J Adv Manuf Eng, Vol. 7, Issue. 1, pp. 31–43, June, 2026

Journal of Advances in Manufacturing Engineering
Web page info: https://jame.yildiz.edu.tr
DOI: 10.14744/ytu.jame.2026.00004

Original Article

Comprehensive analysis of large language model capabilities in face milling operations with virtual twin verification

Uğur ENIŞ, Mehmet Şamil SOYER, Hanife ÜNAL HELVACIOĞLU, Muhammet Mustafa SAVAŞÇI

Research and Development, Siemens A.Ş., İstanbul, Türkiye

ARTICLE INFO

Article history
Received: 12 February 2026
Revised: 20 April 2026
Accepted: 04 May 2026

Key words:
Machining, face milling, artificial intelligence, large language models, generative AI.

ABSTRACT

This paper presents a comprehensive analysis of large language models (LLMs) capabilities in the machine tool domain, specifically focusing on face milling operations. The research evaluates how different prompt techniques (zero-shot, few-shot, and tree-of-thought) affect LLMs' ability to perform tasks traditionally requiring human domain expertise, such as interpreting G-code, recommending appropriate cutting tools, and calculating machining parameters. Performance is evaluated by comparing LLM outputs to industry-standard CAM software and digital twin simulations to verify practical applicability. The findings indicate that current LLM technology shows promise for transforming and optimizing complex engineering tasks in manufacturing but still requires additional operator input and customized approaches to achieve complete operational accuracy. This work contributes to understanding how generative Artificial Intelligence (AI) can be leveraged to optimize, generalize, and standardize machining procedures in industrial applications.

Cite this article as: Eniş, U., Soyer, M. Ş., Ünal Helvacıoğlu, H., & Savaşçı, M. M. (2026). Comprehensive analysis of large language model capabilities in face milling operations with virtual twin verification. *J Adv Manuf Eng*, 7(1), 31–43.

INTRODUCTION

A Large Language Model is a type of artificial intelligence (AI), that is trained with massive amounts of text and purposed to generate humane responses via various natural language processing (NLP) techniques such as, text generation, summarization, and translation. Recent advancements in LLMs have transformed the approach of knowledge share and gathering significantly. Generative AI is also a key value for the machining industry to optimize, generalize and standardize procedures of machining.

In manufacturing, the process of removing unwanted segments of metal workpiece in the form of chips is known as machining. Machining is one of the five groups of manufacturing processes which includes casting, forming, powder metallurgy and joining [1]. The machining process will shape the workpiece as desired, and it is usually done using machine and cutting tools. The machining cutting process can be divided into two major groups which are (i) cutting process with traditional machining (e.g., turning, milling, boring and grinding) and (ii) cutting process with modern machining (e.g., electrical discharge machining (EDM) and abrasive waterjet (AWJ) [2]. Machining processes play a pivotal role in the manufacture of final components, whose outcome, depending on the machining conditions and strategies, can significantly influence the material's functional performances [3].

In recent years, artificial intelligence and machine learning techniques have been increasingly adopted to address various challenges in machining. Soori et al. [4] provided a comprehensive review of machine learning and AI applications in CNC machine tools, covering areas such as

*Corresponding author.
*E-mail address: ugur.enis@siemens.com

process optimization, predictive maintenance, and adaptive control. Similarly, Pimenov et al. [5] critically reviewed AI systems developed specifically for tool condition monitoring, highlighting the potential of data-driven approaches to reduce tool failures and improve machining efficiency. These studies illustrate that AI-based methods have become well-established for specific machining sub-tasks, yet their reliance on structured numerical data and predefined feature engineering limits their generalizability across diverse operational scenarios.

A closely related challenge is the automated selection of cutting tools, which traditionally depends on expert knowledge and catalog-based reasoning. Navaneethan et al. [6] reviewed the state of automated cutting tool selection methods, encompassing rule-based systems, case-based reasoning, and machine learning approaches. Their analysis revealed that despite notable progress, existing methods still struggle to incorporate the breadth of contextual factors — such as workpiece geometry, material properties, and machine constraints — that a human expert naturally considers. This limitation points toward the need for AI systems capable of natural language understanding and multi-factor reasoning.

Traditional machine tool operations involve examining the G-Code, obtaining information about the material to be machined, selecting the appropriate tool for the operation based on this information, and adjusting the federate and spindle speed in accordance with the selected tool material and workpiece material to be machined [7].

These operations represent a set of cognitive tasks that require high knowledge of the discipline in line with the experience of the domain expert. After the transfer of specialized knowledge, training of new personnel, validation of the theoretical knowledge acquired in line with the practices, the capacity and ability to fulfill these tasks are acquired.

However, with LLM technology reaching the capacity to perform cognitive tasks such as natural language understanding, generation, and reasoning, it offers the potential to transform, optimize, and comprehend complex engineering problems for such tasks [8]. Recent research has demonstrated growing interest in applying LLMs to various manufacturing domains. Makatura et al. [9] evaluated LLM capabilities across design and manufacturing tasks, including CNC machining project optimization, revealing both promising performance and notable limitations in spatial reasoning. Ni et al. [10] proposed an LLM-based manufacturing process planning approach aligned with Industry 5.0, utilizing prompt engineering to guide sequential decision-making in process plan generation. In a broader scope, Shahin et al. [11] surveyed the applications of generative AI across the manufacturing lifecycle, documenting case studies in quality control, process optimization, and production planning while identifying critical gaps in domain-specific adaptation. Furthermore, the integration of generative AI with digital twin technology has emerged as a complementary avenue; Mata et al. [12] developed a comprehensive framework combining human expertise with generative AI in digital twin-based manufacturing systems, demonstrating enhanced decision-making capabilities through virtual-physical synchronization.

Within this broader trend, several studies have specifically investigated LLM capabilities for G-code — the standard programming language for CNC machine tools. As noted in the work by Jignasu et al. [13], foundational LLMs showed limited understanding of geometries encoded in G-code for additive manufacturing. Expanding on this direction, Šket et al. [14] compared ChatGPT-3.5 and ChatGPT-4o for G-code generation in CNC machining and found that while newer models produced more accurate code, they still required human verification for practical deployment. Abdelaal et al. [15] addressed this reliability concern by proposing GLLM, a self-corrective G-code generation framework that incorporates user feedback loops to iteratively improve the generated output. These studies collectively establish that while LLMs possess a degree of competence in G-code related tasks, significant challenges remain in ensuring accuracy and completeness.

Beyond G-code generation, researchers have explored LLM deployment for broader CNC operational tasks. Kanimozhi and Sriker [16] demonstrated an explorative deployment of a fine-tuned LLM for on-site CNC machine operator assistance, showing that domain-adapted models can address machine-specific queries more effectively than general-purpose LLMs. Jeon et al. [17] proposed CNC-Talks, a conversational machine monitoring framework that integrates LLMs with real-time data retrieval augmented generation (RAG) to enable natural language-driven interaction with CNC equipment. Stathatos et al. [18] applied LLMs to high-level computer-aided process planning (CAPP) in a distributed manufacturing context, generating alternative process plans from textual part descriptions. While these works demonstrate LLM capabilities in individual aspects of manufacturing, they each focus on a single sub-task rather than evaluating LLMs across the full spectrum of cognitive tasks involved in a machining workflow.

Despite this growing body of research, no study has yet provided a comprehensive evaluation that compares multiple LLMs across different prompt engineering strategies on a complete machining workflow — encompassing G-code interpretation, cutting tool recommendation, and machining parameter estimation — while simultaneously validating the outputs against both industry-standard CAM software and digital twin simulations. In this paper, the domain expertise of foundational LLMs is evaluated across these three key tasks in the machine tool domain to determine whether LLMs can serve as a generic, optimal, and standardized solution for these operations. For this purpose, different foundational LLMs and different prompt techniques will be employed and compared.

MATERIALS AND METHODS

In this section, the traditional approaches for selecting machine parameters with domain expertise will be examined. The discussion begins with face milling as the reference operation.

Face Milling

Face milling is a machining operation that controls the height of the machine part. It is commonly used to generate flat reference surfaces before subsequent finishing operations.

Like all milling operations, face milling employs a cutting tool that rotates while the machined part remains stationary. Face milling requires that a specific amount of material be removed from the top of part, at one or several depth levels, in a single cut or multiple cuts [19].

Based on the machining operation, selection of a face mill cutter must consider several conditions [19].

- CNC machine specifications and condition
- Part material to be machined
- Setup method and work holding integrity
- Method of mounting
- Cutter overall construction
- Face mill diameter
- Number of inserts and insert geometry.

The last two items influenced actual program development at most, because it has a direct relation with the cutting feed and spindle speed. the number of inserts and milling diameter directly affects the cutting speed and spindle speed [20].

Large Language Models (LLMs)

Text generation and summarization with LLMs have increased significantly in recent years for many industries. Foundational LLMs which have been open sourced for use by the general have been trained with a large portion of data that includes many different fields including engineering, art, sports, and science. This way, LLMs can be used well for different types of industries and generate text for different types of tasks.

Also with reinforcement learning, the LLMs capabilities went beyond. Reinforcement learning is a learning type based on a rewarding system for the LLMs to gather feedback directly from humans [21]. This enabled chatbots to establish communication based on user feedback and align the results based on the needs.

As concluded in work by Jignasu et al. [13], while foundational LLMs can perform with reasonable proficiency, they also depict their critical limitations. Their experiments have shown that the understanding of LLMs for geometries with given G-Codes were limited in additive manufacturing domain.

Prompt Engineering

In this paper, three different types of prompting have been examined. Each type is presented with its role in the machine tool evaluation workflow.

Zero-Shot Prompts

Zero-shot prompts refer to instruction based single prompt, that can contain inputs such as:

- Step-by-step guidance
- Options to select
- Question etc.

Task description in a zero shot with instructions plays a crucial role during this type of prompting. E.g. "Translate 'how are you' into German."

As described in the work by Wei [22], a well-constructed zero-shot prompt performs considerably better than others. Theoretically a zero-shot prompt should contain whole information with the instructions for LLM to reason the question, evaluate the data, and format result in an expected format.

Few-Shot Prompts

Few-shot prompts generally require a few task related examples that can lead to an optimum solution for the given task. For better understanding of user need, AI can be adjusted in-context via few-shot prompts [23]. This leads to better generations via LLMs.

Few-shot prompting is a procedure that you give examples regarding to task for a better understanding and limit LLM to work based on the subject. It will lead LLM to a solution.

Few-shot prompting acts like a fine-tuning procedure. With prompts, users can define a context for LLM to understand what is needed, described or wanted for user.

Tree of Thought Prompts

A chain-of-thought is a series of intermediate natural language reasoning steps that lead to final output [24]. Tree of thoughts prompts are a combination of single-shot and few-shot prompting, it works poorly on tasks that require reasoning abilities and often does not improve substantially with increasing language model scale.

In the work by Wei [24], it is concluded that chain-of-thought prompting is a simple method for enhancing the reasoning capabilities of LLMs. It is understood that, with a chain of thought, LLMs can scale way better for arithmetic, symbolic and commonsense reasoning.

Tree of thought prompt contains instruction and input, and the procedure continues with the requirement of more answers. A self-discussion will be led by the LLM resulting in the perfect scenario for the output.

Procedures

The main purpose of this methodological approach is, by using different prompt techniques, to determine to what extent LLMs can effectively perform the tasks mentioned above, which are based on human expertise in the traditional sense. These critical cognitive tasks mostly remain dependent on humans, and LLMs offer a new opportunity to automate these tasks. However, the direct use of outputs can cause situations such as damage to the tool, spoilage of the workpiece, etc., meaning, in terms of practical applicability, verification is needed. Therefore, the methodology includes verification processes by comparing the generated outputs with CAM software (Siemens NX) [25] widely used in the industry and with simulations of the digital twin application (Create MyVirtual Machine) [26] also produced by Siemens for the Sinumerik CNC controller.

To make an evaluation specific to the relevant domain, the following LLM models have been selected. These models provide a balanced comparison across commercial and open-source LLM families.

Claude V3.7 Sonnet: Anthropic' s latest hybrid reasoning model; supports both fast inference and deep thought modes [27].

DeepSeek-R1: DeepSeek's inaugural open-source language model; designed for code and general-purpose tasks with competitive performance across benchmarks. Offers transparency, flexibility, and strong multilingual capabilities [28].

GPT-4.5: OpenAI's transitional model bridging GPT-4 and future iterations. Combines improved reasoning with faster response times and better tool integration. Ideal for both complex problem solving and everyday AI assistance [29].

Ground Truth Definition

In this section, the exact dataset ("ground truth") used as a reference in the comparative evaluation of LLM capabilities is defined in detail. Two face milling operations have been examined to compare and give in context.

Face Milling Operation Example 1

Workpiece material and tool features

The tool configuration for this operation is shown in Figure 1.

Material: C40 Steel (AISI/SAE 1040)
Tool Type: Facing Tool
Tool Diameter: 52 mm
Number of Flutes: 6
Cutting Angle: 45 °

Spindle speed and feed rate

Spindle speed for roughing: 980 RPM
Spindle speed for finishing: 1100 RPM
Cutting speed (Vc) for roughing: 160 m/min
Cutting speed (Vc) for finishing: 180 m/min
Feed rate for roughing: 880 mm/min
Feed rate for finishing: 530 mm/min
Feed per Tooth (Fz): 0.08 mm/rev

Operation Description

A face-milling operation is performed. It contains two groups, for roughing and finishing. The tool approaches the part along the Z axis to the reference plane (G507(Z20)), then moves in the part to 2.2 mm (G1 Z-2.2) and begins cutting by descending 0.2 mm below the surface (G1 Z-2.0). And then the finishing group starts with going first to the reference plane(G507). Then, finishing starts with going down to cutting area with command (G1 Z-2.5). And face milling operation continues.

Geometric Description

Initial setup:

- Program begins with G17 (XY plane), G40 (cutter compensation off), G90 (absolute positioning)
- Tool selection: "ALU_D63" with actual cutting diameter of 52mm and 6 teeth
- Workpiece defined as box shape with dimensions 155x100x20

Approach:

- Rapid traverse (G0) to initial position X5 Y-135
- Spindle starts at S980 rpm with cutting speed VC=160m/min

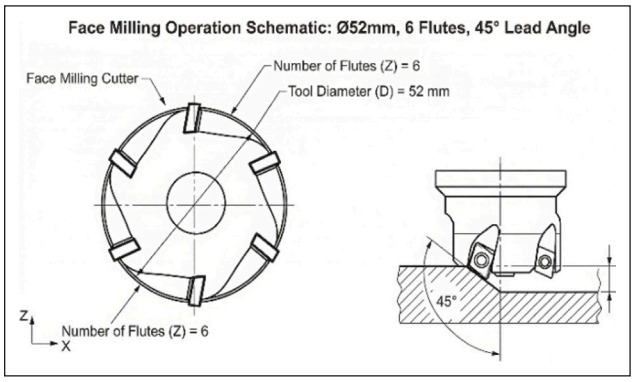

Figure 1. Tool scheme for face milling operation example 1.

- Feed rate F880 with feed per tooth FZ=0.15mm
- Rapid descent to Z1mm followed by coolant activation (M8)

Roughing operation:

- Plunge cut to Z-2.2mm with 3mm corner rounding (RNDM=3)
- Zigzag pattern with 31mm step over (AE):
  - Linear move from Y-135 to Y0
  - 2. Step right to X36
  - Linear move to Y-100
  - Step right to X67
  - Linear move to Y0
  - Step right to X98
  - Linear move to Y-100
  - Step right to X129
- Final move to Y0 Retract to Z2mm at end of roughing

Finishing operation:

- Repositioning to X0 Y-135 with increased spindle speed S1100 (VC=180m/min)
- Reduced feed rate F530 (FZ=0.08mm) for better surface finish
- Six parallel vertical passes with 26mm step over (AE):

1. Pass at X0: Cut from Y-135 to Y35, retract to Z2
2. Pass at X26: Cut from Y-135 to Y35, retract to Z2
3. Pass at X52: Cut from Y-135 to Y35, retract to Z2
4. Pass at X78: Cut from Y-135 to Y35, retract to Z2
5. Pass at X104: Cut from Y-135 to Y35, retract to Z2
6. Pass at X130: Cut from Y-135 to Y35, retract to Z2

- Finishing depth Z-2.5mm (0.3mm deeper than roughing)

Surface Coverage:

- X-axis coverage: 0mm to 130mm (total width of 130mm)
- Y-axis coverage: -135mm to 35mm (total length of 170mm)

Face Milling Operation Example 2

Workpiece material and tool features

The tool configuration for this operation is shown in Figure 2.

Material: Cast iron
Tool Type: Facing Tool
Tool Diameter: 25 mm
Number of Flutes: 3
Cutting Angle: 45 °

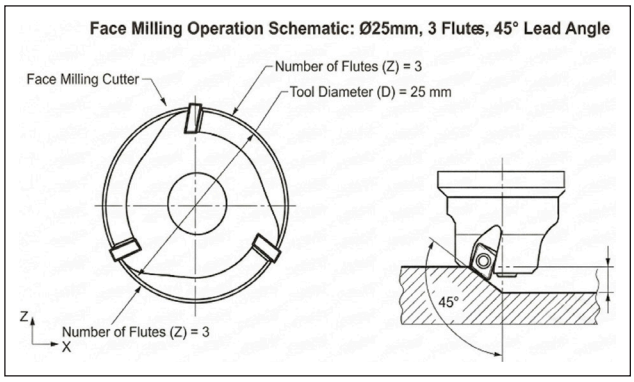

Figure 2. Tool scheme for face milling operation example 2.

Spindle Speed and Feed Rate
Spindle Speed: 600 RPM
Feed rate: 180 mm/min

Operation Description
A face-milling operation is performed. The operation is executed in multiple passes with decreasing depth: 1.5mm, 1.0mm, 0.5mm, and 0.0mm. The tool follows a zigzag pattern with both linear and circular movements. The operation runs at a spindle speed of 600 RPM with a feed rate of 180mm/min. Coolant is active during the cutting operation. The coolant type is emulsion with boron oil.

Geometric Description
Initial setup:
- The program begins with G17 (XY plane), G40 (cutter compensation off), G90 (absolute positioning), G94 (feed per minute)
- Tool selection: "MILL_D25" with 25mm cutting diameter
- Program uses G54 workpiece coordinate system
- Approach:
- Initial rapid positioning to X-360 Y223.9
- Tool change and setup with MILL_D25
- Spindle starts at S600 rpm with coolant activation (M8)
- Initial positioning to X-46.138 Y23.724 with B0 and C0

Cutting operation:
- Four identical passes at different Z depths (1.5mm, 1.0mm, 0.5mm, 0.0mm)
- Each pass follows the same pattern:
  - Linear cuts along X-axis
  - Circular interpolation (G2/G3) for curved features
  - Combined linear and circular movements creating a complex profile

Surface coverage:
- X-axis coverage: approximately -49mm to +37mm
- Y-axis coverage: approximately -20mm to +24mm
- Z-axis levels: 1.5mm, 1.0mm, 0.5mm, and 0.0mm

Prompt Techniques Examined Specific to Machine Tool Domain

In this section, it is explained in detail how the selected prompt techniques -Zero- Shot, Few- Shot and Tree of Thought- are applied to the tasks in the study. The objective is to show how each technique supports operation review, tool selection, and parameter estimation.

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below G-Code and extract the operation type and geometric description of the operation.*
Input: *<GCODE>*
Explanation: *<Area for LLM output>*

Figure 3. Zero shot prompt template for operation examination.

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below G-Code and extract operation type and geometric description of the operation.*
Example:
<GCODE_1>, <Explanation of GCODE_1>
Input: *<GCODE>*
Explanation: *<Area for LLM output>*

Figure 4. Few shots prompt template for operation examination.

Prompt Configurations for Operation Review Task

The main purpose at this stage is to evaluate the LLM's ability to correctly parse and interpret a universally written G Code part for the Face Milling operation. Thanks to the relevant prompt, it will be asked to infer information such as a specific "Operation Type" (e.g.: Face Milling) and "Geometric Description" (e.g.: the process of correcting a 100x80 mm surface by removing 0.5 mm).

Workpiece material and hardness are defined using general information obtained from factory data. Although this information is not directly used in this stage, it is critical for subsequent tasks and is included in the setup.

Zero Shot Prompting Application for Direct Inference in Operation Examination Task

With this prompt technique, it is designed in such a way that the LLM is directly requested to perform a specific task using its general knowledge about CNC machining, G-Code analysis. Within the prompt, a role is usually assigned to the model, and it is ensured that it focuses on the relevant area of expertise. The prompt template is presented in Figure 3.

Few Shots Prompting Application for More Guidance in Operation Examination Task

With this prompt technique, it is structured to include a few examples input and output pairs to obtain the desired output. Within the framework of this study, a two-example learning approach will be used. As an example, a similar G-Code and the operation type and geometric description of this G-Code will be given. The prompt structure is shown in Figure 4.

Tree of Thought Prompting Application for Deliberative Reasoning in Operation Examination Task

With this prompt technique, an attempt has been made to elicit deliberative reasoning for a single interaction with the LLM. Imagining multiple experts encourages the LLM to use various perspectives and generate lines of reason-

Role: *CNC Machine Tool Expert*
Instruction: *Imagine four different experts are analysing this g code, all experts, all experts will write down their own explanation about operation type and geometric description of the corresponding operation, then share it with the group. If any expert realises, they're wrong at any point, then they leave. At the end of this discussion, give the summary of geometric description and operation type via adding experts' conflicts.*
Input: *<GCODE>*
Explanation: *<Area for LLM output>*

Figure 5. Tree of thought prompt template for operation examination.

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below operation and suggest me the suitable tool type, tool material and tool geometry parameters.*
Input: *<Desired machining operation description with geometric explanation and with workpiece material information>*
Explanation: *<Area for LLM output>*

Figure 6. Zero shot prompt template for selecting suitable tool properties.

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below operation and suggest me the suitable tool type, tool material and tool geometry parameters.*
Example:
<Operation Description>,<Tool Type>,<Tool Material>, <Tool Geometry Features>
Input : *<Desired machining operation description with geometric explanation and with workpiece material information>*
Explanation : *<Area for LLM output>*

Figure 7. Few shots prompt template for selecting suitable tool properties.

ing. Each expert sharing their thoughts ensures that these causal steps are clearly expressed. Asking it to realize it is mistaken enables it to perform internal evaluation and eliminate this line of thought. The prompt template is given in Figure 5.

Prompt Configurations for the Task of Selecting Suitable Tool Properties

The main purpose at this stage is to evaluate the LLM's ability to recommend suitable cutting tools for a specific Face Milling operation, considering the process requirements and the properties of the workpiece material. Thanks to the relevant prompt, it will be asked to infer information such as "Tool Type" (e.g.: Indexable Face Mill), "Tool Material" (e.g.: Coated Carbide) and "Tool Geometry Features" (e.g.: Diameter, Number of Flute, Length). The properties of the tool will be compared with

Role: *CNC Machine Tool Expert*
Instruction: *Imagine four different experts are analysing this machining operation description, all experts will write down their own explanation about suitable tool type, tool material and tool geometry features, then share it with the group. If any expert realises, they're wrong at any point, then they leave. At the end of this discussion, give the summary of suggested tool type, tool material and tool geometry features via adding experts' conflicts.*
Input: *<Desired machining operation description with geometric explanation and with workpiece material information>*
Explanation: *<Area for LLM output>*

Figure 8. Tree of thought prompt template for selecting suitable tool properties.

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below tool type, tool material and tool geometry, workpiece material, operation type suggest me the suitable Cutting Speed and Feed rate*
Input: *< Desired machining operation type, workpiece material, tool type tool material and tool geometric features.>*
Explanation: *<Area for LLM output>*

Figure 9. Zero shot prompt template for spindle speed and feed rate estimation.

applied and used data obtained from factory data as it is mentioned in Section Claude Sonnet v3.7 Evaluation and Section Deepseek-R1 Evaluation.

Zero Shot Prompting Application for Direct Inference in the Task of Selecting Suitable Tool Properties

The prompt created using the approach mentioned in Section Claude Sonnet v3.7 is as follows (Fig. 6).

Few Shots Prompting Application for Deliberative Reasoning in the Task of Selecting Suitable Tool Properties

The prompt created using the approach mentioned in Section Deepseek-R1 Evaluation is as follows (Fig. 7).

Tree of Thought Prompting Application for Deliberative Reasoning in the Task of Selecting Suitable Tool Properties

The prompt created using the approach mentioned in Section GPT4.5 Preview Evaluation is as follows (Fig. 8).

Prompt Configurations for the Task of Estimating Suitable Spindle Speed and Feed Rate

The main purpose at this stage is to evaluate the LLM's inference of suitable cutting parameters -cutting speed (RPM) and feed rate (mm/minute)- for a specific workpiece and tool. This allows a direct comparison of model recommendations against reference process values.

Zero Shot Prompting Application for Direct Inference in the Task of Estimating Suitable Spindle Speed and Feed Rate

The prompt created using the approach mentioned Section GPT4.5 Preview Evaluation is as follows (Fig. 9).

Role: *CNC Machine Tool Expert*
Instruction: *Analyse the below tool type, tool material and tool geometry, workpiece material, operation type suggest me the suitable Cutting Speed and Feed rate*
Example:
<Workpiece Material>,<Operation Type>,<Tool Type>,<Tool Material>,<Tool Geometry Features>
Input: *<Desired machining operation type, workpiece material, tool type tool material and tool geometric features>*
Explanation : *<Area for LLM output>*

Figure 10. Few shots prompt template for spindle speed and feed rate estimation.

Few Shots Prompting Application For More Guidance in the Task of Estimating Suitable Spindle Speed and Feed Rate

The prompt created using the approach mentioned in Section Deepseek-R1 Evaluation Figure is as follows (Fig. 10).

Tree of Thought Prompting Application for Deliberative Reasoning in the Task of Estimating Suitable Spindle Speed and Feed Tate

The prompt created using the approach mentioned in Section GPT4.5 Preview Evaluation is as follows (Fig. 11).

EXPERIMENT AND APPLICATION PROTOCOL

Based on data obtained from the factory environment, sample G-Code blocks will be created with the help of NX CAM. This created G-Code will again be simulated on NX CAM, and the final processed output will be recorded and stored for future comparison [25].

Application of Prompts

For each defined prompt – G-Code interpretation, tool selection, and cutting parameter estimation – it will be run on the selected LLM platform. The text-based responses of the LLM will be recorded directly.

Operation Examination

The textual description regarding the G Code produced with the help of NX CAM will be evaluated based on known parameters and purpose. The assessment focuses on whether the generated interpretation matches the intended operation and geometry.

Selecting Tool Properties

A virtual tool corresponding to the "tool geometry features" suggested by the LLM will be modelled in the simulation environment using the Create MyVirtual Machine application. The "Tool Type" and "Tool Material" outputs will be evaluated by comparing them with data obtained from the factory [26].

Estimating Spindle Speed and Feed Rate

The "Cutting Speed" and "Feed Rate" values suggested by the LLM will be applied in the simulation environment using Create MyVirtual Machine application. At this stage,

Role: *CNC Machine Tool Expert*
Instruction: *Imagine four different experts are analysing this desired machining operation type, workpiece material, tool type tool material and tool geometric features, all experts will write down their own explanation about suitable cutting speed and feed rate estimation for given tool, then share it with the group. If any expert realizes, they're wrong at any point, then they leave. At the end of this discussion, give the summary of suggested feed rate and spindle speeds via adding experts' conflicts.*
Input: *<Desired machining operation type, workpiece material, tool type tool material and tool geometric features>*
Explanation: *<Area for LLM output>*

Figure 11. Tree of thought prompt template for spindle speed and feed rate estimation.

the tool is virtually modelled in section Face Milling and the G-Code used for interpretation in section Large Language Models will be run.

Comparison and Verification

The basic verification control will take place by running the created G-Code in the Create MyVirtual Machine environment. The main criterion for success is "the observation that the processed material is compatible with the NX Cam simulation output.". The properties of the virtually processed part (dimensions of the area where Face Milling is applied) include a geometric comparison of the output provided by NX CAM. The estimation of tool selection and cutting parameters will be verified by comparing them with data obtained from the factory.

Performance Evaluation Metrics

A series of metrics will be used to systematically evaluate the performance of LLM outputs throughout different prompt techniques and experimental stages. These metrics are designed by referencing similar academic studies to measure the degree of accuracy and usability of LLM's suggestions [13].

Definition of Success Metrics Related to Operation Examination Task:

- Success: The LLM correctly identifies the operation type and presents an accurate and complete geometric description of the machining features, matching the given G-Code function.
- Partially Success: The LLM correctly identifies the operation type, but the geometric description contains minor deficiencies or errors that do not alter the essence of the operation.
- Unsuccess: The LLM identifies the wrong operation type, or the geometric description is significantly erroneous or misleading.

Definition of Success Metrics Related to Selecting Tool Properties Task:

- Success: All suggested tool parameters ("Tool Type", "Tool Material", "Tool Geometry") represent a valid, optimal, or very close to optimal definition for the relevant machining and workpiece. It will be compared with proven data obtained from the factory.

- Partially Success: The suggested tool is generally suitable but may require minor adjustments for one or more parameters (e.g., if the suggested diameter is close to a standard size) or the suggested tool is a functional alternative, although not optimal.
- Unsuccess: The suggested tool is not suitable for the operation, the material selection is incorrect, or the geometric properties lead to poor performance or failure.

Definition of Success Metrics Related to Estimating Spindle Speed and Feed Rate Task:

- Success: The suggested spindle speed and feed rate values are within an acceptable tolerance range, such as ± 5% of proven values obtained from the factory, and result in smooth machining in the digital twin.
- Partially Success: The parameters are outside the optimal range but do not cause the simulation to fail completely (e.g., longer processing time), the part can still be machined correctly. It remains within the reference value ranges in the datasheet published by the tool manufacturer for the tool used in the factory as it is mentioned in Section Claude Sonnet v3.7 Evaluation and Section Deepseek-R1 Evaluation.
- Unsuccess: The suggested parameters are largely incorrect, leading to errors such as tool breakage, surface defects, or inability to complete the operation. The suggested values are recommended completely independently of the reference ranges in the tool defined factory data as it is mentioned in Section Claude Sonnet v3.7 Evaluation.

RESULTS AND DISCUSSION

At this stage, the results of the experiments conducted in line with the methodology will be discussed within the following scope. The discussion highlights both strengths and limitations observed across models and prompting techniques.

- Relative strengths and weaknesses of each prompt technique for the selected application specific to the machine tool domain.
- The type of error and tendency to error that LLMs repeat when working with machine tool information.
- The extent to which the current LLM outputs are reliable and the level of modification or expert intervention they generally require.
- Inferences for future research regarding its development for use as a domain assistant.

Discussions About the Operation Examination Task

In this section, the operation examination prompts were used for different language models and evaluation results have been discussed. The comparison emphasizes how accurately each model interprets operation type and geometry.

Claude Sonnet v3.7 Evaluation

Demonstrated a significant ability in this task. All prompting techniques achieved a "Partially Success" rating.

The Zero Shot model could not identify the operation as "face milling or surface milling", it identified the operation as a "2D profile milling" operation. But the geometrical description was accurate. So, Zero Shot model correctly identified the 4 depth phases of the operation with 0.5 mm of depth. The path start was correctly identified, feed rate has been detected, and the cutting depth of the phases has correctly also identified. What made zero shot model to identify operation not correctly was the circular arcs in the profile shape. It was a face milling operation with smoothing that is why it contained circular arcs. But Zero Shot model detected these as a profile.

The Few Shots model also could not identify this operation as a "face milling or surface milling" operation, it identified the operation as a "Contour Milling operation." But the geometrical description was also accurate. It correctly identified the four depth passes and in which Z coordinates it started. It correctly identified that these passes have the same pattern, and this operation is a high-precision operation with high-precision compression mode. It also detected that there is smooth motion control.

The Tree of Thought approach, leveraging simulated expert consensus, also could not identify the operation as "Face milling / Surface machining." The geometric description mirrored Zero-Shot's accuracy in most aspects, including the same minor X-limit discrepancy.

Deepseek-R1 Evaluation

Demonstrated a medium rated ability for this task. All prompting techniques could determine geometric descriptions but had failure determining the operation.

The Zero Shot prompting model has detected the operation as "2.5D Milling". It correctly identified the 4 passes at correct depths. It could also identify the feed rate and spindle speed from the given g-code and could describe the contouring path correctly.

The Few Shots model also could not identify this operation as a "face milling or surface milling" operation, it identified the operation as a "Contour Milling" operation. But the geometrical description was also accurate. It correctly identified that there exists multiple depth passes and in which Z coordinates it started. It correctly identified that these passes have the same pattern. It correctly identified that the operation uses step-down milling with positional accuracy tolerances of 0.01 mm on all axes.

Tree of Thoughts approach has determined the operation type as "3-axis symmetrical profile finishing". The multilevel passes were in the expert discussion, but it was not included in the conclusion. The interpolation turning and linear movements were also included in the expert discussion, but it was not declared in conclusion. There was no information about spindle speed or feed rate.

GPT4.5 Preview Evaluation

Demonstrated a medium rate of ability with whole types of prompting. It could be said that GPT4.5 was not generally successful for this type of domain.

The Zero Shot model has determined the given operation as a "contour milling" operation. It could deter-

mine that there exists a there are contour passes from 5 to 0, but the accurate path or a step depth was not determined. The feed rate and spindle speed were also found from operation.

The Few Shots model had determined the operation as a “pocket milling” operation. It has been determined that the operation starts with an approach to 5 mm in Z and proceeds with cutting in different depths. The depths steps were given correctly. The spindle speed and feed rate were present. It also included there exists smooth transitions between segments with features like radii of 7.281 mm.

Tree of Thoughts approach generally failed. It determined the operation as a “roughing operation” with an end mill tool. The operation was made with a facing tool. The steps were also determined in the summary and discussion.

Comparison Between Language Models

With every prompt technique; all language models have failed to determine the operation type. They successfully found the depth step passes and determined the machining parameters that are contained in the operation. Geometrical descriptions were present, but did not cover the whole operation details. It would really need additional comments and information from operators to achieve more.

The best answers were provided from Claude V3.7 Sonnet. Other than operation classification, geometrical descriptions of the Claude V3.7 Sonnet were more detailed and more accurate. The summary of evaluation results is presented in Table 1.

Discussions About the Selecting Tool Properties Task

Claude Sonnet v3.7 Evaluation

All evaluated prompting techniques achieved a "Partially Success" rating. This indicates a general suitability of the suggestions, though none perfectly matched GT’s specific optimal tool.

In Zero Shots, the model correctly recommended a "Face mill cutter with multiple inserts " with material "Coated carbide inserts," which aligns with the GT and workpiece material. For tool geometry parameters, the model suggested a diameter range of 63-80 mm for this operation but has not determined a specific diameter as in ground truth. The number of flutes is also not specified, but a range of 5-8 has been given as a suggestion via model. Cutting angle was specified as 45 °. This was also a tool information that is determined in ground truth.

Few Shots approach also correctly identified the "Facing Tool" type and "Carbide with TiAlN coating." It suggested specific parameters: L=80mm (GT: 100mm), D=63mm (GT: 25mm), and 5 Flutes (GT: 3). While the diameter differed, 63 mm is a standard and functional alternative.

The Tree of Thought summary correctly identified the "Face mill cutter" type and a consensus on "carbide inserts." The geometric discussion was rich, mentioning positive rake angles and lead angles. However, the final summary did not converge on specific dimensions for diameter or flute count matching the GT, instead offering characteristics of a suitable tool class.

Table 1. Operation examination results for selected LLMs

LLMs	Success	Partially success	Unsuccess
Claude v3.7 sonnet		*	
Deepseek-R1			*
GPT 4.5			*

Deepseek-R1 Evaluation

With this model, the few shots prompting technique achieved a "Partially Success" rating. Other techniques have not been so successful for suggesting an alternative functional tool or a similarity with ground truth. This indicates a general suitability of the suggestions, though none perfectly matched GT’s specific optimal tool.

With Zero Shot approach, the model correctly recommended a “Face milling tool with indexable inserts”. The material was “Carbide with TiAlN coating. It suggested s 4-6 numbers of flutes but did not suggest a diameter. It specified a 45 ° lead angle, positive rake angle. There were no diameters or tool length specified. It can be said that a base shape for a cutter has been determined but specified alternative or a ground truth-based example has not been given.

With Few Shots approach, the parameters were specified. It correctly identified the operation as a face milling and gave a suggestion for the tool type as “face milling cutter” with material as” carbide with TiAlN coating”. It gave tool length 80 mm (GT: 100 mm), diameter as 63 mm (GT: 25 mm), number of flutes as 5 (GT: 3). This correctly establishes a functional alternative for the operation as a suggestion.

With the Tree of Thoughts approach, the parameters were not specified in the Summary. It correctly identified the operation as a face milling and gave a suggestion for the tool type as “face milling cutter” with material as” carbide with TiAlN coating”. The model did not suggest a diameter, number of flute or tool length. No similarity or ground truth has been determined by model for this operation with this prompting approach.

ChatGPT-4.5 Preview Evaluation

With ChatGPT-4.5 Preview model, the few shots prompting technique achieved a "Partially Success" rating. Other techniques were not suitable for suggesting an alternative functional tool or a similarity with ground truth. This indicates a general suitability of the suggestions, though none perfectly matched GT’s specific optimal tool.

In Zero Shot approach, the model correctly recommended a “Face milling tool with indexable inserts”. The material was “Carbide with TiAlN coating. No specific parameters for length, diameter or number of flutes have been suggested. It specified a 45 ° lead angle, positive rake angle.

In Few Shots approach, the parameters were specified. It correctly identified the operation as a face milling and gave a suggestion for the tool type as “facing tool” with material as” carbide ISO P-type recommended”. It gave tool length 80 mm (GT: 100 mm), diameter as 50 mm (GT: 25 mm), number of flutes as 4 (GT: 3). This correctly establishes a functional alternative for the operation as a suggestion.

Table 2. Selecting tool properties results for selected LLMs

LLMs	Success	Partially success	Unsuccess
Claude v3.7 Sonnet		*	
Deepseek-R1		*	
GPT 4.5		*	

In the Tree of Thoughts approach, the parameters were not specified in the Summary. It correctly identified the operation as a face milling and gave a suggestion for the tool type as “face milling cutter” with material as” carbide with TiAlN coating”. The model did not suggest a diameter, number of flute or tool length. No similarity or ground truth has been determined by model for this operation with this prompting approach.

Comparison Between Language Models

The Few-Shots approach established good reasoning with different models; however, apart from this approach, the information provided was insufficient to create a functional alternative tool or a similar tool with ground truth. The best scenario was established by Claude V.37 Sonnet. The evaluation results are summarized in Table 2.

Discussions about the Estimating Spindle Speed and Feed Rate Task

Claude Sonnet v3.7 Evaluation

With the Zero Shot technique, parameters derived (n=2292 RPM, Vf=1032 mm/min) based on standard machining parameters (Vc = 180 m/min, fz = 0.15 mm/tooth). While these are common catalogue values for C40 steel with carbide and would likely allow the operation to proceed without immediate tool failure, the given suggestion and actual GT values differ. The federate for finishing in GT was approximately 10 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 4 times bigger than the GT. This significant discrepancy renders the output "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

With the Few Shots technique, parameters derived (n=2000 RPM, Vf=900 mm/min) based on standard machining parameters (Vc = 160 m/min, fz = 0.15 mm/tooth). While these are common catalogue values for C40 steel with carbide and would likely allow the operation to proceed without immediate tool failure, the given suggestion and actual GT values differ. The federate for finishing in GT was approximately 8 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 4 times bigger than the GT. This significant discrepancy renders the output "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

The Tree of Thought's consolidated recommendations (n=2230 RPM, Vf=1000 mm/min) were, like Zero-Shot, based on conservative, widely accepted cutting speeds (Vc≈207 m/min) and feeds per tooth (fz≈0.12 mm/tooth). These deviated from the GT by approximately 3 times greater and 10 times greater. While the reasoning process involved simulated expert discussion, the outcome still fell far bigger of GT’s optimized parameters, leading to an "Unsuccess" rating.

Deepseek-R1 Evaluation

With the Zero Shot technique, parameters derived (n=2546 RPM, Vf=1500 mm/min) based on standard machining parameters (Vc = 200 m/min, fz = 0.2 mm/tooth). While these are common catalogue values for C40 steel with carbide and would likely allow the operation to proceed without immediate tool failure, the given suggestion and actual GT values differ. The federate for finishing in GT was approximately 10 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 4 times bigger than the GT. This significant discrepancy renders the output "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

With the Few Shots technique, parameters derived (n=2000 RPM, Vf=900 mm/min) based on standard machining parameters (Vc = 160 m/min, fz = 0.15 mm/tooth). While these are common catalogue values for C40 steel with carbide and would likely allow the operation to proceed without immediate tool failure, the given suggestion and actual GT values differ. The federate for finishing in GT was approximately 8 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 4 times bigger than the GT. This significant discrepancy renders the output "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

The Tree of Thought's consolidated recommendations (n=1900 RPM, Vf=750 mm/min) were, like Zero-Shot, based on conservative, widely accepted cutting speeds (Vc≈207 m/min) and feeds per tooth (fz≈0.12 mm/tooth). These deviated from the GT by approximately 3 times greater for spindle speed and 7 times greater for feed rate. While the reasoning process involved simulated expert discussion, the outcome still fell far bigger of GT’s optimized parameters, leading to an "Unsuccess" rating.

ChatGPT-4.5 Preview Evaluation

With the Zero Shot technique, parameters derived (n=2290 RPM, Vf=680 mm/min) based on standard machining parameters (Vc = 180 m/min, fz = 0.15 mm/tooth). The suggestion given and actual GT values differ. The federate for finishing in GT was approximately 5 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 3 times bigger than the GT. This significant discrepancy renders the out-

Table 3. Estimating spindle speed and feed rate results for selected LLMs

LLMs	Success	Partially success	Unsuccess
Claude v3.7 Sonnet			*
Deepseek-R1			*
GPT 4.5			*

put "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

With the Few Shots technique, parameters derived (n=2040 RPM, Vf=920 mm/min) based on standard machining parameters (Vc = 140 m/min, fz = 0.15 mm/tooth). While these are common catalogue values for C40 steel with carbide and would likely allow the operation to proceed without immediate tool failure, the given suggestion and actual GT values differ. The federate for finishing in GT was approximately 8 times lower than the suggested feed rate. Spindle speed for this operation has been suggested by model also approximately 4 times bigger than the GT. This significant discrepancy renders the output "Unsuccess" when compared against the factory's proven, high-performance values. Its relative "strength" (if any in this context) was adherence to general safe practices, but this was insufficient for the task's success criteria.

The Tree of Thought's consolidated recommendations (n=1700 RPM, Vf=1000 mm/min) were, like Zero-Shot, based on conservative, widely accepted cutting speeds (Vc≈135 m/min) and feeds per tooth (fz≈0.2 mm/tooth). These deviated from the GT by approximately 3 times greater for spindle speed and 10 times greater for feed rate. While the reasoning process involved simulated expert discussion, the outcome still fell far bigger of GT's optimized parameters, leading to an "Unsuccess" rating.

Comparison Between Language Models

The ground truth values differ a lot with the suggested values with all language models. Spindle speed values were much higher than ground truth for every model, it was also the same for feed rate. It can be said that standard values for calculation of the feed rate and spindle speed are not enough to get optimum cutting speeds. The results are presented in Table 3.

CONCLUSION

This study evaluated three foundational large language models — Claude V3.7 Sonnet, DeepSeek-R1, and GPT-4.5 — across three core machining tasks (operation examination, cutting-tool selection, and spindle-speed / feed-rate estimation) using zero-shot, few-shot, and tree-of-thought prompting. The outputs were verified against industry-standard CAM software (Siemens NX [25]) and digital twin simulations (Create MyVirtual Machine [26]). The results reveal a clear gradient of capability: LLMs can partially interpret machining operations, can propose functional alternative cutting tools under guided prompting, but consistently fail at quantitative parameter estimation — with suggested spindle speeds and feed rates deviating from factory-proven values by factors of three to ten.

A central finding is that apparent improvements in LLM output can mask underlying safety hazards — a pattern appropriately described as "false efficiency." Campean and Pop [30] independently demonstrated this phenomenon: ChatGPT's optimization of a CNC milling program yielded a 37 % cycle-time reduction, yet 83 % of that saving resulted from the silent deletion of a required pocket-milling operation and the removal of safety-critical G43 and G28 commands, producing a non-conforming part. In the present study, the 3–10× parameter overestimates would likewise shorten nominal cycle times while simultaneously pushing the process into regimes that risk tool breakage, workpiece scrap, and machine collision. These results caution against interpreting raw LLM-generated metrics as genuine process improvements without rigorous physical verification.

The experiments further demonstrate that deploying LLMs in the machining domain does not eliminate the need for human expertise — it redistributes it. Every LLM output in this study required expert review: operation descriptions needed correction of misidentified operation types, tool suggestions required dimensional adjustment, and parameter estimates had to be discarded entirely. Rather than reducing the operator's cognitive load, the current generation of LLMs shifts the burden from direct task execution to output verification — a distinction that must be acknowledged when assessing the practical value proposition of LLM-assisted manufacturing. Aghaei and Ansari [31] reached a similar conclusion in their broad review, noting that foundation language models generate linguistically plausible but technically inaccurate answers and that safety-critical validation mechanisms are rarely integrated into existing workflows.

Importantly, the failures observed in parameter estimation cannot be attributed solely to prompt engineering deficiencies. Even the tree-of-thought technique — which structures multi-step deliberative reasoning and represents the most sophisticated prompting strategy tested — produced "Unsuccess" results across all three models. This outcome indicates that the limitation is not in how questions are posed to the model but in the model's fundamental lack of deterministic, physics-grounded reasoning. LLMs generate outputs by statistical next-token prediction over textual corpora; they do not solve the underlying force-balance, thermal, and vibration equations that govern metal-cutting dynamics. Consequently, no refinement of natural-language prompting alone is likely to bridge the gap between catalogue-level heuristics and the process-specific parameter sets that safe, efficient machining demands.

Given this structural limitation, future research should explore hybrid architectures that combine the natural-language interface strengths of LLMs with physics-informed computational methods. Physics-informed neural networks (PINNs) have recently demonstrated the ability to encode governing equations directly into the learning process for machining applications. Darshan et al. [32] developed a physics-informed cutting-force model for end milling that

achieved high predictive accuracy with limited experimental data by embedding force-equilibrium constraints into the network loss function. Zhang et al. [33] proposed a hierarchical Bayesian PINN for the inverse estimation of grinding process parameters grounded in contact-mechanics theory. These advances suggest that coupling an LLM front-end — responsible for G-code interpretation, intent parsing, and report generation — with a PINN back-end — responsible for physics-constrained parameter estimation and process simulation — could yield a system that retains the accessibility of natural-language interaction while delivering the quantitative reliability that standalone LLMs currently lack.

In summary, while foundational LLMs represent a promising interface technology for the machine tool domain, the present results demonstrate that they are not yet a substitute for domain expertise in safety-critical machining decisions. Their deployment should be accompanied by mandatory verification through digital twin simulation and expert oversight, and future development should focus on integrating physics-based reasoning capabilities rather than relying solely on larger training corpora or more elaborate prompting strategies.

Data Availability Statement

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

Author's Contributions

Ugur Eniş: Conception, Design, Methodology, Data Collection and Processing, Prompt Engineering, LLM Evaluation, Analysis and Interpretation, Literature Review, Writer, Critical Review, GitHub Repository Management.

Mehmet Şamil Soyer: Conception, Design, Virtual Twin Simulation, CAM Software Verification, Data Processing, Analysis and Interpretation, Literature Review, Writer, Critical Review.

Hanife Ünal Helvacıoğlu: Conception, Supervision, Methodology Design, Analysis and Interpretation, Literature Review, Critical Review.

Muhammet Mustafa Savasçı: Conception, Supervision, Methodology Design, Analysis and Interpretation, Literature Review, Critical Review.

Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Statement on the Use of Artificial Intelligence

This research evaluates the capabilities of large language models (LLMs) including Claude v3.7 Sonnet, DeepSeek-R1, and GPT-4.5 for machining parameter estimation and G-code interpretation. These AI models were the subject of the study, not tools used in writing the manuscript. The manuscript itself was written by the authors without the assistance of generative AI writing tools.

Ethics

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Parashar, B. S. N., & Mittal, R. (2013). *Elements of manufacturing processes*. PHI Learning Pvt. Ltd.

[2] Yusup, N., Zain, A. M., & Hashim, S. Z. M. (2012). Evolutionary techniques in optimizing machining parameters: Review and recent applications (2007-2011). *Expert Systems with Applications, 39*, 9909-9927. [CroosRef]

[3] Davis, J. R. (Ed.). (1989). *ASM handbook, volume 16: Machining*. ASM International.

[4] Soori, M., Arezoo, B., & Dastres, R. (2023). Machine learning and artificial intelligence in CNC machine tools, a review. *Sustainable Manufacturing and Service Economics, 2*, Article 100014. [CroosRef]

[5] Pimenov, D. Y., Bustillo, A., Wojciechowski, S., Sharma, V. S., Gupta, M. K., & Kuntoğlu, M. (2023). Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. *Journal of Intelligent Manufacturing, 34*, 2079-2121. [CroosRef]

[6] Navaneethan, G., Palanisamy, S., Jayaraman, P. P., Kang, Y.B., Stephens, G., Papageourgiou, A., & Navarro-Devia, J. (2024). A review of automated cutting tool selection methods. *The International Journal of Advanced Manufacturing Technology, 133*, 1063-1092. [CroosRef]

[7] Oberg, E., Jones, F. D., Horton, H. L., Ryffel, H. H., & McCauley, C. J. (2020). *Machinery's handbook* (31st ed.). Industrial Press.

[8] Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N, Song, W., Zhai, X., Song, K., Zhu, D., Li, B., Wang, X., & Liu, T. (2024). Large language models for manufacturing. *arXiv*. Preprint. doi: 10.48550/arXiv.2410.21418

[9] Makatura, L., Foshey, M., Wang, B., Hähnlein, F., Ma, P., Deng, B., Tjandrasuwita, M., Spielberg, A., Owens, C. E., Chen, P. Y., Zhao, A., Zhu, A., Norton, W. J., Gu, E., Jacob, J., Li, Y., Schulz, A., & Matusik, W. (2024). *Large language models for design and manufacturing*. MIT GenAI. https://mit-genai.pubpub.org/pub/nmypmnhs Accessed Jun 03, 2026. [CroosRef]

[10] Ni, M., Wang, T., Leng, J., Chen, C., & Chen, L. (2025). A large language model-based manufacturing process planning approach under industry 5.0. *International Journal of Production Research*, 1-20. [CroosRef]

[11] Shahin, M., Hosseinzadeh, A., & Chen, F. F. (2025). Generative artificial intelligence in manufacturing: Applications, case studies, and future directions for next-generation intelligent production systems. The *International Journal of Advanced Manufacturing Technology, 141*(3-4), 1159-1265. [CroosRef]

[12] Mata, O., Ponce, P., Perez, C., & Ramirez, M. (2026). Digital twin designs with generative AI: Crafting a comprehensive framework for manufacturing systems. *Journal of Intelligent Manufacturing, 37*, 1049-1072. [CroosRef]

[13] Jignasu, A., Marshall, K., Ganapathysubramanian, B., Balu, A., Hegde, C., & Krishnamurthy, A. (2023). Towards foundational AI models for additive manufacturing: Language models for G-code debugging, manipulation, and comprehension. *arXiv*. Preprint. doi: 10.48550/arXiv.2309.02465 [CroosRef]

[14] Šket, K., Potočnik, D., Brezočnik, M., & Ficko, M. (2025). Large language models for G-code generation in CNC machining: A comparison of ChatGPT-3.5 and ChatGPT-4o. *Advances in Production Engineering & Management, 20*(2), 224-238. [CroosRef]

[15] Abdelaal, M., Lokadjaja, S., & Engert, G. (2025). GLLM: Self-corrective G-code generation using large language models with user feedback. *arXiv*. Preprint. doi: 10.48550/arXiv.2501.17584

[16] Kanimozhi, S., & Sriker, Y. S. (2024). Explorative deployment of fine-tuned large language model for on-site computerized numeric control machine operator assistance. *2024 IEEE Silchar Subsection Conference (SILCON)*. IEEE.

[17] Jeon, J., Sim, Y., Lee, H., Han, C., Yun, D., & Kim, E. (2026). CNC-talks: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. *Journal of Manufacturing Systems, 85*, 678-688. [CroosRef]

[18] Stathatos, E., Benardos, P., & Vosniakos, G. C. (2026). Large language models for high-level computer-aided process planning in a distributed manufacturing paradigm. *Robotics and Computer-Integrated Manufacturing, 100*, 103233. [CroosRef]

[19] Smid, P. (2003). *CNC programming handbook: A comprehensive guide to practical CNC programming*. Industrial Press.

[20] Stephenson, D. A., & Agapiou, J. S. (2016). *Metal cutting theory and practice* (3rd ed.). CRC Press. [CroosRef]

[21] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, A., Schulman, J., Hilton, J., Kelton, F., Miller, K., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. Preprint. doi: 10.48550/arXiv.2203.02155

[22] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv*. Preprint. doi: 10.48550/arXiv.2109.01652

[23] Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, & Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv*. Preprint. doi: 10.48550/arXiv.2202.12837

[24] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. Preprint. doi: 10.48550/arXiv.2201.11903

[25] Siemens Digital Industries Software. (n.d.). *NX CAM software: NX for manufacturing*. Siemens Digital Industries Software. https://www.siemens.com/en-us/products/nx-manufacturing/cam-software Accessed on Jun 03, 2026.

[26] Siemens AG. (n.d.). *Run MyVirtual Machine. Retrieved 2024, from Siemens AG*. https://www.dex.siemens.com/industrialsoftware/machine-tool-software/create-myvirtual-machine-operate Accessed on Jun 03, 2026.

[27] Anthropic. (2025, February 24). *Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-7-sonnet Accessed on Jun 03, 2026

[28] DeepSeek. (2025, January 20). *DeepSeek-R1*. https://api-docs.deepseek.com/news/news250120 Accessed on Jun 03, 2026

[29] OpenAI. (2025, February 27). *GPT 4.5*. https://openai.com/index/introducing-gpt-4-5 Accessed on Jun 03, 2026

[30] Campean, E., & Pop, G. (2026). CNC milling optimization via intelligent algorithms: An AI-based methodology. *Machines, 14*(1), 89. [CroosRef]

[31] Aghaei, S., & Ansari, F. (2026). Foundation language models through the lens of manufacturing. *Production & Manufacturing Research, 14*(1), 2632468. [CroosRef]

[32] Darshan, S., Desai, K. A., & Bhattacharyya, A. (2025). Physics-informed experimental design for neural network-based cutting force model in end milling of carbon fiber reinforced polymer composites. *Journal of Manufacturing Processes, 42*, 440-452. [CroosRef]

[33] Zhang, Q., Zhang, Q., Zhao, Y., Liu, Y., Wang, Z., & Ma, Y. (2025). Inverse solution of process parameters in gear grinding using hierarchical Bayesian physics informed neural network (HBPINN). *Scientific Reports, 15*, 18005. [CroosRef]