Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

SteelAgent: An LLM-Orchestrated System for Physics-Informed Steel Property Prediction and Generalization Auditing

Aleksandr Volkov , Roman Sultimov , Mikhail Kuzin , Yury Maximov

Abstract

Machine learning models for steel property prediction routinely report high quality metrics with $R^2 > 0.85$, yet these results rely on random splits that allow similar grades in both train and test sets. We present *SteelAgent*, an interactive system that exposes a *critical generalization gap*: the same models drop from $R^2 > 0.85$ all the way to $R^2 = 0.11$ on unseen steel families revealing more than 7 times higher quality degradation. Similarly, conformal prediction coverage degrades from 91% to 38% under distribution shift induced by holding out substantial data sources. SteelAgent combines physics-informed features grounded in classical metallurgy and interpretable models with conformal uncertainty quantification, and an LLM orchestrator that coordinates six domain-specific tools. The system supports property prediction with specification compliance checking, competitive steel comparison, and cost-aware inverse alloy design over 3,741 heat treatment records spanning 1,234 grades. All predictions are traceable through explicit tool calls, ensuring that all physical quantities are computed, not generated. We made the code and data freely accessible to the community.

1 Introduction

Machine learning models for predicting mechanical properties of steels have achieved impressive accuracy on standard benchmarks [Wen *et al.*, 2019; Xiong *et al.*, 2020; Choudhary *et al.*, 2022]. Gradient-boosted models routinely report $R^2 > 0.85$ for tensile strength prediction, and these numbers are cited to justify deployment in alloy design and qualification pipelines. In practice, however, engineers need models that generalize to *new* steel grades compositions never seen during training because the primary value of a predictive model lies in reducing the cost and time of designing and qualifying novel alloys.

Unfortunately, high accuracy demonstrated in steel property predictions on random splits often reflects data leakage rather than genuine predictive capability [Kapoor and Narayanan, 2023; Kapoor and Narayanan, 2023]. The root cause is compositional similarity leakage: random train/test splits allow nearly identical steel grades (e.g., AISI 4340 and AISI 4330, which differ by only 0.10% of carbon) belonging to the same family to appear in both sets. The models memorize grade-specific composition bundles instead of learning the underlying metallurgical relationships between chemistry, heat treatment, and mechanical response.

When evaluated on *unseen steel families* via Leave-One-Family-Out (LOFO) splitting, where entire categories such as "stainless austenitic" or "CrMo quenched-and-tempered" are held out, the gradient boosting regression quality drops from $R^2 = 0.86$ to $R^2 = 0.11$ representing more than 7.6 times degradation as illustrated in Fig. 1. For safety-critical applications where material qualification decisions depend on predicted properties (aircraft landing gear, nuclear pressure vessels, offshore pipelines) this gap between reported and real-world performance has direct engineering consequences imposing substantial risk for infrastructure and personnel.

We present SteelAgent, an interactive system addressing this gap through three key contributions:

- *No-leak evaluation:* comparison of four split protocols, exposing that conformal coverage degrades from 91% to 38% under distribution shift [Tibshirani *et al.*, 2019].
- *Physics-grounded LLM orchestration:* an agent that enforces tool-mediated reasoning [Yao *et al.*, 2023; Schick *et al.*, 2023], computing physical quantities (M_s martensite start temperature, CE carbon equivalent for weldability, DI ideal critical diameter for hardenability) via verified formulas rather than LLM generation.
- *End-to-end decision support:* from natural-language queries to specification compliance probabilities $P(\text{pass})$ and cost-aware inverse design.

2 System Architecture

SteelAgent[1] follows a three-layer architecture (see Fig. 2): a *physics layer* that computes domain-specific features from classical metallurgical formulas, a *model layer* that runs machine learning algorithms with conformal uncertainty quantification, and an *LLM orchestration layer* that coordinates these components through structured function calling. The

[1]Demo video: https://www.youtube.com/watch?v=BwVBJ-SwuQo. Live demo: https://steelagent.vercel.app

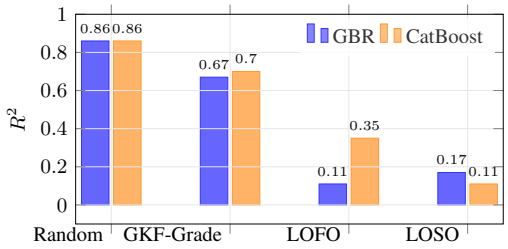


Figure 1: R^2 for tensile strength across four evaluation protocols of increasing stringency. The gap between Random, Leave-One-Family-Out (LOFO; exclude cluster with similar properties), Leave-One-Source-Out (LOSO; exclude single literature source) reveals the extent to which models rely on within-grade similarity. Dataset: $n = 1{,}636$ core heats.

Parameter	Value
Total heats (records)	3,741
Core benchmark heats	1,636
Unique grades	1,234 (594 core)
Steel families	15
Independent sources	27
Composition inputs	9 elements + 2 HT
Physics-derived features	18
Target properties	UTS, YS

Table 1: SteelBench dataset summary (v2.0).

system operates over 3,741 heat treatment records spanning 1,234 grades from 27 independent sources (see Tab. 1), with pre-trained models (XGBoost, LightGBM, CatBoost, Random Forest) for two target properties: ultimate tensile strength (UTS is the maximum stress a material can withstand before fracture) and yield strength (YS is the stress at which the material begins to deform permanently). These two properties are the most commonly specified in industrial standards and drive material selection decisions.

Physics-Informed Features. A key challenge for AI models in metallurgy is that raw composition vectors (e.g., 0.40% C, 0.80% Cr) do not directly encode the physical mechanisms that determine mechanical properties. Two steels with similar compositions can behave very differently depending on how their alloying elements interact during phase transformations. To bridge this gap, SteelAgent computes 18 physics-derived features from 11 raw inputs (C, Mn, Si, Cr, Ni, Mo, V, Cu, Al, T_{aust}, T_{temper}). These features can be grouped as: (i) Martensite start temperature $M_s = 539 - 423C - 30.4\text{Mn} - 12.1\text{Cr} - 17.7\text{Ni} - 7.5\text{Mo} + 10\text{Si}$ [Andrews, 1965] predicts when martensite, the hard phase responsible for strength, begins to form during quenching; (ii) Carbon equivalent $CE_{IIW} = C + \frac{\text{Mn}}{6} + \frac{\text{Cr+Mo+V}}{5} + \frac{\text{Ni+Cu}}{15}$ summarizes the combined effect of all alloying elements on weldability and hardenability into a single scalar; (iii) Ideal critical diameter DI [Grossmann and Bain, 1964] the maximum bar diameter that achieves full hardening at center quantifies how deeply a steel can be hardened, a key property for large structural components; (iv) Ac1/Ac3 transformation temperatures, superheat $\Delta T = T_{\text{aust}} - \text{Ac3}$, tempering resistance index, and maximum martensite hardness. These features serve a dual purpose: they improve Leave-One-Family-Out R^2 by $+0.25$ for gradient boosting regression (see Fig. 3), and they ground

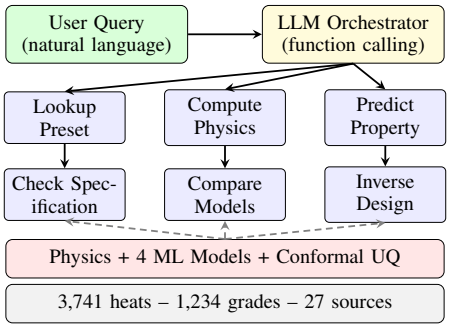


Figure 2: SteelAgent three-layer architecture. The LLM orchestrator dispatches queries to six domain tools; all physical quantities are computed by verified formulas, not generated by the LLM.

the LLM orchestrator in verified physical quantities, preventing hallucination of numerical values.

Uncertainty Quantification. In engineering practice, a point prediction alone (e.g., "UTS = 1,746 MPa") is insufficient for decision-making: the engineer needs to know *how confident* the model is. A prediction interval $[\hat{y}_{\text{lo}}, \hat{y}_{\text{hi}}]$ answers this directly. We apply split conformal prediction [Vovk *et al.*, 2005; Romano *et al.*, 2019; Angelopoulos and Bates, 2023] to construct distribution-free prediction intervals with guaranteed $\geq 90\%$ marginal coverage *under exchangeability* meaning that at least 90% of true values fall within the predicted interval, with no distributional assumptions on the data.

For specification compliance, we convert conformal intervals to pass probabilities via Gaussian approximation: $\sigma = \hat{q}/z_{0.95}$, then $P(\text{pass}) = \Phi((\hat{y} - y_{\min})/\sigma)$, where $y_{\min}$ is the specification minimum. This gives the engineer a single number e.g., $P(\text{pass AMS-6414}) = 0.93$ that directly supports material qualification decisions. Additional unlabeled data may further help with algorithms quality [Amini *et al.*, 2025; Maximov *et al.*, 2018].

LLM Orchestration. A key design decision is to separate *reasoning* (what to compute and how to interpret results) from *computation* (the actual physics and ML calculations). The LLM handles the former; the latter is delegated to verified tools. Following the ReAct paradigm [Yao *et al.*, 2023] and tool-augmented LLM agents [Schick *et al.*, 2023; Bran *et al.*, 2024], SteelAgent implements a multi-turn function-calling loop over six tools: `lookup_preset` (58 reference grades with standard compositions), `compute_physics` (18 derived features), `predict_property` (ML prediction + conformal UQ), `check_spec` ($P(\text{pass})$ for 51 industrial specifications), `compare_models` (4-model consensus with disagreement detection), and `inverse_design` (cost-aware composition search). The system prompt enforces a physics-first pipeline: `compute_physics` must precede `predict_property`, and the LLM must cite tool-computed values in its natural-language synthesis. This constraint ensures that physical quantities (e.g., $M_s = 305°\text{C}$) originate from verified formulas rather than LLM generation, while the LLM provides metallurgical context and interpretation that a non-specialist can understand.

Split	R^2	RMSE	MAE	Cov.
Random K-Fold	0.86	114	73	91%
Group K-Fold (Grade)	0.67	175	105	83%
Leave-One-Family-Out	0.11	218	154	76%
Leave-One-Source-Out	0.17	277	210	38%

Table 2: Generalization gap and UQ degradation across four evaluation protocols (tensile strength, $n = 1{,}636$ core heats, GBR model).

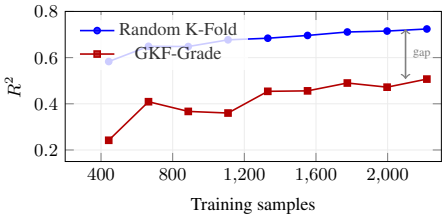


Figure 3: Learning curves (GBR, 5 seeds). Both protocols improve with more data, but the generalization gap persists at all sizes a structural limitation that motivates physics-informed features.

3 Quantitative Results and Demo Scenarios

Table 2 presents two critical findings:

- *The gap is structural, not statistical:* the drop from Random to LOFO is consistent across models (CatBoost: $0.86 \rightarrow 0.35$; GBR: $0.86 \rightarrow 0.11$) and persists even as training data increases (Fig. 3), indicating that additional data alone cannot cover physics-informed features.
- *UQ coverage tracks the gap.* Conformal coverage (guaranteed $\geq 90\%$ under exchangeability [Angelopoulos and Bates, 2023]) degrades proportionally to the distribution shift: $91\% \rightarrow 83\% \rightarrow 76\% \rightarrow 38\%$. This pattern is itself a useful diagnostic: coverage can serve as an automated shift detector, flagging predictions where the exchangeability assumption no longer valid [Tibshirani *et al.*, 2019].

SteelAgent surfaces this information interactively, helping users decide when to trust a prediction.

SteelAgent supports three primary interaction scenarios via a conversational React interface with guided dialogue trees. Users interact in natural language; the system autonomously selects and sequences tool calls, presenting results with physics context and uncertainty.

Scenario 1: Prediction + Specification Compliance. In this scenario, our task is to predict UTS for AISI 4340 (Q&T 845/315°C) and check AMS-6414 compliance. AISI 4340 is a widely used aerospace steel (landing gear, crankshafts, high-strength fasteners). The orchestrator executes a four-tool pipeline: `lookup_preset` $\rightarrow$ `compute_physics` ($M_s = 305$°C, $CE = 0.85$, $DI = 10.7$ in) $\rightarrow$ `predict_property` (UTS = 1,746 MPa, 90% CI [1,546, 1,945] MPa) $\rightarrow$ `check_spec` (P(pass AMS-6414)). The user sees physics context (what transformation behavior to expect), a prediction with calibrated uncertainty (how much to trust it), and a compliance probability (whether the steel will meet the specification).

Scenario 2: Competitive Steel Comparison. In this experiment, we compare AerMet 100 and Ferrium M54 for landing gear redesign. These are premium ultra-high-strength steels with different design philosophies. The system predicts properties for both grades ($\text{UTS}_{\text{AerMet}} = 1{,}927$ MPa; $M_s = 192$°C, $DI = 47$ in) and explains trade-offs through physics signatures: AerMet achieves peak strength via M_2C secondary hardening enabled by 13.4% cobalt, while Ferrium offers better stress-corrosion cracking resistance. This type of comparative analysis typically requires a senior metallurgist and extensive literature review; the system delivers it in seconds with auditable tool calls.

System	Steel-spec.	Inter. UI	Conf. UQ	Physics	Inv. design	LLM agent	Gen. audit
Matbench[a]	✗	✗	✗	✗	✗	✗	∼
JARVIS-ML[b]	✗	✓	✗	✓	✗	✗	✗
Coscientist[c]	✗	✗	✗	✗	✗	✓	✗
ChemCrow[d]	✗	✓	✗	∼	✗	✓	✗
SteelAgent	✓	✓	✓	✓	✓	✓	✓

[a][Dunn *et al.*, 2020] [b][Choudhary *et al.*, 2022] [c][Boiko *et al.*, 2023] [d][Bran *et al.*, 2024]

Table 3: Feature comparison with materials ML platforms and scientific LLM agents. ✓= supported, ✗= not supported, ∼ = partial.

Scenario 3: Inverse Alloy Design We look for compositions achieving UTS $> 1{,}800$ MPa at minimum alloy cost. The system screens 10,000 candidate compositions within physically feasible bounds, evaluating each through the ML pipeline. Results are ranked by estimated alloy cost and include conformal intervals, enabling distinction between alloys that *reliably vs. possibly* achieve the target a critical distinction for qualification.

Additional capabilities include SHAP-based feature importance [Lundberg and Lee, 2017] with metallurgical interpretation (e.g., identifying which alloying elements drive a specific prediction), multi-model consensus with automatic disagreement detection across 4 models, and interactive dataset exploration across 12 steel families and 27 sources.

4 Positioning

Table 3 positions SteelAgent relative to existing materials informatics platforms and scientific LLM agents.

Materials benchmarks. Matbench [Dunn *et al.*, 2020] and Matminer [Ward *et al.*, 2018] provide standardized benchmarks for materials property prediction, but focus on crystal-level properties rather than alloy-level mechanical properties with heat-treatment dependence. SteelAgent addresses this gap with steel-specific evaluation and an interactive interface.

Evaluation pitfalls. [Murdock *et al.*, 2020] [Murdock *et al.*, 2020] show that domain features provide marginal benefit under random splits, while [Kapoor and Narayanan, 2023] [Kapoor and Narayanan, 2023] document leakage across scientific ML. SteelAgent makes this problem interactive and extends it to uncertainty quantification.

Scientific LLM agents. Coscientist [Boiko *et al.*, 2023] and ChemCrow [Bran *et al.*, 2024] demonstrate domain-specific LLM agents for chemistry, [Sultimov *et al.*, 2026b; Sultimov *et al.*, 2026a] for climate and transportation. LLM-based multi-agent systems [Guo *et al.*, 2024] provide theoretical foundations. SteelAgent contributes a *physics-grounded* agent architecture where tool calls enforce correctness of numerical claims, combined with calibrated UQ and generalization auditing features absent from prior scientific agents.

References

[Amini *et al.*, 2025] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. Self-training: A survey. *Neurocomputing*, 616:128904, 2025.

[Andrews, 1965] Kenneth W. Andrews. Empirical formulae for the calculation of some transformation temperatures. *J. Iron Steel Inst.*, 203:721–727, 1965.

[Angelopoulos and Bates, 2023] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, 2023.

[Boiko *et al.*, 2023] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[Bran *et al.*, 2024] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.*, 6(5):525–535, 2024.

[Choudhary *et al.*, 2022] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.*, 8:59, 2022.

[Dunn *et al.*, 2020] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.*, 6:138, 2020.

[Grossmann and Bain, 1964] Marcus A. Grossmann and Edgar C. Bain. *Principles of Heat Treatment*. American Soc. for Metals, 5th edition, 1964.

[Guo *et al.*, 2024] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proc. IJCAI*, pages 8048–8057, 2024.

[Kapoor and Narayanan, 2023] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proc. NeurIPS*, pages 4766–4777, 2017.

[Maximov *et al.*, 2018] Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *J. Artif. Intell. Res.*, 61:761–786, 2018.

[Murdock *et al.*, 2020] Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang, and Taylor D. Sparks. Is domain knowledge necessary for machine learning materials properties? *Integr. Mater. Manuf. Innov.*, 9(3):221–227, 2020.

[Romano *et al.*, 2019] Yaniv Romano, Evan Patterson, and Emmanuel J. Candes. Conformalized quantile regression. In *Proc. NeurIPS*, pages 3538–3548, 2019.

[Schick *et al.*, 2023] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Proc. NeurIPS*, 2023.

[Sultimov *et al.*, 2026a] Roman Sultimov, Mikhail Mozikov, Dmitrii Abramov, Mariia Kovalchuk, Maksim Malykh, Ilya Makarov, Andrei Osiptsov, Aleksandr Volkov, and Yury Maximov. RESPOND: Realistic environment simulation of population and natural disasters with LLM-driven agents. In *Proc. AAAI*, pages 41694–41696, 2026.

[Sultimov *et al.*, 2026b] Roman Sultimov, Aleksandr Volkov, Mile Mitrovic, and Yury Maximov. LLM-guided multi-agent evacuation coordination via episodic memory and cognitive task analysis. In *Proc. AAMAS*, pages 4164–4166, 2026.

[Tibshirani *et al.*, 2019] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Proc. NeurIPS*, pages 2526–2536, 2019.

[Vovk *et al.*, 2005] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[Ward *et al.*, 2018] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.*, 152:60–69, 2018.

[Wen *et al.*, 2019] Cheng Wen, Yan Zhang, Changxin Wang, Dezhen Xue, Yang Bai, Stoichko Antonov, Lanhong Dai, Turab Lookman, and Yanjing Su. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.*, 170:109–117, 2019.

[Xiong *et al.*, 2020] Jie Xiong, TongYi Zhang, and SanQiang Shi. Machine learning of mechanical properties of steels. *Sci. China Technol. Sci.*, 63:1247–1255, 2020.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proc. ICLR*, 2023.