Karshi Multidisciplinary International Scientific Journal

www.kmisj.reapress.com

K. Multidiscip. Int. Sci. J. Vol. 3, No. 2 (2026) 153–164.

Paper Type: Original Article

RAG Pipeline Accuracy Benchmarking in Healthcare Supply Chain: Medical Item Classification Tasks

Syed Mohammad Nasir[1], Aruna Pavate[2,*], Ansari Yusuf[2]

[1] Huron Eurasia India Private Limited, Bengaluru, Karnataka, India; mdnasir020396@gmail.com.

[2] Thakur College of Engineering and Technology, Mumbai, Maharashtra; arunaapavate@gmail.com.

Citation:

Received: 21 October 2025 *Revised: 14 December 2025* *Accepted: 09 February 2026*	Nasir, S. M., Pavate, A., & Yusuf, A. (2026). RAG pipeline accuracy benchmarking in healthcare Supply Chain: Medical item classification tasks. *Karshi Multidisciplinary International Scientific Journal, 3*(2), 153-164.

Abstract

The range of medical products within healthcare Supply Chains is extensive and must be precisely classified into hierarchical categories for purchasing, inventory control, expenditure analysis, and regulatory purposes. Traditional classification methods may not suit the short, mixed nature of product descriptions, or they may produce undue or hallucinated labels (e.g., Large Language Models (LLMs)). Retrieval-Augmented Generation (RAG) introduces a way to ground LLM outputs using an external source of domain knowledge. This work proposes a benchmark for RAG-based pipelines for the hierarchical classification of medical supplies. The performance evaluation of a three-level taxonomy with 214 leaf-level classification nodes was conducted on a benchmark dataset comprising 950 descriptions of medical supply items. We evaluated four pipelining settings by combining dense and BM25 retrieval with two instruction-tuned generation models. Retrieval quality was assessed with MAP@5, Recall@5, NDCG@5, Context Relevance Score (CRS), and Context Coverage (CCO), and the E2E quality was measured by hierarchical classification accuracy, macro-F1, and Answer Fidelity (AF). Dense retrieval outperformed all other methods on retrieval scores. The dense-retrieval models also yield higher hierarchical classification accuracy, with the best configuration achieving 87.3% for the subcategory. The MAP@5 is strongly correlated with subcategory accuracy, as indicated by partial correlation analysis. A human assessment of 240 misclassified queries revealed context dilution, label hallucination, retrieval collapse, and formatting inconsistency as the main failure patterns. The results provide a rigorous framework for assessing and refining RAG pipelines for domain-specific health care supply classification.

Keywords: Retrieval-augmented generation, Healthcare Supply Chain, Medical item classification, Dense retrieval, BM25, Large language models, Retrieval-augmented generation evaluation.

1 | Introduction

The healthcare supply chain includes a wide range of products, from ordinary consumables to complex, application-specific medical devices. When accurately classified within structured taxonomies, these products enable procurement routing, spending analytics, formulary management, inventory structuring, and regulatory reporting. Manual classification, however, is highly expertise-demanding, and scaling with product catalog size and complexity is not feasible. The enclosed research demonstrates that accurate taxonomy classification plays a significant role in healthcare supply chain operations.

Corresponding Author: arunaapavate@gmail.com

https://doi.org/10.22105/kmisj.vi.129

The success of Large Language Models (LLMs) has enabled promising ways to automate classification and other knowledge-intensive tasks. However, general-purpose LLMs may hallucinate information, labels, or classes when pertinent domain knowledge is not adequately embedded in their parameters [1]. Retrieval-Augmented Generation (RAG) overcomes this limitation by retrieving relevant data from an external knowledge base and providing it to the language model during generation [2].

RAG systems, however, depend heavily on retrieval quality. If retrieved passages are irrelevant or insufficient, they can cause the model to generate incorrect outputs, and even relevant passages may be useless if the model cannot make good use of the retrieved evidence [3], [4]. Recent research has thus focused on jointly evaluating retrieval and generation components, rather than evaluating the language model in isolation.

Pavate et al. [5] compared supervised and unsupervised RAG methods and analyzed how choice of retrieval mechanism, training strategy, and quality of retrieved documents impact factuality, relevance, and output consistency. The paper also notes the trade-offs among accuracy, flexibility, computational cost, and output consistency in RAG systems.

These issues are especially relevant in healthcare supply classification, where product descriptions are typically short and may include a mix of abbreviations, part numbers, manufacturer-specific terminology, and technical characteristics. This manuscript therefore focuses on benchmarking retrieval and generation performance for hierarchical medical supply classification. The main contributions of this work are as follows:

I. Development of a reusable RAG benchmarking framework for hierarchical medical supply classification.

II. Comparison of dense and BM25 retrieval using multiple retrieval metrics.

III. Evaluation of retrieval quality, Answer Fidelity (AF), and end-to-end classification accuracy.

IV. Identification of major RAG failure modes.

V. Analysis of practical remediation strategies for improving domain-specific RAG systems.

The remainder of this article is organized as follows. Section 2 presents the related work on RAG, information retrieval, and healthcare-oriented classification. Section 3 describes the dataset, healthcare supply taxonomy, knowledge base, retrieval methods, experimental settings, and evaluation metrics. Section 4 presents and discusses the experimental results, including retrieval performance, hierarchical classification accuracy, AF, and failure-mode analysis. Section 5 concludes the article and outlines directions for future research.

2 | Related Work

RAG combines external retrieval with language-model generation and has become an important architecture for knowledge-intensive NLP tasks. Lewis et al. [2] introduced RAG as a combination of parametric language-model memory and non-parametric external memory. Izacard et al. [6] subsequently demonstrated the value of combining multiple retrieved passages during generation.

An important limitation of RAG is that retrieved information may not always be useful. Shi et al. [3] showed that irrelevant contextual information can negatively affect LLM performance. This observation is especially relevant for medical supply classification, as semantically similar products may fall into different taxonomy classes.

Pavate et al. [5] additionally investigated how the choice of retrieval strategy and learning configuration influences RAG output variability. Their work differentiates between supervised methods, which optimize a specific task using labeled data, and unsupervised methods, which rely on unlabeled data and adaptation.

Information retrieval systems have historically been evaluated using MAP, Recall@K, Precision@K, and NDCG [7]. Recent RAG literature recommends evaluating retrieval and generation simultaneously. For

example, Gao et al. [4] summarized RAG architectures and evaluation techniques, and Es et al. [8] proposed RAGAS to evaluate RAG dimensions such as context relevance and generation quality.

Healthcare NLP has largely focused on clinical data, such as electronic health records, clinical notes, discharge summaries, and radiology reports. Medical supply classification is different, as item descriptions are typically shorter and include product characteristics and procurement language. This paper highlights this linguistic difference as a key driver for domain-specific RAG evaluation.

Recent literature has also explored the generalized use of AI in healthcare Supply Chains, including supplier selection, logistics, inventory-related decision-making, and other operational processes. However, existing studies have mostly focused on optimization and decision-support applications, with relatively few empirical evaluations of data-driven AI methods in real-world healthcare supply-chain settings [9].

2.1 | Research Gap

While RAG has been extensively studied for question answering, biomedical applications, and document analysis, we are not aware of any work that considers hierarchical medical supply classification via RAG while accounting for retrieval quality, fidelity, and end-to-end classification metrics simultaneously.

3 | Materials and Methods

3.1 | Dataset and Taxonomy

The benchmark dataset contains 950 medical supply item descriptions obtained from publicly available procurement and healthcare supply-chain sources. Each description contains between 4 and 94 words, with a median of 21 words. Two domain experts independently annotated the items, achieving Cohen's $\varkappa = 0.89$.

The taxonomy contains three hierarchical levels and 214 leaf-level nodes. Healthcare product classification can also align with standardized classification frameworks and mappings such as the United Nations Standard Products and Services Code (UNSPSC), which support consistent identification and categorization of healthcare products [10]. In this study, we used a task-specific three-level taxonomy to evaluate hierarchical RAG-based classification rather than directly adopting an external standardized taxonomy. The seven first-level service lines are:

I. Orthopedics.

II. Cardiovascular.

III. General Surgery.

IV. Neurology.

V. Oncology.

VI. Imaging and Diagnostics.

VII. Clinical Consumables.

Table 1 shows the dataset distribution.

Table 1. Benchmark dataset distribution across service lines.

Service Line	Item Count	% of Dataset	Avg. Description Length (Words)
Orthopedics	200	21.1%	24.3
General surgery	171	18.0%	19.7
Clinical consumables	162	17.1%	16.2
Cardiovascular	133	14.0%	22.8
Imaging and diagnostics	114	12.0%	18.5
Neurology	95	10.0%	21.1
Oncology	75	7.9%	23.6
Total	950	100%	21.0 (median)

3.2 | Knowledge Base and Retrieval

We constructed the knowledge base from taxonomy definitions, product-classification reference documents, and expert-generated item-to-taxonomy examples. Documents were divided into 256-token chunks with 32-token overlap, resulting in 4,817 indexed chunks.

Two retrieval mechanisms were evaluated. Dense semantic retrieval used transformer-based embeddings with Hierarchical Navigable Small World (HNSW) indexing. BM25 sparse retrieval used an inverted index with the Okapi BM25 ranking function $\mathbf{k_1 = 1.2}$ and $\mathbf{b = 0.75}$. We tokenized query and document texts using whitespace- and punctuation-based tokenization. Stop-word removal, stemming, and lemmatization were not applied so that domain-specific terminology, abbreviations, numerical expressions, and product codes were preserved during retrieval. This preprocessing was selected to retain lexical information that may be important for distinguishing closely related medical-supply items.

Dense semantic retrieval was performed using a transformer-based sentence-embedding model and an HNSW index. The embedding model represented both the supply-item queries and knowledge-base passages in a common vector space, after which nearest-neighbor search retrieved the top-five candidate passages. The embedding model name and version, embedding dimensionality, similarity metric, and HNSW configuration were not recorded in the experimental documentation available for this study; therefore, these implementation-specific parameters are not reported.

Both retrieval approaches used the same 4,817-document knowledge base and retrieved the top five passages for each query, ensuring consistent experimental conditions across the two retrieval methods.

3.3 | Experimental Configuration and Evaluation

Fig. 1 shows the overall pipeline used in this study.

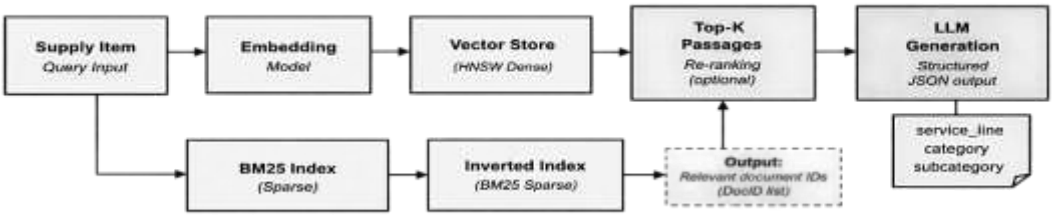

Fig.1. RAG pipeline architecture for hierarchical medical supply classification. Dense (semantic) and sparse (BM25) retrieval paths feed into an optional re-ranking stage before LLM-based structured output generation.

The proposed RAG pipeline with two-stage retrieval revolves around the following: 1) the dense semantic retrieval, and 2) the sparse BM25 retrieval. The supply item acts as the query input and is converted into a dense vector using an Embedding model. The generated embedding is then queried in the HNSW-based Vector Store to find semantically related documents. Simultaneously, the query is executed on a BM25 index built on an Inverted index, fetching documents that are linguistically or lexically closer to the query. Those retrieved document IDs are used to find candidate passages. The fetched candidates are Top-K selected, optionally with re-ranking for better relevance. The chosen passages are fed to the LLM that produces the medical supply classification in a structured JSON format including service line, category, and subcategory attributes.

Among other things, the evaluation framework contrasts semantic similarity-based retrieval with lexical matching to assess their impact on medical-supply classification. Four configurations were evaluated:

Table 2. Experimental pipeline configurations.

Configuration	Retrieval	Generation
C1	Dense	Model A
C2	Dense	Model B
C3	BM25	Model A
C4	BM25	Model B

We combined two retrieval methods and two generation settings, resulting in four experimental setups . The setups paired Dense and BM25 retrieval with two instruction-tuned LMs (Model A and Model B; see *Table 2*). We accessed both models via APIs with temperature-zero sampling. For each query, we supplied the top five retrieved passages to the generation model. The prompt instructed the model to use the retrieved passages as the permitted reference and to produce a structured JSON response containing the service-line, category, and subcategory fields. The experimental documentation available for this study did not record the specific identities, versions, provider details, maximum output-token settings, or complete prompt configurations of Model A and Model B, so we cannot specify them retrospectively. We then measured AF and hierarchical classification accuracy. The hierarchical accuracy requires correct prediction of ancestor categories before a prediction is considered correct at a lower level.

We used Wilcoxon signed-rank tests with Bonferroni correction for comparisons, and partial correlation to examine the relationship between retrieval quality and subcategory accuracy.

Context Relevance Score (CRS): Measures the semantic relevance of the retrieved passages to the input medical-supply description. For each query, the semantic similarity between the query representation and each of the top-five retrieved passages was calculated using their embedding representations. The CRS was computed as the mean semantic similarity across the retrieved passages, producing a normalized score between 0 and 1, where higher values indicate that the retrieved context is more closely aligned with the query.

Context Coverage (CCO): Measures the extent to which the retrieved top-five passages contain the information required to support the target hierarchical classification. Coverage was assessed against the information elements required to determine the correct service line, category, and subcategory. The CCO reflects what fraction of a required set of classification-relevant information elements is included in a retrieved context, and values close to 1 indicate that a retrieved context provides full contextual support. We used MAP@5, Recall@5, NDCG@5, CRS, and CCO to evaluate performance. We then assessed answer fidelity and hierarchical classification accuracy for the generated outputs. For a query q and its retrieved passages $d_1, \ldots, d_K$, where $K = 5$, the CRS is calculated as

$$\mathrm{CRS}(q) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Sim}(q, d_i), \tag{1}$$

where $\mathrm{Sim}(q, d_i)$ denotes the normalized semantic similarity between the query and the i-th retrieved passage.

CCO is calculated as

$$\mathrm{CCO}(q) = \frac{\sum_{j=1}^{M} I_j\,(q, C)}{M}, \tag{2}$$

where C represents the retrieved context, M is the number of classification-relevant information elements required for the query, and $I_j(q, C)$ is an indicator that takes the value 1 when the j-th required information element is supported by the retrieved context and 0 otherwise. The final CCO score is the average coverage over all considered queries.

4 | Results and Discussion

4.1 | Retrieval Performance

The results clearly illustrate the contrast between dense and BM25 retrieval.

Table 3. Retrieval Performance

Metric	Dense	BM25	Difference	p-value
MAP@5	0.831	0.694	0.137	<0.001
Recall@5	0.879	0.731	0.148	<0.001
NDCG@5	0.847	0.712	0.135	<0.001
Context Relevance	0.760	0.610	0.150	<0.001
CCO	0.840	0.670	0.170	<0.001

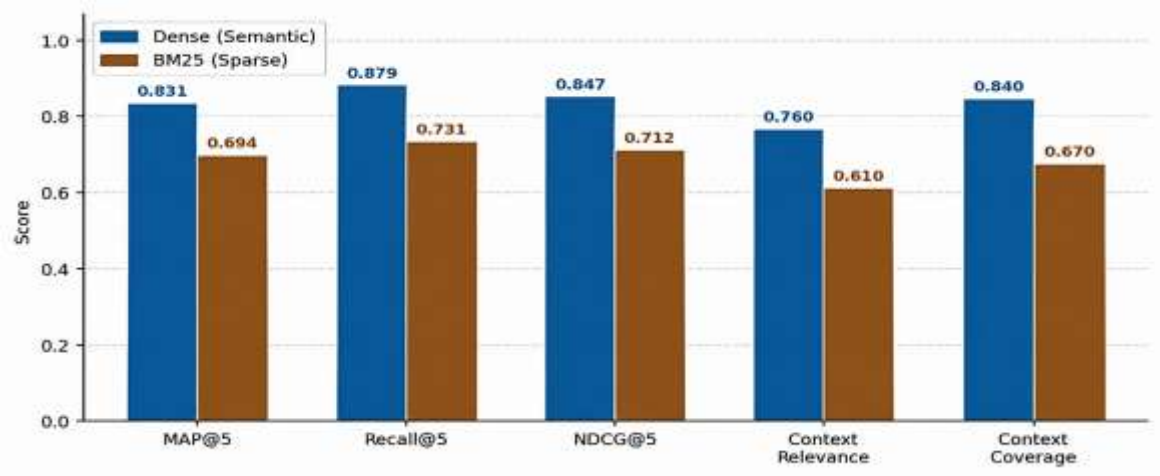

Fig. 2. Comparison of dense and BM25 retrieval performance across MAP@5, Recall@5, NDCG@5, Context Relevance, and CCO.

Fig. 2 compares Dense (semantic) retrieval and BM25 (sparse) retrieval using MAP@5, Recall@5, NDCG@5, Context Relevance, and CCO. Dense retrieval achieves higher scores across all evaluated metrics, with MAP@5 of 0.831, Recall@5 of 0.879, and NDCG@5 of 0.847. It also achieves 0.760 for Context Relevance and 0.840 for CCO. In comparison, BM25 obtains 0.694, 0.731, and 0.712 for MAP@5, Recall@5, and NDCG@5, respectively, while its Context Relevance and CCO scores are 0.610 and 0.670.

The results demonstrate that dense semantic retrieval provides more relevant and comprehensive retrieval results for the medical-supply classification task. The higher Recall@5 and CCO further indicate that semantic retrieval can identify and provide a broader set of useful contextual information to the downstream LLM. These findings support using dense retrieval as an important component of the proposed RAG pipeline.

The attached analysis further reports that BM25 performance decreases for descriptions containing abbreviations, part numbers, and manufacturer-specific terminology, whereas dense retrieval is comparatively stable across vocabulary styles.

4.2 | Classification Performance

Table 4. Hierarchical classification accuracy.

Pipeline	Service-line accuracy	Category accuracy	Subcategory accuracy	Macro-F1
Dense + Model A	96.8%	92.4%	87.3%	0.893
Dense + Model B	95.6%	90.7%	85.1%	0.871
BM25 + Model A	88.4%	80.1%	74.6%	0.741
BM25 + Model B	87.2%	78.6%	72.9%	0.724

The attached manuscript reports that the difference between dense and BM25 configurations is about 12–15 percentage points at the subcategory level, while the difference between the two generation models is about 2 percentage points.

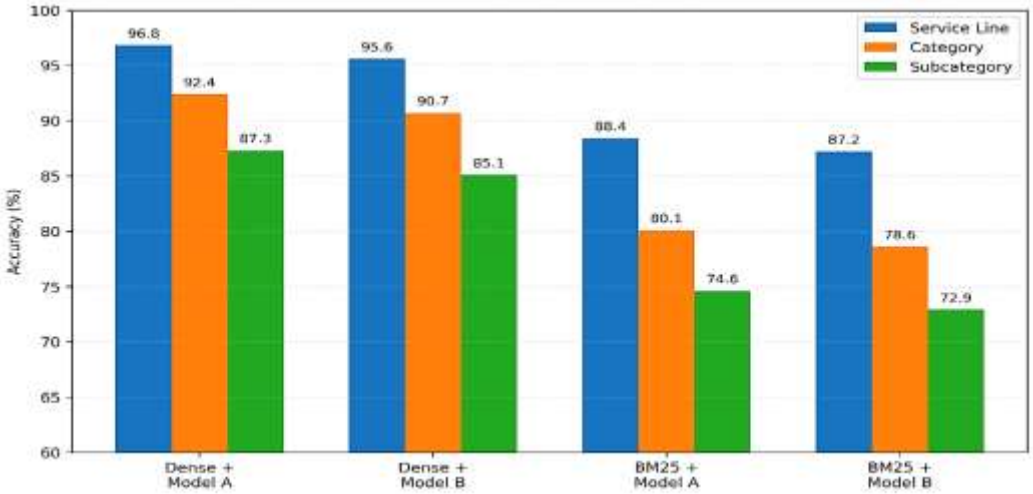

Fig. 3. Hierarchical classification accuracy by pipeline configuration and taxonomy level.

Fig. 3 presents the hierarchical classification accuracy obtained using different retrieval and language-model configurations. The results are reported at three classification levels: Service Line, Category, and Subcategory. The Dense + Model A configuration achieves accuracies of 96.8%, 92.4%, and 87.3%, respectively. The Dense + Model B configuration achieves 95.6%, 90.7%, and 85.1%, respectively. In comparison, the BM25 + Model A configuration achieves 88.4%, 80.1%, and 74.6%, while BM25 + Model B achieves 87.2%, 78.6%, and 72.9%, respectively.

4.3 | Answer Fidelity and Error Analysis

AF was 0.91 for dense retrieval and 0.78 for BM25. Outputs with fidelity below 0.60 occurred in 4.2% of BM25 queries and 1.1% of dense queries. Among these low-fidelity outputs, 83% corresponded to incorrect subcategory predictions.

The partial correlation analysis reported:

$$r = 0.71,\ p < 0.001, \tag{3}$$

between MAP@5 and subcategory accuracy, after controlling for service line and item-description length. The manuscript reports approximately 68% of the variance associated with retrieval quality, with additional explained variance attributed to AF and generation-model choice.

This result complements Pavate et al. [5], who discussed retrieval quality and retrieval strategy as important factors influencing RAG output variability and task performance.

4.4 | Retrieval-Augmented Generation Failure Modes

A manual analysis of 240 misclassified queries identified four major failure modes:

Table 5. Distribution of RAG failure modes based on 240 manually reviewed misclassifications.

Failure Mode	Percentage	Approx. Count*
Context dilution	34%	82
Label hallucination	28%	67
Retrieval collapse	22%	53
Format inconsistency	16%	38
Total	100%	240

*Counts are rounded from the reported percentages and therefore should not be presented as exact observed counts unless you have the original case-level data.

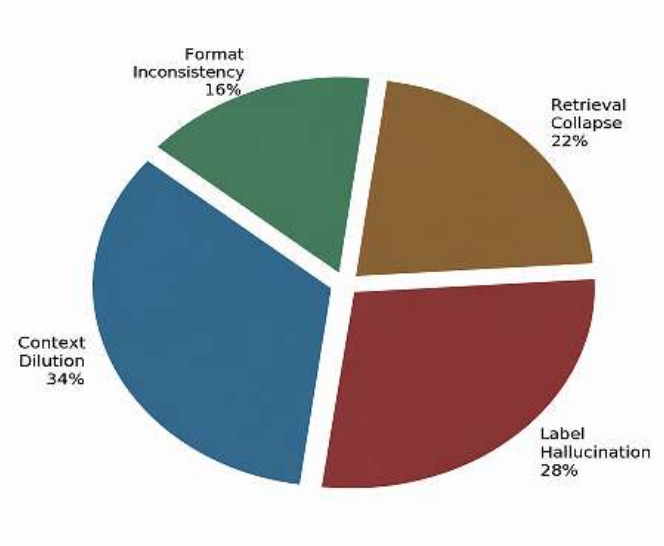

Fig. 4. Distribution of RAG failure modes based on 240 manually reviewed misclassifications.

A manual analysis of 240 misclassified queries identified four major RAG failure modes: Context dilution (34%), label hallucination (28%), retrieval collapse (22%), and format inconsistency (16%). Context dilution was used to refer to the situation where irrelevant or too much information was retrieved, and then the available context for classification was blurred. Label hallucination is when the generated output contains taxonomy labels that are unsupported or incorrect. Retrieval failure resulted from poor retrieval of relevant documents for the target class, especially for difficult/underrepresented classes. Format violation occurred when outputs did not follow the expected structured JSON format. Fig. 4 depicts these malfunction modes. Context dilution was the most commonly observed failure mode. The study reports that cross-encoder re-ranking reduced this failure mode by 41% in a post-hoc experiment. Label hallucination was more prevalent for short item descriptions, while retrieval collapse emerged for infrequent subcategories and queries with many abbreviations. We applied structured JSON validation combined with automatic retries to mitigate format inconsistency.

5 | Discussion

5.1 | Retrieval Quality and Classification Performance

The results show that the quality of retrieval has a very strong influence on the final classification accuracy. Dense retrieval yielded better MAP@5, Recall@5, NDCG@5, context relevance, and CCO, and the hierarchical classification accuracy in the dense setting was also the best.

This observation aligns theoretically with previous work showing that LLMs can be distracted by irrelevant context [3]. It also aligns with RAG evaluation methodologies that promote a joint evaluation of retrieval quality, contextual relevance, and generation quality [4], [8].

5.2 | Knowledge-Base Design

The results indicate that the design of the knowledge base in RAG systems needs to be considered in the RAG system development process. Taxonomy-definition documents need to be explicit about category boundaries (what is and is not in a category), what features differentiate it, and what a stereotypical category member looks like. The chunk size should match the length and information density of a typical query.

Furthermore, it underlines that coverage, even for rare categories in a knowledge base, must be sufficiently maintained, as low representation can cause retrieval collapse.

5.3 | Practical Implications

The findings indicate several implementation considerations for healthcare supply classification systems:

I. evaluate retrieval independently before evaluating generation.

II. explicitly define taxonomy entries in the knowledge base.

III. track retrieval accuracy for rare categories.

IV. potentially incorporate re-ranking for ambiguous retrieved contexts.

V. restrict generated labels to valid taxonomy nodes.

VI. automatically validate structured outputs.

VII. evaluate both retrieval and end-to-end classification performance.

These recommendations are directly informed by the failure-mode analysis and benchmarking results presented in the work.

Automated product classification has historically relied on rule-based methods using predefined terminology, attributes, and classification rules. Machine Learning (ML) techniques offer greater flexibility to address variability in product descriptions, but their effectiveness depends on the availability and quality of relevant training data [11]. These factors are especially relevant in medical supply classification, where item descriptions may involve heterogeneous terminology, technical specifications, abbreviations, and product-specific identifiers.

5.4 | Limitations

This study has potential limitations. First, the gold-standard dataset was constructed by scraping online resources and may not reflect the terms and naming conventions used in hospital purchasing systems. Second, the taxonomy is not standardized, unlike UNPSC and GS1; you may see different performance with other taxonomies. Third, the released experiment documentation does not record the identity and exact configuration of the models (both generation models and the embedding model). Therefore, precise reproduction of the generation and dense retrieval parts may be limited. In future work, it is suggested to report the exact model names and versions, embedding dimensionality, similarity metric, HNSW parameters, API configuration, prompt templates, and decoding settings. Lastly, we did not assess latency or computational expense, even though these may be significant considerations in high-volume production environments.

6 | Conclusion

This study introduces a benchmarking framework for evaluating RAG pipelines for hierarchical medical supply classification. We compared dense and BM25 retrieval strategies using retrieval, fidelity, and end-to-end classification measures on a dataset of 950 medical supply descriptions and a taxonomy of 214 leaf-level nodes.

Dense retrieval outperformed BM25 on all five retrieval metrics. The Dense + Model A configuration achieved 96.8% service-line accuracy, 92.4% category accuracy, 87.3% subcategory accuracy, and a macro-F1 of 0.893. Dense retrieval also had higher AF.

We analyzed 240 misclassified queries and found four main failure types: context dilution, label hallucination, retrieval collapse, and format inconsistency. These results suggest that RAG performance depends on the interplay among knowledge-base coverage, retrieval quality, context selection, and structured generation.

These results also align with recent RAG literature that highlights retrieval quality, contextual grounding, and output variability as key considerations in RAG systems [3–5]. In particular, the review by Pavate et al. [5]

serves as a good reference for discussing retrieval strategy, supervised/unsupervised learning, and output variability in RAG systems.

Future work can explore hybrid dense-sparse retrieval, multi-hop retrieval, cross-encoder re-ranking, automated knowledge-base monitoring, and computational cost/latency evaluation. The attached manuscript also mentions these as future research directions.

Acknowledgments

The authors acknowledge the subject-matter experts who contributed to benchmark annotation and taxonomy design review.

The uploaded manuscript states that no external funding supported the research.

Author Contribution

Conceptualization, M. N. and A. A. P.; methodology, M. N. and A. A. P.; software, M. N.; validation, M. N. and Y. A.; formal analysis, M. N. and A. A. P.; investigation, M. N. and Y. A.; resources, A. A. P. and Y. A.; data curation, M. N.; writing(original draft preparation, M. N.; writing) review and editing, A. A. P. and Y. A.; visualization, M. N.; project administration, A. A. P. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability

The benchmark query dataset and evaluation scripts were derived from publicly available procurement data sources. The curated evaluation dataset, knowledge-base construction scripts, and evaluation code will be made available upon reasonable request to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest. Funders played no role in the study design, data collection, analysis, or interpretation; manuscript writing; or the decision to publish the results.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). NeurIPS. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[3] Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. *International conference on machine learning* (pp. 31210-31227). PMLR. https://proceedings.mlr.press/v202/shi23a.html

[4] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey*. https://arxiv.org/abs/2312.10997

[5] Pavate, A., Mourya, V., Digra, M., Kumavat, K., & Jalalov, M. (2025). Unsupervised vs. supervised RAG: A comparative study of retrieval-augmented language models and their output variability. *International conference on intelligent optimization and big data management* (pp. 418-431). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10400-7_37

[6] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of machine learning research*, *24*(251), 1-43. https://www.jmlr.org/papers/v24/23-0037.html

[7] Manning, C. D., Raghavan, P., & Schütze, H. (2008). XML retrieval. *Introduction to information retrieval*, *1403*, 1404. https://www.internetdownloadmanager.com/register/new_faq/chrome_extension2.html

[8] Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 150-158). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2024.eacl-demo.16

[9] Espinosa, O., Basto, S., Ordóñez, A., Arias, M. L., Sepúlveda, D., & Bhutta, M. F. (2026). A review of artificial intelligence in healthcare Supply Chains: Untapped potential?. *BMC health services research*, *26*(742). https://doi.org/10.1186/s12913-026-14559-2

[10] US., G. H. (2022). *Implementation guideline: Applying the GS1 system of standards for U.S. FDA unique device identification (UDI).* https://documents.gs1us.org/.../Implementation-Guideline-Using-the-GS1-System-for-US-FDA-UDI-R

[11] Roberson, A. (2021). Applying machine learning for automatic product categorization. *Journal of official statistics*, *37*(2), 395–410. https://doi.org/10.2478/jos-2021-0017

Appendix A

Benchmark evaluation protocol summary

The appendix can retain the metric-definition table from your existing manuscript. The KMISJ template permits supplementary methodological details, mathematical material, additional figures, and experimental information in an appendix and requires appendix figures/tables to use the A-prefix, such as *Table A1*.

Table A1: evaluation metric definitions and computation methods.

Metric	Stage	Definition	Computation
MAP@5	Retrieval	Mean average precision of the top-five retrieved passages	Mean of the average precision values computed over the top-five ranked results
Recall@5	Retrieval	Proportion of relevant passages retrieved among the top five results	Number of relevant passages retrieved in the top 5 divided by the total number of relevant passages
NDCG@5	Retrieval	Rank-discounted relevance gain of the top-five retrieved passages	Discounted cumulative gain normalized by the ideal discounted cumulative gain
CRS	Retrieval	Semantic relevance of retrieved passages to the input query	$\mathrm{CRS}(q) = \frac{1}{5}\sum_{i=1}^{5} \mathrm{Sim}(q, d_i)$
CCO	Retrieval	Extent to which the retrieved context contains information required for the target classification	$\mathrm{CCO}(q) = \frac{\sum_{j=1}^{M} I_j(q, C)}{M}$
AF	Generation	Proportion of generated output information traceable to the retrieved context	Supported output information divided by total evaluated output information
Hierarchical accuracy	End-to-end	Correct classification at service-line, category, and subcategory levels	Prediction is counted as correct at a lower level only when the required ancestor levels are also correctly predicted

All retrieval metrics were calculated using the same top-five retrieved passages for each query to maintain consistency across Dense and BM25 retrieval. CRS evaluates the semantic alignment of the retrieved context with the query. In contrast, CCO evaluates whether the retrieved context contains the classification-relevant information necessary to support the target taxonomy decision. Thus, CRS captures relevance, while CCO captures completeness of supporting information.