Automating Operations and Maintenance Manual Data Extraction for Retrofit Applications: Assessing GPT-4o's Performance and Challenges

Haoyu Zuo[1,*], Yuan Yin[1], Kamyar Hazeri[1],Boheng Wang[1],Tom Jennings[2],Julian Buhagiar[2],Ben Cartwright[3], Ruocong Liu[1], and Peter Childs[1]
[1]Imperial College London, London, United Kingdom
[2]BRE, London, United Kingdom
[3]Reusefully, London, United Kingdom
*Email: hz2019@ic.ac.uk

Abstract

Operations & Maintenance (O&M) manuals contain data that is essential to understand ahead of building retrofits. However, their unstructured format makes information retrieval time-consuming and error-prone. This study evaluates GPT-4o's ability to extract information from O&M manuals, assessing accuracy in text extraction, categorization, and numerical data handling. Structured extraction steps and prompt strategies were developed based on industry needs. The extraction performance was compared to human-extracted data based on a dataset of 60 O&M manuals, using quantitative and qualitative evaluation. Results form the study show that text-based extraction is reliable, particularly for component names, types, and manufacturer details, where formatting differences had little impact on accuracy. However, challenges appeared in fields requiring categorization, numerical interpretation, and structured mapping. One issue was category misalignment, where GPT-4o could assign components to broad categories but had difficulty with domain-specific classifications that require specialized knowledge. This inconsistency suggests that rule-based constraints or further adjustments may be needed to improve standardization in retrofit applications. Another challenge was the model's handling of numerical data. Mistakes occurred in quantity extraction, page numbering, and location references, where the model miscalculated totals, misread measurement units, or mismatched printed and digital page numbers. These errors suggest that GPT-4o lacks a way to process numerical information in context, meaning manual checks are still required for numerical fields. These findings highlight GPT-4o's potential to automate O&M data extraction, reducing manual effort and speeding up the retrofit process.

Highlights

- Demonstrates a GPT-4o-based O&M manual data extraction approach for data required for retrofits.
- Assesses GPT-4o's performance in extracting structured O&M data.
- Highlights challenges in classification and numerical accuracy.

Introduction

The New London Plan (Greater London Authority, 2021) requires that larger scale demolition and retrofit projects should achieve at least a 95% reuse or recycling rate for materials. Retrofitting is an important approach for extending building lifespan and reducing waste. A successful retrofit project depends on a clear understanding of the building's current state, which involves reviewing as-built documentation and operational records. (Ma et al., 2012)

Before site assessments, surveyors examine building documentation including Operations & Maintenance (O&M) manuals. These materials provide information on installed systems, materials, and past modifications, which support retrofit planning and execution. O&M manuals contain technical specifications, installation procedures, and maintenance records, making them a key resource for evaluating existing components and identifying necessary upgrades (Chartered Institution of Building Services Engineers, 2020).

Despite their importance, O&M manuals can be difficult to use efficiently (Building Services Research and Information Association, 2016). They include both structured and unstructured data, often spread across multiple sections with inconsistent formatting. Key details may be difficult to locate, requiring surveyors to manually search through lengthy documents. This process is time-consuming, error-prone, and inefficient, especially in projects involving multiple manuals. The lack of a standardized format further complicates data retrieval and comparison, limiting their usability in retrofit planning.

Recent advancements in Large Language Models (LLMs) offer new possibilities (Achiam et al., 2023) for automating O&M data extraction. LLMs can process large volumes of text (Lewis et al., 2020) and extract structured information without relying on predefined rule-based systems (Bommasani et al., 2021). However, applying LLMs to this task presents two main challenges:

How to structure the extraction process to ensure relevant and accurate results. O&M manuals vary in format and terminology, making it unclear how an LLM should be guided to extract the right information. The effectiveness of a predefined prompt structure needs to be examined to determine whether it can consistently produce useful outputs.

How accurately the extracted results match human-verified data. Since O&M manuals do not always follow a consistent structure, errors in retrieval can affect the reliability of extracted information. Extracted values may be incomplete, misinterpreted, or incorrectly assigned to categories, limiting their usefulness in retrofit planning.

To address these challenges, this study applies GPT-4o to automate O&M data extraction. The approach involves designing a two-step framework and prompt stratgies, extracting information into predefined retrofit-relevant categories, allowing follow-up questions and information localization and evaluating accuracy through a benchmark comparison with human-extracted data.

Methods

Data Collection

This study uses an O&M Corpus from Reusefully to evaluate GPT-4o's ability to extract structured data from O&M manuals for building retrofits. A total of 60 manuals were collected from a real-world commercial project, covering different building types, systems, and retrofit-related documents. To protect sensitive information, identifiable project details have been anonymised or removed. The collected manuals were categorized into four primary types based on their content and application, shown in Table 1.

Table 1: Classification and Quantity of O&M Manuals

O&M Manual Category	Number of Manuals	Example Files
Equipment O&M Manuals	15	Meiko Dishwasher FV, Franke Coffee Machine, Williams Upright Double Door Refrigerator G2T-SA, Manitowoc Combi Oven OEB,
Building Systems Manuals	10	Curtain Walling - O&M, Commercial Doors - O&M, O&M West Forvie Building - Flue, O&M XXXX House
Material & Component Manuals	18	Nora Flooring Maintenance, O&M - Ceramic Wall and Floor Tiling, Maxwood Washroom Systems, Milliken Maintenance Instruction - Carpet
Comprehensive O&M Manuals	17	XXXX School O&M - Raised Access Flooring, XXXX School O&M - Sliding Folding Partition, XXXX School O&M - Sub and Superstructure

Defining Extraction Categories and Other Needs

Through discussions with industry experts in building surveying and retrofit planning, we identified the key categories of information required from O&M manuals, shown in Table 2.

Table 2: Definitions of Categories

Category	Purpose in Retrofit Process
Component Name	Identifies the specific part or item in the building to determine if it needs replacement, maintenance, or upgrading.
Component Type	Groups components based on function (e.g., HVAC, lighting) to evaluate system-wide performance and standardization needs.
Element Type	Links the component to a specific building system (e.g., ventilation, flooring) for targeted retrofit interventions.
Amount and Unit of Measurement	Helps in procurement and budgeting by specifying required material quantities.
Location in Building	Determines where interventions are needed, ensuring accurate site assessments and planning.
Date of Installation	Assesses component age and expected lifespan, guiding replacement schedules.
Manufacturer Name	Identifies suppliers for potential reordering, spare parts, or warranty claims.
Warranty Information and Dates	Ensures compliance with warranty terms, reducing unnecessary costs for premature replacements.
Assembly Instructions	Provides guidance for installation, dismantling, or upgrades, minimizing errors during retrofit implementation.

Experienced building surveyors also noted that additional follow-up questions often arise during the retrofit process, requiring clarification beyond these predefined categories. They also emphasized the need to locate page numbers for extracted information, allowing surveyors to verify details efficiently.

O&M Data Standardization

To guide GPT-4o in extracting relevant information, the predefined categories are defined accordingly. Each category represents a key type of data commonly found in O&M documentation:

Component Name - This refers to the exact name of a part or item as it appears in the O&M manual. The name should match the wording used in the original document to ensure consistency in identification.

Component Type - This defines the general classification of a component based on its function within a system. It describes what kind of item it is in a broad sense, such as "metal studwork," "carpet tiles," "HVAC filter," or "light fixture." This classification helps group components according to their purpose, structure, or material.

Element Type - This indicates the building system or area that the component belongs to. Examples include "internal walls & partitions," "floor finishes," "plumbing," or "electrical system." This field provides context for how the component integrates into the overall structure.

Amount and Unit of Measurement - This records the quantity of the component and the corresponding unit of measurement as specified in the O&M manual. Common formats include "200 linear meters," "1500 m^2," or "4 units," ensuring accurate tracking of component volume.

Location in Building - This field refers to the exact location within the building where the component is installed or used. It includes details such as "Offices, 3rd floor," "Mechanical room, basement," or "Lobby ceiling" to facilitate maintenance and inspections.

Date of Installation - This records the precise date when the component was installed, typically in a day/month/year format. This information is critical for tracking the lifecycle and maintenance schedule of building elements.

Manufacturer Name - This specifies the company or brand that produced or supplied the component. The official name should be listed as per the product documentation, such as "Franke Kaffeemaschinen AG" or "GE Appliances."

Warranty Information and Dates - This captures details about the warranty coverage for the component, including the start and end dates of the warranty period. Examples include "2-year warranty from 01/01/2023 to 01/01/2025" or "5-year limited warranty."

Assembly Instructions - This provides step-by-step guidance on assembling, installing, or disassembling the component if necessary. These instructions may include diagrams, tools required, and procedural steps for correct installation.

GPT-4o-Based Extraction Steps

To facilitate the extraction of information from O&M manuals, GPT-4o operates as an assistant into three steps:

Step 1: File Handling and Verification

Before extraction, GPT-4o verifies the uploaded document to ensure it is in a compatible format (PDF, DOCX, XLSX) and contains extractable content. If the document does not conform to expected formats, the assistant informs the user and does not attempt extraction. Users must confirm before proceeding.

Step 2: Structured Data Extraction and Summary

Once a document is verified, GPT-4o extracts relevant information into predefined categories. The extracted information includes component name, type, element type, quantity, installation date, location, manufacturer, warranty details, and assembly instructions. The element type is assigned to one element type, strictly following the New Rules of Measurement (NRM) classification (RICS, 2020*)*, ensuring consistency in categorization. To maintain accuracy, the model strictly limits extraction to predefined categories and avoids generating inferred content.

Step 3: Follow-up Questioning and Information Localization

Beyond extraction, GPT-4o allows users to ask project-specific follow-up questions beyond predefined categories. These may include material specifications, historical maintenance records, or component integration with other systems. Additionally, GPT-4o enhances usability by referencing page numbers and source sections, helping users locate extracted data within the original document.

Prompt Engineering Strategies

To ensure GPT-4o accurately extracts data from O&M manuals and enables follow-up questions, a structured prompt engineering approach was applied (OpenAI, n.d.). The prompt was designed to explicitly define the model's role, skills, and constraints, guiding the extraction process in a controlled manner:

Role Definition – GPT-4o was framed as an O&M manual assistant, responsible for guiding file uploads, verifying documents, and extracting structured data into predefined categories. This ensures the model stays within scope and does not generate unnecessary content.

Skill-Based Structuring – The model was instructed to first verify file compatibility before extraction. It then extracts information strictly within predefined categories, ensuring structured and relevant outputs.

Constraint Enforcement – The model follows strict category definitions, limiting Element Type selection to a fixed NRM-aligned list and ensuring only one extracted value per category. It does not infer missing data or generate content outside the provided framework.

Results

The tool was developed on the Coze (2024) platform, which provides data protection and deployment flexibility, unlike OpenAI's public platform. The final interface is shown in Figure 1. To evaluate the quality of GPT-4o's extractions from O&M manuals, a two-part assessment was conducted, consisting of quantitative accuracy assessment and qualitative assessment of extraction quality. This assessment ensures that the extracted data is both factually correct and useful for retrofit applications.

Quantitative Results

To evaluate the accuracy of GPT-4o's extractions, a manual benchmark dataset was created by an experienced building surveyor. The surveyor independently extracted relevant information from the O&M manuals, ensuring that the benchmark reflected real-world data and industry practices.

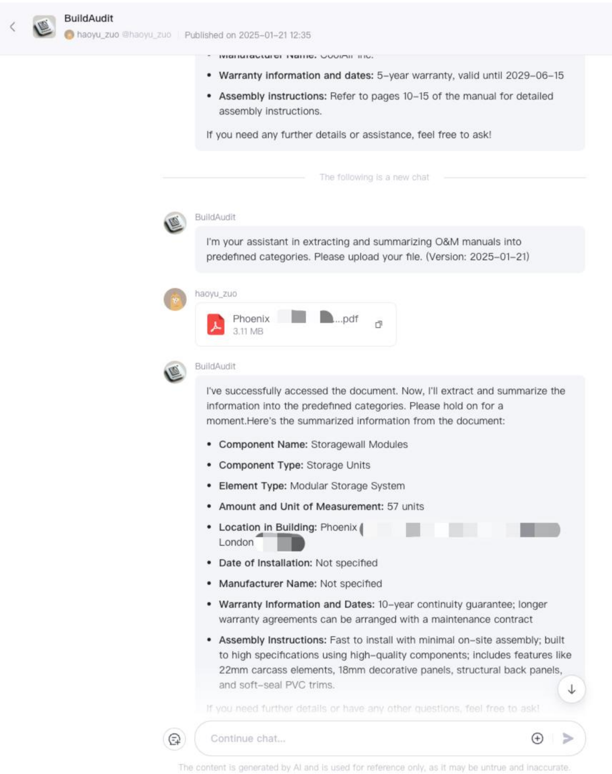

Figure 1. A screenshot of the tool developed.

A human-based comparison was used instead of an automated algorithmic evaluation, as certain data fields require contextual interpretation beyond semantic similarity analysis. For example, in the Location in Building field, the benchmark lists room numbers across multiple floors (e.g., 103, 104, 105 on the first floor; 201, 202, 203 on the second floor; 301, 302, 303 on the third floor). GPT-4o's output was 103 - 105 on the first floor; 201 - 203 on the second floor; 301 - 303 on the third floor. While an algorithm might flag this as incorrect due to formatting differences, a human evaluator would recognize the contextual information as equivalent and accurate. The following guidance are applied during the evaluation

Component Name – The extracted name was compared to the benchmark for an exact match. Minor differences in wording (e.g., "LED Panel Light" vs. "LED Ceiling Panel") were accepted if they did not change the meaning or classification.

Component Type – The extracted classification was checked for functional accuracy. If it aligned with industry terminology, variations in phrasing (e.g., "Air Handling Unit" vs. "HVAC Unit") were considered correct.

Element Type – The extracted category was validated against NRM classification. If multiple valid classifications existed, specialists assessed whether GPT-4o's choice was reasonable in the context of building retrofits.

Amount and Unit of Measurement – The extracted numerical value and unit had to match the benchmark exactly. Small formatting differences (e.g., "10 units" vs. "10.0") were accepted if they did not affect the intended measurement.

Location in Building – The extracted location was evaluated based on meaning rather than exact wording. If it conveyed the same position as the benchmark (e.g., "First Floor, Section A" vs. "Section A on First Floor"), it was considered correct. Missing key details resulted in an incorrect mark.

Date of Installation – The extracted date had to match the benchmark exactly. Minor formatting differences (e.g., "01/01/2022" vs. "1st Jan 2022") were accepted, but missing or misinterpreted dates were marked incorrect.

Manufacturer Name – The extracted name was checked for consistency. Minor formatting variations (e.g., "Siemens Ltd." vs. "Siemens") were accepted if they referred to the same entity. However, extracting a distributor instead of the manufacturer was marked incorrect.

Warranty Information and Dates – The extracted warranty details had to be complete and accurate. Any missing or additional terms beyond what was in the manual were marked incorrect. Partial extractions that omitted key conditions, such as coverage exclusions or expiration dates, were also considered errors.

Assembly Instructions – The extracted instructions were checked for completeness. Minor wording changes were accepted if all key steps were retained. Since some O&M manuals contain lengthy instructions, GPT-4o occasionally summarized steps instead of reproducing them verbatim. As long as all major steps were present and no critical details were omitted, the extraction was considered correct.

An expert in construction data analysis and building operations independently evaluated GPT-4o's extractions against the benchmark. Each extraction was scored using a binary system, with correct results assigned 1 and incorrect ones 0. This method was chosen because the evaluation criteria only allowed for fully correct or incorrect results, with minor formatting differences already accounted for. The binary scoring also simplified statistical analysis and provided a clear basis for comparing extraction accuracy across different data fields. The results are shown in Table 3.

Table 3: Quantitative Results

Column Name	Total Instances	Correct Instances	Accuracy Rate
Component Name	60	60	100%
Component Type	60	60	100%
Element Type	60	56	93%
Amount and Unit of Measurement	60	42	70%
Location in Building	60	52	87%
Date of Installation	60	60	100%
Manufacturer Name	60	58	97%
Warranty Information and Dates	60	58	97%
Assembly Instructions	60	56	93%

Qualitative Results

Qualitative assessment of extraction quality was conducted to ensure usefulness for retrofit applications:

Component Name - GPT-4o consistently extracted component names accurately, with only minor formatting differences that did not affect meaning. For example, "Grundfos CM3-5 Pump" was extracted as "CM3-5 Booster Pump," but the core identification remained unchanged. No significant errors were observed in this category.

Component Type – Most extractions aligned with industry terminology, and minor misclassifications were generally acceptable. For example, "Fire Alarm Control Panel" was sometimes classified as "Electrical Equipment" instead of "Fire and Lightning Protection System." While this classification lacked specificity, it still placed the component in a relevant category and did not significantly affect usability in retrofit assessments.

Element Type – Extractions were expected to strictly follow NRM classifications, but some components were misclassified in ways that could impact retrofit planning. For example, a plumbing valve was sometimes categorized under "Mechanical Services" instead of "Water Distribution Systems", which is the correct classification. Unlike component type variations, errors in element type affect how building systems are mapped and managed, making them unacceptable for structured data applications.

Amount and Unit of Measurement - Numerical values were sometimes extracted incorrectly. In some cases, technical specifications were mistaken for quantities, such as extracting "500 mm pipe diameter" as "500 units." Additionally, when multiple components were listed, GPT-4o sometimes miscalculated totals. For example, for "Air Filters: 6 per floor, across 4 floors," the model extracted "6 units" instead of "24 units." This could lead to incorrect material estimates during procurement.

Location in Building - Extraction was generally accurate when location details were explicitly stated but often incorrect when the information was missing. If a document mentioned "HVAC ducts are installed on all levels," GPT-4 inferred specific floors and listed "Basement, First Floor, Second Floor, Third Floor," even though the document did not specify this. Similarly, if a document mentioned "Main Plant Room," the model added extra details like "adjacent to Boiler Room," which were not present in the original text. These additions could cause errors in site assessments.

Date of Installation - Installation dates were mostly accurate, but when multiple dates were listed (e.g., commissioning vs. last maintenance), GPT-4o sometimes extracted the wrong one. For example, for "Installed: May 2018, Serviced: Jan 2023," the model occasionally returned "Jan 2023" as the installation date instead of "May 2018." This could lead to incorrect assessments of component lifespan.

Manufacturer Name - GPT-4o correctly identified most manufacturers but sometimes confused them with distributors. For example, in "Supplied by XYZ Building Supplies, Manufactured by Schneider Electric," the model extracted "XYZ Building Supplies" instead of "Schneider Electric." This could result in errors when sourcing replacement parts or warranty claims. It is plausible that the addition of a routine to establish whether a company is an OEM (original equipment manufacturer) or distributor could mitigate this.

Warranty Information and Dates - When warranty details were missing, GPT-4o sometimes generated a duration that was not in the original document. For example, if the text stated, "warranty details available upon request," the model sometimes extracted "5-year warranty (2018-2023)" without supporting text. When both product and installation warranties were listed, GPT-4o occasionally combined them (e.g., "2-year product warranty, 1-year installation warranty" was extracted as "3-year total warranty"), which could lead to incorrect warranty claims.

Assembly Instructions - GPT-4o often summarized procedural steps, which was generally acceptable. However, some summaries omitted critical details. For example, "Tighten screws to 30 N m torque using a calibrated torque wrench" was sometimes extracted as "Tighten screws securely," removing an essential technical specification. Missing such details could result in improper installation.

Follow-up questions - GPT-4o could extract relevant details when the information was clearly stated in the manuals. When key details were missing or vaguely described, the model sometimes returned incomplete answers instead of indicating that the information was unavailable. In less structured documents, locating specific details required additional manual review.

Page location - GPT-4o performed well with searchable text but sometimes mismatched printed and digital page numbers. For example, warranty information listed as Page 15 in the manual was extracted as Page 17 due to unnumbered cover pages. Older PDFs or scanned files with unrecognized text also affected accuracy, requiring manual verification in such cases.

Conclusion

This study evaluated GPT-4o's ability to extract structured data from O&M manuals for building retrofits. The results show that the model performs well in text-based extraction, particularly for component names, types, and manufacturer details, where formatting differences had little impact on accuracy. However, challenges appeared in fields requiring categorization, numerical interpretation, and structured mapping.

One issue was category misalignment, especially in the classification of Element Type based on NRM standards. GPT-4o could assign components to broad categories but had difficulty with domain-specific classifications that require specialized knowledge. This inconsistency suggests that rule-based constraints or further adjustments may be needed to improve standardization in retrofit applications. Another challenge was the model's handling of numerical data, which had the highest error rate. Mistakes occurred in quantity extraction, page numbering, and location references, where the model miscalculated totals, misread measurement units, or mismatched printed and digital page numbers. These errors suggest that GPT-4o lacks a way to process numerical information in context, meaning manual checks are still required for numerical fields. This study adopts NRM for classification, though alternative systems such as Uniclass and ISO 12006-2 offer more detailed and hierarchical structures. These may support improved interpretation by language models and could be considered in future work.

Even with these limitations, GPT-4o provides a way to reduce the effort needed for O&M data extraction. Text-based retrieval is reliable, but structured classification and numerical accuracy need improvement. Future work could focus on refining category mapping, adding rule-based validation, and improving numerical extraction to make the process more consistent and accurate for retrofit applications.

Acknowledgement

This research was supported by Innovate UK through their Bridge AI fuding programme via the BuildAudit project. We would also like to acknowledge BRE and Reusefully for their insights and guidance in the development of this study.

Nonmenclature

GPT-4o – Large Language Model used for O&M data extraction

LLM – Large Language Model

O&M – Operations and Maintenance

OEM – Original Equipment Manufacturer

PDF – Portable Document Format

NRM – New Rules of Measurement

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Bommasani, R., Dai, A. M., Liang, P., et al. (2021). On the opportunities and risks of foundation models. Stanford University Report.

Building Services Research and Information Association (BSRIA). (2016). O&M Manuals Best Practice Guide. BSRIA. Bracknell (UK).

Chartered Institution of Building Services Engineers (CIBSE). (2020). Guide M: Maintenance Engineering and Management. CIBSE. London (UK).

Coze. (2024). *Coze AI Development Platform.* Retrieved from https://coze.com

Greater London Authority. (2021). The London Plan 2021: The Spatial Development Strategy for Greater London.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP. Proceedings from Advances in Neural Information Processing Systems (NeurIPS).

Ma, Z., Cooper, P., Daly, D., Ledo, L. (2012). Existing building retrofits: Methodology and state-of-the-art. Energy and Buildings 55, 889–902.

OpenAI. (n.d.). Prompt Engineering Guide. Retrieved from https://platform.openai.com/docs/guides/prompt-engineering.

Royal Institution of Chartered Surveyors (RICS). (2020). NRM - New Rules of Measurement: Order of Cost Estimating and Cost Planning for Capital Works (NRM1). RICS. London (UK).