**FEDERAL UNIVERSITY OF ITAJUBÁ**
**INSTITUTE OF PRODUCTION ENGINEERING AND MANAGEMENT**

# IN-CONTEXT LEARNING VERSUS TRAINED MODELS: GENERATIVE AI AND MACHINE LEARNING FOR SURFACE ROUGHNESS PREDICTION IN DRY Ti-6Al-4V TURNING

ALEX FERNANDES DE SOUZA

ITAJUBÁ – MG
2026

ALEX FERNANDES DE SOUZA

# IN-CONTEXT LEARNING VERSUS TRAINED MODELS: GENERATIVE AI AND MACHINE LEARNING FOR SURFACE ROUGHNESS PREDICTION IN DRY Ti-6Al-4V TURNING

Thesis submitted to the Federal University of Itajubá as a requirement for obtaining the degree of Doctor in Production Engineering.

Advisor: Dr. Pedro Paulo Balestrassi
Co-Advisor: Dr. Filipe Alves Neto Verri

ITAJUBÁ – MG
2026

ALEX FERNANDES DE SOUZA

# IN-CONTEXT LEARNING VERSUS TRAINED MODELS: GENERATIVE AI AND MACHINE LEARNING FOR SURFACE ROUGHNESS PREDICTION IN DRY Ti-6Al-4V TURNING

Thesis submitted to the Federal University of Itajubá as a requirement for obtaining the degree of Doctor in Production Engineering.

**Examining Committee**

---

Prof. Dr. Pedro Paulo Balestrassi
(Advisor – Committee Chair)
Universidade Federal de Itajubá

---

Prof. Dr. Filipe Alves Neto Verri
Instituto Tecnológico de Aeronáutica

---

Prof. Dr. Matheus Brendon Francisco
Instituto Tecnológico de Aeronáutica

---

Prof. Dr. Paulo Henrique da Silva Campos
Universidade Federal de Itajubá

---

Prof. Dr. Ana Carolina Lorena
Instituto Tecnológico de Aeronáutica

---

Prof. Dr. Luiz Célio Souza Rocha
Instituto Federal de Educação Ciência e Tecnologia do Norte de Minas Gerais

ITAJUBÁ – MG
2026

**Acknowledgements**

I would like to express my sincere gratitude to my family for their unconditional support throughout this journey, whose encouragement and presence were fundamental to the completion of this work. I am grateful to my advisor, Prof. Dr. Pedro Paulo Balestrassi, for his guidance, availability, and the trust placed in this research. I also thank my co-advisor at ITA, Prof. Dr. Filipe Alves Neto Verri, for his availability and contributions that enriched this work. I extend my appreciation to Prof. Dr. Paulo Henrique da Silva Campos for his partnership in the laboratory manufacturing process, without which the experimental foundation of this study would not have been possible. I would also like to thank the remaining members of the examination committee, Prof. Dr. Matheus Brendon Francisco, Prof. Dr. Ana Carolina Lorena, and Prof. Dr. Luiz Célio Souza Rocha, for kindly accepting the invitation to evaluate this work and for their availability and valuable contributions. To my colleagues who shared knowledge, challenges, and memorable moments throughout this journey, I offer my sincere thanks for the companionship and mutual support. Finally, I am grateful to CAPES for the financial support provided through the research scholarship, which made this work possible.

**Abstract**

Surface roughness Ra is a critical quality indicator in dry turning of Ti-6Al-4V titanium alloy, a material widely used in aerospace and biomedical applications. Empirical power-law models remain the industrial standard but cannot represent interaction effects or non-linear responses. Machine learning offers greater flexibility, yet its deployment is constrained by small experimental datasets and the need for rigorous validation protocols. Large language models represent an alternative predictive paradigm, potentially encoding implicit knowledge of machining physics that may support quantitative prediction through in-context learning alone. This work compares five predictive models, comprising two conventional mathematical approaches (Power Law and RSM) and three machine learning models (SVR, Gaussian Process Regression, and XGBoost), against three LLMs (Claude Sonnet 4.5, GPT-4o-mini, and Gemini 3 Flash Preview) for Ra prediction in Ti-6Al-4V turning. A Central Composite Design with $k = 3$ factors and two carbide inserts of distinct nose radii yielded $N = 38$ observations. The conventional and ML models were evaluated under nested LOOCV with 5-fold grid search where applicable. LLMs were accessed via API at temperature zero under two prompting strategies, few-shot and chain-of-thought, across three independent runs, with results reported as mean ± standard deviation. Claude Sonnet 4.5 in few-shot mode achieved the best overall performance (RMSE $= 0.288 \pm 0.016$ $\mu$m, $R^2 = 0.985$), outperforming SVR, the best trained model (RMSE $= 0.341$ $\mu$m), by 16% without task-specific training or fine-tuning. The effect of prompting strategy was model dependent: few-shot prompting outperformed chain-of-thought for Claude and GPT, whereas Gemini CoT achieved a substantially lower mean RMSE than Gemini FewShot but exhibited markedly greater run-to-run variability. This variability was especially pronounced for Gemini CoT, whose RMSE ranged from 0.31 to 0.99 $\mu$m across identical executions. The results show that a capable LLM supplied with well-structured experimental demonstrations can surpass optimised trained models for Ra prediction without task-specific training, while also demonstrating that predictive performance and stability depend strongly on both model architecture and prompting strategy.

**Keywords**: Surface roughness prediction; Ti-6Al-4V turning; Machine learning; Large language models; In-context learning; Few-shot prompting; Cross-validation.

## List of Figures

## List of Tables

## List of abbreviations and acronyms

| | |
|---|---|
| API | Application Programming Interface |
| CCD | Central Composite Design |
| CNC | Computer Numerical Control |
| CoT | Chain-of-Thought |
| FS | Few-Shot |
| GPR | Gaussian Process Regression |
| LLM | Large Language Model |
| LOOCV | Leave-One-Out Cross-Validation |
| MAPE | Mean Absolute Percentage Error |
| ML | Machine Learning |
| RMSE | Root Mean Squared Error |
| RSM | Response Surface Methodology |
| SVR | Support Vector Regression |
| XGBoost | Extreme Gradient Boosting |

## List of symbols

| | |
|---|---|
| $V_c$ | Cutting speed |
| $f$ | Feed rate |
| $a_p$ | Depth of cut |
| $r_\varepsilon$ | Tool nose radius |
| $Ra$ | Arithmetic mean surface roughness |
| $Ra_{\text{th}}$ | Theoretical kinematic roughness |
| $C$ | Regularisation parameter (SVR) |
| $\varepsilon$ | Epsilon-insensitive margin (SVR) |
| $\gamma$ | RBF kernel bandwidth (SVR) |
| $\alpha$ | Ridge regularisation strength (RSM) |
| $\ell_j$ | Per-feature length scale (GPR) |
| $\sigma_f^2$ | Signal variance (GPR) |
| $\sigma_n^2$ | Noise variance (GPR) |
| $\sigma(\mathbf{x})$ | Predictive uncertainty estimate (GPR) |
| $\eta$ | Learning rate (XGBoost) |
| $N$ | Total number of experimental observations |
| $n$ | Number of valid predictions |
| $k$ | Number of input factors (CCD) |
| $\alpha_{\text{CCD}}$ | Axial distance parameter (CCD) |
| $\hat{Ra}_i$ | Predicted surface roughness for observation $i$ |
| $\bar{Ra}$ | Sample mean of measured surface roughness |

Contents

# 1 INTRODUCTION

Titanium alloy Ti–6Al–4V has become the default structural material for a remarkable range of demanding applications. Its alpha-beta microstructure delivers tensile strength above 900 MPa at a density roughly 60% that of steel, and its biocompatibility survives the chemical environment of the human body well enough for load-bearing orthopaedic implants (Srivastava *et al.*, 2024; Suresh; Ramesh; Srinath, 2023). Fan blades, turbine discs, femoral stems, and spinal fusion cages are among the components where the same material faces radically different loading conditions yet consistently emerges as the engineering choice (Marin; Lanzutti, 2024; Zhang, C. *et al.*, 2023). What links these applications is not a single outstanding property but the absence of any obviously better alternative: no common alloy combines its strength, weight, and corrosion resistance in quite the same package.

Machining this alloy is another matter. Low thermal conductivity means heat accumulates at the cutting zone instead of leaving with the chip; high chemical affinity with most tool materials means the edge degrades through mechanisms that are not merely mechanical but thermochemical (Ezugwu; Wang, Z.-M., 1997). Tool life is short, cutting temperatures are high, and the transition from controlled wear to sudden tool failure can happen within a few metres of cutting length (Li, Z.; Zhang, P.; Zhou, *et al.*, 2026; Yi *et al.*, 2025). A machinist working with Ti–6Al–4V is essentially managing a process that operates near its own limits most of the time, and getting the surface finish right under those conditions is not straightforward.

The arithmetic mean surface roughness, $Ra$, matters here in ways that go beyond aesthetics. For implantable components, it sits at the intersection of biology and mechanics: surfaces that are too smooth prevent cell adhesion, those that are too rough accumulate biofilm and create stress concentrations at the implant-bone interface (Marin; Lanzutti, 2024; Chowdhury; Arunachalam, 2023). The acceptable $Ra$ window for osseointegrated devices is narrow, and missing it in either direction has clinical consequences. In aerospace, the relationship is more mechanical but equally unforgiving: surface asperities concentrate stress and initiate fatigue cracks, and the literature documents roughness-driven reductions in fatigue life by orders of magnitude in high-cycle loading scenarios (Hu, Y. *et al.*, 2024; Li, W. *et al.*, 2025). Predicting $Ra$ before the first cut is, in both contexts, less about production efficiency than about building in quality from the process plan forward.

The classical tool for this prediction is a power-law regression fitted to experimental data. The model expresses $Ra$ as a product of cutting speed, feed rate, depth of cut, and nose radius, each raised to a calibrated exponent, and it remains the dominant approach in industrial practice because it is interpretable, simple to use, and easy to explain to a process engineer (Wang, X.; Feng, 2002; Butt; Haq; KVVSSN, *et al.*, 2026). Central Composite

Design (CCD) and Response Surface Methodology (RSM) provide the experimental and modelling framework for building predictive relationships with a controlled number of runs (Montgomery, 2017). The problem is structural: a multiplicative form cannot represent interactions between variables, and the model breaks down anywhere outside the region where it was calibrated (Benardos; Vosniakos, 2003). That is a significant limitation when cutting conditions vary and process windows shift.

Machine Learning (ML) addressed this limitation directly. Support Vector Regression (SVR), Gaussian Process Regression (GPR), and gradient-boosted tree models can capture the interaction effects and non-linearities that fixed functional forms miss, and comparative studies show consistent improvements over power-law baselines in these regimes (Das *et al.*, 2022; Alemayehu *et al.*, 2025; Kosarac *et al.*, 2023). The gain is real, but it comes with practical requirements: enough experimental data to train the model, computational infrastructure for hyperparameter optimisation, and disciplined validation to avoid overfitting a small dataset from a designed experiment (Cawley; Talbot, 2010; Arlot; Celisse, 2010). In a machining context where each observation costs time, material, and tooling, accumulating the data needed for a reliable ML model is not always feasible (Deng *et al.*, 2023).

Large Language Models (LLMs) enter this picture from an unexpected direction. Generative Pre-trained Transformer (GPT) models, Claude, and Gemini were trained on vast corpora of scientific and technical text that include machining literature, materials handbooks, and process engineering references (Li, Y. *et al.*, 2024). That training does not produce explicit knowledge of specific experiments, but it does encode statistical associations between cutting parameters, material properties, and surface outcomes that may support prediction without any formal fitting procedure. Two prompting strategies are of particular interest here. Few-shot prompting supplies the model with experimental observations directly in the prompt, allowing it to perform in-context regression without updating its weights (Brown *et al.*, 2020; Dong *et al.*, 2024). Chain-of-thought prompting instead asks the model to reason through the underlying physics before producing a number, a strategy that improves performance on symbolic and mathematical tasks but whose behaviour on continuous physical quantities has not been systematically studied (Wei *et al.*, 2022b; Sprague *et al.*, 2025). Compounding the question is the run-to-run variability of LLM outputs, which is rarely quantified in engineering evaluations but can be substantial (Angermeir *et al.*, 2026; Liang *et al.*, 2023).

What is actually being asked in this work is whether a model that has never seen a lathe can predict surface roughness as well as one that was specifically trained on turning data. The question has practical stakes: process planning workflows increasingly need predictive models at stages of product development where experimental data do not yet exist, and a language model that carries useful quantitative knowledge from its pre-training represents a qualitatively different kind of resource from a conventional regressor (Makatura *et al.*,

2024; Hoang *et al.*, 2025). Whether that resource is reliable enough for manufacturing quality prediction, and under what prompting conditions it fails, is what this study sets out to establish.

## 1.1 Research Problem

The prediction of surface roughness *Ra* in titanium alloy turning processes still relies, in industrial practice, on empirical and mathematical models fitted to experimental data and generally limited to the range of conditions under which they were developed. Machine Learning (ML) models offer greater flexibility for representing non-linear relationships and interactions among process variables, but introduce additional requirements in terms of data availability, hyperparameter optimisation, and rigorous validation. These requirements can be difficult to satisfy in machining applications, where the number of experiments that can be economically conducted is often limited.

Large Language Models (LLMs) represent a different predictive paradigm because they are not explicitly trained on the experimental dataset used for the target problem. Instead, they bring information encoded during pre-training on large collections of technical and scientific text. Recent studies indicate that the use of LLMs in manufacturing remains limited, particularly for quantitative prediction tasks, as discussed by Ko (2026). In this context, Khanghah *et al.* (2025) show that new frameworks are being developed to apply LLMs to extrapolative modelling of manufacturing processes. Nevertheless, their capacity to operate as quantitative predictors in manufacturing engineering remains insufficiently investigated. Existing studies rarely provide systematic comparisons with conventional mathematical and properly optimised ML models, and the stability of LLM predictions across independent runs is also seldom evaluated.

Against this background, the research problem addressed in this work is formulated as follows: Is it possible to use Large Language Models as predictors of surface roughness *Ra* in Ti–6Al–4V turning, and to what extent does their predictive performance compare with that of conventional mathematical and Machine Learning models evaluated under a common cross-validation framework, considering predictive accuracy, run-to-run stability, and industrial applicability?

## 1.2 Objectives

### 1.2.1 General Objective

To compare the predictive performance of conventional mathematical models, ML models, and LLMs in estimating surface roughness *Ra* in Ti–6Al–4V turning under a common Leave-One-Out Cross-Validation (LOOCV) framework, considering predictive accuracy, stability, and practical applicability.

### 1.2.2 Specific Objectives

- To fit and evaluate five predictive models, comprising two conventional mathematical models (Power Law and second-order RSM) and three ML models (SVR, GPR, and Extreme Gradient Boosting (XGBoost)), using nested cross-validation and hyperparameter optimisation where applicable;
- To evaluate three LLMs (Claude, GPT, and Gemini) under two prompting strategies (few-shot and chain-of-thought) as *Ra* predictors without explicit training, through official Application Programming Interface (API) access with temperature set to zero;
- To quantify the run-to-run variability of LLM predictions and report results as mean ± standard deviation across three independent executions, establishing a reproducible evaluation protocol;
- To compare the evaluated predictive approaches in terms of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination ($R^2$), Mean Absolute Percentage Error (MAPE), and stability, and to discuss the practical implications of the findings for industrial applications of surface quality prediction.

## 1.3 Thesis Structure

This work is organised into five chapters. Chapter 2 presents the literature review on Ti-6Al-4V machining, Ra predictive models, and recent applications of machine learning and language models in manufacturing processes. Chapter 3 describes the experimental and computational methodology, including the experimental design, cutting conditions, models evaluated, and the validation protocol adopted. Chapter 4 presents and discusses the results, structured around exploratory analysis, ML model performance, LLM model performance, and a unified comparison between both groups. Chapter 5 presents the conclusions and directions for future work.

# 2 LITERATURE REVIEW

## 2.1 Ti-6Al-4V Machinability and Surface Roughness in Turning

Ti-6Al-4V is difficult to machine for reasons that are rooted in its thermal and chemical behaviour at the cutting zone. Thermal conductivity is roughly one-sixth that of steel, so the heat generated during cutting accumulates at the tool-workpiece interface rather than dissipating through the chip. The alloy retains its strength at elevated temperatures, which means the material resists the plastic deformation that normally aids chip formation (Yao *et al.*, 2023). On top of this, a strong chemical affinity between titanium and most tool materials promotes adhesive wear: a built-up edge forms on the rake face, periodically fractures, and pulls fragments of the cutting edge with it. The combination of thermal loading and adhesion produces wear rates that would be considered severe in steel turning at equivalent cutting parameters (Ezugwu; Wang, Z.-M., 1997; Yi *et al.*, 2025).

These characteristics have a direct consequence for surface quality. In steels and aluminium alloys, the relationship between cutting parameters and Ra is relatively stable across a broad process window. In Ti-6Al-4V it is not: the transition from a stable cutting regime to one governed by chatter or adhesion can happen within a narrow band of cutting speeds, and tool wear progression changes the surface response in ways that are not easily predicted from first principles (Li, Z.; Zhang, P.; Zhou, *et al.*, 2026; Zhang, C. *et al.*, 2023). The process window for acceptable surface finish is narrower than in most engineering alloys, and missing it has real consequences, in implantable devices because surface texture governs osseointegration and fatigue resistance, and in aerospace structures because roughness-driven fatigue crack initiation operates at stress concentrations that even modest increases in Ra can render structurally significant (Marin; Lanzutti, 2024; Hu, Y. *et al.*, 2024).

Surface roughness in turning is quantified by the arithmetic mean deviation Ra, defined as the mean absolute deviation of the surface profile from its centreline over the evaluation length (Shaw, 2005). The kinematic minimum Ra achievable with a given tool geometry follows directly from the geometry of the residual groove left by the tool nose:

$$Ra_{\mathrm{th}} = \frac{f^2}{32\, r_\varepsilon} \tag{2.1}$$

where:

- $f$ is the feed rate;
- $r_\varepsilon$ is the nose radius;
- $Ra$ is the arithmetic mean roughness deviation.

This formula assumes a perfectly sharp and rigid tool cutting without vibration or material side-flow, conditions that are never fully satisfied. In Ti-6Al-4V turning, the departure from this ideal is particularly wide: thermal softening near the cutting edge, elastic springback of the workpiece, and adhesive material transfer all push the measured *Ra* above the geometric prediction, often substantially so at low feed rates, where the geometric contribution is small relative to these secondary effects.

Predicting Ra from cutting parameters has been approached empirically since the earliest systematic studies of metal cutting. The dominant framework relates Ra to cutting speed, feed rate, depth of cut, and nose radius through a multiplicative equation calibrated on experimental data, and it remains widely used in industry because it is compact and easy to interpret (Wang, X.; Feng, 2002; Diniz; Marcondes; Coppini, 2014). The exponents recovered by fitting this model to Ti-6Al-4V data consistently confirm what the kinematic formula already suggests: feed rate dominates, nose radius comes second, and the other two variables contribute little under finishing conditions (Kuntoğlu *et al.*, 2021). A structural limitation of this approach, however, is its inability to represent situations where the effect of one variable depends on the value of another, an issue that machine learning methods are proposed to overcome.

## 2.2 Machine Learning for Surface Roughness Prediction

Artificial neural networks were the first serious attempt to break out of the power-law straitjacket. Benardos and Vosniakos (2003) provided an early systematic review of their application to machining, and later work by Dubey, Sharma, and Pimenov (2022) confirmed that ANNs could handle non-linearities that multiplicative models simply ignored. The catch was predictable: neural networks thrive on data, and in precision machining, where every run consumes time, material, and tool life, the experimental budgets rarely stretch far enough to train them properly. Worse, the resulting models operated as black boxes. In a field where understanding *why* a surface turned out rough matters almost as much as predicting *that* it did, this opacity limited their traction on the shop floor.

As scientific computing libraries matured, attention splintered across a wider set of tools better calibrated to small datasets. Support vector regression, Gaussian processes, and gradient-boosted trees all offered different answers to the same core problem: how to capture the interactions between speed, feed, and tool geometry without overfitting a sparse design matrix. The influx of methods produced a sprawling comparative literature, yet much of it rests on shaky methodological ground. Maitra and Shi (2023) note a recurring weakness: studies evaluating different algorithms on identical Central Composite Design data often report minimal performance gaps, but those gaps are unreliable because the validation protocols differ.

A single train-test split, or even a non-nested cross-validation loop, leaks information from the hyperparameter search into the performance estimate (Cawley; Talbot, 2010).

With fewer than 50 observations, this leakage can easily flip the apparent ranking of two competing models, a problem that Vabalas *et al.* (2019) document specifically in small-sample machine learning evaluation and that Varoquaux (2018) trace to the large variance of cross-validation estimates when $n$ is limited. The result is a body of published work where it is genuinely difficult to tell whether Model A outperformed Model B because of algorithmic merit or because of an optimistic split in the data.

The five models examined in this work were chosen to cut across this methodological ambiguity. They span the full range of available strategies: an interpretable empirical baseline (Power Law) that ignores interactions by design; a second-order polynomial surface (RSM) that can approximate them locally; a support vector regressor (SVR) that relies on structural risk minimization to survive the small-sample regime; a Gaussian process (GPR) that self-calibrates its uncertainty alongside its predictions; and a gradient-boosted ensemble (XGBoost) whose feature importance scores offer a data-driven window into the underlying physics. Evaluating them all under the same nested cross-validation protocol, where the test point is quarantined from both training and hyperparameter tuning, provides a common benchmark that the existing literature largely lacks.

### 2.2.1 Power Law Regression

The power-law model traces its lineage to Taylor's tool life equation, which established the convention of expressing process responses as products of power functions of the cutting parameters. Applied to surface roughness by X. Wang and Feng (2002) in one of the foundational studies of this form, and surveyed broadly by Benardos and Vosniakos (2003), the model takes the form

$$Ra = C \cdot V_c^{\beta_1} \cdot f^{\beta_2} \cdot a_p^{\beta_3} \cdot r_\varepsilon^{\beta_4}, \tag{2.2}$$

where:

- $Ra$ is the arithmetic mean roughness deviation;
- $C$ is a process- and material-dependent constant;
- $V_c$ is the cutting speed;
- $f$ is the feed rate;
- $a_p$ is the depth of cut;
- $r_\varepsilon$ is the nose radius;
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the exponents that encode the sensitivity of $Ra$ to each variable.

Taking logarithms linearises the equation, allowing ordinary least squares to be applied directly for parameter estimation.

What makes this model useful is precisely what makes it limited. The exponents carry physical meaning: a value of $\beta_2 \approx 2$ for feed rate recovers the coefficient implied by the kinematic formula $Ra_{\text{th}} = f^2/(32\, r_\varepsilon)$, and Wey-Hwey Yang and Tarng (1998) confirm that the empirical fit reproduces this geometric relationship reliably across a range of Ti-6Al-4V turning conditions. When exponents deviate from their theoretical counterparts, that deviation is itself informative, it points toward mechanisms, such as adhesive wear or thermal softening, that the geometric model does not account for. In this sense, the power-law fit is as much a diagnostic as a predictor.

The limitation is structural rather than statistical. Because the four variables enter the model as independent multiplicative factors, the effect of feed rate on Ra cannot depend on the current value of nose radius, even though the kinematic formula shows that it does through the term $f^2/r_\varepsilon$. The interaction between cutting speed and feed rate in governing the thermal contribution to Ra is similarly inexpressible. Kosarac *et al.* (2023) make the point bluntly: collecting more data or refitting more carefully will not fix a misspecification that is built into the model form. What more data can do is reveal the extent of the misspecification more clearly, which is a useful but different thing.

Published Ti-6Al-4V power-law models show feed rate exponents ranging from 1.1 to 2.3 and cutting speed exponents near zero, a pattern that Wey-Hwey Yang and Tarng (1998) and Siva Surya (2022) find independently across different experimental designs and cutting tool geometries. The depth of cut exponent is routinely non-significant, consistent with $a_p$ having little direct geometric influence on the residual groove profile in finishing conditions. In the present study, the power-law model is not included as a straw man but as the representive of the method that most industrial users currently rely on, and against which any more complex approach must demonstrate a meaningful return on the added complexity.

### 2.2.2 Response Surface Methodology

Response surface methodology entered the machining literature as a natural complement to designed experiments. When a Central Composite Design is used to collect the data, the second-order polynomial expansion of the input space is the canonical choice for modelling the response, because the CCD is specifically constructed to support estimation of all linear, quadratic, and pairwise interaction terms, as Montgomery (2017) establishes in the foundational treatment of experimental design theory. This alignment between experimental design and model structure is one of the reasons RSM has remained prevalent in machining research for over four decades.

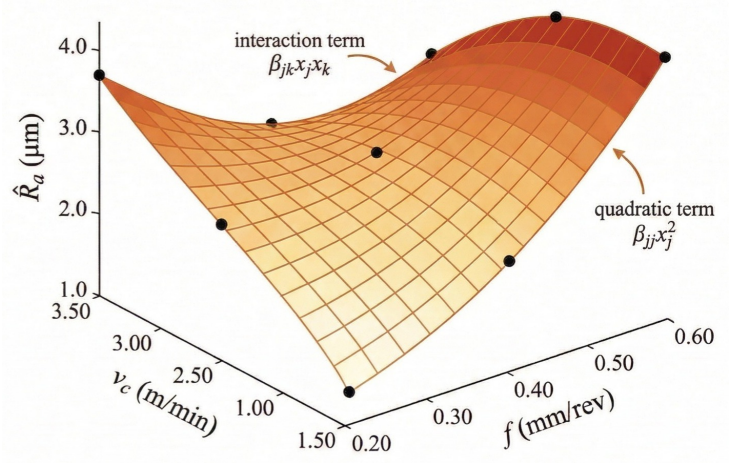
The RSM model is written in coded variables as

$$\hat{Ra} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \beta_{jj} x_j^2 + \sum_{j<k} \beta_{jk} x_j x_k + \varepsilon, \tag{2.3}$$

Where $x_j$ are the standardised input variables, $p = 4$ for the present study, and $\varepsilon$ is the residual error. The full second-order expansion produces 14 regressors (4 linear, 4 quadratic, and 6 pairwise interaction terms), which approaches the sample size of 37 available per LOO fold. Collinearity among the polynomial terms therefore becomes a concern, and ridge regression is applied to stabilise the estimates by penalising the sum of squared coefficients with a regularisation parameter $\alpha$, following the standard treatment described by Noor, Siddiqui, and Iqbal (2023) in the context of turning parameter optimisation under multicollinear conditions (Hoerl; Kennard, 1970).

The advantage of Equation 2.3 over the power-law model is its capacity to represent interaction effects. The term $\beta_{jk} x_j x_k$ allows the model to capture the fact that the influence of feed rate on Ra increases with decreasing nose radius, a relationship the power-law structure cannot express. This additional flexibility translates directly into lower prediction error when the true response surface has significant curvature or cross-variable dependence within the experimental domain. Figure 1 illustrates this geometry: the saddle-shaped surface arises from the quadratic terms $\beta_{jj} x_j^2$, and its asymmetry along the $v_c$ and $f$ axes reflects the interaction term $\beta_{jk} x_j x_k$, structural features that a multiplicative power-law surface cannot reproduce.

**Figure 1** – Second-order response surface as a function of cutting speed $v_c$ and feed rate $f$, with $a_p$ and $r_\varepsilon$ held at their centre-point values. The curvature arises from the quadratic terms $\beta_{jj} x_j^2$ and the asymmetry between axes reflects the interaction term $\beta_{jk} x_j x_k$, the structural feature absent from the power-law baseline. Experimental observations are shown as points (Montgomery, 2017).




In the machining literature, RSM has been successfully applied to Ra prediction in titanium turning by Siva Surya (2022), who consistently report $R^2$ values above 0.90 on training data, and by Grossmann *et al.* (2022), who extend the framework to cryogenic cutting conditions with comparable success.

The limitation of RSM relative to the kernel-based methods discussed below is that the polynomial basis is fixed by construction. If the true response surface deviates from the second-order polynomial form in a way that the chosen terms cannot capture, the model

will exhibit systematic bias in certain regions of the input space. In Ti-6Al-4V turning, the abrupt transitions in surface quality that occur near the boundary between stable cutting and chatter cannot be represented by a smooth polynomial surface, as documented by Ezugwu and Zheng-Min Wang (1997), and RSM models will underperform in datasets that span these transitions, a limitation noted by Du *et al.* (2024) in comparable turning experiments. Within the interior of a well-designed CCD, however, the second-order surface provides a reliable and interpretable approximation that is straightforward to communicate and validate.

### 2.2.3 Support Vector Regression

Support vector regression brings the large-margin framework of support vector machines into problems with continuous responses. Its formulation centres on a tolerance band of width $\varepsilon$: prediction errors that remain within this interval are not penalised, whereas deviations beyond it receive a linear penalty. Roy and Chakraborty (2023) provide a thorough account of the structural risk minimisation principle that supports both the classification and regression versions of the method. Since the fitted model is determined only by the training samples that fall on or beyond the $\varepsilon$-tube—the support vectors—it tends to avoid overfitting, even when the sample size is limited. In machining studies, where experiments are costly and datasets often contain only a few dozen observations, that characteristic is particularly useful.

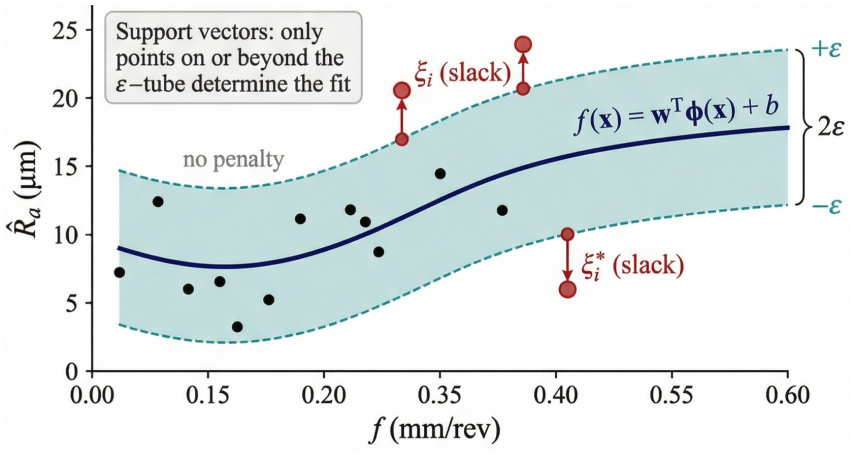
Mathematically, SVR minimises the regularised objective

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right), \tag{2.4}$$

subject to the constraints $|Ra_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b| \leq \varepsilon + \xi_i$, where $\phi(\mathbf{x})$ is a feature mapping induced by the chosen kernel, $C$ governs the penalty for violations, and $\xi_i$, $\xi_i^*$ are slack variables that absorb points falling outside the $\varepsilon$-tube. When the ordinary dot product is replaced by a radial basis function kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, the model can represent non-linear relations without requiring an explicit functional form. The parameter $\gamma$ determines the degree of locality of the kernel, and its calibration is decisive for controlling the bias-variance trade-off. Figure 2 shows what this geometry looks like in practice when $Ra$ is plotted against feed rate.

A practical way to read Figure 2 is to think of the shaded band as a corridor of acceptable error. Any experimental measurement that falls inside that corridor is treated as if the model had predicted it exactly, it contributes nothing to the fitting objective. Only the red points outside the corridor matter, and the model is adjusted just enough to bring their violations within the penalty budget set by $C$. The result is a predictor that is deliberately blind to small fluctuations in the data, which is precisely the behaviour needed when experimental measurements carry their own noise from profilometer positioning and chip re-adhesion. The width of the corridor, $2\varepsilon$, and the stiffness of the penalty, $C$,

**Figure 2** – Geometric interpretation of SVR with an RBF kernel fitted to Ra as a function of feed rate $f$. The shaded $\varepsilon$-tube defines a margin of tolerance around the regression function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$; observations inside the tube incur no penalty. Only the support vectors on or beyond the tube boundary, marked in red with their slack variables $\xi_i$ and $\xi_i^*$, determine the fitted model




together determine how the model balances smoothness against fidelity to the observations, a balance that nested cross-validation sets automatically for each outer fold.

Empirical studies on surface roughness have repeatedly shown that SVR with an RBF kernel can surpass both classical regression models and early neural-network approaches under limited-data conditions. Çaydaş and Ekici (2012) demonstrated this in the hard turning of AISI 304 stainless steel, linking the result to SVR's explicit concern with generalisation error rather than with minimising residuals on the training data. Arnaiz-González, Fernández-Valdivielso, Bustillo, *et al.* (2016) later confirmed the same pattern for Ti-6Al-4V machining, reporting that SVR performed better than neural networks and regression trees on CCD-based datasets with a size comparable to that of the present study.

The advantage appears consistent across materials and experimental designs. Bagga, Patel, Makhesana, *et al.* (2023) found SVR more stable than tree-based ensembles in the prediction of tool life from a limited set of turning experiments, and attributed the difference to the same mechanism: the $\varepsilon$-insensitive loss produces a solution that is anchored to a small subset of the training points and is therefore less sensitive to the particular observations that happen to fall in the training fold.

A central practical difficulty in SVR is the interaction among its three tuneable parameters: $C$, $\varepsilon$, and $\gamma$. A high value of $C$ pushes the model to fit the training observations very closely, while a small $\varepsilon$ shrinks the indifference tube and imposes a stricter fit(Çaydaş; Ekici, 2012). In combination, these choices can generate a highly irregular response function that does little more than reproduce noise. At the opposite extreme, lowering $C$ or widening $\varepsilon$ produces a smoother surface, though with the possibility of overlooking real physical behaviour. The kernel bandwidth $\gamma$ introduces an additional source of sensitivity: if it is

too small, the model becomes overly local; if too large, relevant detail is washed out (Roy; Chakraborty, 2023).

With only 38 observations available in this work, exploring this three-parameter space through conventional split-based tuning is a weak strategy. Maitra and Shi (2023) argue that, in small-sample settings of this kind, the only methodologically sound approach is to embed the hyperparameter search within a nested cross-validation loop, ensuring that the test point has no influence at all on the selection of $C$, $\varepsilon$, or $\gamma$.

For the present study, SVR offers an uncommon combination of properties: strong empirical performance on small machining datasets and enough flexibility to represent the non-linear dependence of $Ra$ on the joint effect of feed rate and nose radius. Its kernel-based formulation can capture the multiplicative interaction $f^2/r_\varepsilon$ without requiring that term to be imposed beforehand. That said, this advantage has a clear cost in interpretability. Unlike a power-law model or a second-order polynomial surface, a fitted SVR does not return a compact equation with coefficients that can be directly examined for physical consistency. Identifying which variables are driving the predictions depends on additional post-hoc analysis, and that concession only makes sense when the improvement in predictive accuracy compensates for the loss of immediate interpretability.

### 2.2.4 Gaussian Process Regression

Most regression models search for a single best-fit function and return a number. Gaussian process regression takes a different route: instead of committing to one function, it maintains a distribution over all functions that are consistent with the observed data (Rasmussen; Williams, 2006). When a new input arrives, the model does not just predict a value, it returns a mean and a variance, where the variance reflects how uncertain the model is about that particular region of the input space (Wang, J., 2023). In manufacturing, that uncertainty estimate is not a statistical curiosity; it tells the process engineer whether a predicted Ra sits comfortably inside the specification window or whether the model is extrapolating into territory it has not seen before (Maitra; Shi, 2023).

In geostatistics, the same framework has been called Kriging for decades. The mathematical foundations that underpin both names are developed in Rasmussen and Williams (2006), whose treatment of covariance functions and marginal likelihood optimisation remains the standard reference for nearly all current implementations. Jie Wang (2023) provide a more accessible entry point for readers coming from an engineering background rather than a statistical one.

The method has been extended considerably since its origins. Lyu, X. Liu, and Mihaylova (2024) survey the variants developed over the past decade for high-dimensional and multi-fidelity problems, while Jones, Schonlau, and Welch (1998) examined its compatibility with designed experiments specifically, noting that the structured spacing of CCD designs aligns naturally with the spatial correlation assumptions that GPR relies on. The prior

over functions is encoded through a covariance kernel, which determines how the model expects the function value at one point to covary with the value at another. In this work a squared-exponential kernel equipped with automatic relevance determination (ARD) is adopted:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left( -\frac{1}{2} \sum_{j=1}^{4} \frac{(x_j - x_j')^2}{\ell_j^2} \right) + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{2.5}$$

Where $\ell_j$ is a separate length scale for each input dimension, $\sigma_f^2$ controls the overall signal variance, and $\sigma_n^2$ accounts for observation noise. The per-variable length scales are the key to ARD: rather than treating all inputs as equally relevant, the model learns from the data itself how sensitive Ra is to each variable. In practice, this means GPR can discover on its own that feed rate matters far more than cutting speed, without the user having to encode that knowledge in advance (Rasmussen; Williams, 2006). Arnaiz-González, Fernández-Valdivielso, Bustillo, *et al.* (2016) confirmed this behaviour in a direct comparison on Ti-6Al-4V CCD data, where the ARD length scales recovered the correct variable ranking without any manual feature weighting.

This automatic calibration matters because it removes a potential source of modelling bias. A fixed, isotropic kernel assumes all inputs contribute equally to the covariance structure, which in a machining dataset with four variables of very different physical relevance is a poor assumption. Souza *et al.* (2026) reached a complementary conclusion from a different direction: when physics-based feature engineering was applied explicitly, the resulting model converged on the same dominant relationships that ARD had already identified implicitly, suggesting the two approaches are encoding the same underlying process information.

All kernel hyperparameters are tuned internally by maximising the marginal likelihood of the training data. This is empirical Bayes in action: the optimisation routine automatically balances model fit against model complexity, guided by a criterion that integrates over the entire parameter space, as Rasmussen and Williams (2006) derive in detail. Consequently, GPR does not require an external grid search or cross-validation loop for hyperparameter selection, a feature that sets it apart from SVR and XGBoost and that identify as particularly valuable when the experimental dataset is small and every observation withheld for validation represents a meaningful loss of training information. L. Hu, Gao, and Guo (2023) exploited exactly this property when developing a transfer-learning extension of GPR for surface roughness prediction, noting that the marginal likelihood provides a principled and computationally efficient way to set the model's effective complexity even when data are scarce.

Tognan, Laurenti, and Salvati (2022) further corroborate the suitability of GPR for small machining datasets, reporting robust predictions of cutting temperature with limited experimental runs and observing that the marginal likelihood consistently identified length

scales that matched the physical understanding of which parameters dominated the process response. Çaydaş and Ekici (2012), in a study that compared SVR and GPR directly in a hard-turning context, found that GPR matched or exceeded SVR accuracy while additionally providing the uncertainty bands that SVR cannot supply.

For a manufacturing engineer, however, the most distinctive output of a GPR model is not the point prediction but the accompanying uncertainty estimate. A prediction of $Ra = 1.8\,\mu$m accompanied by a 95% credible interval of $[1.4,\ 2.2]\,\mu$m tells a richer story than a bare number: it indicates whether the predicted roughness sits comfortably inside the specification window or whether the process capability is marginal. This information is especially valuable at low feed rates, where the empirical correction factor between the theoretical and measured Ra fluctuates widest, as Section 4.1 documents.

Maitra and Shi (2023) have argued that GPR is industry-ready for Ti-6Al-4V surface roughness prediction largely on the strength of these uncertainty bands, reporting root-mean-squared errors that match neural networks at a fraction of the training cost. The ability to flag predictions with high uncertainty is a practical safeguard that deterministic models cannot offer, and Cawley and Talbot (2010) point to exactly this kind of honest uncertainty reporting as the distinguishing feature of Bayesian methods in settings where overfitting risk is high. In safety-critical manufacturing the conversation shifts from asking what the predicted value is to asking how much the prediction should be trusted, and only GPR among the models considered here supports that shift directly.

GPR is the only model in this study that returns a calibrated uncertainty estimate alongside every prediction. That variance is not a byproduct of the method, it is the output that makes GPR useful in quality-critical decisions, where knowing how confident the model is matters as much as knowing what it predicts (Maitra; Shi, 2023). The kernel parameters are optimised internally by maximising the marginal likelihood of the training data, which means no external validation set is needed for hyperparameter selection (Rasmussen; Williams, 2006). This is a meaningful practical advantage when the dataset has 38 observations and every point withheld for validation represents a real loss of training information (Cawley; Talbot, 2010).

The learned length scales $\ell_j$ add a secondary benefit: they provide a direct, data-driven ranking of how much each machining parameter contributes to the surface roughness response. Arnaiz-González, Fernández-Valdivielso, Bustillo, *et al.* (2016) exploit this in their Ti-6Al-4V analysis, where the ARD length scales recovered the expected variable ranking without any manual guidance. The Bayesian formulation also inherently resists overfitting in small samples, since the marginal likelihood penalises model complexity alongside training fit (Cawley; Talbot, 2010).

The one limitation worth acknowledging is computational: the standard GPR implementation scales as $\mathcal{O}(n^3)$ with the number of training points, which becomes prohibitive for large industrial datasets. Lyu, X. Liu, and Mihaylova (2024) discuss sparse approximation

schemes that address this for larger problems. For $N = 38$, the cost is negligible. GPR earns its place in the comparison not because it is the most accurate model on this dataset but because it is the only one that pairs competitive point predictions with a principled uncertainty measure, a combination that Maitra and Shi (2023) and L. Hu, Gao, and Guo (2023) independently identify as the relevant criterion for manufacturing deployment.

### 2.2.5 XGBoost

XGBoost is a gradient-boosted tree ensemble algorithm that constructs its predictor in an iterative manner. Starting from a constant prediction, it adds regression trees sequentially, each one fitted to the negative gradient of the loss function with respect to the current ensemble, as Shangpan Zhang *et al.* (2026) describe in their application to titanium alloy mechanical property prediction. The final model is an additive combination of $M$ trees, each introducing a small correction to the prediction produced by the previous ones:

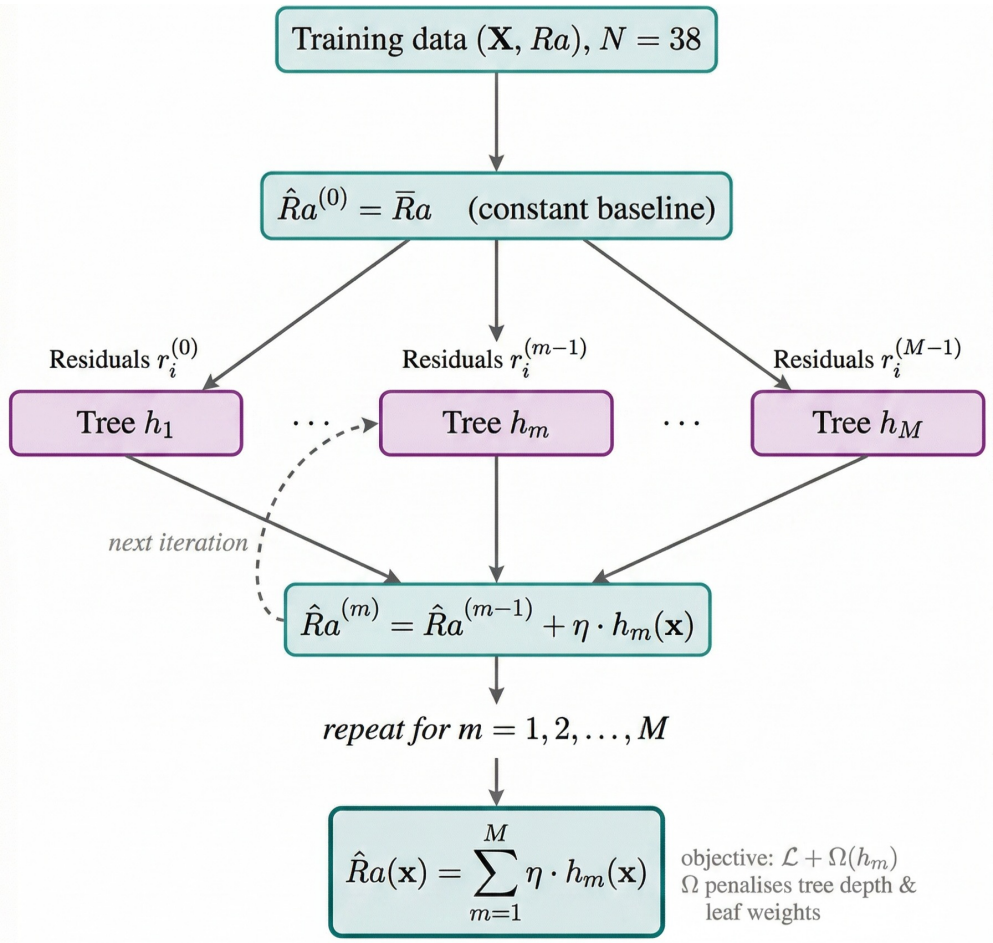
$$\hat{Ra}(\mathbf{x}) = \sum_{m=1}^{M} \eta \cdot h_m(\mathbf{x}), \tag{2.6}$$

Where $h_m$ is a regression tree of bounded depth $d$ and $\eta$ is the learning rate that controls the contribution of each successive step. The regularised objective minimised at each stage includes both the training loss and a complexity penalty on the tree structure. That penalty limits excessive tree growth, which is the regularisation mechanism that Das *et al.* (2022) identify as the main distinction between gradient-boosted trees and earlier ensemble methods in hard turning applications.

XGBoost has become a common baseline in applied machine learning due to its empirical performance across many tabular prediction problems. In the machining literature, Cica *et al.* (2026) apply it to energy consumption prediction in sustainable Ti-6Al-4V milling, where it generally performs well when the training set contains a few hundred observations or more, but shows greater susceptibility to overfitting in the small-sample setting typical of CCD-based experiments. Tree depth is the key hyperparameter in this regard: a tree of depth 2 yields a model with at most four leaf nodes, which is constrained enough to generalise from 37 training observations, whereas a tree of depth 4 can reproduce the training data almost exactly when $n$ is small.

Another parameter interaction that directly affects model behaviour is the relationship between learning rate and the number of boosting rounds. A small $\eta$ reduces the contribution of each individual tree and usually demands a larger ensemble to reach competitive training loss, whereas a larger $\eta$ accelerates fitting but raises the chance that early trees impose corrections that are too specific to a limited dataset. In small-sample conditions, this trade-off must be treated carefully because adding many trees does not automatically improve generalisation when the available observations cover only a narrow experimental region (Zhang, S. *et al.*, 2026; Das *et al.*, 2022).

**Figure 3** – Gradient-boosted tree ensemble (XGBoost). Starting from a constant baseline $\hat{Ra}^{(0)}$, each iteration fits a shallow regression tree $h_m$ to the residuals of the current ensemble and adds it at step size $\eta$. The regularisation term $\Omega$ in the objective function penalises tree depth and leaf weights, preventing overfitting when the training set is small.




XGBoost assigns each input variable an importance score based on the total gain produced by the splits that use that variable across all trees in the ensemble Tianqi Chen and Guestrin (2016), yielding a data-driven estimate of the relative influence of each cutting parameter on *Ra* that does not depend on the physical model. In small-sample turning experiments, Bagga, Patel, Makhesana, *et al.* (2023) showed that this decomposition can recover the same variable rankings suggested by process knowledge, lending it a degree of interpretive credibility beyond mere predictive utility. In the present study, the decomposition may either support or question the relationships suggested by the theoretical formula and the Pearson correlations of Section 4.1; agreement between these independent sources of evidence would amount to a form of cross-validation of the dataset itself.

A further limitation of XGBoost in this problem is that its piecewise constant tree structure does not reflect the continuous physical mechanisms that govern surface roughness formation in turning. Although an additive tree ensemble can approximate smooth response

surfaces with good accuracy when sufficient data are available, that approximation is still built from local partitioning rules rather than from a global functional relation. For that reason, even a strong predictive result would not by itself indicate that the model represents the machining physics more faithfully than simpler parametric or kernel-based approaches, a point that is consistent with the comparative discussion reported by Kosarac *et al.* (2023), Bagga, Patel, Makhesana, *et al.* (2023) and Reeber, Wolf, and Möhring (2024).

Several authors have reported that gradient-boosted trees tend to perform worse than kernel-based methods such as SVR and GPR when the response surface is smooth and the training set is small, a pattern that Kosarac *et al.* (2023), Bagga, Patel, Makhesana, *et al.* (2023) and Reeber, Wolf, and Möhring (2024) document explicitly in their comparison of machine learning methods for Ti-6Al-4V surface roughness prediction. This result matches the theoretical view that boosted trees are better suited to high-complexity problems in which the response surface contains sharp discontinuities that kernel methods do not capture well. For the smooth, physics-driven Ra response of Ti-6Al-4V turning within a CCD domain, the expressive capacity of a deep ensemble is not required. Keeping XGBoost in the comparison despite that expectation allows the empirical results to test this line of reasoning directly, while its feature-importance output adds an interpretive contribution that the other models do not provide.

The five models described in this section differ not only in their mathematical structure but in what they can and cannot express, how they behave when the training set is small, and what information they provide beyond a point prediction. Table 1 consolidates these dimensions into a single comparative view, mapping each model to its functional form, its small-sample behaviour, its interpretability properties, and its primary role in the evaluation that follows. The contrasts visible in the table motivate the model selection: no single method dominates all criteria, and the five together form a complementary set that spans the full range from the classical interpretable baseline to the probabilistic and ensemble alternatives.

## 2.3 Large Language Models — Foundations and Capabilities

Large language models are neural networks trained to predict the next token in a sequence from a large corpus of text. The architecture that underlies virtually all contemporary LLMs is the Transformer, introduced by Vaswani *et al.* (2017) in 2017, whose defining operation is self-attention: each token in a sequence attends to all other tokens, computing a weighted sum of their representations according to learned query-key compatibility scores. Stacking many such attention layers, interspersed with position-wise feed-forward networks and normalisation, produces a model capable of capturing long-range dependencies and compositional structure in ways that previous recurrent architectures could not, a point

**Table 1** – Comparative characteristics and rationale for the five predictive models evaluated in this study.

| **Feature** | **Power Law** | **RSM** | **SVR** | **GPR** | **XGBoost** |
|---|---|---|---|---|---|
| *Model structure* | | | | | |
| Functional form | Power law | Polynomial | RBF kernel | Gaussian proc. | Decision trees |
| Captures interactions | No | Yes | Yes | Yes | Yes |
| Non-linear response | Limited | Quadratic | Arbitrary | Arbitrary | Piecewise |
| *Small-sample behaviour* | | | | | |
| Works with $N < 50$ | Yes | Yes | Yes | Yes | Tends to overfit |
| Hyperparameter tuning | None | $\alpha$ (Ridge) | $C, \varepsilon, \gamma$ | Self-optimising | depth, lr, $n$ |
| *Interpretability* | | | | | |
| Parameter meaning | Exponents | Coefficients | Black box | Black box | Feature import. |
| Uncertainty estimate | No | No | No | Posterior $\sigma$ | No |
| *Rationale for inclusion* | | | | | |
| Primary role | Interpretable baseline | Interaction + curvature | Best LOOCV accuracy | Uncertainty-aware | Feature importance |

that Bommasani *et al.* (2022) develop at length in their survey of the broader implications of this architectural shift.

The scaling behaviour of these models has been the central empirical finding of the past decade. Kaplan *et al.* (2020) showed that model performance on language modelling tasks follows smooth power laws as a function of parameter count, training data volume, and compute budget, with no sign of saturation at scales then considered large. This motivated the training of progressively larger models, GPT-3 at 175 billion parameters, introduced by Brown *et al.* (2020), PaLM at 540 billion, developed by Chowdhery *et al.* (2022), and their successors, and the emergence of capabilities that were not present in smaller models and had not been explicitly trained for. Instruction following, multi-step reasoning, code generation, and quantitative estimation across scientific domains all improved qualitatively, not merely quantitatively, beyond certain scale thresholds, a phenomenon that Wei *et al.* (2022a) term emergent capability and document systematically across a wide range of tasks.

The mechanism that makes LLMs useful for prediction tasks without retraining goes by the name in-context learning, first characterised at scale by Brown *et al.* (2020) in the GPT-3 paper. The idea is disarmingly simple: place a sequence of input-output pairs in the prompt, and the model infers the task from the pattern without touching its weights. No gradient descent, no parameter update, no training loop. The demonstrations act as

a provisional specification of the function the model is expected to approximate (Dong *et al.*, 2024).

What makes this work is not fully understood, but the empirical picture is fairly consistent. Min *et al.* (2022) examined which properties of the demonstrations actually matter and found that the format and the label space carry more weight than the factual correctness of individual examples, a counterintuitive result that suggests the model is reading the structure of the prompt more than its content. Accuracy scales with model size and with how closely the demonstrations resemble the target task, which has direct implications for the prompting strategy used in this work: the closer the few-shot examples are to the actual cutting conditions being predicted, the better the in-context inference is expected to behave (Brown *et al.*, 2020).

Two prompting strategies have attracted particular attention. Few-shot prompting, as defined by Brown *et al.* (2020), provides $k$ examples of the task before the query; the model extends the pattern. Chain-of-thought prompting, introduced by Wei *et al.* (2022b), augments the prompt with explicit intermediate reasoning steps, either as part of the demonstrations or as an instruction to reason before answering. Chain-of-thought was shown to substantially improve performance on arithmetic, commonsense, and symbolic reasoning benchmarks, particularly in models above approximately 100 billion parameters, with Kojima *et al.* (2023) extending this finding to the zero-shot setting by demonstrating that the instruction to reason before answering is alone sufficient to elicit the effect. The improvement is attributed to the model externalising intermediate computations that would otherwise need to be compressed into the residual stream across layers, a form of scratchpad that allows the computation to be distributed over the generated tokens rather than compressed into the hidden state.

Pre-training exposes a language model to an enormous volume of scientific and technical text, machining handbooks, journal articles on surface texture, engineering reference tables, a coverage that Zhao *et al.* (2026) catalogue across major LLM families. That exposure does not produce explicit knowledge of specific experiments.

What it does produce is subtler: statistical associations between cutting parameters, material names, and roughness outcomes, encoded in the model weights and retrievable at inference time without any additional training step. Whether those latent associations are strong enough to support accurate quantitative prediction, and how they interact with the demonstrations supplied in the prompt, is precisely the empirical question this work sets out to answer.

LLM outputs in quantitative tasks are not deterministic. The same prompt submitted twice can return different numbers, due to floating-point non-determinism in distributed inference and stochastic sampling at the output layer (Ouyang *et al.*, 2022). Liang *et al.* (2023) documented sizeable performance swings across repeated runs of the same model on the same benchmark and argued that a single execution tells you very little about what

the model can actually do.

In regression the stakes are higher than in classification: a shift of a few tokens in the output distribution maps directly onto a change in the reported RMSE, with no threshold to absorb it. Hejia Liu, Mochen Yang, and Adomavicius (2025b) document this fragility in data-fitting tasks specifically. The implication is that any study using LLMs as quantitative predictors must run multiple independent evaluations and report the spread, a requirement that shaped the three-run protocol adopted here.

## 2.4 Large Language Models in Engineering and Manufacturing

Large language models began appearing in engineering and manufacturing contexts from 2023 onward, and the applications that emerged first were almost entirely qualitative (Li, Y. *et al.*, 2024; Ouerghemmi; Ertz, 2025). The models were used as conversational interfaces to technical documentation, as process planning assistants, and as code generators, tasks where fluency in language matters more than numerical precision.

Makatura *et al.* (2024) carried out one of the earliest systematic evaluations of this kind, testing GPT-4 across the full computer-aided design and manufacturing workflow. The study covered natural-language-to-CAD translation, design space generation, and manufacturing instruction synthesis. The models handled symbolic and instruction-following tasks reasonably well, but ran into trouble as soon as tasks required exact tolerances or geometric constraints across multiple reasoning steps.

G-code generation for CNC machining attracted its own line of work. Šket *et al.* (2025) put ChatGPT-3.5 and GPT-4o through three phases (independent generation, self-interpretation, and error correction) on three-axis machining tasks. GPT-4o produced syntactically correct toolpaths for simple operations like drilling, but the method depended heavily on iterative user correction and broke down for anything geometrically complex.

Abdelaal, Lokadjaja, and Engert (2025) addressed the correction dependency directly, combining a fine-tuned StarCoder-3B model with retrieval-augmented generation and a self-corrective loop. The result, called GLLM, reduced the need for human intervention during generation. Even so, the framing remained the same across all this G-code work: the LLM acts as a translator between natural language and machine instructions, not as a predictor of process outcomes.

Knowledge retrieval and maintenance formed a second cluster of applications. Yiwei Li *et al.* (2024) surveyed uses across manufacturing, covering quality control, supply chain optimisation, robotic planning, and sensor monitoring. Badini *et al.* (2023) tested ChatGPT for troubleshooting additive manufacturing processes and found it useful for experienced operators who could verify its suggestions, but unreliable for novice users who could not.

Ouerghemmi and Ertz (2025) reviewed 53 papers on LLMs in digital manufacturing and organised the findings into three axes: process optimisation, data structuring and innovation, and human-machine interaction. Across all three, the pattern is the same.

**Table 2** – Research gap analysis across the reviewed literature categories.

| Study category | LLM as quantitative predictor | ML vs LLM direct comparison | Prompting strategy evaluation | Multi-run stability protocol | Nested CV validation |
|---|---|---|---|---|---|
| 1. ML for Ra prediction | – | – | – | – | ✓ |
| 2. LLMs in manufacturing | – | – | – | – | – |
| 3. LLMs for G-code generation | – | – | – | – | – |
| 4. LLMs for tabular regression | ✓ | – | – | – | – |
| 5. LLM benchmarking | – | – | – | ✓ | – |
| 6. CoT prompting studies | – | – | ✓ | – | – |
| **Present study** | ✓ | ✓ | ✓ | ✓ | ✓ |

LLMs are evaluated on how well they process and generate text in a manufacturing context. None of the surveyed work asks whether a language model can predict a quantitative process output directly from numerical experimental data.

The prospect of using LLMs as direct numerical predictors has been explored primarily outside manufacturing. In financial time series, Jin *et al.* (2024) demonstrated that reprogrammed LLMs matched or exceeded specialised forecasting models under few-shot conditions. For molecular property prediction, MolecularGPT demonstrated that a specialised LLM under few-shot in-context learning could outperform standard graph neural network methods on 4 of 7 benchmark datasets with only two demonstrations (Liu, Y. *et al.*, 2024). More broadly, Hejia Liu, Mochen Yang, and Adomavicius (2025a) showed that few-shot in-context learning with GPT-4o-mini could outperform many tabular supervised learning techniques on regression benchmarks; however, the same study identified a fundamental limitation, prediction sensitivity to task-irrelevant variations in the prompt, a form of instability that supervised ML methods are immune to by construction. This finding converges with the benchmarking concerns raised by Liang *et al.* (2023), who documented substantial performance variance across repeated evaluations of the same model and argued that single-run evaluation is insufficient for rigorous quantitative comparison.

Taken together, the literature reviewed in this section reveals a set of open questions that the present work is positioned to address. None of the surveyed studies deployed LLMs as direct predictors of a machining quality variable, compared their performance against properly optimised ML models under the same validation protocol, or examined how prompting strategy and run-to-run stability interact with prediction accuracy in a physical regression task. Table 2 synthesises these gaps; the supporting evidence for each row is developed in the preceding sections.

The gaps identified in Table 2 are not independent of one another. They converge on a single underlying question: can the in-context learning capacity of a large language mode, shaped by pre-training on a corpus that includes machining literature, engineering handbooks, and quantitative scientific text, substitute for the explicit training procedure that conventional ML models require, and do so reliably enough to be useful in industrial practice? Answering this question requires a methodological standard that does not yet exist in the manufacturing literature: a controlled experimental dataset, a rigorous validation protocol shared by both model families, and a multi-run evaluation that distinguishes stable model behaviour from stochastic noise. Establishing that standard, and reporting what it reveals about the relative capabilities of ML and LLM approaches for Ra prediction, is the primary contribution of this work.

# 3 METHODOLOGY

## 3.1 Experimental Setup

All experiments were performed on a ROMI GL240 CNC lathe equipped with a GE FANUC numerical control unit under strictly dry cutting conditions at a controlled ambient temperature of $22 \pm 2\,^{\circ}\mathrm{C}$. The workpiece material was Ti–6Al–4V Grade 5, supplied as cylindrical bars of 25 mm diameter and 150 mm length. A standardised facing and roughing pass was applied before each run to eliminate surface oxide layers and ensure dimensional uniformity across specimens.

Two Sandvik grade 1125 carbide inserts, PVD-coated with submicron-grain cemented carbide, were used with TNMG 1604-MF geometry and MF chipbreaker, differing exclusively in nose radius: $r_\varepsilon = 0.4$ mm (TNMG 160404-MF) and $r_\varepsilon = 0.8$ mm (TNMG 160408-MF). Both inserts were mounted on a MTGNR2020K16 tool holder, providing an entering angle $\kappa_r = 91^{\circ}$, a nominal rake angle of $-6^{\circ}$, and a clearance angle of $6^{\circ}$. A fresh cutting edge was used for every experimental run, isolating the geometric effect of nose radius as the sole varying tool specification and preventing progressive wear from confounding the surface roughness measurements.

## 3.2 Design of Experiments

A Central Composite Design with $k = 3$ factors and five levels was adopted to map the process response space. The three independent variables were cutting speed ($V_c$), feed rate ($f$), and depth of cut ($a_p$); their coded and actual levels are given in Table 3. The CCD structure comprised eight factorial corner points, six axial star points at distance

$$\alpha = \sqrt[4]{2^k} = 1.682 \tag{3.1}$$

The CCD structure comprised eight factorial corner points, six axial star points positioned at a distance $\alpha$ from the design centre, and five centre-point replicates. For a rotatable Central Composite Design with $k = 3$ factors, the axial distance was defined as

$$\alpha = (2^k)^{1/4} = (2^3)^{1/4} = 1.682. \tag{3.2}$$

This value satisfies the rotatability condition of the CCD, meaning that the prediction variance is the same for experimental points located at the same radial distance from the design centre, irrespective of direction in the factor space. The configuration also allows the simultaneous estimation of first-order, quadratic, and two-factor interaction effects. The five centre-point repetitions serve two functions: estimation of pure experimental error independently of the fitted model and quantification of process reproducibility at

the nominal operating condition. Run order within each nose-radius group was fully randomised to reduce systematic bias associated with machine thermal drift.

**Table 3** – CCD factor levels for dry turning of Ti–6Al–4V. The axial star points correspond to $\alpha = 1.682$, calculated as $\alpha = (2^k)^{1/4}$ for $k = 3$ to satisfy the rotatability condition of the Central Composite Design.

| Factor | $-\alpha$ | $-1$ | 0 | $+1$ | $+\alpha$ |
|---|---|---|---|---|---|
| $V_c$ (m/min) | 104 | 135 | 180 | 225 | 256 |
| $f$ (mm/rev) | 0.035 | 0.050 | 0.175 | 0.300 | 0.385 |
| $a_p$ (mm) | 0.032 | 0.150 | 0.250 | 0.330 | 0.368 |

The design yielded $n = 19$ runs per insert and $N = 38$ runs in total. Surface roughness $Ra$ was measured with a Taylor Hobson Surtronic S-128 profilometer with cut-off wavelength $\lambda_c = 0.8$ mm and evaluation length $5\lambda_c$, following ISO 4288, at three equally spaced longitudinal positions per specimen; the reported value is the arithmetic mean of those three measurements. The theoretical kinematic roughness (Equation 2.1) was computed for every experimental point and used as a physical baseline throughout the analysis. Equation 2.1 achieved a Pearson correlation of $r = 0.992$ with measured $Ra$ across the full dataset, indicating that the geometric groove component accounts for most of the observed variation while measurable deviations remain due to thermal, adhesive, and dynamic effects specific to Ti–6Al–4V turning.

### 3.3 Dataset Construction

The experimental data analysed in this study comprise two previously published datasets obtained from dry turning experiments on Ti–6Al–4V. The R04 subset ($r_\varepsilon = 0.4$ mm, $n = 19$) was previously reported by Vilas Boas Pappi Maciel *et al.* (2026), whereas the R08 subset ($r_\varepsilon = 0.8$ mm, $n = 19$) was previously reported by Souza *et al.* (2026). Both subsets were generated from Central Composite Designs involving cutting speed ($V_c$), feed rate ($f$), and depth of cut ($a_p$), and are reused here for a different predictive investigation focused on the comparative evaluation of conventional mathematical models, machine learning models, and large language models.

The two experimental subsets were combined into a single dataset of $N = 38$ observations for the modelling tasks performed in this study. Nose radius was treated as an explicit input feature rather than as a grouping variable because Equation 2.1 establishes an inverse relationship between $r_\varepsilon$ and the theoretical geometric roughness, $Ra_{\text{th}} = f^2/(32r_\varepsilon)$. For identical feed-rate conditions, the theoretical ratio between the two insert geometries is therefore

$$\frac{Ra_{\text{th},0.4}}{Ra_{\text{th},0.8}} = \frac{f^2/(32 \times 0.4)}{f^2/(32 \times 0.8)} = \frac{0.8}{0.4} = 2.0.$$

The inversion of the radius ratio results directly from $r_\varepsilon$ appearing in the denominator of the theoretical expression. Thus, the geometric model predicts that, under the same feed rate, the insert with $r_\varepsilon = 0.4$ mm should produce twice the theoretical geometric roughness of the insert with $r_\varepsilon = 0.8$ mm. Experimentally, the mean roughness values were 2.97 $\mu$m and 1.69 $\mu$m, respectively, giving

$$\frac{\overline{Ra}_{0.4}}{\overline{Ra}_{0.8}} = \frac{2.97}{1.69} \approx 1.76.$$

The experimental ratio is therefore approximately 12% lower than the ideal geometric ratio of 2.0, while preserving the expected direction and approximate magnitude of the nose-radius effect. This difference is consistent with the fact that measured surface roughness also reflects physical mechanisms not represented by the purely geometric formulation. Treating $r_\varepsilon$ as an input feature therefore allows the models to learn both the dominant geometric relationship and its experimental deviations using the complete set of 38 observations.

The feature vector used for all models was

$$\mathcal{F} = \{V_c, f, a_p, r_\varepsilon\}, \tag{3.3}$$

containing only variables directly specified at the machine controller before the cut. This represents the practically relevant prediction scenario: estimating $Ra$ from planned process parameters without requiring any in-process measurement. The target variable was $Ra$ in its original scale ($\mu$m), with no logarithmic transformation, as the models evaluated do not assume Gaussian residuals and the original scale facilitates direct interpretation of RMSE in physically meaningful units.

## 3.4 Machine Learning Models

Five machine learning models of increasing complexity were evaluated. All models except the Power Law were fitted and evaluated under the nested cross-validation scheme described in Section 3.5.

**Power Law:** The power-law model fits a multiplicative relationship in logarithmic space:

$$\ln Ra = \beta_0 + \beta_1 \ln V_c + \beta_2 \ln f + \beta_3 \ln a_p + \beta_4 \ln r_\varepsilon, \tag{3.4}$$

Which is equivalent to the form $Ra = e^{\beta_0} \cdot V_c^{\beta_1} \cdot f^{\beta_2} \cdot a_p^{\beta_3} \cdot r_\varepsilon^{\beta_4}$ in the original scale. Parameters were estimated by ordinary least squares on the log-transformed data. This model is closed-form, has no tuneable hyperparameters, and served as the interpretable baseline against which all other models were compared.

**Response Surface Model (RSM):** The RSM expands the input space with all second-order polynomial terms (main effects, squared terms, and pairwise interactions),

producing 14 regressors for 4 input features. Ridge regression was used to handle the collinearity introduced by the polynomial expansion, with the regularisation strength $\alpha$ selected by the inner cross-validation grid search (Hoerl; Kennard, 1970).

**Support Vector Regression (SVR):** SVR with a Radial Basis Function (RBF) kernel minimises the $\varepsilon$-insensitive loss of Equation 2.4, subject to an $L_2$ penalty on the model weights controlled by $C$ (Cortes; Vapnik, 1995). Features were standardised per fold before fitting. The hyperparameters $C$, $\varepsilon$, and the kernel bandwidth $\gamma$ were all selected by nested grid search.

**Gaussian Process Regression (GPR):** GPR models the target as a draw from a Gaussian process with mean zero and the squared-exponential covariance kernel of Equation 2.5, where $\ell_j$ are per-feature length scales and $\sigma_n^2$ is the noise variance (Rasmussen; Williams, 2006). All kernel hyperparameters were optimised by maximising the marginal likelihood of the training data, an internal procedure that does not require cross-validation, making GPR the only model in this study that self-calibrates its effective complexity. The number of random restarts for the marginal likelihood optimiser was selected by nested grid search to ensure convergence.

**XGBoost:** XGBoost (Chen, T.; Guestrin, 2016) builds an additive ensemble of regression trees by sequentially fitting each tree to the gradient of the loss with respect to the current ensemble prediction. The number of trees, tree depth, learning rate, and subsample ratio were all selected by nested grid search.

## 3.5 Nested Cross-Validation and Hyperparameter Optimisation

All predictive models received the same four input features, $V_c$, $f$, $a_p$, and $r_\varepsilon$, defined previously in the feature vector $\mathcal{F}$. Model configuration involved two distinct types of parameters: fixed settings, which remained unchanged throughout the validation procedure, and tuneable hyperparameters, whose candidate values were evaluated within the inner cross-validation loop.

With $N = 38$ observations, using the same cross-validation procedure for both hyperparameter selection and final performance estimation could produce optimistically biased generalisation estimates because model selection would no longer be independent of model evaluation (Cawley; Talbot, 2010). To reduce this selection bias, a nested cross-validation scheme was adopted for all models requiring hyperparameter optimisation.

The outer loop used LOOCV, which maximised the amount of data available for training in this small-sample setting. For each of the 38 outer folds, one observation was completely withheld for final evaluation, while the remaining 37 observations were used exclusively for model fitting and hyperparameter selection. Within these 37 training observations, a 5-fold inner cross-validation procedure was used to evaluate the candidate hyperparameter configurations. The inner folds were generated with random shuffling and a fixed random seed of 42 (Arlot; Celisse, 2010).

For RSM, SVR, and XGBoost, all combinations specified in the corresponding search grids were evaluated using `GridSearchCV`. The optimisation criterion was negative RMSE, so the configuration producing the lowest cross-validated RMSE was selected. Grid search therefore did not update the models from a single initial hyperparameter configuration. Instead, it compared the predefined candidate combinations and selected the best-performing configuration independently within each outer LOOCV fold.

The selected configuration was subsequently fitted using all 37 observations available in the corresponding outer training set, and the resulting model was used to predict the single held-out observation. This procedure was repeated until every observation had served once as the external test point. Consequently, no held-out observation participated directly or indirectly in the selection of the hyperparameters used to predict it.

The GPR procedure required a slightly different treatment. Its kernel hyperparameters, including the signal magnitude, the per-feature length scales, and the noise level, were optimised internally by maximising the marginal likelihood during model fitting. The length scales were initialised at $\ell_j = 1.0$ for each of the four features, while the initial noise level was set to 0.01. The length-scale search bounds were $[10^{-2}, 10^{2}]$, and the noise level was constrained to $[10^{-5}, 10^{1}]$. The inner cross-validation procedure therefore did not directly select these kernel parameters. Instead, it selected the number of optimiser restarts, `n_restarts_optimizer`, from the candidate values {3, 7, 15}. This parameter controls how extensively the marginal-likelihood optimisation explores different initial conditions.

The Power Law model required no hyperparameter search. Its coefficients were estimated directly by ordinary least squares after logarithmic transformation of the input variables and the response. It was therefore evaluated using the same outer LOOCV structure but without an inner optimisation loop.

Table 4 summarises the fixed model settings and the hyperparameter search spaces used in the nested validation procedure. The candidate ranges were based on configurations previously explored by the authors in a related Ti–6Al–4V surface roughness prediction study (Souza *et al.*, 2026) and were restricted to maintain a feasible computational cost while covering the ranges considered relevant for the present dataset.

All implementations used scikit-learn[1] for RSM, SVR, and GPR, and the `xgboost` library[2] for XGBoost. The Power Law model was implemented in NumPy using ordinary least squares on log-transformed data. The inner cross-validation folds were generated using a fixed random seed of 42 to improve computational reproducibility. All experiments were conducted in Python 3.11 on Google Colab using a standard CPU runtime.

[1] `https://scikit-learn.org` (version 1.3)
[2] `https://xgboost.readthedocs.io` (version 2.0)

**Table 4** – Fixed model settings and hyperparameter search spaces used in the nested cross-validation procedure. Candidate hyperparameter configurations were evaluated by 5-fold inner cross-validation on the 37 training observations of each outer LOOCV fold. RMSE was used as the optimisation criterion.

| Model | Fixed settings | Tuned hyperparameter | Candidate values |
|---|---|---|---|
| Power Law | Log-linear OLS fitting | None | Not applicable |
| RSM | Second-order polynomial expansion; feature standardisation; Ridge regression | Ridge $\alpha$ | {0.01, 0.1, 1, 10, 50} |
| SVR | RBF kernel; feature standardisation | $C$ | {10, 100, 500} |
| | | $\varepsilon$ | {0.05, 0.1, 0.3} |
| | | $\gamma$ | {`scale`, `auto`, 0.1} |
| GPR | Squared-exponential ARD kernel; initial signal magnitude = 1.0; initial $\ell_j = 1.0$; initial noise level = 0.01; `normalize_y=True` | `n_restarts_optimizer` | {3, 7, 15} |
| XGBoost | `colsample_bytree`= 0.8; `reg_alpha`= 0.1; `reg_lambda`= 1.0 | `n_estimators` | {100, 200, 400} |
| | | `max_depth` | {2, 3, 4} |
| | | `learning_rate` | {0.03, 0.05, 0.1} |
| | | `subsample` | {0.7, 0.8, 1.0} |

## 3.6 LLM Evaluation Protocol

### 3.6.1 Model Access and Generation Parameters

Three large language models were evaluated: Claude Sonnet 4.5 (Anthropic, 2025), GPT-4o-mini (OpenAI, 2024), and Gemini 3 Flash Preview (Google DeepMind, 2026). All models were accessed exclusively through their respective official APIs, namely the Anthropic Messages API, the OpenAI Chat Completions API, and the Google Generative AI API, rather than through interactive chat interfaces. All API-based experimental evaluations reported in this study were conducted on January 20, 2026. This access date is reported because API-served language models may be updated over time, and therefore documenting the date of execution contributes to the reproducibility and traceability of the experimental protocol. This choice was deliberate: API access with a fixed generation parameter (`temperature = 0`) provides controlled and reproducible conditions, while also representing the operationally relevant pathway for automated industrial deployment.

The `max_tokens` parameter was set to 32 for few-shot prompts, sufficient for a bare

numeric response, and to 512 for chain-of-thought prompts, to allow adequate space for intermediate reasoning steps before the final numeric answer. No other generation parameters were modified from their API defaults.

### 3.6.2 LOOCV Structure for LLMs

The same outer LOOCV loop used for ML models was applied to the LLMs: for each of the 38 folds, the 37 training observations were embedded in the prompt as demonstrations, and the model was asked to predict *Ra* for the held-out point. This procedure is structurally identical to in-context few-shot learning as described by Brown *et al.* (2020), with the difference that no gradient update occurs: the model's weights are frozen throughout, and the "learning" is purely contextual. Each full LOOCV run therefore consisted of 38 consecutive API calls per model per prompting strategy, for a total of $38 \times 3$ models $\times$ 2 strategies $= 228$ API calls per run.

### 3.6.3 Prompting Strategies

Two prompting strategies were evaluated for each model. Both strategies shared the same underlying structure, comprising a system message and a user message, but differed in the information provided and the type of response requested.

In the few-shot condition, the prompt was designed according to the following principles:

- *Role definition.* The system message assigned the model the role of a precision machining expert, establishing the domain context before any data was presented.
- *Task specification.* The model was explicitly instructed to predict the arithmetic mean surface roughness *Ra* in micrometres for Ti–6Al–4V turning.
- *Output constraint.* To prevent verbose or ambiguous responses, the system message enforced a strict output format: a single positive number with no units, explanation, or surrounding text. This constraint was essential for automated numeric extraction from the API response.
- *Demonstrations.* The user message contained the 37 training observations of the current LOOCV fold, each formatted as a parameter–response pair in the form `Vc, f, ap, re` $\rightarrow$ `Ra`. No additional context, formula, or physical interpretation was provided, following the in-context learning paradigm of Brown et al. (Brown *et al.*, 2020).
- *Query.* The final line of the user message presented the held-out point's parameters with a blank response field, prompting the model to complete the pattern.

In the chain-of-thought condition (Wei *et al.*, 2022b), the prompt retained the role definition and demonstrations but modified three elements:

- *Output constraint relaxed.* The system message permitted brief intermediate reasoning, but maintained the requirement that the final line contain only the numeric prediction. This allowed the model to externalise its reasoning without compromising automated parsing.
- *Physical prior.* Before the experimental demonstrations, the user message presented the theoretical formula $Ra_{\text{th}} = f^2/(32\, r_\varepsilon)$ from Equation 2.1 together with its numerical value pre-computed for the target point of each fold. This prior was included because it encodes the dominant physical relationship in the data (Pearson $r = 0.992$, Section 4.1) and was intended to anchor the model's reasoning to an established engineering formula before it engaged with the experimental examples.
- *Explicit reasoning instruction.* The user message closed with the instruction to think step by step and write only the numeric prediction on the last line, operationalising the chain-of-thought elicitation strategy described by Wei *et al.* (2022b).

### 3.6.4 Response Parsing and Run Repetition

Numeric responses were extracted from model outputs using a post-processing function that retrieved the last numeric value in the response string. This convention handles both the bare numeric outputs of few-shot responses and the reasoned paragraphs of chain-of-thought outputs, where the prediction appears on the final line. Responses outside the physically plausible range $[0.01,\ 50]$ $\mu$m were treated as parse failures and excluded from metric computation. The fraction of valid predictions per run ($n_{\text{valid}}/38$) is reported alongside the RMSE and $R^2$ as an indicator of model reliability.

To quantify the run-to-run variability inherent to LLM inference, each model–strategy combination was evaluated across three fully independent LOOCV executions conducted in separate API sessions. Performance metrics are reported as mean ± standard deviation across runs. This protocol addresses a methodological limitation common in the emerging literature on LLMs applied to engineering prediction tasks: single-run evaluations cannot distinguish stable model behaviour from a fortunate or unfortunate sample from the model's stochastic output distribution (Liang *et al.*, 2023).

## 3.7 Evaluation Metrics

All models, ML and LLM, were evaluated with the same set of metrics computed on the LOOCV predictions:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Ra}_i - Ra_i\right)^2}, \tag{3.5}$$

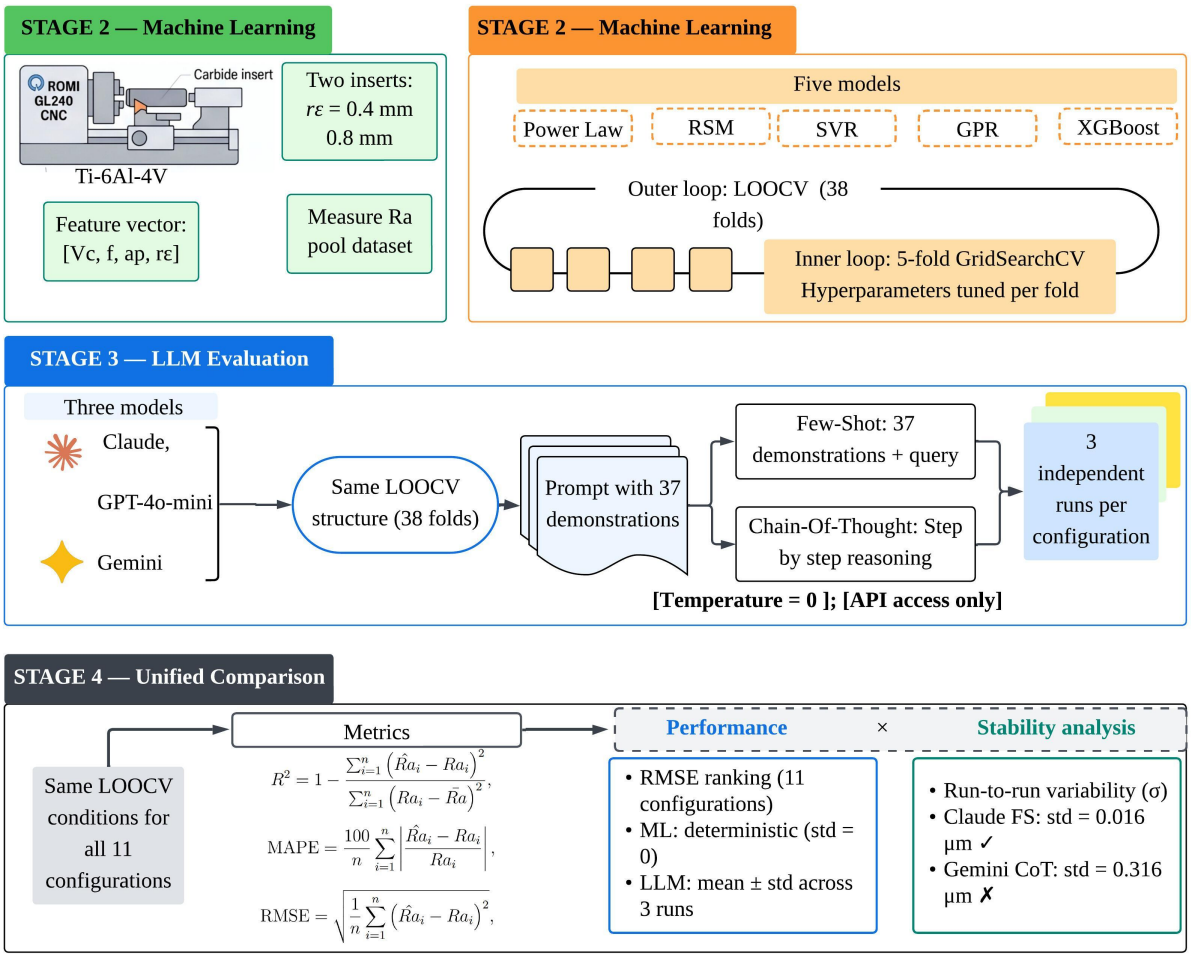
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Ra}_i - Ra_i\right|, \tag{3.6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{Ra}_i - Ra_i\right)^2}{\sum_{i=1}^{n} \left(Ra_i - \bar{Ra}\right)^2}, \tag{3.7}$$

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{Ra}_i - Ra_i}{Ra_i} \right|, \tag{3.8}$$

Where $n$ is the number of valid predictions, $\hat{Ra}_i$ is the predicted value, $Ra_i$ is the measured value, and $\bar{Ra}$ is the sample mean. RMSE was used as the primary ranking criterion because it penalises large errors quadratically, which is consistent with the industrial context where outlier predictions, such as a dramatically incorrect Ra estimate for a specific cutting condition, are more costly than small systematic biases. $R^2$ is reported as a normalised complement to RMSE that facilitates comparison across datasets and with published results. MAPE is included as a scale-independent metric but is not used for ranking, as it becomes unstable for points with very low measured Ra (the denominator approaches zero). For LLM configurations with negative $R^2$, indicating predictions worse than the global sample mean, MAPE is not reported as it loses interpretive value in that regime. The complete experimental and computational pipeline is summarised in Figure 4.

**Figure 4** – Overview of the experimental and computational methodology, comprising experimental design (CCD, $N$=38), nested LOOCV for ML models, LLM prompt comparison (Few-Shot vs. CoT), and unified performance ranking.



# 4 RESULTS AND DISCUSSION

## 4.1 Experimental Dataset and Exploratory Analysis

The combined dataset comprises 38 observations of surface roughness $Ra$ obtained under dry turning of Ti–6Al–4V, equally distributed between two experimental groups: 19 runs associated with $r_\varepsilon = 0.4$ mm (R04) and 19 runs associated with $r_\varepsilon = 0.8$ mm (R08). The experimental domain was defined by a CCD with three controllable factors, cutting speed $V_c$, feed rate $f$, and depth of cut $a_p$. Their levels were established a priori by the experimental design, including factorial, axial, and centre-point conditions. Nose radius was fixed within each experimental group and subsequently included as an explicit input feature in the predictive models.

Before evaluating the predictive models, the experimental response was examined to identify its overall variation, the relative influence of the process inputs, and the extent to which the dominant physical relationship expected from turning theory is reflected in the measured data. Because the levels of $V_c$, $f$, and $a_p$ were predetermined by the CCD rather than obtained as random observational variables, descriptive statistics for these inputs were not used to characterise their variability.

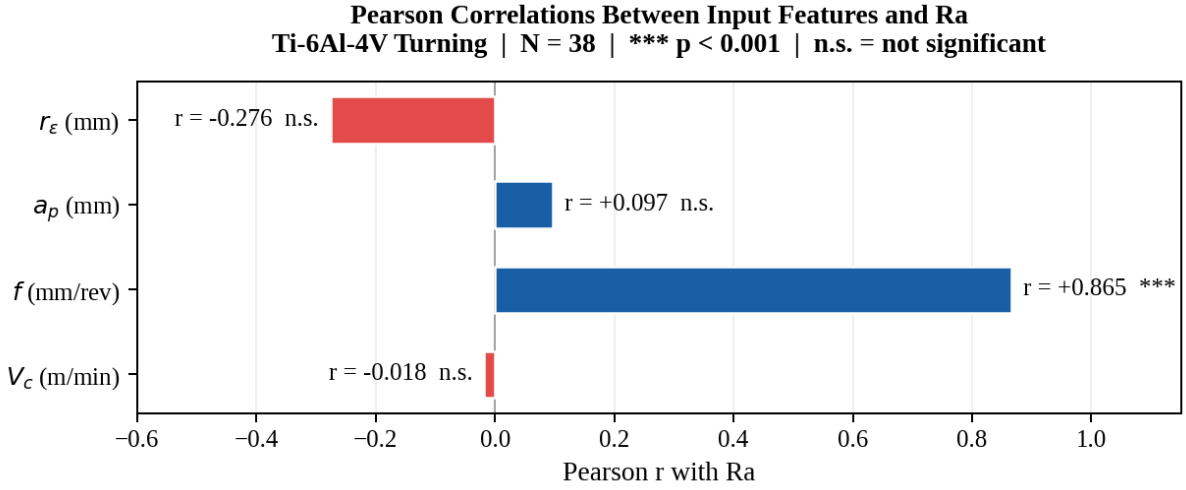
The measured $Ra$ values span from 0.25 to 10.45 $\mu$m across the complete dataset, corresponding to a more than fortyfold difference between the lowest and highest observed roughness values. This broad response range reflects the deliberate exploration of the experimental domain imposed by the CCD rather than uncontrolled variability in the experiments. Factorial and axial conditions extend the observations towards different regions of the process space, while centre-point replicates provide repeated measurements under nominal operating conditions. The resulting dataset therefore combines a wide range of surface responses with a structured experimental design suitable for evaluating predictive models.

The two experimental groups also exhibit different average roughness levels. R04 ($r_\varepsilon = 0.4$ mm) produced a mean $Ra$ of 2.97 $\mu$m, whereas R08 ($r_\varepsilon = 0.8$ mm) produced a mean of 1.69 $\mu$m, resulting in an experimental ratio of approximately 1.76. For identical feed-rate conditions, the geometric relationship $Ra_{\text{th}} = f^2/(32r_\varepsilon)$ predicts a theoretical ratio of $0.8/0.4 = 2.0$. The experimental ratio therefore preserves the direction and approximate magnitude expected from the geometric formulation, while remaining approximately 12% below the ideal value. This difference indicates that measured surface roughness is not determined exclusively by the geometric contribution and may also reflect physical mechanisms and process interactions not represented by the theoretical expression.

### 4.1.1 Influence of Input Features on Ra

Before any model is fitted, it is worth asking which variables in the dataset actually move with Ra and which do not. A first answer comes from the Pearson correlation coefficient, which measures the strength of the linear association between each input and the response. Although a low correlation does not rule out a non-linear or interactive relationship, it does reveal which variables carry direct predictive signal on their own.

**Figure 5** – Pearson correlation coefficients between input features and Ra ($N = 38$). Only feed rate $f$ shows a statistically significant linear association. The non-significance of $r_\varepsilon$ as an isolated variable does not contradict its physical importance, which only manifests in interaction with $f^2$ in the theoretical formula.



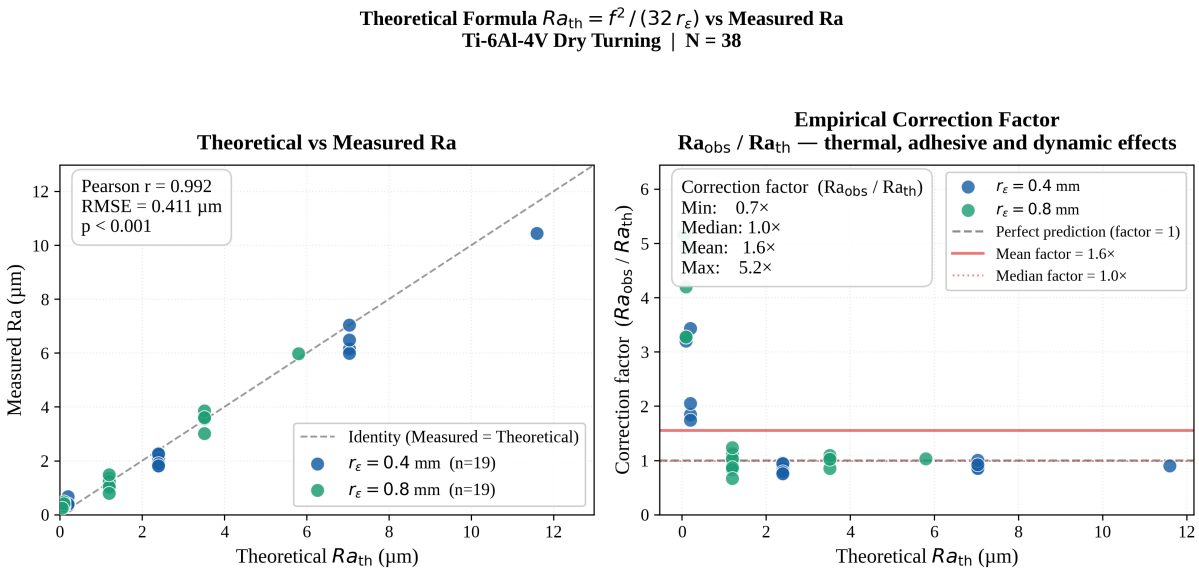

The picture shows that feed rate $f$ is the only variable with a strong and statistically significant association with Ra ($r = +0.865$, $p < 0.001$): when feed increases, roughness goes up, and it does so consistently across the entire dataset. The other three variables tell a different story. Cutting speed $V_c$ ($r = -0.018$) and depth of cut $a_p$ ($r = +0.097$) show no linear association with Ra, consistent with their near-zero exponents in power-law models for Ti-6Al-4V reported in the literature (Yang, W.-H.; Tarng, 1998; Ezugwu; Wang, Z.-M., 1997). Nose radius $r_\varepsilon$ sits at $r = -0.276$, negative but non-significant, which at first glance looks odd for a variable that the kinematic formula places in the denominator of $Ra_{\text{th}} = f^2/32r_\varepsilon$.

The explanation is that Pearson correlation examines one variable at a time. The nose radius effect on Ra does not operate alone — it only becomes visible in combination with feed rate, through the product $f^2/r_\varepsilon$. Analysing $r_\varepsilon$ in isolation strips out exactly the context that makes it influential. As shown in the next section, the combination $f^2/r_\varepsilon$ achieves a correlation of 0.992 with the observed Ra, a result that the bivariate analysis of $r_\varepsilon$ alone cannot detect.

### 4.1.2 The Theoretical Formula as a Physical Baseline

The relationship between the theoretical kinematic roughness (Equation 2.1) and the measured Ra across all 38 experimental points is examined in Figure 6, which presents two complementary views: a predicted-versus-measured scatter plot and the distribution of the empirical correction factor $Ra_{\mathrm{obs}}/Ra_{\mathrm{th}}$. Together, these panels quantify both the predictive strength of the geometric model and the systematic deviations introduced by real cutting conditions in Ti-6Al-4V turning.

**Figure 6** – Left: theoretical $Ra$ from Equation 2.1 versus measured $Ra$ ($n = 38$, $r = 0.992$, RMSE $= 0.411$ $\mu$m). The dashed line represents perfect agreement between theoretical and measured values. Right: empirical correction factor $Ra_{\mathrm{obs}}/Ra_{\mathrm{th}}$, quantifying the deviation of measured surface roughness from the geometric prediction. The largest relative deviations occur at low $Ra_{\mathrm{th}}$ values, indicating that mechanisms not represented by the purely geometric formulation become proportionally more relevant under these conditions.




The left panel of Figure 6 shows that Equation 2.1 captures the dominant trend of the experimental response, achieving a Pearson correlation of $r = 0.992$ ($p < 0.001$) with measured $Ra$. The right panel shows the extent to which the experimental observations depart from the purely geometric prediction. The correction factor ranges from 0.7× to 5.2×, with a median of approximately 1.0× and a mean of 1.6×. The median value indicates that the centre of the correction-factor distribution is close to the theoretical prediction, although individual observations may fall either below or substantially above it. The mean value of 1.6× and the maximum of 5.2× indicate a right-skewed distribution, with the largest relative deviations concentrated at low $Ra_{\mathrm{th}}$ values. These conditions are predominantly associated with low feed rates, where physical mechanisms not represented by the geometric formulation may become proportionally more relevant to the measured surface roughness.

These deviations are physically consistent with the known machinability challenges of Ti-6Al-4V reported by Ezugwu and Zheng-Min Wang (1997), who identify adhesive wear and built-up edge as the dominant surface-altering mechanisms at low cutting speeds and feeds in titanium alloy turning. At low feed rates, the uncut chip thickness approaches the material's grain size, activating plastic deformation mechanisms that the purely geometric formula cannot represent. The result is that $Ra_{\text{th}}$ becomes an increasingly optimistic lower bound as the operating point moves away from the centre of the experimental domain toward the star points with minimum feed rate. This behaviour explains why data-driven models, which learn the empirical correction implicitly from the training examples, are better positioned to handle the full range of the CCD than the theoretical formula alone.

This observation has a direct implication for the chain-of-thought prompting strategy evaluated in Chapter 4: providing Equation 2.1 as a physical prior anchors the model's reasoning to a formula that is unbiased at the median but can deviate by up to 5.2× at the conditions most removed from the centre-point of the design. The few-shot strategy, by contrast, provides no such anchor and instead allows the model to draw directly on the empirical pattern in the demonstrations.

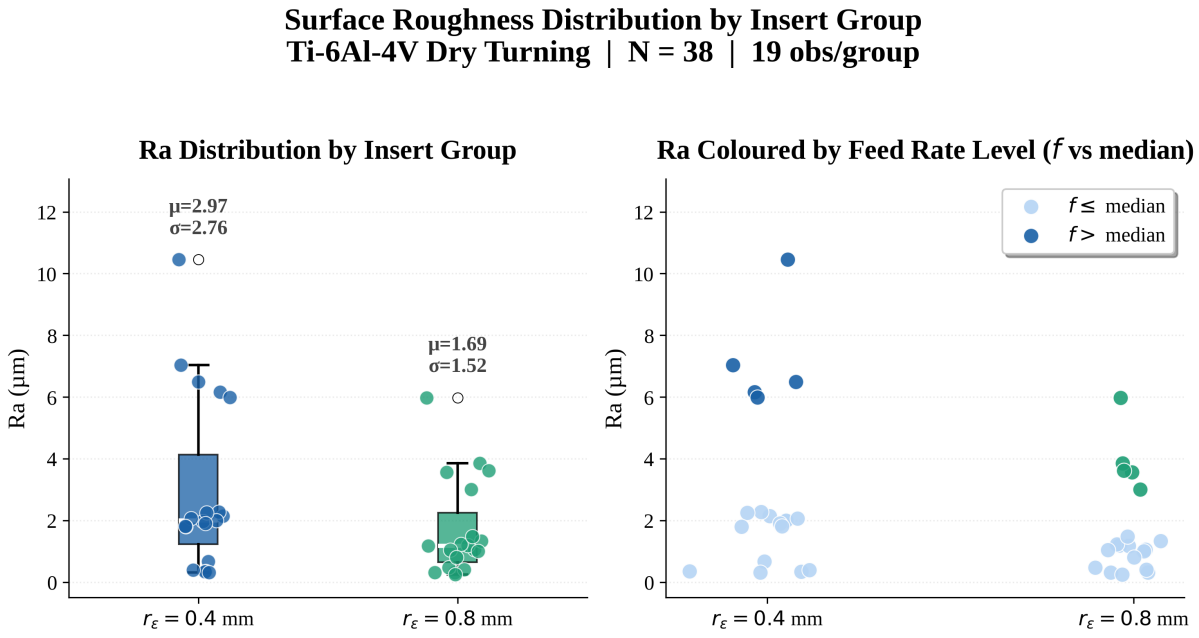
### 4.1.3 Ra Distribution by Insert Group

Combining the two experimental groups into a single modelling dataset requires the systematic difference between the tool conditions to be represented explicitly in the predictor space. In the present analysis, $r_\varepsilon$ was included as an input feature, while $V_c$, $f$, and $a_p$ describe the cutting conditions. The distributional comparison in Figure 7 is therefore used to examine how the response ranges of the two experimental groups overlap and how their within-group variation relates to the variation induced by the process parameters.

Conversely, if the variability within each group, induced by the deliberate variation of $V_c$, $f$, and $a_p$ across the CCD levels, spans a range similar to the gap between groups, then nose radius behaves as a continuous input feature rather than a categorical grouping variable, and the combined dataset is appropriate. Figure 7 addresses this question directly, showing the Ra distribution for each insert group with points coloured by feed rate level relative to the median $f = 0.175$ mm/rev.

Several features of Figure 7 are noteworthy. First, the interquartile ranges of the two groups overlap substantially, which is consistent with combining them into a single dataset: the within-group variability driven by $f$ exceeds the between-group difference driven by $r_\varepsilon$. Second, the right panel makes unambiguous a relationship that the bivariate correlation partially obscures: every observation with Ra above approximately 3 $\mu$m belongs to the high-$f$ subset, regardless of which insert was used. This confirms that feed rate is not merely statistically associated with Ra but is the primary physical lever, and its effect is consistent and dominant across both insert geometries.

**Figure 7** – Left: Ra distributions for the two insert groups. Right: the same observations coloured by feed rate level. Dark points ($f >$ median) consistently produce the highest Ra values within each group, confirming feed rate as the primary driver of surface roughness independent of nose radius.




Third, the five centre-point replicates in each group, visible as tightly clustered points near the median of each distribution, provide a direct estimate of process reproducibility. Their range within R04 (1.80 to 2.25 $\mu$m) and within R08 (0.80 to 1.48 $\mu$m) represents the irreducible noise floor that no model can predict without memorising the run order. This floor sets a practical lower bound on the RMSE that any LOO-validated model can achieve on this dataset.

Although the centre-point range was smaller for R04 (0.45 $\mu$m) than for R08 (0.68 $\mu$m), this difference should not be interpreted as evidence that the 0.4 mm insert is intrinsically more reproducible. Each range is based on only five replicate observations and is therefore sensitive to sampling variability and individual measurements. The observed difference is more appropriately interpreted as normal experimental dispersion under the two nominally identical cutting conditions, rather than as a systematic effect of nose radius.Although the centre-point range was smaller for R04 (0.45 $\mu$m) than for R08 (0.68 $\mu$m), this difference should not be interpreted as evidence that the 0.4 mm insert is intrinsically more reproducible. Each range is based on only five replicate observations and is therefore sensitive to sampling variability and individual measurements. The observed difference is more appropriately interpreted as normal experimental dispersion under the two nominally identical cutting conditions, rather than as a systematic effect of nose radius.

Taken together, the exploratory analysis establishes three properties of the dataset that inform the interpretation of all subsequent modelling results. Feed rate is the dominant and only statistically significant predictor of Ra in isolation. The theoretical formula

captures the trend well but introduces systematic deviations at low-feed conditions that data-driven models are better positioned to correct. The wide dynamic range of the response variable, spanning over two orders of magnitude, provides a demanding but well-structured benchmark for both the ML models and the LLM configurations evaluated subsequently.

## 4.2 Performance of Conventional Mathematical and ML Models

Five predictive models were trained and evaluated for $Ra$ prediction under Leave-One-Out Cross-Validation (LOOCV): a Power Law regression, a second-order Response Surface Model (RSM), a Support Vector Regressor with RBF kernel (SVR), a Gaussian Process Regressor (GPR), and XGBoost. These models comprise two conventional mathematical approaches (Power Law and RSM) and three ML models (SVR, GPR, and XGBoost). To prevent data leakage between hyperparameter selection and generalisation error estimation, all models except the Power Law, which is fitted analytically in logarithmic space and has no tuneable hyperparameters, were optimised through nested cross-validation: for each outer LOOCV fold, a 5-fold grid search was conducted on the 37 available training points to select the best hyperparameter configuration before predicting the held-out observation (Cawley; Talbot, 2010).

The high $R^2$ values obtained under LOOCV do not reflect overfitting. Each prediction was generated for a sample that was never used during training or hyperparameter selection. The elevated predictive accuracy instead reflects the strong physical structure of the dataset: The elevated predictive accuracy instead reflects the strong physical structure of the dataset: the theoretical expression $Ra = f^2/(32\, r_\varepsilon)$ exhibits a very strong linear association with measured $Ra$ (Pearson $r = 0.992$, $r^2 = 0.984$). A well-calibrated model operating on four physically meaningful features in a structured CCD dataset will therefore tend to achieve high $R^2$ under LOOCV. This behaviour is a property of the data structure, not an artefact of the modelling procedure.

Table 5 presents the LOOCV metrics for all five models. Figure 8 visualises the RMSE ranking, and Figure 9 shows the predicted-versus-measured scatter for each model individually.

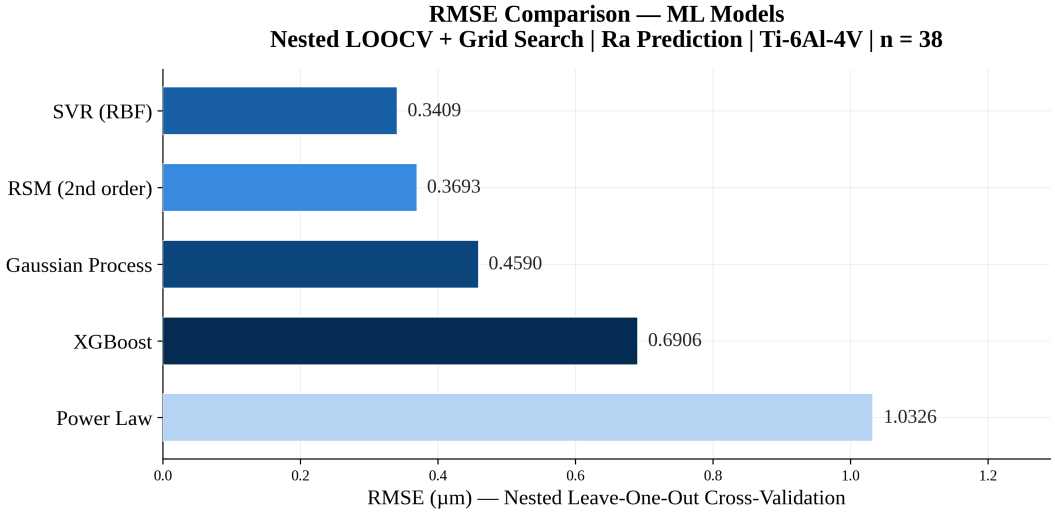
Although RMSE and $R^2$ produce the same model ordering, MAE and MAPE do not necessarily follow the same ranking because the metrics weight prediction errors differently. For a common evaluation set, RMSE and $R^2$ are both directly related to the sum of squared errors, so a lower RMSE is necessarily associated with a higher $R^2$. MAE, in contrast, measures the average absolute deviation and is therefore less sensitive to isolated large residuals. This distinction is visible for GPR, which achieved a lower MAE than RSM (0.283 versus 0.298 $\mu$m) despite having a higher RMSE (0.459 versus 0.369 $\mu$m). This pattern suggests that GPR produced smaller absolute errors for many observations while being more strongly affected by a limited number of larger residuals.

**Table 5** – LOOCV results for the five predictive models with nested cross-validation and grid-search hyperparameter optimisation ($N = 38$). Models are sorted by RMSE. All metrics were computed on held-out observations never used during training or hyperparameter selection.

| Model | RMSE ($\mu$m) | MAE ($\mu$m) | $R^2$ | MAPE (%) |
|---|---|---|---|---|
| SVR (RBF) | 0.341 | 0.269 | 0.978 | 25.0 |
| RSM (2nd order) | 0.369 | 0.298 | 0.975 | 30.2 |
| Gaussian Process | 0.459 | 0.283 | 0.961 | 17.6 |
| XGBoost | 0.691 | 0.431 | 0.912 | 40.5 |
| Power Law | 1.033 | 0.629 | 0.802 | 27.5 |

MAPE provides a different perspective because each absolute error is normalised by the corresponding measured *Ra* value. Since the dataset includes low roughness values, reaching approximately 0.25 $\mu$m, even small absolute prediction errors at these points can produce comparatively large percentage errors. Conversely, larger absolute errors at high-*Ra* observations may have a smaller effect on MAPE. This explains why the MAPE ranking differs from the RMSE and $R^2$ rankings. These differences are mathematically consistent with the definitions of the metrics and do not represent a contradiction among the reported results.

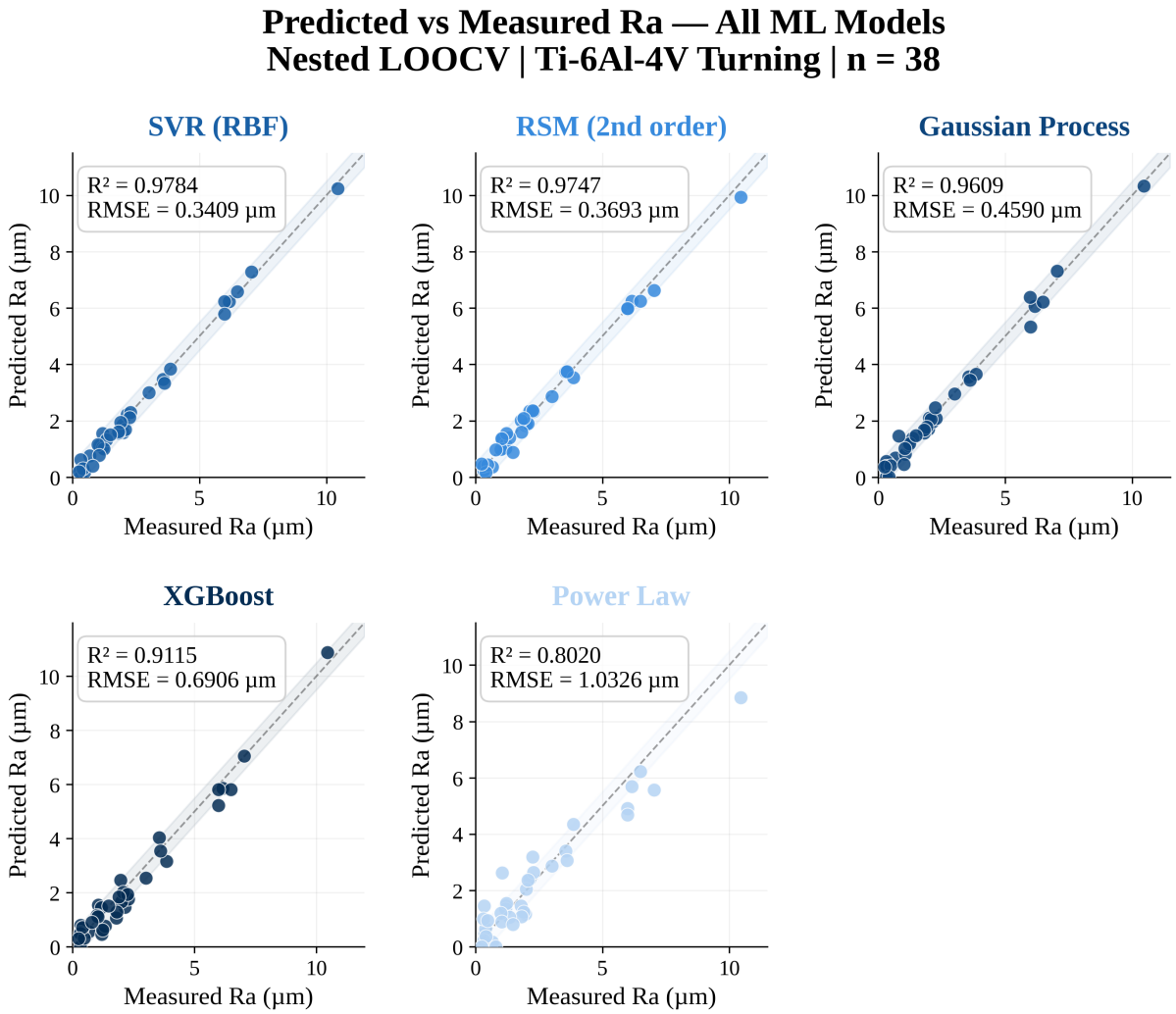
**Figure 8** – RMSE of the five predictive models under nested LOOCV with grid-search hyperparameter optimisation. SVR achieved the lowest RMSE (0.341 $\mu$m), followed by RSM (0.369 $\mu$m) and GPR (0.459 $\mu$m). XGBoost and Power Law showed substantially higher error levels.




As shown in Figure 8 and confirmed by the scatter patterns in Figure 9, SVR, RSM, and GPR form a clearly superior group, with RMSE below 0.46 $\mu$m and $R^2$ above 0.96. XGBoost and Power Law form a second tier, with substantially larger errors and wider scatter at the extremes of the *Ra* range.

The separation between these two tiers is not arbitrary. The three best-performing

**Figure 9** – Predicted versus measured $Ra$ ($\mu$m) for the five predictive models under nested LOOCV ($N = 38$). The dashed line represents the identity ($\hat{Ra} = Ra$), and the shaded band marks the $\pm 0.5$ $\mu$m region. Points clustering near the identity line indicate stronger predictive agreement, whereas deviations at the extremes of the $Ra$ range reveal the systematic limitations of each model.



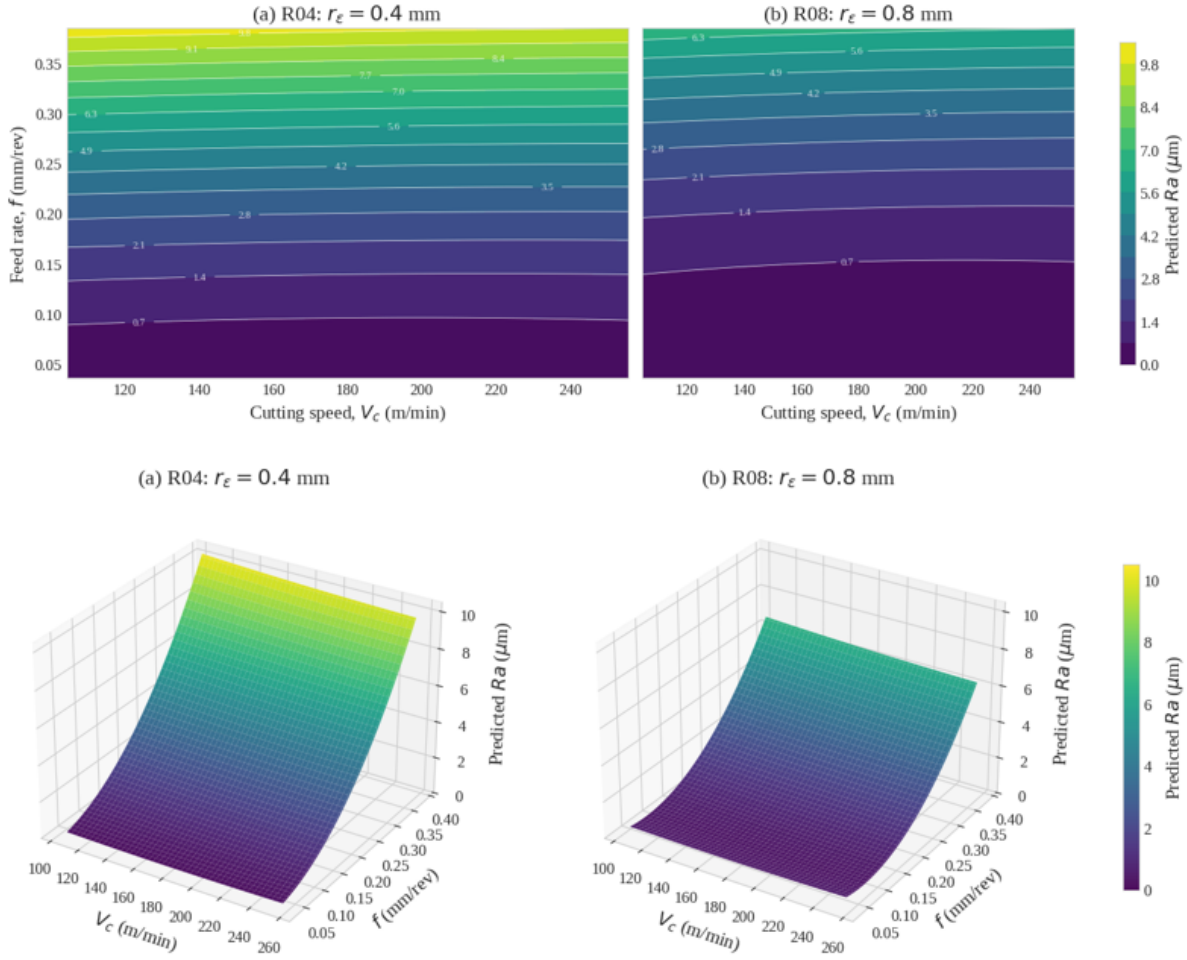

models share an important characteristic: their functional forms can represent continuous and smooth response surfaces with interaction effects, which is consistent with the behaviour of $Ra$ in a CCD dataset. SVR achieved the lowest RMSE through its kernel-based non-linearity and the regularisation built into the $\varepsilon$-insensitive loss. RSM followed closely, benefiting from the direct alignment between the second-order polynomial expansion and the CCD structure. GPR also achieved strong performance, with a modest loss in point-prediction accuracy relative to SVR and RSM, while providing the additional advantage of predictive variance estimation.

By contrast, XGBoost applies a piecewise approximation that is better suited to larger datasets and more irregular response patterns. With only 38 observations and a smooth physical response, its flexibility appears greater than required for the present problem. The

Power Law model occupies the lowest position in the ranking because its multiplicative structure cannot represent interactions among process variables, a limitation that cannot be overcome by coefficient fitting alone.

To further examine the behaviour captured by the RSM, the fitted second-order response was evaluated across the experimental domain while holding the depth of cut at a common intermediate value of $a_p = 0.22$ mm. Figure 10 presents the corresponding response contours and three-dimensional surfaces for the two tool conditions. The model used for this visualisation was fitted to the complete dataset for interpretative purposes only; all predictive performance metrics reported in Table 5 remain those obtained from the nested LOOCV procedure.

**Figure 10** – RSM response contours and surfaces for surface roughness under a common depth of cut of $a_p = 0.22$ mm. Panels (a) and (b) correspond to the tool conditions with $r_\varepsilon = 0.4$ mm and $r_\varepsilon = 0.8$ mm, respectively. A common colour scale is used to permit direct comparison between the predicted response regions. The surfaces were generated from the RSM fitted to the complete dataset for interpretative purposes, whereas predictive performance was assessed using nested LOOCV.



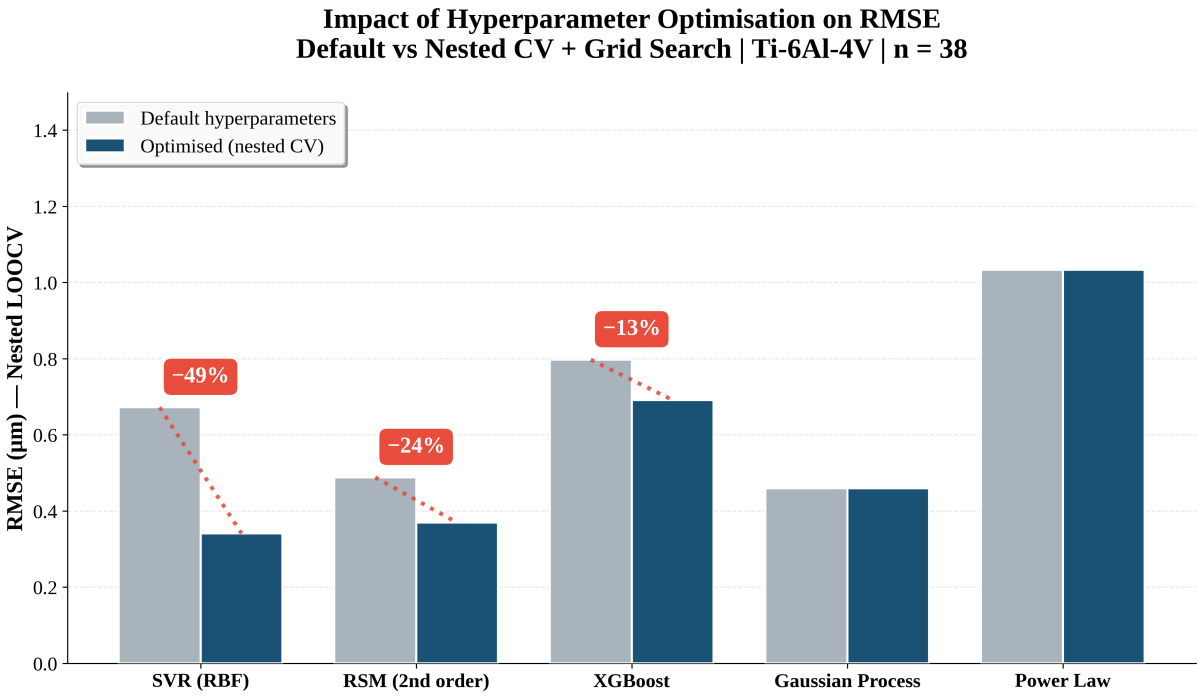

The response surfaces show that feed rate is the dominant factor governing the RSM

predictions across both tool conditions. Surface roughness increases non-linearly as $f$ increases, whereas variation along the cutting-speed axis produces comparatively small changes in the predicted response. This behaviour is consistent with the exploratory analysis, in which feed rate exhibited the strongest association with measured $Ra$. The curvature along the $f$ axis also reflects the second-order polynomial structure of the RSM and the strong dependence of theoretical kinematic roughness on $f^2$. The predicted response is higher for the $r_\varepsilon = 0.4$ mm condition, particularly at larger feed rates, while the $r_\varepsilon = 0.8$ mm condition produces a lower and flatter response over the same process domain.

### 4.2.1 Hyperparameter Optimisation

Before interpreting the final metrics, it is instructive to quantify how much the nested grid search contributed to the reported results. Figure 11 compares the RMSE of each model under default hyperparameter settings against the optimised configuration obtained through nested cross-validation. The improvement varies substantially across models and reveals important characteristics of each method.

**Figure 11** – RMSE reduction from hyperparameter optimisation via nested grid search. SVR gains the most (−49%), reflecting its sensitivity to $C$ and $\gamma$. GPR is unaffected because it self-optimises internally; Power Law has no tuneable parameters.




SVR showed the largest sensitivity to hyperparameter choice. The default configuration ($C = 100$, $\varepsilon = 0.1$, `gamma = scale`) produced RMSE = 0.672 $\mu$m, while the optimised configuration ($C = 500$, $\varepsilon = 0.05$, $\gamma = 0.1$) reduced this to 0.341 $\mu$m, a 49% improvement,

as visible in Figure 11. The consistent preference for a small insensitive tube ($\varepsilon = 0.05$, selected in 100% of the 38 outer folds) reflects the low noise level of the machining data. The higher penalty parameter $C = 500$, selected in 61% of the folds, indicates that the model benefits from stronger regularisation than the conventional default. The fixed bandwidth $\gamma = 0.1$ in 100% of folds confirms that this kernel scale is a robust choice for the present feature space.

RSM improved by 24% through the selection of a lower regularisation strength ($\alpha = 0.1$, selected in 100% of folds, versus the default $\alpha = 1.0$). With 14 polynomial regressors and only 37 training points per fold, the data does not require heavy penalisation to avoid collinearity, and the grid search correctly identified the lighter regularisation as optimal. XGBoost improved by 13% primarily through the constraint on tree depth: `max_depth` $= 2$ was selected in 100% of folds, reflecting that the small sample size does not support deeper decision trees without degrading generalisation. This preference for stumps is consistent with the behaviour reported in the literature for high-variance ensemble methods on small datasets (Probst; Wright; Boulesteix, 2019).

GPR was the only model for which the nested grid search produced no measurable improvement, with RMSE remaining at 0.459 $\mu$m before and after optimisation. This is because GPR self-optimises its kernel hyperparameters, namely the length-scale vector and noise level, by maximising the marginal likelihood of the training data, an internal procedure that already converges to the global optimum independently of the number of random restarts. The grid search for GPR was therefore applied only to `n_restarts_optimizer` ($\in \{3, 7, 15\}$), selecting 7 in 53% of folds, confirming that moderate exploration of the marginal likelihood landscape is sufficient for this dataset. As a byproduct of its probabilistic formulation, GPR provides a predictive uncertainty estimate: the posterior standard deviation across the 38 training points averaged $\sigma = 0.219$ $\mu$m when the model was fitted on the full dataset, a capability unavailable to the other methods and directly applicable to manufacturing quality control (Rasmussen; Williams, 2006).

Table 6 summarises the selected configurations and their effect on RMSE, and the consistency of the grid search choices across folds confirms that the optimal configurations are stable properties of the dataset rather than artefacts of specific training subsets.

### 4.2.2 Best Model: SVR Analysis

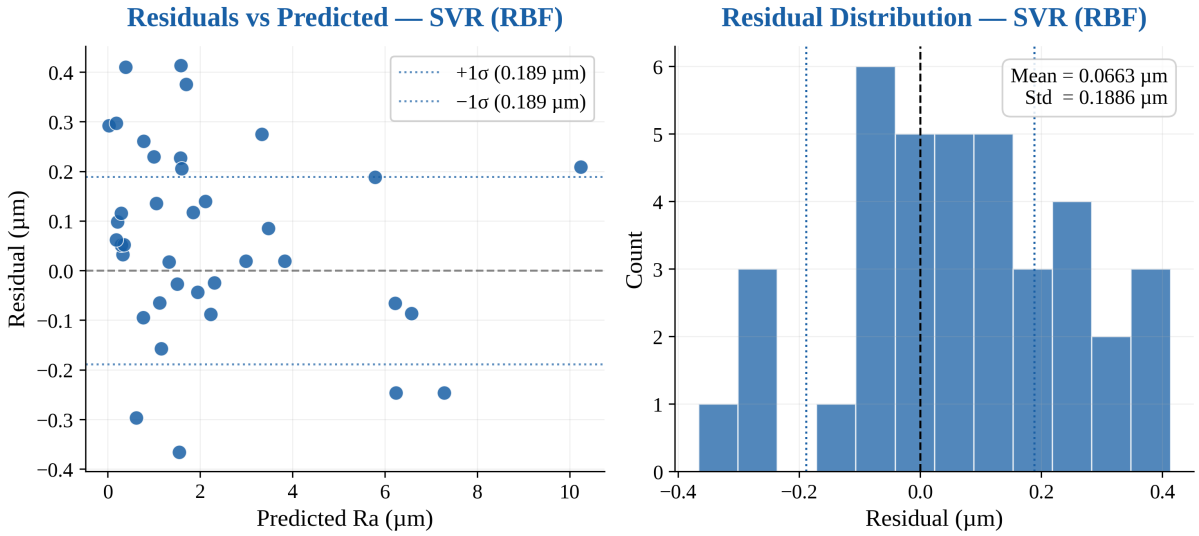
SVR with an RBF kernel achieved the best performance among all ML models (RMSE = 0.341 $\mu$m, $R^2 = 0.978$). As visible in Figure 9, the SVR scatter plot shows tight alignment with the identity line across the entire Ra range, from low-roughness conditions ($Ra <$ 0.5 $\mu$m) to the high-feed axial point ($Ra = 10.45$ $\mu$m). To further characterise the model's error structure, Figure 12 presents the residual analysis.

The residual plot in Figure 12 reveals two important properties of the SVR predictions. First, the residuals are distributed around zero without any systematic trend as a function

**Table 6** – Hyperparameter selection summary from nested cross-validation. The column "Most selected" reports the value or combination chosen in the majority of the 38 outer folds. The final two columns compare RMSE ($\mu$m) before and after optimisation.

| Model | Key hyperparameter(s) | Most selected | Default RMSE | Optimised RMSE |
|---|---|---|---|---|
| SVR (RBF) | $C$ / $\varepsilon$ / $\gamma$ | 500 / 0.05 / 0.1 | 0.672 | **0.341** (−49%) |
| RSM (2nd order) | Ridge $\alpha$ | 0.1 (100% of folds) | 0.488 | **0.369** (−24%) |
| XGBoost | `max_depth` | 2 (100% of folds) | 0.797 | **0.691** (−13%) |
| Gaussian Process | `n_restarts` | 7 (53% of folds) | 0.459 | **0.459** (±0%) |
| Power Law | — | — | 1.033 | **1.033** (±0%) |

**Figure 12** – Residual analysis for SVR (RBF) under nested LOOCV. Left: residuals versus predicted Ra; dotted lines mark $\pm 1\sigma$ ($\sigma = 0.189$ $\mu$m). No systematic trend is visible across the Ra range. Right: residual distribution, centred near zero (mean = 0.066 $\mu$m, std = 0.189 $\mu$m).




of predicted Ra: there is no evidence of the heteroscedastic pattern of increasing error with increasing Ra that typically afflicts multiplicative models. Second, the residual distribution is approximately symmetric and concentrated within $\pm 0.4$ $\mu$m, with only three observations exceeding this range. These three outliers correspond to centre-point replicates, where the five repeated measurements at identical conditions ($V_c = 180$ m/min, $f = 0.175$ mm/rev) introduce an inherent floor on prediction error: no model can predict which specific replicate value will be observed without memorising the run order, and SVR correctly predicts the conditional mean of the replicates rather than any individual observation.

The hyperparameter configuration selected by nested grid search provides additional insight into the data structure. The inner loop converged on $C = 500$, $\varepsilon = 0.05$, and $\gamma = 0.1$ in the majority of folds, with $\varepsilon$ and $\gamma$ fixed across all 38 folds and $C = 500$ selected

in 61% of folds. The small insensitive tube ($\varepsilon = 0.05$ $\mu$m) reflects the low noise level of the profilometer measurements: the data are repeatable enough that the model benefits from fitting closely to the training observations rather than absorbing small residuals into the margin.

The fixed kernel bandwidth $\gamma = 0.1$ across all folds suggests that the response surface is smooth enough to be well-approximated by a single Gaussian kernel without requiring adaptive bandwidth. The elevated $C$ value indicates that the model tolerates few margin violations, consistent with a dataset where the dominant variance arises from genuine physical gradients in the cutting conditions rather than from measurement noise.

The superiority of SVR over RSM and GPR, despite their comparable RMSE values, is partly attributable to the RBF kernel's ability to capture the non-linear interaction between feed rate and nose radius without committing to a fixed polynomial structure. The theoretical formula of Equation 2.1 establishes that Ra scales as $f^2/r_\varepsilon$, a multiplicative interaction that a second-order RSM polynomial can approximate but not represent exactly. The RBF kernel implicitly learns this interaction from the data without being constrained to any specific functional form (Cortes; Vapnik, 1995). GPR achieves a similar flexibility through its kernel structure but incurs a modest accuracy penalty relative to SVR, which Jian Chen *et al.* (2024) attribute to the GPR's tendency to over-smooth predictions in regions with sparse observations, precisely the star-point conditions at extreme feed rates that define the upper end of the Ra range in this dataset.

### 4.2.3 Feature Importance and Physical Consistency

Although XGBoost ranked fourth in predictive accuracy, its feature-importance analysis provides a complementary data-driven perspective on the relative contribution of the input variables. This interpretation complements the second-order response surfaces obtained with RSM rather than replacing their analysis. Figure 13 shows the relative gain assigned to each machining parameter.

As shown in Figure 13, feed rate $f$ dominates with 78.9% of total importance, and nose radius $r_\varepsilon$ contributes 15.2%. Cutting speed $V_c$ and depth of cut $a_p$ account for 2.3% and 3.6%, respectively. This hierarchy is consistent with the theoretical framework established in Section 4.1: the formula $Ra = f^2/(32\, r_\varepsilon)$ assigns influence exclusively to $f$ and $r_\varepsilon$, and the XGBoost importance decomposition indicates that the same two variables dominate the empirical response. The marginal contributions of $V_c$ and $a_p$ detected by XGBoost may reflect deviations from the ideal geometric model associated with additional process mechanisms, including tool–workpiece dynamics, built-up edge formation, and thermal effects, which are not represented by the theoretical formulation.

This convergence between the data-driven importance ranking, the RSM response-surface analysis, and the physics-based formula provides complementary evidence of the physical consistency of the relationships observed in the dataset. Feed rate remains the

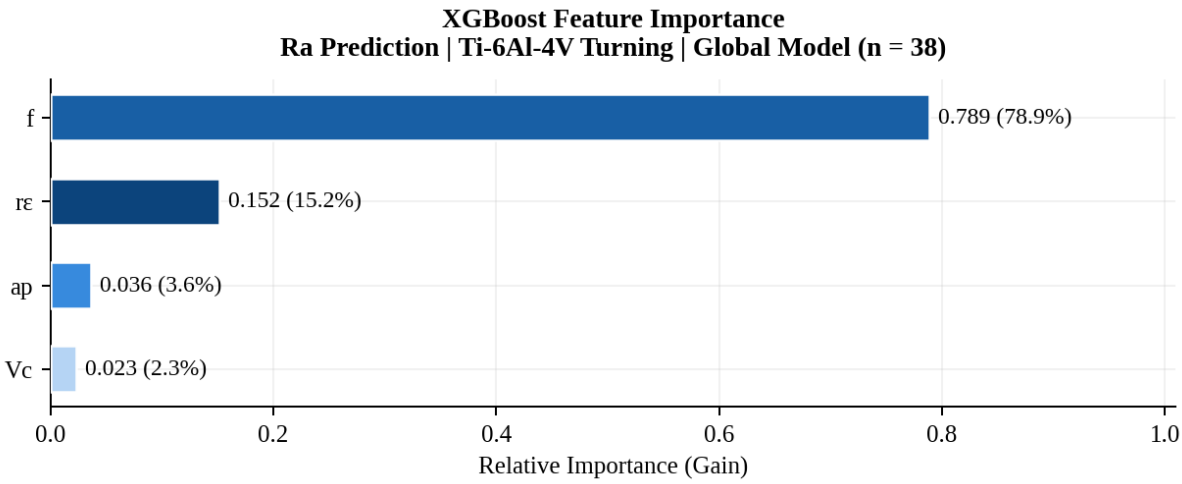



**Figure 13** – XGBoost feature importance by relative gain for Ra prediction in Ti-6Al-4V turning. Feed rate $f$ accounts for 78.9% of the total importance, followed by nose radius $r_\varepsilon$ at 15.2%. Cutting speed $V_c$ and depth of cut $a_p$ together contribute less than 6%, consistent with the theoretical formula $Ra = f^2/(32\,r_\varepsilon)$ and with the Pearson correlation analysis of Section 4.1.

dominant factor governing Ra within the investigated experimental domain, while nose radius provides a secondary but substantial contribution. Cutting speed and depth of cut exhibit comparatively smaller contributions to the response within the tested ranges.

## 4.3 LLM Model Performance

Three large language models were evaluated as zero-training-effort predictors of surface roughness Ra in Ti-6Al-4V turning: Claude Sonnet 4.5 (Anthropic), GPT-4o-mini (OpenAI), and Gemini 3 Flash Preview (Google DeepMind). Each model was accessed exclusively through its official application programming interface (API) rather than through interactive chat interfaces. API access with a fixed generation parameter (temperature = 0) ensures that every prediction is made under controlled, reproducible conditions, which is a prerequisite for any meaningful statistical comparison.

In a chat-based interaction, session history, interface-level post-processing, and user-specific settings introduce uncontrolled variation that makes cross-model comparisons unreliable. API access also mirrors the integration pathway relevant to industrial deployment, where autonomous prediction pipelines must operate without human intervention, as Ouerghemmi and Ertz (2025) document across the surveyed landscape of LLM deployments in digital manufacturing.

Even at temperature zero, residual stochasticity persists as an artefact of floating-point non-determinism in distributed GPU inference (Ouyang *et al.*, 2022). To quantify this effect, each model was evaluated across three fully independent LOOCV runs conducted in separate API sessions. Results are reported as mean ± standard deviation across runs,

following the practice recommended for LLM benchmarking by Liang et al. (Liang *et al.*, 2023), who demonstrated that single-run metrics may not represent stable model behaviour.

### 4.3.1 Prompt Engineering Strategy

Each model received the same prompt structure, constructed programmatically for every LOOCV fold. Two prompting strategies were evaluated in parallel: few-shot learning and chain-of-thought (CoT) reasoning. Listings 1 and 2 present the two prompt templates used throughout the experiments.

In the few-shot condition, the prompt consisted of a system message instructing the model to act as a precision machining expert and to respond with a single positive number, followed by a user message containing the 37 training observations formatted as input–output pairs and a final line requesting the prediction for the held-out point (Listing 1).

```
System:
You are a precision machining expert. Your task is to predict
surface roughness Ra (um) in turning of Ti-6Al-4V titanium alloy.
Always respond with ONLY a single positive number (the Ra value
in um). No explanation, no units, no text - just the number.

User:
Below are experimental measurements of surface roughness Ra (um)
during turning of Ti-6Al-4V:

Vc=135.0 m/min, f=0.0500 mm/r, ap=0.1500 mm, re=0.40 mm -> Ra=0.6700 um
Vc=225.0 m/min, f=0.0500 mm/r, ap=0.1500 mm, re=0.40 mm -> Ra=0.3600 um
[... 35 additional examples ...]

Now predict Ra for:
Vc=180.0 m/min, f=0.1750 mm/r, ap=0.2500 mm, re=0.40 mm -> Ra=?

Respond with only the numeric value.
```

**Listing 1** – Few-shot prompt structure (abbreviated). Each LOOCV fold replaces the 37 examples with the actual training points and the target line with the held-out observation.

This format draws directly on the in-context learning paradigm introduced by Brown *et al.* (2020), in which a language model generalises from input–output demonstrations provided at inference time without any weight updates. The key advantage for the present application is that the entire experimental dataset serves as context in each fold, effectively asking the model to interpolate over a 4-dimensional machining parameter space using pattern recognition alone.

In the chain-of-thought condition, shown in Listing 2, the prompt was augmented with the theoretical surface roughness formula and an explicit instruction to reason step by step before providing a numeric answer on the final line (Wei *et al.*, 2022b; Kojima *et al.*, 2023). The physical prior embedded in this prompt, Equation 2.1, was deliberately chosen because it encodes the dominant physical relationship in the data (Pearson $r = 0.992$ with observed Ra, as established in Section 4.1). The intention was to anchor the model's reasoning chain to this formula and allow the experimental examples to provide the empirical correction for real-world deviations from ideal geometry.

```
System:
You are a precision machining expert predicting Ra (um) for
Ti-6Al-4V turning. You may reason briefly, but your LAST LINE
must be ONLY the numeric Ra prediction - nothing else.

User:
You are predicting Ra (um) for Ti-6Al-4V turning.

PHYSICAL PRIOR: The theoretical roughness formula is:
    Ra_theoretical = f^2 / (32 * re)

For the target point, Ra_theoretical =
    0.1750^2 / (32 * 0.40) * 1000 = 2.3926 um

EXPERIMENTAL DATA (use to correct for real-world deviations):
Vc=135.0 m/min, f=0.0500 mm/r, ap=0.1500 mm, re=0.40 mm -> Ra=0.6700 um
[... 36 additional examples ...]

TARGET:
Vc=180.0 m/min, f=0.1750 mm/r, ap=0.2500 mm, re=0.40 mm

Think step by step, then on your LAST LINE write only the number.
```

**Listing 2** – Chain-of-thought prompt structure (abbreviated). The theoretical prior from Equation 2.1 is computed for each fold and embedded numerically in the prompt before the experimental examples.

The `max_tokens` parameter was set to 32 for few-shot responses and 512 for CoT responses, allowing additional output space for the intermediate reasoning requested in the latter strategy. Numeric outputs were extracted by retrieving the last numeric value in the response string, a convention that handles both the bare numeric outputs of few-shot responses and CoT responses containing intermediate reasoning. Predictions outside the physically plausible range $[0.01, 50]\,\mu$m were treated as invalid responses and excluded from metric computation.

### 4.3.2 Results and Model Behaviour

Table 7 summarises the LOOCV performance of all LLM configurations across three independent runs. Figure 14 presents the mean RMSE with standard deviation error bars for each configuration, and Figure 15 shows the individual run values per model to visualise stability directly.

Table 7 summarises the LOOCV performance of all LLM configurations across three independent runs. For brevity, model–strategy combinations are denoted with the compact labels *FewShot* and *CoT* in tables, figures, and comparative statements (e.g., *Claude FewShot*); the terms *few-shot* and *chain-of-thought* are used in prose to refer to the prompting strategies in general.

**Table 7** – LLM LOOCV results reported as mean ± std across three independent runs. MAPE is omitted for models with negative $R^2$, where the metric becomes uninformative.

| Model | RMSE ($\mu$m) | $R^2$ | MAPE (%) | $n_{\text{valid}}/38$ |
|---|---|---|---|---|
| Claude FewShot | $0.288 \pm 0.016$ | $0.985 \pm 0.002$ | 17.3 | 38 |
| Gemini CoT | $0.542 \pm 0.316$ | $0.927 \pm 0.076$ | — | $\approx 38$ |
| GPT FewShot | $1.239 \pm 0.048$ | $0.715 \pm 0.023$ | 26.0 | 38 |
| GPT CoT | $1.619 \pm 0.000$ | $0.514 \pm 0.000$ | 51.2 | 38 |
| Gemini FewShot | $2.114 \pm 0.021$ | $0.172 \pm 0.023$ | 20.5 | $\approx 22$ |
| Claude CoT | $3.029 \pm 0.513$ | $-0.640 \pm 0.349$ | — | $\approx 28$ |

**Figure 14** – Mean RMSE ± std across three independent LOOCV runs per LLM configuration. The dashed line marks the best ML model (SVR, RMSE = 0.341 $\mu$m) as a reference. Error bar width reflects run-to-run variability; negligible bars indicate near-deterministic behaviour.

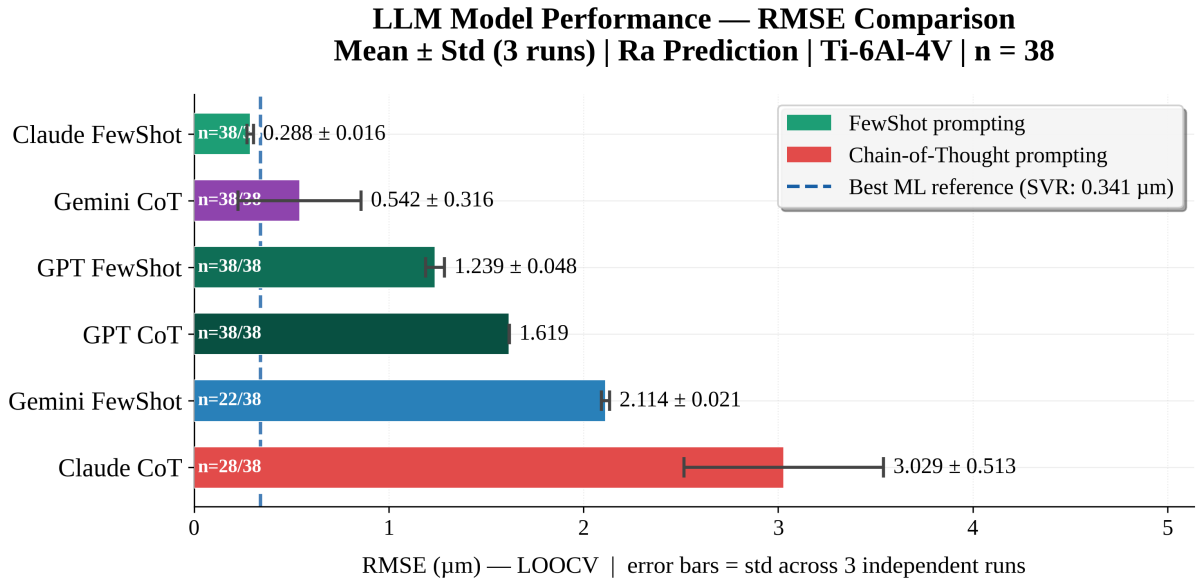
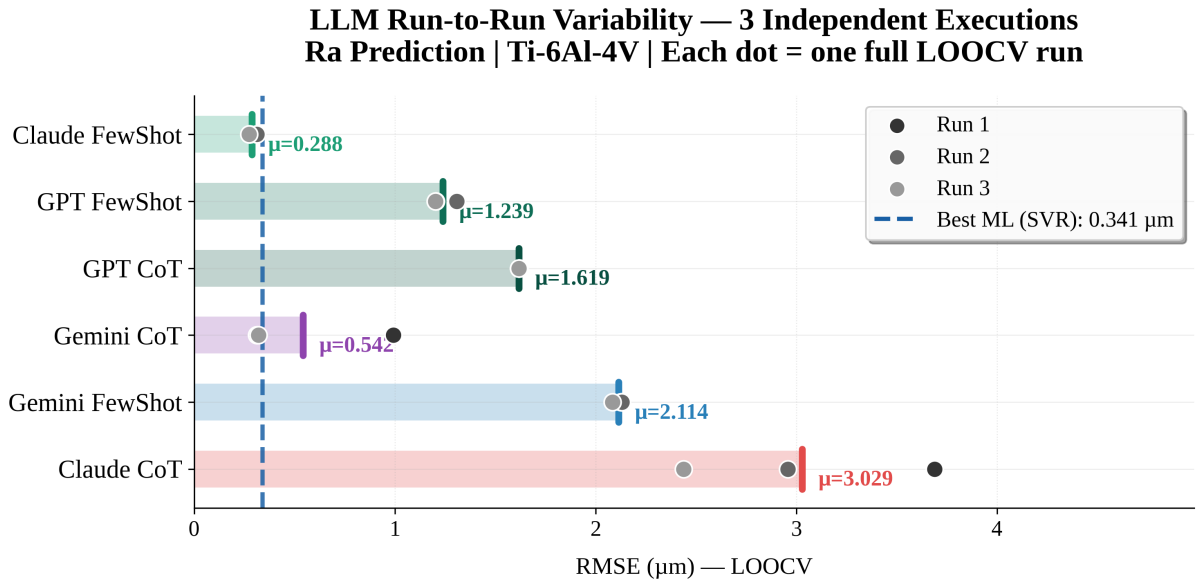


**Figure 15** – Individual RMSE values across three independent runs per LLM configuration. Each dot is one complete LOOCV experiment; the vertical line marks the mean. Tightly clustered dots indicate stable behaviour; widely separated dots reveal sensitivity to inference-time stochasticity.



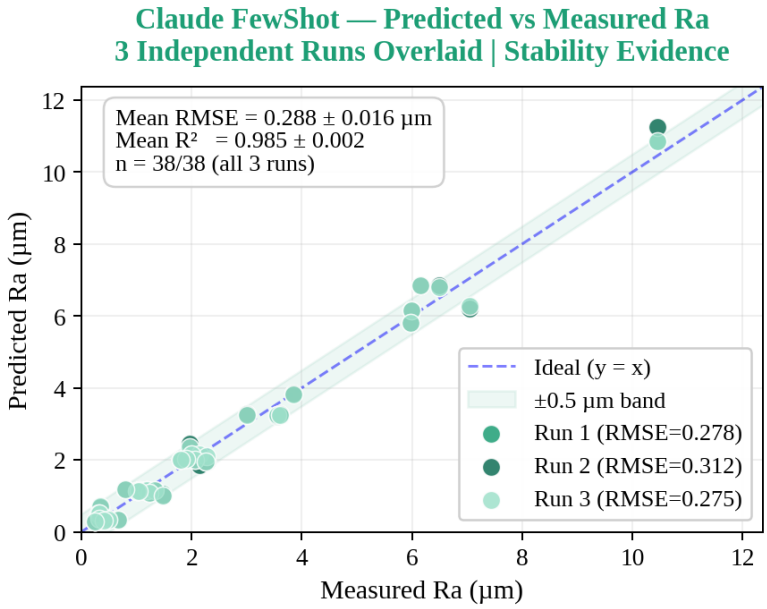

As shown in Figure 14 and Table 7, Claude FewShot achieves the lowest RMSE among all LLM configurations, followed at a considerable distance by Gemini CoT. The remaining models fall below the ML reference line, with Claude CoT producing the worst result overall. The strip plot in Figure 15 reveals that this ranking is reliable for most models but must be interpreted with caution for Gemini CoT, whose three runs span a wide range.

**Claude FewShot.** Claude FewShot achieved a mean RMSE of $0.288 \pm 0.016$ $\mu$m ($R^2 = 0.985 \pm 0.002$), outperforming the best ML model, SVR, by approximately 16%. As visible in Figure 15, the three runs cluster tightly around the mean, with a standard deviation of 0.016 $\mu$m representing less than 6% of that mean — a degree of consistency comparable to a deterministic algorithm. Figure 16 confirms this stability: the predicted-versus-measured scatter plots for the three runs are nearly indistinguishable, with all points clustered tightly around the identity line and the majority falling within the $\pm 0.5$ $\mu$m band. The model returned valid predictions for all 38 hold-out points in every run, indicating no content-filter interference.

Inspection of the raw API responses reveals a consistent prediction strategy: for each held-out point, the model identified the nearest neighbours in the 4-dimensional parameter space and interpolated among them. Feed rate and nose radius dominated the interpolation, consistent with Equation 2.1 and with the feature importance distribution identified by XGBoost ($f$: 78.9%, $r_\varepsilon$: 15.2%, see Section 4.2).

**GPT FewShot and GPT CoT.** GPT-4o-mini in few-shot mode achieved RMSE = $1.239 \pm 0.048$ $\mu$m ($R^2 = 0.715$), placing it below all ML models except Power Law. As seen in Figure 15, the three runs are similarly clustered, confirming stable but limited performance.

**Figure 16** – Predicted vs. measured Ra for Claude FewShot across three independent runs. The near-complete overlap of the scatter clouds confirms near-deterministic behaviour (RMSE $= 0.288 \pm 0.016$ $\mu$m, $R^2 = 0.985 \pm 0.002$).




Inspection of the raw responses reveals a systematic pattern: the model frequently returned an Ra value drawn directly from the training set rather than generating an interpolated prediction. When predicting a point with $f = 0.175$ mm/rev and $r_\varepsilon = 0.40$ mm, for instance, the model consistently predicted 1.90 $\mu$m — the median of the five centre-point replicates visible in the training examples — regardless of the specific held-out value. This nearest-point retrieval strategy degrades performance on axial and factorial points that are far from the design centre.

The GPT CoT configuration was the most deterministic model evaluated (Figure 15), returning identical predictions across all three runs (RMSE $= 1.619 \pm 0.000$ $\mu$m). The reasoning chains revealed that GPT applied the same procedure in every fold: compute the theoretical value from Equation 2.1, identify the most similar training points, and report a weighted average. This locked-in algorithm produced consistent but inaccurate predictions, as averaging discarded the fine-grained interpolation signal present in the examples.

This result aligns with observations by Kojima *et al.* (2023), who noted that chain-of-thought reasoning can lead models to over-formalise a problem and ignore relevant empirical patterns. The model that reasoned most systematically performed worse than the one that reasoned not at all.

**Gemini FewShot.** Gemini 3 Flash Preview in few-shot mode presented a distinct type of limitation. With a mean RMSE of $2.114 \pm 0.021$ $\mu$m and an average completion rate of only 58% (approximately 22 valid predictions per run), the model failed at a more fundamental level. As shown consistently across the three runs in Figure 15, the invalid responses arose from the API returning an empty content block with `finish_reason =`

2, corresponding to safety filter activation in Google's content moderation system (Google DeepMind, 2024). The refusals occurred on the same specific folds across all three runs, predominantly those with low feed rates ($f = 0.035$ and 0.050 mm/rev), indicating that certain prompt patterns consistently triggered the classifier. For industrial deployment, a model that silently refuses to answer roughly one third of all queries is functionally inoperable regardless of its accuracy on valid responses.

**Gemini CoT and Intermittent Instability.** Gemini CoT produced the most complex result pattern of any model evaluated. As visible in Figure 15, the three run values span a far wider range than any other model. The mean RMSE of $0.542 \pm 0.316$ $\mu$m places the model second overall, but the standard deviation exceeds the mean RMSE of most other configurations, making the summary statistic alone an inadequate description of performance.

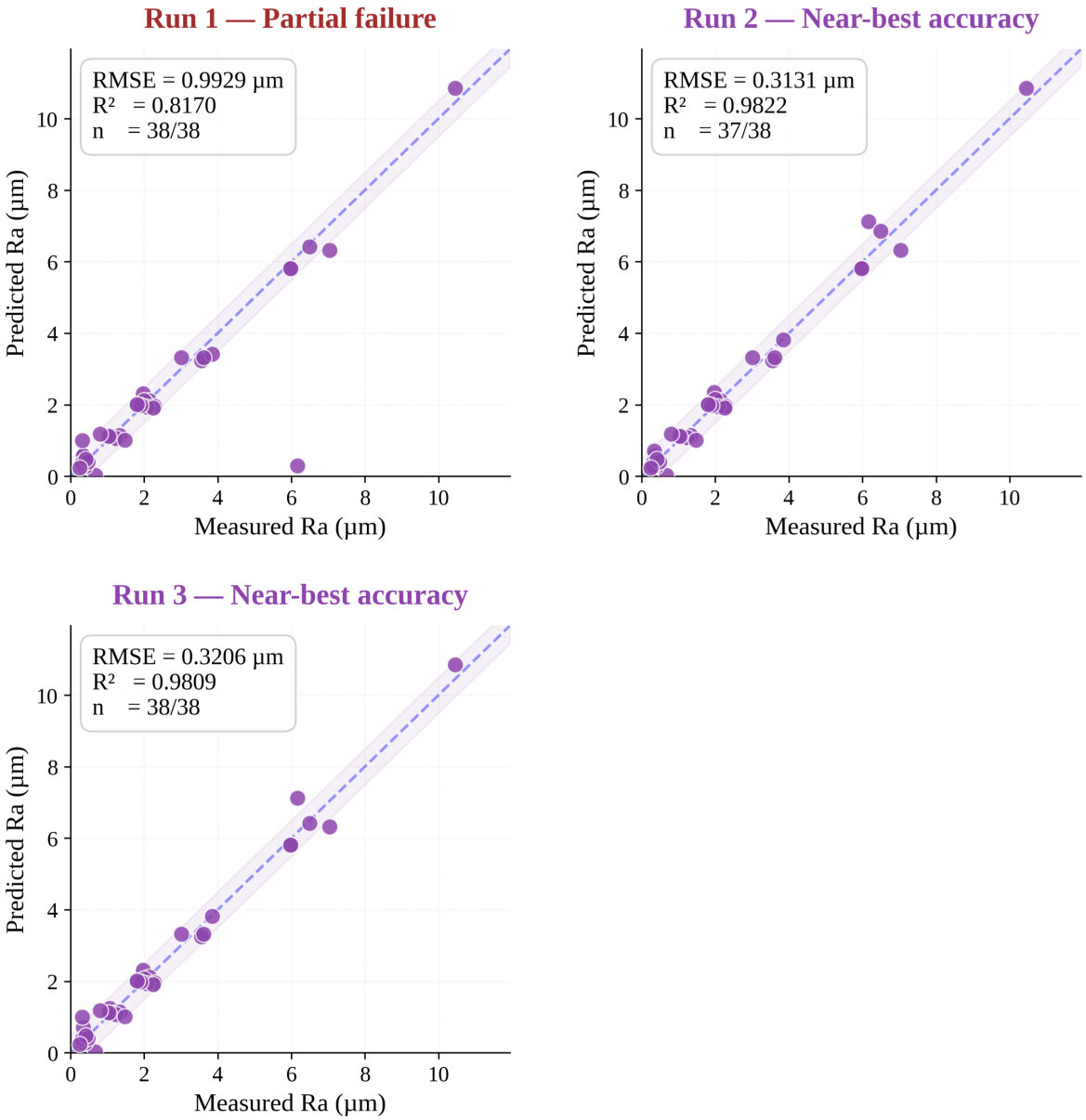
The run-level decomposition shown in Figure 17 reveals the source of this variability. In Run 1, the model achieved RMSE = 0.993 $\mu$m with $R^2 = 0.817$, adequate but clearly degraded relative to the other two runs, in which it achieved RMSE = 0.313 and 0.321 $\mu$m with $R^2 \approx 0.98$. In those two runs, the model approached Claude FewShot's accuracy and outperformed every ML model evaluated. The scatter plots in Figure 17 make the contrast visible: Runs 2 and 3 show near-perfect alignment with the identity line, whereas Run 1 exhibits wider dispersion with several outlying predictions.

Inspection of the raw responses suggests that in Run 1, several reasoning chains were truncated before completion, likely due to internal generation limits, which degraded the numeric output. When the reasoning was allowed to complete, as in Runs 2 and 3, the model produced the most physically detailed responses of any configuration evaluated, including explicit references to Equation 2.1, systematic identification of neighbouring experimental points, and careful correction for real-world deviations.

**Claude CoT.** Claude's chain-of-thought configuration produced the worst performance of any model evaluated (RMSE = $3.029 \pm 0.513$ $\mu$m, $R^2 = -0.640$). A negative coefficient of determination means the predictions are less informative than simply predicting the global sample mean, an outcome worse than a naive baseline. The failure mode can be identified directly from the raw responses. When the held-out point corresponded to one of the five center-point replicates ($V_c = 180$ m/min, $f = 0.175$ mm/rev, $a_p = 0.25$ mm, $r_\varepsilon = 0.40$ mm), the model correctly identified that identical conditions appear in the training set and computed their mean as its prediction. While statistically defensible as a description of the replicates, this is methodologically wrong in the LOOCV framework, where the goal is to predict the specific held-out value, not to summarise the visible ones.

When this averaging heuristic was applied to non-replicate points, it anchored predictions to the wrong reference values and generated large residuals. The root cause is a structural feature of the Central Composite Design: the five center-point replicates created a pattern that made averaging appear locally rational while globally damaging predictive

**Figure 17** – Predicted versus measured Ra for Gemini CoT across three independent runs. Run 1 (top left) shows partial failure (RMSE = 0.993 $\mu$m, $R^2 = 0.817$), while Runs 2 and 3 achieve near-best-in-class accuracy (RMSE $\approx$ 0.32 $\mu$m, $R^2 \approx$ 0.98). Identical prompts and data were used in all three runs, demonstrating the risk of evaluating preview-stage LLMs from a single execution.



accuracy. This failure mode was consistent across all three runs, confirming that it reflects a systematic model behaviour rather than random variance.

### 4.3.3 FewShot vs Chain-of-Thought

A finding that cuts across all three models is that few-shot prompting outperformed chain-of-thought prompting in every case where a meaningful comparison is possible. This is visible in both Figure 14 and Figure 15, where the FewShot bars consistently sit to the left of their CoT counterparts for the same underlying model. Claude FewShot (RMSE = 0.288 $\mu$m) outperformed Claude CoT (RMSE = 3.029 $\mu$m) by a factor of ten. GPT FewShot (RMSE = 1.239 $\mu$m) outperformed GPT CoT (RMSE = 1.619 $\mu$m) by 24%.

This reversal of the expected benefit of chain-of-thought reasoning is consistent with the boundaries of CoT applicability identified in the literature. Wei *et al.* (2022b) demonstrated CoT gains predominantly on tasks requiring multi-step symbolic reasoning, such as mathematical word problems and logical deduction. Surface roughness prediction is a different kind of problem: the answer is found by pattern-matching to nearby experimental observations, not by following a physical derivation to a conclusion.

The CoT prompt anchors the model to Equation 2.1 and instructs it to reason from first principles. As established in Section 4.1, that formula underestimates actual Ra by factors of up to 5.2 in the present dataset, so the anchor is systematically wrong. The few-shot prompt provides no such scaffold, which freed the model to bypass the inadequate theoretical prior and rely directly on the empirical patterns in the context window.

## 4.4 Comparative Analysis: ML vs LLM

The ML and LLM results allow a direct comparison between two fundamentally different approaches to Ra prediction: models that learn exclusively from experimental data through an explicit training procedure, and models that draw on knowledge encoded implicitly during large-scale pre-training on scientific and technical text. The unified RMSE ranking in Figure 18 and the side-by-side scatter comparison in Figure 19 provide the primary visual evidence for the discussion that follows.

### 4.4.1 Performance Hierarchy and the Role of Prompting Strategy

Looking at Figure 18, the eleven configurations do not distribute continuously across the RMSE axis. Three tiers emerge. The first comprises Claude FewShot, SVR, RSM, and GPR, all with RMSE below 0.46 $\mu$m and $R^2$ above 0.96. A second tier groups Gemini CoT, XGBoost, and Power Law. Below these, the remaining LLM configurations produce RMSE values between 1.2 and 3.0 $\mu$m, consistently worse than any trained ML model. That the

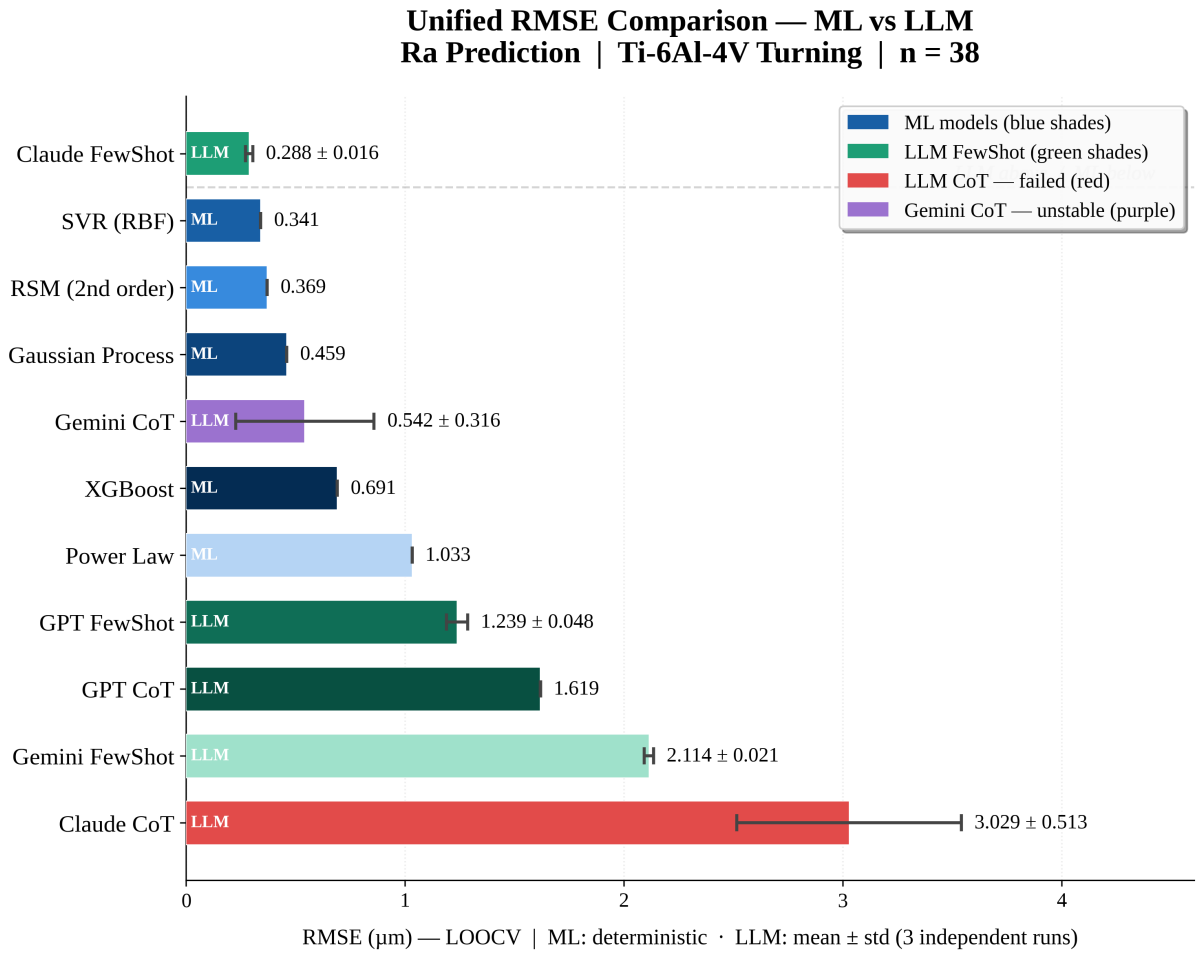


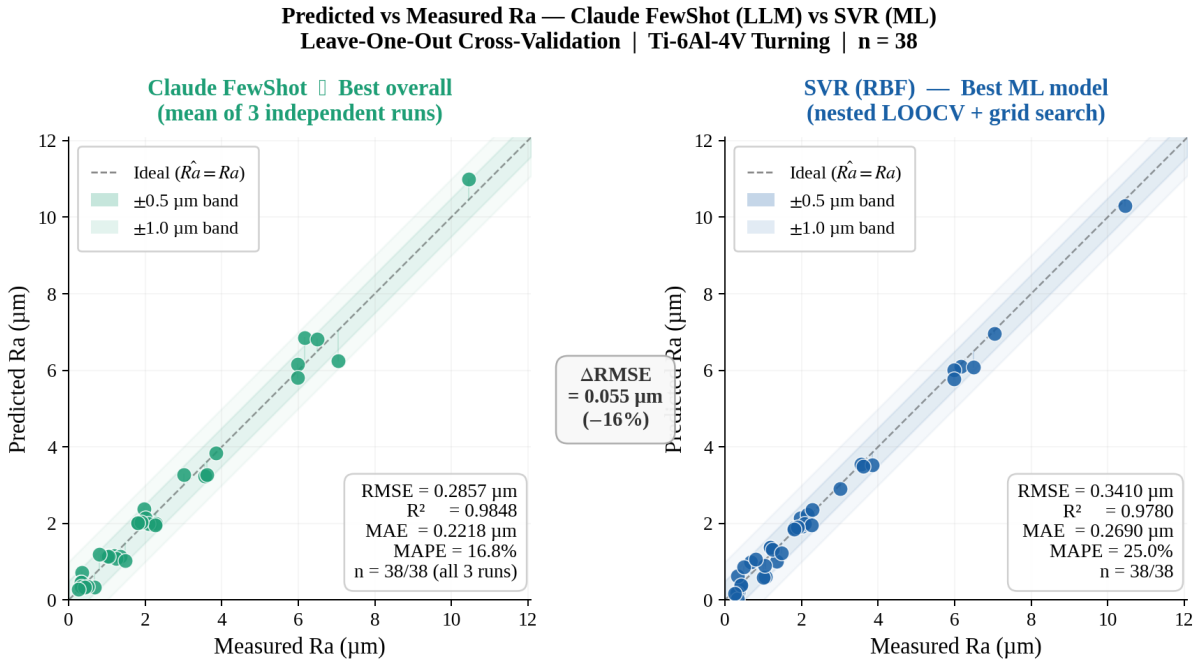

**Figure 18** – Unified LOOCV RMSE ranking for all ML and LLM configurations. ML models are deterministic; LLM results are mean ± std across three runs. The dashed line marks the boundary above which Claude FewShot is the only LLM to outperform all ML models. Gemini CoT shows anomalous run-to-run variability (std = 0.316 $\mu$m).

first tier is led by an LLM, followed by three trained models, is the central finding of this work.

Claude FewShot achieved RMSE = $0.288 \pm 0.016$ $\mu$m against SVR's 0.341 $\mu$m, a 16% reduction in prediction error relative to the best trained model, obtained without a single gradient update, without hyperparameter tuning, and with access only to the 37 training examples in the prompt context. The scatter plots in Figure 19 show that the advantage holds across the full Ra range: both models track the identity line closely, but Claude's residuals are tighter, particularly in the intermediate range ($1 < Ra < 4$ $\mu$m) where XGBoost had shown the largest deviations.

The effect of prompting strategy was model dependent. Few-shot prompting substantially outperformed CoT for Claude, with Claude FewShot achieving an RMSE approximately ten times lower than Claude CoT, and it also provided a 24% reduction in RMSE relative to GPT CoT for the GPT model. Gemini exhibited the opposite behaviour: its CoT configuration achieved considerably lower mean RMSE than Gemini FewShot,

**Figure 19** – Predicted vs. measured Ra for Claude FewShot (left) and SVR (right) under LOOCV ($N = 38$). Both panels share the same axis scale; vertical segments show individual residuals. The ΔRMSE annotation quantifies the advantage of the LLM over the best ML model.




although with substantially greater run-to-run variability. These results indicate that the benefit of explicit reasoning cannot be generalised across LLM architectures for this regression task.

Surface roughness prediction in the present dataset is primarily an interpolation problem governed by strong empirical relationships among nearby experimental conditions rather than a purely symbolic reasoning task. The CoT prompt introduced the theoretical expression $Ra = f^2/(32\, r_\varepsilon)$ as a physical reference. Although this relationship captures the dominant trend in the experimental data (Pearson $r = 0.992$), the empirical correction factor ranges from approximately $0.7\times$ to $5.2\times$, showing that individual observations may fall either below or substantially above the geometric prediction. Consequently, reliance on the theoretical prior can be beneficial when the geometric approximation is close to the measured response, but it may also direct the reasoning away from the empirical pattern when larger deviations occur. In contrast, the few-shot strategy leaves the model free to rely directly on the 37 experimental demonstrations provided in each LOOCV fold. The observed differences between prompting strategies are therefore consistent with the broader finding that CoT is particularly useful for tasks requiring explicit multi-step reasoning, whereas its benefit is less predictable for tasks dominated by interpolation and pattern matching (Sprague *et al.*, 2025).

### 4.4.2 Stability, Reliability, and the Gemini CoT Case

Performance metrics alone do not fully characterise suitability for industrial deployment. A model that achieves low mean error on one run but produces different results on the next is not deployable in a quality-critical process, regardless of how competitive its average RMSE looks.

All five ML models are fully deterministic: given the same data, they produce identical LOOCV predictions on every run. LLMs exhibit residual stochasticity even at temperature zero, an artefact of floating-point non-determinism in distributed GPU inference (Ouyang *et al.*, 2022). The practical significance of this variability differs across models. Claude FewShot's standard deviation of 0.016 $\mu$m is negligible at 5.6% of its mean RMSE. GPT FewShot and GPT CoT are similarly stable (std $< 0.05$ $\mu$m). Gemini CoT presents a qualitatively different situation: its standard deviation of 0.316 $\mu$m exceeds the mean RMSE of SVR, RSM, and GPR, and is larger than the entire performance range of the first tier. The same prompt and dataset produced RMSE values of 0.31, 0.31, and 0.99 $\mu$m across three runs, a spread of 0.68 $\mu$m that cannot be predicted or controlled by the user. A practitioner evaluating this model on a single run would have placed it anywhere from first to sixth in the ranking, depending on which execution they observed.

This raises a methodological point that extends beyond this study. Evaluating LLMs on a single run, still common in the manufacturing literature that has begun adopting language models for prediction tasks (Ouerghemmi; Ertz, 2025; Li, Y. *et al.*, 2024), is insufficient for quantitative applications where reproducibility is a requirement. Angermeir *et al.* (2026) document this concern systematically, finding that non-determinism was a primary barrier to replication across LLM-centric empirical studies. The three-run protocol used here, with results reported as mean $\pm$ standard deviation, should be treated as a minimum standard for LLM evaluation in engineering regression: a single RMSE value for an LLM is a sample from a distribution, not a fixed quantity, and reporting it as such is misleading.

### 4.4.3 Practical Implications

The results of this comparison do not support a wholesale recommendation of one model family over the other. They support a differentiated view of appropriate use cases.

For established manufacturing processes where a documented experimental database exists, hyperparameter optimisation can be conducted rigorously, and deterministic predictions are required, SVR, RSM, and GPR offer reliable, interpretable, and fully reproducible performance. Their RMSE values between 0.34 and 0.46 $\mu$m are well within acceptable bounds for Ra monitoring in Ti-6Al-4V turning, and their deterministic nature means that the same input always produces the same prediction, a property that facilitates auditing, certification, and integration into automated process control loops.

For prototyping stages, new material variants, or field operations where a full nested cross-validation pipeline is impractical and rapid deployment is prioritised, Claude FewShot provides a compelling alternative. It requires no training infrastructure, tolerates small datasets gracefully, as 37 examples proved sufficient to place it at the top of the performance ranking, and can be deployed via a standard API call. The practical limitations are equally concrete: the model requires network access and a commercial API subscription, its behaviour may change between model versions creating longitudinal irreproducibility that does not affect static ML models, and its predictions cannot be certified independently of the model provider. These constraints should be explicitly acknowledged in any deployment protocol.

A more nuanced deployment architecture emerges from these constraints. In the early stages of process qualification, where experimental data is scarce and the cost of a full nested cross-validation pipeline is difficult to justify, Claude FewShot can serve as a rapid screening tool: a single API call with the available observations is sufficient to estimate Ra across unexplored parameter combinations and guide the next experimental run. Once the dataset reaches a size that supports reliable model training, a static ML model such as SVR can be fitted, serialised, and embedded directly into the production control loop without network dependency. In this framing, the LLM and the ML model are not competing alternatives but successive instruments in the same process maturation pathway, with the LLM providing early-stage guidance and the ML model providing certified, auditable predictions at scale. This distinction has practical relevance in aerospace and biomedical manufacturing, where Ti-6Al-4V components are subject to surface integrity requirements that may mandate traceable, deterministic prediction records, requirements that a static ML model can satisfy but that a remotely-hosted LLM currently cannot.

The remaining LLM configurations, namely GPT FewShot, GPT CoT, Gemini FewShot, and Claude CoT, do not appear competitive with either the best ML models or Claude FewShot under the conditions evaluated. GPT FewShot's systematic nearest-value retrieval behaviour, Gemini FewShot's 58% completion rate, and the structural failures of both CoT variants suggest that these configurations would require substantial prompt redesign or model-specific adaptation before reaching first-tier accuracy.

The overarching conclusion is not that LLMs replace ML models for Ra prediction, but that the performance boundary between the two is less clearly drawn than the conventional framing suggests: when a sufficiently capable LLM receives well-structured experimental demonstrations, it can match or surpass a properly optimised ML model within the interpolation region, at least for the process, material, and response variable studied here. Whether this generalises to other machining contexts, response variables beyond Ra, or fundamentally different experimental designs remains an open question and a natural direction for future work.

# 5 CONCLUSIONS

The question motivating this study was whether large language models can serve as competitive predictors of surface roughness Ra in Ti-6Al-4V turning, and to what extent their performance compares with machine learning models optimised through nested cross-validation. The answer is unambiguous: Claude FewShot achieved the lowest RMSE among all eleven configurations evaluated ($0.288 \pm 0.016$ $\mu$m, $R^2 = 0.985$), outperforming the best ML model, SVR with RBF kernel, by 16% under identical Leave-One-Out Cross-Validation conditions.

This result was obtained without any training procedure, without hyperparameter tuning, and with access only to the 37 experimental demonstrations provided in the prompt context. Among the ML models, SVR, RSM, and GPR formed a clearly superior group with RMSE below 0.46 $\mu$m and $R^2$ above 0.96, while XGBoost and the Power Law trailed substantially, the latter reflecting the known limitations of log-linear models in the presence of interaction effects and process-level non-linearities.

The comparison between prompting strategies produced a consistent and mechanistically interpretable pattern: few-shot outperformed chain-of-thought in every model without exception. Ra prediction from a CCD dataset is an interpolation problem, not a symbolic reasoning problem. The theoretical formula embedded in the CoT prompt underestimates actual Ra by factors of up to 5.2 in this dataset, as established in Section 4.1. Anchoring the model's reasoning to that formula introduced a systematic bias that the few-shot prompt simply avoided by providing no reasoning scaffold at all.

The three-run evaluation protocol revealed that run-to-run stability is a critical and underreported dimension of LLM performance. Claude FewShot exhibited near-deterministic behaviour across independent sessions (std = 0.016 $\mu$m), while Gemini CoT produced RMSE values ranging from 0.31 to 0.99 $\mu$m across three identical executions. Single-run evaluations of language models in quantitative engineering tasks are insufficient for drawing reliable conclusions.

Taken together, these findings suggest that the boundary between trained ML models and prompt-based LLMs for manufacturing quality prediction is less clearly drawn than the conventional framing implies. When a sufficiently capable language model receives well-structured experimental demonstrations of a physical process, it can match or surpass a properly optimised ML model within the interpolation region, at least for the process, material, and response variable studied here. This does not position LLMs as universal replacements for ML in machining applications; rather, it identifies a specific operational niche where they offer a genuine advantage: rapid deployment scenarios with small datasets, where the cost of assembling a training pipeline exceeds the cost of a well-designed prompt.

## 5.1 Study Limitations

Several aspects of the experimental and computational design bound the scope of the conclusions and should be considered when interpreting or extending the results.

The dataset comprised 38 observations drawn from a single material, two tool geometries, and one machine tool under dry cutting conditions. The generalisability of the LLM advantage to other titanium alloys, to coated tools under wet or MQL conditions, or to different machine–workpiece configurations has not been established and cannot be inferred from the present results.

The LLM evaluations used pre-release model versions, namely Claude Sonnet 4.5, GPT-4o-mini, and Gemini 3 Flash Preview, that were current at the time of the experiments but are subject to silent updates by their respective providers. Unlike static ML models, LLM weights may change between the submission and publication of this work, which introduces a form of longitudinal irreproducibility that cannot be controlled by the researcher. The model identifiers and API access dates are documented in the methodology to mitigate this limitation. Future work should replicate the few-shot evaluation with successor model generations, particularly GPT model and its successors, to assess whether the performance advantage of few-shot prompting over chain-of-thought is stable across model families and scales with model capability.

The three-run repetition protocol quantified run-to-run variability at temperature zero, but does not constitute a full uncertainty analysis of the LLM predictions. Confidence intervals for individual point predictions were not computed, and the interaction between prompt phrasing variations and predictive accuracy was not systematically explored. Small changes in the system message wording or the input–output formatting convention may produce different results, and the observed performance should be understood as specific to the prompt design documented in the methodology.

Finally, the comparison between ML and LLM models used RMSE as the primary criterion, which weights all prediction errors uniformly across the Ra range. In practice, the cost of a prediction error may be asymmetric, underestimating Ra on a safety-critical aerospace component is more consequential than overestimating it, and an application-specific loss function may reorder the ranking.

## 5.2 Future Work

The results of this study open several concrete research directions.

Extending the experimental database to include wet and MQL cutting conditions, additional titanium grades (Ti–6Al–2Sn–4Zr–2Mo, Ti–5Al–5V–5Mo–3Cr), and a broader range of tool geometries would allow a direct assessment of whether the LLM advantage observed here persists when the demonstrations span a more complex and heterogeneous parameter space. It would also enable investigation of whether the model's implicit

knowledge of Ti–6Al–4V specifically, likely well-represented in pre-training corpora, is the primary driver of the result, or whether the advantage generalises to less-studied alloys.

The prompting design space explored in this work was intentionally minimal: two strategies, one system message per strategy, and a fixed input–output formatting convention. A systematic prompt engineering study, varying the number of demonstrations, the order of examples, the level of physical context provided, and the output format, would clarify how sensitive the LLM performance is to these choices and whether further gains are achievable without model fine-tuning.

Fine-tuning a smaller open-source language model (Llama 3, Mistral, or Phi 3) on a curated dataset of machining experiments would allow a direct comparison between prompt-based in-context learning and parameter-efficient adaptation, potentially combining the deployment simplicity of the LLM approach with the reproducibility and offline capability of locally hosted models. This direction would also address the API dependency limitation identified above, enabling deployment in industrial environments without network connectivity requirements.

## ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## References

ABDELAAL, M.; LOKADJAJA, S.; ENGERT, G. **GLLM: Self-Corrective G-Code Generation using Large Language Models with User Feedback**. [*S. l.: s. n.*], 2025. arXiv: 2501.17584 [cs.SE]. Disponível em: https://arxiv.org/abs/2501.17584.

ALEMAYEHU, H.; ZIYAD, F.; WOGASO, D.; GEBRESLASSIE, M. G.; MANYALA, S. Extreme gradient boosting for prediction of surface roughness of mild steel AISI 1018 under dry and air-assisted machining. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 140, p. 3337–3354, 2025. DOI: 10.1007/s00170-025-16352-7. Disponível em: https://doi.org/10.1007/s00170-025-16352-7.

ANGERMEIR, F. *et al.* Reflections on the Reproducibility of Commercial LLM Performance in Empirical Software Engineering Studies. *In:* PROCEEDINGS of the 48th IEEE/ACM International Conference on Software Engineering (ICSE). [*S. l.: s. n.*], 2026. Disponível em: https://arxiv.org/abs/2510.25506.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v. 4, p. 40–79, 2010. DOI: 10.1214/09-SS054. Disponível em: https://doi.org/10.1214/09-SS054.

ARNAIZ-GONZÁLEZ, Á.; FERNÁNDEZ-VALDIVIELSO, A.; BUSTILLO, A., *et al.* Using artificial neural networks for the prediction of dimensional error on inclined surfaces manufactured by ball-end milling. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 83, p. 847–859, 2016. DOI: 10.1007/s00170-015-7543-y. Disponível em: https://doi.org/10.1007/s00170-015-7543-y.

BADINI, S.; REGONDI, S.; FRONTONI, E.; PUGLIESE, R. Assessing the capabilities of ChatGPT to improve additive manufacturing troubleshooting. **Advanced Industrial and Engineering Polymer Research**, Elsevier, v. 6, n. 3, p. 278–287, 2023. DOI: 10.1016/j.aiepr.2023.03.003. Disponível em: https://doi.org/10.1016/j.aiepr.2023.03.003.

BAGGA, P. J.; PATEL, K. M.; MAKHESANA, M. A., *et al.* Machine vision-based gradient-boosted tree and support vector regression for tool life prediction in turning. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 126, n. 1–2, p. 471–485, May 2023. DOI: 10.1007/s00170-023-11137-2. Disponível em: https://doi.org/10.1007/s00170-023-11137-2.

BENARDOS, P. G.; VOSNIAKOS, G.-C. Predicting surface roughness in machining: a review. **International Journal of Machine Tools and Manufacture**, Elsevier, v. 43, n. 8, p. 833–844, June 2003. DOI: 10.1016/S0890-6955(03)00059-2. Disponível em: https://doi.org/10.1016/S0890-6955(03)00059-2.

BOMMASANI, R. *et al.* **On the Opportunities and Risks of Foundation Models**. [*S. l.: s. n.*], 2022. arXiv: 2108.07258 [cs.LG]. Disponível em: https://arxiv.org/abs/2108.07258.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. **Language Models are Few-Shot Learners**. [*S. l.: s. n.*], 2020. arXiv: 2005.14165 [cs.CL]. Disponível em: https://arxiv.org/abs/2005.14165.

BUTT, M. M.; HAQ, M. ul; KVVSSN, V., *et al.* Machine learning and physics based empirical equations for fatigue life prediction of AM Ti-6Al-4V. **Progress in Additive Manufacturing**, Mar. 2026. Received: 31 October 2025 / Accepted: 22 February 2026 / Published: 13 March 2026. ISSN 2363-9520. DOI: 10.1007/s40964-026-01615-w. Disponível em: https://doi.org/10.1007/s40964-026-01615-w.

CAWLEY, G. C.; TALBOT, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. **Journal of Machine Learning Research**, v. 11, n. 70, p. 2079–2107, 2010. Disponível em: https://www.jmlr.org/papers/v11/cawley10a.html.

ÇAYDAŞ, U.; EKICI, S. Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel. **Journal of Intelligent Manufacturing**, Springer, v. 23, n. 3, p. 639–650, Jan. 2012. DOI: 10.1007/s10845-010-0415-2. Disponível em: https://doi.org/10.1007/s10845-010-0415-2.

CHEN, J.; LIN, J.; ZHANG, M.; LIN, Q. Predicting Surface Roughness in Turning Complex-Structured Workpieces Using Vibration-Signal-Based Gaussian Process Regression. **Sensors**, MDPI, v. 24, n. 7, p. 2117, 2024. DOI: 10.3390/s24072117. Disponível em: https://doi.org/10.3390/s24072117.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *In:* PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [*S. l.*]: ACM, 2016. p. 785–794. DOI: 10.1145/2939672.2939785. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785.

CHOWDHERY, A.; NARANG, S.; DEVLIN, J.; BOSMA, M.; MISHRA, G.; ROBERTS, A.; BARHAM, P.; CHUNG, H. W.; SUTTON, C.; GEHRMANN, S.; SCHUH, P.; SHI, K.; TSVYASHCHENKO, S.; MAYNEZ, J.; RAO, A.; BARNES, P.; TAY, Y.; SHAZEER, N.; PRABHAKARAN, V.; REIF, E.; DU, N.; HUTCHINSON, B.; POPE, R.; BRADBURY, J.; AUSTIN, J.; ISARD, M.; GUR-ARI, G.; YIN, P.; DUKE, T.; LEVSKAYA, A.; GHEMAWAT, S.; DEV, S.; MICHALEWSKI, H.;

GARCIA, X.; MISRA, V.; ROBINSON, K.; FEDUS, L.; ZHOU, D.; IPPOLITO, D.; LUAN, D.; LIM, H.; ZOPH, B.; SPIRIDONOV, A.; SEPASSI, R.; DOHAN, D.; AGRAWAL, S.; OMERNICK, M.; DAI, A. M.; PILLAI, T. S.; PELLAT, M.; LEWKOWYCZ, A.; MOREIRA, E.; CHILD, R.; POLOZOV, O.; LEE, K.; ZHOU, Z.; WANG, X.; SAETA, B.; DIAZ, M.; FIRAT, O.; CATASTA, M.; WEI, J.; MEIER-HELLSTERN, K.; ECK, D.; DEAN, J.; PETROV, S.; FIEDEL, N. **PaLM: Scaling Language Modeling with Pathways**. [*S. l.: s. n.*], 2022. arXiv: 2204.02311 [cs.CL]. Disponível em: https://arxiv.org/abs/2204.02311.

CHOWDHURY, S.; ARUNACHALAM, N. Surface functionalization of additively manufactured titanium alloy for orthopaedic implant applications. **Journal of Manufacturing Processes**, v. 102, p. 387–405, 2023. ISSN 1526-6125. DOI: 10.1016/j.jmapro.2023.07.015. Disponível em: https://www.sciencedirect.com/science/article/pii/S1526612523007004.

CICA, D.; TESIC, S.; SREDANOVIC, B.; VUJASIN, D.; ZELJKOVIC, M.; PUSAVEC, F.; KRAMAR, D. Prediction of Specific Energy Consumption in Sustainable Milling of Ti-6Al-4V with Different Machine Learning Models. **Metals**, MDPI, v. 16, n. 3, p. 266, 2026. DOI: 10.3390/met16030266. Disponível em: https://doi.org/10.3390/met16030266.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer, v. 20, n. 3, p. 273–297, 1995. DOI: 10.1007/BF00994018.

DAS, A.; DAS, S. R.; PANDA, J. P.; DEY, A.; GAJRANI, K. K.; SOMANI, N.; GUPTA, N. K. Machine Learning-Based Modeling and Optimization in Hard Turning of AISI D6 Steel with Advanced AlTiSiN-Coated Carbide Inserts to Predict Surface Roughness and Other Machining Characteristics. **Surface Review and Letters**, v. 29, n. 10, p. 2250137, 2022. DOI: 10.1142/S0218625X22501372. Disponível em: https://doi.org/10.1142/S0218625X22501372.

DENG, C.; YE, B.; LU, S.; HE, M.; MIAO, J. On-line surface roughness classification for multiple CNC milling conditions based on transfer learning and neural network. **The International Journal of Advanced Manufacturing Technology**, v. 128, n. 3-4, p. 1063–1076, 2023. DOI: 10.1007/s00170-023-11997-8. Disponível em: https://link.springer.com/article/10.1007/s00170-023-11997-8.

DINIZ, A. E.; MARCONDES, F. C.; COPPINI, N. L. **Tecnologia da Usinagem dos Materiais**. 9th. São Paulo: Artliber, 2014.

DONG, Q.; LI, L.; DAI, D.; ZHENG, C.; MA, J.; LI, R.; XIA, H.; XU, J.; WU, Z.; LIU, T.; CHANG, B.; SUN, X.; LI, L.; SUI, Z. **A Survey on In-context Learning**. [*S. l.: s. n.*], 2024. arXiv: 2301.00234 [cs.CL]. Disponível em: https://arxiv.org/abs/2301.00234.

DU, J.; WANG, Y.; ZHOU, X.; MEI, X.; LIU, H.; TAO, T. Prediction model and optimization of energy consumption, cutting force, and surface roughness during machine

tool cutting process based on high-order response surface methodology. **Journal of the Brazilian Society of Mechanical Sciences and Engineering**, Springer, v. 46, n. 444, p. 1–18, 2024. DOI: 10.1007/s40430-024-05012-8. Disponível em: https://doi.org/10.1007/s40430-024-05012-8.

DUBEY, V.; SHARMA, A. K.; PIMENOV, D. Y. Prediction of Surface Roughness Using Machine Learning Approach in MQL Turning of AISI 304 Steel by Varying Nanoparticle Size in the Cutting Fluid. **Lubricants**, MDPI, v. 10, n. 5, p. 81, Jan. 2022. DOI: 10.3390/lubricants10050081. Disponível em: https://doi.org/10.3390/lubricants10050081.

EZUGWU, E. O.; WANG, Z.-M. Titanium alloys and their machinability—a review. **Journal of Materials Processing Technology**, Elsevier, v. 68, n. 3, p. 262–274, 1997. DOI: 10.1016/S0924-0136(96)00030-1. Disponível em: https://doi.org/10.1016/S0924-0136(96)00030-1.

GOOGLE DEEPMIND. **Safety Settings and Filters — Gemini API**. [*S. l.: s. n.*], 2024. Google AI for Developers. Accessed: 2025. Disponível em: https://ai.google.dev/gemini-api/docs/safety-settings.

GROSSMANN, F.; BASTEN, S.; KIRSCH, B.; ANKENER, W.; SMAGA, M.; BECK, T.; UEBEL, J.; SEEWIG, J.; AURICH, J. C. Predictive modelling of cryogenic hard turning of AISI 52100 based on response surface methodology for the use in soft sensors. **Procedia CIRP**, Elsevier, v. 108, p. 270–275, 2022. DOI: 10.1016/j.procir.2022.04.070. Disponível em: https://doi.org/10.1016/j.procir.2022.04.070.

HOANG, D.; GORSICH, D.; CASTANIER, M. P.; IMANI, F. Knowledge Graph Fusion with Large Language Models for Accurate, Explainable Manufacturing Process Planning. **arXiv preprint arXiv:2506.13026**, 2025. License: arXiv.org perpetual non-exclusive license. arXiv: 2506.13026 [cs.AI].

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970. DOI: 10.1080/00401706.1970.10488634. Disponível em: https://homepages.math.uic.edu/~lreyzin/papers/ridge.pdf.

HU, L.; GAO, R.; GUO, Y. B. Transfer Learning Enhanced Gaussian Process Model for Surface Roughness Prediction with Small Data. **Manufacturing Letters**, Elsevier, v. 35, p. 1103–1108, Aug. 2023. Supplement. DOI: 10.1016/j.mfglet.2023.08.094. Disponível em: https://doi.org/10.1016/j.mfglet.2023.08.094.

HU, Y.; WANG, Y.; XI, J.; CHEN, A.; NIKBIN, K. Numerical simulation of surface roughness effects on low-cycle fatigue properties of additively manufactured titanium alloys. **Engineering Failure Analysis**, v. 163, p. 108407, 2024. ISSN 1350-6307. DOI: 10.1016/j.engfailanal.2024.108407. Disponível em: https://www.sciencedirect.com/science/article/pii/S1350630724004539.

JIN, M.; WANG, S.; MA, L.; CHU, Z.; ZHANG, J. Y.; SHI, X.; CHEN, P.-Y.; LIANG, Y.; LI, Y.-F.; PAN, S.; WEN, Q. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *In:* INTERNATIONAL Conference on Learning Representations (ICLR). [*S. l.: s. n.*], 2024. Disponível em: https://openreview.net/forum?id=9Ym979uAb1.

JONES, D. R.; SCHONLAU, M.; WELCH, W. J. Efficient Global Optimization of Expensive Black-Box Functions. **Journal of Global Optimization**, v. 13, p. 455–492, 1998. DOI: 10.1023/A:1008306431147.

KAPLAN, J.; MCCANDLISH, S.; HENIGHAN, T.; BROWN, T. B.; CHESS, B.; CHILD, R.; GRAY, S.; RADFORD, A.; WU, J.; AMODEI, D. **Scaling Laws for Neural Language Models**. [*S. l.: s. n.*], 2020. arXiv: 2001.08361 [cs.LG]. Disponível em: https://arxiv.org/abs/2001.08361.

KHANGHAH, K. N.; PATEL, A.; MALHOTRA, R.; XU, H. Large language models for extrapolative modeling of manufacturing processes. **Journal of Intelligent Manufacturing**, June 2025. Received: 05 February 2025 / Accepted: 26 May 2025 / Published: 24 June 2025. ISSN 1572-8145. DOI: 10.1007/s10845-025-02638-w. Disponível em: https://doi.org/10.1007/s10845-025-02638-w.

KO, J. H. Transformer Encoder vs. Mamba SSM: Lightweight Architectures for Machining Stability-Induced Surface-Quality Categorization. **Big Data and Cognitive Computing**, v. 10, n. 1, p. 1, 2026. DOI: 10.3390/bdcc10010001. Disponível em: https://www.mdpi.com/2504-2289/10/1/1.

KOJIMA, T.; GU, S. S.; REID, M.; MATSUO, Y.; IWASAWA, Y. **Large Language Models are Zero-Shot Reasoners**. [*S. l.: s. n.*], 2023. arXiv: 2205.11916 [cs.CL]. Disponível em: https://arxiv.org/abs/2205.11916.

KOSARAC, A.; TABAKOVIC, S.; MLADJENOVIC, C.; ZELJKOVIC, M.; ORASANIN, G. Next-Gen Manufacturing: Machine Learning for Surface Roughness Prediction in Ti-6Al-4V Biocompatible Alloy Machining. **Journal of Manufacturing and Materials Processing**, MDPI, v. 7, n. 6, p. 202, 2023. DOI: 10.3390/jmmp7060202. Disponível em: https://doi.org/10.3390/jmmp7060202.

KUNTOĞLU, M.; ASLAN, A.; PIMENOV, D. Y.; USCA, Ü. A.; SALUR, E.; GUPTA, M. K.; MIKOLAJCZYK, T.; GIASIN, K.; KAPŁONEK, W.; SHARMA, S. A Review of Indirect Tool Condition Monitoring Systems and Decision-Making Methods in Turning: Critical Analysis and Trends. **Sensors**, MDPI, v. 21, n. 1, p. 108, 2021. DOI: 10.3390/s21010108. Disponível em: https://doi.org/10.3390/s21010108.

LI, W.; WANG, S.; YANG, X.; DUAN, H.; WANG, Y.; YANG, Z. Research Progress on Fatigue Damage and Surface Strengthening Technology of Titanium Alloys for Aerospace Applications. **Metals**, v. 15, n. 2, p. 192, Feb. 2025. Special Issue: Mechanical Properties, Fatigue and Fracture of Metallic Materials. ISSN 2075-4701. DOI: 10.3390/met15020192. Disponível em: https://doi.org/10.3390/met15020192.

LI, Y.; ZHAO, H.; JIANG, H.; PAN, Y.; LIU, Z.; WU, Z.; SHU, P.; TIAN, J.; YANG, T.; XU, S.; LYU, Y.; BLENK, P.; PENCE, J.; RUPRAM, J.; BANU, E.; LIU, N.; WANG, L.; SONG, W.; ZHAI, X.; SONG, K.; ZHU, D.; LI, B.; WANG, X.; LIU, T. **Large Language Models for Manufacturing**. [*S. l.: s. n.*], 2024. arXiv: 2410.21418 [cs.AI]. Disponível em: https://arxiv.org/abs/2410.21418.

LI, Z.; ZHANG, P.; ZHOU, F., *et al.* Wear and wear suppression of carbide tool in intermittent cutting: a review. **The International Journal of Advanced Manufacturing Technology**, v. 142, p. 5531–5569, Jan. 2026. DOI: 10.1007/s00170-026-17527-6. Disponível em: https://doi.org/10.1007/s00170-026-17527-6.

LIANG, P.; BOMMASANI, R.; LEE, T.; TSIPRAS, D.; SOYLU, D.; YASUNAGA, M.; ZHANG, Y.; NARAYANAN, D.; WU, Y.; KUMAR, A.; NEWMAN, B.; YUAN, B.; YAN, B.; ZHANG, C.; COSGROVE, C.; MANNING, C. D.; RÉ, C.; ACOSTA-NAVAS, D.; HUDSON, D. A.; ZELIKMAN, E.; DURMUS, E.; LADHAK, F.; RONG, F.; REN, H.; YAO, H.; WANG, J.; SANTHANAM, K.; ORR, L.; ZHENG, L.; YUKSEKGONUL, M.; SUZGUN, M.; KIM, N.; GUHA, N.; CHATTERJI, N.; KHATTAB, O.; HENDERSON, P.; HUANG, Q.; CHI, R.; XIE, S. M.; SANTURKAR, S.; GANGULI, S.; HASHIMOTO, T.; ICARD, T.; ZHANG, T.; CHAUDHARY, V.; WANG, W.; LI, X.; MAI, Y.; ZHANG, Y.; KOREEDA, Y. Holistic Evaluation of Language Models. **Transactions on Machine Learning Research**, 2023. Disponível em: https://arxiv.org/abs/2211.09110.

LIU, H.; YANG, M.; ADOMAVICIUS, G. Just Because You Can, Doesn't Mean You Should: LLMs for Data Fitting. **arXiv preprint arXiv:2508.19563**, Aug. 2025. License: CC BY 4.0. arXiv: 2508.19563 [cs.LG]. Disponível em: https://arxiv.org/abs/2508.19563.

LIU, H.; YANG, M.; ADOMAVICIUS, G. **Robustness is Important: Limitations of LLMs for Data Fitting**. [*S. l.: s. n.*], 2025. arXiv: 2508.19563 [cs.LG]. Disponível em: https://arxiv.org/abs/2508.19563.

LIU, Y.; DING, S.; ZHOU, S.; FAN, W.; TAN, Q. **MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction**. [*S. l.: s. n.*], June 2024. arXiv: 2406.12950 [q-bio.QM]. Disponível em: https://arxiv.org/abs/2406.12950.

LYU, C.; LIU, X.; MIHAYLOVA, L. Review of Recent Advances in Gaussian Process Regression Methods. *In:* PANOUTSOS, G.; MAHFOUF, M.; MIHAYLOVA, L. S. (eds.). **Advances in Computational Intelligence Systems**. Cham: Springer, 2024. v. 1454. (Advances in Intelligent Systems and Computing). ISBN 978-3-031-55568-8. DOI: 10.1007/978-3-031-55568-8_19. Disponível em: https://doi.org/10.1007/978-3-031-55568-8_19.

MAITRA, V.; SHI, J. Evaluating the Predictability of Surface Roughness of Ti-6Al-4V Alloy from Selective Laser Melting. **Advanced Engineering Materials**, v. 25, n. 14,

p. 2300075, 2023. DOI: 10.1002/adem.202300075. Disponível em: https://doi.org/10.1002/adem.202300075.

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. Large Language Models for Design and Manufacturing. **An MIT Exploration of Generative AI**, MIT Case Studies in Social and Ethical Responsibilities of Computing, Mar. 2024. DOI: 10.21428/e4baedd9.745b62fa. Disponível em: https://doi.org/10.21428/e4baedd9.745b62fa.

MARIN, E.; LANZUTTI, A. Biomedical Applications of Titanium Alloys: A Comprehensive Review. **Materials**, v. 17, n. 1, p. 114, 2024. ISSN 1996-1944. DOI: 10.3390/ma17010114. Disponível em: https://doi.org/10.3390/ma17010114.

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?** [*S. l.: s. n.*], 2022. arXiv: 2202.12837 [cs.CL]. Disponível em: https://arxiv.org/abs/2202.12837.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9. ed. [*S. l.*]: Wiley, 2017.

NOOR, K.; SIDDIQUI, M. A.; IQBAL, S. A. Multi-objective Optimization of Parameters in CNC Turning of a Hardened Alloy Steel Roll by Using Response Surface Methodology. **Arabian Journal for Science and Engineering**, Springer, v. 48, n. 3, p. 3403–3423, Mar. 2023. DOI: 10.1007/s13369-022-07117-5. Disponível em: https://doi.org/10.1007/s13369-022-07117-5.

OUERGHEMMI, C.; ERTZ, M. Integrating Large Language Models into Digital Manufacturing: A Systematic Review and Research Agenda. **Computers**, v. 14, n. 8, p. 318, Aug. 2025. Special Issue: AI in Complex Engineering Systems. ISSN 2073-431X. DOI: 10.3390/computers14080318. Disponível em: https://doi.org/10.3390/computers14080318.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C. L.; MISHKIN, P.; ZHANG, C.; AGARWAL, S.; SLAMA, K.; RAY, A.; SCHULMAN, J.; HILTON, J.; KELTON, F.; MILLER, L.; SIMENS, M.; ASKELL, A.; WELINDER, P.; CHRISTIANO, P.; LEIKE, J.; LOWE, R. **Training language models to follow instructions with human feedback**. [*S. l.: s. n.*], 2022. arXiv: 2203.02155 [cs.CL]. Disponível em: https://arxiv.org/abs/2203.02155.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. **WIREs Data Mining and Knowledge Discovery**, Wiley, v. 9, n. 3, e1301, 2019. DOI: 10.1002/widm.1301. Disponível em: https://doi.org/10.1002/widm.1301.

RASMUSSEN, C. E.; WILLIAMS, C. K. I. **Gaussian Processes for Machine Learning**. Cambridge, MA: MIT Press, 2006. ISBN 026218253X. Disponível em: https://direct.mit.edu/books/oa-monograph/2320/Gaussian-Processes-for-Machine-Learning.

REEBER, T.; WOLF, J.; MÖHRING, H.-C. A Data-Driven Approach for Cutting Force Prediction in FEM Machining Simulations Using Gradient Boosted Machines. **Journal of Manufacturing and Materials Processing**, MDPI, v. 8, n. 3, p. 107, May 2024. DOI: 10.3390/jmmp8030107. Disponível em: https://doi.org/10.3390/jmmp8030107.

ROY, A.; CHAKRABORTY, S. Support vector machine in structural reliability analysis: A review. **Reliability Engineering & System Safety**, Elsevier, v. 233, p. 109126, May 2023. DOI: 10.1016/j.ress.2023.109126. Disponível em: https://doi.org/10.1016/j.ress.2023.109126.

SHAW, M. C. **Metal Cutting Principles**. 2nd. Oxford: Oxford University Press, 2005.

SIVA SURYA, M. Optimization of turning parameters while turning Ti-6Al-4V titanium alloy for surface roughness and material removal rate using response surface methodology. **Materials Today: Proceedings**, v. 62, p. 3479–3484, 2022. International Conference on Materials, Processing & Characterization (13th ICMPC). ISSN 2214-7853. DOI: 10.1016/j.matpr.2022.04.300. Disponível em: https://doi.org/10.1016/j.matpr.2022.04.300.

ŠKET, M. *et al.* Large language models for G-code generation in CNC machining: A comparison of ChatGPT-3.5 and ChatGPT-4o. **Advances in Production Engineering & Management**, v. 20, n. 2, p. 155–168, 2025. DOI: 10.14743/apem2025.2.498. Disponível em: https://doi.org/10.14743/apem2025.2.498.

SOUZA, A. F. de; CASTRO CESÁRIO, M. d.; SILVA CAMPOS, P. H. da; PAIVA, A. P. de; VERRI, F. A. N.; BALESTRASSI, P. P. Physics-based feature engineering framework for high-accuracy surface roughness prediction in Ti-6Al-4V dry machining. **The International Journal of Advanced Manufacturing Technology**, v. 143, p. 2631–2648, 2026. DOI: 10.1007/s00170-026-17731-4.

SPRAGUE, Z.; YIN, F.; RODRIGUEZ, J. D.; JIANG, D.; WADHWA, M.; SINGHAL, P.; ZHAO, X.; YE, X.; MAHOWALD, K.; DURRETT, G. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. *In:* INTERNATIONAL Conference on Learning Representations. [*S. l.: s. n.*], 2025. Disponível em: https://openreview.net/forum?id=w6nlcS8Kkn.

SRIVASTAVA, M.; JAYAKUMAR, V.; UDAYAN, Y.; M, S.; S M, M.; GAUTAM, P.; NAG, A. Additive manufacturing of Titanium alloy for aerospace applications: Insights into the process, microstructure, and mechanical properties. **Applied Materials Today**, v. 41, p. 102481, 2024. ISSN 2352-9407. DOI: 10.1016/j.apmt.2024.102481. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352940724004268.

SURESH, G.; RAMESH, M. R.; SRINATH, M. S. Surface Engineered Titanium Alloys for Biomedical, Automotive, and Aerospace Applications. *In:* VIGNESH, R. V.; PADMANABAN, R.; GOVINDARAJU, M. (eds.). **Advances in Processing of Lightweight Metal Alloys and Composites**. Singapore: Springer, Nov. 2023. (Materials Horizons: From Nature to Nanomaterials). p. 85–106. ISBN 978-981-19-7145-7. DOI: 10.1007/978-981-19-7146-4_5. Disponível em: https://doi.org/10.1007/978-981-19-7146-4_5.

TOGNAN, A.; LAURENTI, L.; SALVATI, E. Contour Method with Uncertainty Quantification: A Robust and Optimised Framework via Gaussian Process Regression. **Experimental Mechanics**, Springer, v. 62, p. 1305–1317, Apr. 2022. DOI: 10.1007/s11340-022-00842-w. Disponível em: https://doi.org/10.1007/s11340-022-00842-w.

VABALAS, A.; GOWEN, E.; POLIAKOFF, E.; CASSON, A. J. Machine learning algorithm validation with a limited sample size. **PLoS ONE**, v. 14, n. 11, e0224365, 2019. DOI: 10.1371/journal.pone.0224365. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365#:~:text=November%207%2C%202019-,https%3A//doi.org/10.1371/journal.pone.0224365,-Article.

VAROQUAUX, G. Cross-validation failure: Small sample sizes lead to large error bars. **NeuroImage**, v. 180, p. 68–77, 2018. DOI: 10.1016/j.neuroimage.2017.06.061.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. **Attention Is All You Need**. [*S. l.: s. n.*], 2017. arXiv: 1706.03762 [cs.CL]. Disponível em: https://arxiv.org/abs/1706.03762.

VILAS BOAS PAPPI MACIEL, N.; FERNANDES DE SOUZA, A.; SILVA CAMPOS, P. H. da; ZAMBRONI DE SOUZA, A. C.; VERRI, F. A. N.; BALESTRASSI, P. P. Physics-Informed Symbolic Regression Ensemble (PISRE) for surface roughness prediction in Ti-6Al-4V dry turning. **The International Journal of Advanced Manufacturing Technology**, v. 144, p. 2253–2277, 2026. DOI: 10.1007/s00170-026-17994-x.

WANG, J. An Intuitive Tutorial to Gaussian Process Regression. **Computing in Science & Engineering**, IEEE, v. 25, n. 4, p. 4–11, 2023. DOI: 10.1109/MCSE.2023.3342149. Disponível em: https://doi.org/10.1109/MCSE.2023.3342149.

WANG, X.; FENG, C. Development of Empirical Models for Surface Roughness Prediction in Finish Turning. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 20, p. 348–356, Sept. 2002. DOI: 10.1007/s001700200162. Disponível em: https://doi.org/10.1007/s001700200162.

WEI, J.; TAY, Y.; BOMMASANI, R.; RAFFEL, C.; ZOPH, B.; BORGEAUD, S.; YOGATAMA, D.; BOSMA, M.; ZHOU, D.; METZLER, D.; CHI, E. H.;

HASHIMOTO, T.; VINYALS, O.; LIANG, P.; DEAN, J.; FEDUS, W. Emergent Abilities of Large Language Models. **Transactions on Machine Learning Research**, 2022. Survey Certification. ISSN 2835-8856. Disponível em: https://openreview.net/forum?id=yzkSU5zdwD.

WEI, J.; WANG, X.; SCHUURMANS, D.; BOSMA, M.; ICHTER, B.; XIA, F.; CHI, E. H.; LE, Q. V.; ZHOU, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *In:* ADVANCES in Neural Information Processing Systems. [*S. l.: s. n.*], 2022. v. 35, p. 24824–24837. Disponível em: https://arxiv.org/abs/2201.11903.

YANG, W.-H.; TARNG, Y.-S. Design optimization of cutting parameters for turning operations based on the Taguchi method. **Journal of Materials Processing Technology**, Elsevier, v. 84, n. 1-3, p. 122–129, 1998. DOI: 10.1016/S0924-0136(98)00079-X. Disponível em: https://doi.org/10.1016/S0924-0136(98)00079-X.

YAO, Z.; JIA, X.; YU, J.; YANG, M.; HUANG, C.; YANG, Z.; WANG, C.; YANG, T.; WANG, S.; SHI, R.; WEI, J.; LIU, X. Rapid accomplishment of strength/ductility synergy for additively manufactured Ti-6Al-4V facilitated by machine learning. **Materials & Design**, v. 225, p. 111559, 2023. ISSN 0264-1275. DOI: 10.1016/j.matdes.2022.111559. Disponível em: https://www.sciencedirect.com/science/article/pii/S0264127522011820.

YI, F.; ZHONG, R.; ZHU, W.; ZHOU, R.; GUO, L.; WANG, Y. Comparative Study of Friction Models in High-Speed Machining of Titanium Alloys. **Lubricants**, v. 13, n. 3, p. 113, Mar. 2025. ISSN 2075-4442. DOI: 10.3390/lubricants13030113. Disponível em: https://doi.org/10.3390/lubricants13030113.

ZHANG, C.; ZOU, D.; MAZUR, M.; MO, J. P. T.; LI, G.; DING, S. The State of the Art in Machining Additively Manufactured Titanium Alloy Ti-6Al-4V. **Materials**, MDPI, v. 16, n. 7, p. 2583, 2023. DOI: 10.3390/ma16072583. Disponível em: https://doi.org/10.3390/ma16072583.

ZHANG, S.; CAO, R.; GHALATI, M. K.; DONG, H.; WEI, Y. Integrated machine-learning modelling for mechanical property prediction – A case study on laser-welded TC4 titanium alloy. **Optics & Laser Technology**, Elsevier, v. 199, p. 115019, July 2026. DOI: 10.1016/j.optlastec.2026.115019. Disponível em: https://doi.org/10.1016/j.optlastec.2026.115019.

ZHAO, W. X.; ZHOU, K.; LI, J.; TANG, T.; WANG, X.; HOU, Y.; MIN, Y.; ZHANG, B.; ZHANG, J.; DONG, Z.; DU, Y.; YANG, C.; CHEN, Y.; CHEN, Z.; JIANG, J.; REN, R.; LI, Y.; TANG, X.; LIU, Z.; LIU, P.; NIE, J.-Y.; WEN, J.-R. **A Survey of Large Language Models**. [*S. l.: s. n.*], 2026. arXiv: 2303.18223 [cs.CL]. Disponível em: https://arxiv.org/abs/2303.18223.