pp, $p = 0.0134$). Thus, adding unweighted field copies does not necessarily improve over intact text. The value of field-aware indexing depends on the retrieval task.

Tuned results are reported separately as benchmark-fitted ceilings in Appendix E. We next examine how often $L_1 \to L_2$ recovery coincides with cross-chunk separation.

5 Diagnosing the Fragmentation Failure Mode

Section 4 shows that preserving the complete record recovers most of what chunking loses, sharply so in education. The recovery is consistent with the motivating fragmentation hypothesis, in which query-matching evidence and the record-identifying field fall into different chunks. However, the aggregate Recall@10 gap alone does not establish this mechanism. The $L_1 \to L_2$ gain could also arise from changes in retrieval statistics, such as differences in term frequency or the greater opportunity for lexical overlap in a longer document. The analysis below measures how much of the recovery is attributable to cross-chunk separation.

Classification protocol. We inspect each query for which L_1 misses the gold target in the top 10 but L_2 retrieves it. For each such query q and its gold node r, we identify two sets of chunks among the L_1 chunks belonging to r: the *matching chunks*, those with lexical overlap with q meeting a co-occurrence threshold of $\theta = 2$ distinct token types, and the *identifier chunk*, the chunk containing the record's identifying field (Table 1). To avoid misclassifying Bengali inflectional variants, we apply Unicode token normalization and require co-occurrence across distinct query token types. We then assign each recovered query to one of three categories:

Cross-chunk separation. The matching chunks and the identifier chunk do not overlap: the query-matching evidence in the record was indexed separately from the field that identifies the record. This matches the example in Section 1; preserving an

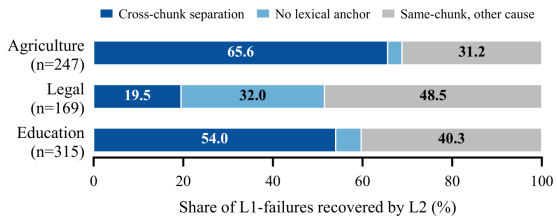


Figure 2: Why L_1 fails and L_2 recovers (n = recovered queries). Cross-chunk separation is the largest category in agriculture and education. In law, most recoveries fall in the same-chunk residual (48.5%) or lack a lexical anchor under the overlap threshold (32.0%).

intact record is one way to reunite this evidence.

No lexical anchor. No chunk of the record, including the identifier chunk, shares meaningful vocabulary with the query ($\max_j$ overlap $< \theta$). This category does not support the separated-evidence explanation. Vocabulary aggregation or changes in ranking may contribute, but the classification does not distinguish them.

Same-chunk, other cause. The matching content and the identifying field are already co-located in the same chunk, yet L_1 still missed the record. The observed matching evidence and identifier are not separated. This category leaves the ranking cause unresolved.

We compute these categories over the full recovered set in each domain, not a sample, since chunk membership and lexical overlap are both fully determined by the indexing pipeline already used to produce Table 2. Because legal L_1 evaluates passage-level chunks while L_2 evaluates act-level nodes, we collapse each act's constituent passages to their parent node before classification, so the legal diagnostic categories reported below are computed at the act level rather than the passage level.

Results. Figure 2 assigns 54.0% of education recoveries and 65.6% of agriculture recoveries to cross-chunk separation. The legal share is 19.5%; its largest category is the same-chunk residual (48.5%), followed by no lexical anchor under the threshold (32.0%). Same-chunk cases also account for 31.2% in agriculture and 40.3% in education. Length normalization, candidate competition, and other ranking effects may explain these cases, but this analysis does not identify their causes. The recovered sets count gains only: L_2 may also lose queries retrieved by L_1.

Counterfactual Condition	Agri	Legal	Edu
Unshuffled Reference	69.90%	77.82%	91.58%
Full Schema Shuffle	0.75%	1.91%	5.09%
Content-Only Shuffle†	56.90%	87.55%	87.02%
Identity-Only Shuffle†	47.55%	84.64%	94.04%

Table 3: Field-value shuffling (Recall@10). †Legal and education partial shuffles use different scoring configurations and are not comparable to their reference values.

6 Field-Level Analysis

The previous results show when fielded indexing helps but not which fields carry that advantage. We examine this at the field level through group ablations and counterfactual permutations, testing whether the candidate query fields identified in Table 1 actually behave as query-aligned anchors.

6.1 Field-Group and Sub-Field Analysis

We group the schema by function (full schema in Appendix A) and report three domain-relevant field groups (Identity, Diagnostic, and Procedural) defined by domain-specific schema roles. Exact group membership and sub-field ablations are in Appendix B.

The strongest reported group differs by domain. Identity reaches 98.09% R@10 in legal retrieval, while Identity and Diagnostic reach 50.15% and 27.25% in agriculture. In education, Procedural fields are strongest at 78.77% versus 64.91% for Identity. The legal and education patterns accord with their query regimes. In agriculture, the disease-name field alone reaches 52.50%, exceeding the Diagnostic group. The education results do not support treating the experiment name as its dominant retrieval anchor.

Legal sub-field results provide a direct illustration of query alignment: `act_title_bn` alone reaches 100.00% R@10, while `act_no_bn` reaches only 2.18%. Both identify the same acts, but only the title appears in the surface form used by queries. We define a *query-aligned anchor* as a field whose content provides a direct lexical link to the query formulation.

6.2 Perturbation Attribution

We permute field values across records while retaining record identifiers (Table 3). Identity-Only shuffles the designated name or title field; Content-Only shuffles the auxiliary full-text field. Full shuffling reduces Recall@10 to 0.75–5.09%, showing that field labels alone are insufficient without the

correct content. In agriculture, shuffling either the name or full text lowers recall. Legal and education partial shuffles use different scoring configurations from the displayed references, so we do not interpret those values as gains or losses against them.

7 Structured Knowledge as Agent Memory

The retrieval experiments show that representation affects which records are found and that this effect may depend on query-aligned fields. We next ask whether representation differences remain when an LLM agent uses the knowledge store as external memory.

7.1 Controlled Agent Protocol

We evaluate a minimal retrieval-grounded agent to hold the memory controller constant across conditions. The agent uses `gemini-2.5-flash-lite` through OpenRouter with one tool, `retrieve_memory(query, k=5)`, and at most two retrieval calls per query. It must retrieve before answering, use only retrieved evidence, cite supporting node identifiers, and return "insufficient evidence" when the memory does not support an answer. The full prompt and tool schema are in Appendix D.

We evaluate four memory conditions: M_1 chunked BM25 (L_1), M_2 full-node BM25 (L_2), M_3^u structured uniform BM25F (L_3, $w_f = 1.0$), and M_3^t structured tuned BM25F (L_3, benchmark-informed weights extending the a priori core; see Appendix E and the ceiling discussion in Section 4). We first sampled 300 queries per domain with seed 42: agriculture by dialect bucket, legal by difficulty, and education by difficulty. Excluding 28 queries without common completed runs leaves 872 (286/289/297 by domain). Abstentions count as retrieval misses. Legal M_1 returns passage identifiers while gold targets use act identifiers; we therefore exclude it from performance comparisons and evaluate $M_1 \rightarrow M_2$ on the remaining 583 queries. Other memory comparisons use all 872. Metrics use the first five unique retrieved identifiers across calls. Citation rate is the fraction of all evaluated queries whose answer contains a `[Node: <id>]` marker; abstentions remain in the denominator, and neither identifier validity nor entailment is checked.

7.2 Representation Transfer to Agent Memory

On agriculture and education ($n = 583$), full-node memory raises Recall@5 from 27.6% to 61.9% (+34.31 pp; paired McNemar $\chi^2 = 186.92$, $p = 1.50 \times 10^{-42}$; discordant pairs $b = 207$, $c = 7$).

Across all three domains ($n = 872$), uniform fielded memory reaches 55.4% Recall@5 versus 56.4% for full-node memory (McNemar $p = 0.292$). Recall@1 is numerically higher (0.304 to 0.322; $p = 0.157$). Citation rate increases from 48.2% to 60.4% (+12.27 pp; McNemar $\chi^2 = 54.78$, $p = 1.35 \times 10^{-13}$, $b = 158$, $c = 51$). These tests are uncorrected exploratory comparisons. The tuned condition reaches 80.5% Recall@5, indicating headroom under benchmark-informed weighting rather than a held-out gain.

Domain patterns differ. In law, Recall@5 is 45.3% for full-node, 48.8% for uniform, and 98.6% for tuned memory; uniform Recall@1 rises from 32.5% to 41.9%. Agriculture drops from 56.3% to 49.3% under uniform fields, then reaches 62.6% with tuning. Education is similar under full-node and uniform memory (67.3% and 67.7%). These observations do not identify a single ranking mechanism shared by all domains.

7.3 Anchor-Field Perturbation in Agent Memory

We conduct an exploratory perturbation study on a separate subset of 100 queries per domain, sampled with seed 42 from pools disjoint from the original agent sample using the same strata. The separate unshuffled run reaches 60.33% Recall@5 and invokes the tuned M_3 backend. It is not pooled with the primary memory comparison.

Here, Anchor shuffles disease names and Bengali aliases in agriculture, act titles and short titles in law, and procedure steps and observation/conclusion fields in education. Content shuffles the configured non-anchor fields, including auxiliary full text. These sets differ from the standalone single-field controls. The perturbed backend also serializes fewer context fields than the unshuffled backend, confounding the answer-level comparison.

Under the same perturbed backend, Recall@5 is 51.33% with content fields shuffled and 35.67% with anchor fields shuffled. This 15.7-point gap is consistent with sensitivity to query-aligned content. Rendering differs from the unshuffled run

Condition	Memory Representation	n	R@1	R@5	MRR@5	Avg. Tokens	Citation Rate
M_1	Chunked BM25 (Agri + Edu)	583	0.158	0.276	0.205	418	50.1%
M_2	Full-node BM25 (Agri + Edu)	583	0.293	0.619	0.418	778	60.7%
M_2	Full-node BM25 (all domains)	872	0.304	0.564	0.404	677	48.2%
M_3^u	Uniform fields (all domains)	872	0.322	0.554	0.411	749	60.4%
M_3^t	Tuned fields (all domains; ceiling)	872	0.548	0.805	0.644	1,015	83.7%

Table 4: Agent memory performance. The first two rows compare agriculture and education only ($n = 583$); legal M_1 is excluded because of incompatible passage/act identifiers. The final three rows use the same three-domain subset ($n = 872$). Tuned fields are an exploratory ceiling. Citation rate measures marker presence, not validated attribution or entailment. Differences use unrounded results.

Memory State (n=300)	R@5	MRR@5	Citation
Unshuffled Structured Run	0.603	0.477	63.3%
Content-Field Shuffled	0.513	0.415	40.3%
Anchor-Field Shuffled	**0.357**	**0.197**	**37.0%**

Table 5: Exploratory agent perturbations on a separate 300-query subset. Bold indicates maximum degradation (lowest performance under perturbation). Anchor-shuffled Recall@5 is 15.7 pp below content-shuffled recall under the same perturbed backend. Field selections and context rendering differ from the primary study; see text. Citation denotes citation rate.

(60.33%), so losses against that baseline are not interpreted. In law, Recall@1 is 58.0% under content shuffling and 9.0% under anchor shuffling. In education, the perturbed anchor content is procedural rather than the experiment name.

8 Discussion and Conclusion

Keeping records intact consistently improves standalone retrieval over the strict chunked baseline, while the additional value of field-aware indexing varies by domain. Cross-chunk separation accounts for most agriculture and education recoveries, while most legal recoveries fall into other categories. Field ablations suggest that useful content follows query formulation: statutory titles in law, disease names in agriculture, and procedures in education. Partial permutations do not establish a unique role for record identifiers.

The agent study also suggests that representation warrants attention in memory design. The chunked-to-intact comparison covers agriculture and education; the three-domain comparison finds a higher citation rate with uniform fields without higher aggregate Recall@5. Neither metric establishes answer correctness. Supplementary dense results are consistent with representation sensitivity but do not evaluate the full ladder (Appendix F).

Fielded records can be viewed as self-contained units with typed attributes. We build no graph and test no graph-based retriever, so these results do not measure the value of links or multi-hop reasoning. Whether graph connectivity adds gains after controlling for record integrity, scoring, and query alignment remains open.

For knowledge already organized into records, indexing intact records is the first design choice to check. Field-aware weighting can help when weighted fields match query vocabulary and can lower recall otherwise. Stronger chunking baselines, canonical BM25F, and other languages are the next tests.

Limitations

Scope. We study three Bengali domains with curated records and make no claim for other languages, domains, or unstructured sources. Dense and hybrid results cover only chunked and structured-flat representations. No graph or inter-record relation is evaluated.

Representation comparisons. Stronger chunking baselines may narrow the reported gaps. Preserving records also changes document length, candidate count, and BM25 normalization; legal relevance changes from passages to acts. At L_1, Recall@10 counts ten ranked chunks without parent-node deduplication, so repeated parents occupy separate slots. Field-aware indexing duplicates overlapping content through auxiliary full text and individual fields. Neither scoring implementation is canonical BM25F, and tuned results are benchmark-fitted ceilings.

Fragmentation diagnosis. The diagnostic categories depend on one co-occurrence threshold ($\theta = 2$); no sensitivity analysis was run.

Queries. Source-derived query formulations may overlap with target-record vocabulary and may not fully represent naturally occurring user inquiries.

Regional formulations have not undergone a dedicated speaker-based authenticity evaluation.

Agent study. We evaluate one backbone with static memory. Legal chunked-agent results are excluded because of an identifier mismatch. Perturbation results are exploratory because configuration and context rendering differ from the primary study. Retrieval and citation presence do not establish answer correctness or semantic support.

Ethical Considerations

The agricultural records include pesticide guidance. These experiments do not evaluate the correctness or safety of chemical advice; preserving record integrity has not been validated as a safety intervention. The agents are research prototypes, and pesticide recommendations require expert review before use. Agricultural query material was reused from KrishokChat and related advisory resources (Reza et al., 2026b,a). Code is released under MIT. Reused datasets, queries, and source texts remain subject to their original licenses and redistribution conditions.

Data and Code Availability

Code and derived benchmark artifacts are available on GitHub (`https://github.com/RaiyaanReza/Beyond-Chunking-Node-Integrity-and-Field-Structure-in-Bengali-Retrieval-Augmented-Agent-Memory`) and Hugging Face (`https://huggingface.co/datasets/RaiyanKhaan/Beyond-Chunking`). Release documentation records provenance and redistribution constraints for third-party materials.

Acknowledgements

We acknowledge assistance from Claude, OpenAI Codex, and Perplexity with manuscript editing, LaTeX preparation, consistency checks, and statistical verification, as detailed in the following declaration.

Declaration on Generative AI

Claude was used for LaTeX editing and manuscript review. OpenAI Codex assisted with LaTeX editing, language revision, manuscript checks, and statistical checks of existing experiment logs. Perplexity assisted with manuscript review, consistency checks, and verification of reported statistics. The evaluated agent backbone is `gemini-2.5-flash-lite`. The authors are responsible for the experimental design, results, and final manuscript.

References

Shuqi Cao, Jingyi He, and Fei Tan. 2026a. HiGMem: A hierarchical and LLM-guided memory system for long-term conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33853–33862. Association for Computational Linguistics.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2026b. Out of style: RAG's fragility to linguistic variation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 280–318, Rabat, Morocco. Association for Computational Linguistics.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Fatema Tuj Johora Faria, Mukaffi Bin Moin, Ahmed Al Wase, Mehidi Ahmmed, Md. Rabius Sani, and Tashreef Muhammad. 2023. Vashantor: A large-scale multilingual benchmark dataset for automated translation of Bangla regional dialects to Bangla language. *Preprint*, arXiv:2311.11142.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Millicent Li, Tongfei Chen, Benjamin Van Durme, and Patrick Xia. 2025a. Multi-field adaptive retrieval. In *International Conference on Learning Representations (ICLR)*. Spotlight.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025b. StructRAG: Boosting knowledge intensive reasoning of LLMs via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.

M. Millat. 2026. BDLawCorpus-Dataset-V1. https://huggingface.co/datasets/millat/BDLawCorpus-Dataset-V1.

Thong Nguyen, Shubham Chatterjee, Sean MacAvaney, Iain Mackie, Jeff Dalton, and Andrew Yates. 2024. DyVo: Dynamic vocabularies for learned sparse retrieval with entities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 767–783.

Khan Raiyan Ibne Reza, Sanjana Aktar Maria, and Sumaiya Tabassum Nimi. 2026a. Where does retrieval fail? Evaluating RAG architectures for agricultural advisory. *Preprint*, arXiv:2608.14886.

Khan Raiyan Ibne Reza, Sumaiya Tabassum Nimi, and Omar Ibne Shahid. 2026b. KrishokChat: A provenance-traceable multi-task Bengali agricultural benchmark with safety-critical chemical advisory. *Preprint*, arXiv:2606.29243.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49.

Yoav Kor Sade, Arvindh Arun, Rishi Puri, Steffen Staab, and Maya Bechler-Speicher. 2026. Ex-GraphRAG: Interpretable evidence routing for graph-augmented LLMs. *arXiv preprint arXiv:2605.21994*.

Adib Sakhawat. 2024. Bangladesh legal acts dataset. https://huggingface.co/datasets/sakhadib/Bangladesh-Legal-Acts-Dataset.

K. M. Jubair Sami, Dipto Sumit, Ariyan Hossain, and Farig Sadeque. 2026. Benchmarking Bengali dialectal bias: A multi-stage framework integrating RAG-based translation and human-augmented RLAIF. *Preprint*, arXiv:2603.21359.

Shalin Shah, Srikanth Ryali, and Ramasubbu Venkatesh. 2026. SRAG: RAG with structured data improves vector retrieval. *arXiv preprint arXiv:2603.26670*.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19336–19352. Association for Computational Linguistics.

Andrea Volpini, Elie Raad, Beatrice Gamba, and David Riccitelli. 2026. Structured linked data as a memory layer for agent-orchestrated retrieval. *arXiv preprint arXiv:2603.10700*.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. CausalRAG: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22680–22693. Association for Computational Linguistics.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38.

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. 2026. Lightweight LLM agent memory with small language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12914–12929. Association for Computational Linguistics.

A Full Knowledge Schema and Field Functional Mapping

To illustrate field structure, semantic roles, and representations across the evaluation ladder, Table 6 summarizes 25 domain fields and three auxiliary full-text entries across the three evaluation corpora.

Functional groups. Identity, Diagnostic, and Procedural are domain-specific ablation labels, not a universal taxonomy. Exact membership for the reported groups is listed in Appendix B; auxiliary full text is separate from these group ablations.

Implementation. Standalone legal and education retrieval apply field weights through text repetition in a single BM25 document, while agriculture and agent memory sum weighted per-field BM25 scores. Both retain auxiliary full text alongside individual fields, duplicating overlapping content rather than isolating field boundaries. The underlying legal corpus also indexes supplementary statutory regime metadata (`legal_system_bn`), evaluated under the Legal Diagnostic group in Appendix B; Table 6 and agent memory (Table 10) focus on substantive statutory provisions. The shuffle controls include different scoring configurations, so not every row is a matched comparison. Exact field inventories and weights are in the released code.

B Supplementary Ablation Details

Table 7 reports BM25 retrieval over concatenated group fields and individual identity fields. The exact group memberships are:

Domain	Field Identifier	Functional Group	Field Semantic Scope and Contents
Agriculture (10 Domain Fields + 1 Auxiliary; 1,028 Knowledge Nodes)			
Agriculture	disease_pest_name_bn	Identity	Canonical Bengali disease or pest name
Agriculture	colloquial_aliases	Identity	Regional and vernacular names
Agriculture	scientific_name	Identity	Latin taxonomic name
Agriculture	symptoms	Diagnostic	Symptoms affecting leaves, stems, and fruit
Agriculture	favorable_conditions_bn	Diagnostic	Temperature, humidity, and rainfall conditions
Agriculture	management	Procedural	Cultural, mechanical, and biological control
Agriculture	prevention	Procedural	Preventive cultivation and seed treatment
Agriculture	chemical_management_bn	Procedural	Chemical ingredients and application guidance
Agriculture	affected_crop_bn	Contextual	Susceptible crop and host species
Agriculture	safety_advisory	Safety / Metadata	Hazard warnings, protective equipment, harvest intervals
Agriculture	content_bn	Auxiliary (full text)	Concatenated record text (auxiliary field)
Legal (10 Domain Fields + 1 Auxiliary; 735 Statutory Acts)			
Legal	act_title_bn	Identity	Official act title
Legal	short_title_bn	Identity	Short title or commonly cited name
Legal	act_no_bn	Identity	Act number and enactment year
Legal	preamble_bn	Contextual	Legislative purpose and policy context
Legal	definitions_bn	Diagnostic	Definitions of legal terminology
Legal	section_titles_bn	Content	Section titles and subject headings
Legal	powers_bn	Procedural	Administrative and executive powers
Legal	procedures_bn	Procedural	Statutory procedures and regulatory protocols
Legal	penalties_bn	Procedural	Fines, penalties, and imprisonment terms
Legal	status_bn	Contextual	In-force status, amendments, and repeal metadata
Legal	content_bn	Auxiliary (full text)	Concatenated record text (auxiliary field)
Education (5 Domain Fields + 1 Auxiliary; 190 Secondary Curriculum Experiments)			
Education	experiment_name_bn	Identity	Textbook experiment title
Education	apparatus_and_chemicals	Identity	Required apparatus and reagents
Education	observation_conclusion	Diagnostic	Observations and scientific conclusions
Education	procedure_steps	Procedural	Experimental procedure
Education	precautions	Safety / Metadata	Laboratory safety and experimental precautions
Education	full_text_bn	Auxiliary (full text)	Concatenated record text (auxiliary field)

Table 6: Semantic summary of domain fields and auxiliary full text. Record identifier fields are named in Table 1; procedures are stronger than the name field in education. Groups follow the domain-specific ablations. The table summarizes core semantic fields, rather than every indexed key (e.g., supplementary statutory regime metadata legal_system_bn is evaluated in Appendix B); exact field inventories are in the released code. Auxiliary text overlaps individual fields.

Agriculture. Identity: disease_pest_name_bn, scientific_name, colloquial_aliases; Diagnostic: symptoms, favorable_conditions_bn; Procedural: management, prevention, chemical_management_bn.

Legal. Identity: act_title_bn, short_title_bn, act_no_bn; Diagnostic: definitions_bn, legal_system_bn; Procedural: procedures_bn, powers_bn, penalties_bn.

Education. Identity: experiment_name_bn, apparatus_and_chemicals; Diagnostic: observation_conclusion; Procedural: procedure_steps.

Functional Group	R@1	R@5	R@10	MRR
Agriculture (Best Tuned BM25F: 72.35%)				
Identity Group	20.90%	41.40%	50.15%	0.296
Diagnostic Group	8.90%	21.00%	27.25%	0.142
Procedural Group	5.20%	10.60%	14.30%	0.076
Legal (Best Tuned BM25F: 100.00%)				
Identity Group	94.70%	97.90%	98.09%	0.962
Procedural Group	36.50%	64.70%	72.73%	0.489
Diagnostic Group	23.40%	39.70%	44.91%	0.305
Education (Best Tuned BM25F: 93.33%)				
Procedural Group	27.00%	68.20%	78.77%	0.449
Diagnostic Group	22.30%	57.90%	67.89%	0.369
Identity Group	14.40%	48.90%	64.91%	0.294
Identity Sub-Field Standalone Retrieval (R@10)				
Agri: disease_pest_name_bn		**52.50%**		0.325
Agri: colloquial_aliases		10.15%		0.055
Agri: scientific_name		1.55%		0.007
Legal: act_title_bn		**100.00%**		0.995
Legal: short_title_bn		91.00%		0.893
Legal: act_no_bn		2.18%		0.015

Table 7: Field-group and identity sub-field BM25 ablation metrics across domains; exact group memberships are listed above.

C Query Construction and Representative Benchmark Examples

To provide concrete visibility into query formulation, linguistic register diversity, and the relation between user queries and memory fields, Table 8 presents nine representative benchmark queries across Agriculture, Legal, and Education.

Corpus provenance and query construction. Benchmark queries were constructed from domain-specific source materials and query-generation procedures designed to reflect the target query regimes:

- **Agriculture (2,000 queries):** Drawn from Bengali agricultural advisory resources and KrishokChat, with Vashantor as a reference for regional formulations (Reza et al., 2026a,b; Faria et al., 2023). Queries span 16 distinct regional dialects alongside Standard Dhaka Bengali. In the symptom-diagnostic regime, queries describe crop symptoms or explicitly name diseases and pests, as illustrated by Agri-02 and Agri-03.
- **Legal (1,100 queries):** Compiled from statutory question-answering resources and citizen inquiries (Millat, 2026). Queries cover 15 regional dialect formulations alongside formal standard legal Bengali. In this known-item regime, inquiries identify target statutes by name or legislative topic while seeking interpretations of specific provisions, administrative duties, or penal liabilities.
- **Education (570 queries):** Developed from secondary science curricula (NCTB textbooks, grades 9–10 physics, chemistry, and biology). Queries cover 5 student/colloquial registers alongside standard textbook prose. In this procedural regime, students articulate hands-on experimental actions, glassware setups, or observed physical changes without recalling formal scientific nomenclature.

Query provenance and target mappings. The evaluation uses the benchmark's fixed query-to-target mappings across representations. Agricultural queries draw on KrishokChat and related advisory resources (Reza et al., 2026b,a); legal and education provenance is described above. Source-derived formulations may share vocabulary with their target records. This study does not establish dialectal authenticity through a dedicated speaker-based evaluation.

Interpretation of the examples. Legal queries explicitly cite act names, supporting the strong title-only result. Cross-chunk separation nevertheless accounts for only 19.5% of legal recoveries, so the examples do not explain the aggregate legal gap. Education queries mention apparatus and actions; Procedural fields outperform Identity in the group ablation. The examples illustrate how query terms can connect to titles, names, and procedural content.

D Agent Memory Prompt and Tool Schema

The retrieval-grounded agent is governed by the system instructions and tool interface shown below:

System Prompt (Agent Instructions):
"You are a factual knowledge assistant. When presented with a question, you must consult your external memory using `retrieve_memory(query, k=5)` before formulating your answer. Answer only using evidence explicitly confirmed by the retrieved records. Cite supporting node identifiers using `[Node: <id>]`. If the retrieved memory does not contain sufficient facts to answer the question, explicitly state 'insufficient evidence'. Do not extrapolate or guess."

Tool Interface:
`retrieve_memory(query: str, k: int = 5) -> List[Dict]`
Returns top-k documents formatted with node identifiers and textual content matching the active memory condition (M1, M2, or M3).

E Field-Weight Tuning Protocol

Configuration	Agriculture	Legal	Education
L_3 Tuned BM25F (Ceiling)†	72.35%	100.00%	93.33%

Table 9: Benchmark-fitted tuned BM25F Recall@10 ceiling. †Weights were selected on the full benchmark without a held-out split; this is a fitted reference, not a formal upper bound.

Protocol and disclosure. Only per-field weights w_f and length-normalization hyperparameters (k_1, b; education best: k_1=3.0, b=0.50) were tuned. Grid search selected the configuration maximizing Recall@10 on each domain's full query pool (ties broken by MRR). Because weights were selected on the full benchmark without a held-out development split, these values represent an empirical ceiling rather than generalizable out-of-sample estimates. In contrast, the standalone L_3 agriculture baseline (69.90% in Table 2) used fixed heuristic weights (`disease`: 3.5, `aliases`: 2.0,

Domain & ID	Dialect / Register	Query Formulation (Romanized Bengali & English)	Gold Target Node	Record Identifier Field & Value
Agriculture: Symptom-Diagnostic and Vernacular Inquiries				
Agri-01	Rajshahi Dialect	*"Gacher pata keno emon holde hoiya kukraiya jaitechhe, eita kiser lokkhon?"* *Gloss:* "Why are the leaves turning yellow and curling up like this? What is this a symptom of?"	`AT_tomato_yellow_leaf_curl_01`	**Field:** `disease_pest_name_bn` **Value:** *Tomator holde pata kukrano rog* (Tomato yellow leaf curl virus)
Agri-02	Barisal Dialect	*"Amne kon dehi, ei tungro rogda kemne chhoray ar gache ki hoi?"* *Gloss:* "Tell me, how does this Tungro disease spread and what happens to the plant?"	`AT_rice_disease_control_040_part1`	**Field:** `disease_pest_name_bn` **Value:** *Dhaner tungro rog* (Rice tungro disease)
Agri-03	Standard Dhaka	*"Ei Hydronomidius molitor pokata kon kon gache beshi dhore?"* *Gloss:* "Which host plants does this Hydronomidius molitor insect commonly infest?"	`AT_extraction_all_pest_books_part363`	**Field:** `disease_pest_name_bn` **Value:** *Hydronomidius molitor*
Legal: Known-Item Statutory and Citizen Consultations				
Legal-01	Standard Formal	*"'Parbotya Zila Sthanio Sarkar Parishad Ain, 1989'-er 7 nombor dharati shohoj bhashay byakhya korun."* *Gloss:* "Explain Section 7 of the 'Hill District Local Government Parishad Act, 1989' in plain language."	`legal_act_0760`	**Field:** `act_title_bn` **Value:** *Parbotya Zila Sthanio Sarkar Parishad Ain, 1989*
Legal-02	Barisal Dialect	*"Hindudharmio Kalyan Trust Ain, 2018 er 16 nombor dharada shohoj koira bujhaiya koben ni?"* *Gloss:* "Could you easily explain Section 16 of the Hindu Religious Welfare Trust Act, 2018?"	`legal_act_1236`	**Field:** `act_title_bn` **Value:** *Hindudharmio Kalyan Trust Ain, 2018*
Legal-03	Bogra Dialect	*"Rangpur Biggan o Projukti Bishwobiddyaloy Ain, 2001-er 58 nombor dharay ki ki aini shomosshya hober pare?"* *Gloss:* "What legal complications might arise under Section 58 of the Rangpur Science and Tech Univ Act, 2001?"	`legal_act_0901`	**Field:** `act_title_bn` **Value:** *Rangpur Biggan o Projukti Bishwobiddyaloy Ain, 2001*
Education: Procedural and Apparatus Descriptions				
Edu-01	Student Colloquial	*"Panite rong meshale ki hoy, eta kibhabe dekhbo?"* *Gloss:* "What happens when dye is mixed in water, and how can we observe this?"	`edu_biology_0001`	**Field:** `experiment_name_bn` **Value:** *Panite tontur byapon porjobekkhon* (Observation of diffusion in water)
Edu-02	Jessore Dialect	*"Macher potka diye chinir drobon upore othar porikkha-ta kemne kore?"* *Gloss:* "How is the experiment of sugar solution rising using a fish bladder conducted?"	`edu_biology_0013`	**Field:** `experiment_name_bn` **Value:** *Otimatraboner porikkha: Thistle fanel o jontobojhilli shahajje obhishrobon* (Observation of osmosis using a thistle funnel and animal membrane)
Edu-03	School Lab Register	*"Onubikkhon jontre peyajer roshal sholkopotro diye kosh dekhar niom ki?"* *Gloss:* "What is the procedure for observing cells from an onion scale leaf under a microscope?"	`edu_biology_0006`	**Field:** `experiment_name_bn` **Value:** *Onubikkhon jontrer shahajje udvidkosh (peyajer kosh) porjobekkhon* (Observation of plant cells (onion cells) under a microscope)

Table 8: Representative benchmark queries across domains, illustrating query regimes, dialectal registers, gold target nodes, and record identifier fields. Queries and identifier values are presented in Romanized transliteration alongside English glosses.

`symptoms`: 1.5, `content`: 1.0) set a priori before evaluation without query tuning. Agent weights (Table 10) start from this a priori set for core fields and add benchmark-informed nonzero weights for secondary fields. The agent settings have no agent-time tuning and differ from standalone configurations.

Candidate grids. Agriculture scanned 9 weight configurations over (`content_bn`, `disease_pest_name_bn`, `colloquial_aliases`, `symptoms`, `management`) spanning identity-boost, diagnostic-boost, uniform, content-heavy, extreme-identity, and scaled-down families; the a priori set sat inside this grid as the baseline candidate (69.90% vs. 72.35% ceiling). Legal and education use domain-specific weight and hyperparameter grids. Table 10 specifies the fixed agent settings separately from the standalone optima.

Domain	Fixed Agent Field Weights (w_f)
Agriculture	`disease_pest_name_bn`: 3.5, `colloquial_aliases`: 2.0, `symptoms`: 1.5, `content_bn`: 1.0, `favorable_conditions_bn`: 1.0, `management`: 0.8, `prevention`: 0.8, `chemical_management_bn`: 0.5, `affected_crop_bn`: 0.5
Legal	`act_title_bn`: 5.0, `short_title_bn`: 2.0, `preamble_bn`: 1.0, `content_bn`: 1.0, `definitions_bn`: 0.8, `section_titles_bn`: 0.8, `act_no_bn`: 0.5, `penalties_bn`: 0.5, `powers_bn`: 0.5, `procedures_bn`: 0.5
Education	`procedure_steps`: 3.0, `experiment_name_bn`: 2.0, `observation_conclusion`: 1.5, `full_text_bn`: 1.0, `apparatus_and_chemicals`: 0.8, `precautions`: 0.5

Table 10: Fixed agent field weights for M_3^t. The backend uses per-field BM25 with default $k_1 = 1.5$, $b = 0.75$ in each domain; the education standalone optimum ($k_1 = 3.0$, $b = 0.50$) is separate. Agriculture agent weights start from the a priori set for core fields and add benchmark-informed nonzero weights for secondary fields. Uniform M_3^u assigns unit weights to these same listed fields. Other serialized fields (including `status_bn` and `legal_system_bn`) are not indexed for agent retrieval. In the agent implementation backend, agriculture fields match the schema in Table 6 and may be keyed with the `_bn` suffix (e.g., `colloquial_aliases_bn`, `symptoms_bn`, `management_bn`, and `prevention_bn`).

F Dense and Hybrid Retrieval Results

Table 11 reports dense, hybrid, and flat unweighted baselines. Structured-flat BM25 concatenates selected record fields, including auxiliary full text, without field weights; it differs from the full-text-only standalone L_2, and field selection differs from Table 2. Dense retrieval embeds that flattened record as a single passage, rather than computing separate field embeddings. Hybrid retrieval fuses sparse and dense rankings with reciprocal rank fusion. While these models do not evaluate the full $L_1 \to L_2 \to L_3$ decomposition across all ablation conditions, their differences are consistent with representation sensitivity beyond sparse retrieval, without separating record integrity from field content.

Level	Retriever	Agri	Legal	Edu
L_1 Chunked	BGE-M3*	62.95%	81.91%	38.42%
L_1 Chunked	mE5-small*	28.65%	n/a†	20.00%
L_1 Chunked	Hybrid RRF	65.60%	n/a†	26.32%
Record fields, flattened	BM25	64.55%	77.82%	91.40%
Record fields, flattened	BGE-M3*	68.40%	96.36%	93.86%
Record fields, flattened	mE5-small*	41.30%	91.73%	77.37%
Record fields, flattened	Hybrid RRF	73.05%	99.00%	87.72%

Table 11: Supplementary dense, hybrid, and flattened-record retrieval results (Recall@10), omitted from the main comparison in Table 2. †BGE-M3 serves as primary dense representative for legal; chunked mE5/Hybrid omitted there. *Evaluated zero-shot without domain fine-tuning.

G Statistical Significance and Bootstrap Confidence Intervals

Table 12 reports the complete paired statistical significance testing matrix for the core representation ladder (Table 2) across all three evaluation domains. For each pairwise comparison ($L_1 \to L_2$, $L_2 \to L_3$, and $L_1 \to L_3$), we report the empirical metric change (Δ in Recall@10 percentage points), the 95% confidence interval estimated via 5,000 paired bootstrap resamples with replacement, and the empirical paired bootstrap tail probability in the direction of the observed change.

All nine reported ladder comparisons have unadjusted $p \leq 0.0134$ and remain significant under Holm correction at $\alpha = 0.05$. In law, $L_2 \to L_3$ is a decrease of 1.73 pp; its 95% interval $[-3.18, -0.27]$ pp excludes zero, with its upper bound below zero. Values at the bootstrap resolution floor are reported as $p \leq 0.0002$; 0.0002 is the resolution floor for 5,000 draws.

Domain	Comparison	Δ (pp)	95% CI (pp)	p-value
Agri (n=2,000)	$L_1 \to L_2$	+5.30	[+3.65, +6.95]	≤ 0.0002
	$L_2 \to L_3$	+2.30	[+0.65, +3.95]	0.0062
	$L_1 \to L_3$	+7.60	[+5.90, +9.30]	≤ 0.0002
Legal (n=1,100)	$L_1 \to L_2$	+6.36	[+3.45, +9.09]	≤ 0.0002
	$L_2 \to L_3$	−1.73	[−3.18, −0.27]	0.0134
	$L_1 \to L_3$	+4.64	[+1.82, +7.45]	0.0004
Edu (n=570)	$L_1 \to L_2$	+52.46	[+47.89, +56.84]	≤ 0.0002
	$L_2 \to L_3$	+1.93	[+0.35, +3.68]	0.0116
	$L_1 \to L_3$	+54.39	[+50.00, +58.60]	≤ 0.0002

Table 12: Complete paired bootstrap significance matrix for Recall@10 (5,000 draws). All nine comparisons remain significant after Holm correction at $\alpha = 0.05$; displayed p-values are unadjusted. The resolution floor for 5,000 draws is 0.0002.