Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1873–1882
https://doi.org/10.54941/ahfe1006998

Shopfloor Terminology for RAG: Aligning Operator Language With Engineering Knowledge

Ludwig Streloke[1], Yannick Rank[2], Freimut Bodendorf[2], Jörg Franke[1], and Patrick Bründl[1]

[1]Friedrich-Alexander-Universitat of Erlangen-Nürnberg, Institute for Factory Automation and Production Systems (FAPS), Nuremberg, Germany
[2]Friedrich-Alexander-University of Erlangen-Nürnberg, Nuremberg, Germany

ABSTRACT

As industrial work becomes increasingly digitalized, integrating human expertise into intelligent systems is essential for reliability and adaptability. This study investigates how curated terminology can improve Large Language Model-based Retrieval-Augmented Generation (RAG) systems for industrial knowledge management. It addresses a key linguistic issue that operators often use colloquial or locally coined terms that differ from standardized terminology found in technical documentation. This can lead to retrieval failures and inconsistent responses. A domain-specific dataset comprising 35 operator questions derived from a wire harness manufacturing manual is used to compare two types of RAG queries: natural-language operator queries and terminology-enhanced queries expanded with curated synonyms. The correctness of the generated answers was assessed by human evaluators. The queries with terminology enhancement achieved on average 67% correct answers in comparison to only 11% for natural-language questions. These results demonstrate the importance of terminology alignment for the reliable and effective use of LLMs in industrial contexts. Curated terminology bridges the gap between operator language and formal documentation, supporting tacit knowledge externalization and improving retrieval reliability. This preliminary study highlights the feasibility and practical relevance of integrating terminology into RAG pipelines and outlines future directions towards adaptive, human-centered knowledge systems in manufacturing.

Keywords: Retrieval-augmented generation (RAG), Large language models (LLMS), Terminology, Shopfloor language, Query expansion, Domain adaptation

INTRODUCTION

The nature of industrial work is undergoing a profound transformation through the increasing digitalization of factories, where operator interact with intelligent systems. Large Language Models (LLMs) represent a key enabling technology to make complex technical knowledge accessible via natural language dialogue (Rank et al., 2025). LLM-based chat assistants are used as intuitive interfaces between humans and technical infrastructures (Zhang et al., 2023), assisting operators through troubleshooting, process adjustments, and knowledge transfer.

Since industrial environments differ from the open-domain settings in which most language technologies are developed. Shopfloor communication is characterized by heterogeneity: operators rely on technical jargon, abbreviations, locally coined synonyms and colloquial expressions which are often transmitted informally and are rooted in tacit, embodied practice (Jose et al., 2024; Freire et al., 2024). While understanding this linguistic richness unlocks valuable experiential knowledge, it poses severe challenges for LLMs (Naqvi et al., 2024). Without an understanding of the nuances of shopfloor terminology, assistants risk misinterpreting queries, overlooking relevant information or providing incorrect responses, which can undermine trust and usability in high-stakes industrial contexts.

Retrieval-Augmented Generation (RAG) is a promising approach that extends LLMs with domain-specific knowledge sources. However, RAG systems typically rely on standardized or curated sources such as technical documentation, which use highly domain-specific language. Since most embedding models are trained on general-purpose corpora, they struggle to represent specialized terminology accurately (Tang and Yang, 2024). This limitation is reflected in benchmarks such as BEIR, where dense embeddings often show reduced performance on domain-specific datasets (Thakur et al., 2021). Approaches to address terminology mismatches via ontologies or knowledge graphs result in significant development and maintenance costs, making them unfeasible for many small and medium-sized enterprises (Barron et al., 2024). Methods are needed that can cope with linguistic variance in a more flexible and scalable way, ideally by leveraging the knowledge that operators already possess.

Against this backdrop, curated terminology emerges as a potential enabler of human-centered AI in industry. By systematically collecting, expanding and integrating the vocabulary used on the shop floor, terminology can bridge the gap between operator language and formal documentation. To examine this, our study sets up a controlled RAG pipeline on a German wire-harness machine manual, compares natural operator queries with terminology-enhanced queries expanded by curated synonyms, and uses expert evaluation to assess answer quality. This approach aims to improve the reliability of LLM-based assistants and preserve tacit employee knowledge, contributing to a more inclusive and trustworthy integration of AI into industrial practice.

RELATED WORK

RAG has emerged as a central research direction in natural language processing in recent years (Guu et al., 2020; Lewis et al., 2020). This development is closely linked to advances in dense retrieval, which enables the retrieval of semantically related documents through vector representations. Although RAG can use sparse or hybrid retrieval, it was primarily developed around dense retrievers (Karpukhin et al., 2020). In recent years a growing body of research is contributing to the developments in RAG. Wu et al. (2024) provide a detailed analysis of underlying retrieval technologies and their application domains, whereas Zhao et al. (2024) extends RAG into multimodal contexts. Other contributions highlight the potential of

structured data, particularly knowledge graphs, to enable more precise and context-aware responses (Peng et al., 2024). Cheng et al. (2025) adopt a knowledge-centric perspective, emphasizing that an accurate understanding of user intent is essential for RAG models to produce responses that are both semantically relevant and contextually appropriate.

Building on this, evidence from domain-specific QA in manufacturing shows that terminological structure is an important factor on RAG effectiveness. Bei et al. (2024) introduce Integrated Term Enhancement Methodology (ITEM), which extracts, normalizes, and explains key domain terms into a curated term dictionary and then conditions retrieval and generation on this term-level context. ITEM improves accuracy over strong dense-retrieval RAG while using fewer tokens. These gains are achieved by making terms explicit, disambiguated, and consistently mapped. This is reducing retrieval noise, aligning queries with corpus language, and stabilizing the interface between retrieval and generation.

Freire et al. (2024) and Jose et al. (2024) highlight that unstructured industrial texts, like shift or issue reports, pose major challenges for RAG systems. In contrast to curated manuals, these texts contain inconsistent jargon, abbreviations, and local language practices, which hinder reliable information extraction and increase hallucination risk. Both studies emphasize that the lack of unified terminology limits the effective use of experience-based knowledge and contributes to a trust gap, as operators still prefer human experts when available.

It should be noted that terminology is closely connected to tacit knowledge. Nonaka and Takeuchi (1995) demonstrate this with their example developing a bread-making machine. Through dialogue between bakers and engineers, the term 'twisting stretch' was coined to describe a baker's intuitive kneading motion. This expression did not exist previously; it emerged from the interaction itself, transforming a practical intuition into shared, communicable knowledge. This example shows that terminology carries essential knowledge about a domain – without it, the domain cannot be fully grasped.

Even within the terminology community, there is a growing recognition that RAG systems struggle with terminological precision. As Di Nunzio (2025) highlights, the lack of control over terminology, multilingual consistency, and definitional accuracy has sparked increasing interest in integrating curated term resources into generative AI workflows. From this discussion emerged the concept of Terminology-Augmented Generation (TAG).

Lee et al. (2025) introduce term-level Retrieval-Augmented Generation (tRAG) as a way to overcome the so-called seen term bias in pseudo-query generation. Their analysis shows that LLMs and document-level RAG approaches tend to over-rely on terms already present in input documents, failing to capture "unseen" but domain-relevant terms that frequently occur in real queries. To address this limitation, tRAG generates domain-specific keywords across the entire corpus, verifies them collectively, and integrates them into refined queries. Experiments on the BEIR benchmark (Thakur et al., 2021) demonstrate that this approach significantly improves

recall of unseen terms and yields consistent performance gains over both standard pseudo-query generation and document-level RAG. This work underscores the importance of optimizing RAG at the term level, since query interpretation depends critically on correctly capturing domain-specific vocabulary.

METHODOLOGY

The aim of this study is to examine whether curated shopfloor terminology can improve the performance of RAG in industrial contexts. To this end, we developed a controlled experimental setup combining the construction of a terminology-sensitive query dataset with a reproducible RAG pipeline applied to a real German-language machine manual. This design allows to isolate the effect of terminology alignment.

The first step was creating an operator question set for a wire harness manufacturing machine. The machine type was selected because it represents a technically complex production process with rich domain terminology and extensive documentation. Through interviews with experienced German operators, we elicit typical shopfloor expressions and slang terms that differ from the standardized terminology used in the manual. Based on these linguistic divergences, we derived a set of paired queries. One set is formulated using the correct technical terminology from the manual, and the other is using colloquial or synonymous variant from operator practice. Each question was embedded in a realistic operational context rather than isolated as a simple definition, ensuring that retrieval performance reflected authentic operator language and task situations.

The machine's operating manual served as the knowledge base for the RAG system. The PDF manual was processed using Docling (Docling Team, 2024). This software applies OCR-based layout analysis to preserve structural information such as headings, tables, and figures when converting into Markdown format. The Markdown document was then segmented into semantically coherent sections using a two-stage splitting strategy implemented in *LangChain* (Chase, 2022). First, the *MarkdownTextSplitter* was applied to preserve structural boundaries; if the resulting chunks exceeded the target length of 950 tokens (200 tokens overlap), *RecursiveCharacterTextSplitter* further refined them. Because technical manuals typically cluster related information in compact sections, the number of retrieved chunks per query was limited to five.

Dense retrieval was deliberately chosen as the sole retrieval method, since bag-of-words approaches such as BM25 rely on lexical overlap and would systematically disadvantage synonym-based queries. Four dense embedding models are compared, based on the assumption that variations in semantic distances representation between models would affect retrieval performance. The embedding models are selected based on their performance on the MTEB benchmark (Muennighoff et al., 2023), which comprehensively evaluates retrieval and semantic similarity tasks: e5-large-v2, multilingual-e5-large, embeddinggemma-300m, and Qwen3-Embedding-8B. Notably, the E5 family was the first dense retriever outperforming BM25 on the BEIR benchmark (Thakur et al., 2021; Wang et al., 2024), covering diverse domains and task types and thus providing an indicator of general retrieval

capability. This aspect is particularly relevant in the present study, where effective domain-specific retrieval is central.

The generative component used *llama3.1:8b-instruct-q8_0* (Llama Team and AI @ Meta, 2024) executed in the *Ollama* (Ollama Inc., 2025) runtime, with a temperature of 0.1 and a fixed random seed to ensure reproducibility and minimize stochastic variation. Each query pair was executed under identical pipeline conditions. The language model was explicitly instructed to formulate answers strictly based on the retrieved context; when relevant information was absent, the system was required to respond with *"I don't know"*. This constraint prevented hallucinated answers and provided a clear criterion for retrieval success.

The system outputs were then evaluated by human annotators. For each response, annotators assessed factual correctness with respect to the machine manual. Since all questions were paired with and without correct terminology, the evaluation followed a pairwise binary design that allowed direct measurement of terminology effects on system output.

In this study, responses were evaluated solely by a domain expert for factual correctness. We adopted this qualitative approach because the dataset is small, the task is domain-specific, and commonly used RAG metrics (e.g., faithfulness/context recall) are not yet well tuned to this setting. Moreover, many quantitative options depend on embedding-based similarity, which risks circularity when the very factor under study is the choice of embeddings. As alternatives, future work will add multi-rater human annotation with inter-rater reliability and complement it with calibrated quantitative metrics (e.g., RAGAS (Es et al., 2024)) or non-embedding baselines (e.g., exact matches to gold facts or citation-anchored checks) once larger data are available.

EXPERIMENTS & RESULTS

In total, the evaluation comprised 35 paired operator-style questions, each formulated once with the correct technical terminology and once with colloquial or synonymous variants. Examples of this terminology and the corresponding synonyms are given in Table 1.

Table 1: Examples of terms in the engineering and shopfloor terminology.

Engineering Terminology		**Shopfloor Terminology**	
Kabelfass (DE)	Cable drum (EN)	Fass, Rolle	Drum, roll
Kabelwechsler	Cable changer	Magazin	Magazine
Klemmeinheiten	Clamping units	Kabelklemmer	Cable clamps
Führungsrohre	Guide tubes	Röhrchen	Small tubes
Richteinheit	Straightening unit	Walzstrecke	Rolling section

Out of the 35 questions, 28 were answered correctly when evaluated under the best-performing embedding model e5-large-v2. The most striking result concerns the role of terminology alignment. When questions were asked with correct technical terms, the system produced on average 67% correct answers in comparison to only 11% for non-terminology-enhanced ones (Figure 1).

This effect provides clear evidence that the use of curated terminology is a decisive factor for RAG performance in industrial contexts. However, our dataset contains a deliberately high density of shopfloor terminology; in more natural settings with sparser or noisier phrasing, we expect the performance gap to be smaller and the absolute accuracies to be lower.

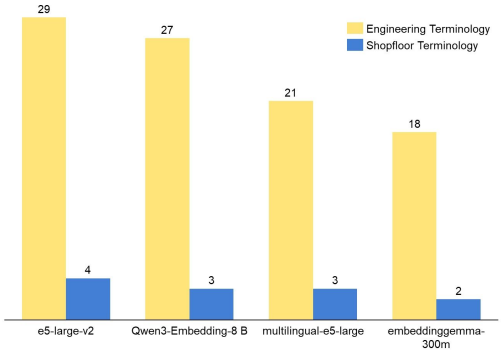

Figure 1: Number of correct answers generated with different embedding-models.

The choice of the model has a substantial influence on the performance as evidenced bythe coefficients of variation of 19% (Engineering Terminology) and 27% (Shopfloor Terminology). It is worth noting that some errors were only indirectly attributable to the embeddings. By retrieving a heterogeneous set of chunks, they reduced the answer stability and occasionally misled the LLM. Finally, although using only German test data may introduce a language-specific bias, a strong language advantage appears unlikely given that multilingual-e5-large performed comparatively poorly, suggesting that other factors predominantly drive the observed differences.

A closer inspection of the results further reveals that there is considerable variance across the embedding models with respect to the correct answering of individual questions. Only 29% of the queries were answered consistent by all four embedding models. In 37% of the cases, at least one model deviated from the others, and in 31% of the cases two models produced divergent results (Figure 2).

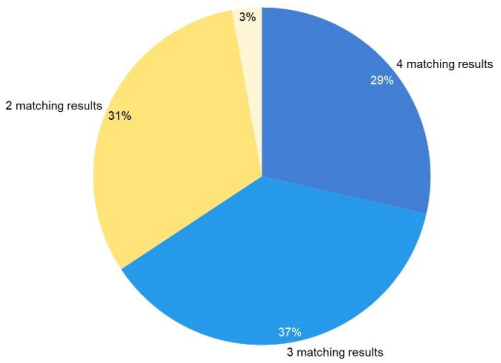

Figure 2: Percentage of matching answers generated with different embedding-models.

The results suggest that synonym-based or colloquial queries disproportionately expose weaknesses in embedding models, which differ substantially in terms of their underlying training data, representational objectives and handling of multilingual and domain-specific language. These differences reveal considerable potential for optimization: by adapting embedding models to the linguistic characteristics of shop floor communication through targeted fine-tuning, using domain-specific corpora or multi-model ensembles, retrieval systems could achieve greater robustness and semantic alignment in real-world language conditions.

CONCLUSION & OUTLOOK

This study provides an initial indication that terminology plays a decisive role in shaping the performance of retrieval-augmented generation in industrial contexts. Our results show that queries using the correct technical terminology were answered more reliably than those relying on synonyms or colloquial variants. In fact, performance improved by a factor of six when the correct terms were used. These findings strongly support and build upon recent research into retrieval-augmented generation. Lee et al. (2025) emphasized the importance of addressing 'unseen terms' to improve domain adaptation, and Tang and Yang (2024) demonstrated that even benchmarks containing some domain-specific data, such as BEIR, fail to capture the full linguistic and contextual variability present in specialized domains. Our results demonstrate that unseen or misaligned shopfloor terms limit retrieval effectiveness. Together, these findings suggest that achieving robustness in highly specialized environments requires more than just benchmark-driven generalization.

At the same time, our experiment revealed clear differences in how embedding models respond to terminological divergence. This results in substantial inconsistencies for the same queries across models. Only a third of all test questions were answered consistently by all four embedding models. Interestingly, in several cases, individual models retrieved for certain queries the correct document segments even when incorrect terminology was used. This suggests that embedding model training still has significant potential for improvement. Consequently, improving the robustness of embeddings to linguistic variation through better training data and domain adaptation is a crucial step towards achieving more reliable RAG systems in industrial contexts. However, as such terms are often unique to specific organizations or even individual teams, training targeted embedding models would be neither scalable nor sustainable. A more promising approach in that case would be to integrate curated terminological resources directly into RAG pipelines, thereby ensuring consistent semantic alignment while preserving adaptability across contexts.

Looking forward, Large Language Models themselves offer a promising avenue for addressing this challenge. Rather than relying solely on static terminological resources such as ontologies or manually curated dictionaries LLMs could be leveraged for semi-automated terminology management in an agentic framework. For example, they could be tasked with actively eliciting

synonyms, contextual features, and operator-specific definitions through targeted prompting, thereby building a broader base of terminological mappings. Such an approach would not only support retrieval alignment but also serve as a mechanism for eliciting tacit knowledge from employees, thus contributing to the formalization of knowledge that is often undocumented but operationally critical.

LLM-based terminology management can be a gateway to capturing tacit knowledge on a large scale as controlled terms form the basis of taxonomies, ontologies and knowledge graphs. Once curated terminology has been integrated into advanced RAG, it can map operator synonyms to standardized terms and their associated knowledge chunks. This creates stable retrieval pathways despite linguistic variation. With continuous, LLM-driven updates, the terminology layer evolves alongside practice and usage, thereby seeding taxonomies, structuring ontologies and instantiating knowledge graphs. Consequently, terminology becomes a component for a sustainable, human-centered knowledge infrastructure in manufacturing, where tacit knowledge is systematically elicited, formalized and incorporated into digital assistance systems.

ACKNOWLEDGMENT

This work was supported by the German Ministry of Economic Affairs and Climate Action under grant 13IK026A.

REFERENCES

Barron, R. C., Grantcharov, V., Wanna, S., Eren, M. E., Bhattarai, M., Solovyev, N., Tompkins, G., Nicholas, C., Rasmussen, K. Ø. and Matuszek, C., et al. (2024), Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization. In *2024 International Conference on Machine Learning and Applications (ICMLA), 18.12.2024–20.12.2024,* Miami, FL, USA, IEEE, pp. 1669–1676.

Bei, Y., Fang, Z., Mao, S., Yu, S., Jiang, Y., Tong, Y. and Cai, W. (2024), Manufacturing Domain QA with Integrated Term Enhanced RAG. In *2024 International Joint Conference on Neural Networks (IJCNN), 30.06.2024–05.07.2024,* Yokohama, Japan, IEEE, pp. 1–8.

Chase, H. (2022), *LangChain*, Available at: https://github.com/langchain-ai/langchain (Accessed 30 September 2025).

Cheng, M., Luo, Y., Ouyang, J., Liu, Q., Liu, H., Li, L., Yu, S., Zhang, B., Cao, J., Ma, J., Wang, D. and Chen, E. (2025), *A Survey on Knowledge-Oriented Retrieval-Augmented Generation*, Available at: http://arxiv.org/pdf/2503.10677v2.

Di Nunzio, G. M. (2025), Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths.

Docling Team (2024), *Docling*, Available at: https://github.com/docling-project/docling (Accessed 30 September 2025).

Es, S., James, J., Espinosa-Anke, L. and Schockaert, S. (2024), *Ragas: Automated Evaluation of Retrieval Augmented Generation*.

Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M.-W. (2020), *REALM: Retrieval-Augmented Language Model Pre-Training*, Available at: http://arxiv.org/pdf/2002.08909v1.

Jose, S., Nguyen, K. T., Medjaher, K., Zemouri, R., Lévesque, M. and Tahan, A. (2024), Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models, *Expert Systems with Applications* 255: 124603.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W. (2020), Dense Passage Retrieval for Open-Domain Question Answering. In Webber, B., Cohn, T., He, Y. and Liu, Y. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 6769–6781.

Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S. and Niforatos, E. (2024), Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking, *Frontiers in artificial intelligence* 7: 1293084.

Lee, D., Kim, J., Kim, J., Hwang, S. and Park, J. (2025), tRAG: Term-level Retrieval-Augmented Generation for Domain-Adaptive Retrieval. In Chiruzzo, L., Ritter, A. and Wang, L. (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 6566–6578.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020), *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, Available at: http://arxiv.org/pdf/2005.11401v4.

Llama Team and AI @ Meta (2024), *The Llama 3 Herd of Models*.

Muennighoff, N., Tazi, N., Magne, L. and Reimers, N. (2023), *MTEB: Massive Text Embedding Benchmark*.

Naqvi, S. M. R., Ghufran, M., Varnier, C., Nicod, J.-M., Javed, K. and Zerhouni, N. (2024), Unlocking maintenance insights in industrial text through semantic search, *Computers in Industry* 157-158: 104083.

Nonaka, I. and Takeuchi, H. (1995), *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, Oxford.

Ollama Inc. (2025), *Ollama: local framework for running language models*, Available at: https://github.com/ollama/ollama (Accessed 30 September 2025).

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y. and Tang, S. (2024), *Graph Retrieval-Augmented Generation: A Survey*.

Rank, Y., Streloke, L., Bründl, P., Bodendorf, F. and Franke, J. (2025), Large Language Models for Tacit Knowledge Elicitation in Industry 5.0: A Literature Review. In *The Human Side of Service Engineering, July 26-30, 2025*, AHFE International.

Tang, Y. and Yang, Y. (2024), *Do We Need Domain-Specific Embedding Models? An Empirical Investigation*, Available at: http://arxiv.org/pdf/2409.18511v4.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. (2021), *BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models*, Available at: http://arxiv.org/pdf/2104.08663v4.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F. (2024), *Multilingual E5 Text Embeddings: A Technical Report*.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N. and Xue, C. J. (2024), *Retrieval-Augmented Generation for Natural Language Processing: A Survey*.

Zhang, C., Chen, J., Li, J., Peng, Y. and Mao, Z. (2023), Large language models for human–robot interaction: A review, *Biomimetic Intelligence and Robotics* 3: 100131.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J. and Cui, B. (2024), *Retrieval-Augmented Generation for AI-Generated Content: A Survey*.