

<https://doi.org/10.1038/s44334-025-00038-9>

# MetalMind: A knowledge graph-driven human-centric knowledge system for metal additive manufacturing



Haolin Fan<sup>1,2</sup>, Zhen Fan<sup>2</sup>, Chenshu Liu<sup>1</sup>, Jianhao Zhu<sup>1</sup>, Tom Gibbs<sup>3</sup>, Jerry Ying Hsi Fuh<sup>2</sup>, Wen Feng Lu<sup>2</sup> & Bingbing Li<sup>1,4</sup>✉

In the Industry 5.0 era, increasing manufacturing complexity and fragmented knowledge pose challenges for decision-making and workforce development. To tackle this, we present a human-centric knowledge system that integrates explicit knowledge from formal sources and implicit knowledge from expert insights. The system features three core innovations: (1) an automated KG construction pipeline leveraging large language models (LLMs) with collaborative verification to enhance knowledge extraction accuracy and minimize hallucinations; (2) a hybrid retrieval framework that combines vector-based, graph-based, and hybrid retrieval strategies for comprehensive knowledge access, achieving a 336.61% improvement over vector-based retrieval and a 68.04% improvement over graph-based retrieval in global understanding; and (3) an MR-enhanced interface that supports immersive, real-time interaction and continuous knowledge capture. Demonstrated through a metal additive manufacturing (AM) case study, this approach enriches domain expertise, improves knowledge representation and retrieval, and fosters enhanced human-machine collaboration, ultimately supporting adaptive upskilling in smart manufacturing.

In the evolving paradigm of Industry 5.0, the focus increasingly shifts toward human-centric innovation, emphasizing skill enhancement and seamless integration of technology into human-driven processes<sup>1</sup>. The proliferation of generative artificial intelligence (Gen AI) innovations necessitates a new approach to reimagine the upskilling and reskilling priorities for employee attraction, engagement, and retention<sup>2</sup>. This shift underscores the development of knowledge systems that support human learning, understanding, development, accessibility, and collaboration, especially in complex processes like metal additive manufacturing (AM)<sup>3</sup>. In this context, when integrated with human-centric knowledge systems, distributed and disparate digital twin (DT) technology is pivotal in improving the usability, interoperability, and effectiveness of human-machine symbiotic manufacturing systems<sup>4</sup>. However, the intrinsic complexity of metal AM presents significant challenges, especially in providing non-expert users with operational insights and actionable knowledge<sup>5</sup>. This challenge is compounded by explicit knowledge often stored in fragmented, unstructured formats like PDF documents, making automated processing and knowledge retrieval difficult. Moreover, implicit knowledge, critical for understanding nuanced processes, often remains difficult to articulate or extract. The lack of accessible, structured knowledge representation methods hinders automated

knowledge extraction, comprehensive and adaptive knowledge system construction, and the broader adoption of Gen AI in metal AM processes.

Recent advancements in large language models (LLMs) have shown promise in addressing these challenges in human-machine symbiotic manufacturing systems<sup>6</sup>. LLMs excel in natural language understanding, summarization, and processing unstructured data, offering opportunities to build comprehensive knowledge systems for manufacturing<sup>7–9</sup>. However, LLMs often struggle with domain-specific knowledge, leading to inaccurate or misleading information, an issue known as hallucination<sup>10</sup>. Retrieval augmented generation (RAG) has emerged as a promising alternative to address this issue. RAG enhances LLM performance by leveraging external knowledge bases as sources and retrieving information as context during the generation process<sup>11</sup>. Unlike fine-tuning to adjust pre-trained weights, RAG employs an external retrieval mechanism to supplement the LLM's internal knowledge, effectively reducing hallucinations when processing domain-specific information<sup>12</sup>. While RAG significantly improves accuracy in domain-specific tasks, its focus on retrieving specific, granular information can limit its ability to capture the broader interconnections and systemic relationships within the metal AM domain, potentially missing important contextual understanding that spans multiple aspects of the field.

<sup>1</sup>Autonomy Research Center for STEAHM (ARCS), California State University Northridge, Northridge, CA, USA. <sup>2</sup>Department of Mechanical Engineering, National University of Singapore, Singapore, Singapore. <sup>3</sup>Nvidia Inc, Santa Clara, CA, USA. <sup>4</sup>Department of Mechanical and Aerospace Engineering, University of California Los Angeles, Los Angeles, CA, USA. ✉e-mail: [bingbing.li@csun.edu](mailto:bingbing.li@csun.edu)

Knowledge graphs (KGs) present a promising solution to address these limitations by providing structured and semantically enriched representations of domain knowledge. KGs organize concepts and their interrelations to facilitate deeper understanding and improved contextual awareness of complex domains like metal AM<sup>13</sup>. Traditional KG construction has relied on named entity recognition (NER) and relation extraction (RE) techniques to define and extract nodes and relationships<sup>14,15</sup>. These methods are labor-intensive and heavily dependent on manual annotation, limiting their scalability. Although recent advancements in natural language processing (NLP) have introduced automated extraction capabilities, the requirement for extensive human-labeled training data remains a significant challenge. LLMs, leveraging their pre-training on extensive datasets, offer a new paradigm for automating KG development by accurately extracting entities and relationships with minimal human input<sup>16</sup>. This automation substantially improves KG construction efficiency, enabling rapid updates and greater adaptability to evolving manufacturing requirements.

Integrating KGs with RAG creates a powerful synergy that enhances knowledge linkage accuracy while promoting logical coherence and technical precision in generated content<sup>17,18</sup>. This combined approach is particularly valuable for capturing and representing different types of manufacturing knowledge that emerge during human-machine interaction (HMI). For instance, implicit knowledge, the expertise professionals possess but haven't formally documented, such as operational experience and intuition critical for successful metal AM operations, can be gradually formalized through the system's interactions. This comprehensive knowledge representation is especially beneficial for training non-expert users in metal AM applications, providing them with structured, reliable tools that bridge the gap between theoretical understanding and practical implementation, ultimately supporting more informed analysis and decision-making processes.

Building upon this synergy, we developed a comprehensive human-centric knowledge system that combines a multi-modal KG with mixed reality (MR) technology to provide more intuitive, efficient, and robust decision support in metal AM. This paper presents a novel methodology for manufacturing knowledge integration through three primary innovations:

1. Enhanced automated KG construction: Existing KG construction methods often rely on rule-based or traditional NLP techniques, which struggle with adaptability and completeness. We introduced a novel automated KG construction pipeline leveraging LLMs. Our collaborative verification mechanism enhances accuracy and mitigates LLM-induced hallucination issues, which are common in existing automated systems. This pipeline captures explicit knowledge from diverse sources through pre-processing, extraction, and collaborative verification, ensuring information accuracy and accessibility.
2. Hybrid knowledge retrieval for contextualized access: Traditional retrieval systems struggle with balancing fine-grained knowledge retrieval and contextual understanding. We implemented a multi-faceted retrieval system that combines vector-based retrieval (for granular details), graph-based retrieval (for contextual relationships), and hybrid retrieval, enabling comprehensive knowledge access tailored to various knowledge levels.
3. Multi-modal knowledge integration and immersive interaction: Manufacturing knowledge systems often fail to capture implicit knowledge from expert interactions. We developed a multi-modal KG framework with embedded entity descriptions and text-to-image retrieval capabilities. The integration with MR technology enables immersive interaction while systematically capturing two types of knowledge: explicit knowledge (from formal sources) and implicit knowledge (unexpressed expert insights, operational expertise, and patterns) during the human-machine and human-human interactions.

## Results

### KG construction results

The KG we constructed offers a comprehensive knowledge representation of the Renishaw AM400 metal AM machine, as depicted in Fig. 1a. This KG

comprises 3102 nodes and 8366 relationships, structured in Neo4j to capture the intricate and interconnected nature of the metal AM machine. We utilized the 'all-MiniLM-L6-v2' model with a 384-dimensional embedding for both text and image descriptions to enhance representation and extract meaningful insights.

To emphasize specific machine elements, such as the 'Chiller' and 'Small Powder Bottle,' we provide a magnified view in Fig. 1b, which highlights their associated nodes and interactions. This multi-layered visualization underscores the KG's ability to represent high-level and component-specific interactions within the AM400 machine.

### Retrieval performance

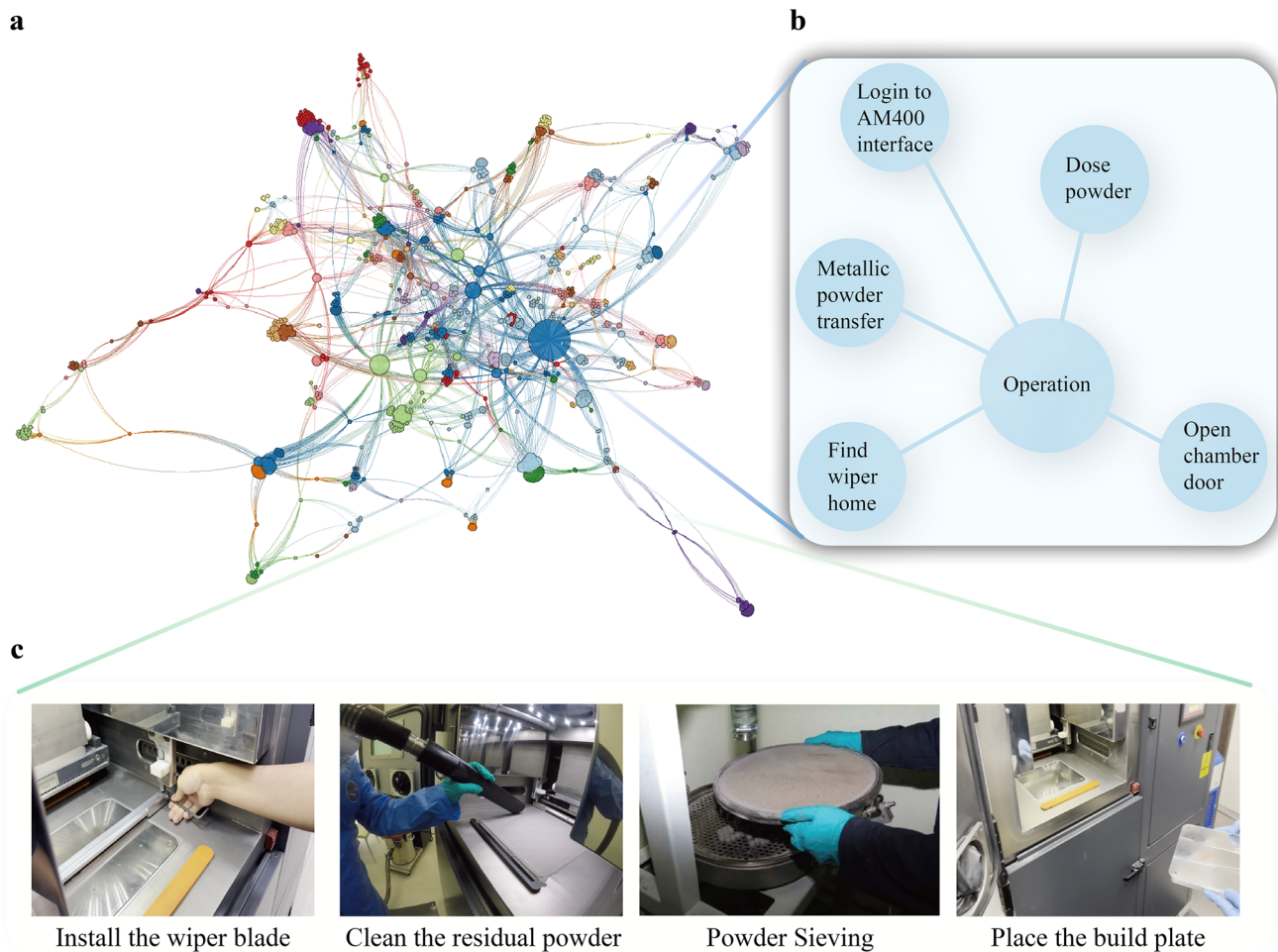
To evaluate the effectiveness of different retrieval modes, we constructed a dataset of 100 queries with manually generated answers serving as ground truth. This dataset includes 70 granular queries and 30 broader system-level queries that require a comprehensive understanding. A variability analysis of the evaluation dataset is provided in the supplementary information. We used several evaluation metrics inspired by<sup>19</sup> to compare performance across different retrieval modes. These metrics include: (i) faithfulness, measuring factual consistency of answers relative to context; (ii) answer relevancy, which penalizes incomplete or redundant information; (iii) context precision, indicating the proportion of relevant information retrieved; (iv) context recall, measuring the successful retrieval of relevant documents; and (v) domain-specific rubrics, a score-based rating from 1 to 5, where experts assess answers based on predefined criteria. Additionally, we tracked token consumption to evaluate the computational efficiency of each mode.

**Multi-faceted RAG performance.** The evaluation results, shown in Fig. 2, indicate that all retrieval modes performed similarly across most metrics for granular retrieval. However, the hybrid mode, which integrates vector and graph-based retrieval, consistently achieved higher scores, particularly in answer relevancy and rubrics score, with a 9% increase in rubric score over the vector mode. This suggests combining multiple retrieval techniques can enhance retrieved answers' quality and contextual relevance, maintaining precision and faithfulness without notable trade-offs.

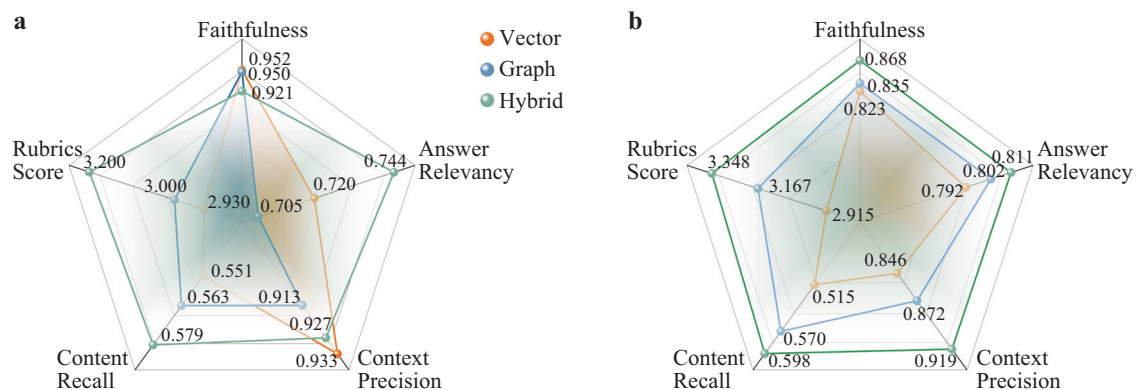
The hybrid approach demonstrated clear global retrieval advantages, outperforming vector and graph modes across all metrics. Notably, the hybrid mode achieved consistent gains in faithfulness, with improvements of around 5% and 4% over vector and graph modes, respectively. Moreover, context recall showed a significant 16% enhancement compared to the vector mode, reflecting its superior ability to retrieve comprehensive contextual information. The rubrics score also favored the hybrid mode, with an overall score of 3.35, compared to 2.92 for the vector mode and 3.17 for the graph mode, highlighting the hybrid model's alignment with quality standards.

We first calculated the average token consumption separately for granular and global contexts to evaluate token efficiency across different retrieval modes. Recognizing the need to balance retrieval quality with token efficiency, we developed a scoring method combining these two aspects. Specifically, we normalized the retrieval accuracy and token consumption values on a 0 to 1 scale, allowing consistent comparison across modes. A weighted composite score was subsequently calculated, assigning 70% weight to normalized retrieval quality to emphasize accuracy and 30% to token efficiency, defined by average token usage. We applied a sigmoid transformation to ensure a more even score distribution and reduce clustering effects, resulting in a non-linear, balanced range from 1 to 5. Table 1.

Token efficiency varied significantly across retrieval modes. Overall, global retrieval consumed more tokens than granular retrieval, indicating a higher computational cost for broader contexts. Among the modes, the graph approach was the most efficient in token usage for both granular and global contexts, demonstrating its strength in managing resources. However, despite its efficiency, the graph mode scored lower overall in granular contexts. In this scenario, the hybrid mode achieved the highest composite score of 4.30, effectively balancing token consumption with retrieval



**Fig. 1 | The constructed knowledge graph (KG).** **a** A global overview of the KG structure. **b** A detailed, zoomed-in view highlighting the Operation domain within the KG. **c** Images represented as entities stored within the KG.



**Fig. 2 | Comparison of retrieval mode performance across granular and global retrieval tasks.** **a** In granular retrieval, the vector and graph modes show similar performance, with the hybrid mode achieving the highest scores. **b** In global

retrieval, the graph mode surpasses the vector mode, while the hybrid mode demonstrates the best overall performance.

accuracy and outperforming both graph (1.83) and vector (2.79) modes. Notably, the vector mode surpassed the graph mode in granular retrieval, suggesting it is suitable for retrieving details. For global queries, the hybrid mode achieved the highest score (4.89), demonstrating its versatility for broad retrieval tasks. It outperformed the graph mode (2.91) by 68.04% and the vector mode (1.12) by 336.61%. The graph mode performed reasonably well, while the vector mode lagged due to its limited global understanding.

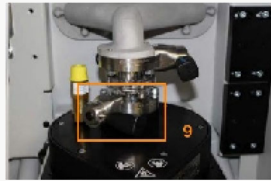


In summary, although each retrieval mode showed strengths in specific areas, the hybrid approach consistently achieved the highest composite scores across granular and global retrieval contexts. This suggests that the hybrid strategy successfully integrates the precision of vector-based retrieval with the comprehensive contextual insights afforded by KG. Such a balanced approach is advantageous in specific domains like metal AM, where accuracy and contextual relevance are essential to understanding the

**Table 1 | Average token consumption (with standard deviation) and corresponding composite scores for each retrieval mode across granular and global contexts**

Aspect	Vector		Graph		Hybrid	
	Token (10 <sup>3</sup> )	Score	Token (10 <sup>3</sup> )	Score	Token (10 <sup>3</sup> )	Score
Granular	5.96 (±1.38)	2.79	5.70 (±1.33)	1.83	5.93 (±1.33)	4.30
Global	6.35 (±1.29)	1.12	6.05 (±1.19)	2.91	6.16 (±1.19)	<b>4.89</b>

Bold text denotes the best performance.

**Table 2 | A qualitative explicit knowledge retrieval result: description and location of the upper safe change filter valve**

Query & Response	Image Reference
<b>Query:</b> What and where is the upper safe change filter valve?  <b>Response:</b> The upper safe change filter valve, referred to as <b>Valve (F1)</b> <sup>1</sup> , is part of the safe change filter assembly on the AM400 system. It is located on the KF flange of the safe change filter assembly. When installing the filter, a <b>centering ring seal</b> <sup>2</sup> is placed on this valve. During maintenance, the upper isolating valve (F1) is opened to <b>fill the filter housing with water</b> <sup>3</sup> .	 <p>1</p>  <p>2</p>  <p>3</p>

complex interdependencies among components and processes. By efficiently managing token usage while maintaining high retrieval quality, the hybrid mode presents a robust solution for meeting the intricate demands of both fine-grained and global understanding.

**Qualitative analysis.** We conduct a qualitative analysis alongside our quantitative findings to offer a more intuitive understanding of the image retrieval process, demonstrated through videos (S1 demo for cleaning the build plate and S2 demo for replacing the filter) in supplementary information. Table 2 highlights key system components and operations in response to the user query, illustrated with corresponding image references. Notably, this image retrieval approach operates without training or fine-tuning a pre-trained vision model, making it suitable for scenarios with limited image references.

We also present a case study demonstrating how implicit knowledge about alcohol concentration requirements in metal AM processes can be systematically captured and formalized. Using KG representation and expert interviews during process operations, we uncovered critical specifications typically undocumented in standard procedures. As shown in Table 3, our analysis reveals the precise concentration requirements for different applications and their underlying rationales. This systematic approach enables the transformation of implicit operational knowledge into explicit, structured information that can be readily shared and implemented across users. Additionally, a user survey was conducted to assess the system's efficiency, with detailed results provided in the supplementary information.

**Table 3 | Representation of implicit knowledge as new tuples in KG**

Process Element	Implicit Knowledge	Knowledge Graph Representation
Wiper Cleaning	Requires 99% alcohol concentration for complete particle removal	WiperCleaning  --hasConcentration: 99%  --hasPurpose: ParticleRemoval  --requiresJustification: ResiduePrevention
Recoater Cleaning	Needs 99% concentration to maintain coating quality	RecoaterCleaning  --hasConcentration: 99%  --hasPurpose: QualityControl  --requiresJustification: UniformityMaintenance
General Surface Cleaning	70–80% concentration sufficient for basic maintenance	SurfaceCleaning  --hasConcentration: 70–80%  --hasPurpose: BasicMaintenance  --requiresJustification: CostEfficiency

## Discussion

Using metal AM as a use case, our research demonstrates how a human-centric knowledge system can effectively bridge the gap between complex technical information and human understanding. The system's fundamental strength lies in its ability to evolve alongside human users, starting with explicit knowledge capture and gradually incorporating implicit knowledge through user interactions.

The knowledge acquisition follows a natural progression that aligns with human learning patterns. The system establishes a foundation through automated KG construction, systematically capturing explicit knowledge from formal technical documentation. The system accumulates implicit knowledge through continued user engagement, uncovering previously unexpressed insights and patterns that emerge from expert usage. Integrating MR technology further enables the system to capture implicit knowledge, addressing the traditionally challenging task of documenting and transferring practical, hands-on expertise.

While our experimental results demonstrate the system's capabilities, several limitations remain and must be acknowledged clearly. First, our current implementation demonstrates the concept using the Renishaw AM400 user guide but has not yet incorporated broader data sources, such as maintenance logs, training materials, and operator feedback. Expanding the knowledge base to include these sources would significantly enrich the implicit knowledge dimension. Second, although functional, our graph-based retrieval approach can benefit from further optimization. Techniques such as advanced node clustering could enhance retrieval accuracy and computational efficiency<sup>18</sup>. Third, the system's current multi-modal capabilities, particularly text-to-image retrieval, represent progress toward more natural HMI. However, achieving full multi-modal functionality and integrating advanced vision-language models<sup>5</sup> still presents significant

practical challenges. Balancing computational demands and practical constraints remains a key area for future exploration.

Looking ahead, we outline several directions to address these limitations and further enhance the MetalMind system. Immediate steps include broadening the knowledge base with additional data sources to capture more comprehensive implicit expertise and insights<sup>20</sup>. Next, implementing advanced graph optimization methods such as hierarchical clustering and embedding-based indexing, can further improve retrieval efficiency and accuracy. Additionally, future work should prioritize the exploration and integration of advanced multi-modal models, carefully evaluating computational trade-offs and practical viability. In the long term, integrating this knowledge system into a DT framework presents compelling opportunities for advancing human-centric manufacturing. Such integration would create a dynamic learning environment where operator actions, system responses, and environmental conditions continuously enrich the knowledge base. Ultimately, this would enable increasingly sophisticated decision support that is adaptable to varying user expertise levels and operational contexts.

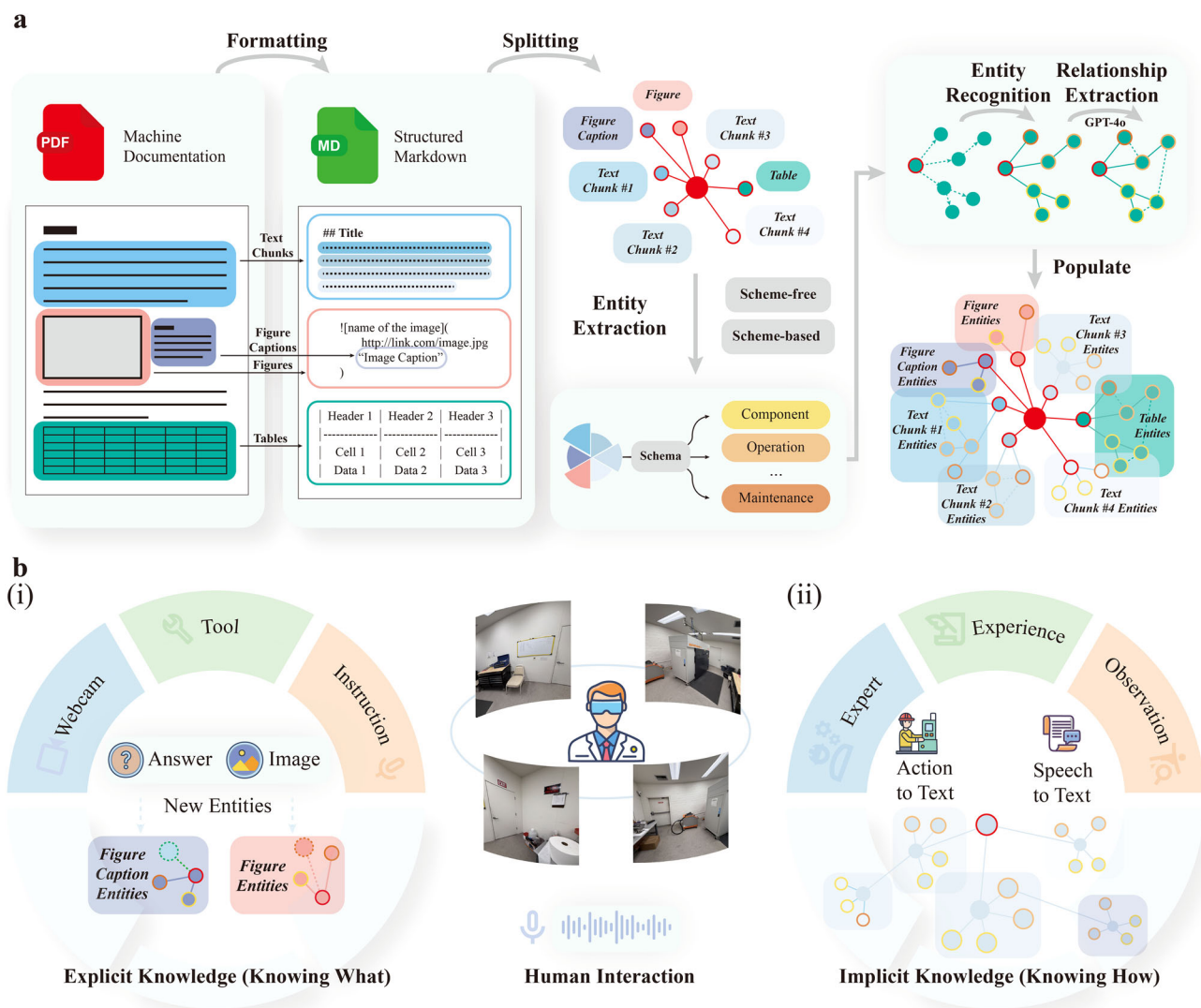
Our work demonstrates that prioritizing human cognitive patterns and learning processes in knowledge system design leads to more effective tools for complex manufacturing environments. The system's ability to capture

and integrate multiple types of knowledge, from explicit documentation to implicit operational expertise, makes advanced manufacturing technologies more accessible and manageable for users across the expertise spectrum. While our implementation focuses on metal AM, the underlying principles and architecture provide a blueprint for human-centric knowledge systems across various manufacturing domains, contributing to the broader evolution of smart manufacturing toward more intuitive and adaptive HMI.

## Methods

Our study focused on developing a human-centric knowledge system using the Renishaw AM400 metal AM machine as a case study. We selected this machine and its accompanying user guide because they comprehensively documented metal AM operations, providing an ideal foundation for developing structured, machine-readable content.

Figure 3 illustrates our proposed framework, which combines automatic KG construction powered by LLM with MR devices to create an immersive, human-centered learning environment. The framework supports explicit and implicit knowledge acquisition, reflecting the natural human learning process. For explicit knowledge (“knowing what”), the system utilizes action recognition tools to provide context-aware assistance,



**Fig. 3 | Overview of a human-centric knowledge system.** The system operates in two phases: **a** Knowledge graph (KG) construction using large language models (LLMs), incorporating expert feedback for quality assurance. **b** Multi-modal knowledge acquisition through mixed reality (MR) (i) direct interaction for

retrieving explicit knowledge with associated images and documentation, including video analysis for operational knowledge, and (ii) expert interaction and hands-on experience for developing implicit knowledge, with the system capturing operational patterns to enhance the KG's practical expertise representation.

enabling users to gain knowledge through natural, intuitive queries that match their thought processes. Implicit knowledge (“knowing how”) is captured through two channels that mirror human learning patterns: expert guidance, where expert instructions are converted to text and integrated into the KG as new entities, and hands-on experience, where user actions are transformed into text descriptions and incorporated as KG entities. This dual approach ensures the system captures theoretical knowledge and practical insights, techniques, and underlying rationales crucial for human understanding and skill development.

The system continuously evolves through user interaction, preserving interaction data such as captured images and their corresponding text descriptions to enrich the KG. These stored interactions serve as ground truth data and provide a foundation for developing multimodal language models<sup>5</sup>, enabling the system to better understand and respond to diverse human learning styles and needs. This adaptive capability ensures the knowledge system remains centered on human users’ evolving requirements and learning preferences.

### Automatic KG construction

Our method of automatic KG construction based on LLMs consists of three critical procedures, as shown in Fig. 3a: (1) preprocessing, which includes format transfer to more structured formats and splitting source documents into small chunks; (2) automatic KG construction through schema population; and (3) post-processing to ensure the quality of the KG.

**Preprocessing.** The first step involved converting the user guide from PDF to Markdown format to improve information organization and accessibility. This conversion allowed us to effectively segment the documentation into sections covering general machine descriptions, operational procedures, and maintenance protocols. To enhance image accessibility and integration with the KG, we extracted all images, uploaded them to a cloud service, and assigned each image a unique URL for efficient referencing and retrieval.

In constructing the KG, each Markdown file corresponds to a central node, with each node representing a specific section of the user guide. To organize the content, each file is divided into text chunks of 600 tokens with an overlap of 100 tokens between consecutive chunks. These text chunks serve as subnodes connecting to the file’s central node. This structure enhances traceability by allowing each chunk to link back to its original section within the file, thereby maintaining context when querying the KG.

**LLM-powered KG construction pipeline.** Following the creation of the node structure, the next step involves extracting entities and relationships from each subnode, as depicted in Algorithm 1. This extraction process operates in two modes: schema-free and schema-based. In the schema-free mode, entities and relationships are identified without predefined categories, resulting in a diverse yet unstructured collection of entities. Although flexible, this approach requires additional human validation to ensure the relevance of extracted entities.

### Algorithm 1. LLM-powered KG construction pipeline (post-processing omitted for clarity)

**Require:** Set of text chunks  $\mathcal{T} = \{T_1, T_2, \dots, T_n\}$   
**Ensure:** Structured knowledge graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- 1: **Initialization:**  $\mathcal{V} \leftarrow \emptyset, \mathcal{E} \leftarrow \emptyset$
- 2: **Phase 1: Schema-free entity extraction to derive schema categories**
- 3: **for all**  $T_i \in \mathcal{T}$  **do**
- 4:   Extract entities:  $\mathcal{V}_i^{\text{free}} \leftarrow \text{LLM} - \text{ExtractEntities}(T_i)$
- 5: **end for**
- 6: Aggregate all schema-free entities:  $\mathcal{V}^{\text{free}} = \bigcup_{i=1}^n \mathcal{V}_i^{\text{free}}$
- 7: Derive schema by clustering entities in  $\mathcal{V}^{\text{free}}$  into categories  $\mathcal{C} = \{C_1, \dots, C_k\} \triangleright$ , e.g., Component, Operation, Maintenance
- 8: **Phase 2: Populate KG using the derived schema**
- 9: **for all**  $T_i \in \mathcal{T}$  **do**

- 10:   Extract       categorized       entities       using  
                   schema:  $\mathcal{V}_i \leftarrow \text{LLM} - \text{ExtractBySchema}(T_i, \mathcal{C})$
- 11:   Extract       relationships       between       enti-  
                   ties:  $\mathcal{E}_i \leftarrow \text{LLM} - \text{ExtractRelations}(T_i, \mathcal{V}_i)$
- 12:    $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_i$
- 13:    $\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_i$
- 14: **end for**
- 15: **Embedding entities and visual nodes**
- 16: **for all** entity node  $v \in \mathcal{V}$  **do**
- 17:   Compute embedding:  $\mathbf{v}_{\text{emb}} \leftarrow \text{Embedding}(v.\text{description})$
- 18:    $v.\text{embedding} \leftarrow \mathbf{v}_{\text{emb}}$
- 19: **end for**
- 20: **for all** image  $I$  linked to text  $T_i$  **do**
- 21:   Create image node  $v_I$  with properties:
- 22:    $v_I.\text{description} \leftarrow \text{caption}(I), v_I.\text{image\_ref} \leftarrow \text{URL}(I)$
- 23:    $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_I\}$
- 24:   Connect node to related text nodes in  $\mathcal{V}$ :
- 25:    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_I, \text{refers\_to}, v_{\text{text}}) | v_{\text{text}} \in T_i\}$
- 26: **end for**
- 27: **return**  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

We initially conduct schema-free extraction to capture a comprehensive range of entities to balance flexibility and structure. Subsequently, we cluster these entities to derive an initial structured schema, categorizing entities into clearly defined types such as Component, Operation, and Maintenance. This resulting schema provides a systematic framework for consistently classifying entities throughout the KG. Once the schema is established, it is uniformly applied across all text chunks, facilitating standardized and structured entity extraction.

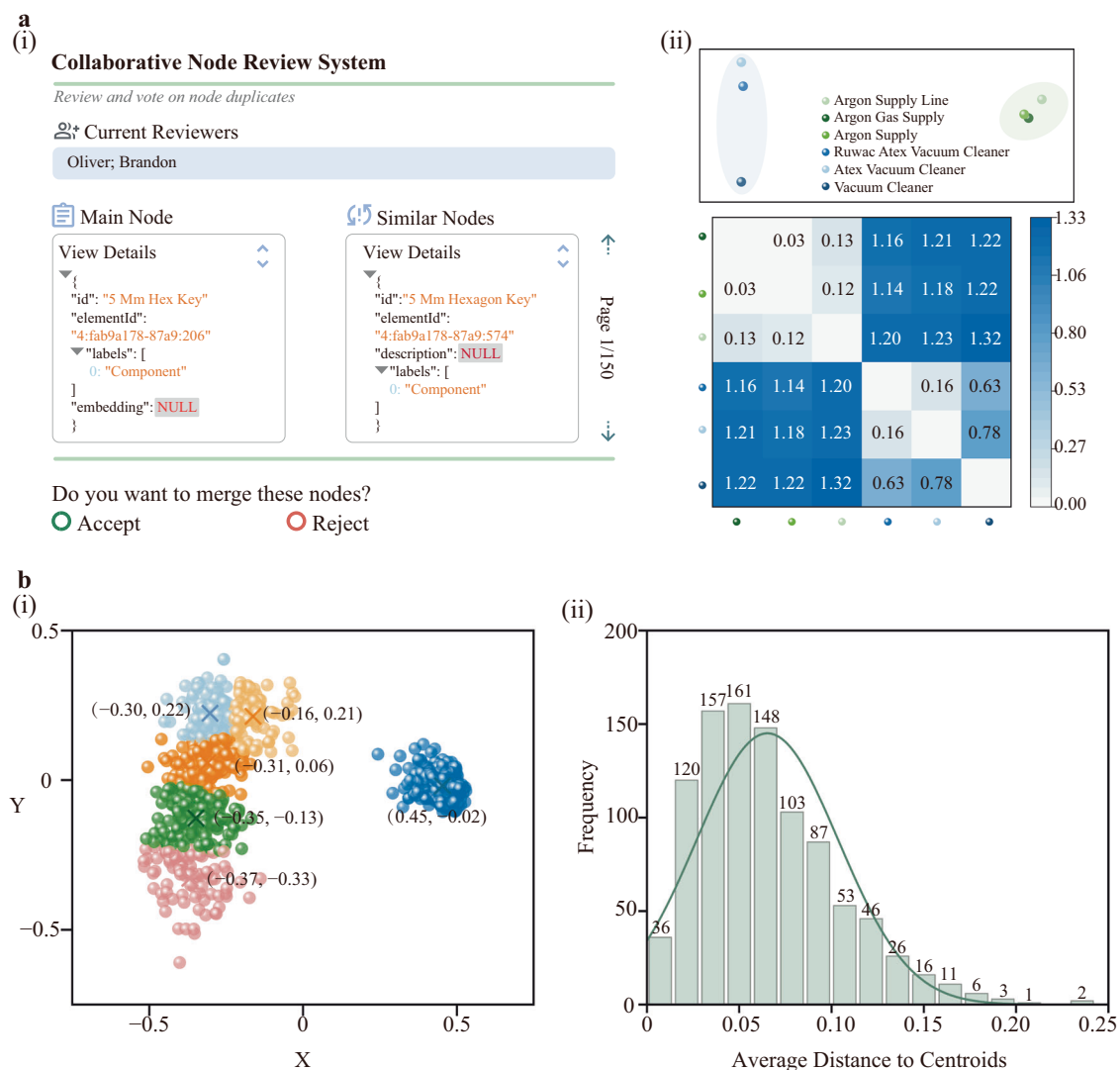
Entity extraction is performed individually for each text chunk, after which the categorized entities are aggregated to create the complete KG. To clearly define relationships among entities, we employ GPT-4o, prompting it to explicitly identify “head” and “tail” entities and specify their semantic relationships using standardized prompts with few-shot examples (an example prompt is presented in the supplementary information).

Each extracted entity includes a textual *description*, which is transformed into an embedding vector stored in the node’s *embedding* property. These embeddings significantly enhance the KG by enabling efficient semantic similarity-based queries and advanced knowledge retrieval. Additionally, visual information integration is achieved by representing images as distinct nodes connected to relevant textual nodes via reference edges. Each image node contains a *description* property storing the image caption and an *image\_ref* property containing the URL link to the image, thus enriching the KG with multimodal context.

**KG post-processing.** The post-processing stage includes removing the standalone nodes, resolving conflicts, and disambiguating duplicated nodes. Standalone nodes are entities with minimal or no connections to other nodes, which can reduce the overall coherence of the network. Given the need for domain-specific and expert knowledge to identify coreferences and resolve ambiguities, we implemented a collaborative approach for these tasks.

To identify duplicate nodes, we calculate the similarity of entities within the embedding space, as shown in Fig. 4b. Entities exceeding a predefined similarity threshold are flagged as potential duplicates. To facilitate the collaborative review process, we developed a web-based application that allows users to evaluate and confirm or reject duplicate nodes, as shown in Fig. 4a. Users can accept or reject the suggested duplicate nodes using their prior knowledge or the provided node descriptions as a reference. The final decision for each node is based on an average of decisions from multiple users, which helps maintain the accuracy and consistency of the review process.

By establishing a robust and accurate knowledge base, we mitigate common LLM issues, such as hallucinations and inaccuracies, thus ensuring generated responses are well-grounded and credible. Additionally, the collaborative nature of the post-processing phase ensures continuous



**Fig. 4 | Overview of the proposed collaborative post-processing method leveraging embedding similarity.** **a** Coreference resolution within the system illustrates the process of identifying and resolving referential ambiguities using embedding similarity. **(i)** The framework of the proposed collaborative node review system. **(ii)**

An example illustrating the underlying principle. **b** A 2D visualization of node embedding distributions. **(i)** PCA and K-means clustering represent the embeddings based on their similarities. **(ii)** A histogram depicting the distances between data points and their respective centroids.

adaptation and refinement of the KG through expert feedback. This adaptability not only minimizes inaccuracies but also reduces the likelihood of the system falling into problematic scenarios, such as question-answering loops, where insufficient or incorrect context might otherwise lead to circular reasoning patterns. Thus, the collaborative review and ongoing refinement of the KG form a critical foundation for building a human-centric knowledge system capable of reliable and contextually appropriate responses.

### Multi-faceted RAG

We developed an RAG framework for explicit knowledge acquisition, incorporating three specialized retrieval modes to enhance retrieval accuracy and improve response quality. Inspired by the principles outlined in GraphRAG<sup>18</sup>, each mode targets different dimensions of query relevance by using distinct retrieval techniques designed to meet diverse information needs within the metal AM printer operation. This tailored approach aims to optimize the performance of the RAG framework in a highly specialized technical context.

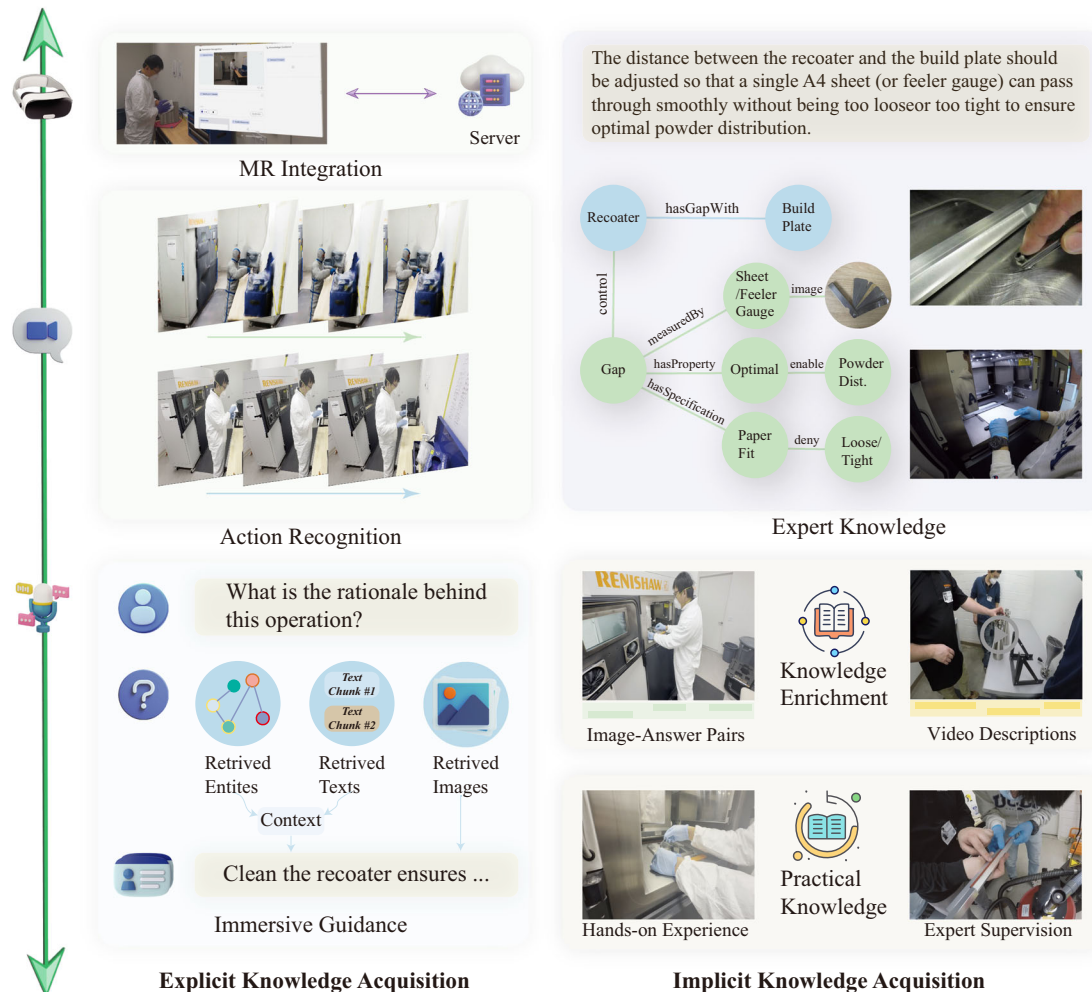
#### Mode 1: vector-based retrieval

The first retrieval mode leverages vector embeddings to represent the user's query and KG nodes specifically labeled as Document,

which contain relevant content extracted from the machine's user guide. Vector embeddings are high-dimensional numerical representations that grasp the semantic meaning of words or phrases, allowing for retrieval based on cosine similarity. This approach is well-suited for retrieving granular, contextually relevant information that aligns with varied user intents, ensuring precise information access within the metal AM domain.

#### Mode 2: graph traversal-based retrieval

The second retrieval mode uses graph traversal techniques to leverage relational insights within the KG. Unlike direct keyword- or vector-based searches focusing on semantic similarity, graph traversal is designed to uncover indirect relationships and contextual dependencies within the network of nodes. This method is particularly valuable for queries requiring understanding connected information, such as multi-step processes or dependencies between components within metal AM operations. The retrieval process begins with an initial set of relevant nodes identified through a keyword or vector search, after which traversal expands this set by exploring nodes linked through meaningful relationships. By examining these connections, this approach uncovers broader, interconnected information that enhances the user's understanding of complex, layered topics in metal AM.



**Fig. 5 | The MR device enables immersive interaction through tools like action recognition, with the model running on a server.** The system builds human-centric knowledge using a KG backend that provides domain-specific explicit knowledge.

This interaction process enriches image-text and video-text pairs, while implicit knowledge is captured via expert guidance and hands-on operations.

### Mode 3: hybrid retrieval

The third retrieval mode, hybrid retrieval, combines the strengths of vector-based and graph traversal methods to deliver a more comprehensive and contextually rich response. This approach begins with vector-based retrieval to identify nodes closely aligned with the semantic meaning of the query. It then expands this set through graph traversal, exploring adjacent nodes and their relationships within the KG. This hybrid approach ensures a robust retrieval process that enhances response quality by integrating the semantic depth provided by vector embeddings with the structured relational insights from graph traversal.

In addition to the three retrieval modes, we implemented an image retrieval component to assist users in understanding various processes in metal AM, such as powder loading and wiper installation. This image retrieval system follows a two-step method. First, we assess whether the retrieved entities have associated entities labeled as **Figure**. In the second step, after generating the final answer, we summarize the response and compute the similarity between each summarized point and the embedding of the original image description. If the similarity score exceeds 0.85, we proceed with image retrieval. This two-step approach ensures all relevant images are retrieved, enriching user comprehension and enhancing the information's contextual relevance.

### Human-centric knowledge system

We integrated a video-based action recognition model and a MR device to evaluate the KG's practical utility as an AI agent tool in metal AM. The

action recognition model, trained using synthetic data derived from a DT of our lab environment built on the NVIDIA Omniverse platform, identifies the current operation being performed. This enables users to ask operation-specific queries, enhancing the precision of decision-making.

The MR device facilitates intuitive interaction by acting as an interface that displays insights and recommendations derived from the KG. In addition, the device supports audio-based input and output, creating a more immersive user experience. The integration ensures that users receive expert-level guidance tailored to their specific tasks. Figure 5 provides an overview of the workflow, highlighting the interaction between the action recognition model, the KG, and the MR device. The flowchart illustrates how the MR device identifies actions, retrieves domain-specific knowledge from the KG, and presents the information to the user in a seamless and accessible manner.

### Data availability

The evaluation dataset is publicly available in the GitHub repository: <https://github.com/FHL1998/MetalMind>.

Received: 5 February 2025; Accepted: 12 May 2025;

Published online: 16 June 2025

### References

- Xu, X., Lu, Y., Vogel-Heuser, B. & Wang, L. Industry 4.0 and industry 5.0— inception, conception and perception. *J. Manuf. Syst.* **61**, 530–535 (2021).

2. Rame, R., Purwanto, P. & Sudarno, S. Industry 5.0 and sustainability: An overview of emerging trends and challenges for a green future. *Innov. Green. Dev.* **3**, 100173 (2024).
3. Fan, H. et al. New era towards autonomous additive manufacturing: A review of recent trends and future perspectives. *Int. J. Extrem. Manuf.* **7**, 032006 (2025).
4. Chow, E. et al. Collaborative moonwalkers. *2024 IEEE Aerospace Conference*. 1–15 (2024).
5. Fan, H. et al. Enhancing metal additive manufacturing training with the advanced vision language model: A pathway to immersive augmented reality training for non-experts. *J. Manuf. Syst.* **75**, 257–269 (2024).
6. Fan, H. et al. AutoMEX: Streamlining material extrusion with ai agents powered by large language models and knowledge graphs. *Mater. Des.* **251**, 113644 (2025).
7. Fan, H., Liu, X., Fuh, J. Y. H., Lu, W. F. & Li, B. Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics. *J. Intell. Manuf.* **36**, 1141–1157 (2025).
8. Gkourmelos, C., Konstantinou, C. & Makris, S. An LLM-based approach for enabling seamless human-robot collaboration in assembly. *CIRP Ann. Manuf. Technol.* **73**, 9–12 (2024).
9. Wang, T., Fan, J. & Zheng, P. An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing. *J. Manuf. Syst.* **75**, 299–305 (2024).
10. Fan, H., Fuh, J., Lu, W. F., Kumar, A. S. & Li, B. Unleashing the potential of large language models for knowledge augmentation: A practical experiment on incremental sheet forming. *Procedia Computer Sci.* **232**, 1269–1278 (2024).
11. Zhao, S. et al. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924* (2024).
12. Li, J., Yuan, Y. & Zhang, Z. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446* (2024).
13. Xiong, C., Xiao, J., Li, Z., Zhao, G. & Xiao, W. Knowledge graph network-driven process reasoning for laser metal additive manufacturing based on relation mining. *Appl. Intell.* **54**, 11472–11483 (2024).
14. Shi, X., Tian, X., Ma, L., Wu, X. & Gu, J. A knowledge graph-based structured representation of assembly process planning combined with deep learning. *Int. J. Adv. Manuf. Technol.* **133**, 1807–1821 (2024).
15. Du, K. et al. Relation extraction for manufacturing knowledge graphs based on feature fusion of attention mechanism and graph convolution network. *Knowl.-Based Syst.* **255**, 109703 (2022).
16. Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F. & Li, B. Knowledge extraction for additive manufacturing process via named entity recognition with LLMs. *Robot. Computer-Integr. Manuf.* **93**, 102900 (2025).
17. Sanmartin, D. KG-RAG: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035* (2024).
18. Edge, D. et al. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).
19. Es, S., James, J., Anke, L. E. & Schockaert, S. RAGAs: Automated evaluation of retrieval augmented generation. *EACL2024*. 150–158 (2024).
20. Fan, H. et al. MaViLa: Unlocking new potentials in smart manufacturing through vision language models. *J. Manuf. Syst.* **80**, 258–271 (2025).

## Acknowledgements

This research was primarily supported by the MUREP High Volume project (Grant no. 80NSSC22M0132) funded through the U.S. NASA Office of STEM Engagement, and the SMART ITAC project (Grant No. DE-EE0009726) funded through the U.S. Department of Energy Office of Manufacturing and Energy Supply Chains. We extend our gratitude to Prof. Larry Smarr and Prof. Thomas DeFanti from the University of California, San Diego for their support in HyperCluster computing through the San Diego Supercomputer Center (SDSC) National Research Platform (NRP) Nautilus, which is sponsored by the National Science Foundation (NSF) under Award Nos. 2100237 and 2120019.

## Author contributions

H.F.: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Z.F.: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. C.L.: Writing – original draft, Visualization, Data curation. J.Z.: Writing – original draft, Formal analysis, Data curation. T.G.: Writing – review & editing, Resources. J.Y.H.F.: Writing – review & editing, Supervision. W.F.L.: Writing – review & editing, Supervision. B.L.: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s44334-025-00038-9>.

**Correspondence** and requests for materials should be addressed to Bingbing Li.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025