ISSN 2456-0286(O)
RUNGTA
International Journal of Mechanical and Automobile Engineering

RESEARCH ARTICLE

A Generative AI-Based CAPP Framework for Real-Time Adaptive Manufacturing

Sameer Singh[1*], Anant Kumar[1], Kush Kumar Dewangan[1], Sanjay G Sakharwade[1], Rahul Mishra[1], Suraj Bandhekar[1]

[1] *Department of Mechanical Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India – 490024.*

***Corresponding Author E-mail:** sameer.singh@rungta.ac.in*

ABSTRACT:

Smart Manufacturing (SM), a cornerstone of Industry 4.0 (I4.0), enhances autonomy across cyber physical systems on the production floor. Connectivity, powered by Digital Twinning (DT), the Industrial Internet of Things (IIoT), Cloud Computing, and advanced sensors, facilitates real-time monitoring and decision-making. DT plays a pivotal role in production and operations management by integrating with a novel CAPP-GPT (Computer-Aided Process Planning–Generative Pretrained Transformer) and production scheduling framework. This hybrid methodology addresses shop floor disruptions by enabling self-healing capabilities at the micro-CAPP level, dynamically adjusting process parameters and toolpaths based on real-time process signature data. Inspired by advancements in NLP and LLMs like ChatGPT, the framework analyzes CAD models and blueprints using a combination of operations research and predictive analytics for setup planning, operation sequencing, and grouping. A hybrid approach incorporating Optimization, Logical Analysis of Data, and Machine Learning (ML) is employed to heuristically generate and modify process plans. At the macro-CAPP level, integration with delayed product differentiation, lotsizing, and transfer line balancing supports batch production in Hybrid Manufacturing (HM) and Smart Assembly systems.

KEYWORDS: CAPP-OSLM; Open-Source Large Language Models; LoRA Fine-Tuning; Macro-CAPP; Micro-CAPP; Digital Twin; Smart Manufacturing; Maintenance 4.0; Quality 4.0.

Received on 24.12.2024 Revised on 29.01.2025 Accepted on 02.02.2025 Published on 07.02.2025 ***Run. Int. J. Mech. and Auto. Engl. 2025; 2(1), 7.***

1. INTRODUCTION

Smart Manufacturing (SM), a cornerstone of Industry 4.0 (I4.0), extends autonomy across the cyber-physical systems that make up the modern shop floor. Connectivity enabled by Digital Twinning (DT), the Industrial Internet of Things (IIoT), cloud and edge computing, and dense sensor networks now supports continuous monitoring and near-real-time decision-making across production systems [1,2]. Within this landscape, Computer-Aided Process Planning (CAPP) remains the critical bridge between product design (CAD) and manufacturing execution (CAM), translating geometric and technological product data into feasible, efficient sequences of manufacturing operations [3,4].

CAPP has progressed through variant, semi-generative, generative, and expert-system paradigms over more than five decades of development [5,6,7], yet fully automated, adaptive process decision-making has remained elusive, chiefly because rule bases and case libraries do not generalize well to novel part

ISSN 2456-0286(O)
RUNGTA
International Journal of Mechanical and Automobile Engineering

geometries, mixed-model production, or unplanned shop-floor disruptions. The recent emergence of Natural Language Processing (NLP) and Large Language Models (LLMs) such as GPT has opened a new avenue for CAPP: reframing process-plan generation as a sequence-prediction or sequence-generation problem that a pretrained transformer can learn directly from CAD and process data [8]. The CAPP-GPT framework, developed at the Production and Operations Management (POM) research lab, University of Windsor, is the most direct precedent for this paper. It integrates a GPT-based generative model with DT feedback to support both macro-CAPP (setup planning, operation sequencing and grouping) and micro-CAPP (real-time, self-healing adjustment of process parameters and toolpaths in response to process-signature deviations), and extends macro-CAPP to delayed product differentiation, lot-sizing, and transfer line balancing for hybrid manufacturing and smart assembly [9,10,11,12,13].

While CAPP-GPT demonstrates the feasibility of GPT-style generative modelling for process planning, its dependence on a proprietary, typically cloud-hosted GPT backbone carries practical limitations for many manufacturers, and particularly for small and medium-sized enterprises (SMEs) common in manufacturing clusters such as Chhattisgarh's engineering sector. Cloud-hosted proprietary LLMs require transmitting potentially sensitive CAD and process data off-premise, incur recurring per-token or per-call costs that scale with shop-floor query volume, are difficult to customize beyond prompt-level adaptation, and introduce network latency that is poorly suited to the sub-second response times required for micro-CAPP self-healing. Open-source LLMs such as Meta's LLaMA and Mistral AI's Mistral now offer competitive language-generation and reasoning performance while being fully inspectable, deployable on-premise, and, critically, fine-tunable on a firm's own data using parameter-efficient techniques such as Low-Rank Adaptation (LoRA), which updates a small fraction of model parameters and can be trained on a single GPU rather than the large compute clusters required for full model retraining [14,15,16].

This paper proposes CAPP-OSLM (Computer-Aided Process Planning using Open-Source Language Models), a conceptual framework that substitutes a LoRA-fine-tuned open-source LLM (LLaMA or Mistral) for the proprietary GPT backbone used in CAPP-GPT, while preserving the same macro-/micro-CAPP architectural split and its integration with DT and IIoT infrastructure. The objective is not to claim superior predictive accuracy over CAPP-GPT, which has not yet been empirically tested for the open-source variant, but to articulate a technically grounded, deployment-oriented alternative that addresses data-sovereignty, cost, latency, and customizability constraints that are especially relevant to manufacturers unable or unwilling to rely on proprietary cloud APIs. The remainder of the paper reviews the relevant literature on CAPP, generative AI in process planning, and macro-CAPP scheduling problems (Section 2); presents the proposed framework's architecture (Section 3); discusses anticipated benefits, risks, and validation requirements (Section 4); and concludes with directions for future empirical work (Section 5).

2. RELATED WORK

2.1. Evolution of computer-aided process planning

Computer-aided process planning was first conceptualised in the mid-1960s as a means of encoding manufacturing process knowledge into computerised systems capable of assisting or replacing manual route-sheet preparation [17]. Subsequent decades produced comprehensive surveys tracking CAPP's evolution from variant systems, which retrieve and adapt plans for previously manufactured, geometrically similar parts, through semi-generative and fully generative systems capable of synthesising new plans from first principles using feature recognition, geometric reasoning, and manufacturing knowledge bases [18,19]. Despite this progress, classical CAPP systems remain limited by static knowledge representations, brittle rule sets, and high configuration effort when applied to new part families or production contexts [4].

2.2. Generative ai and large language models in process planning

ISSN 2456-0286(O)
RUNGTA
International Journal of Mechanical and Automobile Engineering

The success of transformer-based architectures in NLP [20] and the subsequent development of generative pretrained transformer (GPT) models [8] motivated a wave of research applying LLMs to CAPP and adjacent manufacturing decision problems. The CAPP-GPT framework represents the most comprehensive such effort, combining a hybridized Operations Research (OR) and Machine Learning (ML) approach, including Logical Analysis of Data (LAD) and Group Technology (GT), with a GPT-based generative component to automate feature recognition, setup planning, and operation sequencing for both discrete and hybrid (additive-subtractive) manufacturing [9]. A parallel systematic literature review of automated process planning and dynamic scheduling for smart manufacturing confirms that digital twin and AI integration into CAPP and scheduling remains an active but still-maturing research area, with process planning and scheduling functions typically studied in isolation despite their operational interdependence [10].

More recent studies have explored LLMs trained from scratch on synthetic part-encoding datasets for high-level CAPP, achieving high sequence-generation accuracy for feasible process chains in distributed manufacturing settings [21], and knowledge-graph-augmented retrieval-augmented generation (RAG) approaches that explicitly target the numeric hallucination problem inherent to LLM-based process planning by grounding generation in verifiable, structured process knowledge [22]. LLM agents have also been proposed as autonomous mechanical designers capable of iterative design reasoning [23], and domain-adapted open-source LLMs have been explored for additive manufacturing process guidance [24], indicating a broader research trend toward moving away from generic, proprietary chat-style models and toward domain- and firm-specific model adaptation. Physics- and rule-constrained generative approaches in adjacent domains such as semiconductor fabrication further demonstrate that unconstrained LLM outputs are prone to proposing operations that violate process-window or tooling constraints, reinforcing the need for feasibility-checking layers around any generative CAPP component [25].

2.3. Digital twins and the industrial internet of things

Digital Twin technology, defined as a continuously updated virtual replica of a physical asset or process, is now widely recognised as a core enabling technology for real-time monitoring, simulation, and optimisation in I4.0 manufacturing environments [1,2]. Integration of DT with IIoT sensor networks allows manufacturing systems to detect deviations from planned process signatures as they occur, providing the real-time data stream on which self-healing, micro-level process adaptation depends [2]. Predictive maintenance, often described under the label Maintenance 4.0, extends this same sensing and modelling infrastructure to machine health and remaining useful life estimation, and is closely related to Quality 4.0 efforts that use continuous process data to detect quality deviations before they propagate downstream [26,27].

2.4. Macro-capp: Delayed differentiation, lot-sizing, and line balancing

At the macro-CAPP level, hybrid manufacturing and smart assembly systems increasingly rely on delayed product differentiation (DPD), a strategy in which a common product platform is manufactured and held in an undifferentiated state before being converted into specific product variants only once demand is realised, reducing the cost of variety while preserving responsiveness [28]. Extensions of this concept to assembly-line balancing under asymmetric, mixed-model configurations [29], together with recent hybridized machine-learning and optimization approaches to CAPP-integrated group technology and delayed differentiation for hybrid manufacturing [11], multi-period, multi-platform lot-sizing under stochastic demand [12], multi-level capacitated lot-sizing with setup carryover and emission control [13], fuzzy-demand platform design and variant substitution [30], and constrained-clustering approaches to stock design in hybrid manufacturing [31], together constitute the macro-CAPP research base that the proposed framework builds upon rather than replaces.

3. PROPOSED CAPP-OSLM FRAMEWORK

ISSN 2456-0286(O)
RUNGTA
International Journal of Mechanical and Automobile Engineering

3.1. Overall architecture

CAPP-OSLM retains the two-tier macro-/micro-CAPP structure established in prior generative-AI CAPP research [9] but replaces the proprietary GPT-family generative component with a compact, open-source LLM (LLaMA-family or Mistral-family, typically in the 7-8 billion parameter range) that is fine-tuned on the manufacturer's own CAD feature libraries, historical process plans, tooling catalogues, and machine capability data. The architecture comprises four coupled layers: (i) a CAD/feature-ingestion layer that extracts manufacturing features, geometric dimensioning and tolerancing (GD&T) data, and part-family information from CAD models; (ii) a fine-tuned open-source LLM core that generates candidate operation sequences, setup plans, and toolpath adjustments conditioned on the extracted features and on real-time process-signature data; (iii) a feasibility-checking and knowledge-grounding layer, informed by knowledge-graph and retrieval-augmented approaches from the wider literature [22], that filters LLM outputs against hard manufacturing constraints (tool availability, process-window limits, precedence relations) before they are released to the shop floor; and (iv) a DT/IIoT integration layer that closes the loop between planned and actual process execution.

3.2. Macro-capp module

The macro-CAPP module performs setup planning, operation sequencing, and operation grouping using a hybrid Operations Research-Machine Learning approach consistent with the Group Technology and Logical Analysis of Data methods established in prior CAPP-GPT research [9], with the fine-tuned open-source LLM used to propose candidate sequences and groupings that are then evaluated against a mathematical programming formulation of production cost, in line with the two-phase hybrid optimization-ML approach demonstrated for hybrid manufacturing process planning under delayed differentiation [11]. For batch and mixed-model production, the macro-CAPP module further incorporates delayed product differentiation logic [28,29], multi-period lot-sizing under demand uncertainty [12,30], and transfer line/assembly line balancing for hybrid manufacturing and smart assembly systems [13,31], allowing the framework to support both discrete part manufacturing and higher-variety batch production within a single planning layer.

3.3. Micro-capp module and self-healing adaptation

The micro-CAPP module addresses shop-floor disruptions by continuously comparing real-time process-signature data, acquired via IIoT sensors and reflected in the DT, against the planned process trajectory. When a deviation exceeding a defined tolerance is detected, the fine-tuned LLM is queried with the current process state and a constrained set of feasible corrective actions (parameter adjustment, toolpath modification, or operation resequencing), and its proposed correction is passed through the feasibility-checking layer described in Section 3.1 before being applied. This design follows the self-healing concept introduced in CAPP-GPT [9] while explicitly addressing the numeric-hallucination risk documented for unconstrained LLM-based process planning [22,25] by never allowing raw LLM output to reach the machine controller without a feasibility check.

3.4. Fine-tuning methodology: Lora on open-source llm backbones

The proposed fine-tuning approach adapts a pretrained open-source LLM, LLaMA or Mistral, to the CAPP domain using Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning technique that freezes the pretrained model weights and injects small, trainable low-rank matrices into the attention and feed-forward projection layers [16]. Because LoRA typically trains under one percent of total model parameters, it substantially reduces the computational and data requirements relative to full fine-tuning while retaining most of the performance benefit, and has been demonstrated on both LLaMA and Mistral backbones in domain-adaptation settings outside manufacturing [15,16]. In the CAPP-OSLM setting, the fine-tuning corpus would combine: (a) historical process plans and route sheets, reformatted as feature-to-operation-sequence instruction pairs; (b) tool and machine capability tables; and (c) synthetically generated part-encoding and process-chain pairs of the kind used to train LLM-based CAPP systems from scratch [21], supplementing limited proprietary historical

RUNGTA
International Journal of Mechanical and Automobile Engineering

data. Quantized versions of the fine-tuned model (e.g., 4-bit weight quantization) are proposed for on-premise or edge deployment, consistent with recent demonstrations that smaller, locally hosted fine-tuned or retrieval-augmented models can match larger cloud-hosted models on domain-specific manufacturing tasks while running fully on-premise for privacy-preserving, real-time inference [22].

3.5. Integration with quality 4.0 and maintenance 4.0

Because the DT/IIoT layer that feeds the micro-CAPP module already captures continuous process and machine-health signals, the same data stream is proposed to feed parallel Quality 4.0 and Maintenance 4.0 functions: statistical or ML-based quality-deviation detection drawing on the same process signatures used for self-healing, and predictive maintenance models estimating remaining useful life and scheduling condition-based interventions [26,27]. This shared-data design avoids duplicating sensing infrastructure across planning, quality, and maintenance functions, a fragmentation noted as a persistent limitation of I4.0 technology adoption in existing literature [27].

4. DISCUSSION

The central proposition of this paper is that the architectural and algorithmic contributions of GPT-based CAPP research [9] can be decoupled from dependence on a specific proprietary model family. Table 1 summarises the anticipated qualitative differences between a proprietary GPT-based CAPP backbone and the proposed open-source, LoRA-fine-tuned alternative; these are framed as design trade-offs anticipated from the underlying technical literature on LoRA fine-tuning, open-weight model deployment, and RAG-based hallucination mitigation, rather than as measured experimental results, since CAPP-OSLM has not yet been empirically validated.

Table 1. Anticipated qualitative comparison between proprietary GPT-based and open-source LoRA-fine-tuned CAPP backbones.

Dimension	Proprietary GPT-based CAPP (CAPP-GPT)	Proposed Open-Source LoRA-Fine-Tuned CAPP-OSLM
Model hosting	Cloud-hosted, vendor-controlled	On-premise / edge-deployable
Data exposure	CAD/process data sent to third-party API	Data remains within firm's infrastructure
Customization	Prompt-level adaptation only	Weight-level domain adaptation via LoRA
Marginal cost at scale	Recurring per-token/API cost	One-time fine-tuning + local compute cost
Latency for micro-CAPP	Network round-trip dependent	Local inference, lower expected latency
Reasoning capability (as published)	Larger proprietary model class	Smaller open-weight backbone; needs grounding layer
Validation status	Reported in original CAPP-GPT study	Conceptual; empirical validation pending

Several risks and open questions remain. First, open-source LLMs in the 7-8 billion parameter range generally underperform the largest proprietary models on open-ended reasoning benchmarks, so the feasibility-checking and knowledge-grounding layer proposed in Section 3.1 is not optional but a necessary safeguard against infeasible or unsafe process recommendations, consistent with findings that LLMs applied to engineering process planning routinely hallucinate numeric parameters without such grounding [22,25]. Second, the quality of LoRA fine-tuning is bounded by the size and quality of the manufacturer's own historical process-plan corpus; firms with limited digitised process history may need to rely more heavily on synthetic data generation of the kind demonstrated for GPT-2-based CAPP sequence models [21], or on transfer learning from a shared, multi-firm base adapter. Third, the macro-

RUNGTA
International Journal of Mechanical and Automobile Engineering

CAPP extensions to delayed differentiation, lot-sizing, and line balancing proposed here are drawn from an active and still-evolving optimization literature [11,12,13,28,29,30,31]; integrating LLM-proposed sequences with these mathematical programming formulations without degrading their optimality guarantees will require careful interface design between the generative and optimization components, mirroring the two-phase hybrid approach already demonstrated for CAPP-integrated delayed differentiation [11]. Finally, because this paper presents a conceptual framework, empirical validation, comparing CAPP-OSLM against both CAPP-GPT and conventional rule-based CAPP on a shared benchmark of part families and disruption scenarios, is required before any performance claims can be made, and is identified as the immediate next step for this research programme.

5. CONCLUSION

This paper has proposed CAPP-OSLM, a conceptual framework for real-time adaptive Computer-Aided Process Planning that adapts the macro-/micro-CAPP architecture and digital-twin integration established in the CAPP-GPT literature [9] to an open-source, LoRA-fine-tuned LLM backbone (LLaMA or Mistral) in place of a proprietary GPT model. The proposed framework is intended to address data-sovereignty, cost, latency, and customizability constraints associated with cloud-hosted proprietary LLMs, while preserving compatibility with the broader macro-CAPP research base on delayed product differentiation, lot-sizing, and transfer line balancing for hybrid manufacturing and smart assembly. The framework explicitly incorporates a feasibility-checking and knowledge-grounding layer to mitigate the numeric-hallucination risks documented for unconstrained generative process planning, and proposes shared use of digital-twin and IIoT data streams across process planning, Quality 4.0, and Maintenance 4.0 functions. As a conceptual contribution, this work does not yet report experimental validation; the identified next step is implementation and benchmarking of the fine-tuned open-source model against both CAPP-GPT and conventional CAPP baselines on a common set of part families and shop-floor disruption scenarios.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Mechanical Engineering, Rungta College of Engineering and Technology (RCET), Bhilai, for institutional support. No external funding was received for this conceptual study.

REFERENCES

1. Digital Twin applications toward Industry 4.0: A Review. J Ind Inf Integr / Adv Ind Manuf Eng. 2023. https://doi.org/10.1016/j.aime.2023.100114

2. Vital-Soto A, Marzia S, Azab A. Automated process planning and dynamic scheduling for smart manufacturing: A systematic literature review. Manuf Lett. 2023;35:861–872. https://doi.org/10.1016/j.mfglet.2023.07.013

3. Alting L, Zhang H. Computer aided process planning: the state-of-the-art survey. Int J Prod Res. 1989;27(4):553–585.

4. ElMaraghy H. Evolution and future perspectives of CAPP. CIRP Ann Manuf Technol. 1993;42(2):739–751.

5. Ham I, Lu S. Computer-aided process planning: the present and the future. CIRP Ann Manuf Technol. 1988;37(2):591–601.

6. Eversheim W, Schneewind J. Computer-aided process planning — state of the art and future development. Robot Comput Integr Manuf. 1993;10(1):65–70.

7. Kamrani A, Sferro P, Handelman J. Critical issues in design and evaluation of computer aided process planning systems. Int J Prod Res. 1995.

8. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. San Francisco: OpenAI; 2018.

9. Azab A, Osman H, Baki F. CAPP-GPT: A computer-aided process planning-generative pretrained transformer framework for smart manufacturing. Manuf Lett. 2024; 41: 51–62. https://doi.org/10.1016/j.mfglet.2024.09.009

RUNGTA
International Journal of Mechanical and Automobile Engineering

10. Vital-Soto A, Marzia S, Azab A. Automated process planning and dynamic scheduling for smart manufacturing: A systematic literature review. Manuf Lett. 2023;35:861–872.

11. Osman H, Azab A, Hasan RB, Baki MF. Computer-aided process planning group technology and delayed product differentiation for hybrid manufacturing using a hybridized machine learning and optimization approach. Ann Oper Res. 2025. https://doi.org/10.1007/s10479-025-06884-2

12. Sakib MD, Osman H, Azab A, Baki MF. Product-platform design and multi-period, multi-platform lot-sizing for hybrid manufacturing considering stochastic demand and processing time. Manuf Lett. 2023;35:20–27. https://doi.org/10.1016/j.mfglet.2023.07.011

13. Hasan RB, Osman H, Azab A, Baki MF. Improvement to an existing multi-level capacitated lot sizing problem considering setup carryover, backlogging, and emission control. Manuf Lett. 2023;35:28–39.

14. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv:2302.13971 [Preprint]. 2023.

15. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, et al. Mistral 7B. arXiv:2310.06825 [Preprint]. 2023.

16. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR); 2022. arXiv:2106.09685.

17. Niebel BW. Mechanized process selection for planning new designs. ASTME Paper No. 737. 1965.

18. Eversheim W, Schneewind J. Computer-aided process planning — state of the art and future development. Robot Comput Integr Manuf. 1993;10(1):65–70.

19. Kamrani A, Sferro P, Handelman J. Critical issues in design and evaluation of computer aided process planning systems. Int J Prod Res. 1995.

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.

21. Stathatos E, Benardos P, Vosniakos GC, Gross D, Spieker H, Gotlieb A. Large language models for high-level computer-aided process planning in a distributed manufacturing paradigm. Robot Comput Integr Manuf. 2026 (in press).

22. Hoang D, Gorsich D, Castanier MP, Imani F. Knowledge graph fusion with large language models for accurate, explainable manufacturing process planning. Int J Prod Econ. 2026 (in press); preprint: arXiv:2506.13026.

23. Jadhav Y, Barati Farimani A. Large language model agent as a mechanical designer. J Eng Des. 2026:1–37. https://doi.org/10.1080/09544828.2026.2624356

24. Domain adapted large language models for additive manufacturing. arXiv:2603.22017 [Preprint]. 2026.

25. Physics-informed generative AI for semiconductor manufacturing: Enforcing hard physical constraints in generative models by construction. arXiv:2606.11247 [Preprint]. 2026.

26. Achouch M, Dimitrova M, Ziane K, Sattarpanah Karganroudi S, Dhouib R, Ibrahim H, Adda M. On predictive maintenance in Industry 4.0: overview, models, and challenges. Appl Sci. 2022;12(16):8081. https://doi.org/10.3390/app12168081

27. Predictive maintenance in the Industry 4.0: a systematic literature review. Comput Ind Eng. 2020.

28. Lee HL, Tang CS. Modelling the costs and benefits of delayed product differentiation. Manage Sci. 1997;43(1):40–53. https://doi.org/10.1287/mnsc.43.1.40

29. AlGeddawy T, ElMaraghy H. Design of single assembly line for the delayed differentiation of product variants. Flex Serv Manuf J. https://doi.org/10.1007/s10696-011-9074-7

30. Alrahi A, Osman H, Azab A, Hasan RB, Baki MF. Integrated product-platform design and multi-period lot-sizing for hybrid manufacturing with fuzzy demand and variant substitution. Manuf Lett. 2025;44:243–252.