Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077820

REVIEW

A Review of Applications and Challenges of Large Language Models for Foundry Intelligence in the Casting Industry

Yutong Guo[1,2], **Jianying Yang**[1,3] **and Chao Yang**[1,3,*]

[1]Shanghai Key Laboratory of Advanced High-Temperature Materials and Precision Forming, School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]International School of Information and Software, Dalian University of Technology, Dalian, China
[3]Inner Mongolia Research Institute, Shanghai Jiao Tong University, Hohhot, China

*Corresponding Author: Chao Yang. Email: yangchao1987@sjtu.edu.cn

Received: 17 December 2025; Accepted: 23 March 2026; Published: 08 May 2026

ABSTRACT: Large language models (LLMs) and related foundation-model workflows are emerging as promising tools for advancing foundry intelligence across the casting value chain. This review examines their applications in material design and property prediction, process parameter optimization and intelligent control, and defect detection and quality tracing in casting environments. The surveyed studies indicate that LLM-enabled systems can help integrate unstructured technical knowledge with multimodal industrial data. This integration supports composition design, simulation-assisted process optimization, diagnostic reasoning, and knowledge-grounded decision support. However, current evidence shows that the transition from pilot demonstrations to robust industrial deployment remains constrained by several practical barriers, including heterogeneous data integration, insufficient traceability across process stages, reliability under physical and safety constraints, and the latency and resource limitations of shop-floor environments. We further highlight key research directions for real-world foundry applications, including multimodal cognitive systems, lightweight domain-adapted models, trustworthy retrieval-augmented and physics-aware reasoning, and human-in-the-loop validation frameworks. Overall, the review suggests that the future of foundry intelligence will depend not only on model capability, but also on data governance, deployable system design, and reliable integration with metallurgical knowledge and industrial workflows.

KEYWORDS: Large scale language models; foundry industry; material design; process optimization; defect detection; multimodal data fusion

1 Introduction

The integration of artificial intelligence (AI) and machine learning (ML) has transformed research and industrial practice across many domains [1,2], including materials science and advanced manufacturing [3–5]. In materials engineering, data-driven methods have shown strong potential for revealing correlations among composition, processing, microstructural evolution, and macroscopic performance, thereby accelerating the design–build–test cycle [6–9]. Together with intelligent computational frameworks and automated experimentation [10–12], these approaches are increasingly regarded as key enablers for next-generation materials development and process optimization [13].

Casting is one of the most widely used manufacturing routes for metal components, but it remains among the most difficult to control. A typical casting workflow comprises tightly coupled stages, including melting and alloying, inoculation or modification, pouring and filling, solidification, heat treatment, and

subsequent machining. Deviations introduced at any stage can propagate and amplify downstream. Process outcomes are governed by a large set of interacting variables, including composition fluctuation, melt superheat, pouring temperature and velocity, mold temperature, feeding and rising conditions, cooling rate, and heat-treatment schedules. This high-dimensional coupling renders casting sensitive to disturbances and frequently reliant on tacit expertise developed through long-term shop-floor practice.

Modern foundries generate heterogeneous information streams at multiple scales, including operational records, sensor time-series, inspection images, and simulation outputs. In real production environments, these data are often noisy, incomplete, and inconsistent across sources. Misalignment in time and identity is common, particularly when linking heats, pours, molds, and downstream inspection results. Distribution shifts are also pervasive, driven by equipment aging, operator variability, raw-material fluctuations, and environmental changes. Consequently, purely data-driven models may generalize poorly across products and plants and may provide limited interpretability for high-stakes process decisions [14].

Recent progress in foundation models, particularly large language models (LLMs), provides an opportunity to bridge engineering knowledge and industrial data in casting-oriented decision-making. This trend supports the emergence of *foundry intelligence*, which should be distinguished from the broader paradigm of *smart manufacturing*. While smart manufacturing refers to the general digitization, automation, and data-driven optimization of production systems across industries, *foundry intelligence* is its domain-specific realization focused on the casting process chain [15]. It emphasizes the cognitive automation of core foundry tasks—such as alloy design, gating system optimization, and defect diagnosis—through the integration of foundry-specific constraints, including multi-physics phenomena in solidification, complex defect formation mechanisms, cross-process traceability [16], and the codification of tacit shop-floor expertise. Ultimately, foundry intelligence aims to evolve from generic automation toward self-optimizing systems grounded in metallurgical and process-engineering principles.

LLMs can process unstructured technical text, support retrieval-augmented reasoning over historical cases and engineering documents, and serve as interfaces that connect process rules, sensor streams, and enterprise databases into unified workflows. When combined with multimodal perception and process-aware constraints, these systems may support defect diagnosis, root-cause analysis, parameter recommendation, and knowledge management, enabling more systematic optimization than trial-and-error adjustments. Practical deployment nevertheless remains challenging, with critical issues including data standardization, hallucination risk and reliability, enforcement of physical and engineering constraints, and on-site latency requirements.

Although a growing body of work has explored AI applications in casting, many existing reviews remain focused on isolated scenarios and provide limited discussion of how LLM-enabled methods can be integrated into end-to-end casting workflows. In addition, the transition from promising pilot studies to robust industrial deployment remains insufficiently synthesized across the literature. To address this gap, this paper presents a structured review of large language models (LLMs), domain-adapted foundation models, and closely related enabling workflows for foundry intelligence in the casting industry. It synthesizes both direct casting evidence and manufacturing-transferable evidence, and organizes representative studies across three major application domains: materials design and performance prediction, process parameter optimization and intelligent control, and defect detection and quality tracing. In particular, this review treats LLMs not only as potential task-solving models, but also as orchestration, reasoning, and knowledge-integration layers that connect simulation tools, multimodal sensing, engineering documents, and decision-support workflows. On this basis, the review further identifies key technical bottlenecks and outlines future research directions for building reliable, deployable, and scalable foundry intelligence systems.

2 Methodological Foundations and Enabling Workflows

The purpose of this section is to organize the methodological foundations, model categories, and enabling workflows that underpin LLM-enabled foundry intelligence. Rather than focusing on full industrial case studies, this section emphasizes the technical paradigms through which language models, domain-adapted foundation models, and related complementary workflows contribute to materials design, process optimization, and defect-related decision support.

2.1 Review Methodology

To provide a structured, transparent, and application-oriented overview of the relevant literature, this article was designed as a **structured narrative review informed by PRISMA 2020 reporting principles** [17], rather than as a fully systematic review. This approach was adopted to identify, organize, and critically synthesize representative studies on large language models (LLMs), domain-adapted foundation models, and closely related enabling workflows relevant to foundry intelligence and casting-oriented manufacturing applications. The review process comprised three main steps.

1. **Literature search (search execution and query design):** A structured literature search was primarily conducted in Google Scholar, with the final main search completed on **15 December 2025**. The primary publication window was restricted to **2018–2025** in order to capture recent developments in transformer-based language models, foundation-model workflows, and their industrial applications. We primarily considered English-language records, with the review focusing mainly on journal articles and peer-reviewed conference papers. In addition, a limited number of highly relevant recent preprints from recognized repositories (e.g., arXiv) were considered only when they provided sufficient methodological detail and clear relevance to the scope of this review.

The search strategy combined two groups of terms: (i) LLM-related terms, including "LLM", "GPT", "BERT", "transformer", and "large language model"; and (ii) casting- and foundry-related terms, including "casting", "foundry", "material design", "defect detection", "quality inspection", "process monitoring", and "solidification". A combined query string was constructed by integrating these two term groups in order to retrieve studies positioned at the intersection of advanced language-model capabilities and foundry-relevant industrial tasks.

After completion of the primary screening and during subsequent manuscript revision, a small number of **highly relevant additional references** were manually added to reflect rapidly emerging developments that became available after the main search window or were identified during editorial revision. These supplementary references were used primarily to update the discussion and outlook sections, and were not used to redefine the primary screened evidence base.

2. **Screening and selection (PRISMA-style flow and eligibility):** The search returned a large number of potentially relevant records. To maintain a manageable and transparent screening scope, we retrieved and screened the **top 200 results** ranked by relevance from the combined query. All retrieved records were consolidated in a reference management spreadsheet. Duplicate checking was conducted manually based on titles and DOIs, and no duplicate records were identified among the 200 retrieved items. Screening was then performed in two stages: (1) title and abstract screening to remove clearly irrelevant studies, and (2) full-text eligibility assessment.

The selection process can be summarized as follows: **200 records were identified; after deduplication (0 records removed), 200 unique records were screened by title and abstract; 120 full-text articles were assessed for eligibility; and 60 studies met the inclusion criteria for the primary evidence base.**

Studies were included if they satisfied at least one of the following criteria: (a) they explicitly investigated LLMs, transformer-based language models, or domain-adapted foundation models; (b) they addressed foundry-relevant or casting-related tasks, including materials/property knowledge extraction, process monitoring and control, quality inspection, defect diagnosis, or engineering knowledge management; or (c) they described complementary workflows that, although not LLMs in a narrow sense, could be meaningfully integrated into LLM-enabled foundry intelligence pipelines, such as retrieval-augmented systems, multimodal diagnosis frameworks, active learning, or generative design workflows. Studies were excluded if they were (a) unrelated to casting, foundry intelligence, or closely relevant manufacturing scenarios; (b) purely conceptual discussions without analyzable methods or evaluable results; or (c) inaccessible in full text.

3. **Thematic synthesis (data extraction and synthesis):** For each included study in the primary evidence base, we extracted the application domain, model type (general-purpose vs. domain-adapted), data modality, task definition, evaluation metrics, reported performance, and deployment constraints. The selected literature was then organized into three primary application domains that form the core structure of this review: (a) Material Design and Performance Prediction, (b) Process Parameter Optimization and Intelligent Control, and (c) Defect Detection and Quality Tracing. Within each domain, we summarize representative methods and findings while also discussing practical limitations, implementation constraints, and deployment-related challenges relevant to real foundry environments. Where appropriate, the discussion further distinguishes between **direct casting evidence** and **manufacturing-transferable evidence** in order to avoid overstating the current maturity of industrial adoption.

Overall, this methodology was intended to provide a structured, critical, and application-oriented review of LLM-enabled and foundation-model-related developments relevant to foundry intelligence, while acknowledging the heterogeneity and emerging nature of the current evidence base.

2.2 Terminology and Scope of LLM Architectures

Prior to delving into specific applications, it is essential to define the scope of "large language models" and related concepts as used in this review. The suite of LLM-enabled technologies discussed herein encompasses several architectural paradigms, each serving distinct roles within foundry intelligence. Encoder-only architectures, such as BERT and scientific or domain-adapted variants including SciBERT, SteelBERT, MatBERT, and MatSciBERT, leverage bidirectional attention for deep contextual understanding and are primarily employed for structured information extraction, material property prediction, and knowledge encoding from textual sources [18–21]. In contrast, decoder-only or encoder-decoder architectures (e.g., GPT series and LLaMA) excel at autoregressive generation and are more suitable for complex reasoning, planning, code generation, and open-ended dialogue. Furthermore, multimodal large models integrate and reason across text, images, and sensor time-series, enabling visual defect inspection, multi-source data fusion, and cross-modal retrieval-augmented diagnosis. The following sections detail how these model categories are applied and combined to address specific challenges in casting [22]. For this reason, the relevance of LLM architectures in this review is evaluated not only by general manufacturing capability, but by their potential to address casting-specific challenges such as defect causality, solidification-related decision making, process-chain traceability, and the codification of tacit foundry knowledge.

2.3 Material Design and Performance Prediction

By integrating multi-dimensional data resources, including academic literature, experimental data, and simulation results, LLM-enabled methods and related foundation-model workflows are beginning to support more data- and knowledge-driven approaches to casting material composition optimization and performance prediction [23].

2.3.1 Automatic Extraction of Scientific Literature and Knowledge Modeling

Domain-specific foundation models have become an important entry point for applying LLM-enabled methods to materials and foundry intelligence. In contrast to general-purpose language models, these models are pre-trained or adapted on technical corpora from metallurgy, materials science, and manufacturing, allowing them to better capture specialized terminology, composition-processing-property relations, and implicit scientific knowledge embedded in the literature. Within the current materials domain, representative examples include task-oriented BERT variants such as SteelBERT and MatBERT, which have been used for structured information extraction, property prediction, and domain knowledge encoding from scientific texts. In addition to SteelBERT and MatBERT, MatSciBERT further demonstrates the value of materials-specific pre-training, while SciBERT provides a strong scientific-text baseline for assessing the incremental benefit of domain adaptation in downstream materials NLP tasks [18,21].

SteelBERT illustrates the value of domain-specific pre-training in ferrous-alloy applications. By leveraging a curated corpus aligned with steel-related terminology and experimental records, it improves the prediction of mechanical properties compared with a general-purpose BERT baseline. For example, on a 15Cr austenitic stainless steel dataset, SteelBERT improved the R^2 of yield strength prediction from 62.65% for a general BERT baseline to 89.85% [19]. MatBERT further reflects the broader trend toward tighter coupling between language modeling and scientific knowledge. As reviewed by Jiang et al., materials-domain NLP and LLM workflows increasingly seek to incorporate crystallographic, compositional, and physicochemical information, thereby improving the physical plausibility and practical relevance of model outputs [20].

More broadly, the significance of such models extends beyond these two examples. Prior studies have shown that domain-specific pre-training provides measurable advantages over general-purpose models in materials-science language tasks, particularly for extracting entities and relations from technical literature [24]. This suggests that the benefit of specialized foundation models is not limited to a single benchmark, but represents a more general pattern in materials informatics. Recent reviews have also highlighted a wider transition in materials science from narrow task-specific NLP pipelines toward broader LLM and foundation-model ecosystems that support literature mining, scientific reasoning, workflow assistance, and knowledge-grounded decision support [20].

From the perspective of foundry intelligence, these developments are especially relevant because casting engineering depends heavily on fragmented knowledge distributed across handbooks, research articles, process reports, and plant records. Therefore, the most valuable direction is not merely the use of isolated domain BERT models, but the evolution toward knowledge-guided and domain-adapted foundation models that can integrate textual evidence with metallurgical rules, process constraints, and engineering context. Such models may serve as the semantic backbone for downstream foundry applications, including alloy design support, process-window recommendation, and defect-related knowledge retrieval.

Nevertheless, the current evidence base remains heterogeneous. Reported results across SteelBERT, MatBERT, and related materials-domain models are often derived from different corpora, task definitions, and evaluation protocols, making direct quantitative comparison difficult [19,20,24]. Accordingly, Table 1 should be interpreted as a task-specific illustration rather than a universal ranking of domain-specific models.

For completeness, MatBERT-related results discussed in the materials-domain literature were obtained under different datasets, targets, and evaluation protocols; therefore, they are not included in Table 1 to avoid misleading direct comparisons.

The typical architecture of BERT-based models for material informatics is illustrated in Fig. 1.

Table 1: Performance comparison of SteelBERT and traditional baselines on material property prediction tasks. **All results are taken from Tian et al.** [19] (15Cr austenitic stainless steel dataset; evaluated by R^2), and are directly comparable.

Model	YS R^2	UTS R^2	EL R^2
Traditional baselines (from [19])			
General BERT	62.65%	85.97%	76.59%
Random Forest (RF)	78.21%	83.45%	80.12%
Gradient Boosting (GBR)	81.50%	85.10%	82.67%
SteelBERT (fine-tuned) [19]	**89.85%**	**88.34%**	**87.24%**

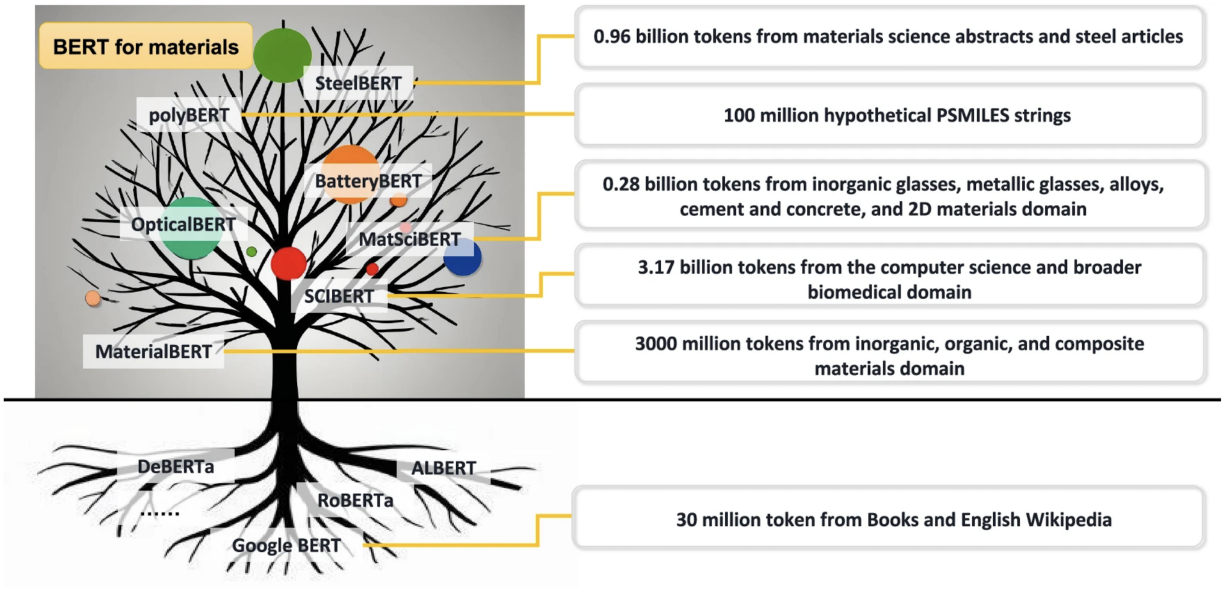

Figure 1: BERT models for materials and the applications in materials design. Reprinted/adapted with permission from Reference [20]. Copyright 2025, The Author(s).

While Table 1 demonstrates the superior performance of SteelBERT over traditional baselines and a general-purpose BERT under one directly comparable setting, it should not be interpreted as a universal ranking of domain-specific models. Their efficacy is contingent upon several stringent prerequisites. First, the requirement for high-fidelity, structured domain corpora is critical. SteelBERT's pre-training depends on a curated, ontology-aligned corpus of ferrous-alloy literature, a resource that may not exist for emerging or non-ferrous alloy systems.

Second, these models embed strong domain-specific priors and scientific constraints. Materials-domain models that incorporate crystallographic, compositional, or physicochemical information can improve extrapolation consistency and physical plausibility, but may also narrow the exploration space and bias model preferences toward established structure-property patterns [20]. Third, the issue of static knowledge representation persists. Domain-specific LLMs are fundamentally limited by the temporal scope of their training data, rendering them unable to autonomously reason about post-training research breakthroughs without explicit retrieval augmentation or continuous fine-tuning protocols.

Consequently, the selection between a general-purpose LLM (greater flexibility, broader but shallower knowledge) and a domain-specific LLM (higher accuracy, stronger domain alignment, and enhanced physical plausibility) involves a fundamental trade-off that must be evaluated against specific application objectives, data availability, and the desired balance between exploratory innovation and iterative optimization [24]. Furthermore, a comprehensive evaluation should extend beyond point estimates like RMSE to include calibrated uncertainty quantification and task-appropriate metrics. In addition, perplexity remains a fundamental metric for evaluating language-model fit on domain-specific corpora. It is important to note that for the concrete industrial applications discussed in this review, task-specific metrics are often more critical for performance assessment, including Precision, Recall, and F1-score for defect classification tasks. For Retrieval-Augmented Generation (RAG) systems, groundedness and faithfulness are key for assessing factual consistency with retrieved evidence. Benchmarking against traditional machine learning baselines (e.g., Random Forests, Gradient Boosting) is also critical. Future progress would benefit from standardized benchmarks spanning materials extraction, property prediction, scientific reasoning, and deployment-oriented evaluation in manufacturing settings.

2.3.2 Complementary Data-Efficient and Generative Workflows

Although active learning, data augmentation, and GAN-based generative design are not themselves large language models, they are discussed here as complementary workflows that may support broader LLM-enabled foundry intelligence systems. In data-scarce alloy design and defect-analysis settings, such methods may improve data availability, mitigate sampling bias, and support the robustness of downstream decision-support pipelines.

Data-driven modeling methods, particularly machine learning (ML), have shown substantial potential for accelerating alloy development. However, their predictive performance is often constrained by limited high-quality datasets and dataset bias, especially under out-of-distribution (OOD) conditions. In this context, active learning provides a practical strategy for small-sample casting-alloy research by iteratively selecting the most informative samples for experimental validation, often in combination with simple augmentation techniques such as noise injection or parameter perturbation [25]. Such workflows are particularly relevant where expensive experiments and sparse labels limit direct model generalization.

Generative methods further extend this data-efficient paradigm. In non-destructive testing (NDT), GAN-based approaches can synthesize industrial X-ray images with controllable defect patterns and paired annotations, thereby alleviating the scarcity of rare defect samples for training inspection models [26]. Related small-sample defect synthesis frameworks, such as DefectGAN and Defect Transfer GAN, have also been reported to improve the generalization of downstream classifiers and may offer potential transferability to casting inspection scenarios with limited defect labels [27]. Similarly, GAN-based synthetic defect injection has been shown to improve automatic defect recognition on public casting radiograph benchmarks [28]. Beyond radiography, conditional GANs have been used to generate microstructure images of castable aluminum alloys conditioned on composition and cooling conditions, supporting data expansion for microstructure–property learning under limited experimental observations [29]. Taken together, these studies illustrate how generative augmentation can strengthen multimodal foundry intelligence in small-sample regimes.

Complementary generative workflows have also been explored for alloy design. Rao et al. combined a multi-objective genetic algorithm (MOGA) with active learning for aluminum alloy design and reported improvements in strength-related properties relative to conventional design strategies [30,31]. More broadly, such frameworks are better understood as enabling workflows rather than direct LLM applications. Their relevance to this review lies in the role that LLM-enabled systems may play as orchestration, specification,

and interpretation layers around generative models, for example by translating natural-language design requirements into formal constraints, documenting search processes, and linking generated candidates to domain knowledge and engineering evidence.

A representative adjacent example is the cardiGAN model, which learns a mapping between composition and properties in high-entropy alloys and supports inverse design [32]. Although this case was demonstrated on high-entropy alloys rather than mainstream casting alloys, it illustrates a broader transition in materials design from empirical trial-and-error toward data-driven, goal-oriented generative search. In the foundry context, similar workflows could in principle be extended to cast irons and Al–Si alloy systems, where future domain-specific generative models trained on curated foundry databases may support inverse design for casting-relevant objectives such as castability, defect susceptibility, mechanical performance, and cost. The architecture of a representative generative model employed for such tasks is shown in Fig. 2.

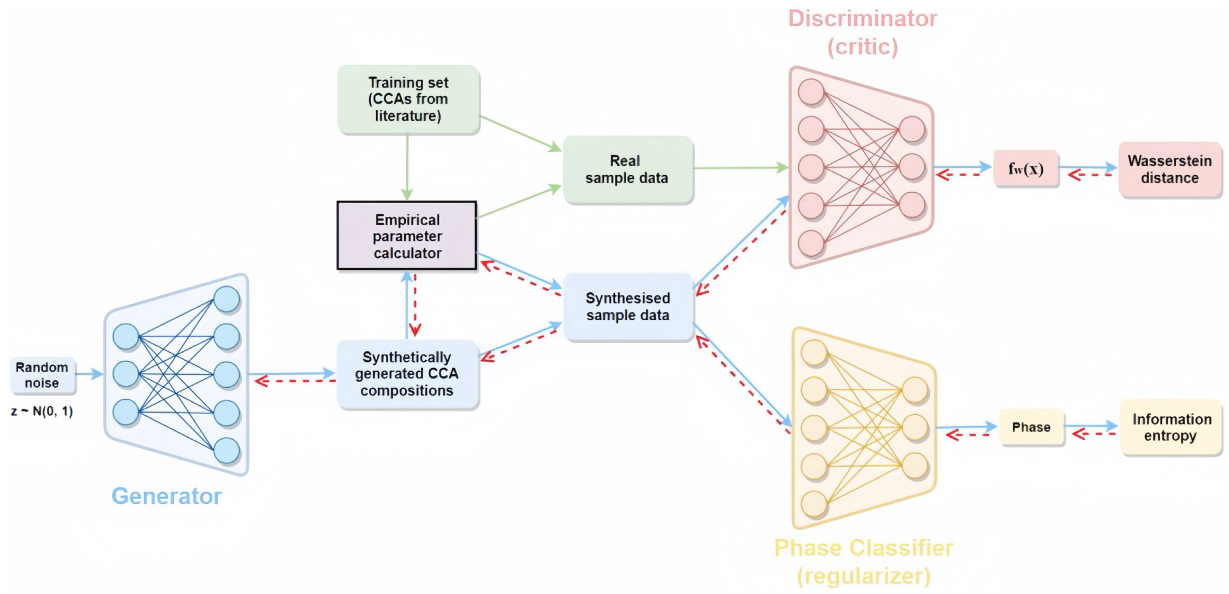

Figure 2: Configuration of the cardiGAN model employed herein. Reprinted/adapted with permission from Reference [32]. Copyright 2022, Elsevier Ltd.

2.4 Process Parameter Optimization and Intelligent Control

To realize the designed materials, optimal and stable casting processes are essential. Large language models (LLMs) enable the dynamic modeling and optimization of casting processes by deeply integrating multi-physical field simulation data with real-time sensor signals, thereby facilitating a shift from experience-driven to data-driven production [33].

2.4.1 Simulation Code Generation and Automatic Modeling

LLM-based agents (e.g., for molecular dynamics) can automatically generate input scripts for simulation software like LAMMPS based on high-level process specifications. This capability allows for the accurate simulation of critical solidification phenomena, including temperature gradients, stress distributions, and grain growth dynamics, thereby aiding in the optimization of gating system designs. Shi et al. demonstrated that this approach can reduce simulation setup time by 40%, significantly accelerating the iteration cycle for process design [34]. The architecture of such an LLM-based agent (MDAgent) is illustrated in Fig. 3.

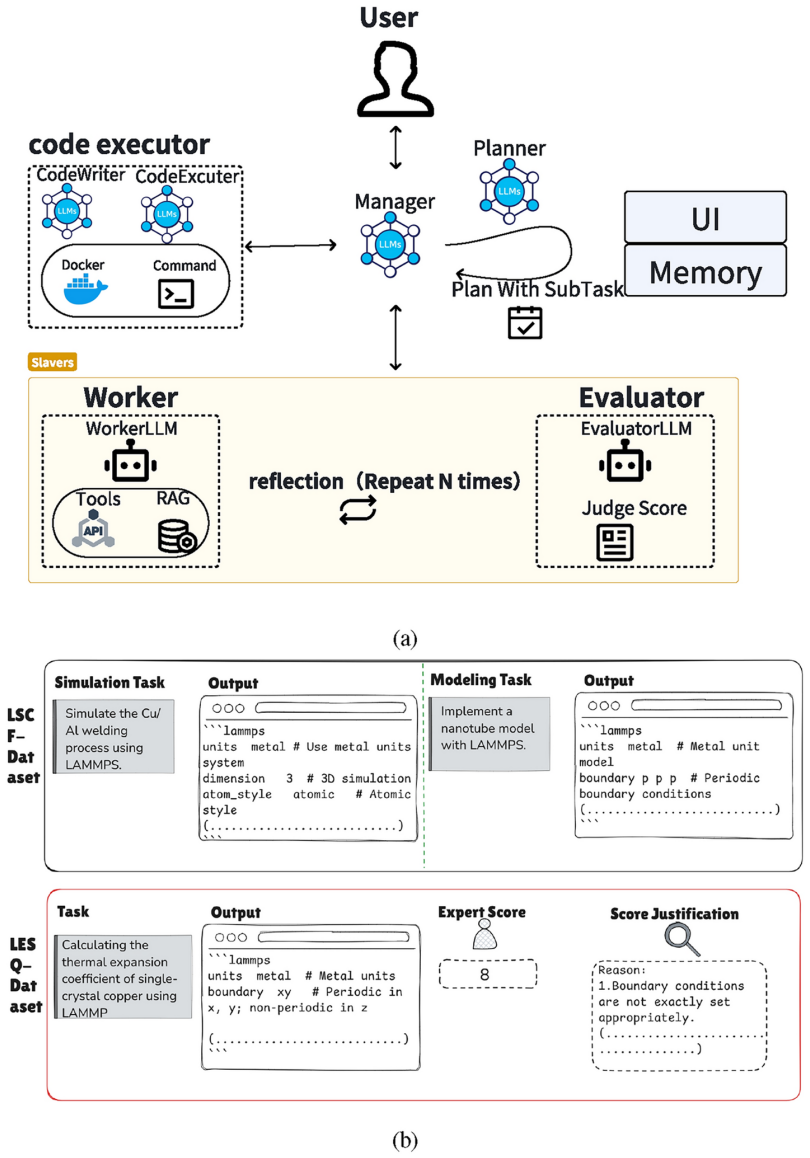

Figure 3: (**a**) Architecture diagram: MDAgent with Manager, Worker, and evaluator powered by large language models (LLMs), interacting through a user interface. (**b**) Example of the dataset used. Reprinted/adapted with permission from Reference [34]. Copyright 2025, The Author(s).

2.4.2 Electronic Laboratory Notebook and Real-Time Monitoring

The Electronic Laboratory Notebook (ELN) framework, powered by LLMs, enables real-time parsing of unstructured data from casting logs, such as operator notes and equipment status updates. By integrating with reinforcement learning (RL) algorithms, the system can dynamically identify process anomalies—including deviations in cooling rate or temperature fluctuations—and autonomously generate corrective adjustment strategies. Jalali et al. reported that such an integrated system can reduce the response time to process deviations to mere seconds, thereby mitigating quality risks associated with parameter drift [35]. An overview of the LLM ecosystem within an ELN platform (eLabFTW) is provided in Fig. 4.

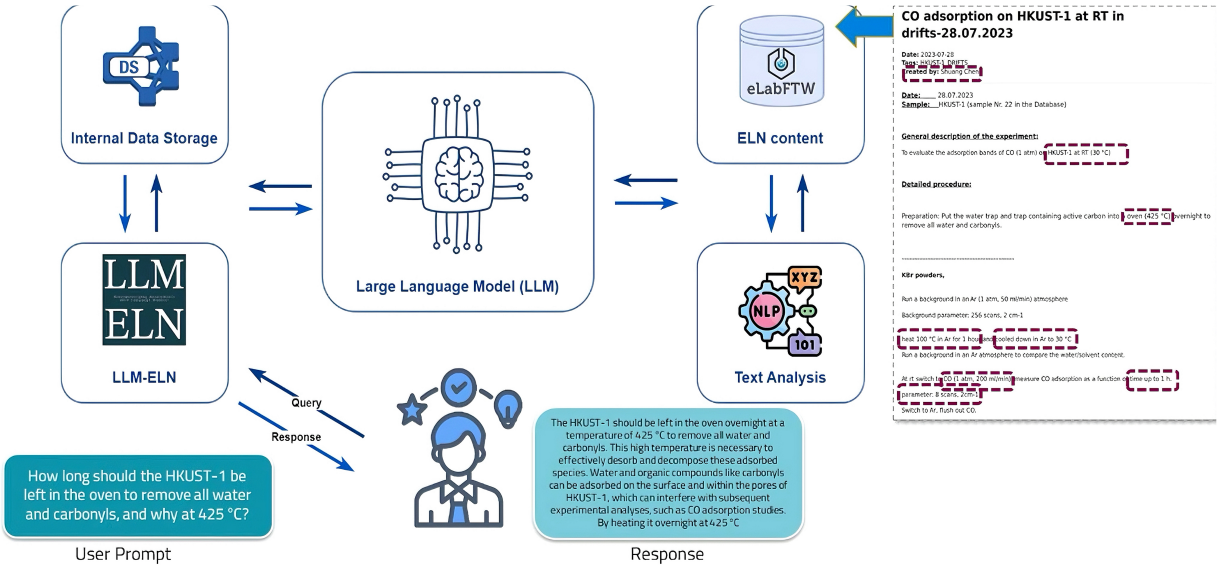

Figure 4: Illustration of Large Language Model Ecosystem in eLabFTW. Reprinted/adapted with permission from Reference [35]. Copyright 2024, The Author(s).

2.4.3 Automation and Standardization of Experimental Process

Inspired by architectures like Coscientist, LLMs can be employed to autonomously plan casting experiments—for instance, tests for melting temperature gradients or composition verification—based on high-level objectives. These models can generate standardized, hardware-agnostic experimental code, enabling full automation from experimental design to execution [36]. This approach not only enhances experimental throughput but also addresses the critical issue of inconsistent data annotation inherent in manual workflows by enforcing a unified data format [37]. The general architecture of such an automated experimental system is depicted in Fig. 5. Nevertheless, process optimization must be complemented by defect detection and tracing to close the quality assurance loop.

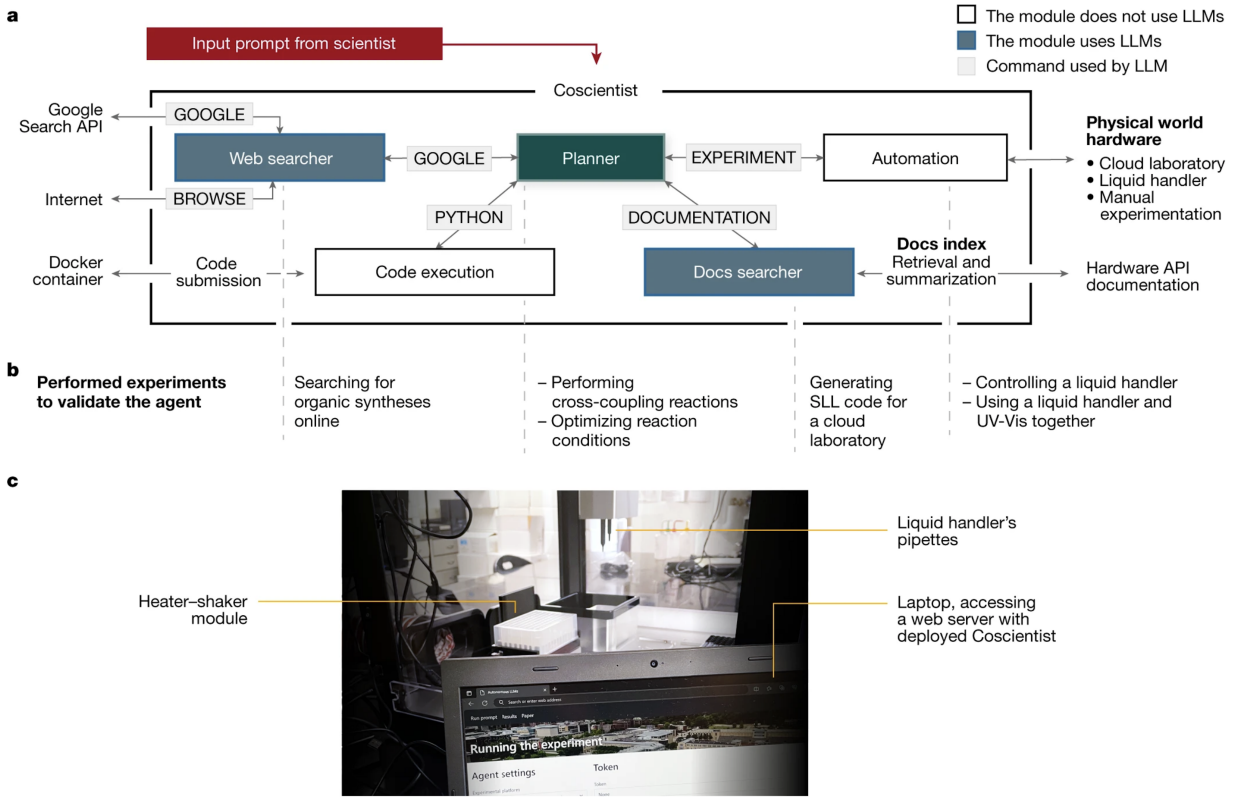

Figure 5: The system's architecture. Reprinted/adapted with permission from Reference [38]. Copyright 2023, The Author(s).

2.5 Defect Detection and Quality Tracing

Casting defects such as shrinkage porosity, gas porosity, inclusions, and hot tearing are major contributors to scrap, rework, and delivery risk. Defect formation is inherently multi-causal and spans multiple stages of the process chain, which makes diagnosis difficult when relying on a single data modality or isolated measurements. Recent studies therefore move toward multimodal learning and knowledge-grounded reasoning frameworks that integrate inspection signals with process histories and domain knowledge to support defect identification, root-cause analysis, and corrective actions.

2.5.1 Multimodal Retrieval-Augmented Diagnosis for Defect Analysis

A central challenge in industrial defect diagnosis lies in aligning heterogeneous data streams captured at varying temporal frequencies and indexed with disparate identifiers. For instance, inspection images are typically tagged by part or batch ID, sensor time-series by equipment and timestamp, and production records by heat, pour, or mold number. Establishing reliable cross-modal traceability thus necessitates an explicit data alignment layer. Common strategies involve: creating a unified traceability key across the production chain; synchronizing timestamps via salient process event markers; and resolving ambiguous links through rule-based or probabilistic matching of production metadata.

Retrieval-Augmented Generation (RAG) offers a robust framework for knowledge-intensive tasks by grounding generation in retrieved evidence. In a typical RAG pipeline, the system first extracts and structures features from each data modality. It then queries a curated knowledge base—containing historical defect cases, engineering manuals, and process guidelines—to retrieve the most relevant precedents. Finally, it

synthesizes a diagnostic report and intervention recommendations that are explicitly grounded in the retrieved evidence, enhancing both accuracy and credibility [39].

For industrial deployment, model evaluation must extend beyond pure prediction accuracy to encompass operational cost-benefit trade-offs. Metrics such as recall and precision are critical, as the cost of a false negative (missed defect) often far exceeds that of a false positive (false alarm). Furthermore, research indicates that classifiers leveraging a compact set of highly discriminative, cross-modal features can achieve competitive performance with high efficiency, frequently surpassing single-modality baselines in real-world defect classification [40]. Model robustness to distributional shift is another paramount consideration. In foundry environments, covariate shift is ubiquitous—driven by changes in alloy grades, equipment wear, operator practices, and raw material properties—and can severely degrade the performance of models calibrated on historical data.

Ensuring Reliability in RAG Systems for Foundry Diagnosis: The practical deployment of RAG for defect diagnosis hinges on its reliability—a paramount concern given the low error tolerance in industrial settings. The correctness of its outputs is not automatic but must be engineered through a multi-layered approach: (1) *Source Quality Control:* The knowledge base must be curated from trusted sources such as historical validated cases, standard operating procedures, and engineering manuals. (2) *Retrieval Confidence Scoring:* Retrieved evidence should be ranked by relevance and assigned confidence scores, allowing engineers to assess provenance and uncertainty. (3) *Constraint Checking:* Generated recommendations must be cross-checked against hard physical and process constraints (e.g., allowable temperature ranges, material safety data sheets) to filter out thermodynamically implausible or unsafe suggestions. (4) *Human-in-the-Loop Verification:* For high-impact decisions, the system should present its reasoning chain and supporting evidence, requiring human confirmation to form a closed-loop verification. Together, these layers mitigate hallucination risks and build the operational trust necessary for shop-floor adoption.

Evaluating RAG Systems: To systematically validate the effectiveness of these reliability mechanisms and the overall system, evaluation must extend beyond traditional metrics such as Precision and Recall. A rigorous assessment requires metrics that characterize both the *retrieval* and *generation* components, as well as their end-to-end interaction:

- **Retrieval:** Ranking-aware measures such as *Mean Reciprocal Rank (MRR@k)* and *Normalized Discounted Cumulative Gain (nDCG@k)* complement *Recall@k* by reflecting whether critical evidence is placed early in the retrieved list.
- **Evidence (context):** *Context Relevance* and *Context Precision/Recall* quantify how much of the retrieved context is useful and whether key supporting snippets are missing.
- **Generation:** *Answer Groundedness/Faithfulness* evaluates whether claims are supported by the retrieved context (mitigating hallucinations), while *Answer Relevance* measures whether the response addresses the query. When verified reference diagnoses are available, *Answer Correctness* can be additionally assessed.
- **Attribution and deployment:** *Citation Precision/Recall* provides a practical proxy for source attribution quality. Deployment-oriented metrics (e.g., latency, token/compute cost, and performance degradation under distributional shift) are essential for benchmarking industrial readiness.

Such multi-dimensional evaluation has been operationalized in recent RAG evaluation frameworks (e.g., RAGAS) for automated benchmarking of retrieval-grounded generation.

In the specific context of foundry defect diagnosis, the imperative for trustworthy, safety-sensitive recommendations makes *Answer Groundedness* and *Citation Precision* particularly critical. High *Recall@k* is also needed to surface rare defect patterns and long-tail failure modes from historical case bases, while

tight shop-floor latency budgets often constrain the feasible complexity of retrieval and generation modules. Ultimately, comprehensive offline test sets composed of historical defect cases with verified diagnoses remain an indispensable engineering practice for iterative model improvement, standardized comparison, and building operational trust for deployment.

2.5.2 Explainable and Actionable Defect Tracing with Process-Aware Reasoning

Interpretability and actionability are paramount for defect diagnosis systems, as corrective actions—such as adjusting process parameters, redesigning gating and risering, or modifying heat treatment cycles—carry significant cost and safety implications. Consequently, recent research focuses on integrating inherently interpretable model components or post-hoc explanation techniques to filter out irrelevant features and bolster the reliability of process-level reasoning.

Beyond merely flagging a defect, an effective system must attribute the root cause to key process variables and prioritize actionable corrective measures, guiding engineers towards efficient troubleshooting [41].

Visual interpretation frameworks and attribution methods can provide practical support for this goal. By analyzing feature importance and attention patterns, these approaches can localize key process parameters associated with defect formation and offer evidence that engineers can validate against process knowledge [42,43]. To reduce hallucination risk and improve deployment reliability, recommendations should be grounded in verifiable evidence retrieved from validated cases and constrained by process specifications and physical principles. Suggestions outside allowable parameter ranges, or incompatible with the process sequence, should be filtered or down-ranked, and high-impact recommendations can be further checked using lightweight simulation or fast surrogate models when feasible.

Overall, multimodal retrieval-augmented systems represent a promising direction for defect diagnosis and quality tracing in casting. Their practical value depends not only on predictive performance, but also on robust traceability, cost-aware evaluation, and the ability to deliver explanations and recommendations that align with process knowledge and on-site constraints.

3 Current Status and Applications

3.1 Typical Application Cases

3.1.1 Efficient Design of High Strength Casting Alloy

Evidence type: materials/manufacturing transferable evidence.

As a transferable example from data-driven alloy design, prior studies have reported that transfer learning combined with genetic algorithms can support inverse design optimization for high-performance alloys [44]. In one representative study, a response-surface model linking creep life with composition and process parameters was used to identify three promising compositional variants, and the best-performing candidate was reported to achieve a 22% increase in stress-rupture strength relative to the baseline alloy under the specific experimental setting of that study. The overall workflow for creep-life prediction and alloy design is illustrated in Fig. 6.

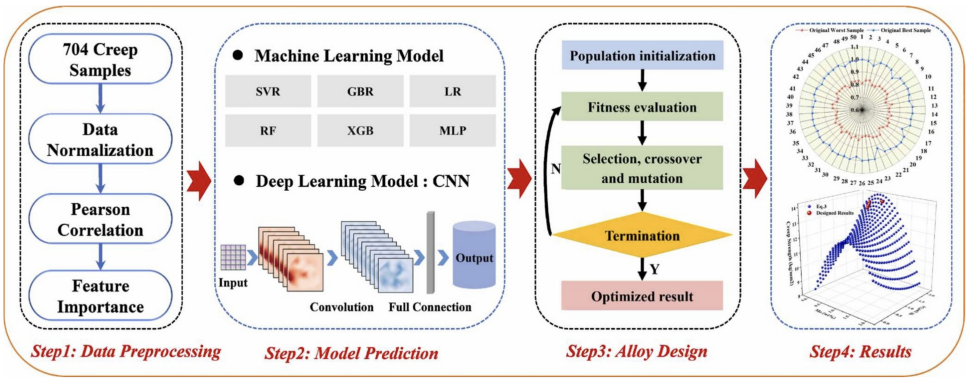

Figure 6: The basic workflow of creep life prediction and alloy design. Reprinted/adapted with permission from Reference [45]. Copyright 2025, Elsevier.

This methodology is consistent with the broader trend toward intelligent materials design. In a complementary study, Wei et al. employed transfer learning from a low-entropy alloy model coupled with active learning to identify key variables in a Cr-Mo-V alloy system, and reported a 30% improvement in fatigue life for the newly designed alloy under their pilot setting [46]. Taken together, these studies suggest that physics-informed and data-driven methodologies can improve the efficiency of alloy design relative to purely empirical trial-and-error workflows, although the reported gains are task-specific and not directly comparable across studies [45].

Critical Notes and Limitations.

The inverse-design methodology, while promising, presents several considerations for industrial deployment:

- **Data Dependency:** The performance of the transfer-learned model and the genetic algorithm is highly contingent on the quality and scope of the underlying alloy composition–property database.
- **Physical Validation Lag:** Computationally identified optimal variants still require subsequent experimental verification, adding time and cost to the design cycle.
- **Narrow Objective Focus:** Single-objective optimization (e.g., creep life) may not capture the multi-property trade-offs required for real foundry components.

These considerations indicate that transferable success in alloy design does not automatically imply immediate deployability in foundry production settings.

3.1.2 Simulation and Verification of Intelligent Gating System

Evidence type: mixed evidence (direct casting evidence and manufacturing-transferable evidence).

Direct casting studies have reported that data-driven methods can assist in the automatic generation of parametric gating-system models. In investment-casting-related work, such models were used to capture the relationships among variables such as gate size, flow rate, and temperature, enabling rapid design verification through simulation tools such as ANSYS. Yu et al. reported that this approach reduced under-pour defects and shortened the overall process design cycle by approximately 30% in their application scenario [47].

Expanding on this idea, a broader LLM-based intelligent design paradigm for manufacturing processes has also been proposed [48]. To address the dual challenges of multi-physics coupling and tacit-knowledge integration in domains including heat treatment and casting, the reported framework follows a “Knowledge Hub–Decision Engine–Verification Closed Loop” architecture. In this framework, the knowledge hub uses

a RAG architecture to integrate domain knowledge bases (e.g., ASM manuals and process databases), while the decision engine supports process optimization by fine-tuning models such as LLaMA-2 and GPT-4 for specific engineering tasks.

Under the reported pilot settings, the framework achieved task-specific gains across different manufacturing scenarios. For example, in heat-treatment applications it was reported to improve evaluation accuracy and reduce deformation out-of-tolerance rates, while on the casting side it supported natural-language generation of APDL command streams and UDF scripts for design verification. Cross-domain demonstrations further reported reductions in design-cycle duration, defect-related cost, and CAE training burden, although these outcomes were obtained under different tasks and conditions and should not be interpreted as directly comparable benchmarks [48].

Critical Notes and Limitations.

The LLM-driven simulation framework, while accelerating design, presents several considerations:

- **Simulation–Reality Gap:** The accuracy of generated simulation scripts and subsequent optimization remains bounded by the fidelity of the underlying physical models (e.g., turbulence and solidification shrinkage) in the CAE software.
- **Knowledge Base Curation:** The effectiveness of the RAG-based decision engine depends on a comprehensively curated knowledge base of defects and solutions, which is non-trivial to build and maintain.
- **Computational Overhead:** Closed-loop verification involving repeated simulation can incur substantial computational costs for complex geometries.

Accordingly, these results should be interpreted as promising pilot evidence rather than as fully mature shop-floor deployment.

3.1.3 Multi-Source Data Diagnosis of Shrinkage Defects

Evidence type: direct casting evidence.

In foundry workshops, multimodal learning models have been reported to fuse process logs or PLC signals with infrared thermal images for defect prediction and root-cause analysis, in some cases leading to measurable defect reduction after targeted interventions [49,50]. In one representative case, an analysis of shrinkage defects suggested that uneven cooling, caused by blocked cooling-system pipelines, was a major contributing factor. Based on insights obtained from multimodal data fusion, cooling-system redesign was reported to reduce defect occurrence in the studied casting process [49]. This case supports the value of integrating heterogeneous data streams, such as thermal imaging and process parameters, for diagnosing complex quality problems in casting.

Within such a pipeline, LLMs could serve as higher-level interfaces for generating diagnostic reports, retrieving relevant historical cases, and organizing engineering knowledge around the outputs of multimodal analysis. Thus, this line of work provides more direct foundry evidence for the practical relevance of multimodal and knowledge-grounded decision support, while the specific contribution of LLMs still requires further direct validation in foundry settings.

Critical Notes and Limitations.

The multimodal diagnosis approach, while effective in the reported case, presents several considerations:

- **Cross-Modal Alignment Complexity:** Establishing reliable, automated traceability between process logs, thermal images, and equipment parameters across multiple production runs remains a significant data-engineering challenge.

- **Root-Cause Ambiguity:** While correlations (e.g., blocked cooling lines) can be identified, definitively isolating a single root cause from interdependent process variables is often difficult.
- **Sensor Dependency:** The method depends on the installation, calibration, and maintenance of dedicated sensing systems such as infrared cameras, increasing infrastructure cost.

These considerations highlight that even direct foundry evidence still requires careful validation before routine deployment across different plants and products.

3.1.4 Intelligent Supply Chain and Predictive Maintenance of Equipment

Evidence type: manufacturing-transferable evidence.

Predictive-maintenance scenarios illustrate another potentially relevant application area for LLM-enabled industrial intelligence, although the available evidence is more transferable from general industrial settings than directly validated in foundry workshops. One reported platform integrates real-time sensor data, including temperature, pressure, and vibration, with LLM-assisted analysis of correlations between equipment operating status and failure modes. Under the reported pilot setting, the system predicted equipment-failure risk up to 72 h in advance, reduced production interruptions by 40%, and lowered maintenance costs by 30% [51]. These results suggest potential value for improving production reliability, but they should be interpreted as context-specific industrial evidence rather than as established foundry-wide performance expectations.

Critical Notes and Limitations.

The predictive-maintenance system, while demonstrating reported cost savings, presents several considerations:

- **High Initial Investment:** Separate capital expenditures for sensor deployment, model development, and system integration can be substantial, affecting the return-on-investment timeline.
- **Model Adaptation Needs:** Models may require frequent retraining or adaptation to maintain accuracy in the presence of equipment wear, process changes, or previously unseen failure modes.
- **Operational Reliance:** Shifting to a predictive-maintenance paradigm requires workflow changes and sustained operator trust in the model's alerts.

Therefore, this case is better regarded as transferable industrial evidence than as direct proof of mature LLM deployment in foundry environments.

3.1.5 Quantitative Performance and Database Scale

The quantitative improvements reported in the aforementioned cases, such as strength increase, cycle reduction, defect-rate reduction, and maintenance-cost reduction, are typically derived from pilot studies or controlled industrial trials and should not be interpreted as directly comparable across tasks, datasets, or operational settings. The scale of the underlying databases also varies substantially:

- The alloy-design examples in Section 3.1.1 relied on datasets ranging from several hundred to a few thousand curated alloy compositions [44–46].
- The predictive-maintenance example in Section 3.1.4 used several months of high-frequency sensor data collected from multiple machines [51].
- The SteelBERT model discussed in Section 2.3.1 was fine-tuned on approximately 50,000 annotated data points derived from ferrous-alloy literature and laboratory measurements [19].

- The reported 30% maintenance-cost reduction is an annualized indicator that mainly reflects savings from reduced downtime and spare-parts waste; initial deployment and retraining costs should be considered separately.

Regarding model comparison, the original SteelBERT study [19] reported a reduction in yield-strength prediction RMSE from approximately 8.4% for a general BERT baseline to 5.2%. However, direct statistical comparison with classical machine-learning models such as Random Forests and Gradient Boosting was not the primary objective of that study. Accordingly, the reported results should be interpreted as evidence that domain-adapted language models can be competitive for tasks involving technical-text understanding and knowledge integration, rather than as a universal ranking across modeling paradigms.

Overall, the quantitative results summarized in this section should be interpreted as context-dependent indicators derived from heterogeneous pilot studies rather than as universally transferable performance benchmarks. Further large-scale, cross-plant, and cross-product validation will be necessary before robust generalization to diverse foundry environments can be claimed.

3.2 Technical Challenges and Limitations

Despite rapid progress, deploying large language models for foundry intelligence remains constrained by several practical bottlenecks. These bottlenecks extend beyond model accuracy and can be grouped into three closely related layers: data integration and traceability, reliability under physical and safety constraints, and real-time deployment in resource-limited shop-floor environments. Addressing these issues is essential for moving from demonstrative prototypes to scalable and dependable industrial systems.

3.2.1 Multi-Source Heterogeneous Data Fusion and Standardization Bottleneck

Foundry production generates multi-source data that are heterogeneous in format, granularity, and semantics, including operational records, sensor time-series, inspection images, and simulation outputs. In many plants, the same concept is recorded with inconsistent terminology, units, and identifiers across departments and devices, which makes it difficult to build end-to-end traceability between heats, pours, molds, and downstream quality outcomes. Such inconsistencies often lead to weak cross-scene generalization and unstable performance when models are transferred across product types, lines, or factories [52].

A central technical difficulty is cross-modal alignment. Images typically correspond to a part or batch, time-series data correspond to equipment and timestamps, and production records correspond to process events and identifiers. Without a reliable alignment layer and standardized metadata schema, multimodal models can learn spurious correlations and fail when the data distribution shifts. Practical solutions should prioritize unified traceability keys, event-based synchronization along the process chain, and systematic data governance that enforces naming conventions, unit normalization, and versioned process definitions [53].

3.2.2 Construction of Trust Generation Mechanism under Physical Constraints

Ensuring model reliability and safety under rigid physical and operational constraints remains a critical barrier. Hallucinations or over-generalizations can produce recommendations that are thermodynamically impossible or breach safety protocols—for instance, suggesting pouring parameters that promote oxide entrainment or heat-treatment cycles that induce quench cracking. Mitigating these risks requires embedding first-principles knowledge (e.g., from thermodynamics, fluid dynamics, and solid mechanics) directly into the model's reasoning process. Promising approaches include constraint-guided decoding, where generation is bounded by hard limits derived from material databases and process windows; verification via lightweight digital twins using reduced-order simulations; and providing explicit, calibrated uncertainty estimates to flag high-risk predictions made in data-sparse regimes [54].

A practical trust mechanism should enforce physical and engineering constraints during generation and decision making. This can be implemented through constraint-aware decoding, rule-based validation against process specifications, and retrieval grounding using verified historical cases and standard operating procedures. Recommendations should be accompanied by confidence indicators and risk levels, and outputs that conflict with hard constraints should be rejected or down-ranked before being presented to engineers. Dynamic boundary control for generated parameters is particularly important in shop-floor scenarios where process windows are narrow and disturbances are frequent [54].

3.2.3 Challenges of Real-Time Edge Deployment

The stringent latency requirements of shop-floor control, often ranging from milliseconds to seconds, conflict sharply with the substantial computational load of foundational LLMs. Deploying these models on resource-constrained edge devices therefore requires co-optimization at both the algorithmic and system levels. Algorithmic strategies include model-compression techniques such as knowledge distillation, quantization, and pruning. However, aggressive compression may lead to non-negligible performance degradation, especially for rare defect patterns or other low-frequency but high-impact events [55]. A representative overview of fine-tuning lightweight LLMs is shown in Fig. 7.

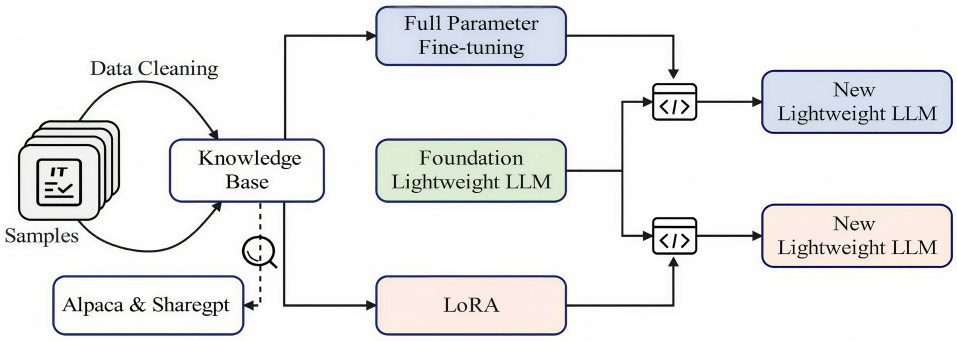

Figure 7: Overview of fine-tuning lightweight LLM. Reprinted/adapted with permission from Reference [55]. Copyright 2025, IEEE.

At the system level, a hierarchical intelligence architecture offers a more practical deployment path. In such a setup, ultra-lightweight models perform continuous monitoring on edge devices and trigger more complex LLM-powered diagnostic queries to an on-premise server only when anomalies are detected. This edge-triggered, cloud-assisted paradigm balances low-latency response with deeper analytical capability and provides a realistic route toward deployment in resource-limited industrial environments [56].

4 Prospects and Conclusions

4.1 Future Work Vision

4.1.1 Multimodal Cognitive Systems for Foundry Intelligence

Real-time, closed-loop process control requires systems that can perceive, reason over, and respond to multimodal data streams with low latency. Future research should therefore focus on multimodal cognitive architectures in which LLMs serve as a high-level reasoning layer that integrates visual data (e.g., cameras and X-ray images), temporal data (e.g., sensor streams), and textual knowledge (e.g., logs, manuals, and engineering records). Key research challenges include efficient cross-modal attention, robust alignment

across heterogeneous data sources, and dynamic knowledge retrieval from continuously updated plant databases to support context-aware decision-making [57,58].

To support industrial use, future multimodal systems should be evaluated not only by perception or reasoning accuracy, but also by their ability to maintain stable end-to-end performance under realistic production conditions. This will require closer integration between multimodal modeling, data engineering, and system-level optimization [59,60]. A conceptual diagram for fine-tuning LLMs for manufacturing is shown in Fig. 8.

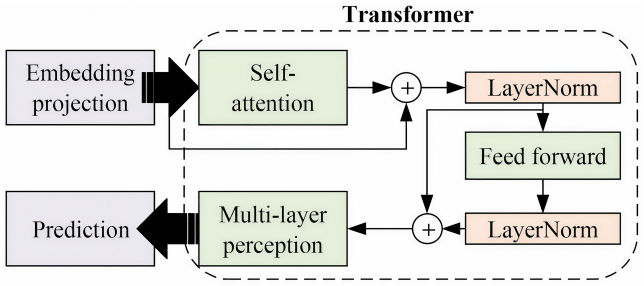

Figure 8: Fine-tuning large-scale language models for the manufacturing domain. Reprinted/adapted with permission from Reference [60]. Copyright 2024, The Author(s).

4.1.2 Engineering Adaptation of Lightweight Domain Models

Given the specialized terminology, narrow process windows, and constrained computational environment of foundries, a promising direction is the development of lightweight, domain-adapted LLMs. Techniques such as Low-Rank Adaptation (LoRA) and other parameter-efficient fine-tuning (PEFT) methods make it possible to customize large base models with limited additional parameters. Coupling such models with metallurgical knowledge graphs or domain-specific process rules may further improve the semantic understanding of foundry concepts such as feeding efficiency, hot tearing susceptibility, and defect causality, thereby enhancing both robustness and engineering relevance [61].

In the longer term, domain-adapted foundry models may provide a more practical balance among accuracy, interpretability, and deployment cost than fully general-purpose models. A comparison of different dataset fusion paradigms relevant to this adaptation is presented in Fig. 9.

4.1.3 Knowledge-Enhanced and Physics-Aware Intelligent Decision Paradigms

Another important direction is the construction of hybrid decision-making systems that combine retrieval-augmented generation (RAG), process-aware reasoning, physics-informed constraint mechanisms, and optimization or reinforcement-learning strategies [62]. Such systems could support a closed-loop decision process of "data retrieval–strategy generation–effect feedback," thereby shifting process optimization from experience-driven practice toward knowledge- and data-driven decision support. By connecting dynamic technical knowledge bases, such as ASM Handbook or process databases, with online optimization modules, future foundry systems may achieve more adaptive and evidence-grounded parameter recommendation [63].

Preliminary studies suggest that such knowledge-enhanced paradigms can reduce trial-and-error costs and shorten process-optimization cycles, although further validation in foundry-specific scenarios remains necessary [64].

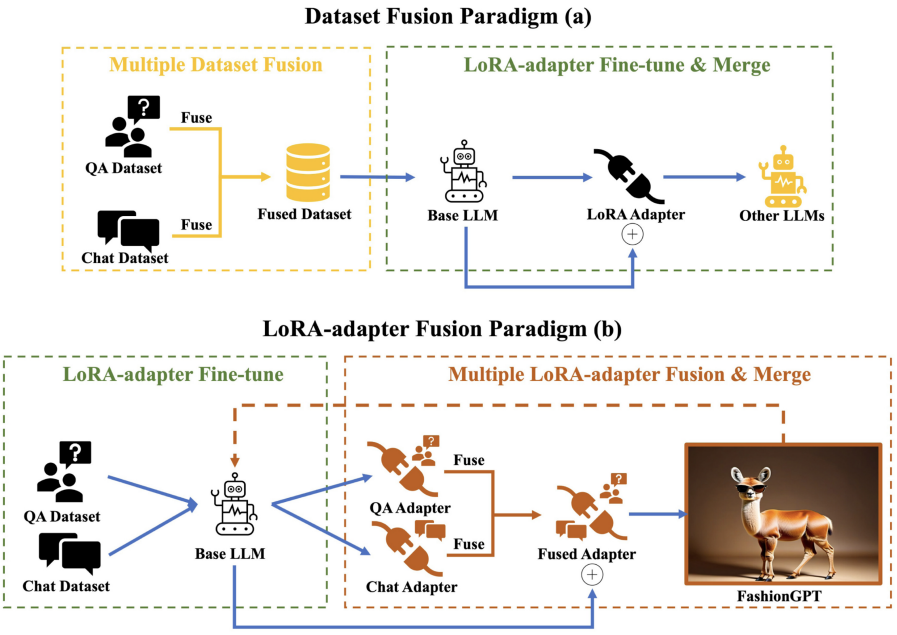

Figure 9: Comparison between dataset fusion paradigm. Reprinted/adapted with permission from Reference [61]. Copyright 2024, Elsevier.

4.1.4 Closed-Loop Autonomous Agents and Verification

A longer-term research frontier is the development of autonomous or semi-autonomous foundry workflows that connect material design, simulation verification, experiment execution, and model updating into a closed loop. Inspired by the paradigm of "AI agent + robotic experiment," such systems could accelerate the iterative cycle of composition design, performance prediction, experimental validation, and model refinement [65]. In foundry settings, this idea is particularly relevant because process optimization often depends on repeated interaction among simulation, testing, and engineering judgment.

Future intelligent casting systems may therefore evolve toward cyber-physical architectures that integrate sensing, embedded simulation, decision support, and control. Such systems could provide the foundation for more adaptive casting processes capable of predicting and optimizing defect formation, microstructure, properties, and service life in a unified framework [66]. Various status-sensing methods applicable to the casting process are summarized in Fig. 10.

4.1.5 Sustainability-Oriented Life-Cycle Optimization

Sustainability should become an increasingly important extension of future foundry intelligence. Rather than optimizing only defect rate, throughput, or process stability, future systems should incorporate material utilization, energy consumption, and environmental impact into a unified optimization framework. One possible direction is a multi-level architecture spanning process optimization, operational scheduling, and system-level life-cycle assessment, so that production efficiency and sustainability objectives can be considered simultaneously.

Such a framework may combine process-level improvement methods, reinforcement-learning-based multi-objective optimization, and life-cycle assessment modules for carbon-footprint tracking. Extending intelligent control strategies across different casting scenarios may further improve the energy efficiency of multi-variety production while preserving quality performance. The material and energy flow of a CRIMSON sand casting process is depicted in Fig. 11.

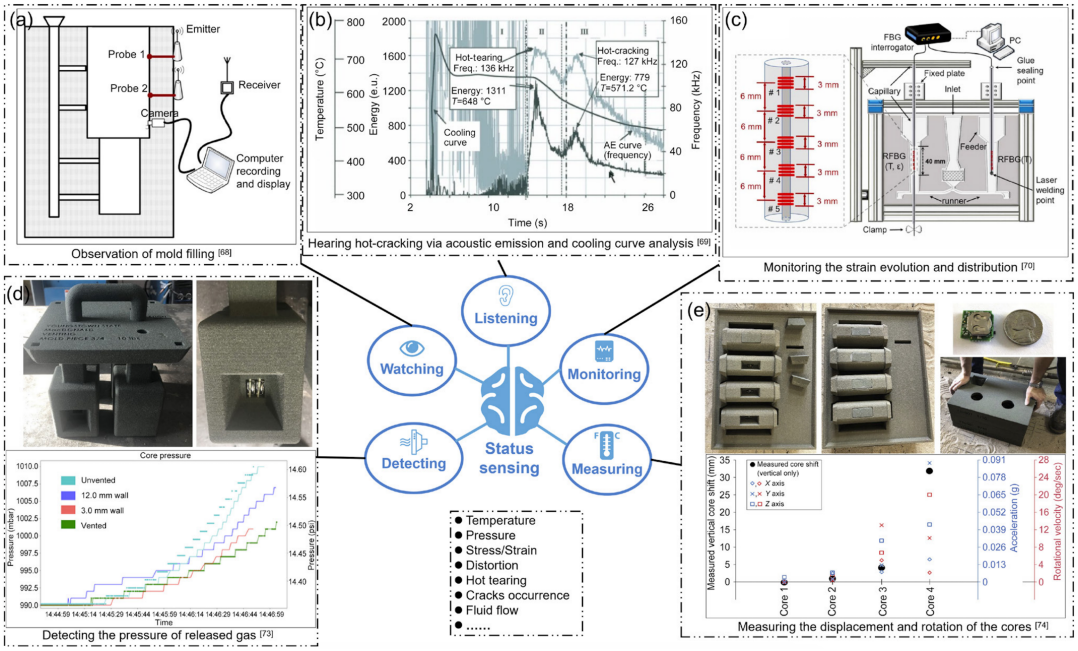

Figure 10: Status sensing methods applied in casting process (**a**–**e**). Reprinted/adapted with permission from Reference [66]. Copyright 2024, Foundry Journal Agency.

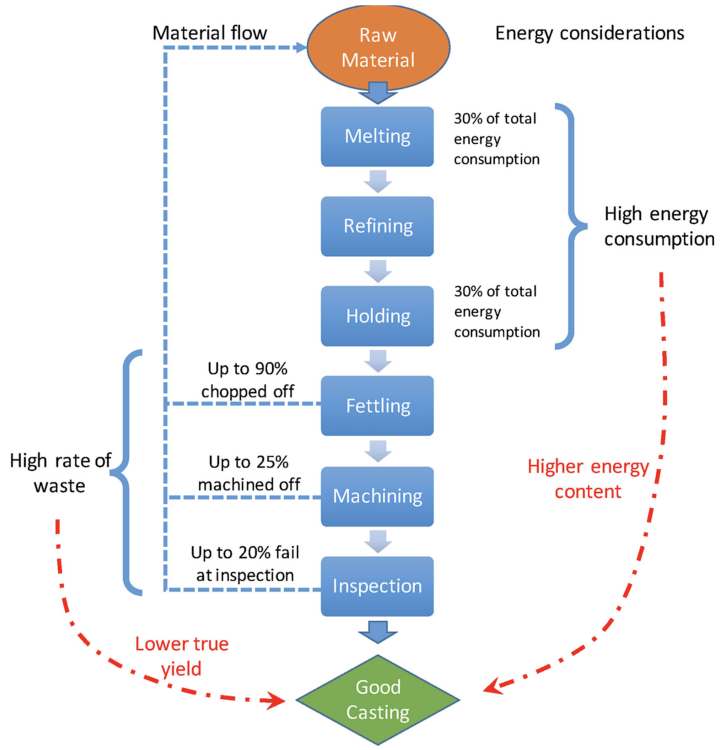

Figure 11: Material and energy flow chart of a CRIMSON sand casting process. Reprinted/adapted with permission from Reference [67]. Copyright 2016, The Authors.

4.2 Concerns and Outlook for Real-World Deployment of LLMs in Foundry Environments

Building on the technical bottlenecks summarized in Section 3.2, this section discusses their implications for the real-world deployment of LLM-enabled foundry intelligence. Industrial adoption should be evaluated not only by the performance gains reported in pilot studies, but also by the ability of these systems to operate reliably under the heterogeneous, safety-sensitive, and resource-constrained conditions of actual foundry workshops. Although recent studies have reported promising results in alloy design, process optimization, defect diagnosis, and predictive maintenance, large-scale deployment remains constrained by a set of interdependent engineering requirements [44–46]. Importantly, these requirements are not merely technical abstractions: in many foundry environments, fragmented records, weak cross-stage identifiers, limited digital infrastructure, and incomplete knowledge-base curation remain practical barriers to implementation [47,49,50]. First, data governance and cross-process traceability remain foundational concerns. Foundry data are typically derived from operational logs, sensor streams, inspection images, simulation outputs, and enterprise records, yet these sources often differ in format, sampling frequency, semantics, and identifier conventions. Without unified traceability keys and standardized metadata schemas linking heats, pours, molds, process events, and downstream quality outcomes, multimodal systems remain vulnerable to spurious correlations, weak transferability, and unstable performance across product types or production lines [52,53]. Accordingly, practical deployment requires stronger data lineage design, naming and unit normalization, and event-synchronized multimodal alignment as prerequisites for trustworthy system integration [52,53].

Second, reliability under physical and safety constraints is a decisive barrier for shop-floor use. In foundry environments, model outputs are not merely informative; they may influence pouring parameters, heat-treatment schedules, gating design, or defect-handling decisions. Hallucinated or poorly calibrated recommendations can therefore become thermodynamically implausible, operationally unsafe, or economically costly, highlighting the need for verification-oriented safeguards in process-control settings [54,68]. A practical deployment pathway requires retrieval-grounded reasoning with explicit provenance, rule-based validation against process windows and operating procedures, and, where feasible, lightweight digital-twin or surrogate-model verification before high-impact recommendations are executed [54,68]. Human-in-the-loop confirmation should therefore remain essential for safety-critical interventions [54]. In foundry settings, the consequences of model failure can be highly asymmetric. For example, an erroneous recommendation on pouring temperature, feeding strategy, or heat-treatment scheduling may lead not only to reduced quality, but also to scrap of an entire batch, unnecessary energy consumption, delayed delivery, or even downstream safety risk if latent defects remain undetected. Similarly, false-negative diagnostic outputs in defect tracing may be more consequential than false positives in many production scenarios. These considerations imply that deployment-oriented evaluation should explicitly incorporate failure severity, intervention cost, and risk tolerance, rather than relying on average predictive performance alone. Third, industrial deployability depends on real-time performance under constrained computing conditions. Shop-floor scenarios often require responses on millisecond-to-second timescales, whereas full-scale foundation models can impose substantial memory and latency overhead. Edge-oriented deployment should therefore rely on tiered architectures in which lightweight local models handle continuous monitoring and anomaly triggering, while more complex reasoning is offloaded to on-premise or cloud resources only when necessary [55,56]. In this context, deployment evaluation should extend beyond isolated model accuracy to include end-to-end latency, tail latency under peak load, throughput, memory footprint, energy consumption, and quality retention relative to cloud-based baselines [55,56,59,60].

Fourth, organizational adoption and life-cycle maintenance should not be underestimated. Several reviewed applications implicitly assume stable workflows, curated knowledge bases, frequent retraining capability, and operator trust in model-generated alerts or recommendations. In practice, however, knowledge-base maintenance, model updating under equipment wear and process drift, and integration into existing engineering decision chains can become major barriers [48,51]. Future systems should therefore be designed as decision-support infrastructures rather than fully autonomous replacements, with clear accountability, transparent evidence presentation, and continuous post-deployment monitoring [48,51].

Overall, progress toward real-world foundry deployment will depend on whether these technical and organizational requirements can be translated into reliable, traceable, and maintainable decision-support infrastructures rather than isolated pilot systems.

4.3 Conclusions

This review has examined the emerging role of large language models (LLMs), domain-adapted foundation models, and related enabling workflows in foundry intelligence across three major application domains: material design and performance prediction, process parameter optimization and intelligent control, and defect detection and quality tracing. The surveyed literature suggests that these methods are most realistically positioned not as universal stand-alone engines for every foundry task, but as knowledge-integration, reasoning, and workflow-orchestration layers that connect unstructured technical knowledge, simulation outputs, sensor data, and engineering decision processes.

At the same time, the current evidence base remains heterogeneous and is still dominated by pilot studies, controlled scenarios, and task-specific implementations. Accordingly, the significance of LLMs for foundry applications should be judged not by isolated performance gains alone, but by their ability to support reliable, traceable, and deployable industrial decision-support systems.

If advances in data governance, domain adaptation, trustworthy reasoning, and deployment engineering continue in parallel, LLM-enabled foundry intelligence may become an important enabling layer for the transition from experience-driven casting practice to more systematic, data- and knowledge-driven foundry manufacturing.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Shanghai Natural Science Foundation [Grant Number 25ZR1401430] and the Science and Technology Cooperation Program of Shanghai Jiao Tong University in Inner Mongolia Autonomous Region—Action Plan of Shanghai Jiao Tong University for "Revitalizing Inner Mongolia through Science and Technology" [Grant Number 2023XYJG0001-01-01].

Author Contributions: Yutong Guo: Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing—Original Draft. Jianying Yang: Investigation, Writing—Review & Editing. Chao Yang: Conceptualization, Writing—Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This is a review article. No new data were generated. All data discussed are available from the corresponding publications cited in the reference list.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang WY, Zhang S, Li G, Lu J, Ren Y, Wang X, et al. Artificial intelligence enabled smart design and manufacturing of advanced materials: the endless Frontier in AI^{+} era. Mater Genome Eng Adv. 2024;2(3):e56. doi:10.1002/mgea.56.
2. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater. 2017;3(1):54. doi:10.1038/s41524-017-0056-5.
3. Hart GLW, Mueller T, Toher C, Curtarolo S. Machine learning for alloys. Nat Rev Mater. 2021;6(8):730–55. doi:10.1038/s41578-021-00340-w.
4. Gu L, Liu Y, Chen P, Huang H, Chen N, Li Y, et al. Bond sensitive graph neural networks for predicting high temperature superconductors. Mater Genome Eng Adv. 2024;2(2):e48. doi:10.1002/mgea.48.
5. Xie J. Prospects of materials genome engineering frontiers. Mater Genome Eng Adv. 2023;1(2):e17. doi:10.1002/mgea.17.
6. Li S, Dong Z, Jin J, Pan H, Hu Z, Hou R, et al. Optimal design of high-performance rare-earth-free wrought magnesium alloys using machine learning. Mater Genome Eng Adv. 2024;2(2):e45. doi:10.1002/mgea.45.
7. Jiang X, Wang Y, Jia B, Qu X, Qin M. Using machine learning to predict oxygen evolution activity for transition metal hydroxide electrocatalysts. ACS Appl Mater Interfaces. 2022;14(36):41141–8. doi:10.1021/acsami.2c13435.
8. Xue D, Balachandran PV, Hogden J, Theiler J, Xue D, Lookman T. Accelerated search for materials with targeted properties by adaptive design. Nat Commun. 2016;7(1):11241. doi:10.1038/ncomms11241.
9. Jiang X, Jia B, Zhang G, Zhang C, Wang X, Zhang R, et al. A strategy combining machine learning and multiscale calculation to predict tensile strength for pearlitic steel wires with industrial data. Scr Mater. 2020;186(10):272–7. doi:10.1016/j.scriptamat.2020.03.064.
10. Wen C, Wang C, Zhang Y, Antonov S, Xue D, Lookman T, et al. Modeling solid solution strengthening in high entropy alloys using machine learning. Acta Mater. 2021;212:116917. doi:10.1016/j.actamat.2021.116917.
11. Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, et al. Machine learning assisted design of high entropy alloys with desired property. Acta Mater. 2019;170:109–17. doi:10.1016/j.actamat.2019.03.010.
12. Zhang Y, Wen C, Wang C, Antonov S, Xue D, Bai Y, et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. Acta Mater. 2020;185(3):528–39. doi:10.1016/j.actamat.2019.11.067.
13. Scharf S, Sander B, Kujath M, Richter H, Riedel E, Stein H, et al. FOUNDRY 4.0: an innovative technology for sustainable and flexible process design in foundries. Procedia CIRP. 2021;98:73–8. doi:10.1016/j.procir.2021.01.008.
14. Wilson LA, Rao GR. Frontiers of materials research: a decadal survey. MRS Bull. 2017;42:537. doi:10.1557/mrs.2017.153.
15. Saxena P, Papanikolaou M, Pagone E, Salonitis K, Jolly MR. Digital manufacturing for foundries 4.0. In: Light Metals 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 1019–25. doi:10.1007/978-3-030-36408-3_138.
16. Uyan TÇ, Otto K, Silva MS, Vilaça P, Armakan E. Industry 4.0 foundry data management and supervised machine learning in low-pressure die casting quality improvement. Int J Met. 2023;17(1):414–29. doi:10.1007/s40962-022-00783-z.
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;371:n71. doi:10.1136/bmj.n71.
18. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3—7; Hong Kong, China. Stroudsburg, PA, USA: ACL; 2019. p. 3613–8. doi:10.18653/v1/d19-1371.
19. Tian S, Jiang X, Wang W, Jing Z, Zhang C, Zhang C, et al. Steel design based on a large language model. Acta Mater. 2025;285:120663. doi:10.1016/j.actamat.2024.120663.
20. Jiang X, Wang W, Tian S, Wang H, Lookman T, Su Y. Applications of natural language processing and large language models in materials discovery. npj Comput Mater. 2025;11(1):79. doi:10.1038/s41524-025-01554-0.
21. Gupta T, Zaki M, Anoop Krishnan NM, Mausam. MatSciBERT: a materials domain language model for text mining and information extraction. npj Comput Mater. 2022;8(1):102. doi:10.1038/s41524-022-00784-w.

22. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv:2303.18223. 2023.
23. Li H, Yang J, Yao J, Sheng C. Digitized material design and performance prediction driven by high-throughput computing. Front Mater. 2025;12:1599439. doi:10.3389/fmats.2025.1599439.
24. Trewartha A, Walker N, Huo H, Lee S, Cruse K, Dagdelen J, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns. 2022;3(4):100488. doi:10.1016/j.patter.2022.100488.
25. Hu M, Tan Q, Knibbe R, Xu M, Liang G, Zhou J, et al. Designing unique and high-performance Al alloys via machine learning: mitigating data bias through active learning. Comput Mater Sci. 2024;244(1):113204. doi:10.1016/j.commatsci.2024.113204.
26. Gosswami BM, Chanda RK, Wittenberg T. Synthetic X-ray image generation for non-destructive testing using generative adversarial networks. EJNDT. 2024;29(6):29930. doi:10.58286/29930.
27. Wang R, Hoppe S, Monari E, Huber MF. Defect transfer GAN: diverse defect synthesis for data augmentation. arXiv:2302.08366. 2023.
28. Garca-Prez A, Gmez-Silva MJ, de la Escalera-Hueso A. Improving automatic defect recognition on GDXRay castings dataset by introducing GenAI synthetic training data. NDT E Int. 2025;151(3):103303. doi:10.1016/j.ndteint.2024.103303.
29. Yin B, Fan Y. Simulating castable aluminum alloy microstructures with AlloyGAN deep learning model. In: TMS, 2024 153rd Annual Meeting & Exhibition Supplemental Proceedings. Cham, Switzerland: Springer Nature; 2024. p. 804–11. doi:10.1007/978-3-031-50349-8_69.
30. Rao Z, Bajpai A, Zhang H. Active learning strategies for the design of sustainable alloys. Phil Trans R Soc A. 2024;382(2284):20230242. doi:10.1098/rsta.2023.0242.
31. Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42. doi:10.1038/s41524-023-01000-z.
32. Li Z, Nash WT, O'Brien SP, Qiu Y, Gupta RK, Birbilis N. cardiGAN: a generative adversarial network model for design and discovery of multi principal element alloys. J Mater Sci Technol. 2022;125(2):81–96. doi:10.1016/j.jmst.2022.03.008.
33. Du Q, Ellingsen K, M'Hamdi M, Marthinsen A, Tveito KO. The integration of neural network and high throughput multi-scale simulation for establishing a digital twin for aluminium billet DC-casting. Mater Trans. 2023;64(2):360–5. doi:10.2320/matertrans.mt-la2022038.
34. Shi Z, Xin C, Huo T, Jiang Y, Wu B, Chen X, et al. A fine-tuned large language model based molecular dynamics agent for code generation to obtain material thermodynamic parameters. Sci Rep. 2025;15(1):10295. doi:10.1038/s41598-025-92337-6.
35. Jalali M, Luo Y, Caulfield L, Sauter E, Nefedov A, Wöll C. Large language models in electronic laboratory notebooks: transforming materials science research workflows. Mater Today Commun. 2024;40(3):109801. doi:10.1016/j.mtcomm.2024.109801.
36. Ruan Y, Lu C, Xu N, He Y, Chen Y, Zhang J, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. Nat Commun. 2024;15(1):10160. doi:10.1038/s41467-024-54457-x.
37. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525–35. doi:10.1038/s42256-024-00832-8.
38. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570–8. doi:10.1038/s41586-023-06792-0.
39. Petrich J, Snow Z, Corbin D, Reutzel EW. Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing. Addit Manuf. 2021;48(10):102364. doi:10.1016/j.addma.2021.102364.
40. Poudel A, Yasin MS, Ye J, Liu J, Vinel A, Shao S, et al. Feature-based volumetric defect classification in metal additive manufacturing. Nat Commun. 2022;13(1):6369. doi:10.1038/s41467-022-34122-x.
41. Birihanu E, Lendák I. Explainable correlation-based anomaly detection for industrial control systems. Front Artif Intell. 2025;7:1508821. doi:10.3389/frai.2024.1508821.

42. Li M, Peng P, Zhang J, Wang H, Shen W. SCCAM: supervised contrastive convolutional attention mechanism for ante-hoc interpretable fault diagnosis with limited fault samples. IEEE Trans Neural Netw Learn Syst. 2024;35(5):6194–205. doi:10.1109/TNNLS.2023.3313728.
43. Humphreys J, Dam HK. An explainable deep model for defect prediction. In: Proceedings of the 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE); 2019 May 28; Montreal, QC, Canada. p. 49–55. doi:10.1109/raise.2019.00016.
44. Wang C, Wei X, Ren D, Wang X, Xu W. High-throughput map design of creep life in low-alloy steels by integrating machine learning with a genetic algorithm. Mater Des. 2022;213:110326. doi:10.1016/j.matdes.2021.110326.
45. Pan C, Wang C, Zhang Y, Wei X, Xu W. High-throughput design strategy for creep-resistance steel using machine learning and optimization algorithm. Mater Today Commun. 2025;46:112467. doi:10.1016/j.mtcomm.2025.112467.
46. Wei X, van der Zwaag S, Jia Z, Wang C, Xu W. On the use of transfer modeling to design new steels with excellent rotating bending fatigue resistance even in the case of very small calibration datasets. Acta Mater. 2022;235:118103. doi:10.1016/j.actamat.2022.118103.
47. Yu J, Wang D, Li D, Tang D, Hao X, Tan S, et al. Engineering computing and data-driven for gating system design in investment casting. Int J Adv Manuf Technol. 2020;111(3):829–37. doi:10.1007/s00170-020-06143-7.
48. Sun Y, Li X, Liu C, Deng X, Zhang W, Wang J, et al. Development of an intelligent design and simulation aid system for heat treatment processes based on LLM. Mater Des. 2024;248(9):113506. doi:10.1016/j.matdes.2024.113506.
49. Kim J, Park C, Park W, Park Y, Cho C, Kim D. A study on high pressure die-casting defect prediction deep learning algorithm for porosity defect detection based on process parameters and thermal image. Korean J Comput Des Eng. 2023;28(3):222–31. doi:10.7315/cde.2023.222.
50. Michno T, Holom R, Schmalzer S, Scampone G, Riegler E, Hartmann M, et al. Porosity classification in high pressure die casting using thermal images and sensor data fusion via fuzzy cognitive maps. In: Proceedings of the 21th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2026); 2026 Mar 9–11; Marbella, Spain.
51. Abbas A. AI for predictive maintenance in industrial systems. Intl J Adv Eng Technol Innov. 2024;1(1):31–51. doi:10.31219/osf.io/vq8zg.
52. Suthar J, Persis J, Gupta R. Critical parameters influencing the quality of metal castings: a systematic literature review. Int J Qual Reliab Manag. 2023;40(1):53–82. doi:10.1108/ijqrm-11-2020-0368.
53. Gao J, Zhong J, Liu G, Zhang S, Zhang J, Liu Z, et al. Accelerated discovery of high-performance Al-Si-Mg-Sc casting alloys by integrating active learning with high-throughput CALPHAD calculations. Sci Technol Adv Mater. 2023;24(1):2196242. doi:10.1080/14686996.2023.2196242.
54. Chakraborty N, Ornik M, Driggs-Campbell K. Hallucination detection in foundation models for decision-making: a flexible definition and review of the state of the art. ACM Comput Surv. 2025;57(7):1–35. doi:10.1145/3716846.
55. Tang X, Liu F, Xu D, Jiang J, Tang Q, Wang B, et al. LLM-assisted reinforcement learning: leveraging lightweight large language model capabilities for efficient task scheduling in multi-cloud environment. IEEE Trans Consum Electron. 2025;71(2):5631–44. doi:10.1109/TCE.2024.3524612.
56. Thapa S, Shiwakoti S, Shah SB, Adhikari S, Veeramani H, Nasim M, et al. Large language models (LLM) in computational social science: prospects, current state, and challenges. Soc Netw Anal Min. 2025;15(1):4. doi:10.1007/s13278-025-01428-9.
57. Wang T, Zhang B, Jiang D, Li D. A multimodal large language model framework for intelligent perception and decision-making in smart manufacturing. Sensors. 2025;25(10):3072. doi:10.3390/s25103072.
58. Mo S, Salakhutdinov R, Morency LP, Liang PP. IoT-LM: large multisensory language models for the Internet of Things. arXiv:2407.09801. 2024.
59. Jiang S, Chen Z, Liang J, Zhao Y, Liu M, Qin B. Infrared-LLaVA: enhancing mmmmmunderstanding of infrared images in multi-modal large language models. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12–16; Miami, FL, USA. p. 8573–91. doi:10.18653/v1/2024.findings-emnlp.501.
60. Fu T, Liu S, Li P. Intelligent smelting process, management system: efficient and intelligent management strategy by incorporating large language model. Front Eng Manag. 2024;11(3):396–412. doi:10.1007/s42524-024-4013-y.

61. Gao D, Ma Y, Liu S, Song M, Jin L, Jiang W, et al. FashionGPT: LLM instruction fine-tuning with multiple LoRA-adapter fusion. Knowl Based Syst. 2024;299:112043. doi:10.1016/j.knosys.2024.112043.
62. Goyal A, Friesen AL, Banino A, Weber T, Rosemary Ke N, Badia AP, et al. Retrieval-augmented reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning (ICML 2022); 2022 Jul 17–23; Baltimore, MD, USA. p. 7740–65.
63. Hu J, Wang H, Tang HK, Kanazawa T, Gupta C, Farahat A. Knowledge-enhanced reinforcement learning for multi-machine integrated production and maintenance scheduling. Comput Ind Eng. 2023;185(1):109631. doi:10.1016/j.cie.2023.109631.
64. Gao Y, Xiong Y, Zhong Y, Bi Y, Xue M, Wang H. Synergizing RAG and reasoning: a systematic review. arXiv:2504.15909. 2025.
65. Yu S, Ran N, Liu J. Large-language models: the game-changers for materials science research. Artif Intell Chem. 2024;2(2):100076. doi:10.1016/j.aichem.2024.100076.
66. Kang JW, Liu BL, Jing T, Shen HF. Intelligent casting: empowering the future foundry industry. China Foundry. 2024;21(5):409–26. doi:10.1007/s41230-024-4056-z.
67. Salonitis K, Jolly MR, Zeng B, Mehrabi H. Improvements in energy consumption and environmental impact by novel single shot melting process for casting. J Clean Prod. 2016;137(4):1532–42. doi:10.1016/j.jclepro.2016.06.165.
68. Galitsky B, Rybalov A. Neuro-symbolic verification for preventing LLM hallucinations in process control. Processes. 2026;14(2):322. doi:10.3390/pr14020322.