ISSN (Print) 1226-1750
ISSN (Online) 2233-6656
Journal of Magnetics 31(3), 467-476 (2026) https://doi.org/10.4283/JMAG.2026.31.3.467

Retrieval-Augmented Question Answering on the Grain-Boundary Diffusion Literature of NdFeB Permanent Magnets: A Comparison of Commercial and Local Large Language Models

Chunghee Nam[1]*, Tae-Hoon Kim[2], and Jung-Goo Lee[2]

[1]*Department of Electrical and Electronic Engineering, Hannam University, Daejeon 34430, Republic of Korea*
[2]*Nano Materials Research Division, Korea Institute of Materials Science (KIMS), Changwon 51508, Republic of Korea*

(Received 13 July 2026, Received in final form 3 September 2026, Accepted 3 September 2026)

The grain-boundary diffusion process (GBDP) is the leading route to high coercivity in NdFeB permanent magnets with reduced heavy-rare-earth (HRE) content, and its rapidly growing literature is increasingly difficult to synthesize. We built a retrieval-augmented-generation (RAG) system over a corpus of 311 GBDP papers that uses lightweight, CPU-only sentence embeddings, section-aware chunking, and source-cited answer generation, and used it to compare three large language models: the commercial GPT-4o and Gemini 3.5 Flash, and the locally served open-weight Llama 3.2. To separate the language model from the retriever, we adopted a controlled design in which retrieval was held constant so that all three models received an identical evidence set for each question, and answers were assessed qualitatively for faithfulness to the retrieved passages and for citation correctness. Across three representative questions (the advantage of GBDP, its diffusion sources and the role of HRE), the models' factual grounding was comparable once the retrieved context was informative; removing duplicate documents from the index improved the retrieval of answer-bearing passages and reduced unsupported additions in the evaluated cases. The principal difference was citation reliability: both commercial models attributed every claim correctly, whereas the local model extracted correct content but omitted or misapplied source citations. Faithfulness thus emerged as a property of the retrieval stage as much as of the model. We conclude that RAG is a practical aid for consolidating dispersed GBDP knowledge, and that a local open-weight model is a viable privacy-preserving option provided its attribution is verified.

Keywords : NdFeB, grain boundary diffusion process, coercivity, heavy rare earth, retrieval-augmented generation, large language models

1. Introduction

Nd–Fe–B (NdFeB) sintered magnets exhibit the highest maximum energy product of any commercial permanent magnet and are indispensable to a wide range of energy and mobility technologies, including traction motors for electric vehicles, direct-drive wind-turbine generators, and a variety of high-efficiency actuators and sensors [1, 2]. Many of these applications operate at elevated temperatures, where the intrinsic coercivity of the magnet decreases sharply; maintaining a sufficiently high coercivity at the working temperature is therefore one of the central challenges in NdFeB magnet design [3-5].

The conventional route to higher coercivity has been the substitution of the heavy rare earth (HRE) elements dysprosium (Dy) and terbium (Tb) for neodymium in the $Nd_2Fe_{14}B$ matrix phase, because the $(Nd,Dy)_2Fe_{14}B$ and $(Nd,Tb)_2Fe_{14}B$ phases possess a substantially larger magnetocrystalline anisotropy field [6-8]. However, this approach is costly and unsustainable: Dy and Tb are among the most supply-constrained and price-volatile of the rare earths, and uniform alloying also lowers the remanence and energy product because the HRE moments couple antiparallel to the iron sublattice [6-8]. Reducing the HRE content while preserving coercivity has thus become a strategic objective for both the magnet industry and materials research.

The grain-boundary diffusion process (GBDP) has emerged as the most effective response to this objective. In GBDP, an HRE-bearing source is applied to the surface

©The Korean Magnetics Society. All rights reserved.
*Corresponding author: Tel: +82-42-629-7472
email: chnam@hnu.ac.kr

This paper was presented at ISAMMA 2026.

© 2026 Journal of Magnetics

of a sintered magnet and diffused along the grain boundaries at elevated temperature, so that the HRE becomes concentrated in a thin shell around each $Nd_2Fe_{14}B$ grain rather than being distributed uniformly through the grain interior [9]. Because coercivity is governed by the nucleation of reverse domains at the grain surface, this core–shell microstructure raises the coercivity with far less HRE than bulk alloying, and with a much smaller penalty in remanence [9]. The breadth of reported GBDP source chemistries reflects the intensity of this research effort, ranging from HRE metals, hydrides and fluorides to low-melting-point alloys that promote grain-boundary wetting and, more recently, light-rare-earth and non-rare-earth sources that aim to reduce or eliminate critical elements [4, 9, 10].

This rapid expansion has produced a large and heterogeneous body of literature. Key design knowledge, including which diffusion source to use, under what process conditions it should be applied, and through what mechanism it acts, is dispersed across hundreds of papers. For a researcher entering the field, synthesizing this knowledge is increasingly time-consuming, and conventional keyword search returns long lists of documents without consolidating the specific compositions, conditions and explanations they contain.

Large language models (LLMs) offer a new way to interrogate such a corpus. In particular, retrieval-augmented generation (RAG) couples an LLM to a fixed, domain-specific document collection: relevant passages are retrieved by semantic similarity and supplied to the model as context, and the model is instructed to answer using only those passages and to cite their sources [11]. By grounding the answer in retrieved evidence, RAG is intended to reduce the unsupported or fabricated statements ("hallucinations") to which unconstrained LLMs are prone, and to provide an audit trail of citations that the reader can verify. This makes RAG attractive for specialized scientific domains such as magnetic-materials research, where answers must be both technically accurate and traceable to primary sources.

A practical question, however, is which LLM to deploy. Commercial models accessed through an API, such as GPT-4o and Gemini, generally offer strong language capability but entail per-query cost, dependence on an

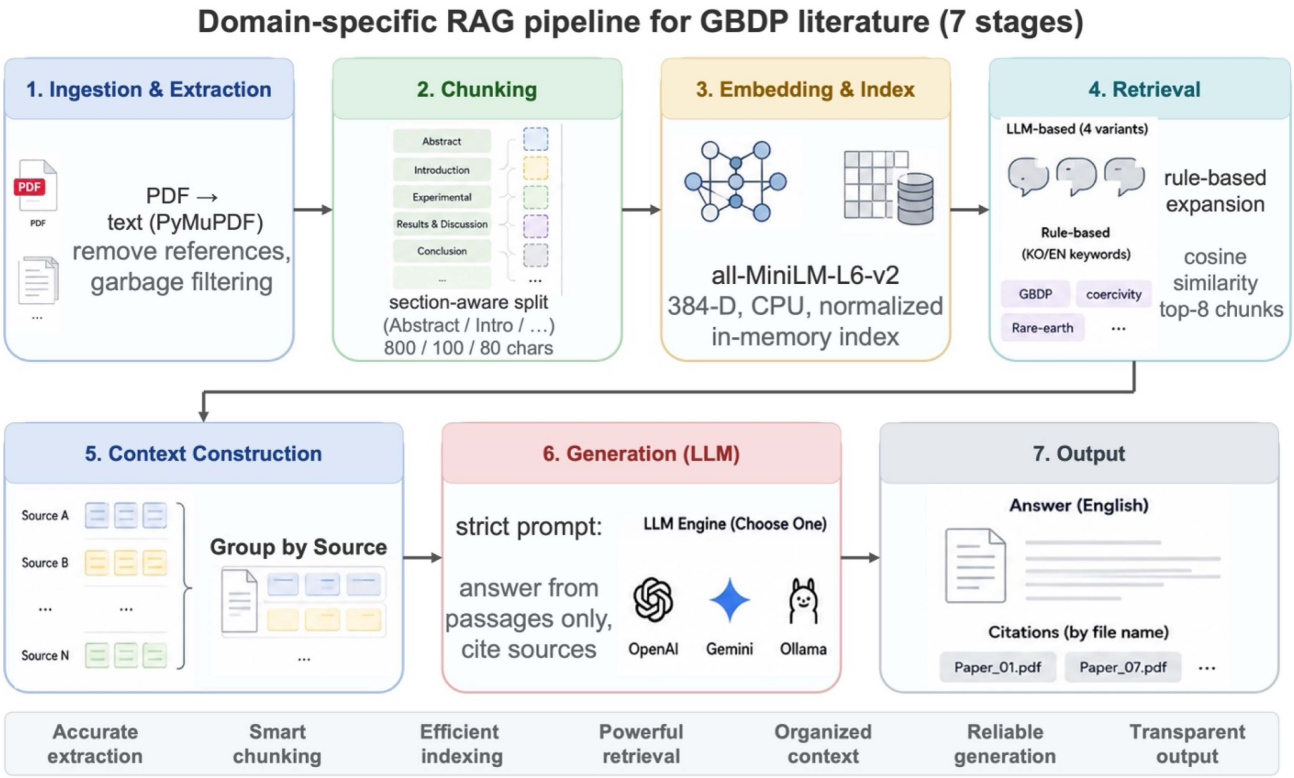


Fig. 1. (Color online) Overall architecture of the retrieval-augmented generation (RAG) system for the GBDP literature, comprising seven stages: (1) ingestion and text extraction from PDFs, (2) section-aware chunking, (3) embedding and indexing with a light-weight CPU sentence-transformer, (4) retrieval by cosine similarity with de-duplication and per-source control, (5) source-grouped context construction, (6) answer generation by the language model under a strict source-citing prompt, and (7) the English, file-cited output.

external service, and the transmission of queries to a third party. Open-weight models such as Llama can be run locally, which removes these constraints and is appealing to laboratories with data-confidentiality or budget limitations, but their reliability on a narrow technical corpus is less well established. For a materials-research setting, the decisive criteria are not fluency in the abstract but whether the model stays faithful to the supplied evidence and attributes its claims correctly, because researchers may act on the extracted information. How commercial and local models compare on these reliability axes for a specialized magnetics corpus has received little attention.

In this work we address that gap. We construct a RAG system over a corpus of 311 GBDP papers for NdFeB magnets, as shown in Fig. 1. We then use this system to compare the commercial GPT-4o, Gemini 3.5 Flash, and the locally served open-weight Llama 3.2 on three representative questions concerning the advantages, diffusion sources, and HRE mechanism of GBDP. To ensure a fair comparison, we adopt a controlled design in which the retrieval stage is held constant so that every model receives an identical evidence set, and any difference between the answers reflects only how each model reads and attributes that evidence, as shown in Fig. 2. The contributions are: (i) a CPU-only, domain-specific RAG system for the GBDP literature with section-aware chunking and source-cited generation; (ii) a controlled, fixed-retrieval protocol that isolates the language model from the retriever; (iii) a qualitative comparison of two commercial models and one local model on faithfulness and citation correctness; and (iv) a consolidated domain reference for GBDP distilled from the corpus. Unlike many previous comparisons of RAG systems or LLMs, which rely on general-purpose benchmarks and often vary both retrieval and generation, this study has three distinctive features. First, the evaluation is based on an expert-curated corpus of 311 papers on GBDP in NdFeB magnets, where terminology, units, and underlying mechanisms are highly domain-specific. Comparisons between commercial and local models in such a specialized scientific domain remain limited. Second, a fixed-retrieval protocol is used so that all models receive the same retrieved evidence, allowing differences in generation to be evaluated independently of retrieval variability. Third, the results indicate that measured faithfulness depends strongly on the quality of the retrieved evidence as well as on the choice of language model. Commercial and local models showed comparable factual grounding, whereas a clearer difference was observed in citation reliability. These results provide a domain-specific perspective that complements previous general RAG/LLM comparisons and studies of materials-focused language models.

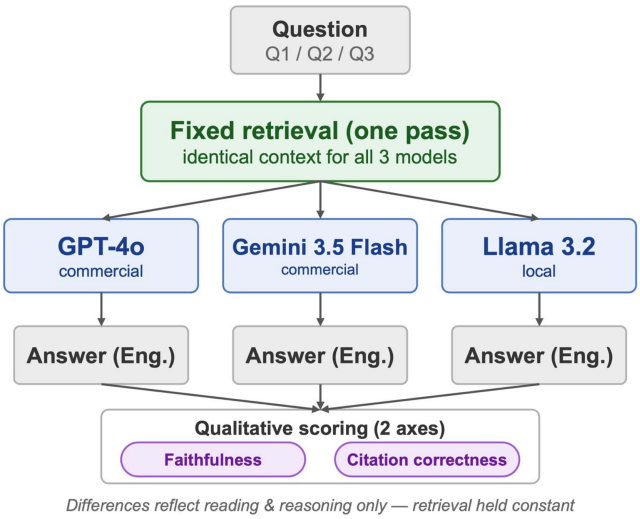


Fig. 2. (Color online) Controlled comparison design. For each question, retrieval is performed once to produce a single shared passage set, which is supplied identically to GPT-4o, Gemini 3.5 Flash and Llama 3.2. The three English answers are then scored qualitatively on faithfulness and citation correctness, so that any difference between answers reflects reading and reasoning only, with retrieval held constant.

2. Methods

2.1. Document corpus

A domain-specific corpus was assembled from the peer-reviewed literature on the grain-boundary diffusion process used to enhance the coercivity of NdFeB permanent magnets. The corpus comprised 311 full-text articles in PDF format, collected from major international publishers and spanning the period from the introduction of GBDP (2005) to 2026 [12]. The papers cover the principal GBDP source chemistries reported to date, including HRE metals, hydrides and fluorides, low-melting-point alloys, and the more recent light-rare-earth and non-rare-earth diffusion sources. The corpus served as the sole knowledge base for the RAG system; no external knowledge was provided to the language models at answer time.

2.2. Retrieval-augmented generation pipeline

The overall architecture is summarized in Fig. 1 and comprises seven stages: document ingestion and text extraction, chunking, embedding and indexing, retrieval, context construction, answer generation and output. All local components of the RAG pipeline, including text processing, embedding, retrieval, and local-model inference,

were executed on CPU without a dedicated GPU. GPT-4o and Gemini were accessed through their respective APIs.

Raw text was extracted from each PDF using the PyMuPDF library. To restrict the knowledge base to substantive scientific content, the reference list and all subsequent material were removed, repeated whitespace was normalized, and a heuristic filter discarded documents whose extracted text was too sparse to be usable. Each cleaned document was segmented into overlapping chunks using a section-aware scheme: section headings (Abstract, Introduction, Experimental, Methods, Results, Discussion, Conclusion and related headings) were detected by regular expression and used as preferred split points, with paragraph and sentence boundaries used otherwise. The default configuration used a target chunk size of 800 characters, an overlap of 100 characters and a minimum chunk length of 80 characters; the complete corpus yielded 3,112 chunks.

Each chunk was encoded into a 384-dimensional, L2-normalized vector using the all-MiniLM-L6-v2 sentence-transformer model, and the vectors were stored in a lightweight in-memory index serialized to disk. Relevance between a query and a chunk was measured by cosine similarity. For a given query, the top-15 chunks were retrieved, de-duplicated, and capped at three chunks per source to prevent any single paper from dominating; the eight highest-ranked chunks were retained as the evidence set. The selected chunks were grouped by source document, each prefixed by its source filename, to form the context. A fixed instruction prompt directed the model to answer using only the supplied passages, to refrain from using prior knowledge, to cite the source filename after each claim, and to answer in English. The maximum generation length was 2,000 tokens (8,192 tokens for the reasoning-based Gemini model; see Section 2.5) and the decoding temperature was set to zero for all models to minimize sampling variability and promote evidence-grounded output.

2.3. Controlled comparison design

The objective is to compare how different language models read and reason over the same retrieved evidence, rather than to compare end-to-end systems whose retrieved evidence may differ. Because the retrieval stage can itself depend on the model used for query expansion, a naive model-by-model comparison would confound differences in retrieval with differences in generation. To remove this confound, retrieval was held constant across all models, as shown in Fig. 2. For each question, query processing and passage retrieval were performed once, using rule-based query expansion only and no model-dependent expansion, producing a single fixed evidence set. This identical context, together with the identical prompt, was supplied to each of the three models. Consequently, any difference between the model answers is attributable solely to how each model reads, synthesizes and attributes the same passages. The index was also de-duplicated so that redundant copies of the same article did not occupy retrieval slots.

2.4. Query set

Three questions were used to probe complementary capabilities, each addressing a core aspect of GBDP science: Q1, “What is the main advantage of GBDP for NdFeB permanent magnets?” (a conceptual question); Q2, “What are the main diffusion sources used in GBDP?” (an extractive, enumerative question); and Q3, “Why is the GBDP effect enhanced when heavy rare earth (HRE) elements are used?” (a mechanistic-reasoning question). By design, Q2 spans three materials-science categories that must be kept distinct: (i) diffusion elements/sources (the heavy rare-earth elements Dy and Tb, and their alloy sources); (ii) source-application/deposition methods (e.g., sputtering, electrophoretic deposition, adhesive coating, screen printing); and (iii) grain-boundary additives that improve grain-boundary-phase wettability (Cu, Al, Ga). These categories are properties of the question and the retrieved evidence, not of the models, and each model’s answer is evaluated against them separately. All questions were posed in English, and all answers were evaluated in English.

2.5. Language models compared

Three models were compared under the controlled design: two commercial models accessed through their APIs, GPT-4o (OpenAI, model gpt-4o) and Gemini 3.5 Flash (Google, model gemini-3.5-flash), and one open-weight model run locally, Llama 3.2 (Meta, served through Ollama as llama3.2), on the same workstation. All three received identical input for each question, with the decoding temperature fixed at zero; no model-specific prompt tuning was applied. The maximum output-token budget was 2,000 tokens for GPT-4o and Llama 3.2. Because Gemini 3.5 Flash is a reasoning model whose internal reasoning tokens are counted against the output-token budget, it was allocated a larger budget of 8,192 tokens so that this internal reasoning did not truncate the visible answer; this affects only the length cap, not the decoding (temperature zero) or the shared retrieved context.

2.6. Evaluation rubric and scoring protocol

Because the study targets the reliability rather than the breadth of generated answers, the comparison focused on

Question	GPT-4o		Gemini 3.5 Flash		Llama 3.2	
	F	C	F	C	F	C
Q1 Advantage of GBDP	✓	✓	✓	✓	✓	✗
Q2 Diffusion sources	✓	✓	✓	✓	✓	✓
Q3 HRE mechanism	✓	✓	✓	✓	✓	△

F = faithfulness C = citation correctness

✓ grounded/correct △ partial ✗ unsupported/missing

Fig. 3. (Color online) Qualitative rubric scoring of the nine answers (three models × three questions) on two axes: faithfulness (F) to the retrieved passages and citation correctness (C). Symbols denote grounded/correct (✓), partial (△) and unsupported/missing (✗). All three models were faithful on every question; the two commercial models cited correctly throughout, whereas the local model omitted citations on Q1 and gave a spurious citation on Q3.

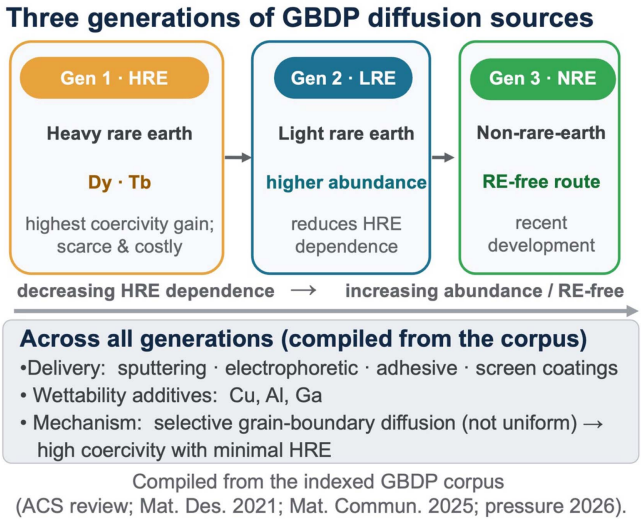


Fig. 4. (Color online) Three generations of GBDP diffusion sources, compiled from the indexed corpus: the first generation based on heavy rare earth (HRE; Dy, Tb), the second on light rare earth (LRE) of higher natural abundance, and the third on non-rare-earth (NRE) elements, with decreasing HRE dependence across the progression. The lower panel lists the cross-cutting facts present in the corpus: deposition methods, the Cu/Al/Ga wettability additives, and the selective grain-boundary diffusion mechanism. This figure also serves as the domain reference against which faithfulness was judged.

two qualitative axes central to RAG evaluation: faithfulness, the degree to which the answer is grounded in the supplied passages and free of unsupported content; and citation correctness, whether each claim is attributed to the correct source filename, as required by the prompt. We further distinguish faithfulness from completeness/relevance, defined as the degree to which an answer addresses the full scientific intent of the question. An answer can be fully faithful to the retrieved passages yet still be incomplete when the corpus does not contain the information required to address the question fully. Q3 is

Table 1. Condensed comparison of the answers from GPT-4o, Gemini 3.5 Flash and Llama 3.2 to the three questions, under the identical retrieved context.

	GPT-4o	Gemini 3.5 Flash	Llama 3.2 (local)
Q1: Advantage of GBDP	Selective diffusion of HRE (Dy/Tb) along grain boundaries rather than uniform modification; significantly enhances coercivity with minimal HRE. Cites ACS, pressure.	Selective diffusion of Dy/Tb (not uniform); coercivity enhancement with minimal HRE, addressing scarcity and cost. Cites ACS, pressure.	Selective diffusion and minimal-HRE gain, then drifts into the diffusion equation and ambient-pressure detail. No source citations.
Q2: Diffusion sources (three categories: • diffusion elements: Dy, Tb • deposition methods • GB additives: Cu, Al, Ga)	Diffusion elements: Dy, Tb. Deposition methods: sputtering, electrophoretic, adhesive, screen. GB additives: not listed. Cites ACS, mtcomm.	Diffusion elements: HRE (Dy, Tb), selectively diffused along the GB phase; Tb as a specific source; plus additional RE/NRE diffusion routes. Deposition methods: not explicitly listed. GB additives: not listed. Cites ACS, MatCharc, pressure, mtcomm, matdes.	Diffusion elements: Dy, Tb. Deposition methods: not listed. GB additives: Cu, Al, Ga (improve GB-phase wettability). Cites ACS, mtcomm.
Q3: HRE mechanism	HRE diffuse substantially faster along grain boundaries than through the bulk; predominantly grain-boundary diffusion. Cites ACS.	Selective diffusion (ACS); HRE diffusion along the GBP (MatCharc); higher grain-boundary diffusion rate vs. bulk. Cites ACS, MatCharc.	Higher grain-boundary diffusion rate (verbatim quote from ACS); also cites a journal-header fragment. Cites ACS (valid), actamat (spurious).

used below as a representative case of this distinction. The nine answers (three models × three questions) were assessed by domain-expert inspection: every substantive claim was checked against the retrieved passages, and each citation was verified against the source from which the claim originated. The consolidated domain reference in Fig. 4 was used as a contextual summary of the corpus rather than as an independent ground truth. Results are reported as a per-question, side-by-side comparison (Table 1 and Fig. 3) rather than as aggregate numerical scores, in keeping with the small, deliberately focused question set.

3. Results

3.1. Retrieved context

For each question, the pipeline retrieved eight chunks from an index comprising 311 papers and 3,112 chunks, and the same evidence set was provided to all three models. To quantify the effect of index de-duplication directly, we compared, for each question, the eight retrieved slots before and after removing content-identical duplicate documents (documents whose full extracted text is identical but stored under different filenames). Before de-duplication, two of the eight retrieved slots (25%) were occupied by content-duplicate documents in all three questions; after de-duplication these duplicate occupations were eliminated (zero of eight) for every question, so that the retrieved evidence for each answer was drawn from distinct source documents (Table 2). This confirms that de-duplication improved the retrieval of answer-bearing passages by preventing redundant copies of the same article from consuming retrieval slots, rather than by changing the language model.

Two properties of the retrieved context are important for interpreting the answers. First, the passages contained explicit statements of GBDP's benefit, namely that the key advantage is the selective diffusion of HRE (Dy and Tb) along grain boundaries rather than uniform modification of the magnet. GBDP also enhances coercivity while consuming only minimal HRE, thereby addressing the scarcity and cost of Dy and Tb. The retrieved passages also described deposition methods, including sputtering, electrophoretic coating, adhesive coating, and screen coating. In addition, they identified Cu, Al, and Ga additives as factors that improve grain-boundary-phase wettability. Second, for Q3 the passages explained the kinetic reason that HRE diffuses predominantly along grain boundaries (a much higher grain-boundary diffusion rate than in the bulk, because diffusion proceeds above the lower-melting Nd-rich phase), but did not contain the magnetocrystalline-anisotropy mechanism that explains why HRE in the grain shell raises coercivity. A strictly passage-grounded answer to Q3 is therefore confined to the diffusion-kinetics interpretation.

3.2. Q1: Advantage of GBDP

The condensed answers are compared in Table 1. All three models produced grounded answers. GPT-4o and Gemini 3.5 Flash both identified the selective diffusion of Dy/Tb along grain boundaries and the enhancement of coercivity with minimal HRE consumption, each attributing the two parts to the correct sources. Llama 3.2 stated the same core advantage and the minimal-HRE point but then drifted into peripheral detail (the diffusion-coefficient expression and the effect of pressure on the grain-boundary network) and, critically, attached no source filenames to any of its claims, failing the citation requirement despite being faithful to the passages.

3.3. Q2: Diffusion sources

With the de-duplicated context, the passages supported all three categories that Q2 spans: the diffusion elements/sources (Dy, Tb); the source-application/deposition methods (sputtering, electrophoretic, adhesive and screen coating); and the grain-boundary additives that improve grain-boundary-phase wettability (Cu, Al, Ga). We assess each model's answer against these three categories separately. GPT-4o named the diffusion elements Dy and Tb and added the deposition methods, citing both relevant sources; it did not list the Cu/Al/Ga additives. Llama 3.2 named the diffusion elements Dy and Tb and the grain-boundary additives Cu, Al and Ga, and on this question it cited both sources correctly, a marked improvement over its citation behavior on Q1; it did not list the deposition methods. Gemini 3.5 Flash, given the larger output-token

Table 2. Effect of index de-duplication on the eight retrieved slots per question. "Duplicate-occupied slots" counts slots filled by content-identical documents stored under different filenames.

Question	Retrieved slots	Duplicate-occupied slots (before)	Duplicate-occupied slots (after)
Q1	8	2 (25%)	0 (0%)
Q2	8	2 (25%)	0 (0%)
Q3	8	2 (25%)	0 (0%)

budget appropriate to a reasoning model (Section 2.5), produced the most complete and fully cited answer: it reported the heavy rare-earth diffusion elements (Dy, Tb), their selective diffusion along the grain-boundary phase, Tb as a specific source, and additional rare-earth/non-rare-earth diffusion routes, attributing each point to the correct source. Its answer interpreted "diffusion sources" more broadly than the others by including the light rare-earth and non-rare-earth diffusion routes, but every claim remained grounded in the passages and was correctly attributed (Table 1). Distinguishing the three categories clarifies that no single model enumerated all three: each model covered the diffusion elements, but the deposition methods and the grain-boundary additives were covered unevenly across models.

3.4. Q3: HRE mechanism

Q3 is the representative case in which an answer can be fully faithful yet incomplete with respect to the scientific intent of the question. Because the magnetocrystalline-anisotropy mechanism was absent from the passages, a faithful answer could only give the diffusion-kinetics reason; the gap therefore originates in the corpus, not in the model. GPT-4o and Gemini 3.5 Flash both gave the kinetic reason correctly: they reported that the grain-boundary diffusion rate of HRE is substantially higher than in the bulk because diffusion occurs above the lower-melting Nd-rich phase, so the behavior is predominantly grain-boundary diffusion, with correct citation; Gemini additionally noted the selective-diffusion advantage and the facilitation of HRE diffusion along the grain-boundary phase. Llama 3.2 gave the same grounded kinetic claim and cited it correctly but additionally cited a second source for which the quoted text is only a journal copyright/header line rather than substantive content, making that citation spurious. All three answers are thus faithful to the passages; none is complete with respect to the anisotropy mechanism a magnetism specialist would expect, because that mechanism is not present in the indexed corpus. Rephrasing Q3 as "Why do HRE elements preferentially diffuse along grain boundaries?" would remove this mismatch, but it would also remove the very limitation of passage-grounded answering that Q3 is intended to expose.

3.5. Rubric summary

Figure 3 summarizes the per-question scoring on faithfulness (F) and citation correctness (C). On the de-duplicated context all three models were faithful on all three questions: no unsupported or fabricated content was observed, in contrast to a preliminary run on a non-de-duplicated index, in which thinner context had led models to supplement the passages with parametric knowledge. The two commercial models also met the citation requirement on every question. The locally served open-weight model remained the weakest on attribution: it omitted source filenames entirely on Q1 and produced a spurious citation on Q3, although it cited correctly on Q2. To complement the qualitative rubric with quantitative descriptors, we also report two per-model summary metrics derived from the same expert assessment (Table 3): the faithfulness rate, defined as the fraction of questions for which all substantive claims in the answer were supported by the retrieved passages, and the citation correctness rate, defined as the fraction of questions for which all substantive claims were attributed to the correct sources. All three models achieved a faithfulness rate of 3/3. The two commercial models achieved a citation correctness rate of 3/3, whereas the local model achieved 1/3, owing to missing citations on Q1 and a partly spurious citation on Q3. Given the deliberately small question set and single-corpus scope, these values are reported as descriptive measures rather than as a statistical benchmark.

Table 3. Per-model quantitative summary across the three questions, based on the expert assessment. Faithfulness was defined as the fraction of questions for which all substantive claims were supported by the retrieved passages. Citation correctness was defined as the fraction of questions for which all claims were attributed to the correct sources.

Model	Faithfulness (Q passed)	Citation correctness (Q passed)
GPT-4o	3 / 3	3 / 3
Gemini 3.5 Flash	3 / 3	3 / 3
Llama 3.2 (local)	3 / 3	1 / 3

Two observations follow. First, improving retrieval quality by de-duplicating the index so that the genuinely informative passages were retrieved directly improved measured faithfulness because the models no longer needed to fill gaps from prior knowledge. Faithfulness is therefore as much a property of the retrieval stage as of the model. Second, with adequate context the principal remaining difference between the commercial and local models was citation reliability rather than factual grounding.

4. Discussion

4.1. Retrieval quality strongly influences measured faithfulness

The most consistent finding of this study is that the measured faithfulness of an answer depended strongly on

the quality of the retrieved evidence, not on the language model alone. In a preliminary run, the retrieved context for several questions was thin, partly because duplicate copies of the same article occupied retrieval slots, causing the models to fill the resulting gaps with parametric knowledge: GPT-4o introduced an HRE-saving and cost argument that was correct in general but absent from the supplied passages, and the local model added neodymium as a diffusion source that the passages did not support. After the index was de-duplicated and the genuinely informative passages were retrieved, these unsupported additions disappeared, and all three models became faithful on all three questions (Table 1 and Fig. 3).

This has a practical implication for anyone building a retrieval-augmented system over a scientific corpus: faithfulness failures should not be attributed to the language model by default. A passage-only instruction can only be honored if the passages contain the answer; when they do not, even a capable model tends to supply what it knows. In the evaluated cases, improving retrieval eliminated the unsupported additions without any change to the language models.

4.2. Commercial versus local models: attribution, not grounding, was the differentiator

With adequate context, the factual grounding of the two commercial models and the local open-weight model was broadly comparable. All three extracted the core HRE-related and diffusion-kinetics information supported by the passages, although their coverage of deposition methods and grain-boundary additives varied. The difference that remained was citation reliability. Both commercial models attached the correct source filename to every claim, whereas Llama 3.2 omitted citations entirely on Q1 and, on Q3, cited a passage whose retrieved text was only a publisher header rather than substantive content.

This distinction is particularly important in materials research. A researcher who acts on an extracted composition or process condition needs to trace it to the originating paper; an answer that is correct but uncited, or cited to the wrong fragment, is difficult to verify and audit. The commercial models' consistent attribution is thus a meaningful advantage for scientific use. At the same time, the local model's competitive grounding is encouraging for laboratories that cannot send queries or proprietary context to an external service: with a stricter prompt or a light post-processing check that every claim carries a valid source, a locally hosted model may be adequate for confidential literature analysis.

4.3. When the corpus does not contain the answer

Question 3 illustrated a second limitation that is independent of the model. The retrieved passages explained the kinetic reason that HRE diffuses preferentially along grain boundaries. However, they did not contain the magnetocrystalline-anisotropy mechanism by which an HRE-enriched shell raises the nucleation field and thus the coercivity, which a magnetism specialist would regard as the core explanation. A strictly passage-grounded system can only report what the corpus contains, so all three models gave the kinetic interpretation rather than the anisotropy mechanism. This is the correct behavior for a faithful system, but it means that the usefulness of the answer is bounded by the coverage of the indexed literature.

4.4. Implications for AI-assisted literature analysis in magnetism

Taken together, these results suggest a realistic role for retrieval-augmented LLMs in NdFeB and GBDP research. The system reliably consolidates dispersed factual content, including diffusion sources, deposition methods, process conditions, and stated advantages, into a single cited answer, which is useful for orientation and for assembling structured summaries from a large body of papers. It is less suited, as configured here, to questions whose answer is a mechanism not explicitly stated in the retrieved text. The principal levers for a trustworthy system are therefore a clean, de-duplicated and sufficiently broad corpus; a retriever tuned so that the informative passages are returned; and, for local models in particular, enforcement of the citation requirement. We do not claim that the system discovers new knowledge in magnetism. Rather, its main contribution is methodological, providing a controlled framework for determining when and how automated synthesis of the GBDP literature can be trusted. Two findings extend beyond a conventional literature review. First, the results indicate that measured faithfulness depends strongly on retrieval quality and that, under the fixed-retrieval conditions examined here, a clearer difference between the commercial and local models lies in citation reliability rather than factual grounding. These findings emerge from a controlled comparison of the models using the same retrieved evidence and therefore do not arise from the literature content alone. Second, the framework clearly identifies the boundary between what the corpus can and cannot support. For example, the absence of the anisotropy mechanism required to fully answer Q3 reveals a limitation in corpus

coverage and can therefore serve as a diagnostic signal for improving the knowledge base. Generating genuinely new scientific insight, such as proposing new compositions or identifying promising process windows, is a separate objective. Achieving this will require structured extraction of composition–process–property data, together with predictive modeling built on top of the retrieval framework, which we identify as a direction for future work.

4.5. Limitations

Several limitations should be noted. The evaluation used three questions and a qualitative, two-axis rubric (faithfulness and citation correctness); it did not score factual accuracy against ground truth or answer completeness, and it did not use multiple independent raters, so the assessment is indicative rather than statistical. The questions were posed once each at a fixed retrieval depth (eight chunks) and decoding temperature (zero), and the conclusions may not transfer to other question types, larger retrieval budgets or non-zero temperatures. The corpus was limited to 311 GBDP papers and was not exhaustive of the field. Finally, the commercial models were accessed through proprietary services whose internal versions may change over time, limiting long-term reproducibility. In contrast, the local model offers greater reproducibility because its weights and inference configuration can be fixed and archived. On a practical note, one model-specific configuration effect deserves mention: the Gemini Flash model is a reasoning model whose internal "thinking" tokens are counted against the output-token budget, so a budget chosen for non-reasoning models would truncate its visible answer (finish reason MAX_TOKENS) unless raised. We therefore allocated a larger output-token budget to this model; this is an operational consideration when comparing reasoning and non-reasoning models, rather than a limitation of the method itself. We emphasize that, given the use of only three questions and a single corpus, the present evaluation should be regarded as a proof of concept. The quantitative metrics reported above summarize a single-rater expert assessment. Standardized RAG evaluation metrics, such as automated faithfulness and answer-relevance scoring, together with multiple independent raters, represent important directions for future work.

5. Conclusions

We developed a CPU-only, domain-specific RAG system for the GBDP literature of NdFeB permanent magnets and used a controlled, fixed-retrieval protocol to compare two commercial language models (GPT-4o, Gemini 3.5 Flash) with a locally served open-weight model (Llama 3.2) on faithfulness and citation correctness. With an informative, de-duplicated context, all three models were faithful to the retrieved passages on every question, and the two commercial models additionally cited their sources correctly throughout; the local model matched the commercial models in factual grounding but applied the citation requirement inconsistently. Two conclusions follow. First, measured faithfulness depends strongly on retrieval quality, highlighting the importance of a clean, de-duplicated, and sufficiently broad corpus in addition to model choice. Second, the results suggest that a locally hosted open-weight model may provide a practical, privacy-preserving alternative to commercial APIs, provided that its source attribution is verified. The system is well suited to consolidating dispersed factual content but is bounded by corpus coverage for mechanism-level questions. Future work includes a larger question set with quantitative, multi-rater scoring, an expanded corpus covering coercivity mechanisms, and improved retrieval to support mechanism-level queries. Beyond expanding the question set, incorporating multiple raters, and broadening the corpus, a promising direction is to extend the framework from factual consolidation toward design insight by extracting structured composition–process–property data from the literature and using these data for constraint-based screening and property prediction. Such an extension could move the system beyond what a conventional literature review can provide.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2025-25464308) and by the Technology Innovation Program (RS-2025-02220764) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

References

[1] D. Brown, B.-M. Ma, and Z. Chen, J. Magn. Magn. Mater. **248**, 432 (2002).
[2] Q. Lu, Y. Shao, Y. Yin, H. Chen, H. Xu, W. Liu, M. Liu, C. Zhong, X. Yu, J. Chen, Y. Liu, X. Yi, and M. Yue, Sustain. Mater. Technol. **36**, e00615 (2023).
[3] C. Zhao, Q. Jiang, S. Ur Rehman, X. Li, Y. Chen, J. Song, and Z. Zhong, J. Rare Earth. **40**, 1894 (2022).
[4] S.-M. Lee, G. Kim, K.-S. Lee, S. Kim, T.-H. Kim, S.-H. Lee, D.-H. Kim, W. Lee, and J.-G. Lee, Acta Mater. **285**, 120660 (2025).

[5] T.-H. Kim, T.T. Sasaki, T. Ohkubo, Y. Takada, A. Kato, Y. Kaneko, and K. Hono, Acta Mater. **172**, 139 (2019).
[6] Y. Wang, H. Qin, N. Liu, Q.-N. Hu, H.-P. Cong, and S.-H. Yu, Adv. Mater. **37**, 2505193 (2025).
[7] J. Yan, Q. Wu, H. Ge, Z. Xi, and J. Liang, Mater. Chem. Phys. **315**, 128901 (2024).
[8] M. Ji, Z. Wang, W. Liu, Y. Li, H. Chen, H. Wu, R. Du, D. Zhang, M. Yue, X. Yi, Y. Liu, and S. Zha, J. Alloys Compd. **970**, 172598 (2024).
[9] D. Zhang, M. Zhu, X. Sun, Q. Sun, X. Song, and J. Zuo, Rare Met. **45**, e70157 (2026).
[10] W. Huang, J. Ran, Y. Liu, T. Li, W. Chen, L. Zhang, and S. Yin, Anal. Chem. **98**, 10913 (2026).
[11] N. A. Krotkov, D. A. Sbytov, A. A. Chakhoyan, P. I. Kornienko, A. A. Starikova, M. G. Stepanov, A. O. Piven, T. A. Aliev, T. Orlova, M. S. Rafayelyan, and E. V. Skorb, J. Chem. Inf. Model. **65**, 11064 (2025).
[12] H. Nakamura, K. Hirota, M. Shimao, T. Minowa, and M. Honshima, IEEE Trans. Magn. **41**, 3844 (2005).