Enabling smart maintenance in SMEs via Domain-Specific RAG: An industrial field study

Mirco Lecce[a]*, Andrea Bioddo[b], Luca Dadone[b], Gaia Falchi[a] and Vittorio Cesarotti[a]

[a]*University of Rome Tor Vergata, Department of Enterprise Engineering Mario Lucertini, Via del Politecnico 1, 00133, Rome, Italy;*

[b]*Neuratio S.r.l., Turin, Italy*

mirco.lecce@uniroma2.it*

Enabling smart maintenance in SMEs via Domain-Specific RAG: An industrial field study

Abstract: Reactive maintenance management in Small and Medium-sized Enterprises (SMEs) is often hindered by technical knowledge fragmentation and reliance on senior experts. Although generative Artificial Intelligence (AI) offers new perspectives for decision support, adopting generalist models introduces operational risks linked to hallucinations and a lack of specific context. This study analyses the effectiveness of a domain-specific Retrieval-Augmented Generation (RAG) framework, designed to encode historical corporate know-how into reliable operating procedures. Through a field validation conducted on a significant sample of real assistance requests, the research compares the proposed system with a generalist Large Language Model. The methodology adopts an evaluation protocol based on the LLM-as-a-judge technique, validated through analyses conducted with a maintenance engineer, to ensure accuracy. The results highlight how targeted context integration reduces factual inaccuracies, ensuring superior technical relevance. Such accuracy emerges as a fundamental enabling factor to speed up fault diagnosis, laying the foundations for intervention time optimization and greater overall maintenance process effectiveness.

Keywords: Retrieval-Augmented Generation; Smart Maintenance; Large Language Models; Industrial AI; LLM-as-a-judge; Human-AI Collaboration.

1. Introduction

The potential of small and medium-sized enterprises (SMEs) is threatened and limited by the difficulty of transferring and developing their people and knowledge. SMEs depend predominantly on the skills acquired over time by senior technicians [1][2][3]. This dependence creates long-term vulnerability for companies due to the high potential turnover of the workforce and the possible retirement of senior employees, leading to the loss of fundamental diagnostic capabilities [4][5]. Furthermore, the lack of structured support based on codes and a unique language for all increases the mental and technical workload of operators, which is strongly positively correlated with increased error rates and prolonged downtime [6]. Because of all this, industrial companies are increasingly researching and investing in Human-in-the-Loop paradigms where digital systems support, rather than replace, human decision-making [7]. Supporting what has been highlighted so far Artificial Intelligence (AI), specifically Large Language Models (LLMs), certainly offer the most likely solution to the transfer of knowledge from senior employees to new hires through a structured approach. Despite their potential effectiveness, the implementation of LLMs in high-risk contexts, such as industrial settings, may be limited in order to ensure high standards of reliability and safety [8][9]. It is highlighted how AI can only be considered “responsible” if it is made “robust” [10]. LLM models, such as GPT-4, are affected by intrinsic hallucinations, providing technical details that are potentially acceptable but in fact incorrect [11]. Rojek et al.[12]point out that AI must act as a supervisor to support maintenance; despite this, LLM systems have levels of variability and accuracy that fail to meet the current demands of industrial engineering activities. However, Retrieval-

Augmented-Generation (RAG) emerges as an architecture for highly complex fields. For example, in the medical sector, where Wu et al. [13] proposed Medical Graph RAG to ensure patient safety, it is possible to replicate the approach in industrial maintenance to ensure a technical basis. Dhami & Goel [14] and Yang et al. [15] highlight how specific RAG applications increase reliability compared to generic applications. By retrieving context from a verified knowledge base (e.g. historical tickets, manuals), these systems transform SMEs' low-quality data into actionable procedures. This approach shifts the paradigm from AI applications for "creative generation" to one based on "evidence-based retrieval", dramatically reducing the risk of error.

Despite the increasing use of applications such as RAG, evaluating the effectiveness of these systems, specifically in industrial contexts, remains a challenge due to the lack of standardized approaches, responses, and data. Recently, research such as RAGBench [16] and [17] highlights the need to define metrics that go beyond simple textual similarity. In this regard, this study aims to present an empirical study on the evaluation of a system, such as Neuratio, a domain-specific RAG system implemented in a real SME environment. Specifically, the research aims to answer the following research questions (RQs):

- RQ1 (Effectiveness): To what extent does a domain-specific RAG architecture improve technical responses for maintenance, compared to a generic LLM, in an SME with limited data availability?
- RQ2 (Trustworthiness): Can the application of RAG in a specific domain mitigate hallucinations, ensuring appropriate levels of decision support?
- RQ3 (Measurement): Is a Reference-Free Evaluation protocol based on an LLM-as-a-judge reliable enough to align with the judgment of a human domain experts?

To answer these questions, this study provides three distinct contributions to the literature on industrial AI and knowledge management:

- System proposal: we introduce Neuratio, an industrial AI platform based on a domain-specific RAG framework and deployed in real SME environments
- Field validation: we conduct a validation on a dataset of 119-real-world maintenance tickets.
- Methodological proposal: we propose and validate a robust evaluation protocol based on LLM-as-a-judge approach, calibrated against human expert judgement, offering a scalable method for monitoring AI reliability.

We compared the Neuratio system with a generic baseline using a dataset of 100 real maintenance tickets. Based on the research of [18] and [19], we adopted an LLM-as-a-judge methodology to quantify and validate metrics such as fidelity and context relevance, providing empirical evidence of the system's impact on knowledge retention.

2. Literature review

2.1 Knowledge management challenges in SMEs maintenance

Maintenance remains a fundamental activity within any industrial setting, although in SMEs it is typically employee-driven, mostly manual, and dependent on the diverse skills of the team [20], creating a level of dependence of the tacit knowledge of more

senior employees, which is difficult to transfer. All this, due to potential turnover levels, would expose companies to the loss of intangible assets and expose their systems to high-impact failures, especially in SMEs [21][22].

Typically, reducing the impact or probability of the risk is complicated because it is managed in a fragmented manner [23][24]. Specifically, the systemic impact that characterizes maintenance interventions forces maintenance personnel to ignore technical manual and best practices, often making it impossible to implement unified standard approaches [25][26]. Consequently, despite the importance of knowledge sharing [27], diagnostic skills and expertise are mostly intrinsic to senior maintenance workers, leading to data impoverishment and limiting the democratization of technical know-how [28].

2.2 Generative AI and the trustworthiness in AI systems

Although Large Language Models open up new development scenarios in natural language processing, integrating them into industrial environments is critical for their integration into safety-critical industrial environments is severely compromised by the phenomenon of hallucinations.
Current LLMs tend not to consider risks specific to the industrial context in which they are supposed to operate, generating responses and leading to plausible but potentially incorrect situations[29], increasing the risk of error, especially in a context such as maintenance, where the safety and effectiveness of interventions are essential [30].
AI in these contexts must guarantee certain requirements:

- Robustness against noisy or incomplete input data.
- Safety in generating instructions that comply with regulation.
- Adherence to technical and operational constraints.

In all those environments that can cause damage to people and systems [31][32]. In this regard, several studies, such as [33] and [34], demonstrate that the current LLMs available do not achieve the necessary accuracy, leading to errors in interventions that require an high level of knowledge. This can have two different consequences:

- Irrelevant actions, increasing resolution time and cognitive load on technicians, instead of providing support and improving the effectiveness of the intervention [35];
- Errors that aggravate conditions, especially in complex systems [36].

2.3 RAG architecture and evaluation protocols

To address the limitations of LLMs, an architecture has been developed for all activities that can be defined as “knowledge-intensive”, known as Retrieval-Augmented-Generation [37]. RAG systems enable access to up-to-date maintenance data without prohibitive costs by combining a generative model and a neural retriever capable of extracting relevant elements from external indexes [38][39]. However, exploiting these architectures in industrial contexts can lead to “bottlenecks” in response evaluations due to the absence of a “Ground Truth” in maintenance or inadequate reference responses according to traditional evaluation methods, such as BLEU and ROUGE, as they are unable to identify the fidelity and accuracy of token resolution for the specific procedure [40][41]. As discussed by [41], these metrics rely on n-gram overlap and fail to capture technical fidelity in specific technical domains. In response to this,

frameworks such as RAGBench [42] are emerging to evaluate the quality and fidelity of generation, leveraging models without knows references, such as “LLM-as-a-judge”. This has been validated by several studies, such as G-Eval [43] and [44], which highlight how it is possible to achieve a strong positive correlation with human judgement without the need for extensive datasets with human labels, identifying a potentially scalable solution to the new evident problem of AI reliability in industrial and high-impact areas.

3. Methodology

3.1 Methodology overview

The methodology is structured in different phases:

(1) System design: two different architectures were evaluated, namely a generic zero-shot LLM and Neuratio, a domain-specific RAG platform deployed in a real SME environment. Both systems are the same generative engine (OpenAI GPT-3.5-Turbo), highlighting the importance of context retrieval.
(2) Data collection: 119 real maintenance tickets were collected from an SME in the metalworking sector. In each case, the RAG knowledge base was built from technical manuals provided by machine manufacturers, external to the 119 tickets.
(3) Evaluation model: Due to high costs, a hybrid evaluation was adopted. Using spearman’s correlation, we validated the significance of the alignment between a senior maintenance engineer and an AI judge in order to calibrate the AI judge. Once the LLM-as-a-judge was validated, we worked on the complete dataset using a different generative engine, in fact the AI Judge is powered by OpenAI GPT-4o.
(4) Statistical analysis: by comparing the scores, we used hypothesis test to quantify the extent of improvements in both safety and utility.

In addition, the experiment pipeline is structured in **Figure 1**: Experimental pipeline.

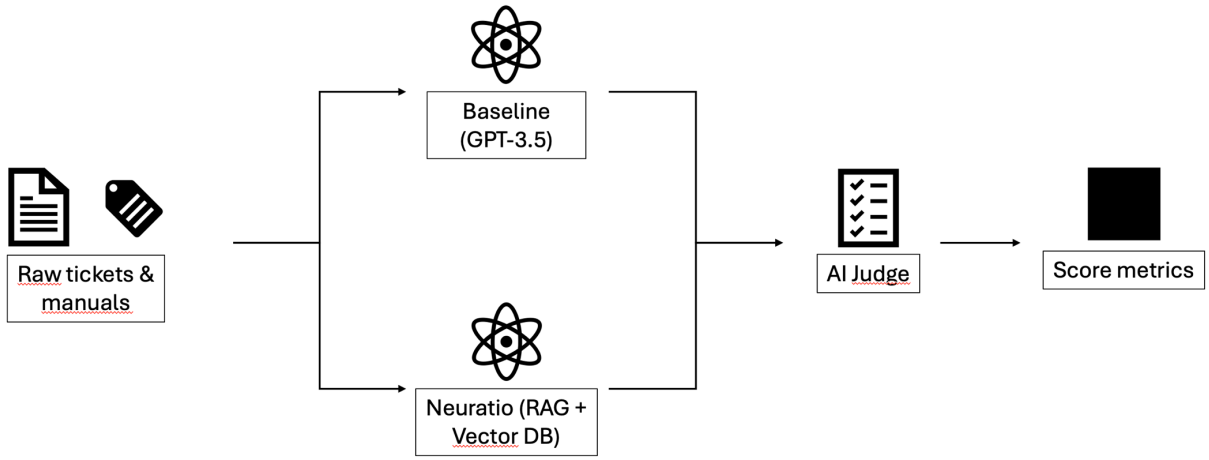


***Figure 1**: Experimental pipeline*

3.2 Neuratio – the proposed framework

Neuratio is a tool-augmented Retrieval-Augmented Generation (RAG) system designed to retrieve and synthesize technical and operational maintenance know-how, from a domain-specific knowledge base. The pipeline is organized starting from a specific knowledge base on the maintenance domain, designed to limit the dispersion of technical knowledge within operational industrial contexts, as shown in **Figure 2** - Neuratio Model.

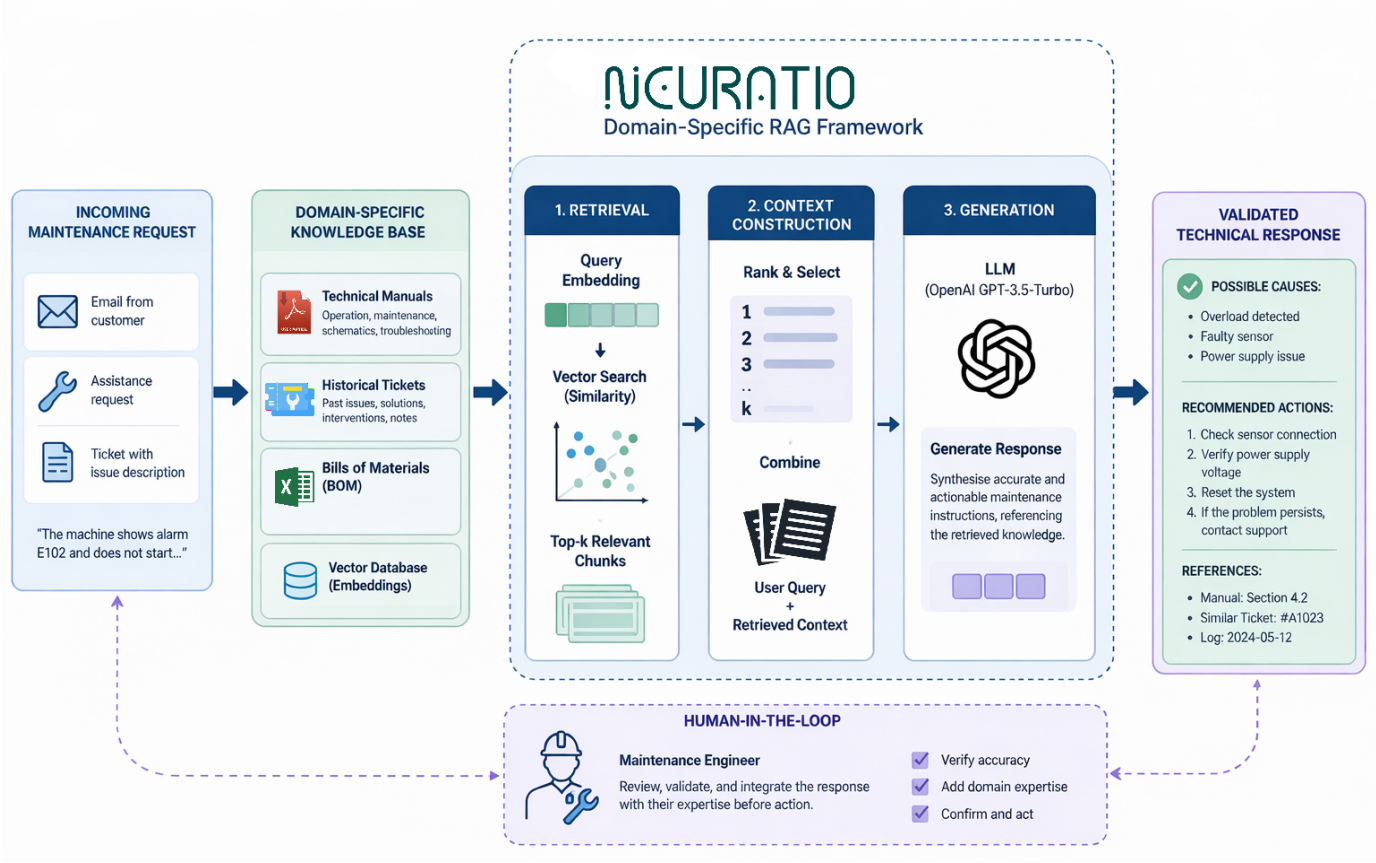


Figure 2 *- Neuratio Model*

The knowledge base is built from many technical manuals, historical maintenance tickets and Bills of Materials (BoM). These sources represent a corpus for a total of more than 526.000 tokens. Additionally, the text is divided into 1500 tokens with an overlap of 100 to ensure textual continuity. In addition, text-embedding-3-large multi-lingual is selected, leveraging a vector database optimized for low-latency similarity search. Finally, when a query is received from the user, Neuratio performs a retrieval on the top k most relevant blocks (k=100).

In this way, the generation process is not based on generic knowledge, but it is closely anchored to specific technical sources from the company and the application domain. The generation model uses the retrieved context to synthesize a coherent technical response, which may include explanations of the problem, operating instructions, and references to documentation. Generation is constrained by the content provided by retrieval, reducing the risk of unsupported or unverifiable information. The result is a structured response that reflect the formal knowledge encoded in manuals.

3.3 Experimental data collection

The model validation was conducted in collaboration with a metalworking SME, on real maintenance tickets collected over a time horizon of approximately 12 months. We collected N=119 tickets, split into N_{cal}=19 for judge calibration and N_{val}=100 for system comparison, where these 119 tickets were identified by the computerized maintenance management system (CMMS). To validate the LLM-as-a-judge approach, we divided the tickets into two groups:

- A calibration set (N_{cal}=19) to calibrate the LLM-as-a-judge with the maintenance engineer, ensuring alignment thought correlation analysis.
- An evaluation set (N_{val}=100) to compare a generic LLM and the Neuratio system.

Specifically, some precautions were made:

(1) The technical manuals and logs used to create the knowledge base are completely separate from the 119 calibration and validation tickets.
(2) All data sets were anonymized for privacy and compliance reasons.

Finally, to fully focus on the impact of the RAG architecture, we:

- Required that generic LLM and Neuratio use the same generation engine (OpenAI GPT-3.5-Turbo);
- Generic LLM presents the same system prompt, but without access to the retrieval module;
- Both systems provide the same "persona" instructions.

3.4 Evaluation model – LLM-as-a-judge

Given the impossibility of having best practices for every complex maintenance scenario and the scalability limitations of human assessment, we decided to adopt a reference-free-evaluation based on an LLM-as-a-judge approach. To this end, the LLM-as-a-judge was validated with an expert maintenance engineer.

(1) Phase I: Calibration (N_{cal}=19)

To validate the judge, the 19 calibration tickets were independently evaluated by an expert maintenance engineer and the AI evaluator, powered by the OpenAI GPT-4o engine, using a 5-point Likert scale for faithfulness and relevance. Subsequently, the alignment between the expert and AI judge's scores was quantified using Spearman's rank correlation (ρ). The expert reviewed the responses without knowing if the responses came from Neuratio or Baseline, to prevent any bias in randomized order. Additionally, despite the limited sample size for calibration, we decided to calculate 95% bootstrap confidence intervals, performing 1.000 resamples, to check the robustness of the model. The AI judge is considered aligned for evaluation when (i) the

correlation is significant ($p<0.05$) and (ii) the bootstrap 95% CI lower bound is > 0 for each metric.

(2) Phase II: Evaluation

Once validation was achieved, the remaining 100 tickets were evaluated for both Neuratio and the generic LLM. The judge was required to evaluate exclusively based on the retrieved context, penalizing any additional information. This evaluation process was repeated three times (temperature = 0), to ensure stability of results, reporting mean and standard deviation.

3.5 Metrics definition

We used two metrics commonly used in the literature to evaluate RAG [43][45]: response faithfulness and answer relevance. Typically, these metrics are evaluated on a Likert scale from 1 to 5, where 5 represents the optimal response. The detailed evaluation grid for the five levels is provided in Appendix A.

- Faithfulness: measures how free the response is from hallucinations and external priori knowledge. It allows us to understand the extent to which, given context X, response A is deduced exclusively from X. The importance of this measure in the industrial setting is linked to the generation of "useful but invented procedures", which would increase the risk to people's health and safety. The scoring scale can be interpreted as follows:
    - Score 1 (hallucinations): the response is incorrect, or contains false information, or is not aligned with the reference context.
    - Score 5 (completeness of information): the statement is aligned with the context, perfectly centered on the reference context.

- Relevance: measures the applicability of the response to the specific situation. It allows us to understand how effectively, given user X's request, the generated response Y allows us to address X, thus reducing the diagnostic impact (time and effort) on the maintenance technician.
    - Score 1 (irrelevant): the response is vague and refers to an unrelated topic.
    - Score 5 (highly relevant): the response guides the technician through the solution in detail, step-by-step.

3.6 Statistical analysis

We performed an analysis of the RAG model scores compared to the generic LLM, in our evaluation set of 100 tickets. We calculated the mean and standard deviation for both model evaluation indicators, faithfulness and relevance.

To demonstrate the significant difference between Neuratio and the generic model, we adopted the non-parametric Wilcoxon hypothesis test because of the likert scale (1-5) data is based on an ordinal distribution. The result is statistically significant for p-value < 0.05. Additionally, we measured the Win Rate, which is the percentage of

tickets in which Neuratio's response was superior to the generic model.

4. Results

In this section, we present the results derived from the validation of the Neuratio model. We will first measure the reliability of the LLM-as-a-judge protocol (RQ3). We will then compare the Neuratio model with a generic model (RQ1, RQ2) through non-parametric hypothesis test to demonstrate the significance of the experiment. The section will conclude with an analysis of different real-world maintenance scenarios to highlight the importance and impact of the Neuratio model in industrial contexts.

4.1 Reliability of LLM-as-a-judge

The validation of LLM-as-a-judge was performed by calibrating the N_{cal} sample (n=19) through comparison between the senior maintenance engineer and GPT-4o. The Spearman correlation calculation (ρ) measures a positive correlation for faithfulness and relevance, demonstrating a link between the judgment of the maintenance engineering expert and LLM-as-a-judge. Specifically, the results are a correlation $\rho=0.612$ ($p<0.01$) for faithfulness and a strong correlation $\rho=0.717$ for relevance ($p<0.001$).

Due to the limited sample size, we had to create confidence intervals with a significance level of 5% and 1,000 resampling, observing lower and upper limits of:

- Faithfulness CI: [0.108; 0,829]
- Relevance CI: [0.289; 0,917]

While the confidence intervals reflect the limited sample size of the pilot calibration, thanks to the results obtained from the correlation analysis, the LLM-as-a-judge evaluations were deemed calibrated and reliable, like the maintenance engineer's responses, and therefore reliable for the evaluation of the remaining tickets.

4.2 Comparative and quantitative analysis

There is a clear difference in the performance of the Neuratio RAG system and the generic LLM in terms of performance analysis on the dataset under consideration (Nval = 100). "Table 1 - Comparative performance on evaluation set" describes some statistics of the results.

In detail, Neuratio obtained $\mu=3.77$ ($\sigma=1.26$) in the evaluation scores, compared to $\mu=1.51$ ($\sigma=0.61$) for generic LLM. Although the standard deviation is significantly lower in the generic LLM, due to values of 1-2 in most tickets, the Neuratio average is significantly higher, with a positive value greater than +148.3%. We performed a Wilcoxon hypothesis test, obtaining a statistically significant difference between the two technologies (Faithfulness, W=35.5, Z=-8.10, p-value < 0.001; Relevance, W=73.0, Z=-8,17, p<0,001).

From a qualitative analysis of the results obtained, we can affirm the effectiveness of retrieving information from technical manuals using RAG technology, ensuring a limited impact for plausible but invented procedures.

To evaluate the results, we calculated a success rate, which is the percentage of tickets where Neuratio scored higher than the generic LLM. Neuratio clearly won 85% of tickets based on faithfulness and 88% of tickets based on relevance of answers.

Tickets where Neuratio scored lower than or equal to the generic LLM included ambiguous questions and inappropriate documentation, negatively impacting the RAG model.

Table 1 - Comparative performance on evaluation set

Metric	System	Mean (SD)	Mdn	Win Rate (%)	Wilcoxon test	Effect size (*r*)
Faithfulness	Generic LLM (GPT-3.5)	1.51 (0.61)	1.0	15%	-	
	Neuratio (RAG)	3.77 (1.26)	4.0	85%	W = 35.5, p<0.001*	0.864
Relevance	Generic LLM (GPT-3.5)	1.49 (0.57)	1.0	12%	-	
	Neuratio (RAG)	3.75 (1.24)	4.0	88%	W = 73.0, p<0.001*	0.847

4.3 Qualitative analysis

To highlight the impact and differences between the two models in industrial contexts, below you can see the details of one of the tickets, for which Neuratio obtained a score significantly higher than the baseline.

Customer request:" *We had noticed the cutter head stuttering after stopping and just before a cut as if something was loose – we were looking into this but the attached drive belt wheel failed before we found a solution. We believe the bolts on the clamp failed which caused the wheel to move damaging our guide plate and belt. I assume our bolts were changed on our end at some point as they are 8.8- we are planning to change these to 12s but I wanted your opinion on what could have caused the had to shudder when it came to a stop just before cutting. Also, I would like to see our options to get 2 new belts, 1 new pulley, and 3 new pully guide plates – especially if we can expedite those.*"

Neuratio response:" *Dear customer, Thank you for your detailed description. Based on your report of cutter head stuttering and subsequent drive belt wheel failure, we recommend the following steps: Inspect the head drive assembly for any damage, deformation, or wear, focusing on the clamp and pulley interface. Ensure all pins and screws are intact and properly tightened. Replace any damaged components and use hardware matching the original specifications (see Manual §3.1.2, p.20 and §8.2, p.84). Check and adjust the drive belt tension using the two adjustment screws as shown below*

(Manual §7.8, p.79). Incorrect belt tension or pulley movement can cause the shuddering you described. If the issue persists after these checks, please send us close-up photos of the clamp/pulley area and the damaged components so we can further assess for deformation or other issues. Regarding spare parts, to ensure we supply the correct items and can expedite delivery, please provide:

- *Photos and any markings/codes from the damaged belts, pulley, and guide plates*
- *Belt specifications (profile/pitch, length, width, manufacturer/code)*
- *Pulley details (number of teeth, pitch/profile, bore/shaft size, any stamped code)*
- *Guide plate dimensions and confirmation if they match our part P/N E001019424*
- *A photo of your machine nameplate (Model EDP-W-1650, Machine ID 112)*
- *Confirmation of the assembly location for each part and the required quantities (belts x2, pulley x1, guide plates x3)*
- *Your required delivery timeframe for expediting*

Once we receive this information, we will verify availability and provide the fastest shipment option."

Baseline response:" *If the head was stuttering specifically after stopping and just before a cut, you were likely dealing with a resonance or positioning struggle. Here are the most probable culprits:*

- *Loss of Belt Tension/Backlash: If the drive belt was already worn or the pulley clamp was slightly loose, the motor's inertia upon stopping would cause the head to "overshoot" and then "hunt" for the correct position. This creates a vibrating or shuddering effect.*
- *PID Tuning / Servo Gain: If the "stop" is too aggressive, the servo motor may oscillate trying to hold its position. If the mechanical connection (the pulley) has any play, this oscillation is magnified.*
- *Mechanical Obstruction or Binding: A slightly bent guide plate or debris in the tracks can cause the motor to fight through a "tight spot." When it stops, the stored energy in the belt snaps back, causing a stutter.*
- *The "Bolt Failure" Theory: If the bolts were already fatigued, the pulley might have been "clunking" on the shaft keyway or clamp surface. This microscopic movement translates to a visible shudder at the head."*

Neuratio integrates section numbers and manual pages (e.g., Manual #7.8, p. 79) into its responses to help resolve issues. It can also correctly identify the machine, model, part number, and any customizations made by the industrial machinery manufacturer. This ensures that Neuratio can be used by a highly pragmatic junior maintenance engineer.

5. Discussion

The results show that generic LLMs are not suitable for detailed industrial maintenance without external maintenance support due to a series of hallucinations that could have a negative impact on the machinery and people. The Neuratio model, based

on RAG technology, provides a precise interface for reading and navigating complex technical documentation. The results demonstrate the model's superiority over generic LLMs without access to company systems. Integrating a RAG model like Neuratio within industrial companies is expected to yield significant operational benefits:

- Potential reduction in mean time to repair (MTTR): eliminating maintenance personnel's time spent consulting highly technical manuals.
- Increased effectiveness in identifying the correct machinery and/or system component: detailed extraction of product codes, machinery, and page references from the technical manual, simplifying procurement processes.
- Increased knowledge dissemination: The Neuratio model allows for standardized intervention quality, reducing misunderstandings and increasing the percentage of interventions that can be handled by junior technicians, while reducing the burden on senior maintenance engineers.

Despite this, some limitations may arise:

- Neuratio lacks the spatial understanding of the shop floor, which is necessary to fully diagnose industrial issues (e.g., vibration levels and noise levels of industrial motors). If the user's query is not extremely detailed, the proposed solutions may be limited.
- Updating technical manuals or making changes to machinery, if not properly updated in Neuratio's databases, may result in inaccurate or outdated responses.
- Neuratio's impact on industrial machinery maintenance depends on the quality of the underlying documentation and the technicians' skills.

6. Conclusion

The results obtained have highlighted how Neuratio provides a significantly improved operational interface to a generic LLM. By integrating the RAG framework, we are able to reduce the risk of hallucinations in generic AI models. Neuratio's architecture represents an industrial AI solution that, by ensuring high factual accuracy and relevance, lays the operational groundwork for reducing maintenance times and standardizing support processes for maintenance technicians. Thus, future scientific research will focus on evolving the framework to reduce the dimensional gap with respect to the shop floor, through a multimodal RAG architecture that integrates audio and video processing capabilities, using specially designed sensors and digital twins, to bridge the gap between user prompts and physical dimensions.
Finally, the possibility of integrating Neuratio with ERP and CMMS systems will be evaluated to verify updates to manuals and machinery against historical data.

6. References

[1] A. O. Adesina, D. N. Ocholla, and A. Adesina, “South African Journal of Information Management,” 2024, doi: 10.4102/sajim.

[2] J. Oranga, “Tacit Knowledge Transfer and Sharing: Characteristics and Benefits of Tacit & Explicit Knowledge.” [Online]. Available: https://jaruda.org

[3] G. Ni, Z. Zhang, Z. Yuan, H. Huang, N. Xu, and Y. Deng, “Transformation paths and influencing factors of tacit knowledge into explicit knowledge in real estate companies: a qualitative study,” *Engineering, Construction and Architectural Management*, vol. 29, no. 3, pp. 1319–1342, Mar. 2022, doi: 10.1108/ECAM-09-2020-0677.

[4] M. Kamps, M. van den Boomen, J. van den Bogaard, and M. Hertogh, “Intergenerational transfer of engineering expertise: knowledge continuity management in storm surge barrier engineering,” *Built Environment Project and Asset Management*, vol. 14, no. 6, pp. 874–891, Oct. 2024, doi: 10.1108/BEPAM-10-2023-0179.

[5] S. R. Pillai and D. D. Iyer, “Knowledge Retention Strategies in the Face of Workforce Turnover,” *International Academic Journal of Innovative Research*, vol. 11, no. 4, pp. 21–26, Dec. 2024, doi: 10.71086/iajir/v11i4/iajir1128.

[6] H. Shin, L. Rothrock, and V. Prabhu, “Evaluating Technicians’ Workload and Performance in Diagnosis for Corrective Maintenance,” *Sensors*, vol. 24, no. 6, Mar. 2024, doi: 10.3390/s24061943.

[7] A. Nikitin and S. Kaski, “Human-in-the-Loop Large-Scale Predictive Maintenance of Workstations,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2022, pp. 3682–3690. doi: 10.1145/3534678.3539196.

[8] A. Brintrup, G. Baryannis, A. Tiwari, S. Ratchev, G. Martínez-Arellano, and J. Singh, “Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions,” 2023.

[9] A. Sottana, B. Liang, K. Zou, Z. Yuan, and N. Ai, “Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks.” [Online].

[10] J. Li *et al.*, “The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.03205

[11] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2305.11747

[12] I. Rojek, M. Jasiulewicz-Kaczmarek, M. Piechowski, and D. Mikołajewski, “An Artificial Intelligence Approach for Improving Maintenance to Supervise Machine Failures and Support Their Repair,” *Applied Sciences (Switzerland)*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084971.

[13] J. Wu *et al.*, “Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation,” Oct. 2024, [Online]. Available: http://arxiv.org/abs/2408.04187

[14] A. Dhami and L. Goel, “Optimizing retrieval augmented generation pipelines for domain specific applications,” *International Journal of Research in Modern Engineering & Emerging Technology*, vol. 13, no. 3, pp. 55–72, Mar. 2025, doi: 10.63345/ijrmeet.org.v13.i3.4.

[15] Y. Yang, Y. Tang, and K. Y. Tam, “InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning,” Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.13064

[16] R. Friel, M. Belyi, and A. Sanyal, “RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems,” Jan. 2025, [Online]. Available: http://arxiv.org/abs/2407.11005

[17] D. Ru *et al.*, “RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation,” Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.08067

[18] U. Patel *et al.*, “THELMA: Task Based Holistic Evaluation of Large Language Model Applications-RAG Question Answering,” Jun. 2025, [Online]. Available: http://arxiv.org/abs/2505.11626

[19] T. Vu *et al.*, “Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation,” 1710. [Online].

[20] M. Reslan, E. M. Hastings, M. P. Brundage, and T. Sexton, “A Data-Driven Framework for Team Formation for Maintenance Tasks.”

[21] S. Durst and H. Aisenberg Ferenhof, “Knowledge Leakages and Ways to Reduce Them in Small and Medium-Sized Enterprises (SMEs),” *Information (Switzerland)*, vol. 5, no. 3, pp. 440–450, Sep. 2014, doi: 10.3390/info5030440.

[22] S. R. Pillai and D. D. Iyer, “Knowledge Retention Strategies in the Face of Workforce Turnover,” *International Academic Journal of Innovative Research*, vol. 11, no. 4, pp. 21–26, Dec. 2024, doi: 10.71086/iajir/v11i4/iajir1128.

[23] V. A. Alexandru *et al.*, “Knowledge management approaches of small and medium-sized firms: a cluster analysis,” *Kybernetes*, vol. 49, no. 1, pp. 73–87, Jan. 2020, doi: 10.1108/K-03-2019-0211.

[24] J. I. Igartua-López, J. Retegi-Albisua, M. Ruiz, and M. Mindegia, “Evaluation of a strategy-oriented method to identify and prioritise knowledge management initiatives in smes,” *Journal of Industrial Engineering and Management*, vol. 14, no. 1, pp. 3–14, 2021, doi: 10.3926/jiem.3277.

[25] L. O. Iheukwumere-Esotu and A. Y. Kaltungo, “Assessment of barriers to knowledge and experience transfer in major maintenance activities,” *Energies (Basel).*, vol. 13, no. 7, 2020, doi: 10.3390/en13071721.

[26] L. O. Iheukwumere-Esotu and A. Yunusa-Kaltungo, “Knowledge criticality assessment and codification framework for major maintenance activities: A case study of cement rotary kiln plant,” *Sustainability (Switzerland)*, vol. 13, no. 9, May 2021, doi: 10.3390/su13094619.

[27] T. Yıldız, B. Balkan Akan, Ü. Sığrı, and M. Dabić, “Impact of knowledge-sharing culture on organizational creativity: integrating explicit and tacit knowledge sharing as mediators,” *Journal of Knowledge Management*, vol. 29, no. 4, pp. 1248–1277, Mar. 2025, doi: 10.1108/JKM-05-2024-0633.

[28] J. Cárcel-Carrasco, J. A. Cárcel-Carrasco, and E. Peñalvo-López, “Factors in the relationship between maintenance engineering and knowledge management,” *Applied Sciences (Switzerland)*, vol. 10, no. 8, Apr. 2020, doi: 10.3390/APP10082810.

[29] T. Franzmeyer *et al.*, “High Accuracy, Less Talk (HALT): Reliable LLMs through Capability-Aligned Finetuning,” Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.04051

[30] M. Ahmad, I. Yaramic, and T. D. Roy, “Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI,” Oct. 26, 2023. doi: 10.20944/preprints202310.1662.v1.

[31] C. Jeong, S. Lee, S. Jeong, and S. Kim, “A Study on the Framework for Evaluating the Ethics and Trustworthiness of Generative AI.”

[32] K. Alshehri and M. Sayed Mouchaweh, “Generative AI for the Power Sector: Methodology, Safety, Trustworthiness, and Policy Implications,” Jul. 2025. doi: 10.30573/KS--2025-DP35.

[33] Z. Song *et al.*, “Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey,” Aug. 2025, [Online]. Available: http://arxiv.org/abs/2502.10708

[34] A. Vinayan Kozhipuram, S. Shailendra, and R. Kadel, “Retrieval-Augmented Generation vs. Baseline LLMs: A Multi-Metric Evaluation for Knowledge-Intensive Content,” *Information (Switzerland)*, vol. 16, no. 9, Sep. 2025, doi: 10.3390/info16090766.

[35] C. Pei *et al.*, “Flow-of-Action: SOP Enhanced LLM-Based Multi-Agent System for Root Cause Analysis,” in *WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025*, Association for Computing Machinery, Inc, May 2025, pp. 422–431. doi: 10.1145/3701716.3715225.

[36] “Applied Research on an Aircraft Maintenance Assistant Based on a Large Language Model”.

[37] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Apr. 2021, [Online]. Available: http://arxiv.org/abs/2005.11401

[38] D. Wampler, D. Nielson, and A. Seddighi, “Engineering the RAG Stack: A Comprehensive Review of the Architecture and Trust Frameworks for Retrieval-Augmented Generation Systems,” Nov. 2025, [Online]. Available: http://arxiv.org/abs/2601.05264

[39] M. Klesel and H. F. Wittmann, “Retrieval-Augmented Generation (RAG),” *Business and Information Systems Engineering*, vol. 67, no. 4, pp. 551–561, Aug. 2025, doi: 10.1007/s12599-025-00945-3.

[40] C. Sharma, “Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers,” May 2025, [Online]. Available: http://arxiv.org/abs/2506.00054

[41] A. Jadon, A. Patil, and S. Kumar, “Enhancing Domain-Specific Retrieval-Augmented Generation: Synthetic Data Generation and Evaluation using Reasoning Models,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.15854

[42] R. Friel, M. Belyi, and A. Sanyal, “RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems,” Jan. 2025, [Online]. Available: http://arxiv.org/abs/2407.11005

[43] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment,” May 2023, [Online]. Available: http://arxiv.org/abs/2303.16634

[44] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, “ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2311.09476

[45] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, and E. Gradients, “RAGAS: Automated Evaluation of Retrieval Augmented Generation.”