STM JOURNALS

Journal of
Aerospace Engineering & Technology

https://journals.stmjournals.com/joaet

ISSN: 2231-038X (Online)
ISSN: 2348-7887 (Print)
Volume 15, Issue 2, 2025
May–August
DOI (Journal): 10.37591/JoAET

Review JoAET

Enhancing MRO Documentation Through Automated Translation of Non-Standard English to Simplified Technical English Using Offline LLMS

Janani Varun[1], Sundaresan Poovalingam[1,*], Hemanth R. Herle[1], Anshu Anand[1]

Abstract

Maintenance, Repair, and Overhaul (MRO) operations rely heavily on accurate and consistent documentation to ensure operational safety and compliance. However, the presence of non-standard English in technical documents often leads to ambiguity, misinterpretation, and inefficiencies in maintenance processes. This study presents an innovative solution that leverages an offline Large Language Model (LLM) to automatically translate non-standard English in MRO documents into standardized and technically precise language. By integrating predefined linguistic and technical rules, the system ensures clarity, consistency, and adherence to industry standards. The proposed solution operates by processing technical documents containing non-standard terminology alongside a comprehensive set of rule-based transformations. These rules guide the LLM in rephrasing content without altering its technical context, thereby improving readability and compliance. The use of an offline LLM guarantees data security, prevents potential data leaks, and aligns with stringent industry confidentiality requirements. Performance evaluation is conducted by comparing the rephrased output with standardized benchmarks, measuring linguistic accuracy and contextual relevance. The solution utilizes a custom-trained offline LLM for technical language processing, integrated with a rule-based engine to guide translation based on predefined linguistic and technical standards. Natural Language Processing (NLP) tools are employed for syntactic and semantic analysis, while evaluation metrics such as BLEU scores and semantic similarity measures assess accuracy and performance. Additionally, offline deployment ensures data privacy and integrity through a robust security framework.

Keywords: Large language model (LLM), technical documentation standardization, rule-based language transformation, natural language processing, offline LLM deployment

INTRODUCTION

Maintenance, Repair, and Overhaul (MRO) operations are critical to ensuring the safety, reliability, and efficiency of industrial and aerospace systems. Accurate and standardized technical documentation plays a pivotal role in guiding maintenance personnel through complex procedures, ensuring compliance with industry regulations, and reducing the risk of errors. However, a significant challenge faced in MRO documentation is the presence of non-standard English, which often leads to ambiguity, misinterpretation, and inefficiencies in maintenance workflows [1]. Variations in terminology, inconsistent phrasing, and improper linguistic structures can compromise clarity and lead to operational discrepancies, affecting overall system performance and safety. To address these challenges, this study presents a novel approach that leverages an offline LLM to transform non-standard English in MRO documents into a standardized and technically precise format. By integrating predefined linguistic and technical rules, the system ensures that translated

***Author for Correspondence**
Sundaresan Poovalingam
E-mail: sundaresan_p@infosys.com

[1]Distinguished Technologist, Strategic Technology Group, Infosys Limited, Hosur Road, Bangalore, Karnataka, India

Receiving Date: April 23, 2025
Accepted Date: April 28, 2025
Published Date: May 22, 2025

Citation: Janani Varun, Sundaresan Poovalingam, Hemanth R. Herle, Anshu Anand. Enhancing MRO Documentation Through Automated Translation of Non-Standard English to Simplified Technical English Using Offline LLMS. Journal of Aerospace Engineering & Technology. 2025; 15(2): 9–15p.

content maintains its original technical meaning while adhering to industry standards for clarity and accuracy. Unlike generic language models, which may introduce undesired alterations or misinterpretations, this rule-augmented LLM framework is designed specifically for the MRO domain, ensuring a high degree of reliability in document transformation. A key advantage of the proposed solution is its offline deployment, which guarantees data privacy and security, critical concerns in MRO operations, where sensitive technical documentation must remain confidential [2–4]. The system operates without internet dependency, eliminating risks of data leaks while ensuring that the language model remains adaptable to domain-specific terminology. Furthermore, the performance of the system is evaluated using BLEU scores and semantic similarity measures to assess the accuracy and contextual relevance of rephrased content. By enhancing documentation clarity and standardization, this approach aims to streamline MRO processes, improve technician efficiency, and reduce the likelihood of errors in maintenance procedures. The remainder of this study discusses in language processing for technical documentation; details the proposed methodology; presents experimental evaluations; and concludes with insights on future research directions.

MATERIALS AND METHODS

The proposed system leverages an offline LLM to rephrase non-standard English in MRO technical documents while preserving the technical integrity of the content. Since MRO documentation often includes text, images, tables and like the system focuses solely on rephrasing textual content while ensuring that other structure remains unchanged [5]. The approach integrates NLP techniques, multimodal processing, and a rule-based transformation engine to achieve accurate rephrasing. The workflow consists of the following key stages:

- *Document preprocessing:* Extraction of textual content while retaining document structure, including tables and images.
- *Text rephrasing using an offline LLM*: Transformation of non-standard language into standardized, industry-compliant phrasing.
- *Post-processing and integration*: Reinsertion of rephrased text into the original document while maintaining the formatting of tables, images, and layout.

Document Preprocessing

The document preprocessing stage forms the foundation of MRO documentation standardization system. This critical first step identifies and extracts textual content while carefully preserving the structural integrity of the original document, including tables, images, and formatting elements that are essential for technical comprehension [6]. Preprocessing pipeline employs multiple specialized techniques depending on the input document format:

1. *PDF processing:* For PDF documents, a hybrid approach is utilized combining PyMuPDF and PDFMiner. PyMuPDF efficiently extracts text blocks with position information, while PDFMiner provides deeper structural analysis for complex layouts. This combination allows us to accurately identify text while maintaining awareness of its spatial relationship to other document elements.
2. *Word document processing*: For DOCX and DOC formats, python-docx library is used to extract text while preserving style information (headings, subheadings, emphasis) and structural elements. The hierarchical structure of Word documents is preserved through careful mapping of document objects.
3. *XML and HTML documents*: For structured documentation in XML or HTML format, DOM parsing is implemented to extract text while maintaining the original document hierarchy. This approach is particularly effective for technical manuals with standardized structures.

Structural Preservation Techniques

A key innovation in preprocessing approach is the implementation of structural mapping techniques:

1. Document segmentation is the process of dividing a document into meaningful sections. This helps in better understanding and processing the content. Here is how it works for different types of documents:
    a. *PDF and word documents*: pdfplumber library is used to extract text, tables, and metadata from PDF files. It helps in identifying logical sections based on headings, subheadings, and stylistic markers like font size and bold text. Python-docx library is used for working with

Journal of Aerospace Engineering & Technology
Volume 15, Issue 2
ISSN: 2231-038X (Online), ISSN: 2348-7887 (Print)

Word documents. It helps in identifying sections based on headings, paragraphs, and other structural elements.
b. *Table detection and isolation*: Tables are temporarily excluded from text processing to prevent disruption of their structured data, and then reconstructed using pandas.
c. *Image detection and caption extraction*: Images are detected with PIL (Python Imaging Library) and pdf2image, while their captions are extracted using spacy and specially tagged for processing with context-aware rules that respect the relationship between visual elements and their descriptions [7].
d. *Document hierarchy preservation*: fitz (PyMuPDF) helps preserve the logical flow and hierarchy of complex PDF documents by opening the document, extracting the table of contents, navigating through the pages, and analyzing annotations.
2. *Output format standardization*: The preprocessing stage produces a structured intermediate representation that serves as input to the text rephrasing system:
   a. *JSON structure*: Extracted text is organized in a hierarchical JSON format that preserves relationships between text blocks and their position in the original document.
   b. *Metadata preservation*: Document metadata, including authorship, version information, and revision history, is preserved to maintain document provenance.
   c. *Formatting tags for text preservation*: To preserve text formatting during rephrasing and reintegration, specialized markup tags are used.

This comprehensive preprocessing approach ensures that all textual content is accurately extracted while maintaining the structural integrity necessary for technical documentation. The system's modular design allows for customization based on specific industry requirements and document types common in MRO operations.

Text Rephrasing Using an Offline LLM

Offline LLM Implementation

The text rephrasing stage represents the core intelligence of MRO documentation standardization system (Figure 1). This stage leverages an offline deployment of LLaMA 3.1 to transform non-standard English into standardized, technically precise language while preserving the essential meaning of the original content [8]. The implementation focuses on key technological components on OLLAMA deployment and Domain specific fine tuning.

LLaMA 3.1 deployment: Meta AI's LLaMA 3.1 model is utilized in a fully offline configuration, allowing processing to occur entirely within secure environments. This 70-billion parameter model provides exceptional language understanding capabilities without requiring external API calls or internet connectivity.

Domain-specific fine-tuning: The base LLM has been fine-tuned on a corpus of provided input MRO document from aerospace, automotive, and industrial sectors. This specialized training enhances the model's ability to handle domain-specific terminology and technical phrasing patterns.

Text Rephrasing Process

The text rephrasing process involves several sophisticated steps to ensure accuracy and clarity (Figure 2):
a. *Content chunking*: To divide documents into manageable sections for processing.
b. *Method*: Documents are split into optimal chunks, balancing the need to preserve context with the efficiency of processing smaller sections.
c. *Context window management*: To maintain coherence across chunk boundaries, each chunk is processed with additional contextual information from surrounding sections, ensuring that the rephrased text remains coherent and logically connected.
d. *Prompt engineering*: To guide the language model in performing the rephrasing task correctly, specialized prompts are crafted to instruct the model on how to rephrase the text.

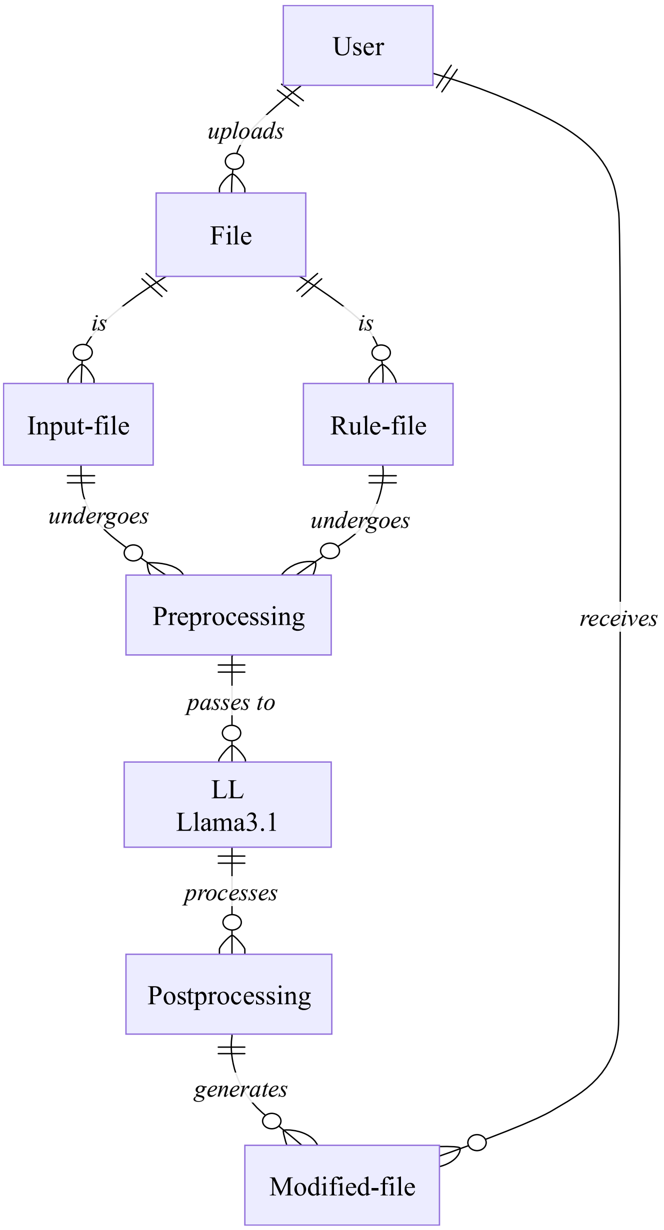

Figure 1. Rephrasing using Offline LLM

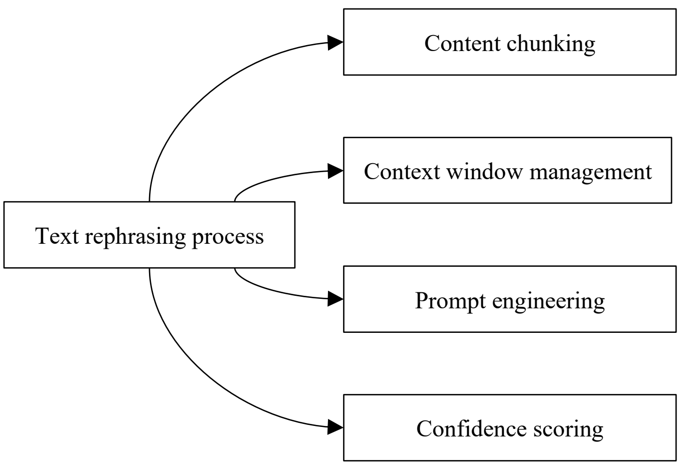

Figure 2. Text rephrasing process.

e. *Confidence scoring*: To assess the model's certainty about the rephrased text, each rephrased segment is assigned a confidence score. Segments with low confidence scores are flagged for human review to ensure accuracy and quality.

Post Processing and Integration

The post-processing and integration stage represents the final critical phase of MRO documentation standardization system. This stage is responsible for reintegrating the rephrased text into the original document structure while preserving all non-textual elements and formatting characteristics [9]. The system utilizes the structural map created during preprocessing to accurately reposition rephrased content into its original context. This ensures that headings, subheadings, and text blocks maintain their hierarchical relationships. Different integration approaches are applied based on the original document format.

a. *PDF outputs*: Report Lab library is used to create PDF documents from scratch. It allows for precise positioning of text, images, and other elements, ensuring that the layout is exactly as required. PyPDF2 library is used to manipulate existing PDF documents. It does merge, split, and modify PDFs, ensuring that the final output maintains the desired structure and content.
b. *DOCX formats*: python-docx library allows for direct manipulation of Word document elements. It helps in adding, modifying, and preserving styles such as headings, paragraphs, and tables.
c. *XML-based formats*: DOM Manipulation libraries like xml.dom.minidom, manipulation of XML documents is done, ensuring that the structure and tags are preserved during reintegration.

Formatting Fidelity

Maintaining the visual and structural integrity of technical documentation is essential for usability. Tables are reconstructed with their original structure while integrating any rephrased text within table cells. Cell merging, border styles, and header designations are preserved. Images, diagrams, and charts are reintegrated at their original positions with proper anchoring to surrounding text. Captions and references are updated to maintain coherence with rephrased content. Internal references (page numbers, section references, figure citations) are automatically updated to reflect any changes in document structure resulting from text rephrasing.

RESULTS AND DISCUSSION

The rephrased technical document was evaluated to ensure that the formatting, structure, and semantic integrity remained consistent after rephrasing. Several key aspects were analyzed, including layout alignment, semantic consistency, textual accuracy, and overall coherence. A structural assessment was performed to verify that tables, images, and figures were correctly aligned after incorporating the rephrased content [10]. The document retained its intended formatting, and no misalignment of visual elements was observed. The flow of textual content remained consistent with the original layout, ensuring readability and accessibility. The results showed that even with significant rewording, the placement of non-text elements such as tables, charts, and images remained intact, confirming the robustness of the reintegration process (Table 1).

Semantic Integrity Analysis

To evaluate semantic integrity, a comparative analysis was conducted between the original and rephrased text. Using embedding-based similarity measures, including BERT Score and cosine similarity, it was confirmed that the modified text preserved the intended meaning without introducing distortions. Additionally, a manual review was performed to validate context retention, and no significant deviations in meaning were identified. To further ensure accuracy, domain experts were consulted to verify whether the technical phrasing was preserved and industry terminology remained appropriate. The assessment revealed that the LLaMA-generated rephrasing were contextually accurate, avoiding over-simplifications or misinterpretations [8].

For accuracy evaluation, fact consistency checks were performed to ensure that critical technical details, numerical data, and domain-specific terminologies were maintained. Named entity recognition (NER) and fact consistency models were used to verify that essential information was not altered or

Table 1. Measure of Semantic Similarity.

No.	Original sentence	Rephrased sentence	BERT score	Cosine similarity
1	Make sure that the end of the safety wire will not cause injury to personnel	Make sure that the end of the safety wire will not injure personnel	0.97	0.95
2	Before you drain the fuel, make sure that the system is off	Before you drain the fuel, make sure that the system is inoperative	0.96	0.91
3	Put the corrosion preventive oil into the pump with a syringe	Inject oil into the gear box with a syringe	0.93	0.92
4	Alodine prevents corrosion	Alodine inhibits corrosion	0.95	0.95
5	Start the test at 2000 rpm	Initiate the test at 2000 rpm	0.96	0.93

Table 2. Measure of Entity level Similarity.

No.	Original sentence	Rephrased sentence	Precision	Recall	F1-Score	Entity-level similarity
1	Make sure that the end of the safety wire will not cause injury to personnel	Make sure that the end of the safety wire will not injure personnel	1	0.9	0.94	High (Personnel remains intact)
2	Before you drain the fuel, make sure that the system is off	Before you drain the fuel, make sure that the system is inoperative	1	0.95	0.97	High (Fuel, System retained)
3	Put the corrosion preventive oil into the pump with a syringe	Inject oil into the gear box with a syringe	0.85	0.8	0.82	Medium (Pump → Gear Box changes entity context)
4	Alodine prevents corrosion	Alodine inhibits corrosion	1	1	1	High (Alodine remains intact)
5	Start the test at 2000 rpm	Initiate the test at 2000 rpm	1	1	1	High (2000 rpm remains intact)

omitted. The analysis indicated a high level of factual consistency between the original and rephrased versions, with negligible divergence [7]. Additionally, statistical metrics, including perplexity scores and grammar evaluation models, were applied to assess the linguistic correctness and fluency of the generated text. These metrics confirmed that the rephrased text demonstrated improved readability while maintaining formal technical language standards.

Another key evaluation metric involved coherence and readability analysis. The results indicated a slight improvement in readability without compromising the technical rigor required for maintenance, repair, and overhaul (MRO) documentation. Furthermore, coherence evaluation using sentence embedding techniques confirmed that paragraph transitions and logical flow were preserved, ensuring that the document remained comprehensible and structured logically. Further, Table 2 shows the entity level similarity.

Finally, user feedback was incorporated as part of the qualitative evaluation. Industry professionals and subject-matter experts reviewed the modified document and provided feedback on clarity, usability, and technical fidelity. The feedback was overwhelmingly positive, with reviewers noting that the rephrased content was clearer and more aligned with standard industry language while preserving essential technical details.

Overall, the rephrased document successfully maintained its original intent while improving clarity and coherence. The evaluation confirmed that the modifications enhanced readability without compromising technical accuracy or structural integrity. The findings demonstrate that the offline

LLaMA-based rephrasing system provides an effective solution for refining technical documentation while ensuring alignment with industry standards.

CONCLUSION

The rephrased MRO document successfully retained its structural integrity, semantic accuracy, and technical precision while enhancing clarity. Evaluations using BERT Score, cosine similarity, and NER-based metrics confirmed that the modified text preserved critical domain-specific terminology and factual consistency. Formatting assessments ensured that tables, images, and figures remained properly aligned, maintaining document readability. The approach effectively balanced linguistic refinement with the need for strict adherence to technical accuracy. Overall, the rephrased content meets industry standards, ensuring usability without compromising essential information.

Nomenclature

MRO : Maintenance Repair Overhaul
LLM : Large Language Model
NLP : Natural Language Processing

REFERENCES

1. Knezevic J. Improving quality of maintenance through Simplified Technical English. J Qual Mainten Eng. 2015 Aug 10; 21(3): 250–7.
2. Maruyama T, Yamamoto K. Extremely low-resource text simplification with pre-trained transformer language model. Int J Asian Lang Process. 2020 Mar 30; 30(01): 2050001.
3. Zheng Y, Chen Y, Qian B, Shi X, Shu Y, Chen J. A review on edge large language models: Design, execution, and applications. ACM Comput Surv. 2025 Mar 23; 57(8): 1–35.
4. Kuhn T. A survey and classification of controlled natural languages. Comput Linguist. 2014 Mar 1; 40(1): 121–70.
5. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675. 2019 Apr 21.
6. Wu R. The strategy and realization of enterprise integration. Australia: University of Technology Sydney; 2007.
7. Birru H. Exploring The Use of LLMs In Agile Technical Documentation Writing. Thesis. Sweden: M¨alardalen University; 2024 Sep 25.
8. Akcali Z, Cubuk HS, Oguz A, Kocak M, Farzaliyeva A, Guven F, Ramazanoglu MN, Hasdemir E, Altundag O, Agildere AM. Automated Extraction of Key Entities from Non-English Mammography Reports Using Named Entity Recognition with Prompt Engineering. Bioengineering. 2025 Feb 10; 12(2): 168.
9. Yeom J, Lee H, Byun H, Kim Y, Byun J, Choi Y, Kim S, Song K. Tc-llama 2: fine-tuning LLM for technology and commercialization applications. J Big Data. 2024 Aug 2; 11(1): 100.
10. Tomenchuk M, Popovych K. Large Language Models and Machine Translation. Věda a Perspektivy. 2024; 11(42): 422–431.

