https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Retrieval-Augmented Large Language Model for Wind Turbine Maintenance Work-Order Classification

Yoke W. Tan[1], James Carroll[1]

[1] Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XQ, UK

Correspondence to: Yoke Wang Tan (tan-yoke-wang@strath.ac.uk)

Abstract. Systematic analysis of operation and maintenance (O&M) data may play a significant role in reducing the cost of energy (CoE) of wind power. However, the maintenance work order text data remained comparatively underexplored due to the linguistic variability and complexity of unstructured text. While the applications of Large Language Models (LLMs) to automate manual maintenance-text analysis are emerging, smaller open-source models generally exhibited lower performance than larger proprietary models. In this study, a framework incorporating retrieval-augmented generation (RAG) is proposed to reduce the performance gap between small scale open-source LLMs and larger proprietary LLMs in wind turbine maintenance-text classification. Using a pre-labelled dataset, the proposed framework incorporates a compact, label-specific knowledge corpus containing class definitions, relevant terminology, and label boundaries to provide task-specific contextual information during inference. The findings show that the RAG-enhanced open-source model Qwen3.5-9B achieved a macro-F1 of 0.72, exceeding the 0.66 achieved by the baseline proprietary model GPT-5.6-luna, while remaining competitive with GPT-5.6-luna when both models were evaluated under the RAG configuration. Large inter-model variations observed in class-level F1-scores were also reduced for most label classes following the application of RAG. Despite the additional input-token requirements introduced by RAG, only marginal trade-off in inference throughputs was observed among the locally deployed models. These findings indicate that task-specific retrieval using a compact, label-structured knowledge corpus can improve the competitiveness of smaller open-source LLMs for wind-turbine maintenance work-order classification, although further evaluation across additional and larger maintenance datasets, retrieval corpora and LLMs is required to establish the generalisability and scalability of the proposed framework.

1 Introduction

Wind turbine (WT) operation and maintenance (O&M) costs remain a major challenge in the wind sector, accounting for up to 30% of the overall cost of energy (CoE) of wind power (Marzouk, 2026). Numerous studies have therefore exploited information-rich O&M data, including supervisory control and data acquisition (SCADA) measurements and maintenance records to assess turbine condition and reliability aimed at reducing WT downtime (Dao et al., 2025; Mucchielli et al., 2021; Turnbull et al., 2021). While structured numerical SCADA data have been widely investigated in WT fault studies due to their compatibility with computational analysis, unstructured maintenance data, which is predominantly recorded as free text remain comparatively underexplored in wind O&M literature. Unlike SCADA data which tracks the operational measurements of the turbines, maintenance data typically comprise maintenance logs and reports documenting maintenance work performed throughout a turbine's operational life. These records, commonly referred to as work-orders (WO) often contain typographical errors, abbreviations, and incomplete descriptions (Salo and Mcmillan, 2019; Sexton et al., 2019). Such linguistic variability and inconsistency are further compounded by the differences in data-recording practices, terminology and documentation standards across WT operators (Lutz et al., 2023). Consequently, maintenance-text analysis often relies on long and labour-intensive manual processing, which substantially reduce the timeliness and operational value of the information obtained (Salo and Mcmillan, 2019). Despite these limitations, maintenance records provide valuable

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

chronological evidence of component-level interventions and failure-related events when analysed systematically (Navinchandran et al., 2022).

The emergence of Large Language Model (LLM) in recent years has substantially expanded the capabilities of NLP in handling linguistic variability and contextual interpretation, providing a potential solution to the labour-intensive process outline in the previous paragraph. In contrast to conventional ML approaches, LLMs are pretrained on diverse corpora and do not rely on manually engineered features and labelled training data. This enables them to capture semantic relationships and contextual information across sentences, supporting the complex language task such as text classification and information extraction in both zero-shot and few-shot settings. Despite their capabilities in NLP, LLMs may exhibit reduced performance in specialised domain application due to the underrepresentation of domain-specific terminology in their pre-training data (Shen et al., 2024).

Recent studies have adapted general purpose LLM by updating model parameters using proprietary domain knowledge (J et al., 2024; Lin et al., 2025). Although this approach can improve task-specific performance, its computational requirements, associated costs, and dependence on sufficiently large, quality labelled datasets may limit its practical applicability (Lutz et al., 2023). To reduce the computational and data requirements associated with fine-tuning, model-preserving adaptation methods such as prompt engineering and retrieval-augmented generation (RAG) have been explored for adapting general-purpose LLMs. Prompt engineering provides structured instructions while contextual information retrieved through RAG help guide an LLM towards responses within a defined scope. In (Ai et al., 2026), a RAG pipeline was proposed for a knowledge-to-action question answering task in aviation safety O&M. This study demonstrated that open-source LLMs can be effectively adapted to domain-specific applications when coupled with RAG.

The increase in application of LLMs across specialised technical domains has also extended to wind-energy research, where recent studies have started to investigate their use for WT maintenance work-order text analysis. One of the pioneering studies in LLM adaptation work in this domain, (Fan et al., 2026), investigated the performance of fine-tuned small-scale model for the generation of optimal maintenance strategies for wind farms. Although the model achieved strong performance under static-input conditions, its performance under dynamic scenarios was skewed towards cases resembling its training data, indicating the potential overfitting and reduced model generalisability. As an alternative to fine-tuning to address quality training data scarcity, (Tian et al., 2026) investigated the use of RAG in wind power system fault detection. The study presented a fault-causality analysis and the proposal of maintenance strategies support, which is supported by external knowledge retrieved through RAG. However, the industrial applicability of the framework remains limited because the framework was evaluated exclusively on synthetic fault cases derived from a reference textbook rather than real-world industrial data.

While the preceding studies demonstrated the value of LLM for domain adaptation in wind-energy applications, (Malyi et al., 2026) examined the more directly relevant task of WT maintenance-text classification using off-the-shelf LLMs. The study compared API-hosted proprietary models with locally hosted open-source models on a high-cardinality classification task defined by turbine-component labels. Although the larger proprietary models achieved stronger overall performance, the comparatively small gap observed for several open-source models highlighted their potential as more accessible alternatives. The study also introduced reliability-enhancing measures, including constrained output formatting, dynamic label-space definition, and model-generated confidence scores, which reduced invalid or hallucinated responses across the evaluated models. Despite establishing a strong benchmark for LLM-based WT maintenance-text-classification, the evaluated dataset was filtered and pre-processed to remove textual noise, resulting in substantially reduced dataset relative to the raw maintenance records. This leaves the scalability of the framework to larger and less curated maintenance dataset insufficiently explored. In addition, the comparatively small performance gap between open-source models and proprietary models raises the question of whether task-specific adaptation with RAG can further improve smaller locally hosted models.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Building on these findings and methodological direction, the contribution and novelty of the present study is to investigate whether the adaptation of retrieval augmented generation (RAG) can provide task-specific contextual information during inference to improve classification performance and further narrow the performance gap between small-scale open-source LLMs and large proprietary models in WT maintenance-text classification.

2. Methodology

An LLM-based analytical workflow is proposed in this study to analyse a labelled real-world maintenance WO dataset. Using prompt engineering in combination with RAG technique, a structured prompt incorporating additional label-rule context and definitions was constructed for each inference iteration. This enabled locally deployed small-scale LLMs and API-accessed proprietary models to classify each WO text into a predefined label space comprising 18 component categories. The design and contribution of each stage of the proposed workflow are described in detail in the following sections.

2.1 Dataset and Pre-processing

A real-world maintenance work-order dataset comprised of approximately 100,000 entries of maintenance records is used for this study. Each entry includes turbine ID, turbine operational year, WO ID, total WO cost, cost category, WO text, long WO text and failure category assigned manually by domain experts. The original dataset consisted of 100,000 entries which combined work order text records and material usage database. As a result, a proportion of the original dataset were records documenting different material usage for the same work order.

Duplicated entries sharing WO ID and WO text were removed to reduce redundant repetitions that would not provide additional information for evaluating the text-classification performance. The two features selected for this study – the failure category and WO text were subsequently pre-processed to ensure consistent data quality across entries. Although the long WO text field was intended to provide additional details for each record, most entries were either empty or a repetition of the corresponding WO text. This field was therefore excluded from the analysis to reduce textual redundancies.

Following rule-based cleaning adopted in industrial maintenance-text analysis (Salo and Mcmillan, 2019) , inconsistencies in the WO text were further reduced by correcting typographical errors and abbreviations and removing whitespace, symbols, and numbers that were not part of known recognised component identifiers. The final resulting dataset contains 6016 unique entries of maintenance WO text and it's corresponding manually labelled failure category.

Table 1: Examples of Work Order Text and Label

	Work-Order Text	Assigned Label
1.	Generator fan replacement	Generator
2.	Defect blade bearing lube	Blades
3.	No communication to turbine	Control system
4.	feedback error vent fans	Heaters and coolers
5.	rotating hydraulic union oil leak	Pitch and hydraulic

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

2.2 Task definition and Case-study Design

Maintenance-text classification can have different task formulations depending on the information that users intend to extract from the maintenance records. Previous studies have shown that classification objectives may vary according to the target categories and levels of classification granularity adopted for a given dataset (Lutz et al., 2023). Given that the LLM performance in this study is evaluated against an existing labelled dataset, the task definition should therefore reflect the objective represented by the assigned labels.

The dataset used in this paper was previously labelled by domain experts through the labour-intensive manual process involving identification of matching keywords and interpretation of the maintenance work described in the work-order text (Carroll et al., 2016). Although keyword matching was part of the original labelling process and reflects the rule-based approaches commonly used in prior data mining efforts (Salo and Mcmillan, 2019), the final category was not determined solely by the presence of individual terms. Instead, the assigned label reflected the interpreted maintenance context and the principal wind-turbine component affected by each maintenance activity. Each WO text was therefore assigned to the highest-level affected component rather than to the subcomponents explicitly mentioned in the text. However, subcomponent labels were also included in the classification scheme and were assigned only when the principal affected component could not be inferred from the maintenance context. For example, data entry no.1 in Table 1, generator fan replacement, was assigned to the generator category rather than the heaters and coolers because the maintenance activity was attributed to the generator as the principal affected component.

With these elements considered, the objective of the evaluated models in this study was to assign the most appropriate predefined label based on the principal affected component and the hierarchical labelling criteria described in each maintenance WO text. The list of predefined labels and their corresponding label context is attached in Appendix A – Table A1.

2.3 Prompt Engineering and LLM Inference Framework

LLM inference in this study was performed independently for each pre-processed WO text, with the models were tasked with assigning the input record to one of the predefined labels. For each inference instance, a standardised prompt template was used to incorporate the WO text and retrieved contextual information. The prompt contained engineered classification instructions designed to guide the LLM towards consistent interpretation of unstructured maintenance text. A predefined label space and an additional output instruction were also included at the end of the prompt, restricting the LLMs to generating only predefined labels as their responses. The maximum output token was limited according to the longest predefined label to reduce generation of invalid responses.

As the response behaviour may vary across the LLMs for the same input, adherence to the output instruction may also differ between models (Qin et al., 2024). Although output instruction was included in the prompts, hallucinated labels and invalid responses remained observable during the experiments. An optional standardised output schema was therefore implemented to further constrain the structure of the generated response. When enabled, the models were restricted to generating responses in JSON format, thereby further reducing invalid responses during inference. An additional output-processing layer was also applied regardless of output-schema implementation to systematically extract and validate predicted labels. Invalid outputs, including parsing failures, malformed JSON and hallucinated labels, were reassigned to *"Review_Required"* label to flag the for subsequent review. The compatibility of this processing layer was maintained across different model configurations. Notably, all raw LLM outputs were preserved and compiled into the same output file, allowing users to inspect the responses associated with invalid predictions. The inference parameters and model configurations used in this study are summarised in Appendix A – Table A2.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

2.4 Retrieval Augmented Generation (RAG)

Retrieval-augmented generation is a retrieval-based approach used to improve LLM response generation without requiring additional parametric fine-tuning (Tural et al., 2024). In this study, the RAG component was designed to supplement the LLM input with additional label-specific contextual information retrieved from a structured knowledge corpus prior to classification. To reflect the classification objective and labelling rules defined in Section 2.2, the retrieval corpus was structured by label, with each entry detailing the corresponding label definition, relevant keywords and label boundaries, as illustrated in Appendix A – Table A1.

For each WO text, relevant contextual information was retrieved by measuring its similarity to the entries in the knowledge corpus using cosine similarity (Han et al., 2012) presented in Eq. (1). As the corpus contained explicit label-specific terminology and keywords, Term Frequency–Inverse Document Frequency (TF-IDF) was adopted as the primary text-representation approach. TF-IDF primarily captures lexical similarity between the work-order text and corpus entries, whereas contextual embedding models can represent broader semantic relationships between texts. To assess whether the choice of retrieval representation influenced downstream classification performance, a limited sensitivity comparison was conducted using TF-IDF and two contextual embedding models, all-MiniLM-L6-v2 and bge-base-en-v1.5 (Xiao et al., 2023). All other experimental conditions, including the retrieval corpus and inference configurations, were kept consistent for this comparison.

$$Cosine\ similarity\ (A, B) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

The proposed RAG component was designed to be configurable, allowing the number of retrieved contexts to be adjusted according to the context-window limitations of each LLM. For the present study, the number of retrieved contexts, k, was fixed at $k = 3$ to maintain consistent input-token requirements across the evaluated models, particularly given the more limited context windows of the locally deployed models. The same retrieval corpus and method were applied across all evaluated models, except where the retrieval representation was varied for the sensitivity analysis.

2.5 Model Selection and Computational Setup

To provide a proprietary-model baseline, a recent GPT model – GPT-5.6-luna available through the OpenAI API at the time of experimentation was selected as the proprietary reference model. Notably, access to OpenAI API models is subject to platform-specific rate limits that constrain the number of requests or tokens that can be processed within a specified period (GPT-5.6 Luna Model | OpenAI API, 2026) . Given the size of dataset, classification of individual records through iterations using standard API requests was relatively inefficient. Therefore, the OpenAI Batch API (Batch API | OpenAI API, 2026) was used to process the dataset in batches, enabling large-scale inference under the service constraints applicable to batch processing.

In contrast, the open-source LLMs were selected based on their ability to run efficiently through Ollama (Ollama documentation - Ollama, 2026) on the available hardware. All locally hosted models were executed on a workstation equipped with an Intel Core Ultra 9 185H processor (2.30 GHz), 32 GB of system memory, and an NVIDIA GeForce RTX 4070 Laptop GPU with 8 GB of dedicated VRAM. The selected models are listed below:

Proprietary LLM: GPT-5.6-luna

Open-Source LLM: Mistral-7B, Qwen3.5-9B, Llama3-8B, Gemma3-4B

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

2.6 Evaluation Metrics

The performance of the selected models was evaluated using standard classification metrics, with the labelled dataset serving as the ground truth for evaluation. The investigated dataset exhibited an imbalanced distribution; however, each predefined label class was considered equally important within the classification objective. Therefore, accuracy and macro-averaged precision, recall and F1-score were used to assess the overall classification performance, with particular emphasis on the macro-averaged metrics to account for performance across all classes. Following the observation of invalid or hallucinated outputs observed across the evaluated models, the hallucination rate was also reported as an additional measure of response validity. Invalid outputs were excluded from the calculations of classification metrics exclude invalid labels as they did not correspond to any label within the predefined classification space. Notably, *True Positive, True Negative, False Positive, False Negative* were denoted in the equations by *TP, TN, FP* and *FN,* respectively, and the subscript *label* denotes the individual label classes. The evaluation metrics (Grandini et al., 2020) are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP_{label}}{TP_{label}+FP_{label}} \quad (3)$$

$$Recall = \frac{TP_{label}}{TP_{label}+FN_{label}} \quad (4)$$

$$class\text{-}level\ F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (5)$$

$$Macro\ Average\ Precision = \frac{\sum_{k=1}^{K} Precision_{label}}{K} \quad (6)$$

$$Macro\ Average\ Recall = \frac{\sum_{k=1}^{K} Recall_{label}}{K} \quad (7)$$

$$Macro\ F1 = 2 * \left(\frac{Macro\ Average\ Precision*Macro\ Average\ Recall}{Macro\ Average\ Precision^{-1}+Macr\ \ Average\ Recal\ ^{-1}}\right) \quad (8)$$

$$Hallucination\ Rate = \frac{\sum Invalid\ Output}{\sum Output} \text{ x } 100\% \quad (9)$$

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3. Results

The classification performance of the LLMs was evaluated using the metrics defined in Section 2.6. Given that the task defined in Section 2.2 requires reliable classification across all predefined labels, the overall evaluation should not be dominated by the majority classes within the imbalanced dataset. Accordingly, model performance is presented with primary emphasis on macro-F1 Score, supplemented by class-level F1-Score and accuracy when applicable. The results are presented across different experimental configurations to examine the effects of model choice, structured output constraints, retrieval-augmented generation and retrieval representation on classification performance.

3.1 Overall Classification Performance

A summary of the model performances is presented in Table 3 and 4. The results presented in this section aims to provide an overview of the baseline for each model, where the classifications were performed without the RAG component.

Table 3: Base Prompt Classification Results without Formatted-Output Constraints

Model	Hallucination rate (%)	Accuracy	Macro Average		
			Precision	Recall	F1-Score
Mistral-7B	66.10	0.63	0.57	0.62	0.54
Qwen3.5-9B	0.80	0.67	0.68	0.69	0.67
Llama3-8B	25.10	0.57	0.60	0.59	0.53
Gemma3-4B	5.10	0.51	0.66	0.51	0.50
GPT-3.5-luna	0.10	0.67	0.64	0.76	0.65

Table 4: Base Prompt Classification Results with Formatted-Output Constraints (Baseline)

Model	Hallucination rate (%)	Accuracy	Macro Average		
			Precision	Recall	F1-Score
Mistral-7B	0.00	0.50	0.57	0.60	0.53
Qwen3.5-9B	0.00	0.68	0.69	0.67	0.65
Llama3-8B	0.00	0.58	0.65	0.57	0.57
Gemma3-4B	0.00	0.49	0.51	0.60	0.49
GPT-3.5-luna	0.00	0.68	0.64	0.77	0.66

At base prompt configuration without formatted-output constraints, Qwen-3.5 was the only open-source model to achieved comparable performance to GPT-5.6-luna, with accuracy and macro F1-score of 0.68 and 0.67, respectively. While all evaluated models produced some degree of invalid or hallucinated responses, GPT-5.6-luna exhibited the lowest hallucination rate, and Mistral-7B exhibited the highest hallucination rate at 66.1%. Because invalid responses were excluded from the calculation of the classification metrics, models with higher hallucination rates were evaluated on substantially fewer valid predictions. The resulting classification scores should therefore be interpreted alongside the corresponding hallucination rates.

When a JSON-format constraint was applied to the generated output, no invalid response was observed across all models. Under this configuration, all open-source models besides Llama3-8B showed reduced macro-F1, whereas GPT-5.6-luna showed marginal improvement in overall classification performance. While GPT-5.6-luna achieved the strongest

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

classification result across all models, Qwen-3.5-9B remained the strongest-performing open-source model and achieved only marginally weaker result than the proprietary model. Overall, under Base Prompt configuration, the proprietary GPT-5.6-luna model demonstrated superior classification performance than the evaluated open-source models, whereas Gemma-4B exhibited the weakest overall classification performance. Notably, Qwen3.5-9B achieved comparable results to GPT-5.6-luna despite only under Base Prompt configuration.

Given the effectiveness of formatted-output constraint in reducing invalid responses, only results generated under the formatted-output configuration are presented in the subsequent sections to ensure consistent comparison across the evaluated models. The formatted-output results generated under Base Prompt configuration are therefore considered as the baseline performance for all subsequent analyses.

3.2 RAG-enhanced Classification Performance

3.2.1 Performance Improvement for Open-source Models

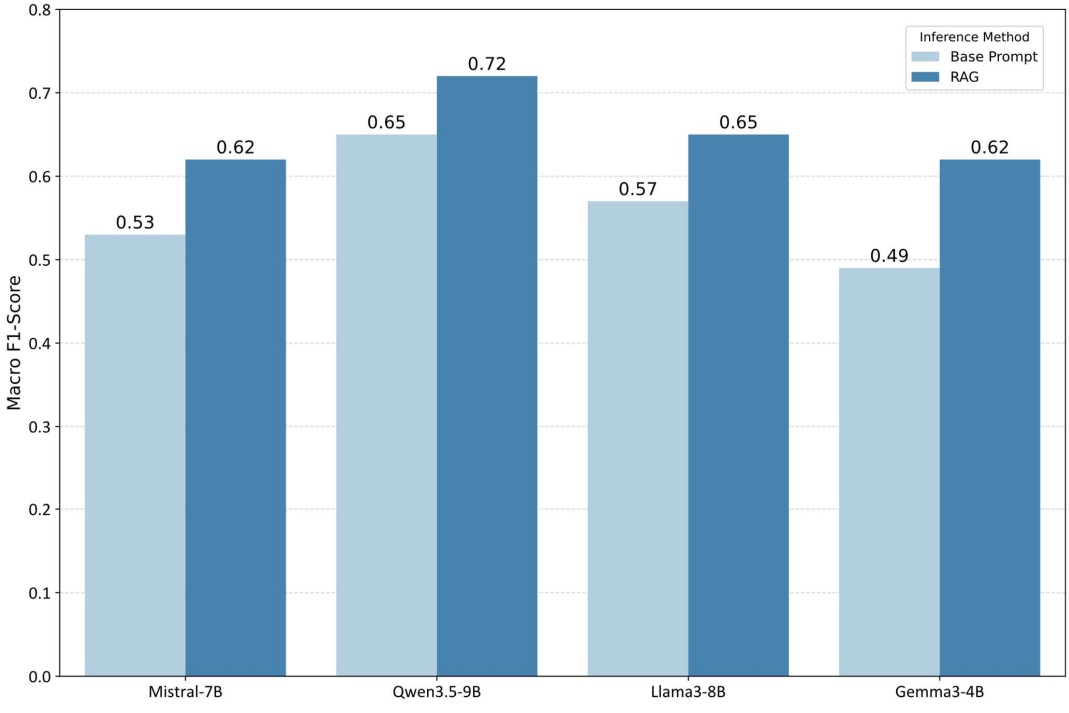


Figure 1: Performance Comparison between Configurations for Open-source Models

The open-source models achieved considerable improvement in classification performance following the implementation of RAG. The largest absolute increase was observed for Gemma3-4B at 0.13, whereas Qwen3.5-9B showed the smallest improvement at 0.07. Despite the larger performance gains observed for Mistral-7B, Llama3-8B and Gemma3-4B, their highest RAG-enhanced macro F1-score only reached 0.65, equivalent to the baseline performance of Qwen3.5-9B. Qwen3.5-9B remained the best performing open-source model, achieving the highest macro-F1 score of 0.72 under the RAG configuration. Additional illustrations for the evaluated metric results comparison between Base Prompt and RAG configuration for each open-source model were included in Appendix B – Figure B1.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3.2.2 Best Open-source Model vs Proprietary Model Baseline

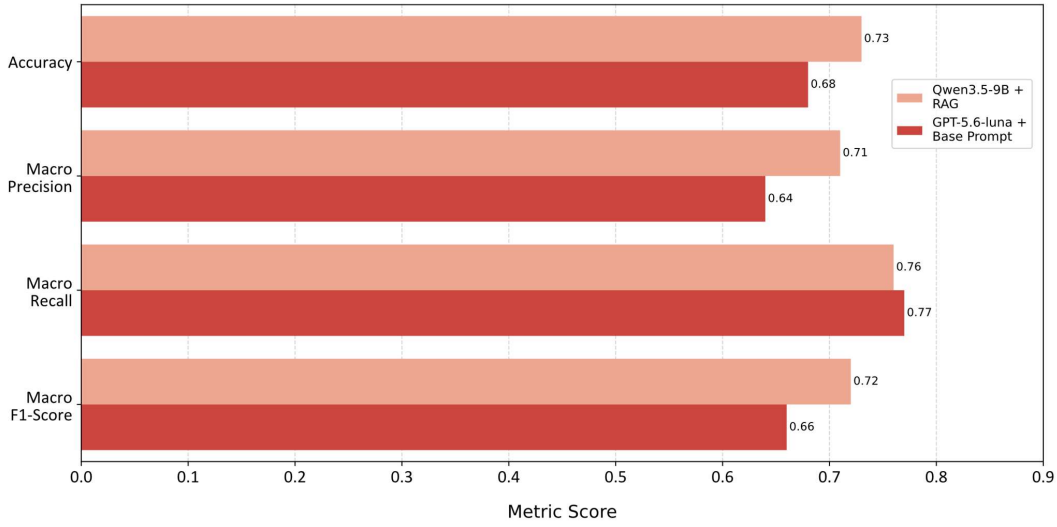


Figure 2: Performance Comparison between Qwen3.5-9B+RAG and Baseline GPT-5.6-luna

Figure 2 presents a direct comparison for the overall performance between Qwen3.5-9B under the RAG configuration and GPT-5.6-luna under the Base Prompt configuration. Across the four evaluated metrics, Qwen3.5-9B achieved higher performance in three. The greatest difference was observed for macro precision, where Qwen3.5-9B exceeded GPT-5.6-luna by 0.07, followed by macro-F1 and accuracy with differences of 0.06 and 0.05, respectively. However, the macro recall of Qwen3.5-9B remained marginally weaker than GPT-5.6-luna by 0.01 even with RAG.

3.2.3 Performance of Proprietary Model with RAG configuration

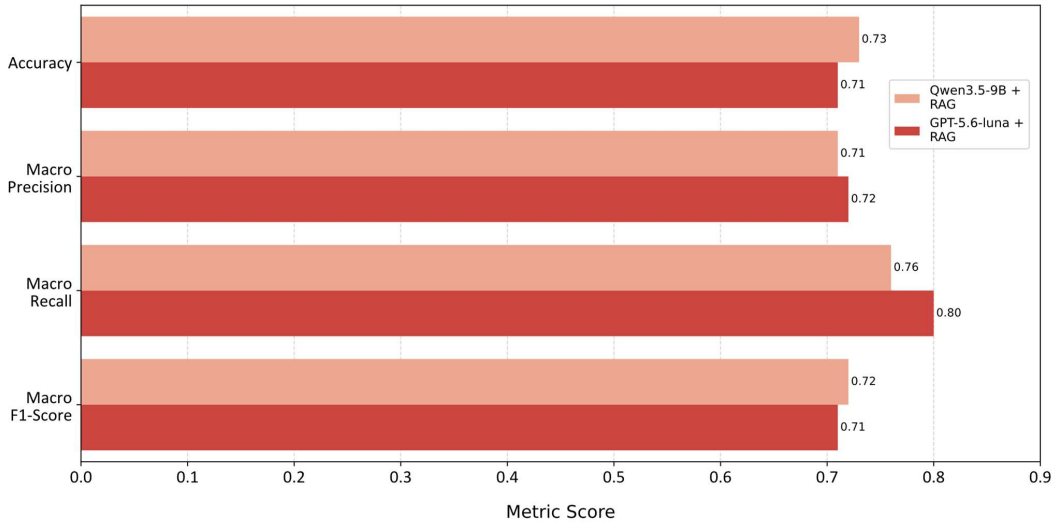


Figure 3: Performance Comparison between Qwen3.5-9B and GPT-5.6-luna under RAG configuration

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

To determine whether the observed performance advantage with retrieval augmentation was retained when the same approach was applied to the proprietary model, Qwen3.5-9B and GPT-5.6-luna were compared under the RAG configuration, illustrated in Fig. 3. Following RAG application, GPT-5.6-luna exhibited improvements in overall classification performance similar to Qwen3.5-9B. While both models achieved comparable results, Qwen3.5-9B retained a marginal advantage in accuracy and macro-F1 at 0.73 and 0.72, respectively. However, GPT-5.6-luna achieved marginally higher macro precision, and exceeding Qwen3.5-9B in macro recall 0.04.

3.3 Class-Level Classification Performance

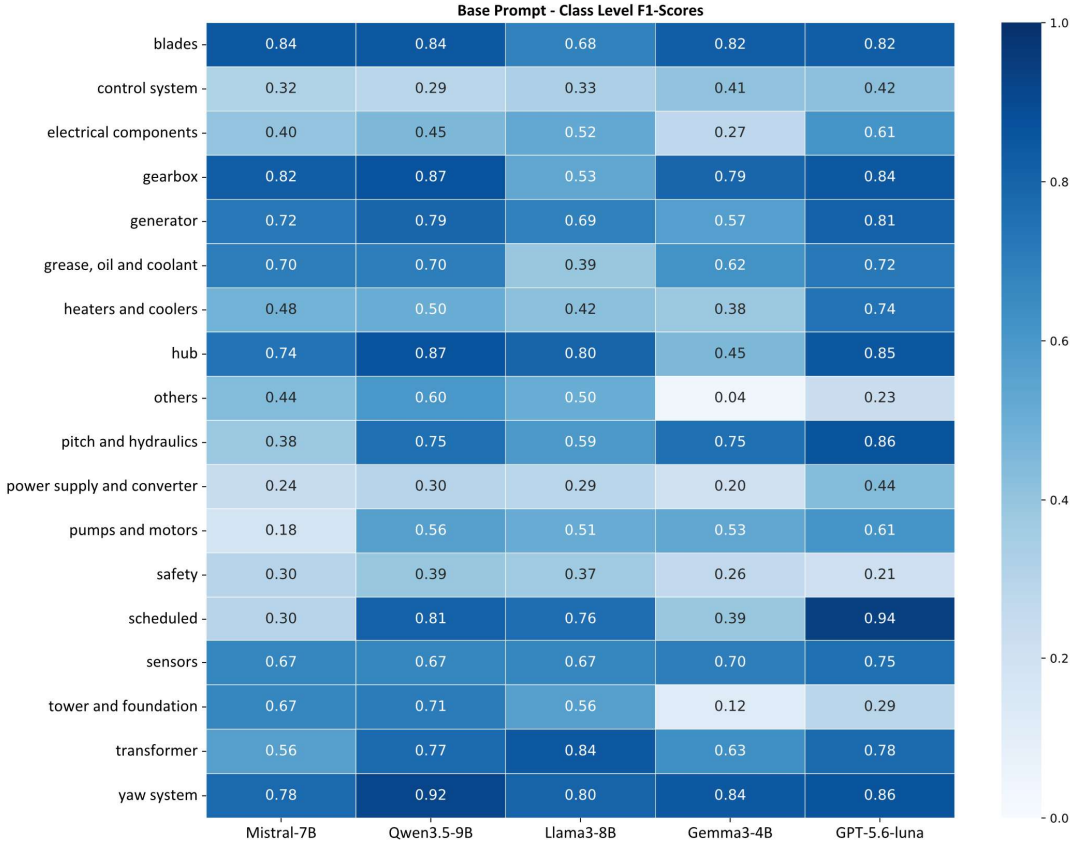


Figure 4: Baseline Class-Level F1 Scores

Despite using the same prompt template, considerable variation in F1-score was observed across both models and label classes. Among the 18 predefined label classes, *"yaw system"* exhibited consistently strongest F1-range across all evaluated models. Similarly, the *"blades"* and *"sensors*" class showed relatively consistent performance across models, although at *"sensors"* achieved lower-overall F1-scores. Large inter-model variations were observed for several classes, particularly "*scheduled"*, "*tower and foundation"*, "*others"*, "*pitch and hydraulics"*, and "*pumps and motors"*. Classes including *"gearbox"* and "*hub"* generally achieved comparatively high F1-scores, but their performance was reduced for individual models.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



WIND ENERGY SCIENCE DISCUSSIONS

Consistently weak classification performance was observed for *"safety"*, *"control system"* and *"power supply and converter"*, for which all models achieved F1-score equal or below 0.50 under Base Prompt configuration. Other classes including "*others"*, "*pitch and hydraulics"*, and "*pumps and motors"*, exhibited substantial model-dependent performance. Notably, Gemma3-4B exhibited the greatest class-to-class variability among the evaluated models, which was aligned with its overall performance.

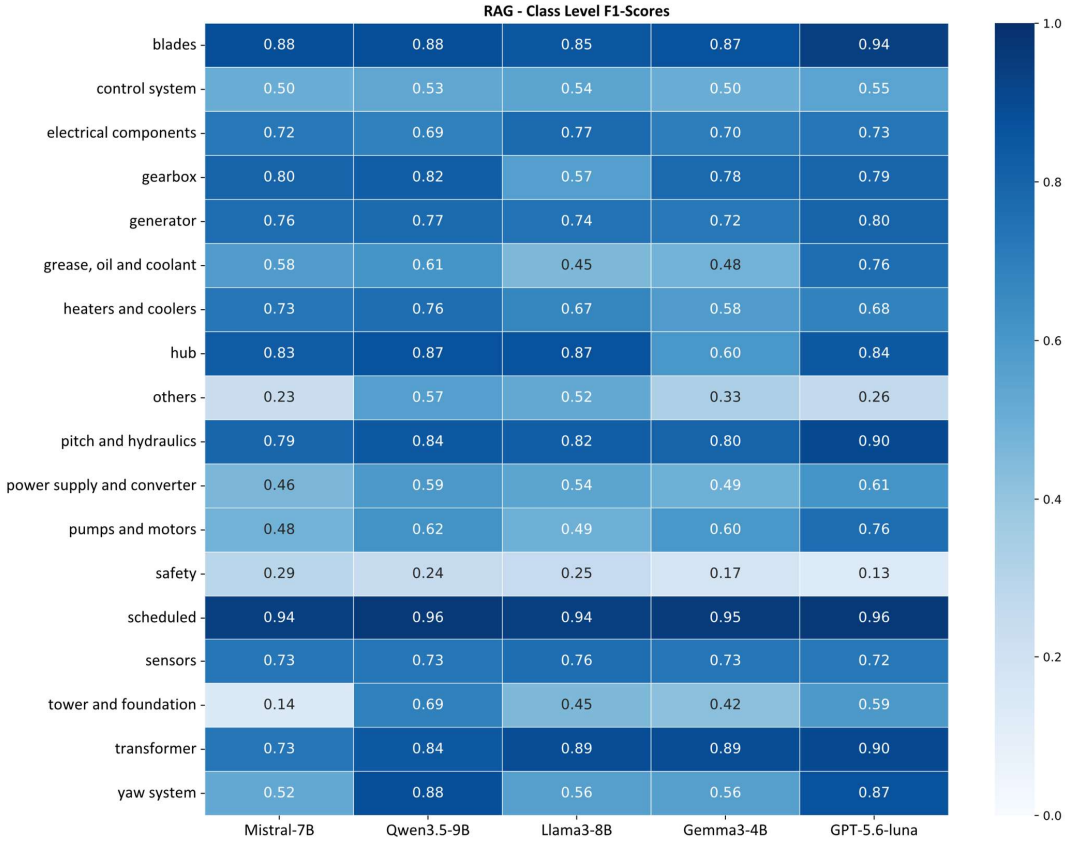


Figure 5: RAG-enhanced Class-Level F1 Scores

Following the application of RAG, inter-model variation was reduced for most label classes, regardless of their overall F1-scores. This effect was particularly pronounced for "scheduled", for which the inter-model F1-score range decreased from 0.64 under the Base Prompt configuration to 0.02 under RAG. Similar reductions were observed for "pitch and hydraulics" and "electrical components". Greater consistency was also observed for several other classes, including "*generator"*, "*control system"*, and "*sensors"*. However, "tower and foundation" remained highly model-dependent despite RAG, while "*yaw system"* was the only class for which model variation increased substantially.

The effect of RAG on class-level performance was not uniformly positive. Improvements were observed across many classes, but lower F1-scores were observed for *"safety"* and *"yaw system"* across all evaluated models following RAG. The response was particularly model-dependent for "*tower and foundation*", where its F1-score decreased substantially for Mistral-7B from 0.67 to 0.14.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3.4 LLM Inference Framework Efficiency

Table 5: Base Prompt without Formatted-Output Constraints

Models	∑ Input Tokens	∑ Output Tokens	Throughput (record/s)
Mistral-7B	1,065,883	34,934	5.88
Qwen3.5-9B	1,004,627	19,001	4.35
Llama3-8B	969,928	24,736	7.69
Gemma3-4B	955,901	25,868	7.69
GPT-5.6-luna	943,872	340,242	5.90

Table 6: Base Prompt with Formatted-Output Constraints

Models	∑ Input Tokens	∑ Output Tokens	Throughput (record/s)
Mistral-7B	1,065,883	80,522	3.13
Qwen3.5-9B	1,004,627	68,316	2.17
Llama3-8B	969,928	73,998	2.94
Gemma3-4B	955,901	75,050	2.78
GPT-5.6-luna	1,551,488	472,884	2.79

Table 7: RAG with Formatted-Output Constraints

Models	∑ Input Tokens	∑ Output Tokens	Throughput (record/s)
Mistral-7B	2,265,660	81,637	2.44
Qwen3.5-9B	2,060,418	70,763	1.72
Llama3-8B	2,019,703	71,463	2.56
Gemma3-4B	2,013,973	81,798	2.50
GPT-5.6-luna	2,600,668	372,510	4.01

Computational efficiency is an important consideration when selecting LLMs for large-scale classification. Given the available computational resources, the inference efficiency varied across models and experimental configuration, as presented in Table 5 - Table 7. As GPT-5.6-luna-inference was performed through the OpenAI Batch API, submitted batch files were completed within a 24-hours batch-processing window rather than providing directly comparable per-record inference times. Thus, the throughput values reported for GPT-5.6-luna should be interpreted as approximate estimates and not as direct measures of inference efficiency comparable with the locally deployed models.

Under the first configuration in Table 4, Mistral-7B recorded the highest input-token usage, whereas GPT-5.6-luna recorded the lowest. In contrast, GPT-5.6-luna produced the highest output-token usage despite its lower input-token count. Among the locally deployed models, Llama3-8B and Gemma3-4B achieved the highest throughput while processing comparatively fewer total tokens. Following the implementation of the formatted-output constraint, input-token usage remained unchanged across the open-source models, whereas GPT-5.6-luna showed a large increase. Output-token usage also increased across all models, with the Mistral-7B achieving the highest throughput among the locally deployed models under this configuration.

Under the RAG configuration with formatted-output constraints, input-token usage increased considerably across all models due to the additional retrieved context. Despite this increase, only marginal reductions in throughput were observed across the open-source models. GPT-5.6-luna exhibited a higher estimated throughput under the RAG configuration, reaching 4.01

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

records/s; however, this value should be interpreted in the context of the different Batch API timing mechanism described above. Among the open-source models, Llama3-8B and Gemma3-4B achieved marginally higher throughputs while Qwen3.5-9B exhibited the lowest throughput.

3.5 Sensitivity to Retrieval Representation

The choice of retrieval representation may influence the contextual information retrieved under RAG due to differences in how textual similarity is represented. This section compares the overall classification performance obtained using TF-IDF, all-MiniLM-L6-v2 and bge-base-en-v1.5 as retrieval representations. TF-IDF represents a sparse lexical retrieval approach, whereas MiniLM and BGE provide dense semantic retrieval representation. Differences in the retrieved contextual information were examined to assess their influence on classification outcomes under the present experimental setting. Results generated using Qwen3.5-9B were used for this comparison.

Table 8: Overall Classification Performance based on Retrievers

Model	Accuracy	Macro Average		
		Precision	Recall	F1-Score
TF-IDF	0.73	0.72	0.74	0.71
all-MiniLM-L6-v2	0.70	0.70	0.76	0.70
bge-base-en-v1.5	0.71	0.71	0.77	0.71

Only marginal differences in overall classification performance were observed across the three retrieval approaches. Under the present experimental setting, TF-IDF achieved the strongest overall classification performance, whereas both semantic embedding approaches achieved higher macro recall. The semantic retrieval approaches also required additional processing time due to additional model initialisation and context-encoding, which were not required for TF-IDF.

Table 9: Retrieval Example and Corresponding Classifications

Work Order Text	Retrieval Representations	Retrieved Context Ranks	Predicted Labels
Troubleshoot signal error position transducer	TF-IDF	1. sensors 2. Electrical components 3. safety	Sensors
	all-MiniLM-L6-v2	1. sensors 2. safety 3. control system	Sensors
	bge-base-en-v1.5	1. yaw system 2. sensors 3. safety	Sensors

Table 9 presents an ambiguous work-order text – *Troubleshoot signal error position transducer* with true label of sensors, together with the contextual labels retrieved by each method and the resulting predicted label. Although all three retrieval approaches successfully retrieved the context corresponding to the true label, differences were observed in the ranking and composition of the retrieved contextual entries. Despite these variations Qwen3.5-9B correctly classified the example under all three retrieval approaches. Overall, the choice of retrieval representation had only limited effects on the classification performance under the conditions evaluated in this study.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

4. Discussion

4.1 Key Findings

The key findings of this study indicate that the application of retrieval-augmented generation can substantially improve the performance of smaller open-source LLMs in wind-turbine maintenance text classification. Under the Base Prompt configuration with formatted-output constraints, the findings were consistent with the benchmark results reported in (Malyi et al., 2026), where the proprietary model achieved stronger classification performance and lower hallucination rates than the evaluated open-source models. Additionally, this study also revealed that smaller open-source model may achieve performance close to that of proprietary models. In particular, Qwen3.5-9B achieved only marginally lower macro F1-score than GPT-5.6-luna under Base Prompt configuration, highlighting the potential of smaller open-source models as candidates for domain-specific adaptation even under zero-shot condition.

Following the application of RAG, the classification performance improved across all evaluated open-source models, with Qwen3.5-9B exceeding the baseline performance of the proprietary model. The remaining open-source models also achieved performance within a similar range to the proprietary baseline. Under the evaluated experimental conditions, the results demonstrate that RAG can substantially reduce the performance gap between smaller open-source and larger proprietary LLMs in wind turbine maintenance text classification. Furthermore, the application of RAG to GPT-5.6-luna in Section 3.2.3 indicated the benefits of task and domain-specific contextual information may also improve the performance of off-the-shelf proprietary models in specialised classification tasks.

4.2 Implications for Model Selection and Knowledge Corpus Design

The baseline performance of all evaluated models presented in Section 3.1 highlights the limitations of both smaller open source and proprietary models in wind turbine maintenance text classification under zero-shot conditions. Without additional domain specific contextual information, the models showed varying degree of difficulty in aligning their predictions with the classification objective defined in Section 2.2. Despite the use of a consistent prompt template across the evaluated models, the weaker baseline performance of several open-source models reflected differences in their ability to interpret the classification instructions and apply the predefined labelling criteria. The observed inter-model variation may partly arise from intrinsic model characteristics, including differences in model architecture, model scale, pre-training, and post-training procedures, which can influence how individual models interpret the same maintenance text and classification instructions.

It might reasonably be expected that the additional contextual information provided through RAG would have a broadly comparable effect across the evaluated models. In contrast, the magnitude of improvement varied considerably between models. The results presented in Section 3.2.1 showed that models with weaker baseline performance generally exhibited larger gains following RAG implementation than models with stronger baseline performance. This pattern suggests that the task specific information retrieved from the knowledge corpus may provide greater benefit to models with more limited baseline capability for the classification task. Nevertheless, a more systematic evaluation across a wider range of models would be required to establish the relationship between baseline capability, intrinsic model characteristics and the effectiveness of retrieval augmentation.

A similar pattern was also observed at the class-level, where the retrieved contextual information benefited some predefined labels more than the others. While class-level classification performance improved across label classes following RAG, substantial decrease in F1-score was also observed for already strong performing class such as *"yaw system"*, indicating that retrieval augmentation did not uniformly improve every class. Persistently weak performance for certain classes such as *"safety"* and *"others"* may indicate residual ambiguity in the corresponding label definitions, overlap between classification categories, or insufficient contextual information to distinguish closely related labels. This variation suggests that the

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

effectiveness of the knowledge corpus may depend on how well the retrieved information represents the terminology and label boundaries associated with each class. Additional curation of the contextual information corresponding to persistently weaker or those showing performance reductions may therefore help improve class-level consistency and performance.

4.3 Implication of Retrieval Representation

Under the experimental conditions evaluated in this study, the representation sensitivity analysis in Section 3.5 showed only marginal differences in classification performance between sparse-lexical and dense semantic retrieval approaches. As illustrated by the example in Table 8, the downstream classification remained unchanged despite differences in the ranking of retrieved contextual information across the evaluated representation approaches. This finding suggests that the proposed framework was relatively robust to the choice of retrieval representation under the present experimental conditions. However, the relative effectiveness of each retrieval approach may depend on the structure of the knowledge corpus and the characteristics of the maintenance records being classified. Notably, the knowledge corpus used in this study was relatively small compared with larger documents such as web-based or proprietary databases.

The limited downstream differences observed across the evaluated representations may also reflect the structure of the knowledge corpus used in this study. As the corpus contained concise label definitions, explicit keywords and clearly defined label boundaries, relevant information could be identified effectively through both lexical and semantic similarities. Under these conditions, the simpler TF-IDF representation remained competitive with the contextual embedding approaches despite differences in retrieval ranking. This suggests that the complexity of the retrieval representation should be considered in relation to the structure and scale of the knowledge corpus rather than assuming that semantic embedding methods will necessarily provide superior downstream classification performance.

4.4 Computational and Practical Considerations

The trade-off between classification performance and inference efficiency was presented in Section 3.4 by comparing the differences in the model throughput across the experimental configurations. Across the evaluated models, configurations associated with stronger classification performance generally exhibited lower throughput. Although relatively high throughput was achieved when inference was performed using only the base prompt without additional output constraints, the resulting classification contained higher proportion of invalid or hallucinated response. This reduces the practical reliability of the classification workflow, as additional manual review may be required to validate such outputs. Consequently, the computational advantage associated with higher inference throughput may not necessarily translate into greater overall workflow efficiency when substantial post-processing or manual validation is required.

Through the application of formatted-output constraint, the hallucination rate in LLM responses were reduced to zero across all models, improving output validity at the cost of some reduction in inference throughput. The application of RAG also introduced a further but comparatively modest computational trade-off, where the classification performance improved across the evaluated models while throughput reduced only marginally. These findings indicate that the additional computational requirements associated with structured output and retrieval augmentation should be considered alongside their improvements in classification reliability and performance. For API-based proprietary models, the additional input-tokens introduced through RAG may also increase inference costs where API usage is charged according to token consumption, representing an additional practical consideration for large-scale deployment.

For the evaluated retrieval representation approaches, the dense semantic retrieval models required additional processing time to initialise the embedding model, encode the knowledge corpus and generate embeddings for each WO text during retrieval. Given the relatively small knowledge corpus used in this study, this additional computational overhead remained

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



limited. However, the processing requirements may become more substantial when the approach is applied to larger datasets or knowledge corpora.

4.5 Limitations

The main limitation of this study is associated with the use of manually labelled data generated by domain experts as the ground truth for evaluating LLMs classification performance. As expert interpretation may vary between individuals (Lutz et al., 2023), such variation can introduce uncertainty into the reference labels against which model predictions are evaluated. Although the proposed framework attempted to reduce this uncertainty by incorporating user-defined label boundaries and task-specific contextual information, these measures were insufficient to completely mitigate the ambiguities across all predefined label classes.

The second limitation is associated with the computational hardware resources available for local deployment. The hardware used in this study limited the range and scale of the open-source models that could be evaluated efficiently, resulting in the selection of comparatively smaller models than those typically available through proprietary platforms. Access to proprietary models was also constrained by the available API service, and GPT-5.6-luna did not represent the most capable proprietary model available at the time of the study. Consequently, the proprietary-model results should not be interpreted as representing the maximum performance achievable using commercially available LLMs.

Another limitation concerns the prompt design used to describe the classification task at model inference. To ensure consistent inter-model comparison, the prompt was designed to remain compatible with the smaller context-window limits of the open-source models to run efficiently. This constraint may have restricted the amount of task-specific instruction and contextual detail that could be incorporated into the prompt. Further investigation of alternative prompt designs may identify other configurations that yields better results, given the computational and context-window constraints of the evaluated models.

Although the proposed framework achieved strong overall classification performance, some degree of human interventions is still required to validate uncertain or invalid model outputs. The framework automatically flagged invalid responses that did not correspond to any predefined class, but the interpretation and subsequent handling of these invalid outputs remained dependent on the user. Further improvements in model reliability and output consistency are therefore necessary before the workflow could operate with minimal human supervision.

Finally, the findings are limited to the maintenance dataset investigated in this study. Evaluation using datasets from additional wind-turbine operators is necessary to assess the reproducibility and generalisability of the proposed framework across different documentation practices, maintenance terminologies, and operational contexts in future studies. The scalability of the framework to larger datasets, larger knowledge corpora and higher-cardinality label spaces was not evaluated in this study and therefore remain an area for future investigation.

5. Conclusion and Future Work

This study demonstrated the effectiveness of retrieval-augmented generation (RAG) in reducing the performance gap between small scale open-source LLMs and large proprietary LLMs for wind turbine maintenance-text classification. The proposed framework incorporated RAG during model inference, providing user-defined label boundaries and task-specific contextual information to guide model predictions according to predefined classification objective. Additional formatted-output constraints were implemented to substantially reduce label hallucinations and invalid responses across the evaluated models.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Under baseline evaluation, although most open-source models achieved considerably weaker performance than the proprietary model, Qwen3.5-9B achieved only marginally lower macro F1-score than GPT-5.6-luna. Following the application of RAG, classification performance improved across all evaluated open-source models, with Qwen3.5-9B achieving the highest macro F1-score of 0.72 and exceeding the baseline performance of the proprietary model. Qwen3.5-9B also remained competitive to GPT-5.6-luna when both models were evaluated under the RAG configuration.

At the class-level, RAG reduced inter-model variation across the predefined labels, resulting in greater consistency in classification performance across the evaluated models. However, the retrieved contextual information showed varying effects across individual labels. Persistently weak performance and minor reductions observed for certain classes indicate that further refinement of the knowledge corpus may be beneficial for improving the performance consistency across all the predefined label space.

The computational evaluation indicated a trade-off between classification performance and inference efficiency among the locally deployed open-source models, with stronger-performing configurations generally exhibiting lower throughput. This relationship was not directly extended to the proprietary model because it was deployed under a different inference configuration. The retrieval-representation sensitivity analysis further revealed that the choice between sparse lexical and dense semantic retrieval representations had only a limited influence on downstream classification performance under the evaluated conditions. This suggests that the knowledge corpus used in this study could be effectively retrieved using both lexical and semantic similarity approaches.

Although the strong classification performance across the evaluated models in the proposed framework may potentially address the labour-intensive nature of manually labelling maintenance work order, the full automation of wind turbine maintenance-text classification was not demonstrated, as some degree of human intervention remained necessary to validate and handle invalid responses under the framework. The findings were also limited to the maintenance dataset investigated and may not directly generalise to datasets from different wind turbine operators.

Building on these findings, several areas under the adaptation of LLM for maintenance text classification remain for further investigation:

- Evaluate the performance of the proposed framework using alternative maintenance dataset to investigate its generalisability across documentation practices, maintenance terminologies, and operational contexts.
- Investigate the effect of model scale and context-window capacity on prompt optimisation strategies and overall classification performance for open-source LLMs.
- Conduct more extensive retrieval benchmarking to examine the how text representation and retrieval accuracy influence downstream classification performance.
- Investigate the scalability of the proposed framework to larger dataset, larger knowledge corpus and higher-cardinality label spaces.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Appendix A

Table A1: Predefined Labels and Corresponding Context

Predefined Labels	Keywords	Label Definition	Label Boundaries
Pitch and hydraulic	hydraulic, pitch, valve, brake	Maintenance work related to the pitch system and hydraulic system of the wind turbine.	Prefer this when hydraulic oil, hydraulic pumps and brakes are involved, even if it also involves other components.
Blades	blade, edge, trailing edge	Maintenance work related to the blades of the wind turbine.	Prefer this when blades are involved, even if it involves other components.
Generator	generator, rotor, bearing	Maintenance work related to the generator of the wind turbine.	Prefer this when the generator is involved, even if it involves subcomponents that are part of the generator.
Gearbox	gearbox, gear, swarf, gear oil	Maintenance work related to the gearbox of the wind turbine.	Prefer this when the gearbox is involved, even if it also involves other components such as pumps or oil.
Transformer	transformer, traf, trafo	Maintenance work related to the transformer of the wind turbine.	Prefer this when the transformer is involved, even if it also involves other components such as electrical parts or ventilation.
Yaw system	yaw	Maintenance work related to the yaw system of the wind turbine.	Prefer this when the yaw system is involved, even if it also involves other components such as temperature error or motors.
Power supply and converter	power, supply, converter, ups, skii, skii-pack, vcs, skiip	Maintenance work related to the power supply and converter of the wind turbine.	Prefer this when the power supply or converter is involved, even if it also involves other electrical components.
Electrical components	voltage, current, cable, contactor, relay, breaker, q6, q7, q8, vcp, f60, breaker, lctu, capacitor, inverter, coil, circuit, hv, grid	Maintenance work related to the electrical components of the wind turbine.	Maintenance work that involves electrical components, even if it also involves other mechanical components.
Control system	control, ctrl, frequency, logic, synchronize, communication, software, upgrade, controller, module, feedback	Maintenance work related to the controls of the wind turbine.	Prefer this label when the control of wind turbine is involved, especially when it involves software or communication aspects.
hub	hub	Maintenance work related to the hub of the wind turbine.	Prefer this label when the maintenance work is carried out at the hub.
Pumps and motors	Pump, motor	Maintenance work related to the pumps and motors in the wind turbine.	This label should be used when the maintenance work mainly involves pumps or motors, but not when it involves other components.
Grease, oil and coolant	Grease, oil, coolant	Maintenance work related to the grease, oil and coolant of the wind turbine	This label should be used when the maintenance work mainly involves grease, oil or coolant, but not when it involves other components.
scheduled	punch, list, annual, service, audits	Maintenance work and inspections that have been scheduled or planned, either by the operator, preventive maintenance or third party.	Any punch list, annual services, main services and third-party inspections should be labelled as scheduled.
sensors	sensor, measurement, transducer, counter, cms, accelerometer, temperature, level	Maintenance work and fault involved with sensors and measurements.	This label is prioritized for maintenance work that involves sensors or measurement devices, even if the sensor belongs to a different component.
Heaters and coolers	heater, cooler, ventilation, vent, fan, dehumidifier, cooling,	Maintenance work related to the ventilations, heating and cooling systems	Prefer this label when maintenance involves ventilations, heating or cooling systems, unless

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

	heating	in the wind turbine.	other components were mentioned.
Tower and foundation	tower, foundation, monopile, platforms	Maintenance work related to tower and foundation structures of the wind turbine.	Prefer this label even if it also involves other components such as electrical components or yaw platforms.
safety	safety, guide, support, emergency, light, rail, kit, access, aid	Maintenance work related to the safety features of the wind turbine internal or external structures.	Prefer this label when the work involves the repair or maintenance of safety features surrounding the wind turbine, even if electrical components were involved.
others	-	This label is used when the maintenance work does not fit into any of the other categories, or when the category cannot be determined from the text.	This label should only be used when the maintenance work cannot be classified into any of the other categories, or when the text is too vague to determine the category. Avoid using this label if there is any indication of a specific category in the text.

535

Table A2: LLM Inference Parameters

	Inference Parameters	Values
1.	temperature	0
2.	Maximum output token (unformatted output)	10
3.	Maximum output token (JSON formatted output)	30
4.	Think	False
5.	Output Format (if applicable)	JSON

540

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Appendix B

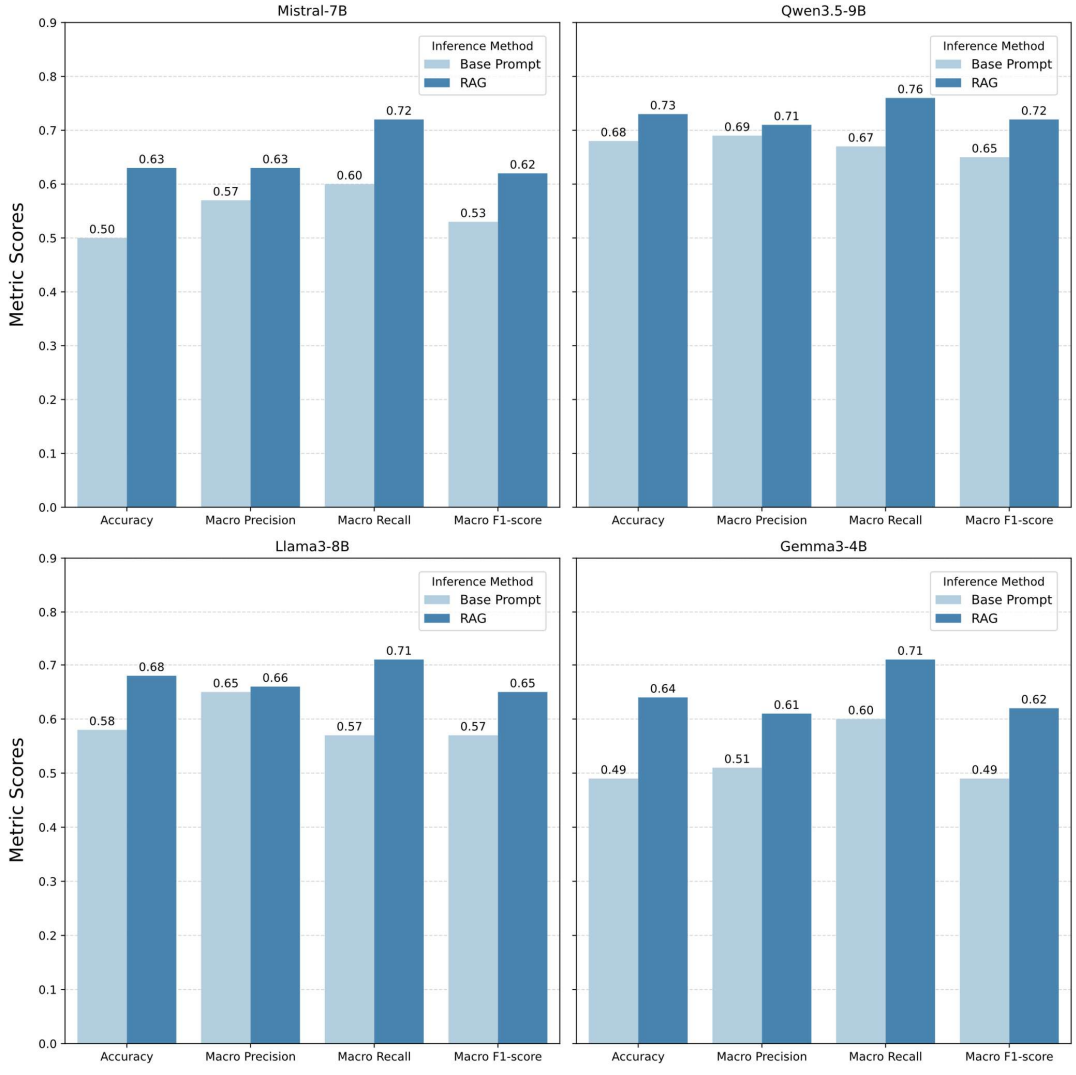


Figure B1: Overall Classification Performance for Each Open-source Model under Base Prompt and RAG Configurations

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Code and data availability

While the full dataset investigated in this work is not available to the public due to confidentiality, the code repository is provided on github: https://github.com/TanYokeWang/MWO-RAG-LLM.git to allow reproducibility using other available datasets containing maintenance text which records interventions performed on affected wind turbines.

Author contributions

Y.W.T contributed to the formal analysis, investigation, methodology, validation, visualization and the writing of the original drafts, whereas JC contributed to data curation, resources and supervision for this work. The conceptualization and writing (review and editing) were performed by both authors.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications adds a standard disclaimer: Copernicus Publications remains neutral with regards to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

The usage of generative AI in this paper had been limited to only the formulation of paper structure, sentence constructions and language corrections where applicable.

Acknowledgements

This work was lead and supported by Centre for Doctoral Training in Wind and Marine Energy Systems and Structures—WAMSS CDT: https://www.wamss-cdt.co.uk/ (accessed on 30 Sep 2026), a comprehensive doctoral training programme at Strathclyde, Oxford and Edinburgh Universities.

Financial support

This work was supported by grant EP/S023801/1 from the UK Engineering and Physical Sciences Research Council (EPSRC).

Review statement

The review statement will be added by Copernicus Publications listing the handling editor as well as all contributing referees according to their status anonymous or identified.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

References

Ai, K., Karr, J. A., Jiang, M., Chawla, N. V., and Wang, C.: KEO: Knowledge Extraction on OMIn via Knowledge Graphs and RAG for Safety-Critical Aviation Maintenance, 2026.

Carroll, J., McDonald, A., and McMillan, D.: Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines, Wind Energy, 19, 1107–1119, https://doi.org/10.1002/we.1887, 2016.

Dao, C. D., Husayn, I., and Dao, P. B.: Analysis of Failures, Downtimes and Reliability of Wind Turbines From SCADA Data, Wind Energy, 28, https://doi.org/10.1002/we.70073, 2025.

Fan, D., Liu, M., Shao, Y., Yang, L., Liu, Y., Zhang, Y., Ren, Y., and Wang, Z.: Domain-Specific Large Language Model for Maintenance Decision-Making on Wind Farms by Labeled-Data-Supervised Fine-Tuning, Engineering, 60, 343–361, https://doi.org/10.1016/j.eng.2025.12.019, 2026.

Grandini, M., Bagli, E., and Visani, G.: Metrics for Multi-Class Classification: an Overview, 2020.

Han, J., Kamber, M., and Pei, J.: 2 - Getting to Know Your Data, in: Data Mining: Concepts and Techniques (Third Edition), Morgan Kaufmann, Boston, 39–82, https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00002-2, 2012.

J, M. R., VM, K., Warrier, H., and Gupta, Y.: Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations, 2024.

Lin, L., Zhang, S., Fu, S., and Liu, Y.: FD-LLM: Large language model for fault diagnosis of complex equipment, Advanced Engineering Informatics, 65, https://doi.org/10.1016/j.aei.2025.103208, 2025.

Lutz, M. A., Schäfermeier, B., Sexton, R., Sharp, M., Dima, A., Faulstich, S., and Aluri, J. M.: KPI Extraction from Maintenance Work Orders—A Comparison of Expert Labeling, Text Classification and AI-Assisted Tagging for Computing Failure Rates of Wind Turbines, Energies (Basel)., 16, https://doi.org/10.3390/en16247937, 2023.

Malyi, M., Shek, J., McDonald, A., and Biscaya, A.: A Comparative Benchmark of Large Language Models for Labelling Wind Turbine Maintenance Logs, IET Renewable Power Generation, 20, https://doi.org/10.1049/rpg2.70212, 2026.

Marzouk, O. A.: Progress of the International Energy Agency Tracking Clean Energy Progress (IEA TCEP), https://doi.org/10.13140/RG.2.2.14098.62402, 2026.

Mucchielli, P., Bhowmik, B., Ghosh, B., and Pakrashi, V.: Real-time accurate detection of wind turbine downtime - An Irish perspective, Renew. Energy, 179, 1969–1989, https://doi.org/10.1016/j.renene.2021.07.139, 2021.

Navinchandran, M., Sharp, M. E., Brundage, M. P., and Sexton, T. B.: Discovering critical KPI factors from natural language in maintenance work orders, J. Intell. Manuf., 33, 1859–1877, https://doi.org/10.1007/s10845-021-01772-5, 2022.

Qin, Y., Song, K., Hu, Y., Yao, W., Cho, S., Wang, X., Wu, X., Liu, F., Liu, P., and Yu, D.: InFoBench: Evaluating Instruction Following Ability in Large Language Models, in: Findings of the Association for Computational Linguistics: ACL 2024, 13025–13048, https://doi.org/10.18653/v1/2024.findings-acl.772, 2024.

Salo, E. and Mcmillan, D.: Work Orders-Value from Structureless Text in the Era of Digitisation, https://doi.org/https://doi.org/10.2118/195788-MS, 2019.

Sexton, T., Hodkiewicz, M., and Brundage, M. P.: Categorization Errors for Data Entry in Maintenance Work-Orders, https://doi.org/https://doi.org/10.36001/phmconf.2019.v11i1.790, 2019.

Shen, J., Tenenholtz, N., Hall, J. B., Alvarez-Melis, D., and Fusi, N.: Tag-LLM: Repurposing General-Purpose LLMs for Specialized Domains, 2024.

Tian, Y., Zhang, L., Chang, F., Wang, N., Su, B., and Gu, Z.: A deep reasoning large language model designed for troubleshooting faults in wind turbine generators, in: Proceedings of 2025 2nd Symposium on Big Data, Neural Networks, and Deep Learning, BDNNDL 2025, 1–8, https://doi.org/10.1145/3784013.3784014, 2026.

Tural, B., Örpek, Z., and Destan, Z.: Retrieval-Augmented Generation (RAG) and LLM Integration, in: 2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS), 1–5, https://doi.org/10.1109/ISAS64331.2024.10845308, 2024.

https://doi.org/10.5194/wes-2026-203
Preprint. Discussion started: 5 November 2026
© Author(s) 2026. CC BY 4.0 License.



Turnbull, A., Carroll, J., and McDonald, A.: Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure, Wind Energy, 24, 197–211, https://doi.org/10.1002/we.2567, 2021.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N.: C-Pack: Packaged Resources To Advance General Chinese Embedding, 2023., 2023.

620 Batch API | OpenAI API, https://developers.openai.com/api/docs/guides/batch, last access: 2 October 2026., 2026.

GPT-5.6 Luna Model | OpenAI API, https://developers.openai.com/api/docs/models/gpt-5.6-luna, last access: 2 October 2026., 2026.

Ollama documentation - Ollama, https://docs.ollama.com/, last access: 2 October 2026., 2026.