Journal of Machine Engineering, 2026, Vol. 26
ISSN 1895-7595 (Print) ISSN 2391-8071 (Online)

Received*: 05 February 2026* / Accepted: *02 March 2026* / Published online: *16 April 2026*

manufacturing,
artificial intelligence,
information retrieval,
large language models

Tom KELLER[1*],
Carsten WOHLGEMUTH[1],
Eike PERMIN[1]

EVALUATING DATA SUITABILITY FOR RAG SYSTEMS IN MANUFACTURING: A COMPARISON BETWEEN HUMAN AND LLM JUDGMENTS

Retrieval-Augmented Generation (RAG) is widely used for manufacturing assistance, but its effectiveness depends on selecting retrievable text units. We test whether humans or Large Language Models (LLMs) can judge which case descriptions are better suited as RAG inputs. We constructed 100 synthetic manufacturing service cases, each paired with a realistic query and two comparable problem–solution variants differing in contextual completeness, granularity, and quality. Five engineers and five LLMs chose the variant expected to be more retrievable and useful. As a reference, both variants were indexed in a minimal retrieval setup with one chunk per case and evaluated with MRR@3, treating the case-matching chunk as the only relevant item among distractors. LLMs showed much higher within-group agreement than humans, yet neither cohort consistently matched retrieval-derived winners. Ties were frequent; on non-tied cases, majority decisions fell below chance and were significantly worse than random guessing in one embedding setting, while no individual rater achieved above-chance performance. Overall, the findings indicate that perceived RAG-fitness is not a reliable proxy for retrieval performance and should be grounded in retrieval-based evaluation under the target deployment setup.

1. INTRODUCTION

Retrieval-Augmented Generation (RAG) has become a prominent approach for deploying Large Language Models (LLMs) in manufacturing assistance, mainly because it is described as mitigating key limitations of standalone LLM use. By grounding responses in retrieved, domain-specific context, RAG is positioned as a way to reduce hallucinations, handle outdated knowledge, and improve practical usability in real-world settings [1, 2]. In manufacturing-oriented troubleshooting, RAG systems are described as retrieving from technical manuals, expert databases, and organizational knowledge sources to synthesize

[1] Institute of General Mechanical Engineering, Faculty of Computer Science and Engineering Science, TH Köln University of Applied Science, Germany
[*] E-mail: tom.keller@th-koeln.de
https://doi.org/10.36897/jme/218709

context-specific guidance, with the stated goal of reducing downtime and improving resolution performance [3]. Industrial machine-tool systems increasingly aim to support operators and reduce downtime by combining sensor and control data with digital assistance functions [4]. RAG is also increasingly used to query proprietary internal knowledge through secure pipelines, which is discussed as particularly relevant for industries handling sensitive information and requiring privacy compliance [5]. Related manufacturing work on data spaces emphasizes that controlled, sovereignty-preserving data exchange is a central prerequisite for deploying AI services on proprietary industrial knowledge [6]. Empirical results from an automotive case study additionally suggest efficiency benefits, as RAG-based search is reported to reduce task completion time and page visits compared to traditional search tools [7].

However, the practical effectiveness of these systems depends heavily on the quality of the indexed text units. In industrial environments, where repositories often contain redundant, overlapping, or inconsistently formulated content, selecting the most suitable textual variants for the retriever remains a significant challenge.

Against this background, it becomes relevant to ask who can make suitability decisions reliably at scale. Human judgment is an intuitive reference point but may vary across individuals and does not scale easily. LLMs are increasingly used for text-centric evaluation and curation tasks, and current LLMs were trained predominantly on human-generated text and are typically initialized from large pre-trained models before downstream fine-tuning [8]. This suggests potential alignment with human decision patterns, but it remains unclear whether LLMs can match or exceed human performance when judging which problem-solution descriptions are more suitable inputs for RAG.

This paper therefore investigates whether humans or LLMs more accurately assess the RAG-fitness of manufacturing service-case descriptions. Our analysis focuses on four aspects that matter for scalable data selection: (1) the internal consistency of human judgments, (2) the internal consistency of LLM judgments, (3) performance relative to a random baseline, and (4) whether the best-performing individual human or the best-performing individual LLM achieves superior performance. The scientific contribution of this study is empirical evidence on the reliability of human and LLM judgment as a proxy for retrieval performance, advancing the understanding of how RAG-oriented data selection for manufacturing assistance should be operationalized and validated.

2. LITERATURE REVIEW

A central reason why RAG is sensitive in practice is its architecture. RAG is commonly described as a retriever-generator system in which a retriever selects relevant information from a data store and the generator produces an answer conditioned on the user query and the retrieved text [1, 9]. In practical deployments, this retrieval step is preceded by indexing, which includes extracting and cleaning heterogeneous enterprise sources, for example PDF, HTML, Microsoft Word, or Markdown, and converting them into a uniform text format before chunking, vectorization, and storage in a vector database [1]. Manufacturing knowledge sources also include process-near data from tool-integrated sensors used for

monitoring and maintenance decisions, complementing document-centric corpora [10]. Typical implementations therefore combine indexing and preprocessing with chunking, embedding into a vector store and generation [1, 5]. Figure 1 summarizes this basic RAG setup and the interaction between indexing, retrieval, augmentation, and generation.

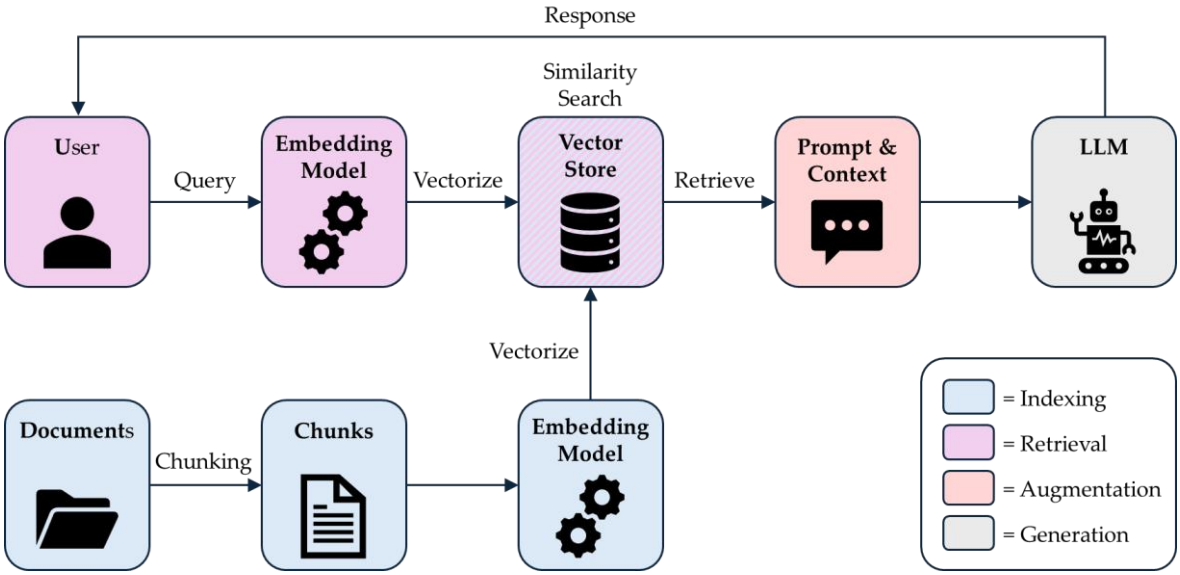

Fig. 1. RAG Architecture

Recent research therefore highlights data quality as a prerequisite for robust RAG development. Practitioners report that data quality issues occur across multiple RAG stages and can propagate downstream, which motivates early and step-specific quality controls rather than only evaluating the final generated output [11]. Industry perspectives similarly treat “RAG quality” as comprising the quality of the underlying data, the retriever, and the generated answers, explicitly emphasizing data quality as a key determinant [12]. From a trustworthiness perspective, retrieved context that is redundant, outdated, or noisy can directly harm answer accuracy and robustness, strengthening the case for corpus curation before deployment [13].

This requirement directly collides with the reality of enterprise text corpora. Organizations often maintain large repositories of operational text, including tickets, issue reports, and maintenance documentation, where overlap and duplication are common. Studies on large bug and issue repositories report substantial duplicate shares, frequently in the 10–30% range and in some cases higher [14–16]. Duplication increases triage effort and becomes harder to handle when semantically similar cases are phrased differently, because surface-level similarity is insufficient. A manufacturing-adjacent analogue appears in maintenance logs for predictive maintenance, which are reported to contain near-duplicate entries alongside other data problems such as typos and missing fields [17]. At the same time, such repositories are operationally attractive for retrieval-based assistance, since retrieving from historical cases can reduce resolution time in production settings [18]. This creates a practical selection problem: which textual variants are better suited as knowledge units for RAG systems, given that corpora may include redundant and inconsistently formulated content.

3. METHODOLOGY

The study followed a controlled A/B design to compare human and LLM judgments of data suitability for RAG in manufacturing. We constructed 100 synthetic manufacturing service cases in German. Each case consisted of a problem–solution description intended to serve as a retrievable knowledge unit and a corresponding query phrased to resemble how practitioners would realistically ask for support in production environments, with case structure and wording aligned to the style of real-world practitioner case descriptions to ensure typicality. For every case, two variants (A/B) of the problem–solution description were authored. Variants were designed to differ systematically in contextual completeness, granularity, and overall textual quality (e.g., presence of boundary conditions and parameters, degree of stepwise operational detail versus compressed technical shorthand, clarity and structure). These aspects were varied across cases to avoid a single dominant manipulation pattern. To prevent length from becoming a confounder, variant lengths were controlled using token counts; both variants were kept approximately equal in length, so that observed differences primarily reflect contextual completeness, granularity, and quality rather than document size.

Human judgments were collected from five engineers via a Google Forms survey. The number of five raters was chosen as a pragmatic minimum for computing majority decisions with an unambiguous outcome (minimum 3/5). Given the extremely detailed nature of the questionnaire with 100 elements, requiring a high time investment, the focus was placed on recruiting five highly suitable experts to ensure evaluation quality. This sample size also ensures a symmetric comparison with the LLM cohort. The participants included two mechanical engineers and three industrial engineers with professional experience in manufacturing companies. They are currently active in sustainability management, production planning, R&D, quality inspection, and project management. For each case, participants reviewed the query alongside Variant A and Variant B and selected the variant they considered more suitable as a RAG knowledge unit for answering the query. Participants were explicitly instructed to focus on expected retrievability and usefulness for downstream answer generation rather than on which variant was more factually correct. To contextualize familiarity with LLM-based tooling, participants reported their frequency of LLM use; three indicated daily use and two at least weekly use. The same A/B decision task was then administered to five leading LLMs (ChatGPT-5.2, Gemini 2.5 Flash, Claude Sonnet 4.5, DeepSeek-V3.2, Grok 4). These specific models were selected as they represented both highly powerful and simultaneously freely available models from leading LLM providers at the time of the investigation. All models were prompted and evaluated in German to match the language of the authored cases and queries. Each model was evaluated using an identical prompt template that constrained outputs to a binary choice without explanation, preventing post-hoc rationalization and ensuring comparability across models: “Your task is to select, for each case, the variant (A or B) that is better suited as a basis for a RAG system to answer the given QUERY. The goal is not to decide which variant is more factually correct, but which one is more likely to be retrievable, usable, and helpful for generating an appropriate answer. Output only “A” or “B”, without any additional text. QUERY: [query]. Variant A: [text]. Variant B: [text].” For both cohorts, the odd number of raters enabled unambiguous majority

decisions per case (minimum 3/5). We further quantified within-cohort convergence by computing the per-item agreement level, defined as the fraction of raters selecting the majority option (60% for a 3–2 split, 80% for 4–1, 100% for 5–0). To assess whether human and LLM judgments exhibit consistency beyond chance, we compared the observed distribution of agreement levels against the distribution expected under independent random A/B choices ($p = 0.5$) for five raters and applied a chi-square goodness-of-fit test.

To derive an operational, retrieval-based reference for which variant is objectively more suitable for RAG, we implemented all cases in a minimal RAG retrieval setup. Chunking was configured such that each problem–solution description constituted exactly one chunk, isolating the effect of text formulation from segmentation artefacts. Retrieval embeddings were computed using Gemini embeddings in two dimensionalities (768 and 3072) to test robustness across embedding capacity. For each embedding setting and for each case, retrieval was performed against a corpus of 100 case chunks, and the relevant item for a given query was defined as the chunk corresponding to that same case; all other case chunks in the corpus served as non-relevant distractors. Retrieval performance was quantified using Mean Reciprocal Rank at three (MRR@3). MRR@3 was selected as a practice-oriented indicator because many RAG pipelines forward only a small number of retrieved chunks (often top-3) to downstream generation due to context window and cost constraints; thus, performance differences outside the top-3 may not translate into practical answer improvements. For each case and embedding setting, MRR@3 was computed separately for the A-variant corpus and for the B-variant corpus, yielding an objective retrieval-based winner (A better, B better) or a tie when both variants achieved the same MRR@3. Because MRR@3 is discrete and top-k truncated (here, $k = 3$), ties were explicitly tracked and excluded from winner-based accuracy analyses, as no objective preference can be inferred under the chosen metric.

Judgment accuracy was computed by comparing the majority choice of each cohort (humans & LLMs) to the retrieval-based winner on the subset of non-tied cases. In addition to cohort-level majority performance, we evaluated individual-level performance for each human and each LLM to identify best-performing raters within each cohort. To quantify whether observed performance deviated from a random baseline, we used exact one-sided binomial tests with chance level $p = 0.5$ on the non-tied subset, with the direction of the alternative hypothesis chosen according to the observed accuracy (above-chance or below-chance). Finally, we reported whether the best-performing individual human and best-performing individual LLM achieved superior performance under each embedding setting; these best-performer results were treated as exploratory, as selecting the maximum performer introduces selection bias and inflates apparent performance.

4. RESULTS

This section first reports within-cohort agreement patterns for humans and LLMs. It then evaluates cohort- and individual-level accuracy against retrieval-derived winners under MRR@3, including comparisons to a 0.5 chance baseline on the non-tied subset (with effective sample sizes differing by embedding setting). Best-performing raters within each cohort are identified and reported as part of the individual-level analysis.

4.1. INTERNAL CONSISTENCY OF HUMAN AND LLM JUDGMENTS

Across the 100 cases, the LLM cohort exhibited substantially higher convergence than the human cohort (Fig. 2). Full agreement (100%, i.e., 5–0) occurred in 63/100 cases for LLMs but in only 6/100 cases for humans. Conversely, minimal majorities (60%, i.e., 3–2) dominated the human cohort (70/100 cases) but were comparatively rare for LLMs (15/100 cases).

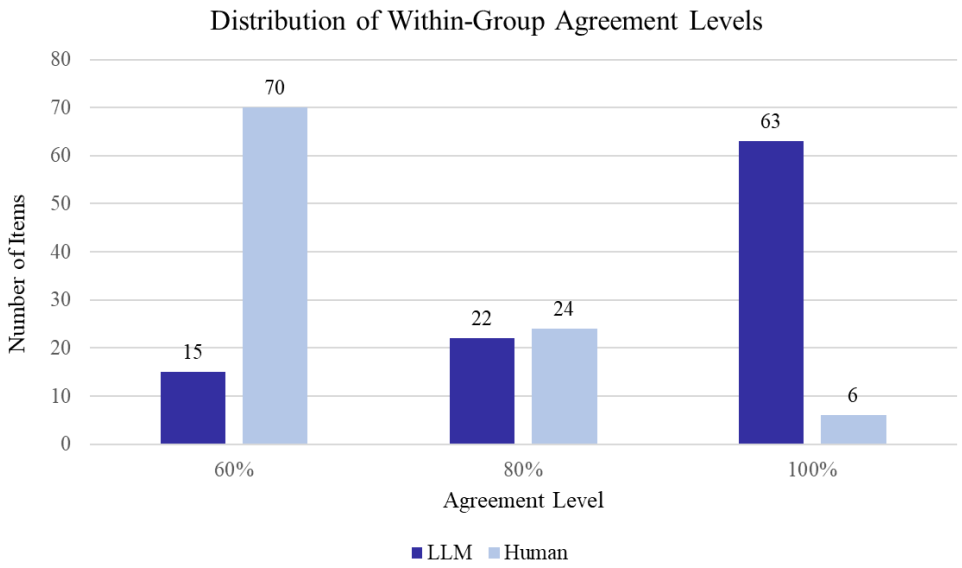

Fig. 2. Agreement Levels by Cohort

To test whether the observed agreement distributions exceed what would be expected from independent random A/B voting ($p = 0.5$, five raters), we applied a chi-square goodness-of-fit test. The LLM cohort shows a strong deviation from chance ($\chi^2(2) = 554.13$, $p = 4.7 \times 10^{-121}$); the human cohort does not ($\chi^2(2) = 2.59$, $p = 0.274$). This indicates that LLM judgments are significantly more internally consistent than expected under random voting, whereas human agreement does not show detectable above-chance convergence.

4.2. RETRIEVAL-DERIVED WINNERS UNDER MRR@3 AND BASELINE COMPARISON

MRR@3 comparisons between Variant A and Variant B resulted in a high proportion of ties, yielding embedding-dependent effective sample sizes for winner-based evaluation. Under the 768-dimensional embedding, 52/100 cases were ties, leaving 48 non-tied cases. Under the 3072-dimensional embedding, ties increased to 73/100 cases, leaving 27 non-tied cases. These values are depicted below in Fig. 3.

Cohort-level accuracy was computed by comparing each cohort's majority decision to the retrieval-based winner on the non-tied cases. In both embedding settings, the human-majority and LLM-majority decisions produced identical hit counts. For the 768-dimensional

embedding, the cohorts matched the retrieval-based winner in 15 of 48 non-tied cases (31.25%). For the 3072-dimensional embedding, they matched the winner in 10 of 27 non-tied cases (37.04%). Thus, at the cohort-majority level, neither humans nor LLMs achieved alignment with the retrieval-derived winner above chance; instead, both cohorts were substantially below the 0.5 baseline in both embedding conditions.

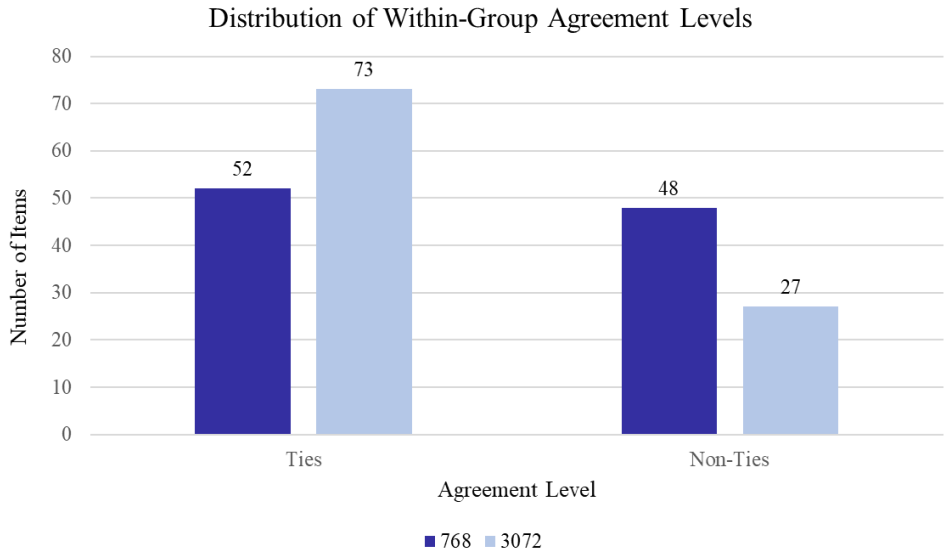

Fig. 3. Ties and Non-Ties by Embedding Setting (MRR@3)

To assess whether this below-chance performance is statistically distinguishable from random guessing, we applied exact one-sided binomial tests against $p = 0.5$ on the non-tied subsets (testing for worse-than-chance performance, given observed accuracies < 0.5). For the 768-dimensional embedding, the result is significantly below chance ($p = 0.00664$), indicating that random guessing would outperform the observed cohort-majority decisions at $\alpha = 0.05$. For the 3072-dimensional embedding, the result is not significantly below chance ($p = 0.12389$), meaning that the observed below 0.5 accuracy cannot be clearly distinguished from chance performance at $\alpha = 0.05$. Overall, these results show that both cohorts – humans and LLMs – fail to reliably identify the retrieval-derived winner under MRR@3 at the majority-vote level, and that the evidence for worse-than-chance performance is embedding-dependent, reaching significance only in the 768-dimensional setting.

4.3. INDIVIDUAL-LEVEL PERFORMANCE AND BEST-PERFORMING RATERS

Individual-level performance was evaluated for each of the five LLMs and five human raters on the same non-tied subsets used for the retrieval-based winner comparison. Performance is reported as hit rate (accuracy) relative to a 0.5 chance baseline, with exact one-sided binomial tests provided for both worse-than-chance and better-than-chance performance. For the 768-dimensional embedding ($n = 48$ non-tied cases), individual hit rates and p-values are summarized in Table 1.

Table 1. Individual Performance vs. Chance (768-Dim Embedding)

Rater	Hit rate	*p* (worse than chance)	*p* (better than chance)
ChatGPT-5.2	31.25%	0.00664	0.99724
Gemini 2.5 Flash	29.17%	0.00276	0.99896
Claude Sonnet 4.5	31.25%	0.00664	0.99724
DeepSeek-V3.2	43.75%	0.23544	0.84384
Grok 4	43.75%	0.23544	0.84384
Human 1	39.58%	0.09671	0.94430
Human 2	39.58%	0.09671	0.94430
Human 3	58.33%	0.90329	0.15616
Human 4	52.08%	0.66727	0.44272
Human 5	47.92%	0.44272	0.66727

Under the 768-dimensional embedding, three models (ChatGPT, Gemini, and Claude) perform significantly below chance in one-sided worse-than-chance tests ($p < 0.05$), whereas DeepSeek and Grok do not differ significantly from chance. Human raters achieve higher peak hit rates than the best LLM in this setting; however, even the top human result does not reach one-sided significance for above-chance performance at $\alpha = 0.05$. Overall, this embedding condition yields several statistically detectable below-chance outcomes but no statistically confirmed above-chance performance for either cohort. For the 3072-dimensional embedding ($n = 27$ non-tied cases), corresponding individual hit rates and p-values are reported in Table 2.

Table 2. Individual Performance vs. Chance (3072-Dim Embedding)

Rater	Hit rate	p (worse than chance)	p (better than chance)
ChatGPT-5.2	25.93%	0.00958	0.99704
Gemini 2.5 Flash	37.04%	0.12389	0.93896
Claude Sonnet 4.5	37.04%	0.12389	0.93896
DeepSeek-V3.2	48.15%	0.50000	0.64945
Grok 4	40.74%	0.22103	0.87611
Human 1	37.04%	0.12389	0.93896
Human 2	33.33%	0.06104	0.97388
Human 3	66.67%	0.97388	0.06104
Human 4	66.67%	0.97388	0.06104
Human 5	29.63%	0.02612	0.99042

In this embedding setting, the LLM results cluster around chance, apart from ChatGPT, which falls significantly below chance. The human results show a similarly mixed picture: one rater is significantly below chance, while the strongest human performances remain just short of one-sided above-chance significance.

Accordingly, although the highest individual hit rates in the 3072-dimensional condition are achieved by human raters, neither cohort contains an individual whose performance can be statistically confirmed as above chance on the non-tied subset at $\alpha = 0.05$. Notably, one-sided above-chance significance would require at least 19 correct decisions ($\geq 70.4\%$), whereas the best human results reach 18 (66.7%), i.e., one correct decision short of this threshold.

5. DISCUSSION

The results highlight a consistent gap between perceived RAG-fitness and retrieval-measured RAG-fitness under a concrete operationalization. Although the LLM cohort showed markedly higher within-group convergence than the human cohort, this convergence did not translate into higher alignment with the retrieval-derived winner under MRR@3. At the majority-vote level, both cohorts performed below the 0.5 baseline in both embedding settings, and in the 768-dimensional embedding the below-chance deviation was statistically significant. At the individual level, several raters also exhibited statistically detectable below-chance performance, while no rater, human or LLM, could be confirmed as above chance under the strict winner-based evaluation on the non-tied subset. This pattern suggests that both humans and LLMs may apply selection heuristics that appear plausible for well-written documentation or useful context, but that these heuristics are not reliably predictive of the retrieval behaviour produced by a specific embedding model and a top-3 truncation. The findings therefore caution against equating internal agreement, particularly for LLMs, with correctness relative to an operational retrieval objective and extend the work of Müller and Holstein [11], who identified that data quality issues propagate through multiple RAG stages. Our results demonstrate that even expert-based selection fails to serve as a reliable early-stage control for retrieval performance. Furthermore, while industry perspectives treat data quality as a key determinant of RAG success, our study clarifies that "quality" in this context is a functional property tied to retrieval behavior that often eludes subjective human or LLM intuition.

A key methodological factor is the high tie rate under MRR@3, which reduced the effective sample size for winner-based testing to 48 and 27 cases, depending on embedding dimensionality. Substantively, ties indicate that, for a large portion of cases, Variant A and Variant B were indistinguishable with respect to top-3 rank behaviour in the chosen retrieval stack. This is practically relevant: even when variants differ in contextual completeness, granularity, and textual quality with length controlled, these differences often do not change whether the relevant chunk is retrieved early enough to be forwarded to generation in many top-k RAG settings. At the same time, fewer decisive cases limit statistical power and raise the threshold for detecting above-chance performance, as seen in the 3072-dimensional setting where the best human results approached but did not reach one-sided significance. Accordingly, the absence of statistically confirmed above-chance performance should be interpreted cautiously, reflecting both task difficulty under this operationalization and the prevalence of effective no-difference cases under MRR@3. The high tie prevalence also suggests that increasing retrieval depth, for example via MRR@10, would not necessarily change the inferential picture here, because the dominant driver is frequent equality of the two variants' MRR@3 outcomes rather than missing the relevant chunk in the top-3.

From an organizational perspective, the results imply that manual curation based on intuition alone, even by technically trained engineers, may be a weak proxy for how a text unit behaves in an embedding-based retriever, and that LLM-based screening, while scalable and internally consistent, is not automatically aligned with the retrieval objective either. For companies facing the selection problem described in the introduction, namely large heterogeneous service and maintenance corpora with overlap, redundancy, and inconsistent

formulation, this suggests a shift from choosing the better text to measuring the better text. A pragmatic workflow treats humans and LLMs as generators or normalizers and validates alternatives through retrieval evaluation on representative queries. The tie results further suggest that firms may not need to optimize every record aggressively. If many items are retrieval-equivalent at top-k, curation effort can be prioritized toward cases that are retrieval-sensitive or safety-critical. In practice, this points toward tiered curation strategies: enforce a minimal quality and structure bar globally and run targeted retrieval diagnostics where the business impact of wrong or delayed retrieval is highest.

The observed embedding dependence also has implications. The effective sample size dropped substantially from the 768- to the 3072-dimensional embedding due to higher tie prevalence, and statistically significant below-chance majority performance appeared only in the 768-dimensional setting. This shows that retrieval-derived winners are not invariant, but depend on the embedding model and dimensionality, corpus composition, and evaluation metric. Accordingly, claims that one variant is objectively better for RAG must be scoped to the deployment configuration. The same holds for LLM behaviour: model choice mattered, with heterogeneous outcomes across the five LLMs tested. Given rapid model updates and continued progress in both foundation and embedding models, these findings should be read as evidence about current behaviour under a specific setup, not as a static humans-versus-LLMs verdict. This motivates periodic recalibration as embeddings, retrieval configurations, or curation models evolve.

Time and scalability considerations reinforce this point. Human evaluation of many items is costly and difficult to scale in industrial settings where domain expertise is scarce, while LLMs can deliver immediate and consistent decisions at minimal marginal cost. However, the present findings indicate that speed and internal agreement do not guarantee retrieval-aligned selection quality. This motivates hybrid approaches: LLMs reduce human workload by pre-structuring texts and standardizing terminology, while retrieval-based tests serve as an objective gate and human effort is reserved for auditing edge cases and high-risk content where incorrect retrieval has safety or downtime implications. Such a workflow reflects the architecture-sensitive nature of RAG outlined in the introduction. Because retrieval performance is a primary determinant of downstream answer quality, curation should be evaluated at the retriever, not only at the level of perceived text quality.

Finally, the results suggest several research directions to strengthen the scientific understanding of RAG-fitness as a measurable property of text. First, both cohorts worked under relatively generic instructions. Future studies can test whether structured rubrics for humans, for example requiring boundary conditions, parameter values, component identifiers, and stepwise actions, increase both agreement and retrieval alignment. Similarly, LLM prompts could be engineered to enforce explicit consideration of retrievability, such as terminology overlap with likely queries and discriminative identifiers, while still yielding a forced A/B choice. Second, future work can isolate concrete textual features, such as standardized part numbers, error codes, synonym coverage, abbreviation handling, explicit parameter ranges, and consistent naming, and quantify their marginal effects on retrieval rank and tie likelihood across embeddings. Third, moving beyond a minimal one-chunk-per-case setup toward enterprise pipelines, including multi-chunk documents, metadata filters, hybrid lexical-semantic retrieval, and re-ranking, will clarify whether formulation differences

become more consequential once chunking and ranking refinements are introduced. Overall, the findings imply that scalable RAG-oriented data selection is best treated as an engineering and measurement problem: humans and LLMs can generate and normalize case descriptions, but selection should be validated against the retrieval behaviour of the intended deployment stack.

6. CONCLUSION

This study compared human and LLM judgments of data suitability for RAG in manufacturing using 100 controlled A/B pairs of problem–solution descriptions with length held approximately constant while contextual completeness, granularity, and textual quality were varied. LLMs exhibited substantially higher internal agreement than humans, indicating stable and shared selection heuristics within the model cohort. However, when judgments were evaluated against retrieval-derived winners under a minimal RAG setup using MRR@3, neither cohort reliably identified the better-retrieving variant. Majority-vote performance was below the 0.5 baseline in both embedding settings and was significantly worse than chance in the 768-dimensional condition, while individual-level analyses likewise showed no statistically confirmed above-chance performance for any human or LLM under the non-tied winner-based evaluation.

A high proportion of retrieval ties under MRR@3 indicates that many formulation differences do not translate into top-3 retrieval advantages in this setup but also reduces the sample size for winner-based inference and sharpens significance thresholds. Overall, the findings suggest that perceived RAG-fitness, whether assessed by humans or LLMs, is not a reliable proxy for retrieval performance in an embedding-based retriever without validation. For practice, this supports retrieval-grounded curation workflows in which alternative formulations are generated or normalized by humans or LLMs, but selection is validated using the retrieval configuration intended for deployment. Future research should isolate textual features that influence retrieval outcomes, and test whether structured rubrics for humans or improved prompting for LLMs can increase alignment with retrieval-based objectives under realistic RAG pipelines.

ACKNOWLEDGEMENTS

This paper was prepared in the context of the project "KI-ssist", which is funded within the innovation competition NEXT.IN.NRW by the European Union and the State of North Rhine-Westphalia through the EFRE/JTF Programme NRW 2021–2027 (EFRE-20800990).

REFERENCES

[1] GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG M., WANG H., 2024, *Retrieval-Augmented Generation for Large Language Models: a Survey*, arXiv.

[2] SHAN R., SHAN T., 2025, *Retrieval-Augmented Generation Architecture Framework: Harnessing the Power of RAG*, Cognitive Computing - ICCC 2024, 88–104, https://doi.org/10.1007/978-3-031-77954-1_6.

[3] CHENG M., LUO Y., JIE O., LIU Q., LIU H., LI L., YU S., ZHANG B., CAO J., MA J., WANG D., CHEN E., 2025, *A Survey on Knowledge-Oriented Retrieval-Augmented Generation*, ArXiv, abs/2503.10677, https://doi.org/10.48550/arxiv.2503.10677.
[4] FLEISCHER J., PUCHTA A., GÖNNHEIMER P., 2021, *Seamless and Modular Architecture for Autonomous Machine Tools*, Journal of Machine Engineering, https://doi.org/10.36897/jme/141565.
[5] OCHE A.J., FOLASHADE A.G., GHOSAL T., BISWAS A., 2025, *A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress*, Gaps, and Future Directions, arXiv.
[6] FRIEDRICH C., VOGT S., RUDOLPH F., PATOLLA P., GRÜTZMANN J.M., HOHMEIER O., RICHTER M., WENZEL K., REICHELT D., IHLENFELDT S., 2024, *Enabling Federated Learning Services Using OPC UA, Linked Data and GAIA-X in Cognitive Production*, Journal of Machine Engineering, 24/2, 18–33, https://doi.org/10.36897/jme/188618.
[7] MAYAT N., WACHTER C., SPATZENEGGER S., HINRICHS M.P., WEISSER T., SCHMITT R.H., 2025, *Performance of Rag-Based Systems in Industrial Organizations: A Case Study in the Automotive Industry*, IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS), 1–6, https://doi.org/10.1109/icps65515.2025.11087842.
[8] SHUMAILOV I., SHUMAYLOV Z., ZHAO Y., PAPERNOT N., ANDERSON R., GAL Y., 2024, *AI Models Collapse when Trained on Recursively Generated Data*, Nature, 631/8022, 755–759, https://doi.org/10.1038/s41586-024-07566-y.
[9] ZHAO P., ZHANG H., YU Q., WANG Z., GENG Y., FU F., YANG L., ZHANG W., JIANG J., CUI B., 2024, *Retrieval-Augmented Generation for AI-Generated Content: A Survey*, arXiv.
[10] BLEICHER F., RAMSAUER C., LEONHARTSBERGER M., LAMPRECHT M., STADLER P., STRASSER D., WIEDERMANN C., 2021, *Tooling Systems with Integrated Sensors Enabling Data Based Process Optimization*, Journal of Machine Engineering, 5–21, https://doi.org/10.36897/jme/134244.
[11] MÜLLER J., HOLSTEIN, 2025, Data Quality Challenges in Retrieval- Augmented Generation, *https://doi.org/10.48550/arXiv.2510.00552*.
[12] BREHME L., DORNAUER B., STRÖHLE T., EHRHART M., BREU R., 2025, *Retrieval-Augmented Generation in Industry: an Interview Study on Use Cases, Requirements, Challenges, and Evaluation*, Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 110–122, https://doi.org/10.5220/0013739500004000.
[13] ZHOU Y., LIU Y., LI X., JIN J., QIAN H., LIU Z., LI C., DOU Z., HO T.-Y., YU P.S., 2024, *Trustworthiness in Retrieval-Augmented Generation Systems: A Survey*, arXiv.
[14] CAVALCANTI Y.C., DA MOTA SILVEIRA NETO P.A., LUCRÉDIO D., VALE T., DE ALMEIDA E.S., DE LEMOS MEIRA S.R., 2013, *The Bug Report Duplication Problem: an Exploratory Study*, Software Qual J, 21/1, 39–66, https://doi.org/10.1007/s11219-011-9164-5.
[15] EBRAHIMI N., TRABELSI A., ISLAM MD.S., HAMOU-LHADJ A., KHANMOHAMMADI K., 2019, *An HMM-Based Approach for Automatic Detection and Classification of Duplicate Bug Reports*, Information and Software Technology, 113, 98–109, https://doi.org/10.1016/j.infsof.2019.05.007.
[16] JAHAN S., RAHMAN M.M., 2022, *Towards Understanding the Impacts of Textual Dissimilarity on Duplicate Bug Report Detection*, arXiv.
[17] DIMIDOV V., HAWLADER F., JAFARNEJAD S., FRANK R., 2025, *Cleaning Maintenance Logs with LLM Agents for Improved Predictive Maintenance*, arXiv.
[18] XU Z., CRUZ M.J., GUEVARA M., WANG T., DESHPANDE M., WANG X., LI Z., 2024, *Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering*, Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2905–2909, https://doi.org/10.1145/3626772.3661370.