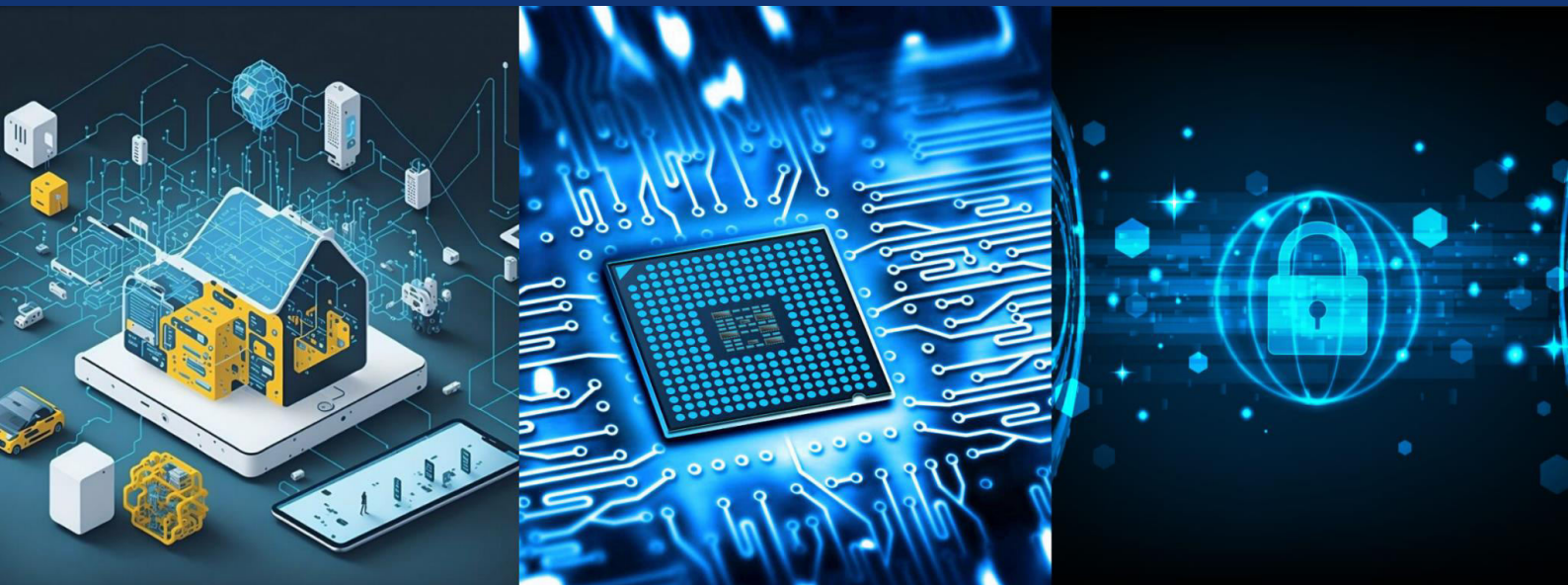
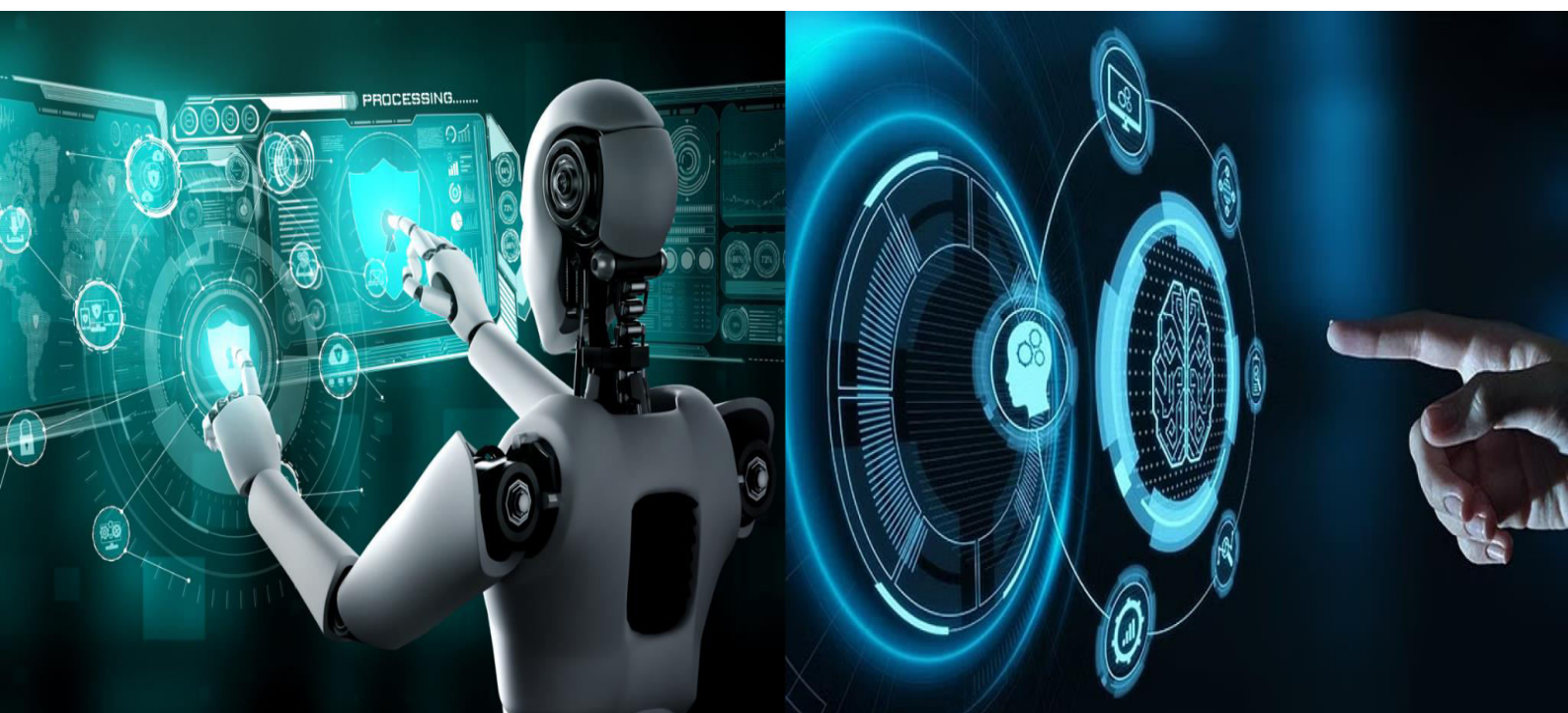




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





A Survey of Retrieval-Augmented Generation over Heterogeneous, Multi-Source Knowledgebases

Srujan Zanjali¹, Rimanshu Sonule¹, Pratik Bhoyar¹, Tanisha Shingne¹, Prof. Ayaz Khan²

B.Tech Student, Department of Computer Science and Engineering, St. Vincent Pallotti College of Engineering and Technology, Nagpur, Maharashtra, India¹

Assistant Professor, Department of Computer Science and Engineering, St. Vincent Pallotti College of Engineering and Technology, Nagpur, Maharashtra, India²

ABSTRACT: Retrieval-augmented generation (RAG) improves a language model's output by conditioning it on passages retrieved from an external corpus, but most of this work assumes a single, roughly homogeneous document collection. Practical knowledgebases violate that assumption: websites, PDFs, video transcripts, and code repositories are extracted, chunked, cited, and queried in different ways, and one query may need to draw on several at once. This survey organizes the techniques required for grounded question answering over such heterogeneous, multi-source knowledgebases: dense retrieval and RAG foundations, hybrid lexical-dense reranking, chunking strategies for long and structured documents, modality-specific retrieval for web, video, and code, multi-document fusion, and confidence estimation with abstention. Rather than treating each technique in isolation, we examine how each is affected once a corpus is assembled from structurally different sources at query time. One system, SAGE, which ingests and answers questions over crawled websites, PDFs, video transcripts, and GitHub repositories through a shared pipeline, appears as a brief case study showing how the surveyed ideas instantiate in a working implementation. We close with open problems in cross-modal fusion, confidence calibration, incremental re-indexing, and evaluation that remain unresolved across the literature.

KEYWORDS: Retrieval-augmented generation; multi-source retrieval; multi-modal question answering; survey; reranking; chunking

I. INTRODUCTION

Retrieval-augmented generation (RAG) has become the dominant pattern for grounding language-model output in an external corpus: a retriever selects passages relevant to a question, and a generator conditions its answer on those passages rather than on parametric memory alone [1]. Since its introduction, RAG research has expanded rapidly along several axes, including retriever architecture, passage fusion, iterative retrieval, and calibration, nearly all of it measured on a single, relatively uniform corpus such as a Wikipedia dump, a set of internal wiki pages, or a collection of PDF manuals. That uniformity is analytically convenient: it fixes one chunking strategy, one citation format, and one notion of a document.

That convenience disappears once the material a user wants to query spans structurally unrelated sources — web pages, formatted documents, spoken-language transcripts, and source code — each with its own extraction process, natural retrieval unit, and citation semantics. Answering one query against whichever source holds the relevant evidence, with a citation identifying both the source and the location within it, requires composing techniques largely designed and evaluated one modality at a time. This composition problem, rather than any single retrieval or generation advance, is what motivates a survey dedicated to heterogeneous knowledgebases.

The survey is organized around that composition. Sections III–V review dense retrieval foundations, hybrid reranking, and chunking; Section VI surveys modality-specific retrieval for web, video, code, and multi-document fusion; Section VII addresses confidence and abstention, where a system must judge not only *what* to retrieve but whether it retrieved



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

enough. Section VIII compares the techniques, Section IX presents SAGE as a case study, Section X collects open problems, and Section XI concludes.

Scope and relation to prior surveys. General RAG design is already surveyed at length [2], and hallucination in generation has its own dedicated survey [3]; we duplicate neither. Instead of surveying RAG technique by technique over an assumed-uniform corpus, we organize technique choices around the points of strain that emerge once a corpus is assembled from structurally different sources at query time, which is why chunking, reranking, and confidence calibration are reconsidered in Sections IV–VII. The literature below was selected by starting from the four modality-specific retrieval areas of Section VI and working back to the shared machinery each assumes, rather than from any single venue’s proceedings; we claim no exhaustive coverage of any one area on its own terms.

II. PROBLEM DEFINITION AND TERMINOLOGY

Much RAG terminology implicitly assumes a single corpus, so we first state precisely what changes when a system moves to several heterogeneous ones.

We use *source* to mean one ingested unit with its own identity and extraction process (one crawled website, one PDF, one video, or one code repository), and *knowledgebase* to mean a set of sources a user has grouped together. A *chunk* is the unit actually embedded and retrieved, produced from a source by some modality-specific chunking function. A *query* may target one source, an entire knowledgebase, or an explicit, user-selected subset of sources spanning one or more knowledgebases; we call the last case a *combined-source query*, and it is the case that introduces requirements a single-corpus system does not have.

Three properties distinguish multi-source retrieval from ordinary multi-document retrieval within one corpus:

- **Heterogeneous extraction.** Each source type has its own extraction pipeline (Section VI) that yields text of different lexical character (crawled prose, PDF-extracted prose with layout artifacts, ASR-transcribed speech, and source code) before any retrieval model sees it.
- **Heterogeneous chunking.** As Section V discusses, the natural retrieval unit differs by modality (a paragraph, a timestamped transcript segment, a function-scoped code block), so a single global chunking function is a simplification rather than a given.
- **Source-level accountability.** A citation in a multi-source system must identify which *source*, not just which passage, produced a piece of evidence. A combined-source query also carries an implicit expectation that every source the user explicitly selected was actually considered, a requirement with no analogue when the corpus is one undifferentiated collection.

III. DENSE RETRIEVAL AND RAG FOUNDATIONS

The move from sparse lexical matching to learned dense representations began with semantic-matching architectures such as DSSM [4] and matured into dual-encoder retrievers, of which Dense Passage Retrieval (DPR) [5] is the canonical instance: queries and passages are embedded separately and compared by inner product or cosine similarity. Because the passage side is fixed at index time, its embeddings can be precomputed offline and served through approximate nearest-neighbor search, leaving query-time cost independent of corpus size, which is what makes one shared index across web, PDF, transcript, and code chunks feasible.

Retrieval-augmented generation [1] pairs such a retriever with a generator that conditions its output on the retrieved passages, marginalizing over retrieved evidence rather than depending on parametric memory alone. Fusion-in-Decoder [6] refined the coupling by having a sequence-to-sequence model encode each retrieved passage separately, so the decoder attends across all of them at once, scaling better with passage count than concatenating them into a single encoder pass. The design space around this core, what to retrieve, when, how many passages, how to fuse them, and whether to iterate, is already surveyed at length [2]. Its relevance here is specific: once a system ingests more than one kind of source, every one of those choices must be resolved per modality rather than once globally, and that requirement is this survey’s throughline.

Every dense retriever rests on a sentence- or passage-embedding model. Sentence-BERT [7] made dense retrieval practical by producing embeddings whose cosine geometry is directly usable for ranking and clustering, sidestepping



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the quadratic cost of scoring every query-passage pair with a full cross-encoder. Distillation pushes the same idea toward deployment: MiniLM [8] transfers a large teacher's attention behavior into a compact student, trading a little accuracy for much lower latency and memory use, which is decisive for interactive retrieval. The distilled checkpoint all-MiniLM-L6-v2 has become a common default in open-source RAG tooling.

Two design choices here take on added weight once a system indexes several source types together. One is the approximate nearest-neighbor index: fixing its distance metric once for a shared index presumes the encoder's geometry behaves comparably across prose, transcribed speech, and code — a presumption rarely tested directly. The other concerns training data: DPR's original recipe [5] draws negatives from a single corpus's question-passage pairs, so a system reusing an off-the-shelf encoder inherits a negative distribution unlikely to match the mixed prose/transcript/code negatives it meets at query time. Fine-tuning one encoder against a deliberately multi-modal negative distribution appears absent from the literature, an open problem revisited in Section X.

IV. RERANKING AND HYBRID RETRIEVAL

Dense retrieval alone systematically underperforms on queries dominated by exact terms, identifiers, proper nouns, error codes, which a bi-encoder's semantic similarity underweights. Two broad families address this. One re-scores an initial dense-retrieved list with a more expensive, jointly-encoded model: cross-encoder rerankers [9] score a query and passage together, and late-interaction models such as ColBERT [10] compute token-level similarity, keeping much of a cross-encoder's accuracy at lower cost. The other family blends dense and lexical signals directly. CLEAR [11] trains a dense retriever to capture the residual a classical scorer such as BM25 [12] misses; Reciprocal Rank Fusion [13] merges independently produced ranked lists by rank position alone, with no need for calibrated, comparable scores. Concretely, RRF scores a document d as

$$\text{RRF}(d) = \sum_i \frac{1}{k + \text{rank}_i(d)} \quad (1)$$

summed over each ranked list i in which d appears, with $\text{rank}_i(d)$ its position in list i and k a small constant (commonly 60) damping the influence of any single list's very top ranks; a document absent from a list simply contributes zero rather than requiring an imputed score. This rank-only property is what makes RRF attractive for fusing across sources with different lexical and semantic characteristics (Section VI): it needs no assumption that a 0.73 semantic score from a prose retriever means the same thing as a 0.73 score from a code retriever, only that each retriever's own ranking is locally meaningful.

This blending family suits multi-source retrieval because its candidate pool mixes prose, code fragments, and transcript segments with dissimilar embedding and lexical statistics (Section V), which rank-based or explicitly weighted fusion tolerates better than a score premised on one uniform passage distribution. A structural consequence follows too: the output must decide not only *which* passages are best but *whether at least one passage from each requested source* reaches the final answer, so ranking becomes a two-stage problem, scoring followed by selection under a source-coverage constraint, rather than scoring and truncating alone.

The cost asymmetry between these two families determines which one a system can afford. A cross-encoder or ColBERT-style pass [9], [10] runs a neural forward pass per candidate — easy to justify over the top 100 candidates once per query, but proportionally more expensive precisely when a multi-source system widens its pool to guarantee per-source representation. A closed-form hybrid score behaves differently: computed from quantities retrieval already produced, so widening the pool for fairness leaves its cost unchanged.

V. CHUNKING STRATEGIES

Retrieval granularity is usually decided before any retrieval model is chosen: documents are split into fixed-size, overlapping chunks, and the chunk, not the document, becomes the unit that is embedded, indexed, and retrieved. Fixed-size chunking with modest overlap remains the dominant practical choice because it is cheap, deterministic, and works passably across many document types. Chunk-size sensitivity varies by domain and document structure, as shown for financial reports, where headers, tables, and dense terminology interact with the chosen chunk boundary in ways that measurably affect downstream retrieval quality [14].



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

“Late chunking” [15] inverts the usual order of operations: it encodes a long document with a long-context embedding model first and derives chunk-level embeddings from token-level representations afterward, so each chunk’s embedding still reflects surrounding context that a naively-chunked-then-embedded passage would lose. Comparative studies [16] suggest no single strategy dominates across document types, precisely the situation a multi-source system is in by construction: a website’s prose, a PDF’s mixed prose-and-table layout, a video transcript’s spoken, punctuation-light stream, and a source code file’s line-oriented structure differ enough that a strategy tuned for one is unlikely to suit another. In practice, multi-source systems tend to adopt a shared default chunker for prose-like sources and a distinct unit for whichever modality least resembles prose (most often code, by line, or transcript, by time/character band), rather than forcing one chunking function onto every modality.

Chunk overlap deserves separate comment from chunk size, since the two are often conflated as one hyperparameter. Overlap keeps a fact straddling a boundary recoverable, at a fixed, easily budgeted storage and embedding cost, unlike chunk size, whose effect on retrieval quality is corpus- and query-dependent. A further wrinkle specific to multi-source systems is that overlap’s benefit is not uniform across modalities: a transcript chunked by a time/character band, where segment boundaries already align with natural pauses in speech, needs overlap less than dense, boundary-agnostic prose extracted from a PDF’s continuous text flow. That suggests, though we know of no study measuring it directly, that a single global overlap parameter, like a single global chunk size, is a simplification a genuinely multi-modal chunking strategy would eventually need to relax.

An alternative to fixed-size chunking is to place boundaries adaptively, at points where adjacent sentences’ embeddings diverge beyond some threshold (“semantic chunking”), so a boundary tends to coincide with an actual topic shift rather than an arbitrary character count. Late chunking [15] is one instance of this adaptive-boundary family, and comparative evaluations [16] generally find adaptive strategies competitive with, but not uniformly better than, well-tuned fixed-size chunking, at the cost of an extra embedding pass over candidate boundaries. For a multi-source system this tradeoff is itself modality-dependent: sentence-level boundaries are cheap for prose but ill-defined for source code, where syntax-aware boundaries (function or class scope) are the more natural criterion.

A different response to the same problem is to stop treating chunk granularity as a single choice at all. RAPTOR [17] recursively clusters and summarizes chunks into a tree, so retrieval can draw on a fine-grained leaf chunk, a paragraph-level summary, or a whole-document summary depending on the abstraction level a query needs — a broad question answered from a high node, a specific one from a leaf, without forcing a single chunk size to serve both. The motivation resembles the overview chunk of Section IX — broad queries are poorly served by fine-grained chunks — but RAPTOR builds a full tree through recursive LLM summarization, a heavier device than one deterministic overview chunk derived from source statistics.

VI. MODALITY-SPECIFIC RETRIEVAL CHALLENGES

A. Web Documents

Retrieval over crawled web content inherits two problems documents-at-rest avoid. Content must be extracted from a live, often JavaScript-rendered page rather than a static file, and the crawl frontier must be bounded by domain, depth, or page count to avoid indexing unbounded linked content. Both sit upstream of retrieval yet decide what becomes retrievable: a page that fails to render or falls outside the crawl’s limits is silently absent from the corpus. Near-duplicate content is a subtler hazard: templated navigation and boilerplate repeated across a site’s pages can dominate a naive chunk’s lexical signature, so without explicit stripping a retriever may return several near-identical, low-information chunks in place of one genuinely relevant one.

B. Video and Transcript QA

Two features set transcript-based video QA apart. The retrieval unit is inherently temporal: a useful citation names a timestamp range rather than a document title, and benchmarks such as TVQA [18] enforce this by scoring temporally localized evidence rather than whole-video labels. The text is also less clean when it comes from an automatic transcript rather than a human-authored one, carrying transcription errors, missing punctuation, and imprecise sentence boundaries that other sources’ text usually avoids, bearing directly on how far a system should trust a confidence score computed uniformly across all its sources (Section VII). LLoVi [19] handles long-form video QA by converting video to language and reasoning over it with a long-context model, the same reduction to text plus timestamps a transcript-based multi-source system relies on. Chunk size interacts awkwardly with timestamp labeling: a short transcript chunk



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

yields a precise timestamp but a weak, fragmentary retrieval signal, while a longer chunk strengthens the signal at the cost of a timestamp that only loosely locates the answer within a multi-minute span.

C. Code-Repository Retrieval

Code differs from prose in ways that shape both extraction and retrieval. A repository is a directory tree of files linked by import and dependency relationships, and a chunk boundary aligned to function or class scope serves retrieval better than an arbitrary character cut, motivating line- or syntax-aware chunking over prose's character-budget chunking. Semantic code search was an early target for dense methods, exemplified by CodeSearchNet [20], and has since grown into retrieval-augmented code completion that must decide not just what to retrieve but *where in the repository* to look. RepoCoder [21] interleaves retrieval and generation so an initial, imperfect generation drives better subsequent retrieval, while Repoformer [22] learns when repository-level retrieval is worth its cost. A repository's own scaffolding, README, directory layout, and conventions, is itself retrievable signal distinct from any single file's contents. One point receives only partial treatment in the literature: an import or call-graph edge between two files is direct evidence of relatedness that text-similarity retrieval cannot perceive, so a chunk-level retriever blind to the dependency graph discards structure a compiler already has.

D. Multi-Document and Multi-Source Fusion

Retrieving well from any single source is necessary but not sufficient once a query may be legitimately answered by evidence spread across several sources at once. End-to-end training of a reader and retriever jointly over multiple documents [23] treats multi-document evidence aggregation as a first-class training objective rather than an artifact of independently ranking single documents and concatenating the top results. Reciprocal Rank Fusion [13], discussed above as a reranking technique, is equally a multi-source fusion technique when its inputs are ranked lists retrieved independently *per source*. A distinguishing requirement of multi-source (as opposed to merely multi-document) retrieval is source-level fairness: a user who explicitly selects several sources jointly expects each to be actually considered, which pure top-*k*-by-score selection does not guarantee if one source's evidence is systematically weaker or sparser than another's. This requirement is orthogonal to whichever fusion technique is used: unweighted rank fusion, learned re-weighting, or a joint reader-retriever objective can all be layered with an explicit source-coverage constraint applied after scoring, since it acts on provenance metadata that fusion techniques carry through but rarely use.

VII. CONFIDENCE, CALIBRATION, AND ABSTENTION

A grounded QA system that always answers, regardless of retrieval quality, will eventually answer confidently from irrelevant or insufficient evidence. Selective prediction under distribution shift [24] frames this as a decision the system should make explicitly, answer or abstain, rather than an emergent property of whatever the generator happens to produce. Whether a language model's own reported confidence can be trusted for this is itself contested: [25] finds larger models can be reasonably well calibrated on some self-assessment tasks, but hallucination surveys [3] document broad and persistent failure modes where fluent, confident-sounding generation is simply wrong — arguing for grounding the abstention decision in retrieval evidence rather than the generator's self-report wherever that evidence is available. Retrieval augmentation has itself been shown to reduce hallucination in open-ended generation [26], further supporting gating on retrieval quality.

Two recent systems fold the abstention decision into the model's own reasoning rather than an external gate. Self-RAG [27] trains a single model to emit reflection tokens that decide at generation time whether to retrieve and whether its own output is supported by what was retrieved, rather than computing that decision upstream from a fixed formula. Corrective RAG [28] takes a middle path, adding a lightweight retrieval evaluator that labels evidence as correct, incorrect, or ambiguous and triggers a distinct action per label (using the evidence, discarding it for a web search, or blending the two) — closer to the confidence-gate pattern described here, with a learned classifier in place of a closed-form heuristic. Neither has been evaluated on a multi-source, multi-modal corpus, though both suggest a learned calibration component that could be applied per source type.

Multi-source retrieval complicates calibration in a way single-corpus systems never encounter: when sources have systematically different retrieval-quality distributions, the noisier text of an automatic video transcript against the cleaner text of a crawled webpage, say, a single fixed threshold applied uniformly will be miscalibrated for at least one of them. A system must either normalize confidence per source type before thresholding, or make an explicit, stated exception where uniform thresholding is known to misfire.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In practice abstention is not binary: a system can refuse outright, answer with an attached caveat, or ask a clarifying follow-up. The caveat option weakens the guarantee from “no unsupported claims reach the user” to “flagged but still delivered,” which can be the more defensible choice for noisy transcripts, where a strict refusal threshold would decline a disproportionate share of genuinely answerable questions. Most selective-prediction work [24] treats abstention as binary, leaving a graded, per-source-type policy to individual builders’ judgment.

VIII. COMPARATIVE SYNTHESIS

Sections III–VII discussed each technique family on its own terms; Table I lays them side by side across the same five dimensions, to make explicit what each does and, just as importantly, does not address. The “multi-source limitation” column is our own assessment, not a claim from the cited papers: none of these techniques targeted a combined web/PDF/video/code corpus, so each entry marks a gap our framing exposes rather than a shortcoming of the original work.

Three patterns emerge reading the table across rather than down. First, confidence/abstention handling is almost entirely absent (two of five rows address it, and only as a within-modality mechanism), consistent with Section VII’s point that multi-source calibration is largely unaddressed. Second, every technique touching chunking granularity does so for one modality at a time; no row varies its strategy by source type, precisely the “heterogeneous chunking” property Section II identified as distinguishing multi-source retrieval. Third, of the families surveyed, only rank fusion and the multi-document reader-retriever approach (Section VI) engage with fusing evidence across more than one source at all, and neither was evaluated with genuinely heterogeneous sources.

TABLE I
REPRESENTATIVE TECHNIQUES COMPARED ACROSS THE DIMENSIONS THIS SURVEY ORGANIZES AROUND

Technique	Retrieval	Chunking granularity	Modality	Confidence handling	Multi-source limitation (our assessment)
DPR / RAG / FiD [5], [1], [6]	Dense bi-encoder + generator	Passage-level, not addressed	Text (single corpus)	Not addressed	Assumes one homogeneous corpus, no provenance concept
ColBERT / CLEAR / RRF [10], [11], [13]	Late-interaction / learned hybrid / rank fusion	Passage-level / N/A	Text / modality-agnostic	Not addressed	Scales per candidate or fuses lists; neither guarantees per-source coverage
RepoCoder / Repoformer [21], [22]	Iterative / selective dense-sparse	Code-block (line/function scale)	Code only	Repoformer: binary retrieve-or-not gate	Single-repository scope; no cross-modality fairness
TVQA / LLoVi [18], [19]	Dense over transcript	Timestamp-aligned segment	Video only	Not addressed	Single modality; no cross-source combination
Self-RAG / CRAG [27], [28]	Adaptive (learned) / dense + web fallback	Passage-level	Text	Learned self-reflection / retrieval-evaluator classes	Learned components trained/evaluated on one modality only

IX. CASE STUDY: SAGE

SAGE (Semantic Analysis and Generation Engine) ingests four source types, crawled multi-page websites, PDFs, YouTube video transcripts, and GitHub repositories, through modality-specific extractors terminating in a common chunk representation. It embeds all chunks with the same sentence encoder [7], [8] and answers questions against one, several, or all of a user’s sources through a single retrieval, reranking, confidence-gating, and generation pipeline. Rather than repeat its full architecture, we use three of its design decisions as concrete instances of ideas introduced earlier.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Overview chunks as a chunking-strategy instance. Section V noted that no single chunking strategy dominates across the modalities a multi-source system ingests, and that broad, non-specific queries pose a particular retrieval problem: a question like “what is this about” has poor lexical and semantic overlap with any single content chunk, because the answer lives in a source’s overall structure rather than any one passage. SAGE addresses this without tuning the chunk size at all. It adds a synthetic, deterministic *overview chunk* to every source at ingestion time, built from the source’s own statistics and excerpts (crawled page titles, transcript timestamps, or, for a GitHub repository, a live-introspected on-disk directory tree and detected project conventions), with no additional generative model call. In the terms of Section V this is a chunking-strategy decision, not generation-time summarization: an extra chunk that changes what is retrievable and extends to a new modality without training data or fine-tuning.

The diversity floor as a multi-source-fusion instance. Section VI identified source-level fairness as a requirement specific to multi-source (not merely multi-document) retrieval: a query naming several sources should not silently omit one because its single best-matching chunk scored lower than another source’s several good chunks. SAGE addresses this with a diversity floor: before a fixed number of citation slots is filled by reranked score, each explicitly combined source has its single best-scoring chunk reserved, and only the remaining slots are filled by score across all sources — a direct instance of the score-then-select-under-a-coverage-constraint pattern anticipated in Sections IV and VI. A live run confirmed the behavior beyond unit tests: a query spanning a website and a topically unrelated code repository still cited both, and a three-source query cited each source rather than dropping the weakest-scoring one. Such fairness failures look benign when a reranker is examined alone and surface only under genuinely combined queries.

Confidence gating with a stated per-modality exception. Consistent with Section VII, SAGE computes a single retrieval-derived confidence score from its top-ranked and top-three blended scores plus a small evidence-count bonus, refusing to answer in grounded mode below a fixed threshold. Video is the one exception: its chunks are exempted from hard refusal and instead receive a lower-confidence warning, since automatically transcribed text is measurably noisier than the system’s other sources. This is the per-source exception anticipated in Section VII.

Evidence, not a benchmark. SAGE’s test suite (95 tests across 12 files, covering chunking, URL normalization, reranker and confidence sanity checks, per-modality overview generation, and an end-to-end API suite) shows the mechanisms above are implemented and exercised, not a quantitative retrieval- or answer-quality result. SAGE reports no benchmark score against a labeled dataset, and none is constructed here, since a combined-modality benchmark able to support one does not yet exist (Section X).

At the product level, SAGE sits between two familiar categories: consumer “chat with your sources” tools such as NotebookLM [29], and composability frameworks such as LangChain [30] and LlamaIndex [31], matching neither, being an opinionated pipeline rather than a finished application or a general library. Table II compares them on modality coverage and whether a fairness mechanism is built in; NotebookLM’s is marked undisclosed since it is closed-source.

TABLE II
MODALITY COVERAGE AND FAIRNESS-MECHANISM COMPARISON

System	Web	PDF	Video	Code	Multi-source fairness mechanism
SAGE	✓	✓	✓	✓	Built-in: diversity floor (Section IX)
NotebookLM [29]	✓	✓	✓	—	Undisclosed (closed-source)
LangChain [30]	✓†	✓†	✓†	✓†	Not built-in; assemblable from library components
LlamaIndex [31]	✓†	✓†	✓†	✓†	Not built-in; assemblable from library components

† Available via the framework’s connector/loader ecosystem rather than as a single default pipeline; each modality requires the developer to select and wire in the relevant integration. — = not offered as a documented first-party feature as of this writing.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

X. OPEN CHALLENGES

Several problems recur across the surveyed systems and papers without a settled answer.

Cross-modal fusion at the scoring level remains largely heuristic. Blended semantic-lexical scores, rank fusion, and learned combination methods (Section IV) each rest on different, usually untested assumptions about whether a score from one modality's chunk is comparable in magnitude to a score from another's. SAGE's diversity floor avoids the question by guaranteeing representation instead of asserting cross-modality score comparability, yielding a fairness guarantee at the expense of a true relevance ordering across sources.

Confidence calibration across heterogeneous corpora is equally unsettled. Per-source-type exceptions, such as the video-transcript exception in Section IX, are practical patches rather than calibrated solutions; no published method jointly calibrates one retrieval-confidence score across genuinely heterogeneous modalities using a single labeled dataset.

Incremental re-indexing has received little attention in the multi-source case. Chunking strategies (Section V) are typically presented as one-time operations, yet a source that changes over time, a recrawled website or an updated repository, forces a decision about which chunks to preserve and which to re-embed. Content-addressed chunk identifiers make that tradeoff explicit but do not resolve it.

Evaluation for multi-source, multi-modal RAG trails single-corpus evaluation. Most retrieval and QA benchmarks, including several in Section VI, are modality-specific, and no widely adopted benchmark jointly measures retrieval quality, citation correctness, and confidence calibration across web, document, video, and code sources combined in one query, which would let systems like SAGE be compared on equal terms rather than by architecture description alone. Standard metrics (Recall@ k , MRR, nDCG) presume a labeled query-document pair from one corpus. Extending them to a combined-source query first requires deciding whether relevance judgments are pooled across sources or required per source, the latter capturing the fairness property of Sections VI and IX, and no standard convention yet governs that choice.

Fine-tuning against a multi-modal negative distribution, raised in Section III, is a training-side gap: widely used open-source sentence encoders, including the one in our case study, were trained on single-corpus question-passage pairs, and fine-tuning one encoder against hard negatives spanning prose, transcript, and code appears unexplored; the common alternative is separate per-modality encoders reconciled afterward.

Dependency-aware retrieval for structured sources, raised in Section VI's discussion of code, generalizes beyond code: a website's link graph and a PDF's cross-references are structural signals most chunk-level dense retrievers discard just as they discard a repository's import graph. GraphRAG [32] takes structure this seriously within a single corpus, extracting an entity-relationship graph via LLM and pre-summarizing its communities so broad questions can be answered without any single retrieved chunk; extending that across genuinely different modalities, linking a code symbol to the PDF paragraph documenting it, for instance, remains open.

XI. CONCLUSION

Grounded question answering over heterogeneous, multi-source knowledgebases relies on the same core techniques as single-corpus RAG: dense retrieval, hybrid reranking, chunking, and confidence-gated generation. Each must be reconsidered once the corpus is assembled from structurally different sources at query time. Chunking must accommodate prose, transcripts, and code without assuming one size fits all; reranking must pair with an explicit source-fairness mechanism as soon as queries can name several sources at once; confidence estimation must either normalize across source-type-specific quality distributions or state its exceptions explicitly. SAGE's overview chunks and diversity floor served as checkable instances of these adjustments.

The highest-leverage next steps follow from the open problems of Section X: a standardized multi-modal benchmark should score retrieval quality, citation correctness, and confidence calibration across web, document, video, and code sources together rather than one at a time; retrieval confidence should be calibrated jointly across heterogeneous corpora instead of patched per source type; and retrieval/reranking pipelines should choose their chunking and fusion strategy adaptively by source modality rather than applying one global configuration to every ingested source.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2005.11401.
2. Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," 2023. arXiv:2312.10997.
3. Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, art. 248, 2023.
4. P.-S. Huang *et al.*, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2013, pp. 2333–2338.
5. V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020, pp. 6769–6781. arXiv:2004.04906.
6. G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proc. EACL*, 2021. arXiv:2007.01282.
7. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019. arXiv:1908.10084.
8. W. Wang *et al.*, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *NeurIPS*, 2020. arXiv:2002.10957.
9. R. Nogueira and K. Cho, "Passage re-ranking with BERT," 2019. arXiv:1901.04085.
10. O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf.*, 2020. arXiv:2004.12832.
11. L. Gao, Z. Dai, and J. Callan, "Complement lexical retrieval model with semantic residual embeddings," 2020. arXiv:2004.13969.
12. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
13. G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. 32nd Int. ACM SIGIR Conf.*, 2009, pp. 758–759.
14. A. Jimeno-Yepes *et al.*, "Financial report chunking for effective retrieval augmented generation," 2024. arXiv:2402.05131.
15. M. Günther *et al.*, "Late chunking: Contextual chunk embeddings using long-context embedding models," 2024. arXiv:2409.04701.
16. C. Merola and J. Singh, "Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation," in *2nd Workshop on Knowledge-Enhanced Information Retrieval, ECIR*, 2025. arXiv:2504.19754.
17. P. Sarthi *et al.*, "RAPTOR: Recursive abstractive processing for tree-organized retrieval," in *Proc. ICLR*, 2024. arXiv:2401.18059.
18. J. Lei *et al.*, "TVQA: Localized, compositional video question answering," in *Proc. EMNLP*, 2018. arXiv:1809.01696.
19. C. Zhang *et al.*, "A simple LLM framework for long-range video question-answering," in *Proc. EMNLP*, 2024. arXiv:2312.17235.
20. H. Husain *et al.*, "CodeSearchNet challenge: Evaluating the state of semantic code search," 2019. arXiv:1909.09436.
21. F. Zhang *et al.*, "RepoCoder: Repository-level code completion through iterative retrieval and generation," in *Proc. EMNLP*, 2023. arXiv:2303.12570.
22. D. Wu *et al.*, "Repoformer: Selective retrieval for repository-level code completion," in *Proc. ICML*, 2024. arXiv:2403.10059.
23. D. S. Sachan *et al.*, "End-to-end training of multi-document reader and retriever for open-domain question answering," in *NeurIPS*, 2021. arXiv:2106.05346.
24. A. Kamath, R. Jia, and P. Liang, "Selective question answering under domain shift," in *Proc. ACL*, 2020. arXiv:2006.09462.
25. S. Kadavath *et al.*, "Language models (mostly) know what they know," 2022. arXiv:2207.05221.
26. K. Shuster *et al.*, "Retrieval augmentation reduces hallucination in conversation," in *Findings of EMNLP*, 2021, pp. 3784–3803. arXiv:2104.07567.
27. A. Asai *et al.*, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. ICLR*, 2024. arXiv:2310.11511.
28. S.-Q. Yan *et al.*, "Corrective retrieval augmented generation," 2024. arXiv:2401.15884.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

29. Google, "NotebookLM: How to try Google's experimental AI-first notebook," Google Blog, 2023. [Online]. Available: <https://blog.google/technology/ai/notebooklm-google-ai/>
30. LangChain, Inc., "LangChain documentation." [Online]. Available: <https://docs.langchain.com/>
31. LlamaIndex, "LlamaIndex documentation." [Online]. Available: <https://developers.llamaindex.ai/python/framework/>
32. D. Edge *et al.*, "From local to global: A graph RAG approach to query-focused summarization," 2024. arXiv:2404.16130.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details