Foundation-Model-Based Agents in Industrial Automation: Purposes, Capabilities, and Open Challenges

Vincent Henkel[1*†], Felix Gehlhoff[1†], David Kube[2], Asaad Almutareb[3], Luis Cruz[4], Bernd Hellingrath[5], Philip Koch[6], Christoph Legat[7], Florian Mohr[8], Michael Oberle[9], Felix Ocker[10], Thorsten Schoeler[11], Mario Thron[12], Nico Andre Töpfer[6], Lucas Vogt[13], Yuchen Xia[14]

[1*]Institute of Automation Technology, Helmut Schmidt University / University of the Federal Armed Forces Hamburg, Hamburg, Germany.
[2]Siemens AG, Nuremberg, Germany; Institute for Technologies and Management of Digital Transformation, Bergische Universität Wuppertal, Germany.
[3]Artiquare GmbH, Ingolstadt, Germany.
[4]Facultad de Ingeniería Mecánica, Electrónica y Biomédica (FIMEB), Universidad Antonio Nariño, Bogotá, Colombia.
[5]Chair of Information Systems and Supply Chain Management, University of Münster, Münster, Germany.
[6]Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM, Stade, Germany.
[7]Research Group on Cognitive Autonomy & Predictive Intelligence, Faculty of Electrical Engineering, Technical University of Applied Sciences Augsburg, Augsburg, Germany.
[8]Birkenfeld Institutes of Technology, Trier University of Applied Sciences, Birkenfeld, Germany.
[9]Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany.
[10]Honda Research Institute Europe, Offenbach am Main, Germany.
[11]Faculty of Computer Science, Augsburg Technical University of Applied Sciences, Augsburg, Germany.

[12]Institute for Automation and Communication (ifak), Magdeburg, Germany.
[13]Process-to-Order Group, TUD Dresden University of Technology, Dresden, Germany.
[14]Institute for Industrial Automation and Software Engineering, University of Stuttgart, Stuttgart, Germany.

*Corresponding author(s). E-mail(s): vincent.henkel@hsu.hamburg;
Contributing authors: felix.gehlhoff@hsu.hamburg;
david.kube@siemens.com; asaad.almutareb@artiquare.com;
luicruz@uan.edu.co; bernd.hellingrath@ercis.uni-muenster.de;
philip.koch@ifam.fraunhofer.de; christoph.legat@tha.de;
f.mohr@umwelt-campus.de; michael.oberle@ipa.fraunhofer.de;
felix.ocker@honda-ri.de; thorsten.schoeler@tha.de; mario.thron@ifak.eu;
nico.andre.toepfer@ifam.fraunhofer.de; lucas.vogt@tu-dresden.de;
yuchen.xia@ias.uni-stuttgart.de;
†These authors contributed equally to this work.

Abstract

Foundation models, particularly large language models, are increasingly integrated into agent architectures for industrial tasks such as decision support, process monitoring, and engineering automation. Yet evidence on their purposes, capabilities, and limitations remains fragmented across domains. This work examines how mature foundation-model-based agent systems are in industrial contexts, how their functional profile differs from conventional agent systems, and which limitations persist. A systematic literature survey following the PRISMA 2020 guideline is presented, screening 2 341 publications and synthesising a corpus of **88** publications through a structured coding scheme. The results show that reported systems are predominantly at prototype and early validation stages (75.0 % at TRL 4–6), with deployment-oriented evidence remaining rare (9.1 %). Operational goals are most frequently positioned in user assistance, monitoring, and process optimisation, while conventional production-control purposes such as planning and scheduling are less prominent. Compared with an established baseline for industrial agent systems, the capability profile reveals substantial gains in human interaction (+37 %) and dealing with uncertainty (+35 %), but a pronounced deficit in negotiation (−39 %). The most widely reported limitations concern lack of generalization, hallucination and output instability, data scarcity, and inference latency. A working definition of *foundation-model-based industrial agents* is also proposed, bridging conventional agent theory, automation-engineering standards, and the foundation-model paradigm.

Keywords: foundation models, large language models, multi-agent systems, industrial automation, systematic literature review

Abbreviations

DT Digital Twin
FM Foundation Model
HMI Human–Machine Interaction
LLM Large Language Model
MAS Multi-Agent System
MCP Model Context Protocol
MLLM Multimodal Large Language Model
PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG Retrieval-Augmented Generation
TRL Technology Readiness Level

1 Introduction

Software agents and Multi-Agent Systems (MASs) have been studied for decades as a design paradigm for distributed decision-making in industrial domains such as production control, logistics, and process engineering [1, 2]. In conventional settings, i.e. agent systems that predate the integration of Foundation Models (FMs), agents are typically rule-based or optimisation-driven entities that pursue locally specified objectives under predefined interaction protocols such as the Contract Net Protocol [3]. With the emergence of FMs, and Large Language Models (LLMs) in particular, a new class of agent systems has gained momentum [4]. Unlike their conventional counterparts, FM-based agents can interpret unstructured and noisy information such as natural-language instructions, maintenance logs, and visual or sensor-stream inputs, interact with human operators through conversational interfaces, and orchestrate heterogeneous tool chains by generating executable code or application programming interface calls through flexible reasoning [5]. According to Ren et al. [6], FMs-based industrial agents comprise different levels of technological capabilities: LLM-agents primarily extend language-centric reasoning and tool use, Multimodal Large Language Model (MLLM)-agents add multimodal perception across textual, visual, and sensor data, whereas Agentic AI refers to a further step toward self-directed, goal-driven autonomy in dynamic environments. The shift from conventional MAS to FM-based agent systems is not only a shift in realised capabilities but also in the underlying coordination logic. Ali et al. [7] point out the constrast between symbolic coordination through explicit protocols such as the Contract Net Protocol or blackboard systems with neural coordination based on structured conversation, role-based workflows, and prompt-driven orchestration.

Thus, even though there have been attempts to address challenges such as interaction with human operators, tool orchestration, and general optimisation applications in the pre-FM era, these new technologies make such systems much more accessible, easier to develop and maintain, as well as considerably more capable and adaptive.

These developments have led to rapid adoption across a broad range of industrial applications, from engineering design automation and shop-floor control to energy-system operation and information-technology infrastructure management [8–10]. At the architectural level, LLM-based agents are commonly integrated as central cognitive components that interpret context, generate plans or recommendations, and invoke external tools, often augmented by retrieval mechanisms to ground decisions in domain-specific knowledge [5, 10]. Recent surveys on LLM-based agents in general-purpose settings have mapped

these architectural patterns, reasoning strategies, and tool-use mechanisms [4, 5, 11]. However, the degree to which these general findings transfer to industrial contexts, where safety, determinism, and integration with legacy systems impose additional constraints, remains an open question [12].

Despite this growing adoption, there is no established working definition that bridges conventional agent concepts and standards, such as autonomous action in an environment [1] or encapsulated entities with control objectives [13], with the FM paradigm, resulting in possible confusion with approaches that merely use LLMs for text generation or conversational interaction. Recent surveys acknowledge this gap: Jin et al. [14] note that among researchers there is no "clear distinction between LLMs and LLM-based agents" and that "unified standard and benchmarking" remain in an early stage, while Zhou et al. [15] call for a "unified taxonomy" to address the currently "fragmented approach" to classifying FM-based agent architectures. A related problem refers to "conceptual retrofitting", i.e., the tendency to describe modern LLM-based systems using classical agent concepts such as Belief-Desire-Intention (BDI) or perceive–plan–act–reflect loops, despite substantial differences in their operational mechanisms [7]. This lack of conceptual consolidation is also evident in recent manufacturing-focused literature, which explicitly notes that the definitions, capability boundaries, and interconnections of LLM-agents, MLLM-agents, and Agentic AI remain insufficiently clarified [6]. Compact operational characterisations have been proposed, e.g., defining agentic LLMs as systems that "reason, act, and interact" [16], yet no consolidated definition exists that integrates these perspectives with automation-engineering standards. Without such a definition, a systematic literature review lacks a reproducible inclusion criterion.

At the same time, the level of maturity of FM-based agents is unclear, since evidence on such systems is fragmented across application domains, levels of technological maturity, and evaluation practices. Domain-specific surveys confirm that the integration of LLMs in manufacturing is "still in its initial stages" [17], and a recent meta-survey on agent evaluation describes the field as a "complex and underdeveloped area" and aims to bring "clarity to the fragmented landscape of agent evaluation" [18]. Individual publications report prototypes in manufacturing [8, 19], logistics [20], energy systems [10], or engineering design [9], but no consolidated overview maps the capabilities and maturity of these approaches to a common framework. Another study points out that existing industrial FM-based agent approaches appear to cluster around assistive and task-oriented roles, whereas stronger forms of self-directed goal formulation and system-level orchestration are still largely discussed as an emerging Agentic AI vision rather than as established industrial practice [6]. As a consequence, it remains difficult to assess the overall technological maturity of the field, to identify recurring limitations across domains, and to establish a comparable evaluation basis, leaving practitioners in doubt about the capabilities and merits of adoption of FM-based agent systems in industrial contexts.

Furthermore, it is unclear how FM integration changes the functional profile of industrial agents compared to classical MAS. Earlier systematic work on software agents in industrial production [2] provides a baseline of purposes (e.g., planning, scheduling, control) and capabilities (e.g., negotiation, coordination, reactivity). Initial evidence suggests a shift: Greis et al. [21] explicitly "contrast the capabilities of classic autonomous software agents and LLM software agents" in a digital-twin-enabled manufacturing context, and Zhao et al. [22] observe that conventional agent negotiation relies on "pre-defined and fixed heuristic rules" that are ill-suited to dynamic disturbances, motivating a multimodal LLM-based alternative. This suggests that the two paradigms may be complementary rather than FM-based agents being the successor of conventional approaches. However, this has not been systematically examined across domains.

Against this background, the guiding research questions (RQs) of this work are:

- **RQ1:** How can FM-based industrial agents be defined, and what is the current maturity of such systems in industrial and industrially relevant research, including their technology readiness, application domains, and use cases?
- **RQ2:** Which system purposes and capabilities do FM-based agents exhibit, and how does their functional profile differ from conventional industrial agent systems?
- **RQ3:** Which limitations, challenges, and future work directions are most frequently reported for FM-based agent systems in industrial contexts?

To address these questions, this work follows a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style systematic review (Section 3). For each included publication, technology readiness, application domain, system purposes, capabilities, reported limitations, and future work directions are assessed and consolidated.

The main contributions of this work are threefold. First, it proposes a working definition of FM-based technical agents that bridges conventional agent theory, automation-engineering standards, and the FM paradigm, and evaluates whether this definition adequately captures the systems reported in the corpus (Section 5.1). Second, building on this definition, it provides a cross-domain synthesis of technological maturity, system purposes, and capability profiles of FM-based agents, including a descriptive comparison with an established baseline for industrial agent systems. Third, it consolidates recurring limitations and future work directions into structured themes that can inform the design, evaluation, and deployment of FM-based agent systems in industrial contexts.

2 Related work

This section relates the present work to three streams of prior research: industrial software agents and MASs (Section 2.1), FM-based agents in industrial contexts (Section 2.2), and existing taxonomies for purposes and capabilities (Section 2.3). Section 2.4 introduces the baseline used for comparative analysis.

2.1 Industrial software agents and multi-agent systems

Industrial software agents and MASs have long been studied as a design paradigm for distributed decision-making and control in production and related industrial domains [23]. In conventional

settings, agent-based approaches are typically motivated by the need to decompose complex objectives into locally manageable sub-problems (e.g., planning and scheduling, dispatching decisions, shop-floor control, and diagnosis and fault management) and to coordinate these decisions across heterogeneous resources under operational constraints [24, 25]. At the same time, industrial applications impose strong non-functional requirements, including determinism, safety, and real-time suitability, which shape how autonomy and interaction mechanisms can be realized in practice [26, 27].

From a conceptual perspective, Russell and Norvig [28] provide a broad characterization of an agent as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators", a definition that is deliberately paradigm-agnostic and accommodates both symbolic and subsymbolic realizations. Within the more specific MAS tradition, Wooldridge [1] emphasizes autonomous action, reactivity, and coordination among multiple entities, including interaction patterns such as negotiation where applicable. In industrial survey work, these conventional expectations are operationalized as concrete purposes and capabilities that can be evidenced from application reports in production contexts [2]. Traditional applications include, for example, heuristics-based decentralised process planning and scheduling as well as controlling resources and diagnosis tasks on the shop floor [25].

2.2 Foundation-model-based agents in industrial contexts

Recent FM-based industrial agent systems build on these foundations but often repurpose them towards assistive and decision-support-centric roles (e.g., maintenance decision support or manufacturing assistance) rather than towards fully decentralised negotiation-heavy control [8, 9, 19]. FMs, and in particular LLMs, provide generic language understanding and generation capabilities that can be adapted to various downstream tasks [29]. In industrial and industrially relevant agent systems, LLMs are commonly integrated as central cognitive components that (i) interpret human instructions and contextual information, (ii) generate plans or recommendations, and (iii) orchestrate tool-mediated actions by interacting with external software components (e.g., databases, simulation models, code generators, or domain services) [30]. A recurring architectural pattern is to complement LLMs with explicit grounding mechanisms such as Retrieval-Augmented Generation (RAG) to improve factuality and to connect agent decisions to domain-specific knowledge sources [10]. Ren et al. [6] also emphasize that industrial tasks also depend on the joint interpretation of, e.g., enterprise data, maintenance records, sensor streams, and machine-vision inputs, which is why multimodal LLM-based agents are increasingly discussed as a relevant architectural extension that supports context-aware perception, diagnosis, and decision support.

Prior work has noted limitations related to robustness, reliability, and external validity, particularly when LLMs are used for decision-making under

incomplete information or when generated outputs must be executable in technical systems. Practical constraints such as latency, cost, and integration effort are also discussed [8, 9, 19, 31].

2.3 Taxonomies for purposes, properties, and capabilities

Systematic comparison across agent systems requires a representation that separates *what* a system is intended to achieve from *how* it achieves it. Müller et al. [32] characterise industrial autonomous systems through four higher-level dimensions—*systematic process execution*, *adaptability*, *self-governance*, and *self-containedness*—and relate these to lower-level abilities such as learning ability, decision-making capacity, cooperability, reactivity, and self-explanation. Since these relations are many-to-many rather than one-to-one, their framework suggests that higher-level system qualities should be analysed separately from the concrete functional means by which they are realised. Kaber [33] provides an additional conceptual basis for separating higher-level system qualities from concrete functionality by distinguishing *automation* from *autonomy*. Rather than treating autonomy as a higher level of automation, he defines it through three conditions—*viability*, *independence*, and *self-governance*—and further operationalises the distinction in terms of the demands a system places on its environment, human collaborators, and task protocols. In this work, the analysis follows the purpose–property–capability framing employed by Reinpold et al. [2], which builds on the work by Müller et al. [32]. In this framing, *system purposes* describe operational goal categories (e.g., planning, scheduling, control, monitoring, user assistance), *capabilities* describe concrete functional skills evidenced by the system (e.g., interaction, communication, coordination, reasoning), and *properties* aggregate capability evidence into higher-level dimensions that support corpus-level comparison (see Section 3.1 for the working definition used in this review). This distinction is particularly useful for FM-based agent systems because LLMs can simultaneously shift the operational emphasis towards assistive purposes and affect the evidencing of capabilities such as human interaction and uncertainty handling.

2.4 Baseline for comparative analysis

Reinpold et al. [2] systematically compare industrial software agents and Digital Twins (DTs) in production contexts through a PRISMA-aligned literature review covering 145 publications. Their work provides the purpose and capability taxonomy adopted in this review (see Section 2.3) together with composite scores that summarize capability profiles at property level, thereby establishing a quantitative reference point for corpus-level comparison. Given the different temporal scopes of the two reviews, with Reinpold et al. [2] largely predating the broad adoption of FMs in industrial agent research, substantial overlap between the two corpora is unlikely, and the two studies can be viewed as largely complementary samples of conventional and FM-based agent research, respectively.

In this work, Reinpold et al. [2] is used as a baseline for descriptive comparison, enabling the computation of comparable coverage profiles and differences in percentage points. Differences in corpus composition and inclusion criteria

between the two studies should be considered when interpreting the comparison (see Section 5.3).

3 Method

This work follows the PRISMA 2020 guideline for systematic literature reviews [34] and adapts it to the analysis of FM-based agent systems in industrial and industrially oriented research. The procedure comprises the standard PRISMA stages of (1) identification of potentially relevant records via a structured multi-database query and (2) relevance filtering according to predefined eligibility criteria, followed by (3) a consolidation step in which the resulting corpus is transformed into a consistent, machine-readable representation for quantitative analysis.

3.1 Working definition

Agents have a long-standing history within academia [1]. While these classical foundations provide the essential basis for this review, the modern literature often uses the term "agent" interchangeably with related concepts such as agentic workflows or simply "LLMs". To maintain a clear scope, this work distinguishes its focus from these adjacent terms and provides a working definition that is (i) grounded in established agent concepts, (ii) compatible with automation-engineering notions of technical agents, and (iii) explicit about what it means for a system to be "FM-based".

In the agent literature, Wooldridge [1] defines an agent as "a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives". In the context of automation engineering, the VDI/VDE 2653 guideline defines a technical agent as "an encapsulated hardware or software entity with specified objectives regarding the control of a technical system or a part thereof" [13]. Foundation models, in turn, are characterized as models "trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" [29]. According to Ren et al. [6], systems with high degrees of autonomy that use these FMs exhibit capabilities such as semantic retrieval and context-awareness, adaptive reasoning, and autonomous decision-making, which includes the realisation of selected actions, i.e. influencing the actual process.

Combining these streams and concepts, the following definition is proposed that guides the focus of this work:

Working definition used in this review

A foundation-model-based industrial agent is an encapsulated hardware or software entity acting in the context of an industrial system that is capable of autonomous action in order to meet its specified design objectives, and that uses a foundation model as a central component for context interpretation, decision-making, as well as action selection and execution.

Throughout the review, each category is assessed based on paper-level evidence: if the publication does not provide sufficient detail, the corresponding category is treated as not evidenced.

3.2 Data sources and search strategy

Records were retrieved from four bibliographic and preprint sources: Scopus, Semantic Scholar, arXiv, and OpenAlex. The search covered publications from 2020 onwards and targeted LLMs and related FMs in agent or multi-agent settings within industrial and industrially relevant domains, including manufacturing, logistics, energy systems, and the engineering life cycle (product development and process engineering), as well as cross-domain functionalities such as maintenance, quality management, and Human–Machine Interaction (HMI). The core query combined (i) terms for LLMs and FMs, (ii) terms for agents and MASs, and (iii) terms for industrial contexts. Across all sources, 3025 valid results and 2341 unique records were obtained.

The composite search query is shown in Table 1. The search and export were executed on September 8, 2025.

3.3 Screening and eligibility

Eligibility is defined at the level of individual publications. A record is included if it meets all of the following criteria:

1. The publication describes a concrete application, implementation, or case study (not a purely conceptual contribution, high-level vision paper, survey, or review without practical realisation or empirical results).
2. The system uses one or more LLMs, MLLMs, or related FMs for control or decision-making within an agent or agent-based architecture.
3. The application context is industrial or industrially relevant, covering the above-mentioned domains, engineering life cycle phases, or cross-domain functionalities.

Given the size of the initial corpus, the title and abstract screening was supported by an LLM, which assigned each record a relevance score against the eligibility criteria stated above. The inclusion threshold was calibrated on manually annotated samples, and a subset of below-threshold records was reviewed manually to control for false negatives. Records above the threshold were forwarded to full-text assessment, where final eligibility was determined by the authors.

After title and abstract screening, a full-text assessment determines whether the use of an agent or MAS qualifies as a *foundation-model-based technical agent* according to the working definition in Section 3.1. At this stage, publications that constitute surveys or reviews without an original application contribution were excluded. The remaining full texts were assessed against the working definition, verifying that the system exhibits autonomous behaviour, beyond pure text generation and that a FM is involved in the decision loop (e.g., interpreting context, planning, selecting actions, or orchestrating tool calls).

Table 1 Search query used for multi-database retrieval. The query is composed of three conceptual blocks connected by Boolean AND.

Block	Terms
FM/LLM	`"LLM" OR "large language model*" OR "foundation model*" OR "generative AI" OR "agentic AI" OR AI OR "Diffusion Model" OR "Vision Language Action Model" OR VLA OR "vision-language-action model" OR "World Model"`
Agent	`agent OR "intelligent agent" OR "multi-agent system" OR MAS`
Industrial context	`industr* OR manufactur* OR production OR factory OR shopfloor OR automat* OR logistic* OR mainten* OR quality OR "process engineering" OR "process industry" OR energy OR "energy system*" OR develop* OR "human-machine interaction" OR HMI OR "cyber-physical system*" OR "Simulation Model"`

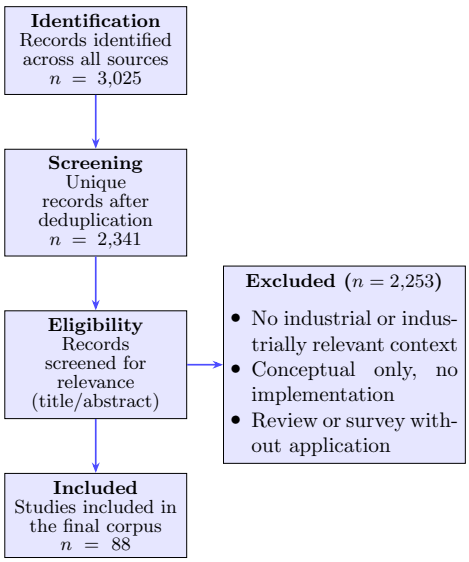

Fig. 1 PRISMA-style flow of records for the identification, screening, and inclusion of publications on FM-based agent systems in industrial and industrially relevant contexts.

The overall identification, screening, and inclusion process is summarized in the PRISMA flow diagram in Figure 1. The title and abstract screening have been supported by an AI tool chain, i.e. an LLM was used to assess the probability of a title matching the above-stated criteria. Afterwards, the threshold has been determined by looking at different samples, resulting in a corpus comprising $N = 88$ included publications. By publication type, conference papers dominate (49/88, 55.7%), followed by preprints (20/88, 22.7%) and journal articles (19/88, 21.6%). Nearly all publications date from 2024 (38/88) or 2025 (42/88), confirming that FM-based agent systems in industrial contexts are a rapidly emerging research area.

3.4 Taxonomies and operationalization

Following the purpose–property–capability framing introduced in Section 2.3, this work adopts the taxonomy of Reinpold et al. [2] to enable a transparent and comparable characterization of agent systems across publications.

Based on this distinction, each included publication is mapped to a fixed taxonomy of operational purposes (planning, scheduling, dispatching, control, diagnosis and fault management, user assistance, monitoring, virtual commissioning, and process optimisation). In addition, reported system capabilities are classified according to the capability

taxonomy and aggregated into higher-level property dimensions (sociability, autonomy, intelligence, and fidelity) to support corpus-level comparison. Classification is performed at paper level: each category is marked as *supported* if explicit textual evidence is provided in the publication, and as *not evidenced* otherwise. Because a single system can serve multiple operational goals and exhibit multiple capabilities, purpose and capability assignments are treated as multi-label indicators.

3.5 Comparative analysis and aggregation

Comparative analysis against the baseline survey [2] is performed in a descriptive manner. Using the shared taxonomy, relative occurrence shares are compared by reporting differences in percentage points. Higher-level property dimensions are computed as weighted averages of the underlying capability indicators, following the weighting scheme defined by Reinpold et al. [2]. Reported limitations and future work directions are grouped into thematic categories.

4 Results

This section details the results of the corpus analysis ($N = 88$), structured along the three research questions introduced in Section 1. Section 4.1 addresses RQ1 by establishing the descriptive statistics of the field, focusing on Technology Readiness Levels (TRLs) and application domains. Section 4.2 addresses RQ2 through a comparative analysis of system purposes and capabilities. Finally, Section 4.3 addresses RQ3 by consolidating the reported limitations and future work directions.

4.1 Technological maturity and application domains

Addressing RQ1, this section characterises the technological maturity of the corpus along the TRL scale and its distribution across application domains.

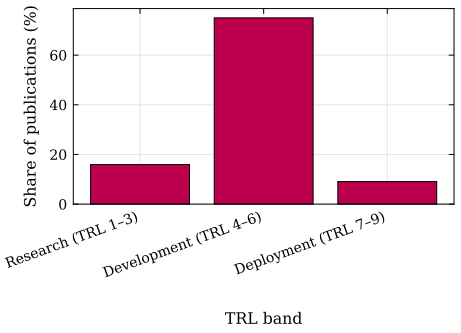

Fig. 2 Aggregated technology readiness levels (TRL bands) of FM-based technical agents across the corpus.

Technology readiness is assessed using the nine-level TRL scale [35], aggregated into three bands: research (TRL 1–3, basic principles to proof of concept), development (TRL 4–6, laboratory validation to demonstration), and deployment (TRL 7–9, system prototype to operational use).

As shown in Figure 2, research-stage contributions (TRL 1–3) account for only 14/88 (15.9%). This low share should be read in light of the inclusion criteria (Section 3.3), which require a concrete application, implementation, or case study and therefore tend to filter out purely conceptual contributions. Beyond this, the relatively high share of TRL 4+ systems suggests that laboratory-stage prototypes are commonly reached and feasible with FM-based approaches. Accordingly, the majority of reported systems fall into the development band (TRL 4–6; 66/88,

75.0%), indicating that most contributions present laboratory prototypes or early-stage validations. Deployment-oriented evidence (TRL 7–9), however, remains rare (8/88, 9.1%), showing that despite the ability to quickly develop lab-scale demonstrators, the step to deployment in operation is more difficult and has rarely been achieved.

To characterize the industrial landscape addressed by the corpus, each publication is assigned to one of four application domains. The first three domains are inspired by the application scenarios of *Plattform Industrie 4.0* [36]; the fourth was added inductively to accommodate energy systems, IT operations, and transportation use cases that fall outside the original manufacturing-centric scope:

Production & Manufacturing Control covers planning, control, and optimisation of production and manufacturing processes.
Logistics & Supply Chain covers coordination, decision support, and automation in logistics and supply chain operations.
Engineering & Design Automation covers technical development, simulation, design automation, and engineering workflows.
Industrial Operations & Networks covers energy systems, grid and network operations, IT and cloud operations, transportation, and cross-domain industrial decision support.

As shown in Figure 3, the largest cluster is *Industrial Operations & Networks* (34/88, 38.6%), covering power grid operation and voltage control [37–42], energy management and charging infrastructure [43–46], cloud and IT operations [47–51], transportation and autonomous vehicles [52–54], building management [55, 56], telecommunications [57, 58], and cross-domain decision support for industrial knowledge systems [59–69]. The second-largest cluster is *Engineering & Design Automation* (28/88, 31.8%), including electronic design automation and analog circuit layout [70–74], computational simulation and modelling [75–79], product and mechanical design [80–85], process automation and knowledge-assisted workflows [86–94], and maintenance decision support and risk assessment [95–97]. *Production & Manufacturing Control* accounts for 19/88 papers (21.6%), spanning shop-floor control and multi-agent manufacturing systems [98–104], robotic manipulation and human–robot collaboration [105–107], semiconductor manufacturing [108, 109], process planning [110–112], predictive maintenance [113], mining [114], agricultural data management [115], and materials discovery and industrial applications [116, 117]. Finally, *Logistics & Supply Chain* forms a smaller but distinct group (7/88, 8.0%), addressing fleet dispatching [20], warehouse management [118], heterogeneous multi-agent coordination for delivery and rescue [119–122], and procurement [123].

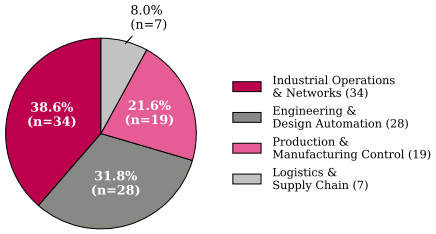

Fig. 3 Distribution of application domains across the corpus ($N = 88$).

4.2 System purposes and capabilities

Addressing RQ2, this section compares the purpose and capability profile of FM-based agents with the baseline survey [2].

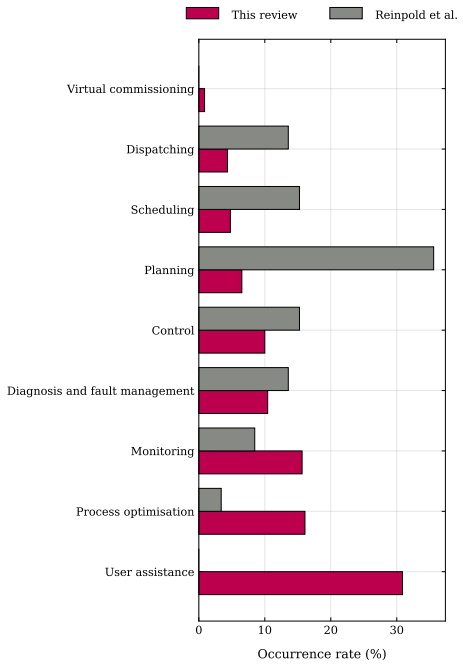

Fig. 4 Relative occurrence of system purposes in the coded corpus compared to the reference taxonomy from Reinpold et al. [2].

Figure 4 compares the purpose profile of FM-based agents with the baseline survey [2]. Because each publication may serve multiple operational goals, purposes are coded as multi-label indicators; the corpus yields 230 purpose assignments in total. User assistance accounts for the largest share (71 assignments, 30.9%), whereas it is absent in the baseline (+30.9 pp). This indicates a notable shift in the orientation of agent-based applications. Classical agent surveys have not reported user assistance as a system purpose. The integration of FMs into industrial agents appears to have enabled this category, thereby enlarging the set of applications that can be addressed. Process optimisation (37 assignments, 16.1%; baseline 3.4%) and monitoring (36 assignments, 15.7%; baseline 8.5%) follow as the second and third most frequent purposes. Here as well, a considerable increase in the utilisation of agent-based systems is observed. Note that process optimisation here refers to the improvement of resource design and performance, not to the application of, e.g., optimisation heuristics to optimize a whole production process [2]. The findings imply a convergence of the agent and DT paradigm as traditionally, these purposes have largely been fulfilled by DT applications [2]. It appears that FMs enable agents to process and leverage domain-specific information, thus facilitating their use for these knowledge-intensive applications.

Conversely, conventional production-control purposes are markedly less prominent than in the baseline. Planning drops from 35.6% to 6.5% (15 assignments), scheduling from 15.3% to 4.8% (11), and dispatching from 13.6% to 4.3% (10). Here, classical heuristics and optimisation-focused approaches have been more pronounced. A possible reason for this drop in the number of planning-oriented approaches with FM-based agents is the limited reliability of agents with FMs (Figure 7) due to hallucinations and instabilities. Control accounts for 23 assignments (10.0%; baseline 15.3%); while the absolute number of control-related papers is comparable, its relative share decreases because the overall set of purpose assignments is larger due to the emergence

of new purposes such as user assistance. Diagnosis and fault management (24 assignments, 10.4%; baseline 13.6%; −3.1 pp) and virtual commissioning (2, 0.9%; baseline 0.0%; +0.9 pp) are retained at near-baseline levels, indicating that knowledge-intensive and verification-oriented purposes are preserved in FM-based systems, unlike the sharper decline in planning, scheduling, and dispatching.

Overall, the purpose profile indicates a repositioning of agent functionality from embedded, optimisation-centred production control towards assistive, monitoring-oriented roles.

The capability profile in Figure 5 reveals both continuities and marked shifts relative to the baseline. The most frequently evidenced capabilities are encapsulation (88/88, 100%; set to *supported* by design, as the working definition requires encapsulated entities and the prompt-based interaction paradigm inherently hides internal states, see Section 5.1), decision-making (80/88, 90.9%), reactivity (79/88, 89.8%), context awareness (74/88, 84.1%), external system communication (73/88, 83.0%), and coordination (60/88, 68.2%). Within this set of capabilities, the findings indicate a similar yet more pronounced capability pattern compared to classical industrial agents, where these capabilities are likewise the most frequently reported. Condition monitoring (28/88, 31.8%), robustness and resilience (30/88, 34.1%), prioritisation (25/88, 28.4%), and mobility (1/88, 1.1%) also show no major differences to the baseline. The similar score in robustness and resilience, i.e. the capability to react to disturbances and disruptive events, can be attributed to similar vulnerabilities within both architectural paradigms: in neither case do the majority of reported systems demonstrate graceful degradation or continued operation under partial component failure. It should be noted that the results reflect the empirical evidence reported in the publications, not the potential capabilities of the systems.

Marked differences emerge in capabilities that are directly enabled by FMs. Human interaction (58/88, 65.9%; +37.1 pp) marks the largest positive deviation, which is expected as FMs, especially LLMs, provide a much more accessible human–machine interface through interacting with the user in natural language. Dealing with uncertainty (37/88, 42.0%; +35.3 pp) shows the second-largest increase, indicating that FMs enable agents to reason and act under incomplete or ambiguous information, a capability that classical rule-based and optimisation-driven agents rarely exhibit. Adaptivity (49/88, 55.7%; +23.5 pp), learning (37/88, 42.0%; +20.0 pp), and proactivity (26/88, 29.5%; +14.3 pp) are likewise substantially more prevalent, implying that FMs as a data-driven approach, provide the technological means to enable these complex capabilities. Visualisation (34/88, 38.6%; +23.4 pp) and simulation (29/88, 33.0%; +12.6 pp), capabilities commonly associated with DTs, can be found more often in FM-based approaches, where tool and model access is facilitated by standardised interfaces such as Model Context Protocol (MCP). Flexibility (57/88, 64.8%) and co-operation (48/88, 54.5%) remain at comparable levels to the baseline.

Interestingly, real-time capability (27/88, 30.7%; +15.4 pp) scores higher in FM-based approaches. Here, real-time capability refers to the ability to react under tight temporal restrictions, not

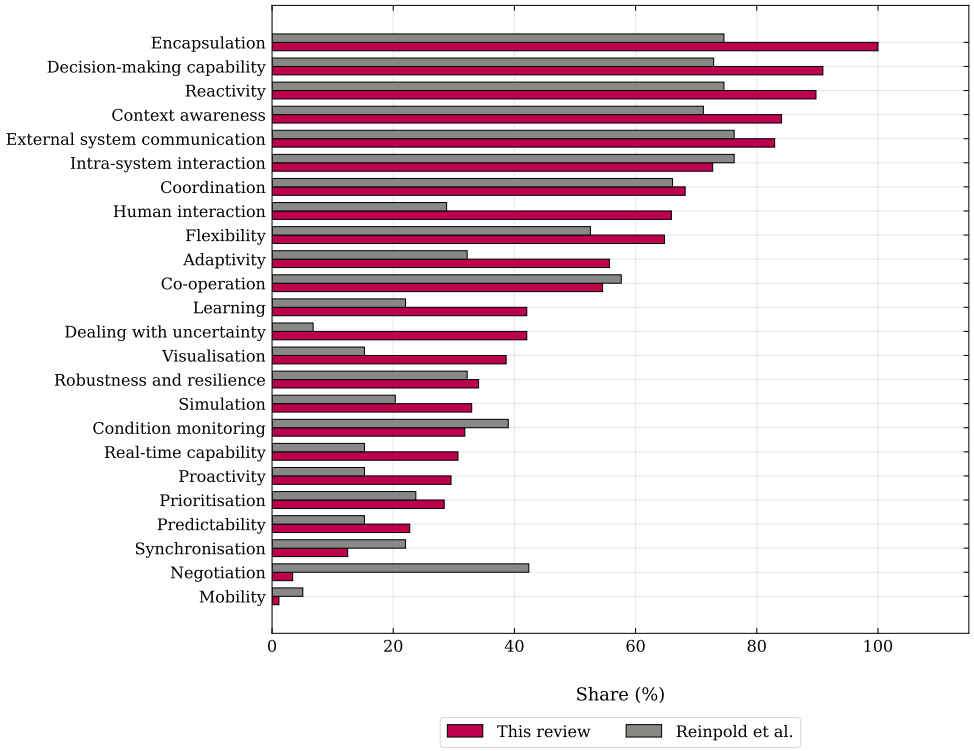

Fig. 5 Capability coverage of FM-based agent systems in this work compared with Reinpold et al. [2].

necessarily hard real-time constraints. The latter requires further investigation into the actual hard-real-time capabilities of these systems.

Notably, classical approaches score considerably higher in synchronisation (11/88, 12.5%; −9.5 pp) and especially negotiation (3/88, 3.4%; −39.0 pp). Synchronisation, i.e. the capability to react to changes in the system and update the agent's internal model, has been more prominent in classical approaches, likely because many optimisation algorithms require a synchronised model of the underlying system. The major difference in negotiation corresponds to the shift in application focus identified in the purpose analysis, showcasing the transition from negotiation-based optimisation heuristics to a more user-centric, assistive paradigm. The typically small number of agents observed in the corpus may additionally reduce the opportunities for negotiation by construction.

The composite score comparison in Figure 6 aggregates these capability-level observations into higher-level property dimensions, following the weighting scheme defined in Reinpold et al. [2]. Autonomy (median 0.50 vs. 0.42, +0.08), intelligence (median 0.48 vs. 0.40, +0.08), and fidelity (median 0.28 vs. 0.17, +0.11) are consistently higher in this work. Sociability, by contrast, shows a lower median (0.55 vs. 0.64, −0.09), which is

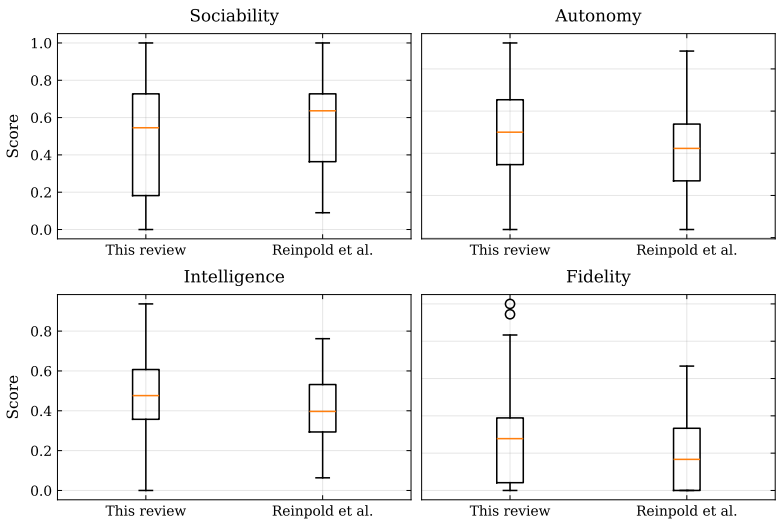

Fig. 6 Composite scores for sociability, autonomy, intelligence, and fidelity comparing this work with Reinpold et al. [2].

consistent with the reduced negotiation capability identified above.

4.3 Limitations and future work

Addressing RQ3, this section consolidates the most frequently reported limitations and future work directions across the corpus.

Figure 7 summarises the most frequently reported limitation themes across 308 total limitation mentions. Limited generalisation and real-world validity is the dominant concern (44 mentions, 14.3%). This theme subsumes simulation-only evaluations, proof-of-concept scope, and restricted domain coverage: many systems are validated in controlled laboratory settings, on narrow benchmarks, or in single-site case studies, leaving open how well they transfer to different operational conditions or industrial environments. This finding is consistent with the TRL distribution reported in

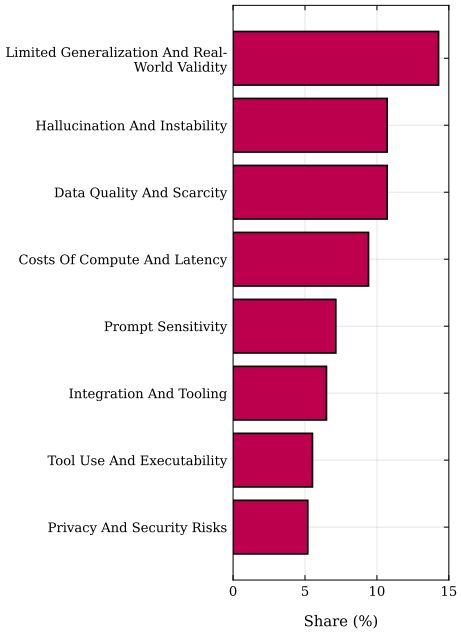

Fig. 7 Top limitation themes for FM-based agent systems in industrial and industrially relevant settings, measured by the share of papers in which each theme is reported.

Section 4.1, where the majority of systems remain at development stage. Hallucination and instability (33, 10.7%) ranks second and encompasses not only factual hallucinations but also non-deterministic behaviour across repeated runs, numeric and logical errors, and the generation of non-executable outputs. Data quality and scarcity (33, 10.7%) captures domain-specific data deficits, reliance on synthetic or biased datasets, and challenges in obtaining accurate system parameters. Costs of compute and latency (29, 9.4%) completes the top cluster and reflects inference latency, high computational cost, and limited scalability under concurrent workloads.

The remaining themes point to integration-layer and governance challenges. Prompt sensitivity (22, 7.1%) highlights the brittleness of natural-language-mediated control: model behaviour is highly sensitive to prompt phrasing and ordering, decoding parameters, and formatting conventions, often requiring handcrafted or few-shot prompts and additional guardrails to produce valid and stable outputs. Integration and tooling (20, 6.5%) addresses the complexity of retrofitting FMs into industrial toolchains, including vendor dependencies, format conversions between simulation, CAD, or DT data and natural-language representations, and inconsistent interfaces across automation layers. Tool use and executability (17, 5.5%) is a related but distinct concern: FM-based agents frequently misuse tools or APIs, generate syntactically invalid code or structured outputs (e.g., JSON, SQL, domain-specific languages), and produce non-executable plans that require verification and retries. Privacy and security risks (16, 5.2%) round out the profile and encompass data leakage through model queries, dependence on external API providers, and vulnerability to prompt injection and adversarial manipulation of FM-based decision loops.

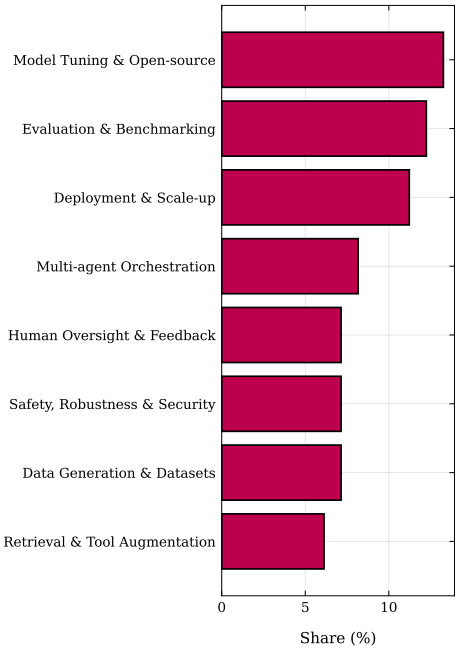

Fig. 8 Top future work themes for FM-based agent systems, measured by the relative share of papers mentioning each theme.

The future-work distribution in Figure 8 is based on 98 total mentions and mirrors the limitation profile. Model tuning and open-source models (13, 13.3%) leads the ranking; this theme covers domain-specific adaptation through fine-tuning, distillation, or open-weight alternatives, as well as calls for locally runnable models and public artifact releases. Evaluation and benchmarking (12, 12.2%) addresses the external-validity gap through shared benchmarks, multi-criterion validation

frameworks (combining expert evaluation, consistency analysis, and functional accuracy), user studies, and broader evaluation scopes. Deployment and scale-up (11, 11.2%) calls for transitioning proofs-of-concept to real operational environments, broadening coverage across geographic regions and application domains, and enabling enterprise-level integration. Multi-agent orchestration (8, 8.2%) indicates ongoing interest in advancing coordination strategies, domain-specialised agent roles, and adaptive orchestration mechanisms.

The remaining future-work themes address trustworthiness and infrastructure. Human oversight and feedback (7, 7.1%) proposes structured approaches to balance agent autonomy with operational safety, including dynamic confidence thresholds, human review gates, and explicit feedback loops for in- and out-of-the-loop designs. Safety, robustness and security (7, 7.1%) encompasses verification and observability mechanisms, robustness to adversarial prompt perturbations, privacy protections, and methods for interpretable and trustworthy decision-making. Data generation and datasets (7, 7.1%) calls for physics-based or synthetic data generation, broader and more diverse benchmark datasets, and automated data ingestion pipelines to overcome the data scarcity identified above. Retrieval and tool augmentation (6, 6.1%) advocates strengthening agent reasoning and correctness through domain-specific RAG, integration of external computation tools and data sources, and structured context provision. As with limitations, mention frequency should not be equated with practical severity, since papers differ substantially in reporting depth and focus.

5 Discussion

Across the corpus, the evidence base for FM-based technical agents in industrial and industrially relevant settings is characterised by prototype-oriented contributions and early-stage validation, while reports of operational deployments remain comparatively rare (RQ1). In terms of operational goals, most systems focus on user-assistance, monitoring, and resource-related process optimisation rather than in conventional production-control roles such as dispatching, scheduling, and planning (RQ2). Consistent with this purpose profile, the reported capability coverage emphasizes context handling, decision support, interaction, and communication with external tools and systems, whereas conventional multi-agent mechanisms such as negotiation are less frequently evidenced. Finally, the most frequently reported limitations and future-work directions point to practical deployment barriers and trustworthiness concerns: external validity and evaluation scope, reliability (including instability and hallucination), data constraints, latency and cost, and challenges at the interface between FMs and industrial automation infrastructures (RQ3).

5.1 Adequacy of the working definition

The working definition introduced in Section 3.1 requires (i) encapsulation and specified objectives in an industrial context, (ii) autonomous action beyond pure text generation, and (iii) an FM as a central component for decision-making and action selection. The corpus-level results allow a retrospective assessment of whether this three-part criterion adequately captures the systems reported in the literature.

The capability profile (Figure 5) supports the definition's emphasis on encapsulation, decision-making, and autonomous action: encapsulation, decision-making capability, and reactivity are evidenced in over 90% of the included systems, and external system communication, a proxy for actionable tool use beyond text output, exceeds 80%. Context awareness is similarly prevalent, indicating that the included systems do not merely generate text but actively interpret situational information to select actions. These observations support the view that criterion (ii) effectively separates FM-based MAS from purely generative LLM applications.

The encapsulation and objective criterion (i) is harder to validate empirically, because no publication in the corpus explicitly discusses encapsulation as a design property. However, encapsulation arguably arises *by design* in FM-based agents: external interaction is mediated exclusively through prompts, and the internal states, reasoning traces, and strategy of the model are not directly accessible to other system components at runtime. Combined with the prevalence of architectures in which individual agents are assigned distinct roles and objectives (Section 4.1), this suggests that encapsulation is a de facto design principle of FM-based agents, not because authors deliberately engineer it, but because the prompt-based interaction paradigm inherently hides internal states from the external world.

A potential limitation of the definition concerns its scope: the strong focus on user assistance and monitoring in the corpus (Section 4.2) raises the question of whether some assistive systems, for instance conversational knowledge-retrieval interfaces, satisfy the autonomy criterion or are better characterised as interactive tools. The conservative screening applied in this work mitigates this risk by requiring autonomous behaviour, but borderline cases remain. Future work could refine the autonomy threshold, for example by distinguishing degrees of autonomy along established taxonomies.

5.2 Implications for research and practice

For industrial deployment, the observed maturity distribution and the limitation profile suggest that integration, latency, and safety assurance are currently key bottlenecks. In particular, the recurrent focus on tool use, executability, and interaction with existing automation layers indicates that many systems are not limited by the underlying FM alone, but by the end-to-end engineering of reliable action execution and monitoring in production environments.

For evaluation, the predominance of early-stage evidence and the strong future-work emphasis on benchmarking and validation highlight the need for more consistent reporting and more deployment-near evaluation setups. In addition to task-level performance, evaluation should explicitly cover robustness, failure modes, and operational constraints (e.g., latency, compute costs, and data availability) that are decisive in industrial contexts.

For agent architecture design, the capability profile points to language-mediated interaction and knowledge-intensive reasoning as prominent value propositions, while negotiation- and synchronisation-heavy MASs patterns appear less central in the current corpus. This suggests that practical designs often rely on tool-augmented, workflow-integrated architectures with explicit grounding, observability, and

human oversight rather than on fully decentralised negotiation protocols.

5.3 Threats to validity

This work is subject to selection effects and coverage bias induced by the chosen data sources, query formulation, and the focus on publications from 2020 onwards. In addition, the screening procedure is intentionally strict with respect to excluding purely conceptual work; consequently, the resulting corpus reflects research that reports implementations or case studies and may under-represent design-only contributions which, even though they lack empirical proof, still might provide valuable insights into existing capabilities.

Two selection effects at opposite ends of the TRL scale may influence the maturity distribution reported in Section 4.1. At the lower end, contributions at TRL 1–3 are likely underrepresented because the inclusion criteria (Section 3.3) require a concrete application, implementation, or case study; purely conceptual work, position papers, and early theoretical proposals are not considered. Early-stage research on FM-based agent concepts that has not yet been accompanied by an implementation is therefore not reflected in the corpus, even if such work exists in the literature. At the upper end, systems operating at higher readiness levels in industrial practice may not always be reported in scientific publications, for instance due to intellectual property concerns, confidentiality agreements, or weaker publishing incentives for industrial practitioners once a system has left the prototype stage. The TRL distribution can therefore be understood as reflecting the maturity landscape as visible in the scientific literature under the applied inclusion criteria, which should be kept in mind when interpreting the observed shares at the extremes of the scale.

Coding ambiguity is a second threat to validity. The analysis is generally performed at paper level and relies on explicit evidence in publications. One exception is encapsulation, which is set to *supported* for all included systems by design rather than by textual evidence (see Section 5.1). If a paper omits architectural detail, a capability or purpose may be present in the implemented system but not evidenced for coding. Conversely, some categories may be described as intended functionality without strong empirical validation. These issues are particularly relevant for maturity (TRL) assessments, which are often inferred from evaluation descriptions rather than stated explicitly. A related concern refers to the coding of real-time capability: several publications declare their system or individual actions as "real-time" without explicitly specifying the temporal resolution or response-time guarantees. In some cases, real-time capability is attributable to specific sub-components (e.g., sensor-based monitoring modules) rather than to the FM-based agent itself. These claims were accepted at face value during coding, which may lead to an overestimation of real-time capability in the corpus. This coding ambiguity reflects a broader validity challenge in empirical synthesis: conclusions depend on the completeness, operational clarity, and triangulation of the reported evidence, while omitted detail may weaken the basis for stronger inferences [124].

The descriptive comparison against the baseline survey [2] is limited by differences in corpus composition, inclusion criteria, and reporting depth. The

baseline covers industrial software agents and DTs broadly, whereas this work focuses specifically on FM-based technical agents. In addition, differences in reporting depth and evidence operationalisation can affect whether a capability or purpose is counted as supported. Although the same purpose–capability framing is used, observed differences cannot solely be attributed to the FM integration but might also stem from corpus-dependent coverage deviations.

Finally, the quantitative synthesis is largely based on mention frequencies and paper-level counts. Mention frequency does not necessarily correspond to practical severity or importance: papers differ in reporting depth and focus, and, for example, a limitation theme can be decisive for deployment even if it is mentioned infrequently.

5.4 Outlook

Based on the consolidated limitation and future-work themes, immediate next steps should include (i) more deployment-near evaluations and longitudinal studies that move beyond prototype evidence, (ii) shared evaluation protocols and benchmarking frameworks, and reporting templates for FM-based agent systems in industrial contexts, (iii) robust tool and digital-infrastructure integration patterns (including monitoring, debugging, and auditability), and (iv) safety- and security-oriented engineering measures (e.g., fallback strategies, constraint enforcement, and governance processes). In addition, domain-relevant datasets remain important prerequisites for comparing approaches and for improving external validity across application areas. Looking at the distribution of application domains, logistics and supply chain appears to have been studied only rarely, which is another avenue of potential future work.

Beyond these consolidation-oriented directions, the results show a particularly interesting dynamic around diagnosis and fault management. As noted in Section 4.2, this purpose is retained at near-baseline levels (24 assignments, 10.4%; baseline 13.6%; −3.1 pp), unlike the sharper decline observed for planning, scheduling, and dispatching. This persistence is consistent with the capability profile of FM-based agents: they can process and interpret heterogeneous failure symptoms, parse unstructured maintenance logs and operator reports, and leverage patterns learned during training about system behaviour and failure modes to support root-cause identification and the determination of adequate corrective actions. The combination of context awareness, dealing with uncertainty, and learning, which show the largest positive deviations relative to the baseline, aligns conceptually with diagnostic tasks. At the same time, the limitations identified in Section 4.3, hallucination, output instability, and inference latency, remain particularly consequential for this purpose, since diagnostic tasks in industrial settings can be safety-relevant and time-critical. Translating the observed potential into reliable deployments therefore represents a promising direction for future work.

A further research opportunity concerns the systematic analysis of task decomposition in hybrid agent architectures. The capability profile shows that FM-based agents excel in language-mediated interaction, context interpretation, and reasoning under uncertainty, yet conventional algorithmic methods remain superior for well-defined computational tasks such as scheduling and path

planning. A principled investigation into which sub-tasks genuinely benefit from FM reasoning and which are better delegated to established domain algorithms, e.g. based on negotiations and established protocols, would help practitioners design more efficient architectures that combine the strengths of both paradigms. Such an analysis could build on the purpose and capability dimensions introduced in this work and extend them with a task-level granularity that distinguishes generative, interpretive, and computational sub-tasks within a single agent workflow. Moreover, integrating classical and FM-based agents with the DT paradigm provides another possibility for investigation. DTs, that for example exhibit capabilities like simulation and prediction, can further enhance the spectrum of possible applications.

6 Conclusion

This work provides a systematic literature review of FM-based technical agents in industrial and industrially relevant contexts and synthesizes a corpus of $N = 88$ publications using a transparent, evidence-linked coding scheme. The results indicate that the reported systems are predominantly evaluated at prototype and early validation stages, with comparatively few publications providing deployment-oriented evidence (RQ1). Across the corpus, operational goals are most frequently positioned in assistive, monitoring, and resource-related process optimisation-centric roles, and the reported capability profile emphasises interaction, context handling, decision support, and tool-mediated communication, while conventional negotiation-heavy MASs mechanisms are less frequently evidenced (RQ2). The most widely reported limitations and future work directions highlight external-validity gaps, reliability concerns, data constraints, and practical deployment barriers related to latency, cost, and system integration (RQ3). Overall, the proposed coding representation and comparative framing can serve as a reusable basis for future surveys and for more standardised evaluation and reporting of industrial FM-based agent systems.

Acknowledgements. This work was carried out within the VDI/VDE-GMA Technical Committee 3.35 on Industrial Agents. The authors thank the committee members for the valuable discussions and contributions that shaped this survey.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** The authors declare no competing interests.
- **Ethics approval:** Not applicable.
- **Data availability:** Not applicable.
- **Author contributions:** The first three authors are listed in order of contribution; the remaining authors are listed alphabetically by surname. V.H.: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. F.G.: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. D.K.: Data curation, Formal analysis, Writing – review & editing. A.A., B.H., C.L., F.M., F.O., L.C., L.V., M.O., M.T., N.A.T., P.K., T.S., Y.X.: Formal analysis, Writing – review & editing.

References

[1] Wooldridge, M.: An Introduction to Multi Agent Systems. John Wiley & Sons, Chichester (2002)

[2] Reinpold, L.M., Wagner, L.P., Gehlhoff, F., Ramonat, M., Kilthau, M., Gill, M.S., Reif, J.T., Henkel, V., Scholz, L., Fay, A.: Systematic comparison of software agents and digital twins: differences, similarities, and synergies in industrial production. Journal of Intelligent Manufacturing **36**, 765–800 (2025) https://doi.org/10.1007/s10845-023-02278-y . Published online: 2024-01-04

[3] Smith, R.G.: The contract net protocol: High-level communication and control in a distributed problem solver. IEEE Transactions on Computers **C-29**(12), 1104–1113 (1980) https://doi.org/10.1109/TC.1980.1675516

[4] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Qin, W., Zheng, Y., Qiu, X., Huang, X., Zhang, Q., Gui, T.: The rise and potential of large language model based agents: a survey. Science China Information Sciences **68**, 121101 (2025) https://doi.org/10.1007/s11432-024-4222-0

[5] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**, 186345 (2024) https://doi.org/10.1007/s11704-024-40231-1

[6] Ren, Y., Liu, Y., Ji, T., Xu, X.: Ai agents and agentic ai–navigating a plethora of concepts for future manufacturing. Journal of Manufacturing Systems **83**, 126–133 (2025) https://doi.org/10.1016/j.jmsy.2025.08.017

[7] Ali, M.A., Dornaika, F., Charafeddine, J.: Agentic ai: a comprehensive survey of architectures, applications, and future directions. Artificial Intelligence Review **59**(11) (2026) https://doi.org/10.1007/s10462-025-11422-4

[8] Xia, Y.: Integrating Large Language Model Agents with Digital Twins for Industrial Autonomous Systems. IAS-Forschungsberichte, vol. 2026,2. Dissertation of University of Stuttgart, Shaker Verlag, (2026)

[9] Deng, H., Namoano, B., Zheng, B., Khan, S., Erkoyuncu, J.A.: From prediction to prescription: Large language model agent for context-aware maintenance decision support. PHM Society European Conference **8**(1), 10 (2024) https://doi.org/10.36001/phme.2024.v8i1.4114

[10] Gamage, G., Mills, N., Silva, D.D., Manic, M., Moraliyage, H., Jennings, A., Alahakoon, D.: Multi-agent rag chatbot architecture for decision support in net-zero emission energy systems. In: 2024 IEEE

International Conference on Industrial Technology (ICIT), pp. 1–6. IEEE, (2024). https://doi.org/10.1109/icit58233.2024.10540920

[11] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Survey Track, pp. 8048–8057 (2024). https://doi.org/10.48550/arXiv.2402.01680

[12] Pérez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F.J., Englund, C., Tauber, M., Nikolakopoulos, G., Flores, J.L.: Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. ACM Computing Surveys **56**(7), 176–117640 (2024) https://doi.org/10.1145/3626314

[13] VDI/VDE 2653-1 - Multi-agent systems in industrial automation: Fundamentals. DIN Media GmbH, Berlin (2018). https://www.dinmedia.de/de/technische-regel/vdi-vde-2653-blatt-1/282864028

[14] Jin, H., Huang, L., Cai, H., Yan, J., Li, B., Chen, H.: From LLMs to LLM-based agents for software engineering: A survey of current, challenges and future. arXiv preprint (2024) arXiv:2408.02479. Preprint

[15] Zhou, J., Lu, Q., Chen, J., Zhu, L., Xu, X., Xing, Z., Harrer, S.: A taxonomy of architecture options for foundation model-based agents: Analysis and decision model. In: arXiv Preprint (2024). https://doi.org/10.48550/arXiv.2408.02920 . Preprint

[16] Plaat, A., Duijn, M., Stein, N., Preuss, M., Putten, P., Batenburg, K.J.: Agentic large language models, a survey. Journal of Artificial Intelligence Research **84**, 29 (2025) https://doi.org/10.1613/jair.1.18675

[17] Zhang, C., Xu, Q., Yu, Y., Zhou, G., Zeng, K., Chang, F., Ding, K.: A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing. Robotics and Computer-Integrated Manufacturing **92**, 102883 (2025) https://doi.org/10.1016/j.rcim.2024.102883

[18] Mohammadi, M., Li, Y., Lo, J., Yip, W.: Evaluation and benchmarking of LLM agents: A survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, (2025). https://doi.org/10.1145/3711896.3736570

[19] Lim, J., Vogel-Heuser, B., Kovalenko, I.: Large language model-enabled multi-agent manufacturing systems. In: 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), pp. 3940–3946. IEEE, (2024). https://doi.org/10.1109/case59546.2024.10711432

[20] Kalantari, S., Wang, Y., Sun, S., Wang, X.: Fleetwiz: An intelligent platform for spatio-temporal

multi-resource truckload fleet dispatching. In: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, pp. 665–668. ACM, (2024). https://doi.org/10.1145/3678717.3691272

[21] Greis, N.P., Cherukuri, H.P., Outeiro, J.C.M.: Multi-agent systems for manufacturing digital twins: A perspective on agency and large language models. In: IFAC-PapersOnLine, vol. 59, pp. 1612–1617 (2025). https://doi.org/10.1016/j.ifacol.2025.09.271

[22] Zhao, Z., Tang, D., Liu, C., Wang, L., Zhang, Z., Zhu, H., Chen, K., Nie, Q., Ji, Y.: A large language model-based multi-agent manufacturing system for intelligent shopfloors. Advanced Engineering Informatics **65**, 103888 (2026) https://doi.org/10.1016/j.aei.2025.103888

[23] Xie, J., Liu, C.-C.: Multi-agent systems and their applications **7**(1), 188–197 (2017) https://doi.org/10.1080/22348972.2017.1348890

[24] Huckert, J.L., Sidorenko, A., Wagner, A.: Analysis and assessment of multi-agent systems for production planning and control. In: Silva, F.J.G., Pereira, A.B., Campilho, R.D.S.G. (eds.) Flexible Automation and Intelligent Manufacturing: Establishing Bridges for More Sustainable Manufacturing Systems. Lecture Notes in Mechanical Engineering, pp. 687–698. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-38241-3_77

[25] Gehlhoff, F.: Agent-based decentralised architecture for integrated process planning and scheduling of transport and production processes. PhD thesis, Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg (2023). https://doi.org/10.24405/15181

[26] Massouh, B., Danielsson, F., Lennartson, B., Ramasamy, S., Khabbazi, M.: Safe and reconfigurable manufacturing: safety aware multi-agent control for plug & produce system. The International Journal of Advanced Manufacturing Technology **134**, 529–544 (2024) https://doi.org/10.1007/s00170-024-14112-7

[27] Cruz Salazar, L.A., Vogel-Heuser, B.: A CPPS-architecture and workflow for bringing agent-based technologies as a form of artificial intelligence into practice. at – Automatisierungstechnik **70**(6), 580–598 (2022) https://doi.org/10.1515/auto-2022-0008

[28] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Pearson, Hoboken, NJ (2020)

[29] Bommasani, R., *et al.*: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021) https://doi.org/10.48550/arXiv.2108.07258

[30] Chen, C., Zhao, K., Leng, J., Liu, C., Fan, J., Zheng, P.: Integrating large language model and digital twins in the context of industry 5.0: Framework, challenges

and opportunities. Robotics and Computer-Integrated Manufacturing **94**, 102982 (2025) https://doi.org/10.1016/j.rcim.2025.102982

[31] Jee, T.K.: Llm-based overlay issue classification and solution optimization in semiconductor manufacturing. In: DTCO and Computational Patterning IV, p. 48. SPIE, (2025). https://doi.org/10.1117/12.3050976

[32] Müller, M., Müller, T., Talkhestani, B.A., Marks, P., Jazdi, N., Weyrich, M.: Industrial autonomous systems: a survey on definitions, characteristics and abilities. at - Automatisierungstechnik **69**(1), 3–13 (2021) https://doi.org/10.1515/auto-2020-0131

[33] Kaber, D.B.: A conceptual framework of autonomous and automated agents. Theoretical Issues in Ergonomics Science **19**(4), 406–430 (2018) https://doi.org/10.1080/1463922X.2017.1363314 https://doi.org/10.1080/1463922X.2017.1363314

[34] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., *et al.*: The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ **372**, 71 (2021) https://doi.org/10.1136/bmj.n71

[35] Mankins, J.C.: Technology readiness levels: A white paper. Technical report, NASA, Office of Space Access and Technology (April 1995). Advanced Concepts Office

[36] Bundesministerium für Wirtschaft und Energie (BMWi): Fortschreibung der anwendungsszenarien der plattform Industrie 4.0. Technical report, Plattform Industrie 4.0 (2018). AG Forschung und Innovation. https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/hm-2018-fb-landkarte.pdf

[37] Zhang, W., Yao, T., Zhou, F., Jin, H., Liu, J., Wan, Z., Liu, C., Wang, Y., Chai, B., Chen, X.: A conversational agent based on large language models for fault recovery planning generation. In: 2025 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE, (2025). https://doi.org/10.1109/iscas56072.2025.11043523

[38] Yang, X., Lin, C., Liu, H., Wu, W.: Rl2: Reinforce large language model to assist safe reinforcement learning for energy management of active distribution networks. IEEE Transactions on Smart Grid **16**(4), 3419–3431 (2025) https://doi.org/10.1109/tsg.2025.3568226

[39] Arnautov, K.V., Akimov, D.A.: Application of large language models for optimization of electric power system states. In: 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), pp. 314–317. IEEE, (2024). https://doi.org/10.1109/elcon61730.2024.10468377

[40] Yan, L., Cheng, C.: Voltage control for distribution networks based on large language model-assisted deep reinforcement learning. IEEE Access **13**, 76072–76084 (2025)

https://doi.org/10.1109/access.2025.3565280

[41] Gao, Q., Shen, L., Shi, J., Gu, X., Gu, S., Ge, Y., Xie, Y., Zhu, X., Zang, B., Zhang, M., Nazir, M.S., Ji, J.: Transformer-enhanced intelligent microgrid self-healing: Integrating large language models and adaptive optimization for real-time fault detection and recovery. Energy Engineering **122**(7), 2767–2800 (2025) https://doi.org/10.32604/ee.2025.065600

[42] Badmus, E.O., Sang, P., Stamoulis, D., Pandey, A.: PowerChain: A Verifiable Agentic AI System for Automating Distribution Grid Analyses. Preprint (2025). https://doi.org/10.48550/arXiv.2508.17094

[43] Ou, P., Wang, Y., Lin, W., Wu, J.: An llm-based modeling and decision optimization for user-centric electric vehicle charging. In: 2024 IEEE 8th Conference on Energy Internet and Energy System Integration (EI2), pp. 4078–4083. IEEE, (2024). https://doi.org/10.1109/ei264398.2024.10991378

[44] Mongaillard, T., Lasaulce, S., Hicheur, O., Zhang, C., Bariah, L., Varma, V.S., Zou, H., Zhao, Q., Debbah, M.: Large Language Models for Power Scheduling: A User-Centric Approach. Preprint (2024). https://doi.org/10.48550/arxiv.2407.00476

[45] Matharaarachchi, A., Mendis, W., Randunu, K., Silva, D.D., Gamage, G., Moraliyage, H., Mills, N., Jennings, A.: Optimizing generative ai chatbots for net-zero emissions energy internet-of-things infrastructure. Energies **17**(8), 1935 (2024) https://doi.org/10.3390/en17081935

[46] Gamage, G., Mills, N., Rathnayaka, P., Jennings, A., Alahakoon, D.: Cooee: An artificial intelligence chatbot for complex energy environments. In: 2022 15th International Conference on Human System Interaction (HSI), pp. 1–5. IEEE, (2022). https://doi.org/10.1109/hsi55341.2022.9869464

[47] Wang, Z., Liu, Z., Zhang, Y., Zhong, A., Wang, J., Yin, F., Fan, L., Wu, L., Wen, Q.: Rcagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 4966–4974. ACM, (2024). https://doi.org/10.1145/3627673.3680016

[48] Patel, D., Lin, S., Rayfield, J., Zhou, N., Vaculin, R., Martinez, N., O'donncha, F., Kalagnanam, J.: AssetOpsBench: Benchmarking AI Agents for Task Automation in Industrial Asset Operations and Maintenance. Preprint (2025). https://doi.org/10.48550/arXiv.2506.03828

[49] Shi, B., Luo, Y., Wang, J., Zhao, Y., Zhang, S., Hao, B., Zhao, C., Sun, Y., Zhang, Z., Sun, R., Li, H., Song, W., Chen, X., Miao, J., Pei, D.: Flowxpert: Expertizing troubleshooting workflow orchestration

with knowledge base and multi-agent coevolution. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, pp. 4839–4850. ACM, (2025). https://doi.org/10.1145/3711896.3737221

[50] Jeong, C., Sim, S., Cho, H., Kim, S., Shin, B.: E2E Process Automation Leveraging Generative AI and IDP-Based Automation Agent: A Case Study on Corporate Expense Processing. Preprint (2025). https://doi.org/10.48550/arXiv.2505.20733

[51] Paulose, R., Neelanath, V., George, M.: Domain agnostic agentic ai: Enabling autonomous automation with smartgenie copilot. In: 2025 Emerging Technologies for Intelligent Systems (ETIS), pp. 1–6. IEEE, (2025). https://doi.org/10.1109/etis64005.2025.10961403

[52] Sezgin, A.: Scenario-driven evaluation of autonomous agents: Integrating large language model for uav mission reliability. Drones **9**(3), 213 (2025) https://doi.org/10.3390/drones9030213

[53] Yu, J., Wang, Y., Ma, W.: Large language model-enhanced reinforcement learning for generic bus holding control strategies. Transportation Research Part E: Logistics and Transportation Review **200**, 104142 (2025) https://doi.org/10.1016/j.tre.2025.104142

[54] Xu, Z., Chen, T., Huang, Z., Xing, Y., Chen, S.: Personalizing driver agent using large language models for driving safety and smarter human–machine interactions. IEEE Intelligent Transportation Systems Magazine **17**(4), 96–111 (2025) https://doi.org/10.1109/mits.2025.3551736

[55] Yang, H., Siew, M., Joe-Wong, C.: An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems. Preprint (2024). https://doi.org/10.48550/arxiv.2403.16809

[56] Sawada, T., Mizuno, M., Hasegawa, T., Yokoyama, K., Kono, M.: Office-in-the-loop: an investigation into agentic ai for advanced building hvac control systems. Data-Centric Engineering **6**, 31 (2025) https://doi.org/10.1017/dce.2025.10010

[57] Wang, Y., Afzal, M.M., Li, Z., Zhou, J., Feng, C., Guo, S., Quek, T.Q.S.: Large language model as a catalyst: A paradigm shift in base station siting optimization. IEEE Transactions on Cognitive Communications and Networking **11**(6), 4313–4327 (2025) https://doi.org/10.1109/tccn.2025.3548615

[58] Wang, D., Wang, Y., Jiang, X., Zhang, Y., Pang, Y., Zhang, M.: When large language models meet optical networks: Paving the way for automation. Electronics **13**(13), 2529 (2024) https://doi.org/10.3390/electronics13132529

[59] Wanna, S., Parra, F., Valner, R., Kruusamäe, K., Pryor, M.:

Unlocking underrepresented use-cases for large language model-driven human-robot task planning. Advanced Robotics **38**(18), 1335–1348 (2024) https://doi.org/10.1080/01691864.2024.2366974

[60] Gamage, G., Mills, N., Silva, D.D., Manic, M., Moraliyage, H., Jennings, A., Alahakoon, D.: Multi-agent rag chatbot architecture for decision support in net-zero emission energy systems. In: 2024 IEEE International Conference on Industrial Technology (ICIT), pp. 1–6. IEEE, (2024). https://doi.org/10.1109/icit58233.2024.10540920

[61] Chen, R., He, C.: Fostering collective intelligence in cpss: an llm-driven multi-agent cooperative tuning framework. Frontiers in Physics **13**, 1613499 (2025) https://doi.org/10.3389/fphy.2025.1613499

[62] Wang, S., Liang, C., Gao, Y., Liu, Y., Li, J., Wang, H.: Decoding urban industrial complexity: Enhancing knowledge-driven insights via industryscopegpt. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 4757–4765. ACM, (2024). https://doi.org/10.1145/3664647.3681705

[63] Su, J., Cardie, C., Nakov, P.: Adapting fake news detection to the era of large language models. In: Findings of the Association for Computational Linguistics: NAACL 2024, pp. 1473–1490. Association for Computational Linguistics, (2024). https://doi.org/10.18653/v1/2024.findings-naacl.95

[64] Wu, Y., Wang, H., Zhang, Y., Li, X., Wu, H., Fan, M., Liu, T.: Business compliance detection of smart contracts in electricity and carbon trading scenarios. In: 2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 177–178. IEEE, (2024). https://doi.org/10.1109/issrew63542.2024.00074

[65] Zhou, R., Yang, Y., Wen, M., Wen, Y., Wang, W., Xi, C., Xu, G., Yu, Y., Zhang, W.: Trad: Enhancing llm agents with step-wise thought retrieval and aligned decision. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–13. ACM, (2024). https://doi.org/10.1145/3626772.3657788

[66] Mai, K., Ghate, N., Lee, J., Beuran, R.: Llm-based fine-grained abac policy generation. In: Proceedings of the 11th International Conference on Information Systems Security and Privacy, pp. 204–212. SCITEPRESS - Science and Technology Publications, (2025). https://doi.org/10.5220/0013225500003899

[67] Kalian, A.D., Lee, J., Johannesson, S.P., Otte, L., Hogstrand, C., Guo, M.: Fine-Tuning and Prompt Engineering of LLMs, for the Creation of Multi-Agent AI for Addressing Sustainable Protein Production Challenges. Preprint (2025). https://doi.org/10.48550/arXiv.2506.20598

[68] Chen, X., Zhang, L.: Revolutionizing Bridge Operation and Maintenance with LLM-based Agents: An Overview of Applications and Insights. Preprint (2024). https://doi.org/10.48550/arxiv.2407.10064

[69] Kar, I., Ralte, Z., Shivakumara, M., Roy, R., Kumari, A.: Agents are all you need: Elevating trading dynamics with advanced generative ai-driven conversational llm agents and tools. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), pp. 1–6. IEEE, (2024). https://doi.org/10.1109/i2ct61223.2024.10543356

[70] He, Z., Wu, H., Zhang, X., Yao, X., Zheng, S., Zheng, H., Yu, B.: Chateda: A large language model powered autonomous agent for eda. In: 2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD), pp. 1–6. IEEE, (2023). https://doi.org/10.1109/mlcad58807.2023.10299852

[71] Shen, J., Chen, Z., Zhuang, J., Huang, J., Yang, F., Shang, L., Bi, Z., Yan, C., Zhou, D., Zeng, X.: Atelier: An automated analog circuit design framework via multiple large language model-based agents. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **45**(1), 31–44 (2026) https://doi.org/10.1109/tcad.2025.3573228

[72] Chang, C.-C., Ho, C.-T., Li, Y., Chen, Y., Ren, H.: Drc-coder: Automated drc checker code generation using llm autonomous agent. In: Proceedings of the 2025 International Symposium on Physical Design, pp. 143–151. ACM, (2025). https://doi.org/10.1145/3698364.3705347

[73] Liu, B., Zhang, H., Gao, X., Kong, Z., Tang, X., Lin, Y., Wang, R., Huang, R.: Layoutcopilot: An llm-powered multiagent collaborative framework for interactive analog layout design. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **44**(8), 3126–3139 (2025) https://doi.org/10.1109/tcad.2025.3529805

[74] Wu, H., Zheng, H., He, Z., Yu, B.: Divergent Thoughts toward One Goal: LLM-based Multi-Agent Collaboration System for Electronic Design Automation. Preprint (2025). https://doi.org/10.48550/arxiv.2502.10857

[75] Lykov, A., Dronova, M., Naglov, N., Litvinov, M., Satsevich, S., Bazhenov, A., Berman, V., Shcherbak, A., Tsetserukou, D.: LLM-MARS: Large Language Model for Behavior Tree Generation and NLP-enhanced Dialogue in Multi-Agent Robot Systems. Preprint (2023). https://doi.org/10.48550/arxiv.2312.09348

[76] Xia, Y., Dittler, D., Jazdi, N., Chen, H., Weyrich, M.: LLM experiments with simulation: Large Language Model Multi-Agent System for Simulation Model Parametrization in Digital Twins. Preprint (2024). https://doi.org/10.48550/arxiv.2405.18092

[77] Kim, S., Yu, Y., Seo, H.: Artificial intelligence orchestration for

text-based ultrasonic simulation via self-review by multi-large language model agents. Scientific Reports **15**(1), 12474 (2025) https://doi.org/10.1038/s41598-025-97498-y

[78] Dong, Z., Lu, Z., Yang, Y.: Fine-tuning a large language model for automating computational fluid dynamics simulations. Theoretical and Applied Mechanics Letters **15**(3), 100594 (2025) https://doi.org/10.1016/j.taml.2025.100594

[79] Jia, M., Cui, Z., Hug, G.: Enhancing llms for power system simulations: A feedback-driven multi-agent framework. IEEE Transactions on Smart Grid **16**(6), 5556–5572 (2025) https://doi.org/10.1109/tsg.2025.3589114

[80] Liu, J., Lin, F., Li, X., Lim, K.H., Zhao, S.: Physics-Informed Autonomous LLM Agents for Explainable Power Electronics Modulation Design. Preprint (2024). https://doi.org/10.48550/arxiv.2411.14214

[81] Lin, J., Zhao, D., Lu, S., Li, R., Xu, X., Wang, Z., Li, W., Ji, Y., Zhang, C., Shi, L., Jin, X., Gao, H., Wang, G.: Conversational large-language-model artificial intelligence agent for accelerated synthesis of metal–organic frameworks catalysts in olefin hydrogenation. ACS Nano **19**(26), 23840–23858 (2025) https://doi.org/10.1021/acsnano.5c04880

[82] Ataei, M., Cheong, H., Grandi, D., Wang, Y., Morris, N., Tessier, A.: Elicitron: A framework for simulating design requirements elicitation using large language model agents. In: Volume 3B: 50th Design Automation Conference (DAC), pp. 03–03056. American Society of Mechanical Engineers, (2024). https://doi.org/10.1115/detc2024-143598

[83] Du, C., Nousias, S., Borrmann, A.: Towards a copilot in BIM authoring tool using a large language model-based agent for intelligent human-machine interaction. Preprint (2024). https://doi.org/10.48550/arxiv.2406.16903

[84] Ghafarollahi, A., Buehler, M.J.: Automating alloy design and discovery with physics-aware multimodal multiagent ai. Proceedings of the National Academy of Sciences **122**(4), 2414074122 (2025) https://doi.org/10.1073/pnas.2414074122

[85] Jadhav, Y., Farimani, A.B.: Large Language Model Agent as a Mechanical Designer. Preprint (2024). https://doi.org/10.48550/arxiv.2404.17525

[86] Ye, Y., Cong, X., Tian, S., Cao, J., Wang, H., Qin, Y., Lu, Y., Yu, H., Wang, H., Lin, Y., Liu, Z., Sun, M.: ProAgent: From Robotic Process Automation to Agentic Process Automation. Preprint (2023). https://doi.org/10.48550/arxiv.2311.10751

[87] Meyer, F., Freitag, L., Hinrichsen, S., Niggemann, O.: Potentials of large language models for generating assembly instructions. In: 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation

(ETFA), pp. 1–8. IEEE, (2024). https://doi.org/10.1109/etfa61755.2024.10710806

[88] Liu, J., Cui, C., Lin, P., Hui, P., Ghias, A.M.Y.M., Zhang, C.: A large language model based multi-agent framework for goal oriented controller design in power electronics. In: 2025 8th International Conference on Energy, Electrical and Power Engineering (CEEPE), pp. 1499–1502. IEEE, (2025). https://doi.org/10.1109/ceepe64987.2025.11034293

[89] Tomkou, D., Fatouros, G., Andreou, A., Makridis, G., Liarokapis, F., Dardanis, D., Kiourtis, A., Soldatos, J., Kyriazis, D.: Bridging industrial expertise and xr with llm-powered conversational agents. In: 2025 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), pp. 1050–1056. IEEE, (2025). https://doi.org/10.1109/dcoss-iot65416.2025.00158

[90] Singh, S., Fore, M., Karatzas, A., Lee, C., Jian, Y., Shangguan, L., Yu, F., Anagnostopoulos, I., Stamoulis, D.: Llm-dcache: Improving tool-augmented llms with gpt-driven localized data caching. In: 2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 1–4. IEEE, (2024). https://doi.org/10.1109/icecs61496.2024.10848749

[91] Zhou, L., Huang, Y., Yan, T., Li, D., Cai, H., Pan, L., Shi, J.: Research on the automatic bim modeling method of substation based on llm agent. IET Conference Proceedings **2024**(21), 686–692 (2025) https://doi.org/10.1049/icp.2024.4301

[92] Wang, R., Gou, J.: Human-ai collaboration empowered knowledge-oriented agent in large language model. In: 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 486–491. IEEE, (2025). https://doi.org/10.1109/icaibd64986.2025.11082074

[93] Ruan, Y., Lu, C., Xu, N., Zhang, J., Xuan, J., Pan, J., Fang, Q., Gao, H., Shen, X., Ye, N., Zhang, Q., Mo, Y.: Accelerated end-to-end chemical synthesis development with large language models (2024). https://doi.org/10.26434/chemrxiv-2024-6wmg4

[94] Crawford, N., Duffy, E.B., Evazzade, I., Foehr, T., Robbins, G., Saha, D.K., Varma, J., Ziolkowski, M.: BMW Agents – A Framework For Task Automation Through Multi-Agent Collaboration. Preprint (2024). https://doi.org/10.48550/arxiv.2406.20041

[95] Deng, H., Namoano, B., Zheng, B., Khan, S., Erkoyuncu, J.A.: From prediction to prescription: Large language model agent for context-aware maintenance decision support. PHM Society European Conference **8**(1), 10 (2024) https://doi.org/10.36001/phme.2024.v8i1.4114

[96] Xia, Y., Jazdi, N., Weyrich, M.:

Enhance fmea with large language models for assisted risk management in technical processes and products. In: 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4. IEEE, (2024). https://doi.org/10.1109/etfa61755.2024.10710996

[97] Tao, L., Huang, Q., Wu, X., Zhang, W., Wu, Y., Li, B., Lu, C., Hai, X.: LLM-R: A Framework for Domain-Adaptive Maintenance Scheme Generation Combining Hierarchical Agents and RAG. Preprint (2024). https://doi.org/10.48550/arxiv.2411.04476

[98] Lim, J., Vogel-Heuser, B., Kovalenko, I.: Large language model-enabled multi-agent manufacturing systems. In: 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), pp. 3940–3946. IEEE, (2024). https://doi.org/10.1109/case59546.2024.10711432

[99] Romero, M.L., Suyama, R.: Agentic AI for Intent-Based Industrial Automation. Preprint (2025). https://doi.org/10.48550/arXiv.2506.04980

[100] Zhao, Z., Tang, D., Liu, C., Wang, L., Zhang, Z., Zhu, H., Chen, K., Nie, Q., Ji, Y.: A Large Language Model-based multi-agent manufacturing system for intelligent shopfloor. Preprint (2024). https://doi.org/10.48550/arxiv.2405.16887

[101] Xia, Y., Shenoy, M., Jazdi, N., Weyrich, M.: Towards autonomous system: flexible modular production system enhanced with large language model agents. In: 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–8. IEEE, (2023). https://doi.org/10.1109/etfa54631.2023.10275362

[102] Wang, Z., Qin, H.: Intelligent industrial production process automatic regulation system based on llm agents. In: 2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), pp. 133–137. IEEE, (2024). https://doi.org/10.1109/aiea62095.2024.10692701

[103] Ren, L., Wang, H., Dong, J., Jia, Z., Li, S., Wang, Y., Laili, Y., Huang, D., Zhang, L., Li, B.: Industrial foundation model. IEEE Transactions on Cybernetics **55**(5), 2286–2301 (2025) https://doi.org/10.1109/tcyb.2025.3527632

[104] Kiangala, K.S., Wang, Z.: An experimental hybrid customized ai and generative ai chatbot human machine interface to improve a factory troubleshooting downtime in the context of industry 5.0. The International Journal of Advanced Manufacturing Technology **132**(5-6), 2715–2733 (2024) https://doi.org/10.1007/s00170-024-13492-0

[105] Li, Y., Wang, H., Fei, B.: Sortingbot: Leveraging large language models and 3d vision for multi-category material sorting. In: 2024 IEEE 15th Annual Information Technology, Electronics and Mobile Communication Conference

(IEMCON), pp. 346–355. IEEE, (2024). https://doi.org/10.1109/iemcon62851.2024.11093526

[106] Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Kristensen, L.B., Li, A.Z., Zhao, Y., Xu, H., Kuramshin, A., Aspuru-Guzik, A., Shkurti, F., Garg, A.: Large language models for chemistry robotics. Autonomous Robots **47**(8), 1057–1086 (2023) https://doi.org/10.1007/s10514-023-10136-2

[107] Ranasinghe, N., Mohammed, W.M., Stefanidis, K., Lastra, J.L.M.: Large language models in human-robot collaboration with cognitive validation against context-induced hallucinations. IEEE Access **13**, 77418–77430 (2025) https://doi.org/10.1109/access.2025.3565918

[108] Lin, C.-Y., Tsai, T.-H., Tseng, T.-L.: Generative ai for intelligent manufacturing virtual assistants in the semiconductor industry. IEEE Robotics and Automation Letters **10**(4), 4132–4139 (2025) https://doi.org/10.1109/lra.2025.3544506

[109] Wang, L.-C.: Llm-assisted analytics in semiconductor test (invited). In: Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD, pp. 1–7. ACM, (2024). https://doi.org/10.1145/3670474.3685974

[110] Holland, M., Chaudhari, K.: Large language model based agent for process planning of fiber composite structures. Manufacturing Letters **40**, 100–103 (2024) https://doi.org/10.1016/j.mfglet.2024.03.010

[111] Liu, Z., Zeng, R., Wang, D., Peng, G., Wang, J., Liu, Q., Liu, P., Wang, W.: Agents4PLC: Automating Closed-loop PLC Code Generation and Verification in Industrial Control Systems using LLM-based Agents. Preprint (2024). https://doi.org/10.48550/arxiv.2410.14209

[112] Xia, Y., Dittler, D., Jazdi, N., Chen, H., Weyrich, M.: Llm experiments with simulation: Large language model multi-agent system for simulation model parametrization in digital twins. In: 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4. IEEE, (2024). https://doi.org/10.1109/etfa61755.2024.10710900

[113] Jiang, W., Hu, F.: Artificial intelligence agent-enabled predictive maintenance: Conceptual proposal and basic framework. Computers **14**(8), 329 (2025) https://doi.org/10.3390/computers14080329

[114] Sun, Y., Liu, X.: Research and application of a multi-agent-based intelligent mine gas state decision-making system. Applied Sciences **15**(2), 968 (2025) https://doi.org/10.3390/app15020968

[115] Pan, Y., Sun, J., Yu, H., Luck, J., Bai, G., Chamara, N., Ge, Y., Awada, T.: Building multi-agent copilot towards autonomous agricultural data management and analysis. In: 2024 IEEE International Conference on Big Data (BigData), pp. 4384–4393. IEEE,

(2024). https://doi.org/10.1109/bigdata62323.2024.10826038

[116] Liang, G., Tong, Q.: Llm-powered ai agent systems and their applications in industry. In: 2025 IEEE World AI IoT Congress (AIIoT), pp. 0463–0471. IEEE, (2025). https://doi.org/10.1109/aiiot65859.2025.11105299

[117] Wang, W.Y., Zhang, S., Li, G., Lu, J., Ren, Y., Wang, X., Gao, X., Su, Y., Song, H., Li, J.: Artificial intelligence enabled smart design and manufacturing of advanced materials: The endless frontier in ai¡sup¿+¡/sup¿ era. Materials Genome Engineering Advances **2**(3), 56 (2024) https://doi.org/10.1002/mgea.56

[118] Berlec, T., Corn, M., Varljen, S., Podržaj, P.: Exploring decentralized warehouse management using large language models: A proof of concept. Applied Sciences **15**(10), 5734 (2025) https://doi.org/10.3390/app15105734

[119] Yang, T., Feng, P., Guo, Q., Zhang, J., Zhang, X., Ning, J., Wang, X., Mao, Z.: Autohma-llm: Efficient task coordination and execution in heterogeneous multi-agent systems using hybrid large language models. IEEE Transactions on Cognitive Communications and Networking **11**(2), 987–998 (2025) https://doi.org/10.1109/tccn.2025.3528892

[120] Feng, P., Yang, T., Liang, M., Wang, L., Gao, Y.: Oc-hmas: Dynamic self-organization and self-correction in heterogeneous multiagent systems using multimodal large models. IEEE Internet of Things Journal **12**(10), 13538–13555 (2025) https://doi.org/10.1109/jiot.2025.3545496

[121] Tian, Y., Lin, F., Zhang, X., Ge, J., Wang, Y., Dai, X., Lv, Y., Wang, F.-Y.: Logisticsvista: 3d terminal delivery services with uavs, ugvs and usvs based on foundation models and scenarios engineering. In: 2024 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), pp. 2–7. IEEE, (2024). https://doi.org/10.1109/soli63266.2024.10956119

[122] Jin, A., Ye, Y., Lee, B., Qiao, Y.: Decoagent: Large language model empowered decentralized autonomous collaboration agents based on smart contracts. IEEE Access **12**, 155234–155245 (2024) https://doi.org/10.1109/access.2024.3481641

[123] Manno, C., Manno, G., Tramontana, E.: Decentralized e-bidding for b2b procurement using blockchain and ai autonomous agents. In: 2025 33rd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 1–6. IEEE, (2025). https://doi.org/10.1109/wetice67341.2025.11091982

[124] Yin, R.K.: Validity and generalization in future case study evaluations. Evaluation **19**(3), 321–332 (2013) https://doi.org/10.1177/1356389013497081