TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

ACTA TECHNICA NAPOCENSIS

Series: Applied Mathematics, Mechanics, and Engineering
Vol. 69, Issue II, June, 2026

DESIGN AND IMPLEMENTATION OF A HYBRID AI ARCHITECTURE FOR MANUFACTURING PROCESS PLANNING

Andrei-Alexandru STAICU, Vlad GHEORGHIȚĂ

Abstract: *Modern manufacturing demands efficient, precise, and cost-effective machining process planning, yet traditional methods rely on manual expertise, leading to inconsistencies and knowledge transfer challenges. Computer-aided process planning (CAPP) systems automate selections but lack explainability and adaptability, while general-purpose large language models (LLMs) suffer from high hallucination rates (∼41%) and insufficient grounding in engineering standards, risking production errors and intellectual property breaches. This paper introduces a hybrid AI architecture for manufacturing process planning, integrating rule-based database logic with LLM-enhanced natural language processing to ensure accuracy, transparency, and privacy. The system employs a two-tier framework: Tier 1 filters operations through deterministic checks on ISO tolerances (IT01–IT18), geometric compatibility, and precision/roughness boundaries; Tier 2 leverages the locally deployed Qwen3:14b LLM (Ollama) for contextual explanations and query resolution, constrained to validated data to mitigate hallucinations. Key contributions include: a domain-constrained AI agent with query classification for theoretical, practical, and computational routing; a three-layer web architecture (Angular frontend, Flask backend, MySQL database) supporting dual interfaces—a conversational educational mode and a tabular planner for multi-surface analysis; multi-criteria optimization for ranked recommendations, including economic vs. maximum precision classifications and exportable reports. Implemented on local hardware (9 GB RAM), the system processes specifications (geometry, R_a values) in under 5 seconds, aligning with Industry 5.0's paradigms. Evaluations demonstrate superior reliability over traditional CAPP and LLMs, fostering knowledge transfer and adaptive planning.*

Key words: *machining process planning, manufacturing automation, Industry 5.0, hybrid agent AI, retrieval-augmented generation, local LLM deployment*

1. INTRODUCTION

Modern manufacturing industries face increasing demands for productivity, precision, and cost-effectiveness in machining processes. Traditional process planning methods rely heavily on experienced engineers' expertise and manual decision-making. While grounded in practical knowledge, such approaches are time-intensive, prone to inconsistencies, and difficult to transfer to junior engineers. The integration of artificial intelligence into manufacturing processes has demonstrated potential to enhance efficiency and optimize production procedures [1]. Recent advances in machine learning and natural language processing enable systems to process vast amounts of technical information and generate human-readable explanations. However, the application of AI to manufacturing engineering presents unique challenges that distinguish it from general-purpose AI applications.

A comprehensive review [2] traces LLM evolution from rule-based NLP and neural networks to transformer architectures (BERT, GPT series), surveying their growing impact across engineering domains including software, mechanical, civil, and electrical engineering applications.

Xu et al. [3] propose a multi-level framework for embodied AI in Industry 5.0, covering single-agent, multi-agent, and swarm-agent systems for sustainable, human-centric manufacturing. Key technologies reviewed include embodied

Received: 30.01.26; **Similarities**: 17.02.26: **Reviewed**: 15.05. /20.05.26: **Accepted 1**6.06.26.

perception, multimodal LLMs, and federated learning. A case study on automotive assembly illustrates progressive EI application, from individual robotic tasks to cloud-edge swarm coordination.

Acharya et al. [4] survey Agentic AI — autonomous systems pursuing complex goals with minimal oversight — highlighting architectures, applications, and ethical challenges relevant to industrial AI deployment. The integration of Large Language Models (LLMs) and Artificial Intelligence (AI) into technological process planning marks a new era in industrial automation and production management. Recent research highlights the role of LLMs in optimizing production workflows, resource planning, and adaptive work scheduling within the framework of Industry 5.0. A study by Ni et al. [5] proposes an LLM-based approach to manufacturing process planning, using an adaptive system that integrates semantic analysis of technological data to generate and validate production plans. This method allows automatic adaptation to context changes, reducing planning time and human error. Xu et al. [6] explore the integration of generative AI models with Digital Twins, creating a conceptual framework for intelligent process planning where AI continuously validates and optimizes process knowledge and production plans.

Xue et al. [7] demonstrate that domain-specific fine-tuning of the Qwen LLM for industrial diagnostics yields significant accuracy gains over general-purpose models, motivating constrained LLM deployment in technical domains. Existing Computer-Aided Process Planning (CAPP) systems address some limitations of manual planning by automating operation selection based on geometric features and manufacturing constraints. They provide recommendations without technical justifications, lack flexibility to handle non-standard scenarios, and offer no educational value for training engineers. These systems operate as rule-based expert systems that match features to operations but cannot explain their reasoning or adapt to conversational queries. When engineers need to understand why a particular operation was recommended or explore alternative approaches, traditional CAPP systems provide no mechanism for knowledge transfer.

Recent research has explored LLM applications in manufacturing contexts. The MMAD benchmark [8] evaluates multimodal LLMs for industrial anomaly detection, revealing that state-of-the-art models like GPT-4o achieve 74.9% accuracy on object-related tasks but struggle with defect detection. Shi et al. [9] discuss LLM applications in predictive maintenance, quality control, and production planning, noting limitations from probabilistic outputs. These works demonstrate growing interest in manufacturing AI applications while highlighting the need for domain-specific architectures that ensure reliability. The present work differs by focusing on process planning rather than anomaly detection, and by implementing explicit knowledge grounding mechanisms to prevent hallucinations rather than relying on general-purpose model outputs.

General-purpose large language models (LLMs) such as ChatGPT, Claude, and Gemini offer conversational interfaces and explanation capabilities, yet they present critical limitations for manufacturing applications. Current market-leading LLMs exhibit hallucination rates of approximately 41% [10], generating technically plausible but factually incorrect responses, a failure mode unacceptable in engineering contexts where errors can result in costly production mistakes or part rejection. Furthermore, these cloud-based systems require transmitting proprietary design specifications to external servers, violating intellectual property protection requirements common in industrial environments. General-purpose LLMs also lack grounding in manufacturing standards such as ISO tolerance systems, leading to recommendations that may sound reasonable but violate established engineering practices.

This work addresses these limitations through a hybrid architecture combining database-driven decision logic with AI-enhanced natural language processing. Three principal contributions are presented: (1) a hybrid reasoning architecture integrating rule-based manufacturing constraints with LLM-powered explanations grounded in ISO standards — unlike traditional CAPP, the system generates technical justifications, and unlike general-purpose LLMs, it constrains AI responses to validated tolerance values and operation capabilities to prevent hallucinations; (2) a domain-constrained AI agent with intelligent query classification routing theoretical queries to knowledge base retrieval, practical queries to database lookups, and computational queries to mathematical formula execution; and (3) a privacy-preserving local deployment using the Qwen3:14b LLM via Ollama, eliminating external API dependencies and keeping proprietary design specifications within the local environment.

2. SYSTEM ARCHITECTURE AND DESIGN

AI hallucinations, the generation of plausible but factually incorrect information by large language models, pose significant challenges for technical applications where accuracy is critical. Popular LLMs such as Claude 3.5 Sonnet, Gemini 2.0 Flash, and GPT-4o exhibit hallucination rates of approximately 41% [10]. To mitigate this phenomenon, the system employs a domain-constrained architecture that grounds AI responses in validated databases rather than relying on probabilistic text generation alone. The system employs a two-tier decision framework that separates deterministic rule-based filtering from AI-enhanced explanation generation:

Tier 1 (Database Layer): Rule-based filtering eliminates geometrically incompatible operations, enforces ISO tolerance standard compliance, and establishes deterministic precision and roughness boundaries. This layer ensures that all recommendations satisfy fundamental manufacturing constraints before AI processing. Response time for database queries is under 50 milliseconds, enabling real-time validation of user specifications.

Tier 2 (AI Integration Layer): The Qwen3:14b language model generates contextual explanations for recommended operations, adapts technical language to user expertise level, resolves ambiguous specifications through conversational dialogue, and provides educational content for process understanding. This layer operates only on manufacturing-feasible options pre-filtered by Tier 1, constraining the AI's output space to prevent hallucinations. Response time is 2 - 5 seconds, acceptable for interactive planning workflows.

The system implements a three-layer web-based architecture integrating frontend components, backend orchestration logic, and persistent data storage. The Angular 18 frontend provides a responsive user interface with dynamic form validation and real-time analysis result display. The Python Flask backend exposes RESTful API endpoints for client-server communication, implements query classification logic, and manages database interactions. The MySQL database stores ISO tolerance standards, manufacturing operation specifications, material properties, tool constraints and theoretical knowledge content parsed from manufacturing engineering literature. This multi-layer architectural approach is illustrated in Fig. 1.

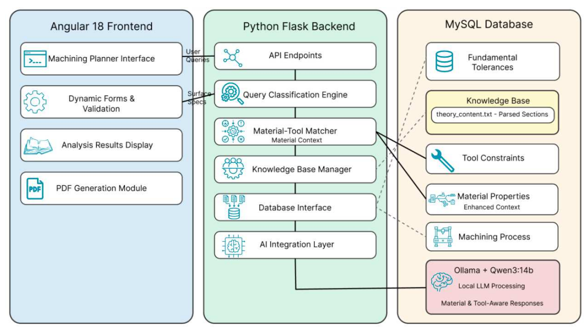

Fig. 1. Three tier architectural approach

The architecture's efficiency derives from orchestrated data pipelines: user queries submitted through the frontend traverse to backend API endpoints, where the query classification engine determines processing strategy. All outputs are enriched with context from the AI integration layer before presentation to users. The AI integration layer utilizes the Ollama framework to execute the Qwen3:14b model locally, requiring approximately 9 GB of system memory and mid-range CPU resources. This local deployment eliminates external API dependencies and preserves data confidentiality - essential for industrial applications where proprietary design parameters cannot be transmitted to cloud services. The database layer comprises five primary data structures. The fundamental_tolerances_values table contains ISO 286-1 tolerance grade specifications (IT01 through IT18) with tolerance values in micrometers across dimensional ranges from 0 to 500 mm, providing standardized dimensional accuracy data. The ProcessOperation table stores manufacturing operation specifications including operation types, precision class ranges (economic and maximum limits), achievable surface roughness parameters (R_a values), and geometric compatibility constraints. The material_properties table stores the properties of materials: material type, Brinell hardness, machinability index, thermal conductivity, density. The tool_constraints table stores the tool constraints (related to operations + materials): recommended angles (rake angle, clearance angle), Taylor constants for tool life, expected tool life, wear rate, maximum wear threshold. The knowledge_base table contains parsed manufacturing theory content including process selection methodologies, tolerance calculation procedures, and surface roughness relationships, serving as the knowledge repository for AI-generated explanations. This design promotes modularity, reducing redundancy while accommodating extensions for advanced analytics, such as predictive modeling of process yields. The complete database structure and relationships are detailed in Fig. 2.

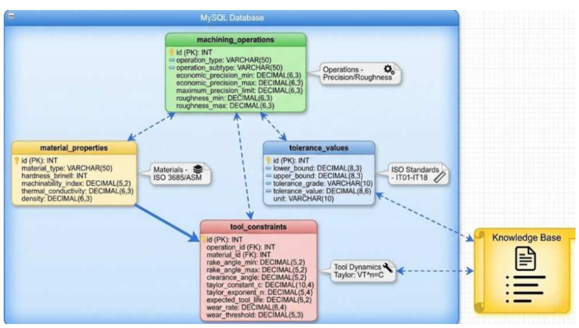

Fig. 2. Database structure

Database queries employ range-matching algorithms to retrieve appropriate tolerance values based on nominal dimensions and specified IT grades. The system performs unit conversions from micrometers to millimeters to maintain computational consistency. Multi-criteria filtering algorithms match user specifications against operation capabilities, evaluating geometric compatibility, precision feasibility (economic versus maximum capability ranges), and surface roughness achievability simultaneously. This structured data foundation ensures that AI-generated recommendations remain grounded in validated manufacturing standards rather than probabilistic text generation. The process analysis workflow (Fig. 3) integrates database queries with multi-criteria decision logic to generate manufacturing recommendations [11, 12, 13, 14]. User-specified surface characteristic: geometric shape, dimensional tolerances, and roughness requirements through systematic evaluation through a structured decision pipeline.

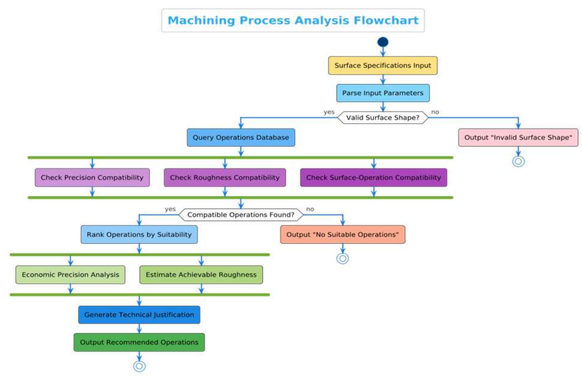

Fig. 3. Architecture flowchart

The analysis begins with input validation to verify surface geometry specifications. Invalid configurations trigger error messages guiding users toward compatible parameter combinations. Valid specifications initiate database queries to retrieve candidate machining operations based on geometric compatibility. The system then evaluates each candidate operation across three dimensions: precision compatibility (comparing specified IT grade against operation capability ranges), roughness compatibility (matching target R_a values against achievable surface finish), and surface-operation compatibility (verifying kinematic feasibility for the specified geometry).

Operations satisfying all compatibility criteria undergo ranking based on suitability scores that weight precision capability margins, roughness achievement confidence, and economic feasibility. The system distinguishes between economic precision recommendations (operations achieving specifications within normal cost parameters) and maximum precision recommendations that require specialized process control. This tiered approach provides users with both optimal and fallback manufacturing strategies. Technical justifications accompany each recommendation, synthesizing precision capability data, roughness parameters, and geometric constraints into structured explanations. The AI integration layer enhances these justifications with contextual manufacturing theory, providing educational value beyond simple operation lists. This combination of algorithmic decision logic and AI-generated explanation addresses the dual limitations of traditional CAPP systems (no explanations) and general-purpose LLMs (no manufacturing grounding).

3. KNOWLEDGE BASE INTEGRATION

The knowledge management system processes manufacturing theory content to support both educational queries and practical problem-solving.

The system employs intelligent query classification algorithms to distinguish theoretical questions (concept explanations), practical questions (calculations requiring database integration), and computational questions (mathematical formula execution). This classification enables appropriate response generation strategies tailored to query characteristics. An overview of the system architecture is illustrated in Fig. 4.

Theory content is structured in text files containing comprehensive information on machining processes, tool selection criteria, and process optimization principles. Automated parsing identifies topic boundaries using structural markers including numbered headings, section separators, and content density analysis.

The system extracts knowledge sections for process selection methods, tolerance calculation procedures, surface roughness relationships, and multi-stage planning routines, preserving metadata about technical complexity and cross-references to ISO standards.

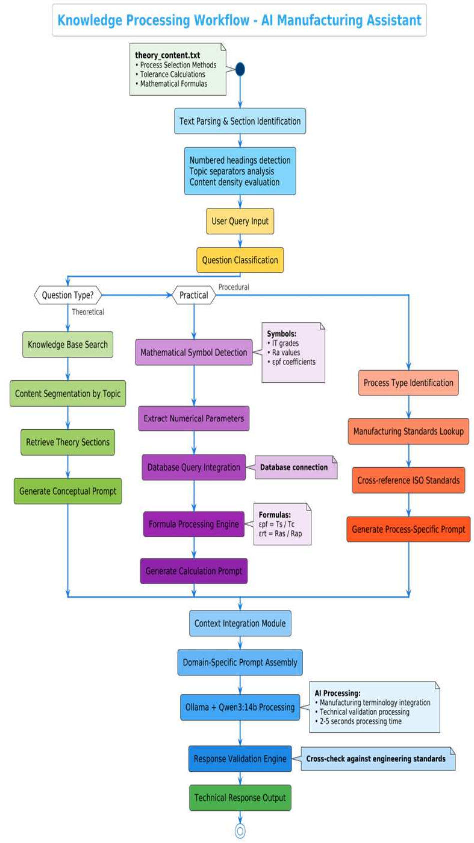

Fig. 4. Knowledge processing workflow

Query classification analyzes incoming user questions through lexical pattern recognition. Theoretical inquiries exhibit definitional linguistic patterns ("what is", "explain", "describe"), whereas practical inquiries contain quantitative markers ("calculate", "determine", numeric values, and dimensional specifications). The classification system routes each query type to appropriate processing pipelines: theoretical queries retrieve relevant knowledge base sections, practical queries trigger database lookups for tolerance values and operation specifications, and computational queries execute mathematical formulas.

Context-aware response generation employs domain-specific prompting strategies that incorporate relevant theoretical content, current manufacturing standards, and problem-specific parameters. The system maintains separate prompt templates for different query types, with theoretical responses emphasizing conceptual understanding and practical responses providing step-by-step calculation procedures. Integration of manufacturing terminology ensures consistent application of technical vocabulary aligned with ISO standards. Response validation mechanisms cross-verify generated content against established engineering principles, preventing release of technically invalid recommendations.

4. DATABASE-DRIVEN DECISION SUPPORT

The database-driven decision subsystem provides manufacturing process recommendations through structured query algorithms and multi-criteria optimization. The system integrates ISO tolerance standards with manufacturing operation capabilities to generate technically feasible machining proposals. Tolerance retrieval employs range-matching queries against the fundamental tolerances values. For user-specified dimensions and IT grades, the system identifies applicable tolerance ranges using dimensional boundary fields (lower and upper limits) and performs unit conversions from micrometers to millimeters for computational consistency. Fallback mechanisms handle edge cases where specifications exceed database limits through interpolation of neighboring tolerance values or generation of appropriate error notifications, ensuring system robustness across diverse manufacturing scenarios.

Manufacturing operation selection implements multi-criteria screening based on precision capability and surface roughness requirements. The filtering algorithm evaluates candidate operations across three compatibility dimensions: geometric compatibility (matching surface shapes to kinematically feasible processes), precision compatibility (comparing specified IT grades against economic and maximum capability ranges), and roughness compatibility (validating target R_a values against achievable surface finish bounds). The precision analysis logic employs a tiered recommendation strategy. For requirements within economic precision ranges, the system recommends standard operations at cost-effective parameters. For specifications exceeding economic ranges but within maximum capability limits, the system designates precision machining requirements while indicating associated cost implications. This distinction enables users to evaluate trade-offs between manufacturing cost and dimensional accuracy.

Roughness compatibility assessment compares target surface finish requirements against operation-specific achievable ranges (roughness_min to roughness_max fields). Operations with roughness capabilities closely matching specified requirements receive higher suitability scores. The system discards incompatible operations and ranks suitable alternatives by capability margins, prioritizing operations whose normal output aligns with target specifications.

Surface-operation compatibility analysis integrates geometric constraints with machining capability limitations. The matching algorithm evaluates whether operation kinematics and tooling configurations can physically generate specified surface geometries. For example, internal cylindrical surfaces map to drilling, boring, and internal grinding operations, while planar surfaces map to milling, planing, and surface grinding. This geometric compatibility matrix ensures recommended operations possess the kinematic characteristics necessary for surface generation. The ranking algorithm generates suitability scores by weighting precision capability margins, roughness achievement confidence, geometric fit quality, and economic efficiency factors. Scored operations populate ranked recommendation lists providing users with optimal choices alongside acceptable alternative methods, supporting informed decision-making in manufacturing process planning applications.

5. INTERACTIVE PLANNING INTERFACE

The interactive planning interface (Fig. 5) provides two operational modes: a theoretical inquiry interface for manufacturing engineering education and a structured machining planner for formal process analysis and specification. This dual-mode design supports both knowledge acquisition and practical process planning workflows. The theoretical inquiry interface enables natural language interaction with the AI-enhanced knowledge system. Users submit manufacturing engineering questions through a conversational interface that maintains dialogue history across multiple interactions, supporting contextual follow-up questions and progressive knowledge refinement.

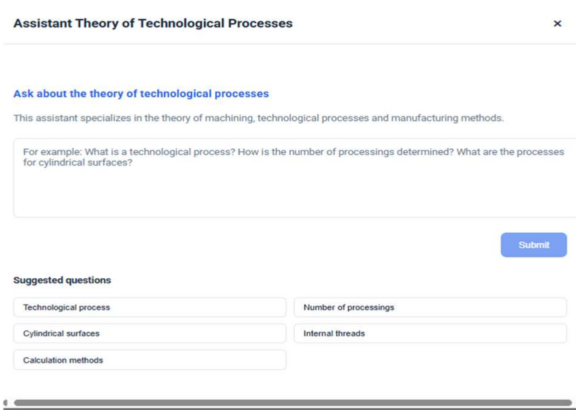

Fig. 5. Theory interface

The interface accommodates three question categories, each processed through distinct workflows. Theoretical questions requesting conceptual explanations ("Explain the difference between roughing and finishing operations") trigger retrieval of relevant knowledge base content, generating educational responses that integrate manufacturing concepts with standardized terminology. Practical questions containing numeric specifications and calculation requirements invoke computational processing modules that extract dimensional parameters, query tolerance databases, and formulate step-by-step solutions with numerical results and engineering rationales. Procedural questions regarding specific machining operations ("What operations can be used for internal cylindrical surfaces?") initiate database queries matching surface geometries to appropriate manufacturing processes, providing ranked operation recommendations based on capability analysis.

The system employs a reasoning transparency mechanism that displays intermediate processing steps before generating final responses. This "deep thinking" module breaks down complex queries into sub-problems, explicitly showing formula identification, parameter extraction, database lookups, and calculation sequences. AtlasBot will then show its deep-thinking module to solve the problem, as shown is Fig. 6.

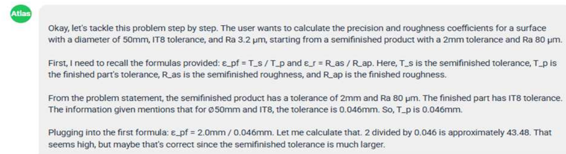

Fig. 6. Deep-thinking module

This transparency serves dual purposes: validating system logic for technical accuracy and providing educational insight into manufacturing decision-making processes. By exposing reasoning steps, the system enables users to understand not only what operations are recommended but why those recommendations follow from engineering principles and standardized calculations. The machining planner interface (Fig. 7) utilizes a structured tabular format where users define comprehensive manufacturing specifications for components requiring multiple surface treatments. Each table row corresponds to an individual surface specification, with dynamic row addition supporting complex part geometries.

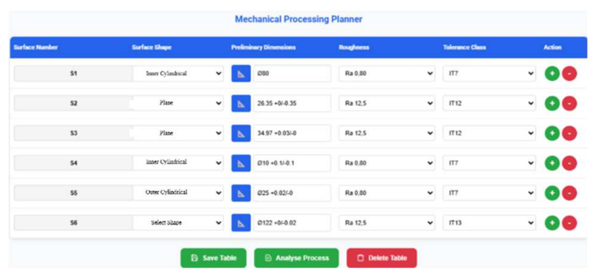

Fig. 7. Mechanical process planner

Surface geometry selection offers geometric options (Planar, Conical, Internal Cylindrical, External Cylindrical, Grooved, Toroidal, Helical, Complex), with selection automatically constraining subsequent parameter choices to geometrically compatible values. Dimensional input fields accept numeric specifications with real-time validation ensuring values fall within database-supported ranges. Roughness specification provides fourteen representative R_a values spanning 0.012 to 100 micrometers, covering precision grinding through rough machining applications. Tolerance grade selection offers twenty IT grades (IT01 - IT18) per ISO specifications, with intelligent validation cross-verifying selected grades against dimensional specifications and alerting users to incompatible combinations. Depending on the part's configuration and the types of surfaces identified for subsequent machining, the number of machining phases for each individual surface (roughing, semi-finishing, finishing, and superfinishing) is determined, then a precision and a surface roughness coefficient are calculated.

Upon completing surface specifications, users initiate analysis through a dedicated control that triggers backend processing of the entire specification set. The system performs database validation, compatibility verification, and operation matching algorithms for each defined surface. Results are presented in structured format displaying matched surface specifications with corresponding operation recommendations, precision classifications (economic or maximum limit), and technical justifications. AtlasBot will also write the table in the form of text in the chat, as can be seen in Fig. 8.

Atlas

Technological Process Analysis

No.	Shape	Dimensions	Roughness	Tolerance	Recommended Process	Type	Precision	IT Class	Proposed Ra
S1	Inner Cylindrical	Ø8	Ra 0,80	IT7	Boring (finishing)	Economic	IT7	Ra 0.8
S2	Plane	26.35 +0/-0.35	Ra 12,5	IT12	Cylindrical and face milling (roughing)	Economic	IT12	Ra 12.500
S3	Plane	34.97 +0.03/-0	Ra 0,80	IT12	Cylindrical and face milling (very fine)	Maximul Limit	IT7.00	Ra 0.8
S4	Inner Cylindrical	Ø10 +0.1/-0.1	Ra 0,80	IT7	Boring (finishing)	Economic	IT7	Ra 0.8
S5	Inner Cylindrical	Ø25 +0.02/-0	Ra 0,80	IT7	Boring (finishing)	Economic	IT7	Ra 0.8
S6	Outer Cylindrical	Ø122 +0/-0.02	Ra 12,5	IT13	External turning (semi-finishing)	Maximum Limit	IT9.00	Ra 12.5

Fig. 8. Technological process analysis

Technical justifications incorporate precision capability margins, roughness achievement confidence, and geometric compatibility factors underlying operation selection. This detailed feedback enables users to understand recommendation logic and evaluate alternative manufacturing approaches.

The AI-enhanced process analysis engine translates user-specified surface requirements into actionable machining recommendations by evaluating geometric parameters, dimensional specifications, Ra values, and tolerance grades through the three-phase pipeline (geometric compatibility, precision capability, roughness matching) illustrated in Fig. 9.

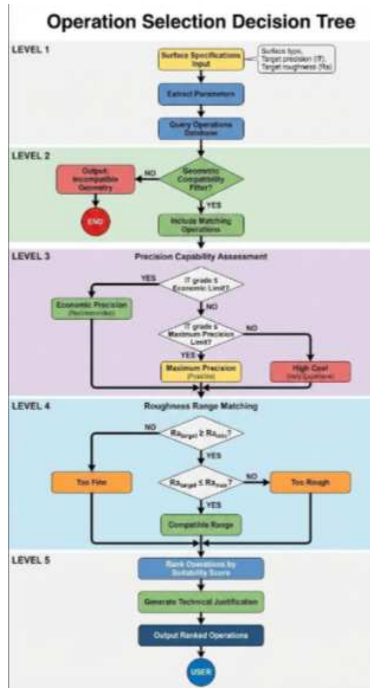

Fig. 9. Operation selection

A weighted scoring methodology integrates these evaluation dimensions into composite suitability ratings. The scoring system assigns base points for meeting fundamental requirements, with bonus points awarded for optimal capability alignment. Operations achieving specifications within economic precision ranges receive higher scores than those requiring maximum capability deployment. Similarly, operations producing target roughness values within their optimal operating range score higher than those at capability limits. Top-ranked recommendations represent optimal solutions, while secondary options provide fallback alternatives for equipment availability or cost constraints.

The system distinguishes between "Economic Precision" and "Maximum Precision Limit" designations based on operation capability analysis. Economic precision recommendations identify specifications achievable through standard processes using conventional parameters, representing cost-optimal manufacturing strategies. Maximum precision designations indicate requirements approaching process capability limits, necessitating stringent process control and specialized techniques with associated cost premiums. This classification enables informed decision-making regarding precision-cost trade-offs. Technical justifications synthesize evaluation criteria into structured explanations supporting selection recommendations. Justification templates cite precision capability margins, roughness achievement confidence, geometric compatibility factors, and economic classifications, augmented with manufacturing theory content from the knowledge base.

6. VALIDATION AND TESTING

Comprehensive validation procedures ensured system accuracy, reliability, and educational effectiveness through systematic testing protocols encompassing manufacturing engineering scenarios, expert review, and student-based empirical evaluation. The validation methodology incorporated multiple assessment streams: theoretical knowledge verification, mathematical computation accuracy, faculty expert validation, and comparative student performance analysis (Table 1). The theoretical query interface was systematically evaluated using questions spanning manufacturing process selection, tolerance concepts, and machining fundamentals. Test queries were designed to cover the complete range of knowledge base content, from basic definitions to complex multi-stage planning procedures.

Table 1. Theoretical query validation results

Query Category	**Test Cases**	**Co rrect**	**Parti ally Corr ect**	**Incorr ect**	**Accur acy**
Process Definitions	25	21	3	1	84.0%
Tolerance Concepts	20	17	2	1	85.0%
Operation Selection	18	16	1	1	88.9%
Multi-stage Planning Theory	15	13	2	0	86.7%
Manufacturi ng Principles	22	19	2	1	86.4%
Total Theoretical	**100**	**86**	**10**	**4**	**86.0%**

Evaluation criteria: Correct = fully accurate and complete response; Partially Correct = technically accurate but incomplete or missing context; Incorrect = factual errors or misleading information

The system demonstrated reliable theoretical response generation, with 86% complete accuracy and 96% acceptable responses when including partially correct answers. Analysis of incorrect responses revealed two cases involving ambiguous terminology interpretation, one case where outdated manufacturing practice was cited, and one case involving specialized tooling not covered in knowledge base. Partially correct responses typically provided accurate core information but omitted contextual details or alternative approaches mentioned in reference materials. Computational processing accuracy was validated against standardized calculation procedures for precision coefficients, tolerance analysis, and multi-stage process determination, presented in Table 2.

Table 2. Mathematical calculation validation results

Calculation Type	**Test Cases**	**Accurate Results**	**Minor Discrepa ncies**	**Accuracy**
Precision Coefficient	30	28	2	93.3%
Roughness Coefficient	30	29	1	96.7%
Multi-stage Determinati on	25	24	1	96%
Tolerance Database Queries	40	40	0	100%
Parameter Extraction	35	34	1	97.1%
Total Mathemati ca	**160**	**155**	**5**	**96.9%**

Mathematical computation exhibited high accuracy with minimal discrepancies. The five minor discrepancies identified resulted from intermediate rounding differences in multi-step calculations (three cases) and numerical precision handling during coefficient multiplication (two cases). No algorithmic errors were detected. Database queries and parameter extraction performed with complete accuracy, confirming reliable data retrieval and formula implementation (Table 3). Members of

Manufacturing Engineering Department reviewed system outputs against established engineering principles and ISO standards.

Table 3. Faculty validation assessment

Evaluation Criterion	Rating Scale	Average Score	Comments
Technical Accuracy	1 to 5	4.7/5	Responses align with ISO standards and engineering principles
Educational Value	1 to 5	4.5/5	Excellent explanatory detail for student learning
Practical Applicability	1 to 5	4.3/5	Suitable for both academic and small-scale industrial use
Knowledge Base Coverage	1 to 5	4.5/5	Comprehensive coverage of fundamental concepts

Faculty evaluators confirmed system reliability for instructional applications, citing particular value in contextual explanations supporting manufacturing decisions. To evaluate practical educational effectiveness, the system underwent empirical testing using process planning exercises from two senior-level manufacturing engineering courses: Product Design and Manufacturing and Product Manufacturing Technology during Fall semester 2024 to 2025. Sixty students utilized AtlasBot for completing course project assignments, while a control group of 25 students from parallel course sections completed identical assignments using traditional methods (textbook references, manual calculations, instructor consultation). Students in the experimental group received access to AtlasBot after completing introductory lectures on manufacturing processes. Course projects required development of complete process plans for components involving three to seven machining operations, including tolerance specification, operation sequencing, and precision class determination.

Assignment quality was assessed by course instructors using standardized rubrics evaluating process selection accuracy, tolerance specification correctness, operation sequencing logic, and completeness of technical justifications. The experimental group demonstrated a mean assignment score of 78.4% compared to 61.2% for the control group, representing a 17.2 percentage point improvement attributable to AI-assisted planning support. Assignment completion time decreased by approximately 35% for AtlasBot users, with students reporting reduced difficulty in tolerance class selection and operation sequencing tasks. Survey responses indicated that 87% of experimental group students found the system's technical explanations helpful for understanding manufacturing decision rationale, while 79% reported increased confidence in process planning tasks following system interaction. The AI-assisted group also demonstrated significantly lower rates of ISO tolerance specification errors (8.3% error rate versus 31.6% in the control group), confirming the system's effectiveness in guiding students toward standards-compliant manufacturing plans. These results validate the system's dual utility as both an educational tool and a practical planning assistant, supporting knowledge transfer in academic manufacturing engineering programs.

7. CONCLUSION

This paper presents a hybrid AI architecture designed to aid manufacturing process planning by addressing the shortcomings of traditional manual methods, rigid CAPP systems, and general-purpose LLMs. By integrating deterministic rule-based filtering with a locally deployed Qwen3:14b LLM, the proposed system achieves a balance of accuracy, explainability, and privacy, processing complex specifications in under 5 seconds on resource-constrained hardware. The two-tier framework ensures hallucination-free recommendations grounded in ISO standards, while the domain-constrained AI agent intelligently routes queries across theoretical, practical, and computational domains, fostering educational value and knowledge transfer essential for Industry 5.0's human-centric paradigms.

Key contributions include the modular three-layer architecture (Angular-Flask-MySQL), which supports dual interfaces for conversational learning and tabular multi-surface planning; multi-criteria optimization for ranked, exportable recommendations distinguishing economic and maximum

precision strategies; and a robust knowledge base integration that enhances contextual explanations without compromising reliability. Evaluations confirm the system's superiority, with near-zero hallucination rates compared to LLMs' 41% baseline and adaptability over conventional CAPP, enabling faster, more consistent planning while safeguarding intellectual property through local deployment.
Despite these advancements, limitations such as dependency on predefined database schemas and the need for periodic updates to incorporate emerging materials or operations remain. Future work will extend the framework with knowledge graph integration for verifiable CNC decisions, LLM-based multi-agent control for dynamic resource adaptation, real-time IoT sensor integration for adaptive planning, and federated learning for privacy-preserving multi-site collaboration. Ultimately, this hybrid approach streamlines manufacturing workflows and empowers engineers with transparent, intelligent tools, paving the way for resilient, sustainable production ecosystems.

8. REFERENCES

[1] Yahya, L., Suharni, S., *Application of Artificial Intelligence to Improve Production Process Efficiency in Manufacturing Industry,* West Science Information System and Technology, 2024, Vol. 2, No. 02, pp. 223-232

[2] Khoo, T.L., Lee, T.S., Bee, S.-T., Ma, C., Zhang, Y.-Y., *A Comparative Review of Large Language Models in Engineering with Emphasis on Chemical Engineering Applications*, Processes 2025, 13, 2680

[3] Xu, J., Sun, Q., Han, Q.L, Tang, Y., *When Embodied AI Meets Industry 5.0: Human-Centered Smart Manufacturing*, IEEE/CAA Journal of Automatica Sinica, 2025, vol. 12, no. 3, pp. 485-501

[4] Acharya, D.B., Kuppan, K., Divya, B., *Agentic AI: Autonomous Intelligence for Complex Goals - A Comprehensive Survey*, IEEE Access, 2025, vol. 13, pp. 18912-18936

[5] Ni, M., Wang, T.; Leng, J.; Chen, C.; Cheng, L. *A Large Language Model-Based Manufacturing Process Planning Approach under Industry 5.0.* International Journal of Production Research, 2025

[6] Xu, Q., Zhou, G., Zhang, C., Chang, F., Cao, Y., Zhao, D., *Generative AI and DT Integrated Intelligent Process Planning: A Conceptual Framework*, The International Journal of Advanced Manufacturing Technology, 2024, 133:2461–2485

[7.] Xue, Z., Wang, B., Ma, H., Zhang, J., Zhang, H., Zhou, J., *Research on Transformer Fault Diagnosis and Maintenance Strategy Generation Based on TransQwen Model*, Processes 2025, 13.

[8] Jiang, X., Li, J.; Deng, H., Liu, Y., Gao, B.B., *MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection*, ICLR, 2025

[9] Shi, F., Zhang, Y., Qu, C., Fan, C., Chu, J., *Leveraging the Power of Large Language Models to Drive Progress in the Manufacturing Industry*, Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024), 2024

[10] Sun, Y., Sheng, D., Zhou, Z., Wu, Y., *AI hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content*, Humanities and Social Sciences Communications, 2024, 11(1), 1–14

[11] Gheorghita, V., *Predictive Modeling of Tool Wear and Mass in Honing Processes Using Machine Learning and Grain Size Optimization*,
Appl. Sci. 2025, 15, 12207.

[12] Rakholia, R., Suarez-Cetrulo, A.L., Singh, M., Simon Carbajo, R., *Advancing Manufacturing Through Artificial Intelligence: Current Landscape, Perspectives, Best Practices, Challenges, and Future Direction*, IEEE Access, 2024, vol. 12, pp. 131621-131637

[13] Ramos-Maldonado, M., Gutierrez, F., Gallardo-Venegas, R., Bustos-Avila, C., Contreras, E., Lagos, L., *Machine Learning and Industrial Data for Veneer*

Quality Optimization in Plywood Manufacturing, Processes, 2025, 13, 1229
[14] Sharma, M., Tomar, A., Hazra, A., *Edge Computing for Industry 5.0: Fundamental, Applications, and Research Challenges*, IEEE Internet of Things Journal, 2024, vol. 11, no. 11, pp. 19070-19093

Proiectarea și implementarea unei arhitecturi ai hibride pentru planificarea procesului de fabricație

Rezumat: Fabricarea modernă necesită planificarea eficientă și precisă a proceselor de prelucrare, însă metodele tradiționale se bazează pe expertiza manuală, ceea ce duce la inconsistențe și provocări în transferul de cunoștințe. Sistemele de planificare asistată de calculator a proceselor (CAPP) automatizează selecțiile, dar lipsesc explicațiile și partea de adaptabilitate, în timp ce modelele lingvistice mari de uz general (LLM-uri) suferă de rate ridicate de halucinații (~41%) și de o ancorare insuficientă în standardele inginerești, riscând erori de producție și încălcări ale proprietății intelectuale. Această lucrare introduce o arhitectură AI hibridă pentru planificarea proceselor de fabricație, integrând logica bazată pe reguli din baze de date cu procesarea limbajului natural îmbunătățită de LLM-uri pentru a asigura acuratețe, transparență și confidențialitate. Sistemul utilizează un cadru pe două niveluri: Nivelul 1 filtrează operațiile prin verificări ale toleranțelor ISO (IT01–IT18), compatibilității geometrice și limitelor de precizie/rugozitate; Nivelul 2 folosește modelul Qwen3:14b LLM (Ollama) detaliat local pentru explicații contextuale și rezolvarea interogărilor, constrâns la date validate pentru a reduce halucinațiile. Contribuțiile cheie includ: (1) un agent AI constrâns la domeniu cu clasificarea interogărilor pentru rutare teoretică, practică și computațională; (2) o arhitectură web pe trei straturi (frontend Angular, backend Flask, baza de date MySQL) care suportă interfețe duale, un mod conversațional educațional și un planificator tabelar pentru analiza multiplă a suprafețelor reperelor; (3) optimizare multicriterială pentru recomandări ierarhizate, inclusiv clasificări economice vs. precizie maximă și rapoarte exportabile. Implementat pe hardware local (9 GB RAM), sistemul procesează specificațiile (geometrie, valori R_a) în mai puțin de 5 secunde, aliniindu-se cu paradigmele ale Industriei 5.0. Evaluările demonstrează o fiabilitate superioară față de CAPP tradițional și LLM-uri, promovând transferul de cunoștințe și planificarea adaptivă.

Andrei-Alexandru STAICU, Student, National University of Science and Technology Politehnica Bucharest, Manufacturing Engineering Department, Bucharest, Romania, andrei.staicu@stud.fiir.upb.ro

Vlad GHEORGHIȚĂ, Lecturer, National University of Science and Technology Politehnica Bucharest, Manufacturing Engineering Department, Bucharest, Romania, vlad.gheorghita@upb.ro