Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating the Use of LLM-based Multi-Agent Systems for Internal Information Sharing and Access Within Manufacturing

Ngoc Bach Nguyen

nbn26@cantab.ac.uk

University of Cambridge

Sam Brooks

Zhengyang Ling

Research Article

Keywords: Large Language Models, Multi-Agent Sys- tems, Internal Information Sharing and Access, Smart Manufacturing, Retrieval-Augmented Generation, Data Quality

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9860546/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at The International Journal of Advanced Manufacturing Technology on November 6th, 2026. See the published version at https://doi.org/10.1007/s00170-026-19467-7.

The International Journal of Advanced Manufacturing Technology manuscript No.
(will be inserted by the editor)

Evaluating the Use of LLM-based Multi-Agent Systems for Internal Information Sharing and Access Within Manufacturing

Ngoc Bach Nguyen · Sam Brooks · Zhengyang Ling

Received: date / Accepted: date

Abstract Internal information sharing in manufacturing remains constrained by data silos, heterogeneous systems, poor data quality, and limited interoperability across ISA-95 levels. While Large Language Models (LLMs) are increasingly proposed for manufacturing information access and for integrating information with operational technology, there is limited empirical evidence comparing their performance with unassisted human users on representative manufacturing tasks.

This paper evaluates a modular LLM-based multi-agent system for retrieving and synthesising information across heterogeneous data sources from a 3D-printing factory demonstrator. The system was benchmarked against human participants on traceability and certification processing tasks involving packing lists, relationship-tracking records, machine logs, location-tracking data, and Authorised Release Certificates. Three task types under four data-quality conditions were tested, including: unaltered baseline data, whitespace injection, truncated identifiers, and severed relational links.

Under the bounded experimental conditions, the highest-performing LLMs achieved higher mean accuracy and shorter completion times than the average unassisted participant, with Claude Sonnet-4 achieving 94.1% mean accuracy compared with 62.2% for humans. However, this advantage was conditional: some smaller models fell below the human average, and the best human participants outperformed all models. The system was robust to whitespace errors and severed relational links, but vulnerable to truncated identifiers and verbose prompt scaffolding. These findings indicate that multi-agent LLM systems can support bounded manufacturing information-access tasks as a decision-support layer, but require careful model selection, data validation, prompt design, and human oversight.

Keywords Large Language Models · Multi-Agent Systems · Internal Information Sharing and Access · Smart Manufacturing · Retrieval-Augmented Generation · Data Quality

N.B. Nguyen · S. Brooks · Z. Ling
Institute for Manufacturing, Department of Engineering, University of Cambridge, Alan Reece Building, 17 Charles Babbage Road, Cambridge CB3 0FS, UK
E-mail:
nbn26@cantab.ac.uk, sjb351@cam.ac.uk, zl461@cam.ac.uk

1 Introduction

Modern manufacturing enterprises confront escalating complexity driven by rapid technological advances, disrupted global supply chains, and increasing international competition [1]. Seamless internal information sharing within manufacturing companies is a key enabler of Industry 4.0 and Industry 5.0. This requires significant internal integration between digital systems, which is severely impeded by organisational data silos and heterogeneous digital systems [2].

The ISA-95 functional hierarchy (ANSI/ISA-95.00.01; IEC 62264-1) provides the canonical reference frame which describes the different levels of information within a manufacturing business. ISA-95 partitions manufacturing operations into five levels, each with characteristic systems, data types, and operating time-scales (Figure 1 shows a reproduced example) [3].

Level 0 focuses on the physical production process: machines and material flow. Level 1 is sensing and actuation, with PLCs, barcode readers, and machine outputs used to collect data from operations. Level 2 is the supervisory control with SCADA and HMI systems often sitting here, as well as higher-level multi-machine control. These Operational Technology (OT) layers (levels

0, 1, and 2) are often well integrated in modern factories, and many studies focus on collecting or extracting data from machines for control [4, 5].

Level 3 focuses on manufacturing operations management, which often includes functions such as quality assurance, work order scheduling, compliance, and certification documentation. Above this at Level 4 sit enterprise-level functions and data, such as business planning, financial modelling and supply chain planning. These are referred to as the Information Technology (IT) layers, and the most persistent integration challenge lies between the IT and OT layers.

Information can be shared or transferred down or up levels of ISA-95. For example, companies want to share machine and tracking data collected on the shop floor (levels 1 and 2) for use in MES software (level 3) [6]. Alternatively, there is a need to relay key measurements from order information in ERP software (Level 4) to shop floor operators setting up machines [7]. At the core of these integration challenges is the need to process and synthesise information across the OT/IT divides [8].

Changes to architecture [9] or standardised interoperability connectivity layers [10] have been proposed as possible solutions, but these require extensive changes to existing software and systems. Furthermore, *data quality* variations, such as extra characters or scanning errors, regularly corrupt these records [11, 12]; these approaches do not address this issue, and further work is needed to clean or preprocess the data before it is useful internally.

Large Language Models (LLMs) have emerged as a candidate solution for processing natural language and unstructured data [1], but there is a pronounced lack of empirical research quantitatively evaluating integrated LLM systems against unassisted human users on representative manufacturing information-access tasks [13, 14]. This study addresses that gap by empirically evaluating a multi-agent LLM-based system against an initial unassisted human baseline. The study aims to address the following research questions:

Central Research Question: How effectively can a multi-agent LLM-based system assist in retrieving and synthesising data and information from systems on different ISA 95 levels?

Sub-Question: How do specific data quality conditions and task complexity affect the performance of both LLM-aided and non-aided tasks?

The remainder of this paper is structured as follows. Section 2 reviews internal information sharing barriers, and the current state of LLM applications within manufacturing and this space. Section 3 presents the multi-agent system architecture and the experimental design. Section 4 reports the comparative performance findings across task complexity and data quality conditions. Section 5 discusses results, deployment implications and limitations, with Section 6 summarising future work.

2 Literature Review

This section reviews the literature on internal information-sharing barriers in manufacturing and on current LLM applications and system architectures, and identifies the specific research gaps.

2.1 Internal Information Sharing and Access Barriers in Manufacturing

Achieving seamless internal information flow across the factory is fundamental to the agility and responsiveness required for Smart Manufacturing [15]. Many companies have moved away from paper-based forms or information already in manufacturing; however, three persistent categories of barriers impede this flow: interoperability, data quality or context, and information silos.

Interoperability is the most structurally entrenched barrier. Manufacturing environments comprise heterogeneous mixes of automation equipment, ERP, MES, and PLC systems that employ incompatible data formats and lack standardised interfaces. For example, MES software may lack the necessary APIs to receive raw PLC data that has not been processed into the required format. This forces manual data re-entry and reliance on third-party middleware, such as Kepware or OPC Router, which bridges PLC tag data into MES/ERP schemas through hand-configured mappings [2, 16]. However, these configurations are brittle in the sense that each new sensor, format change, or downstream consumer typically requires manual remapping; they cannot interpret unstructured artefacts (PDFs, operator notes, free-text logs) and must be re-engineered when the underlying schema shifts. Substantively identical findings have emerged across more than three decades of literature [2, 8, 16–20], reflecting the intractability of this class of problem rather than a lack of awareness or effort.

Data quality or a lack of data context presents a second critical challenge. Poor quality and inconsistency are prevalent across lots of different types of manufacturing data [2, 11, 21]. Poor data quality can manifest in many ways. Common factors considered include timeliness (is the data there when needed), accuracy (does the data have errors), consistency (is the data in the same format or structure each time), and completeness (is there data missing) [11, 22]. Lastly, data must retain



Fig. 1 ISA-95 functional hierarchy (ANSI/ISA-95.00.01; IEC 62264-1). The standard partitions manufacturing operations into five levels, each with its own characteristic systems, data types, and time scale. The diagram highlights the IT/OT barrier in orange, which this study focuses on.

their intended meaning; for example, a stream of numbers without context for what they relate to has no use [16, 17].

Informational silos form a third major barrier, which can be caused by technical or organisational structures. Functional departments develop domain-specific vocabularies and optimise local performance without adequate regard for system-wide integration, with information typically handed off in fixed batches between functions such as design, maintenance and production rather than shared as a continuous, queryable record [17]. This produces chronic information silos between planners, designers, and shop-floor operators, and has been shown to have direct, measurable negative effects on both operational and business performance [17, 19].

2.2 LLMs and Agentic Architectures in Manufacturing

Advances in transformer-based Large Language Models have opened new avenues for addressing common data challenges that traditional software cannot resolve [15, 23]. Three architectural patterns now dominate the literature: standard Retrieval-Augmented Generation (RAG), agentic and multi-agent frameworks, and grounded hybrid architectures (knowledge-graph or cognitive-architecture overlays). Table 1 groups the identified studies by these three architectural groups and by the ISA-95 levels they primarily address.

RAG grounds LLM outputs in curated domain documents to reduce hallucination and improve factual consistency [24, 25]. Kernan Freire et al. [24] deploy RAG over factory manuals and SOPs and evaluate via user perception surveys at the operations-management level. Wellsandt et al. [26] extend this pattern to a conversational assistant over internal documents. Liu et al. [25] combine RAG with LLM-based named entity recognition for the additive-manufacturing process and to answer questions about documents. In all these studies, the document set is known and indexed in advance of queries to the LLM, meaning new documents or data need to be integrated before use. Evaluations are also focused on metrics such as precision and recall or on expert scores, rather than on comparisons with human operators.

Generic foundation LLM models lack the specialised vocabulary and procedural knowledge of a given factory, and are prone to confidently hallucinating plausible-sounding but incorrect outputs in response to domain-specific queries. A second cluster of architectures explicitly addresses this by constraining LLM reasoning to a curated ontology represented as a knowledge graph. Ma

et al. [27] propose a fault-diagnostic pipeline in which a knowledge graph constrains LLM reasoning to verifiable maintenance decisions, mitigating hallucination and black-box deployments. Fan et al. [28] apply the same grounding pattern in metal additive manufacturing.

A third cluster of recent work in manufacturing shifts from passive retrieval to *agentic* orchestration, in which LLMs leverage zero-shot and few-shot reasoning to drive multi-step autonomous workflows [29, 30]. Zhao et al. [31] demonstrate a five-module agent architecture (Machine Server, Bid Inviter, Bidder, Thinking, Decision agent) in which LLM-driven agents negotiate task allocation across shop-floor machines and reduce makespan compared with classical heuristics. Huang et al. [32] apply locally hosted LLMs for scheduling in human-robot collaborative flexible manufacturing systems. Ni et al. [33] transform user descriptions into structured manufacturing task flows via task splitting, step generation, and holistic process synthesis. Elbasheer et al. [34] use a similar pattern to enable conversational generation of executable simulation models. The agentic paradigm extends upstream into engineering design as well: Guo et al. [35] position the LLM as an autonomous collaborator that plans, executes, and adapts computer-aided engineering (CAE) workflows. These works collectively begin to push LLM use across both upper enterprise, shop floor and manufacturing execution layers, but are often constrained to one layer. Evaluations against a domain-internal yardstick (makespan, schedule quality, simulation fidelity, solver fidelity) rather than against human performance.

From Table 1, there are two studies which cross L2 to L3 [31, 32]. Both studies use multi-agent LLM systems to process shop-floor data as input for agent-based scheduling. However, challenges like interoperability, poor data quality, and information silos are not accounted for. Other studies predominantly focus on processing manufacturing operations documents at L3 or planning and ordering documents at L4. Secondly, the table highlights a lack of comparison between human operators and LLM systems.

2.3 Identification of Research Gaps

These three key gaps are noted in the literature, which directly motivate this study:

Gap 1: Neglect of lower ISA-95 levels. Research has concentrated on Levels 3–4, with fewer studies looking at data from L2 to higher levels (L3 and 4). Communication within lower levels (L1–L2), where raw, unprocessed shop-floor data reside, also remains underexplored [2, 16].

Gap 2: Lack of empirical human benchmarking. Quantitative evaluations comparing the performance of integrated LLM systems against unassisted human baselines in representative manufacturing workflows are absent. Existing studies either evaluate isolated Natural Language Processing (NLP) components or rely on qualitative perception metrics, stopping short of operationally grounded, head-to-head performance comparisons [14, 24, 36].

Gap 3: Data quality errors are untested. Manufacturing data is notoriously imperfect [11]. Yet no study reviewed above empirically evaluates LLM system performance under variable data quality conditions. The quantitative impact of data corruption on the accuracy of automated retrieval and synthesis is therefore unknown, thereby constraining confidence in the viability of real-world deployment [12].

This study targets these three gaps by deploying a multi-agent LLM system on heterogeneous, cross-level manufacturing data (Levels 1–3) and benchmarking it against an unassisted human-participant baseline across four controlled data quality conditions and tasks.

3 Methodology

This section outlines the methodology used in the study, including the demonstrator used to collect data, the developed multi-agent LLM system, and the experimental process.

3.1 Demonstrator Factory Operations

The Variable Organisation and Operations Management (VOOM) Smart Factory Demonstrator is a research-grade demonstrator of a discrete-manufacturing production line, designed to reproduce multiple interacting smart factory sites [37]. Data collected over one day of operation at the 3D-printing factory, producing 100 parts from ten 3D printers, was used. Figure 2 shows an overview of the factory facilities and the different stations.

The factory had a basic manufacturing execution system (MES) for handling and scheduling orders on the shop floor. Alongside this, tracking systems monitored when parts or items entered a location and which items were added or linked to others, using barcode scanners. Other digital solutions for monitoring, power use, operator activities, and quality issues were also available, but these were not used in this study.

Figure 3 shows a Business Process Modelling Notation (BPMN) diagram for the 3D printing factories' operations, information systems, and flow. The various

Table 1 Representative LLM System Architectures in Manufacturing Research, grouped by architectural pattern. Within each pattern, the ISA-95 level addressed and the evaluation approach are recorded; the final row situates the present study against the cohort.

Architecture used	Reference	Data Type Used	Evaluation Method	Related ISA-95 Level	Human comparison
Standard RAG / fixed-corpus retrieval	[24]	Manuals, SOPs (unstructured)	User perception survey	L3	Partial (qualitative)
	[26]	Internal documents	Conceptual / expert review	L3–L4	No
	[25]	AM process literature (text)	Precision/recall, F1	L3	No
	[14]	Structured process data	User acceptance study	L3–L4	Qualitative only
Agentic / multi-agent	[31]	Machine state, job descriptors	Makespan vs. classical heuristics	L2–L3	No
	[32]	Schedule + disruption events	Schedule quality benchmark	L2–L3	No
	[33]	Natural-language task descriptions	Process-flow conformity	L3–L4	No
	[34]	NL planning specifications	Simulation-model fidelity	L3–L4	No
	[35]	Parametric PDEs, CAE specs	Solver fidelity / reduced-order model accuracy	L3	No
Grounded hybrid (KG / cognitive)	[27]	Maintenance logs + KG	Diagnostic accuracy vs. KG ground truth	L3	No
	[21]	Maintenance logs (unstructured)	Precision/recall vs. annotators	L3	Annotator comparison
	[36]	Quality/defect reports	F1-score benchmark	L3	No
	[28]	AM literature + KG, MR interface	Retrieval quality vs. vector/graph baselines	L3	No

tasks and collected information are broken down into ISA-95 levels. Within this study, there are five key types of data collected by the factory used:

1. **Location Tracking Data:** This data is collected by scanners and software on the shop floor at set locations or states in production. Data is stored in an SQLite database and exported to CSV files for the study.
2. **Relationship Tracking Data:** This data is collected by scanners and software on the shop floor and records when two items are linked, such as material onto a printer or product gears into an order. Data is stored in an SQLite database and exported to CSV files for the study.
3. **Packing list documents:** When an order is created, the MES generates a packing list that includes the order number and the quantity of parts to pack in the final order. These are stored as PDFs.
4. **Machine logs for printers:** The printer API is used to extract data on the parts produced and job completion times. Data is stored in an SQL database but exported to CSV files for the study.
5. **Authorised Release Certificate:** This document is generated automatically by the MES upon confirmation that the part has been made by shop floor operators. These are stored as PDFs.

This data was chosen because it spans different levels of the ISA-95 structure, and currently, these sources of information are all kept separately in different software or storage systems. The formats also vary depending on the software used to collect the data.

3.2 Multi-Agent LLM System Architecture

To process heterogeneous manufacturing data from different systems, a modular multi-agent system was developed based on the LangChain framework [38]. Rather than a monolithic RAG implementation, the system employs *adaptive* RAG combined with specialised tool usage, building on prior adaptive retrieval work [39, 40]. Adaptive RAG allows the system to determine dynamically which data sources to traverse in response to a given query, without pre-loading experimental data prior

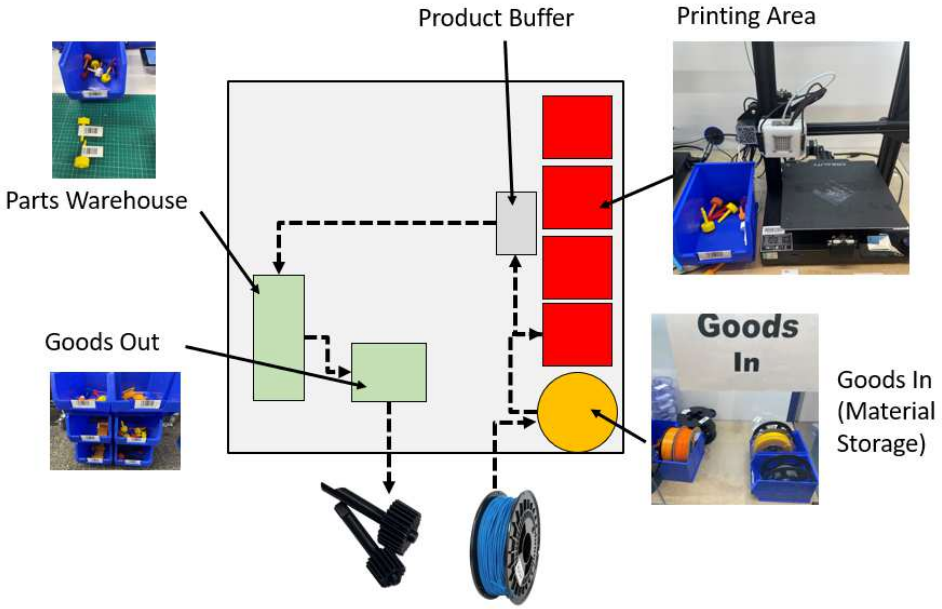


Fig. 2 Diagram showing a plan view of the VOOM 3D printing factory set up.

to evaluation (as in traditional RAG). This gives the system better adaptability to new data or information. The architecture decomposes complex queries across four specialised agents, illustrated in Figure 4:

- **Master Agent:** Orchestrates the complete workflow. Parses the user query, formulates a step-by-step execution plan, delegates tasks to specialist agents, and triggers re-planning when critical issues arise.
- **Data Retrieval Agent:** Interfaces directly with the raw factory data sources via a domain-specific Tool Suite providing functions for data retrieval, entity lookup, and relationship mapping [41].
- **Reconciliation Agent:** Cross-validates retrieved information from disparate sources and resolves conflicts. Critical issues trigger iterative re-planning by the Master Agent.
- **Synthesis Agent:** Generates the final human-readable report from validated data, using role-based prompting and structured stepwise reasoning prompts to support interpretable outputs [42].

The modular design allowed the underlying foundation model to be substituted without architectural changes, enabling systematic benchmarking across six LLMs from three providers: **Anthropic** (Claude Sonnet-4, Claude Haiku), **DeepSeek** (DeepSeek Reasoner, DeepSeek Chat), and **OpenAI** (O4-Mini, GPT-4o Mini). This selection spans standard and advanced reasoning models available at the time of the study in 2025.

3.3 Experimental Process

3.3.1 Tasks and Complexity

The core experiment of this study focused on getting humans and the agent-based LLM system to complete tasks using the data and information collected in VOOM. Three different complexity tasks were designed in collaboration with the 3D Printing lab operators, designed to replicate the kind of questions a factory manager might want to answer with their data. Task complexity is defined in this study by the number of discrete look-up steps an operator must perform across the heterogeneous data sources to assemble the answer. The tasks were:

1. **E-Tasks (3 operator steps): Information retrieval about parts IDs.** *"For Packing List `[ID]`, identify all associated gears."* Steps: (1) Locate the PDF packing list and extract the Order ID; (2) filter a relational CSV for that Order ID; (3) list the associated part identifiers.
2. **M-Tasks (4 operator steps): Trace other parts made of a machine** *"For Part ID `[ID]`, identify the 3D printer that produced it, and count the total number of parts printed on that machine today."* Steps: (1) Locate the part in the relational CSV; (2) read off its printer assignment; (3) query the machine-log CSV with that printer ID; (4) count rows.
3. **H-Tasks (6 operator steps): Complex data verification.** *"For Order ID `[ID]`, verify that the completion date on the Authorised Release Certifi-*

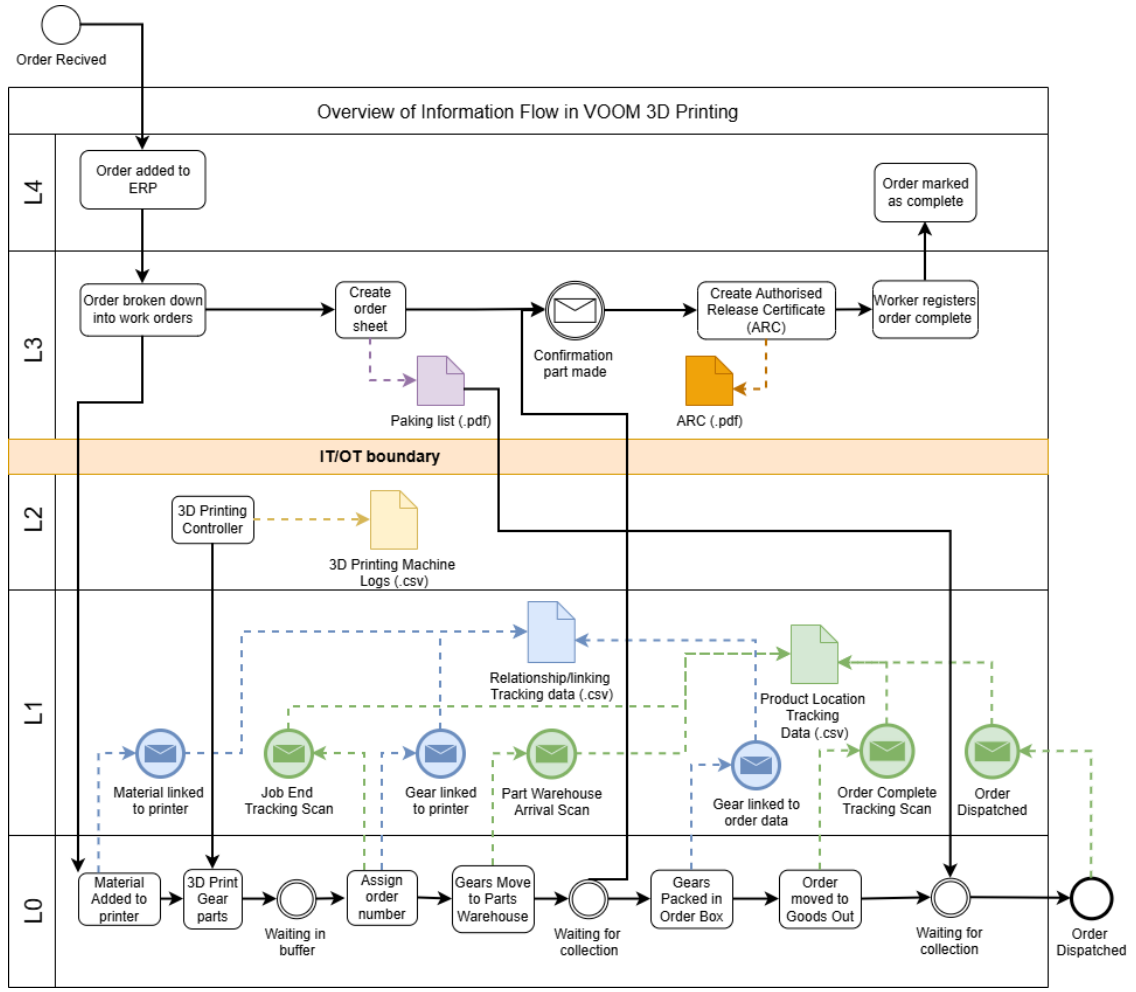


Fig. 3 BPMN model of the information flow and operations occurring in the 3D printing factory. Activities are broken down into ISA-95 levels. Information used in this study is shown in different colours, including the machine logs (yellow), ARCs (orange), packing list (purple), relationship or linking tracking data (blue) and location tracking system (green).

cate matches the date the order arrived at the Parts Warehouse." Steps: (1) locate the corresponding ARC PDF; (2) extract the certified completion date; (3) locate the order in the location-tracking CSV; (4) filter for the warehouse-arrival event; (5) extract the arrival timestamp; (6) compare the two dates.

Performance in each task was recorded on two primary metrics: **binary accuracy** (correct/incorrect, scored algorithmically against the script-generated reference answer) and **task completion time** (recorded to the nearest second by the researcher using a stopwatch or outputted from the LLM API interface).

3.3.2 Data and Quality Conditions

Common data quality issues were simulated by adding random errors to the original dataset. Four different data conditions are created:

- **Q0 (Baseline):** Uncorrupted operating data; the reference condition.
- **Q1 (Whitespace Injection):** inserts random space characters into barcode identifiers (e.g., `3DOR100081` → `3D OR100081`). This mimics scanning or data-entry errors; as a result, an exact string match against a clean record fails.
- **Q2 (Truncated Identifier):** removes characters from identifiers (e.g., `3DOR100081` → `3DR100081`). This mimics data truncation, which could happen in transit between systems.
- **Q3 (Severed Relational Link):** Removes one direction of a bilateral parent–child relational mapping. The other direction of the link still exists, so the underlying record is not strictly absent, but the forward-direction lookup that the operator would normally use returns no results, forcing an indirect, reverse-direction inference.

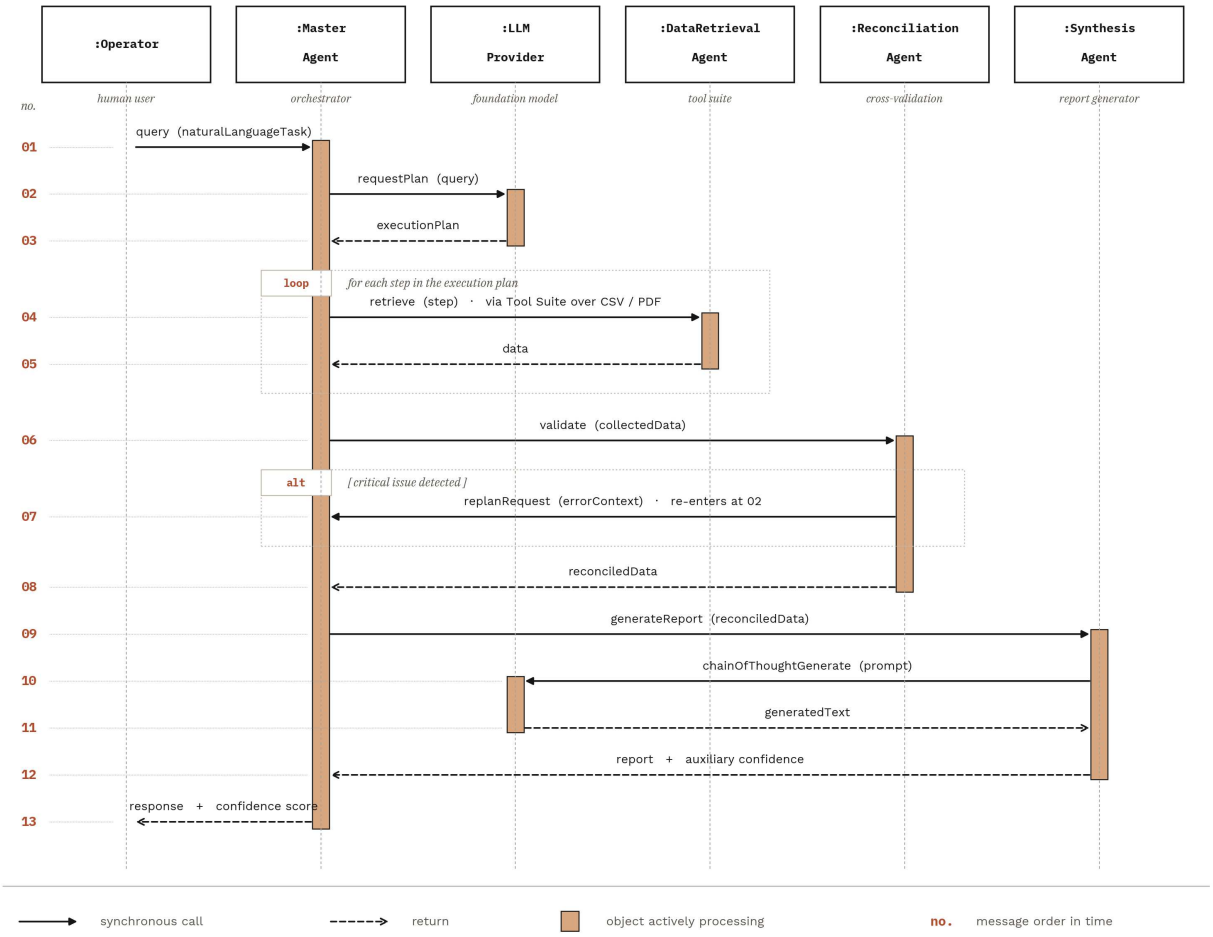


Fig. 4 Multi-agent LLM system architecture as a UML sequence diagram. Lifeline heads at the top are objects only: `:Operator`, `:MasterAgent`, `:LLMProvider`, `:DataRetrievalAgent`, `:ReconciliationAgent`, `:SynthesisAgent`. The *loop* frame covers the iterative retrieval over each step in the execution plan; the *alt* frame shows the re-planning path that re-enters at message 02 when the Reconciliation Agent detects a critical issue.

It should be noted that all errors (Q1, Q2 and Q3). Would cause issues and disrupt a fixed deterministic system or pipeline currently used in many digital systems. For example, if MES and an ERP system were sharing data with incorrect ID, white spaces or missing links, this would prevent accurate answers or processing within the software.

3.3.3 Human Participant Study

The human-participant study was conducted with nine participants ($n = 9$), recruited from the engineering student population at the University of Cambridge's Department of Engineering, with general technical literacy but no prior operational experience with the VOOM environment or its data sources. Before each session, every participant received a short standardised briefing of approximately 15 minutes covering VOOM's operations and the data collected; the format of the data and the contents of each data source; the file-system structure for locating the data; and a worked example of one task per complexity tier. Participants used only standard desktop tools with a search function (e.g., file explorer, Excel spreadsheet viewer, PDF viewer, text search) to perform the tasks; no specialised middleware or query tools were provided.

Before the main study, a pilot test with two non-participant volunteers was used to confirm that the task wording was clear and assignments could be completed even with data errors added.

A **counterbalanced design** was used to ensure each participant had a variety across task type and data quality condition [43]. Each participant completed 10 tasks; the order of task type (E/M/H) and data quality condition (Q0–Q3) was rotated across participants in three groups of three, so that no single position in the

sequence was disproportionately associated with any type or condition. Across the 90 participant-tasks, the allocation was 45 tasks under Q0 (Baseline with no error) and 15 tasks each under Q1, Q2, and Q3, and 30 tasks each of E-, M-, and H-types. The counterbalancing equalises type and condition independently, although their interaction is not fully crossed (addressed in Section 5.1).

3.3.4 LLM Implementation

The multi-agent system subsequently executed the same 90 tasks as the human participants, using standardised short prompts (Table 2), to keep the comparison with the human-participant study as consistent as possible. To support reproducibility and account for stochastic variation in LLM outputs, each task was executed three times per model. The raw binary outcomes from these repetitions were retained for inferential modelling, while per-task means across the three runs were used for descriptive tables and figures. This yielded 270 evaluations per model (90 tasks × 3 repetitions). Extending this across all six LLMs produced 1,620 LLM evaluations in total.

To assess the system's sensitivity to prompt scaffolding, a secondary experiment tested the three top-performing models (Claude Haiku, Claude Sonnet-4, DeepSeek Reasoner) under three prompt-length conditions, in which the task being asked was kept identical and only the surrounding descriptive scaffolding was expanded:

- **Short** – the prompt template used in the main study: a one-line task statement followed by three numbered procedural steps and no surrounding context (36–67 tokens across the E/M/H task types and the base/data-quality variants).
- **Medium** – the same task wording, wrapped in a labelled `CONTEXT` sentence, a longer `REQUIRED STEPS` list (4–7 numbered items depending on task type and variant), and an explicit `OUTPUT FORMAT` specification (104–195 tokens; Medium/Short ratio 2.4–3.4×, mean 2.9×).
- **Long** – the Medium scaffolding plus two further labelled sections (`DATA SOURCES AVAILABLE` listing the five VOOM trails, and `VALIDATION REQUIREMENTS` or `SUCCESS CRITERIA` listing cross-source checks) (138–371 tokens; Long/Short ratio 2.1–6.9×, mean 4.8×, with the upper end driven by an addition of explicit data-source enumeration).

Each condition preserves the underlying lookup required by the question and the correct answer; the additional length comes from the labelled `CONTEXT`, the expanded `REQUIRED STEPS` list, and the cross-validation requirements introduced in Medium and Long. The full Medium and Long prompt templates for each task type are provided in Appendix A (Tables 8 and 9). Each model was tested on nine representative tasks per prompt-length condition, spanning all three task types (E/M/H) and data-quality conditions; Short-condition results for these nine tasks were reused from the main study; only the Medium and Long conditions were newly executed. This sub-study is intended as an indicative probe of prompt-length sensitivity on the multi-agent system rather than a confirmatory result.

3.3.5 Statistical Analysis

Accuracy was recorded as a binary outcome for each task attempt (correct/incorrect), and completion time was recorded in seconds. Welch's t-tests [44], Mann–Whitney U-tests [45], and one-way ANOVA [46] are retained as supplementary descriptive comparisons because they are readily interpretable and allow comparison with conventional benchmarking studies. However, these tests should be interpreted with caution because they do not fully represent the repeated task–model and participant structures. Effect sizes are reported alongside significance: Cohen's d for pairwise comparisons and η^2 for ANOVA-style summaries [47], with the conventional small/medium/large thresholds of 0.01/0.06/0.14 for η^2 and 0.2/0.5/0.8 for d.

Completion time was strongly right-skewed for several LLMs, especially where a small number of long-running responses inflated the mean. For this reason, both mean and median completion times are reported. Pairwise comparisons of time used Mann–Whitney U-tests [45] alongside Welch's t-tests [44].

API token consumption and associated costs were tracked per model to support the cost analysis in Section 5.

4 Results

This section reports the comparative findings across three task complexity levels (E, M, H) and four data quality conditions (Q0–Q3). Section 4.1 reports overall accuracy and completion time and sets out the per-model results. Sections 4.2 and 4.3 then report the effect of task complexity and data quality, respectively, and Section 4.4 reports the prompt-length sub-study.

In this section, t denotes the Welch's t-test statistic, d denotes Cohen's effect size, and U denotes the Mann–Whitney U statistic. Statistical significance is indicated by p values, with $p < 0.05$ treated as significant.

Table 2 Short Prompt Templates by Task Type

Type	Prompt Template
E	Find all gears for Packing List {packing_list_id}. Steps: (1) Get Order ID from the packing list. (2) Find all gears for that order. (3) List the gears.
M	For Part {part_id}, find which printer was used and count the total parts on that printer. Steps: (1) Find the printer for this part. (2) Count all parts printed on that printer. (3) Report printer and count.
H	For Order {order_id}, check if the ARC document date matches the warehouse arrival date. Steps: (1) Find the ARC document date. (2) Find warehouse arrival date. (3) Compare dates and report match status.

4.1 Overall Accuracy and Completion Time

Under the study conditions, the top-performing LLM models achieved higher accuracy and shorter completion times than the average unassisted human participant. Per-model figures are summarised in Table 3, and average model performance versus human performance is shown in Figure 5. The mean accuracy across all six LLM models was 80.2%, compared with 62.2% for human participants, a difference of 18.0 percentage points. A supplementary Welch comparison on the per-(task, model) means indicates that this difference is statistically significant (Welch's $t(103.1) = 3.37$, $p = 0.001$, Cohen's $d = 0.50$, medium effect; $n_{\text{LLM}} = 540$ vs. $n_{\text{human}} = 90$). The result does not imply that all LLM configurations outperformed all humans. GPT-4o Mini achieved 50.7% accuracy, below the human average, and the best human participant achieved 100% accuracy, exceeding the best LLM model in this benchmark (DeepSeek Reasoner at 97.4% and Claude Sonnet-4 at 94.1%).

LLMs were also faster overall. The pooled-mean and pooled-median completion times tell complementary stories: the pooled mean across the raw 1,620 LLM observations is 52 s, compared with 166 s for the 90 human observations (supplementary Welch's $t(96.6) = -11.66$, $p < 0.001$); the corresponding pooled medians are 14 s versus 152 s (Mann-Whitney $U = 19{,}090$, $p < 0.001$). Both statistics are reported because the LLM time distributions are right-skewed (Table 3); the largest mean-vs-median gap is GPT-4o Mini (88 s mean vs. 9 s median), driven by a small number of long-running outliers that inflate the mean.

The wide range of values for human participants in Figure 5 warrants closer examination. Figure 6 plots each participant's mean accuracy against mean completion time, revealing substantial individual differences: two participants (P4, P9) achieved 100% accuracy at below-average completion times, while two others (P2, P8) achieved only 30%. Participant familiarity with manual search strategies, spreadsheets, PDFs, and the task structure was not independently measured, so this variation should not be overinterpreted. It does, however, show that the human baseline was heterogeneous and that the LLM advantage is relative to the average unassisted participant rather than to every human user.

4.2 Impact of Task Complexity

Figure 7 shows the results across different task types. Human participants and several top-performing LLM models had the lowest accuracy on E-type tasks (with three operator steps) and higher accuracy on the nominally more complex M- and H-type tasks (see Figure 7a). This is the reverse of the assumption embedded in the task design that step count would be a reliable proxy for difficulty. Within this benchmark, step count alone was therefore not a reliable indicator of task difficulty. A similar counter-intuitive result is visible in the completion time in Figure 7 b), where both humans and LLMs were faster on H-tasks than on E-tasks.

Restricting the analysis to the uncorrupted Q0 subset alone preserves the pattern (E = 50.0%, M = 66.7%, H = 77.8%), so the reversal is not an artefact of the data condition.

A supplementary ANOVA-style comparison on the LLM per-task means showed that the multi-agent system held its accuracy roughly constant across task types, with no evidence of a significant performance change ($F(2, 537) = 0.118$, $p = 0.889$, $\eta^2 < 0.001$). Top models such as DeepSeek Reasoner and Claude Sonnet-4 achieved mean accuracies of 100.0% and 97.8%, respectively, on H-type tasks, demonstrating the high accuracy possible with LLMs in this bounded task setting. Per-model accuracy and completion time by task type are reported in Tables 4, 5, and 6; both metrics are also visualised in Figure 7.

Table 3 Overall accuracy and completion time compared with human participants. Models are grouped by provider family (DeepSeek, Anthropic, OpenAI). Both Mean and Median completion times are reported. Bold values mark the top three across the cohort for each of the accuracy and time metrics.

Family	Model	Accuracy	Mean Time (s)	Median Time (s)	Acc. supp. test	Time supp. p
–	Human Participants	62.2%	166	152	–	–
DeepSeek	DeepSeek Reasoner	**97.4%**	160	146	$t = 6.63^{***}$, $d = 0.99$	p=0.580 (ns)
	DeepSeek Chat	77.4%	30	24	$t = 2.42^{*}$, $d = 0.36$	p<0.001***
Anthropic	Claude Sonnet-4	**94.1%**	**14**	**14**	$t = 5.67^{***}$, $d = 0.85$	p<0.001***
	Claude Haiku	**93.0%**	**8**	**7**	$t = 5.36^{***}$, $d = 0.80$	p<0.001***
OpenAI	O4-Mini	68.5%	**10**	10	$t = 0.96$ (ns), $d = 0.14$	p<0.001***
	GPT-4o Mini	50.7%	88	**9**	$t = -1.85$ (ns), $d = -0.28$	p<0.001***

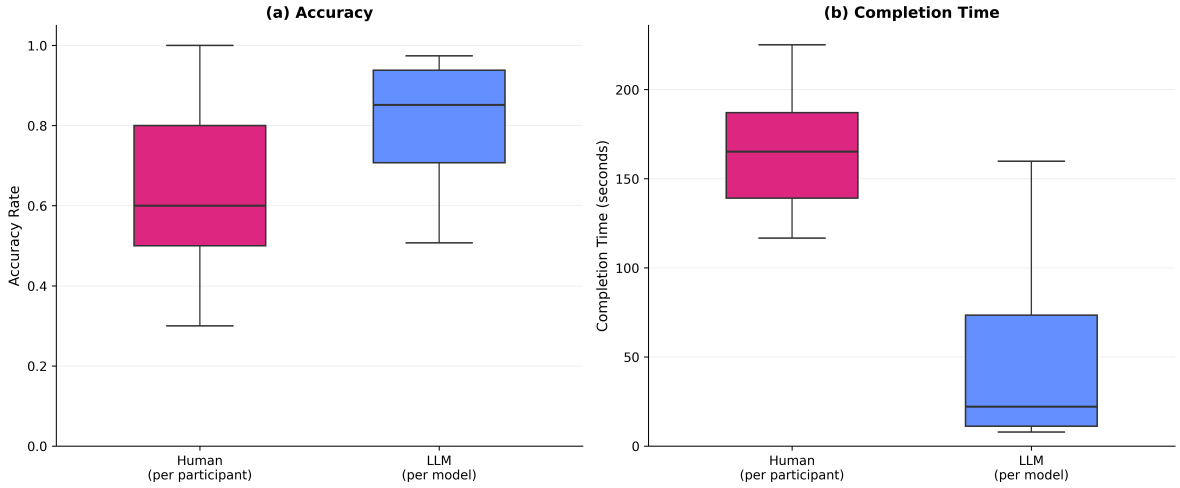


Fig. 5 Boxplot comparison of (a) accuracy and (b) completion time between human participants and LLM models.

Table 4 Mean accuracy and per-cell Mean and Median completion time by group on E-tasks (3-step). Bold values mark the top three in the LLM cohort for each metric (accuracy, descending; time, ascending). The human participant result is included for direct comparison but is not part of the LLM ranking. Rows ordered by accuracy descending within the LLM cohort.

Rank	Model	Accuracy	Mean Time (s)	Median Time (s)
–	Human Participant	43.3%	186	174
1	DeepSeek Reasoner	**93.3%**	171	147
2	O4-Mini	**86.7%**	**10**	**10**
3	Claude Sonnet-4	**85.6%**	**14**	14
4	DeepSeek Chat	84.4%	29	20
5	Claude Haiku	81.1%	**8**	**7**
6	GPT-4o Mini	54.4%	97	**9**

Table 5 Mean accuracy and per-cell Mean and Median completion time by group on M-tasks (4-step). Bold values mark the top three across the LLM cohort for each metric (accuracy descending; time ascending). The human participant result is included for direct comparison but is not part of the LLM ranking.

Rank	Model	Accuracy	Mean Time (s)	Median Time (s)
–	Human Participant	63.3%	179	182
1	DeepSeek Reasoner	**98.9%**	174	174
1	Claude Sonnet-4	**98.9%**	**15**	15
3	Claude Haiku	**97.8%**	**8**	**7**
4	GPT-4o Mini	65.6%	79	**7**
5	O4-Mini	62.2%	**10**	**10**
6	DeepSeek Chat	52.2%	31	23

4.3 Resilience to Data Quality Degradation

Data-quality condition produced a measurable shift in LLM accuracy. In the supplementary ANOVA-style comparison on per-task means, the hypothesis that mean LLM accuracy was the same across all four data-quality conditions was rejected ($F(3, 536) = 15.96$, $p < 0.001$, $\eta^2 = 0.082$, medium effect). The same test in the human cohort failed to reject the null hypothesis ($F(3, 86) = 1.06$, $p = 0.370$, $\eta^2 = 0.036$). This should not be interpreted as evidence that humans are unaffected by data quality; rather, the small number of human observations per corrupted condition ($n = 15$) limits the power of this comparison. The main descriptive pattern is that truncated identifiers caused the clearest degradation, particularly for LLM retrieval.

Neither the LLM nor the human completion-time distribution showed a data condition linked change ($p = 0.389$ and $p = 0.621$, respectively). Per-condition

Table 6 Mean accuracy and per-cell Mean and Median completion time by group on H-tasks (6-step). Bold values mark the top three across the LLM cohort for each metric (accuracy descending; time ascending). The human participant result is included for direct comparison but is not part of the LLM ranking.

Rank	Model	Accuracy	Mean Time (s)	Median Time (s)
–	Human Participant	80.0%	134	98
1	DeepSeek Reasoner	**100.0%**	135	133
1	Claude Haiku	**100.0%**	**8**	**8**
3	Claude Sonnet-4	**97.8%**	**13**	13
4	DeepSeek Chat	95.6%	31	26
5	O4-Mini	56.7%	**10**	**10**
6	GPT-4o Mini	32.2%	87	**10**

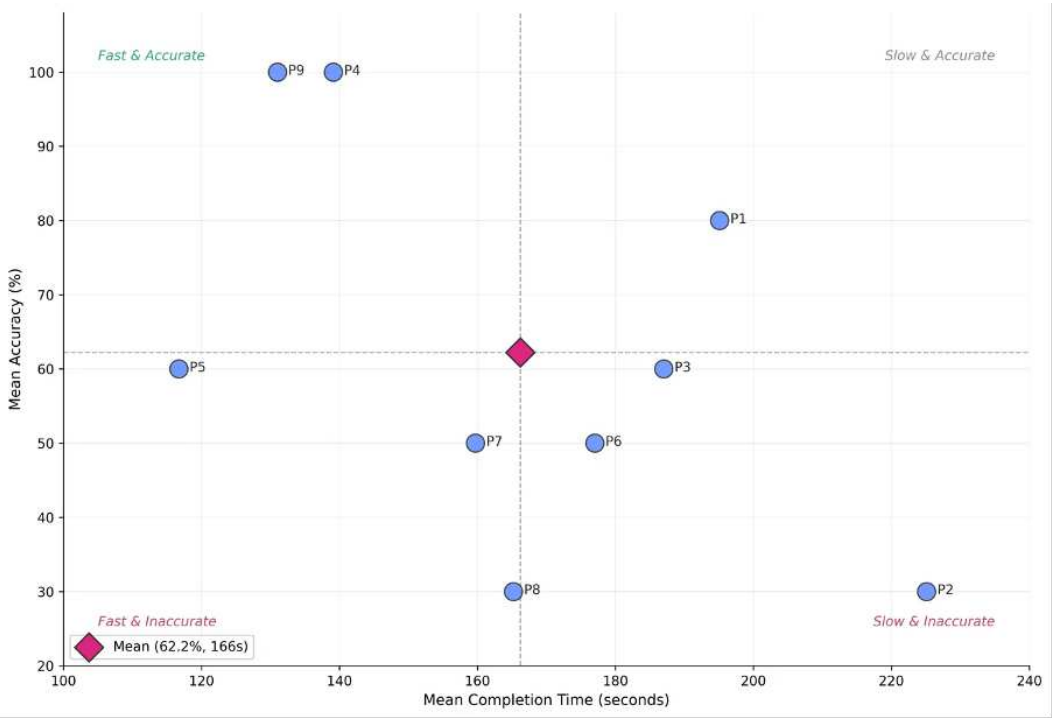


Fig. 6 Individual participant average accuracy and completion time performance across their 10 tasks. The diamond marker and dashed crosshairs indicate the total mean (62.2%, 166 s).

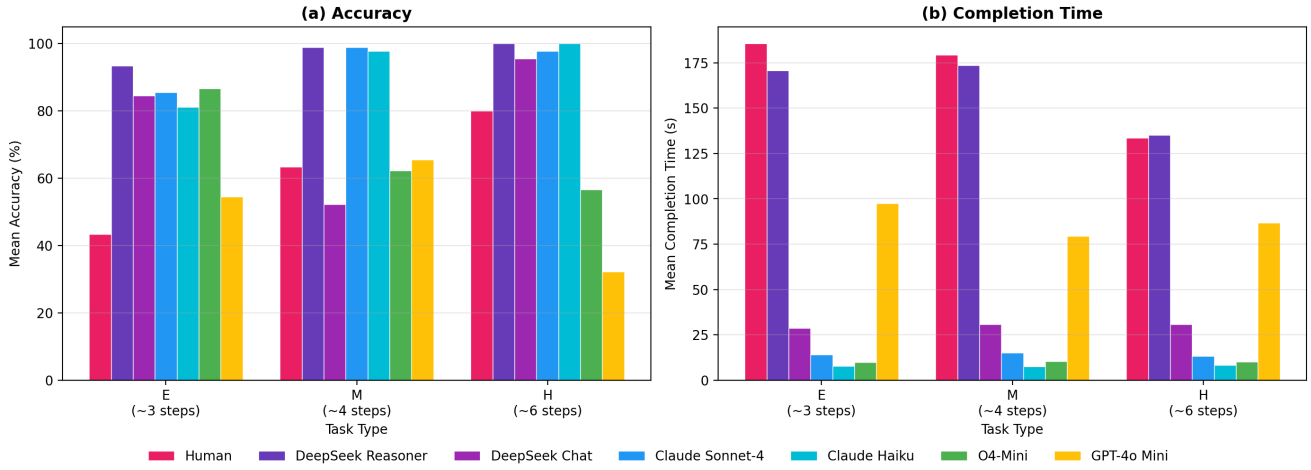


Fig. 7 Mean (a) accuracy and (b) completion time by task type (E, M, H) for the human participants and all six LLM models. Each bar is the mean across the per-(task, model) cells in that task type ($n = 90$ task observations per human; $n = \{30$ easy, 30 medium, 30 hard$\} \times 3$ repeats per LLM). Per-model results are also reported in Tables 4–6.

performance is summarised in Figure 8 and detailed below.

Q0 (Baseline). Under uncorrupted conditions, top LLMs (DeepSeek Reasoner and Claude Haiku) achieved 100.0% accuracy, while human participants achieved a mean of 66.7%. The Q0 gap establishes the reference condition against which the corrupted conditions are compared.

Q1 (Whitespace Injection). The top-3 LLMs (DeepSeek Reasoner, Claude Sonnet-4, Claude Haiku) maintained essentially 100% accuracy under Q1; per-model Q0→Q1 accuracy change within ±1.5 percentage points (pp). Across all six models, the Q0→Q1 change ranged from 16.3 % loss for O4-Mini to 1.5 % increase for Claude Sonnet-4; none of these per-model differences was significantly different (Q0 vs. Q1: $p_{\text{adj}} = 0.297$). Human participants showed a marginal increase in accuracy to 73.3% under this condition compared with 66.7% under Q0; this difference was not statistically significant (Welch's $t = -0.48$, $p = 0.633$). This pattern

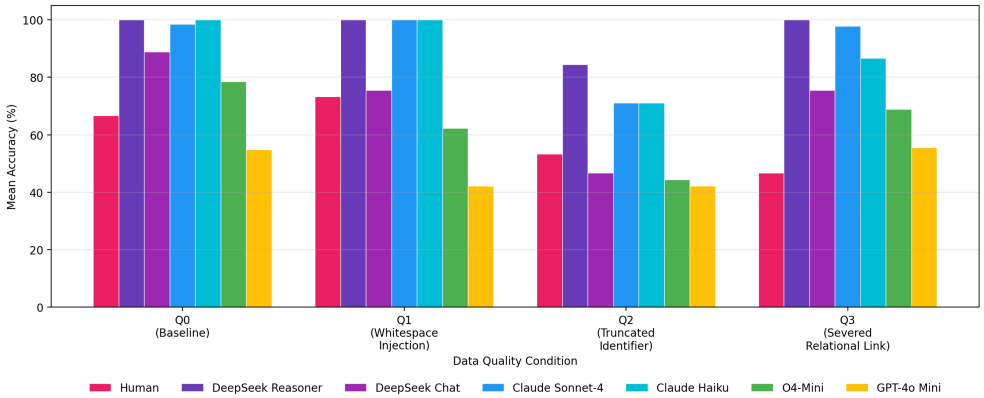


Fig. 8 Mean accuracy by data quality condition (Q0 Baseline, Q1 Whitespace Injection, Q2 Truncated Identifier, Q3 Severed Relational Link) for the human participants and all six LLM models. Each bar is the mean per task and model experiments.

is likely because the vectorisation used by LLMs is not as vulnerable to extra white spaces as precise string matching or searches would be.

Q2 (Truncated Identifier). This condition was the most damaging across the four, producing per-model accuracy drops of 12.6 to 42.2 percentage points relative to Q0; the cohort-level Q0→Q2 mean drop of 26.8 pp was the largest observed degradation (supplementary Tukey HSD post-hoc on the deduplicated $n = 540$ descriptive cells: $p_{\text{adj}} < 0.001$). Character deletion alters the alphanumeric identifiers required for deterministic data lookups, so exact-match retrieval becomes unreliable. Even under this condition, DeepSeek Reasoner held 84.4% accuracy against the human participants, which dropped to 53.3%.

This per-model gap of 31.1 % does not reach conventional significance (Welch's $t = 2.06$, $p = 0.052$, Cohen's $d = 0.75$), reflecting the wide spread of the $n = 15$ human Q2 observations. The cohort-level LLM-vs-human Q2 difference is similarly not significant ($p = 0.642$), so the Q2 condition narrows the LLM advantage to the point where individual-model superiority is detectable only with larger samples.

Q3 (Severed Relational Link). Removing one direction of a bilateral relational link reduced human accuracy to 46.7%, but top LLMs largely retained their performance: DeepSeek Reasoner and Claude Sonnet-4 reached 100.0% and 97.8% accuracy, respectively. In supplementary pairwise tests, both model differences against the human participants were statistically significant (DeepSeek Reasoner vs. human: Welch's $t = 4.00$, $p = 0.001$, $d = 1.46$; Claude Sonnet-4 vs. human: $t = 3.78$, $p = 0.002$, $d = 1.38$). At the cohort level, LLM Q3 accuracy (80.7% across the $n = 90$) was also higher in a supplementary Welch comparison than human Q3 accuracy (46.7%, $n = 15$): Welch's $t = 2.47$, $p = 0.025$, $d = 0.92$. This is likely because LLMs perform a wide search of the data and can still find the remaining link (even with one removed), whereas human participants could not find the other link as easily or assumed there was only one.

4.4 Sensitivity to Prompt Scaffolding

To probe a potential deployment sensitivity, a sub-study tested the three top-performing models (Claude Haiku, Claude Sonnet-4, and DeepSeek Reasoner) on the Medium and Long prompt variants defined in Section 3.3.4 (full templates in Appendix A). In brief: Short is the one-line task prompt used in the main study (36–67 tokens across the 42 unique (task, data-quality) combinations the sub-study covers); Medium wraps the same task wording in a labelled `CONTEXT` sentence, an expanded `REQUIRED STEPS` list, and an explicit `OUTPUT FORMAT` specification (104–195 tokens; Medium/Short ratio 2.4–3.4×, mean 2.9×); and Long adds two further labelled sections listing the available VOOM data sources and the cross-source validation criteria (138–371 tokens; Long/Short ratio 2.1–6.9×, mean 4.8×). The underlying question and correct answer were held constant across the three conditions; only the descriptive scaffolding around the question expanded. The sub-study involved nine tasks per model per prompt-length condition (54 prompt-length evaluations in total), a distinct quantity from the nine human participants in the main study. The reported numbers are intended as an indicative probe of prompt-length sensitivity rather than confirmatory findings.

Although short prompts yielded 93.0–97.4% accuracy across the three models, the introduction of more verbose scaffolding was associated with a substantial accuracy decline (Table 7). Claude Sonnet-4 exhibited the largest single drop, falling from 55.6% (Medium) to 22.2% (Long), a 33.3 percentage-point decline. For

Table 7 Performance Breakdown by Prompt Length for Top Three Models (Short-prompt results from the main 90-task evaluation for comparison; Medium/Long results from a separate 9-task sub-study per model per condition)

Model	Prompt	Accuracy	Avg. Time (s)
Claude Haiku	Short	93.0%	8
	Medium	55.6%	8
	Long	33.3%	7
Claude Sonnet-4	Short	94.1%	14
	Medium	55.6%	11
	Long	22.2%	13
DeepSeek Reasoner	Short	97.4%	160
	Medium	55.6%	222
	Long	44.4%	451

DeepSeek Reasoner, the mean completion time roughly doubled from 222 s (Medium) to 451 s (Long), due to longer generated reasoning and reporting outputs.

A potential explanation is that tasks with a short prompt ask "find printer, count parts," and get the right answer. The same task with a longer prompt asks the planner to go through a verbose, multi-step process; the planner generates a fragile, multi-step plan consisting of validation, cross-validation, and timeline-consistency steps. One step in the longer generated plan could fail to return meaningful context for the data or make a false assumption, causing a cascade of failures through the rest of the chain.

5 Discussion

The empirical results indicate a substantial advantage in accuracy and completion time for the best-performing multi-agent LLM configurations over average unassisted human performance in this controlled study. This section discusses the findings in relation to the literature, examines the unexpected results, assesses economic viability, and reviews the deployment landscape and its limitations.

5.1 Task Complexity Pattern

Mean accuracy rose as the number of task steps increased: 43.3% on E-type (3 steps), 63.3% on M-type (4 steps), and 80.0% on H-type (6 steps) for human participants. This contradicts the assumption made during task design that step count would be a reasonable proxy for difficulty. The result should be interpreted cautiously because task type, identifier structure, and source format were not fully separable in the experimental design. Nevertheless, it suggests that manufacturing information-access difficulty may depend less on the number of procedural steps alone and more on other factors such as search ambiguity and the structure of the underlying data sources. This interpretation is consistent with prior arguments that information silos and interface friction, rather than analytical complexity alone, are major drivers of human error in manufacturing data workflows [2, 17].

The LLM agent-based system showed only a minor, non-significant change in performance. The Master Agent decomposes any query into a structured execution plan, so a three-step E-task and a six-step H-task receive the same systematic treatment, and the operator context-switching burden does not apply. Increasing the number of steps to accommodate larger context windows could degrade system performance, but this is less of an issue as context windows grow for new models.

5.2 Resilience to Manufacturing Data Realities

Manufacturing data is notoriously imperfect [11]. Top LLMs performed strongly across the data quality issues introduced and were able to adapt when spaces were added to identifiers (Q1) or when one direction of a relational link was removed (Q3). However, LLM accuracy degraded sharply when identifiers were truncated (Q2) and characters were missing. This asymmetry mirrors patterns reported in related LLM-on-noisy-data studies in adjacent industrial domains, where retrieval components fail abruptly under string-level corruption while rule-based human checking degrades only gradually [11, 12, 21].

Q2 (Truncated Identifier) was the most damaging condition because character deletion alters the alphanumeric identifiers required for deterministic lookup. However, the vectorisation used by LLMs to store and process data looks for similar characters, so most tasks are still answered correctly. However, this could be an issue when two similar IDs are present and mistaken for one another (e.g., 3D1003 and 3D100), a condition not tested in this study. Ma et al. [27] report a comparable abrupt degradation in their LLM diagnostic pipeline when maintenance-log identifiers are perturbed, indicating that the structural failure mode generalises across studies.

5.3 Architectural Comparison to Prior Work

The architecture presented in this study diverges from the predominant approaches in the literature in two key respects. First, while existing manufacturing LLM deployments predominantly employ standard RAG over a fixed document corpus at Levels 3–4 [24, 26], the present system employs *adaptive* RAG with real-time data ingestion, enabling it to operate without pre-indexed knowledge bases. This is practically significant for environ-

ments where data changes continuously, such as live production logs.

Second, the decomposition into discrete Master, Data Retrieval, Reconciliation, and Synthesis agents, informed by agentic AI paradigms [29, 48], enables multi-step reasoning and cross-system validation. The empirical superiority on H-type tasks, where top models achieved 97–100% accuracy on six-step cross-document synthesis, provides direct evidence of the practical utility of this decomposition for the class of tasks that most closely resemble real vertical integration challenges.

5.4 Economic Viability and Industrial Adoption

Beyond technical performance, the adoption of any integration technology requires a credible return on investment (ROI). The direct operational expenditure (OpEx) for the LLM-based system is small compared with that for human participants. Using per-task token consumption from the evaluation logs, combined with each provider's published API list prices as of the experimental period, mean per-task cost across the six models is on the order of 10^{-3} USD; the cheapest configuration (Claude Haiku) sits below $0.005 per task and the most expensive (DeepSeek Reasoner under maximum-token outputs) below $0.03 per task.

For comparison, the human participants' OpEx is estimated at $1.05 per task, benchmarked at a nominal hourly wage of $22.82/hour for a manufacturing quality-control inspector from the U.S. Bureau of Labour Statistics 2024 [49]. The capital expenditure (CapEx) for system development is estimated at $15,400, representing seven weeks of development at a nominal hourly rate of $55/hour for a data scientist [49]. Assuming a conservative operational volume of 100 queries per day by the LLM (a plausible estimate given documented query-intensive manufacturing workflows [16]), the direct labour cost comparison yields a breakeven point of approximately 13,400 tasks, or roughly six to seven working months. Beyond this horizon, the system's negligible per-task cost generates compounding savings relative to the manual baseline. True deployment could be higher with digital infrastructure and maintenance costs [26].

5.5 Limitations

The first limitation is that the human study relied on a small sample of $n = 9$ participants drawn from the engineering student population at the University of Cambridge's Department of Engineering. This group has significant technical skills and familiarity with LLM tools, but less industry and operations experience than most factory operators. A larger replication with practitioners drawn from manufacturing, quality, traceability, or shop-floor populations would improve the generalisability of the study results. Future studies should use a larger set of participants and quantify different skill or experience levels.

Secondly, participants encountered the data sources used for the first time during the study. In a real deployment, an operator would work with the same data and system for weeks or months, gaining more experience with the data structure and therefore achieving faster processing times. Future studies should investigate longer-term use by having participants use the same datasets over weeks or months.

Lastly, the evaluation was conducted using data from a smart factory demonstrator (VOOM) [37] and tasks devised by the authors and operators. Real manufacturing data introduces greater volume and a wider range of readings than this simulated factory does. The errors in the data may also differ from those reported here and be harder for the LLM system to overcome. Future research should focus on deploying the multi-agent system used here in a real-world manufacturing context, leveraging operational data.

6 Conclusion and Future Work

This paper evaluated whether an LLM-based multi-agent system can support internal information access and sharing across heterogeneous manufacturing data sources. The study was motivated by three gaps in the existing literature: the limited empirical evaluation of LLM systems at lower ISA-95 levels and crossing the IT-OT divide, the lack of quantitative benchmarking against unassisted human users, and the limited understanding of how data quality issues affect the use of LLM-enabled systems in manufacturing.

To address these gaps, a multi-agent system combining adaptive retrieval-augmented generation with specialised data-access tools was developed and tested using data from the VOOM 3D-printing factory demonstrator. The system was evaluated for traceability and certification tasks that required information to be retrieved and synthesised from tracking records, machine logs, packing lists, and Authorised Release Certificates. Its performance was compared with that of human participants across different task types and under controlled data-quality conditions.

The results show that, within the bounded task environment studied, high-performing LLM models can substantially outperform the average unassisted human participant in both accuracy and completion time. This

indicates that multi-agent LLM systems have strong potential as an information-access layer for manufacturing workflows, where users currently need to manually navigate fragmented documents, spreadsheets, and databases.

However, the results also show that this advantage is conditional. Performance varied considerably between LLM models, with smaller models performing below the average of human participants. The best human participants also achieved perfect accuracy, demonstrating that LLM systems should not be viewed as a straightforward replacement for skilled users or operators. Instead, the findings support a human-centred deployment model in which LLM-based systems act as decision-support tools.

The study also highlights important design implications for future manufacturing applications of LLMs. First, nominal step-count was not a reliable proxy for task difficulty. Shorter tasks involving apparently simple identifier lookups yielded lower accuracy than longer verification tasks, suggesting that error-prone information workflows may be driven more by interface fragmentation and search ambiguity than by the number of procedural steps. Second, the system's robustness to data-quality issues was uneven. It handled whitespace errors and missing relational links well, but performance declined more sharply when identifiers were missing characters. Third, the prompt-length experiment showed that additional context substantially reduced accuracy, reinforcing the need for disciplined prompt design and controlled input structuring in industrial deployments.

Overall, the study provides empirical evidence that multi-agent LLM systems can support internal information sharing and access in manufacturing.

Future work should extend this evaluation beyond the controlled VOOM demonstrator to live industrial datasets. Testing the system in an operational manufacturing environment would expose it to larger data volumes, more varied data formats, access-control constraints, and less predictable information structures. Studies with a larger number of participants should also be conducted.

A second direction for future work is to examine longitudinal human–AI collaboration. The present study compared one-off human performance against system performance, but did not evaluate how operators might learn to use, challenge, or correct the system over time.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics approval and consent to participate

The human user study reported in this paper was reviewed and approved by the Department of Engineering Ethics Committee at the University of Cambridge. All participants provided written informed consent prior to participation, having been informed of the study purpose, the voluntary nature of participation, and their right to withdraw at any time without penalty. No personally identifiable data were collected; task responses and completion times were recorded anonymously using participant identifiers.

Data Availability

The VOOM simulation dataset, along with others, used in this study, is available from the corresponding author upon reasonable request.

Authors' contributions

Ngoc Bach Nguyen: Conceptualisation; Methodology; Software; Investigation; Formal analysis; Data curation; Visualisation; Writing & Editing. **Sam Brooks**: Conceptualisation; Methodology; Supervision; Resources; Review & Editing. **Zhengyang Ling**: Conceptualisation; Collecting Data for study.

References

1. Ziwei Wan, Zhen Gao, Marco Di Renzo, and Lajos Hanzo. The road to industry 4.0 and beyond: A communications-, information-, and operation technology collaboration perspective. *IEEE Network*, 36 (6):157–164, 2022. doi: 10.1109/MNET.008.2100484.
2. Pramod Kumar, Jaiprakash Bhamu, and Kuldip Singh Sangwan. Analysis of Barriers to Industry 4.0 adoption in Manufacturing Organizations: an ISM Approach. *Procedia CIRP*, 98:85–90, 2021. ISSN 22128271. doi: 10.1016/j.procir.2021.01.010. URL `https://linkinghub.elsevier.com/retrieve/pii/S2212827121000330`.
3. Charlotta Johnsson. Isa 95 - how and where can it be applied. In *Proceedings of the 2004 IEEE International Conference on Emerging Technologies and Factory Automation*, pages 1–8, 10 2004.

4. Zhengyang Ling, Sam Brooks, Duncan McFarlane, Alan Thorne, and Gregory Hawkridge. Vision-based extraction of industrial information from legacy programmable logic controllers. *Robotics and Computer-Integrated Manufacturing*, 97:103088, 2026.
5. Gregory Hawkridge, Anandarup Mukherjee, Duncan McFarlane, Yedige Tlegenov, Ajith K Parlikad, Nicholas J Reyner, and Alan Thorne. Monitoring on a shoestring: Low cost solutions for digital manufacturing. *Annual Reviews in Control*, 51:374–391, 2021.
6. Soujanya Mantravadi, Jagjit Singh Srai, and Charles Møller. Application of mes/mom for industry 4.0 supply chains: A cross-case analysis. *Computers in Industry*, 148:103907, 2023.
7. Dan Li, Åsa Fast-Berglund, and Dan Paulin. Current and future industry 4.0 capabilities for information and knowledge sharing: Case of two swedish smes. *The International Journal of Advanced Manufacturing Technology*, 105(9):3951–3963, 2019.
8. National Research Council, Committee to Study Information Technology and Manufacturing, and National Research Council/Computer Science and Telecommunications Board. *Information Technology for Manufacturing*. National Academies Press, 1995. ISBN 978-0-309-17671-2. OCLC: 1480558441.
9. Patrick Denzler and Wolfgang Kastner. Reference architectures for closing the it/ot gap. In *Digital Transformation: Core Technologies and Emerging Topics from a Computer Science Perspective*, pages 95–123. Springer, 2023.
10. Lütfi Apilioğulları. Digital transformation in project-based manufacturing: Developing the isa-95 model for vertical integration. *International Journal of Production Economics*, 245:108413, 2022.
11. Benjamin Hazen, Christopher Boone, L.A. Jones-Farmer, and Jeremy Ezell. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 08 2014. doi: 10.1016/j.ijpe.2014.04.018.
12. Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F. Chen. Resilience of large language models for noisy instructions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11939–11950, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.697. URL `https://aclanthology.org/2024.findings-emnlp.697/`.
13. Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. A study on chatgpt for industry 4.0: Background, potentials, challenges, and eventualities. *Journal of Economy and Technology*, 1:127–143, 2023. ISSN 2949-9488. doi: https://doi.org/10.1016/j.ject.2023.08.001. URL `https://www.sciencedirect.com/science/article/pii/S2949948823000033`.
14. Melanie Steinhüser, Lena Waizenegger, Shahper Vodanovich, and Alexander Richter. Knowledge management without management – shadow it in knowledge-intensive manufacturing practices. In *Proceedings of the 25th European Conference on Information Systems*, pages 1–10, 06 2017.
15. Chao Zhang, Qingfeng Xu, Yongrui Yu, Guanghui Zhou, Keyan Zeng, Fengtian Chang, and Kai Ding. A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing. *Robotics and Computer-Integrated Manufacturing*, 92:102883, April 2025. ISSN 07365845. doi: 10.1016/j.rcim.2024.102883. URL `https://linkinghub.elsevier.com/retrieve/pii/S0736584524001704`.
16. Julian M. Müller, Johannes W. Veile, and Kai-Ingo Voigt. Prerequisites and incentives for digital information sharing in Industry 4.0 – An international comparison across data types. *Computers & Industrial Engineering*, 148:106733, October 2020. ISSN 03608352. doi: 10.1016/j.cie.2020.106733. URL `https://linkinghub.elsevier.com/retrieve/pii/S0360835220304563`.
17. James F. Lardner. *Manufacturing Systems: Foundations of World-Class Practice*. National Academies Press, Washington, D.C., January 1992. ISBN 978-0-309-04678-7. doi: 10.17226/1867. URL `http://www.nap.edu/catalog/1867`. Pages: 173 and onwards.
18. Zude Zhou, Peng Hu, and Quan Liu. Towards a Performance-Aware Model of Manufacturing Information Sharing. In *2008 IEEE International Conference on Industrial Technology*, pages 1–4, Chengdu, China, April 2008. IEEE. ISBN 978-1-4244-1705-6. doi: 10.1109/ICIT.2008.4608472. URL `http://ieeexplore.ieee.org/document/4608472/`.
19. Barbara Flynn, Baofeng Huo, and Xiande Zhao. The impact of supply chain integration on performance: A contingency and configuration approach. *Journal of Operations Management - J OPER MANAG*, 28:58–71, 01 2010. doi: 10.1016/j.jom.2009.06.001.

20. Joakim Kembro, Dag Näslund, and Jan Olhager. Information sharing in extended supply chains: why does it have to be so difficult? *Production Planning & Control*, pages 1–15, April 2024. ISSN 0953-7287, 1366-5871. doi: 10.1080/09537287.2024.2331742. URL https://www.tandfonline.com/doi/full/10.1080/09537287.2024.2331742.
21. Marvin Matthes, Oliver Guhr, Martin Krockert, and Torsten Munkelt. Leveraging LLMs for Information Extraction in Manufacturing. In Matthias Thürer, Ralph Riedel, Gregor Von Cieminski, and David Romero, editors, *Advances in Production Management Systems. Production Management Systems for Volatile, Uncertain, Complex, and Ambiguous Environments*, volume 732, pages 355–366. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-71636-2 978-3-031-71637-9. doi: 10.1007/978-3-031-71637-9_24. URL https://link.springer.com/10.1007/978-3-031-71637-9_24. Series Title: IFIP Advances in Information and Communication Technology.
22. Claire Hannibal, Jack Rowan, Olatunde Durowoju, David Bryde, Jake Holloway, Omolola Adeyemi, and Saira Shamim. Who shares wins? understanding barriers to information sharing in managing supply chain risk. *Continuity & Resilience Review*, 4(2): 161–175, 2022.
23. Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large Language Models for Supply Chain Optimization, July 2023. URL http://arxiv.org/abs/2307.03875. arXiv:2307.03875 [cs].
24. Samuel Kernan Freire, Chaofan Wang, Mina Foosherian, Stefan Wellsandt, Santiago Ruiz-Arenas, and Evangelos Niforatos. Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. *Frontiers in Artificial Intelligence*, 7:1293084, March 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1293084. URL https://www.frontiersin.org/articles/10.3389/frai.2024.1293084/full.
25. Xuan Liu, John Ahmet Erkoyuncu, Jerry Ying Hsi Fuh, Wen Feng Lu, and Bingbing Li. Knowledge extraction for additive manufacturing process via named entity recognition with LLMs. *Robotics and Computer-Integrated Manufacturing*, 93:102900, June 2025. ISSN 07365845. doi: 10.1016/j.rcim.2024.102900. URL https://linkinghub.elsevier.com/retrieve/pii/S073658452400187X.
26. Stefan Wellsandt, Mina Foosherian, Alexandros Bousdekis, Bernhard Lutzer, Fotis Paraskevopoulos, Yiannis Verginadis, and Gregoris Mentzas. *Fostering Human-AI Collaboration with Digital Intelligent Assistance in Manufacturing SMEs*, pages 649–661. Springer Nature Switzerland, 09 2023. ISBN 978-3-031-43661-1. doi: 10.1007/978-3-031-43662-8_46.
27. Yunfei Ma, Shuai Zheng, Zheng Yang, Hongcheng Pan, and Jun Hong. A knowledge-graph enhanced large language model-based fault diagnostic reasoning and maintenance decision support pipeline towards industry 5.0. *International Journal of Production Research*, 2025. doi: 10.1080/00207543.2025.2472298. URL https://doi.org/10.1080/00207543.2025.2472298.
28. Honghong Fan, Zhengyi Fan, Chao Liu, et al. MetalMind: a knowledge graph-driven human-centric knowledge system for metal additive manufacturing. *npj Advanced Manufacturing*, 2:25, 2025. doi: 10.1038/s44334-025-00038-9. URL https://www.nature.com/articles/s44334-025-00038-9. Three innovations: LLM-driven KG construction with collaborative verification; hybrid (vector + graph) retrieval; MR-enhanced interface for continuous knowledge capture.
29. Alexander Sukharevsky, Dave Kerr, Klemens Hjartar, Lari Hämäläinen, Stéphane Bout, and Vito Di Leo. Seizing the agentic AI advantage. https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage, June 2025. Accessed: 2025-6-24.
30. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.
31. Zhen Zhao, Dunbing Tang, Changchun Liu, Liping Wang, Zequn Zhang, Haihua Zhu, Kai Chen, Qingwei Nie, and Yuchen Ji. A large language model-based multi-agent manufacturing system for intelligent shopfloors. *Advanced Engineering Informatics*, 69:103888, 2026. doi: 10.1016/j.aei.2025.103888. URL https://doi.org/10.1016/j.aei.2025.103888.
32. Jin Huang, Yue Teng, Qihao Liu, Liang Gao, Xinyu Li, Chunjiang Zhang, and Guoqing Xu. Leveraging large language models for efficient scheduling in human–robot collaborative flexible manufacturing systems. *npj Advanced Manufacturing*, 2:47, 2025. doi: 10.1038/s44334-025-00061-w. URL https://doi.org/10.1038/s44334-025-00061-w.

33. Mingzhe Ni, Tao Wang, Jiewu Leng, Chong Chen, and Lianglun Cheng. A large language model-based manufacturing process planning approach under industry 5.0. *International Journal of Production Research*, 2025. doi: 10.1080/00207543.2025.2469285. URL https://doi.org/10.1080/00207543.2025.2469285.
34. Mohaiad Elbasheer, Yuanjun Laili, Francesco Longo, Vittorio Solina, Yiran Tao, Pierpaolo Veltri, Yuteng Zhang, and Lin Zhang. Natural language-driven production planning: Integrating large language models with automatic simulation model generation in manufacturing systems. *Journal of Intelligent Manufacturing*, November 2025. doi: 10.1007/s10845-025-02732-z. URL https://doi.org/10.1007/s10845-025-02732-z.
35. Jiachen Guo, Chanwook Park, Dong Qian, Thomas J. R. Hughes, and Wing Kam Liu. Large language model-empowered next-generation computer-aided engineering. *Computer Methods in Applied Mechanics and Engineering*, 450:118591, 2026. doi: 10.1016/j.cma.2025.118591. URL https://www.sciencedirect.com/science/article/abs/pii/S0045782525008631. LLM-empowered CAE agent that plans, executes, and adapts PDE solver workflows for model order reduction; evaluated against solver fidelity rather than against engineers.
36. Juan Pablo Usuga-Cadavid, Samir Lamouri, Bernard Grabot, and Arnaud Fortin and. Using deep learning to value free-form text data for predictive maintenance. *International Journal of Production Research*, 60(14):4548–4575, 2022. doi: 10.1080/00207543.2021.1951868. URL https://doi.org/10.1080/00207543.2021.1951868.
37. Sam Brooks, Duncan McFarlane, Alan Thorne, Zhengyang Ling, Gregory Hawkridge, Darius Danaei, Sławomir Tadeja, Svetan Ratchev, Sudhir Rama Murthy, Nikolai Kazantsev, Sebastian Pattinson, Chander Velu, and Thomas Bohné. Design of a demonstrator environment for investigating multi-factory production and operation challenges. In *Advances in Manufacturing Technology XXXVI*, pages 99–109. IOS Press, 2023. ISBN 978-1-64368-256-8. doi: 10.3233/ATDE230908. URL https://doi.org/10.3233/ATDE230908.
38. Wenqi Guan and Yang Fang. Optimizing web-based ai query retrieval with gpt integration in langchain: A cot-enhanced prompt engineering approach. In *Proceedings of the 2nd International Conference on Machine Learning and Automation, CONF-MLA 2024, November 21, 2024, Adana, Turkey*, CONF-MLA. EAI, 2025. doi: 10.4108/eai.21-11-2024.2354589. URL http://dx.doi.org/10.4108/eai.21-11-2024.2354589.
39. Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity, 2024. URL https://arxiv.org/abs/2403.14403.
40. Ayman Asad Khan, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report, 2024. URL https://arxiv.org/abs/2410.15944.
41. Zhuocheng Shen. Llm with tools: A survey, 2024. URL https://arxiv.org/abs/2409.18807.
42. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.
43. Hayne W. Reese. Counterbalancing and other uses of repeated-measures latin-square designs: Analyses and interpretations. *Journal of Experimental Child Psychology*, 64(1):137–158, 1997. ISSN 0022-0965. doi: https://doi.org/10.1006/jecp.1996.2333. URL https://www.sciencedirect.com/science/article/pii/S0022096596923334.
44. Tae Kyun Kim. T test as a parametric statistic. *Korean J Anesthesiol*, 68(6):540–546, 2015. doi: 10.4097/kjae.2015.68.6.540. URL http://ekja.org/journal/view.php?number=8123.
45. H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar 1947. doi: 10.1214/aoms/1177730491.
46. Eva Ostertagova and Oskar Ostertag. Methodology and application of one-way anova. *American Journal of Mechanical Engineering*, 1:256–261, 11 2013. doi: 10.12691/ajme-1-7-21.
47. Gail M. Sullivan and Richard Feinn. Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282, 09 2012. ISSN 1949-8349. doi: 10.4300/JGME-D-12-00156.1. URL https://doi.org/10.4300/JGME-D-12-00156.1.

48. Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.08435.
49. U.S. Bureau of Labor Statistics. May 2024 Occupational Employment and Wage Statistics: National Occupational Employment and Wage Estimates, April 2025. URL https://www.bls.gov/oes/2024/may/oes_nat.htm. Reference period: May 2024; release date: 2 April 2025. National median hourly wages cited herein: SOC 15-2051 Data Scientists $54.13/hour (annual median $112,590); SOC 51-1011 First-Line Supervisors of Production and Operating Workers $34.23/hour (annual median $71,190); SOC 51-9061 Quality Control Inspectors $22.82/hour (annual median $47,460).

A Prompt Template Variants

Tables 8 and 9 present the Medium and Long prompt templates used in the context overload sub-study (Section 4.4).

Table 8 Medium-Length Prompt Templates by Task Type

Type	Prompt Template
E	Task: Find all gears for Packing List {packing_list_id} CONTEXT: You are analyzing manufacturing data to identify gears associated with a specific packing list. REQUIRED STEPS: 1. Locate the specified Packing List document 2. Extract the Order ID from the packing list 3. Use the Order ID to find all associated gears in the system 4. Provide a complete list of gears for the order OUTPUT FORMAT: – Order ID found; Total number of gears; Complete gear list; Any issues encountered.
M	Task: For Part {part_id}, determine which 3D printer was used and count total parts on that printer. CONTEXT: You need to trace a specific part back to its manufacturing origin and analyze printer utilization. REQUIRED STEPS: 1. Identify the specific part in the tracking system 2. Trace the part to its assigned 3D printer through relationship data 3. Access machine logs to verify printer assignment 4. Count all parts printed on the identified printer 5. Validate consistency between data sources OUTPUT FORMAT: – Part ID confirmation; Assigned printer identification; Total parts count; Data source validation results.
H	Task: For Order {order_id}, verify that ARC document completion date matches warehouse arrival date. CONTEXT: You are performing compliance verification by cross-referencing manufacturing certification documents with logistics tracking data. REQUIRED STEPS: 1. Locate the ARC (Acceptance/Release Certificate) document for the order 2. Extract the completion/certification date from the document 3. Find the warehouse arrival timestamp in location tracking data 4. Compare the dates and determine if they match within acceptable tolerance 5. Assess compliance status and any regulatory implications OUTPUT FORMAT: – Order ID confirmation; ARC document date; Warehouse arrival date; Date match status; Compliance assessment; Risk level evaluation.

Table 9 Long-Length Prompt Templates by Task Type

Type	Prompt Template
E	Task: Find all gears for Packing List {packing_list_id}. CONTEXT: You are analyzing a 3D printing factory's order-fulfillment data. Lookup the Packing List to find the corresponding Order ID, then trace all associated gears. REQUIRED ANALYSIS: 1. Locate and parse the specified Packing List document. 2. Extract the Order ID from the document. 3. Use the Order ID to find all associated gears in the relationship data. 4. Provide a complete gear list for the order. DATA SOURCES AVAILABLE: – Packing List documents (PDF); Relationship tracking data (CSV); Location tracking data; Machine operation logs. VALIDATION REQUIREMENTS: – Cross-reference findings across multiple sources; verify gear-to-order relationships; ensure completeness; document any discrepancies. EXPECTED OUTPUT: Structured report with Order ID, total gear count, complete gear list, and validation notes.
M	Task: For Part {part_id}, determine which 3D printer was used and count the total number of parts printed on that printer. CONTEXT: Cross-system validation between relationship data and machine logs to trace a part back to its manufacturing origin and analyze utilization. REQUIRED ANALYSIS: 1. Identify the specific part in the system. 2. Trace the part to its assigned 3D printer through relationship data. 3. Access machine logs to verify printer assignment. 4. Count all parts printed on the identified printer. 5. Validate temporal consistency between systems. DATA SOURCES AVAILABLE: – Relationship data (worker-printer-part associations); machine logs (start/end times); location data (barcode tracking); worker data (RFID activity logs). SUCCESS CRITERIA: – Correctly identify the 3D printer; provide accurate part count; cross-validate assignment; report any discrepancies.
H	Task: For Order {order_id}, verify that the ARC (FAA 8130-3) document completion date matches the date the order arrived at the Parts Warehouse. CONTEXT: Critical compliance verification in aerospace manufacturing, FAA certificates must align with actual completion to ensure airworthiness. REQUIRED ANALYSIS: 1. Locate and parse the ARC certificate. 2. Extract the certificate completion date. 3. Trace the order through location data to find warehouse arrival. 4. Compare certificate date with arrival date. 5. Identify discrepancies and assess compliance implications. DATA SOURCES AVAILABLE: – FAA certificates (PDF); barcode tracking data; relationship data; worker certification logs. SUCCESS CRITERIA: – Extract and confirm dates; verify date alignment; provide compliance assessment and corrective recommendations.