Physiscore: An Autonomous Expert-Agentic Framework for Multimodal Material Characterization in Optoelectronic Research

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce Physiscore, an autonomous system designed to bridge the gap between raw experimental signals and scientific insight in perovskite research. Although high-throughput testing has scaled data generation, interpretation remains a manual rate-limiting step, introducing subjective bias and compromising reproducibility. To address this, we implement a pragmatic AI-Coscientist framework using a natural-language interface, empowering experimentalists to orchestrate complex analytical pipelines without programming proficiency. Physiscore spans multimodal techniques including photovoltaic, optical, and structural/spectroscopic analysis. The platform bifurcates the workflow into a deterministic computational path and an intelligent reasoning path. By internalizing physical constraints—such as Scherrer analysis in XRD—directly into the algorithms, we enforce rigor that neutralizes the hallucination risks of generative models. The system automates error-prone derivations and data curation, achieving a dramatic increase in efficiency. Experimental validation demonstrates that Physiscore accelerates processing speeds by up to 131.4-fold (notably in J-V and FE-SEM analysis) and reduces human labor by an average of over 30-fold across all modalities. These outputs are synchronized with OriginLab for instantaneous visualization. This establishes a robust paradigm for materials discovery, refocusing human intelligence on experimental design rather than repetitive processing.

1 Introduction

Recent advances in materials research are undergoing a fundamental paradigm shift, driven by the expansion of the Materials Genome Initiative and the integration of physics-informed AI and automated characterization frameworks [1, 2]. Within this shifting landscape, high-throughput experimentation and machine learning-assisted acceleration have transitioned from emerging strategies to indispensable infrastructure for modern science [3]. However, while the velocity of experimental data generation has increased exponentially, the capacity for "scientific interpretation"—the extraction of fundamental insight—remains tethered to manual human effort [4]. This disparity between quantitative data expansion and qualitative interpretive stagnation represents a structural bottleneck in contemporary experimental science [5].

Specifically, core analytical tasks such as J–V (Current density-voltage) curve fitting, XRD (X-ray diffraction) phase identification, and XPS (X-ray photoelectron spectroscopy) chemical state analysis are still largely confined to analyst-dependent workflows relying on fragmented software ecosystems [6, 7]. Such practices not only stifle research productivity but also introduce subjective bias and reproducibility crises. Systematic audits have revealed that approximately 30% of manually interpreted XPS studies in the literature contain physically inconsistent or erroneous conclusions, highlighting

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

an urgent requirement for intelligent, constraint-aware analytical infrastructure [8]. Furthermore, the conversion of raw experimental signals into computer-readable datasets—specifically metadata labeling—remains one of the most time-intensive, low-value phases of the research lifecycle [9].

To alleviate these human-centered bottlenecks, Large Language Model (LLM)-based intelligent agent systems have recently emerged as promising tools [10, 11]. Pioneering systems such as Coscientist and ChemAgents have demonstrated that LLMs can meaningfully contribute to literature mining, protocol design, and autonomous experimentation [12–15]. Parallel efforts have shown potential for extracting structured numerical knowledge from unstructured literature and automating analytical code generation [16]. However, most current systems rely heavily on cloud-based infrastructures or focus primarily on text-centric extraction. Consequently, these frameworks exhibit significant limitations in supporting the integrated interpretation of heterogeneous multimodal datasets—including J–V, EQE (External quantum efficiency), XRD, and FE-SEM (Field emission scanning electron microscope)—routinely encountered in the laboratory, and they lack robust interfaces that bridge human domain expertise with AI-generated outputs [10, 17, 18].This limitation is especially critical in semiconductor device research, where reliable mechanism elucidation requires the direct correlation of crystallinity with device performance (J–V) and carrier dynamics [19, 20].

Current practices depend on inefficient tools and fragmented documentation, leading to context loss and collaborative friction [21, 22]. Moreover, the inherent hallucination behavior of LLMs remains a substantial barrier in the physical sciences, where physical rigor and interpretability are non-negotiable [23]. Beyond simple automation, there is a pressing need for an AI Co-scientist that internalizes physical constraints within its algorithms and integrates seamlessly into the researcher's daily development environment [13, 17, 22].

We present Physiscore, a Gemini-assisted agentic platform implemented within a VS Code-embedded environment to optoelectronic research. The system integrates automated pipelines for analysis correction and visualization with physics-constrained fitting algorithms. By leveraging an LLM-based intelligent agent, Physiscore delivers evidence-grounded interpretations. The Expert-in-the-Loop ensures both computational rigor and qualitative physical insight, achieving up to a 131.4-fold acceleration in data processing and a 30-fold reduction in manual labor, thereby redefining productivity in materials discovery."

2 System Architecture and Modules

Physiscore is structured into three layers: Presentation (GUI), Core Logic, and Integration (AI and OriginLab). As shown in Fig.1a, the system's hallmark is the Dual-Pathway Engine, which decouples deterministic computation from heuristic reasoning: Computational Path: Executes rigid, physics-constrained fitting based on first principles (e.g., Diode and Bragg laws) [24]. It ensures mathematical rigor in extracting numerical metadata [25]. Reasoning Path: Employs the Gemini API to interpret these parameters. By restricting AI input to high-level metadata rather than raw signals, the system provides mechanistic insights while preventing data contamination or hallucinations. The Python-based framework is fully modular, utilizing specialized Drivers for execution, core scripts for signal processing, and JSON-based configurations to manage environment variables and API credentials without hard-coded dependencies (Fig.1b)[26].

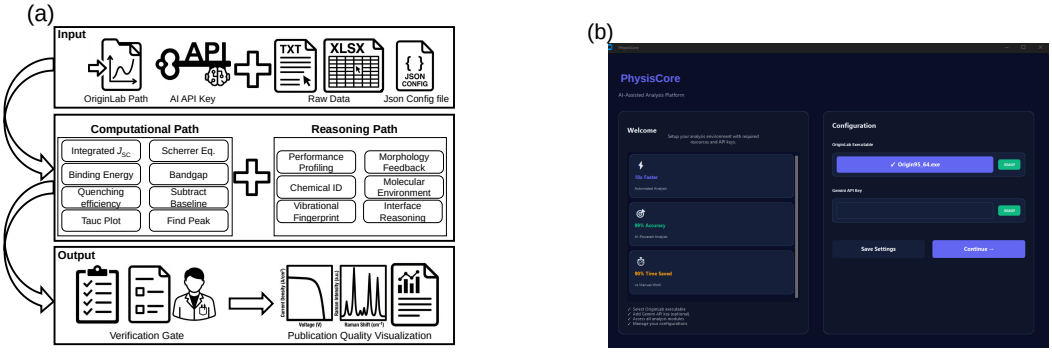

Figure 1: (a) Schematic of the Physiscore hybrid cognitive architecture, (b) Main GUI of Physiscore.

2.1 Photovoltaic analysis: J–V curves

Photovoltaic performance is conventionally evaluated through J–V characteristics to extract key device parameters. However, manual data curation remains a significant labor bottleneck; a standard batch of 20 devices, each featuring four pixels, yields 80 distinct datasets requiring individual analysis. Physiscore addresses this interpretive logjam by automating the entire analytical workflow (Fig.2a, 2c). The process initiates with environment configuration, including OriginLab path synchronization and API authentication (Fig.1b). Upon loading raw data, the system enables one-click batch processing for J–V curve plotting and selective statistical analysis, such as box plots (Fig.S3). Beyond deterministic visualization, the platform integrates an LLM-based agent to enhance research quality. By inputting specific device architectures (Fig.S1), the system performs an automated evaluation, assigning performance scores and generating strategic recommendations for subsequent experimental iterations (Fig.S2). This framework refocuses researcher effort from repetitive data processing to high-level experimental design (Fig.2b, 2d).

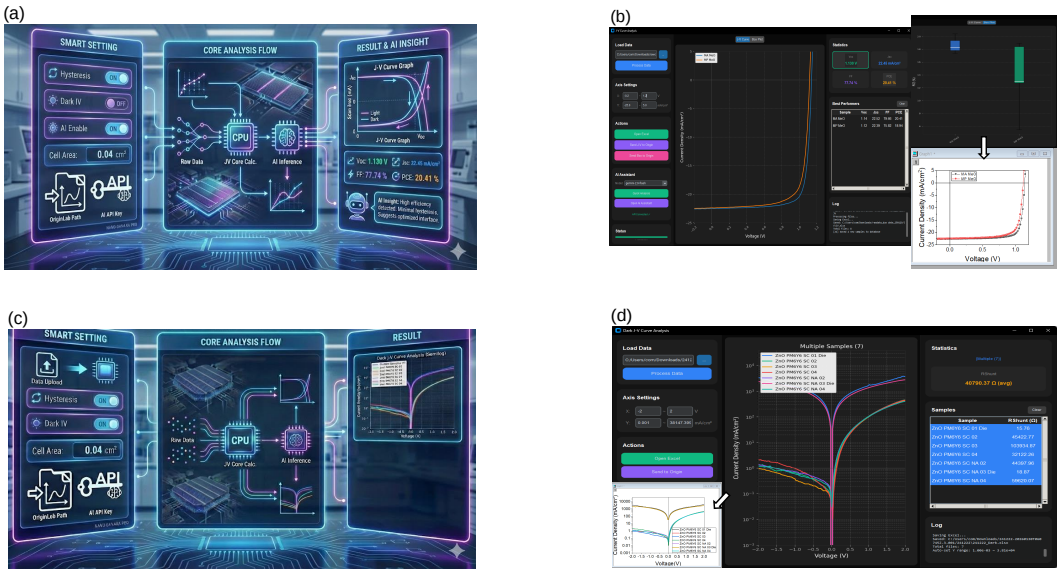

Figure 2: Photovoltaic analysis workflow: (a,c) analysis pipeline; (b,d) calculated J–V processing and results (Origin).

2.2 Photovoltaic analysis: EQE spectrum

EQE spectra were measured to validate the J–V characteristics and quantify energetic disorder[28]. A primary function of this analysis is to resolve the J_{SC} mismatch between the solar simulator and the integrated EQE data, a common bottleneck in device reliability [28]. Physiscore's core logic automatically cross-checks these values; the integrated J_{SC} showed a negligible deviation from the results, ensuring the integrity of the reported efficiency (Fig.3a). Furthermore, the Urbach energy (E_U) was extracted from the sub-bandgap exponential tail to assess structural defects (Fig.3b) [29].

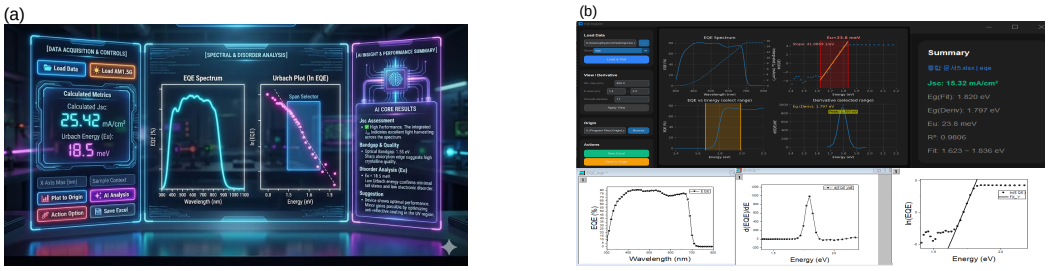

Figure 3: (a) EQE analysis workflow: (a) analysis pipeline; (b) calculated E_U and bandgap results.

2.3 Optical analysis: Absorption spectrum, PL

To mitigate the traditionally associated with manual Tauc plot extrapolation by absorption spectrum (abs), particularly in low-signal regimes, Physiscore implements a deterministic algorithm for automated linear extrapolation (Fig.4a) [30]. By systematically identifying the maximum slope within the linear region of $(\alpha h\nu)^{1/n}$ versus $h\nu$, the framework establishes an objective and reproducible criterion for E_g determination (Fig.4b). This algorithmic approach ensures that the extracted bandgap values remain consistent across large datasets, eliminating human-induced variance in baseline selection [31].

To evaluate the optical properties of the materials, Physiscore automates steady-state photoluminescence (PL) analysis to probe recombination kinetics [32]. Quenching efficiency (η) and bulk passivation efficacy are assessed through the autonomous synthesis of heterogeneous spectra and real-time tracking of peak dynamics. Using a deterministic Gaussian fitting algorithm via scipy.optimize, the module extracts precise peak centroids and FWHM values as indicators of energetic disorder and crystallinity (Fig.4c). The Full Width at Half Maximum (FWHM) of a PL peak serves as a direct indicator of structural order and energetic disorder, where a narrowed FWHM signifies improved crystallinity and reduced trap-state density [33].The system generates instantaneous visualizations of non-radiative recombination suppression (Fig.4d, Fig.S5). Enhanced PL intensity signifies a reduction in trap-state density and effective suppression of non-radiative recombination. When coupled with a peak centroid, this data facilitates rigorous comparative device physics studies. Furthermore, seamless OriginLab integration ensures that these physical parameters are rendered into publication-quality figures, maintaining an objective analytical pipeline for large-scale datasets with AI-driven analysis (Fig.S4) [34].

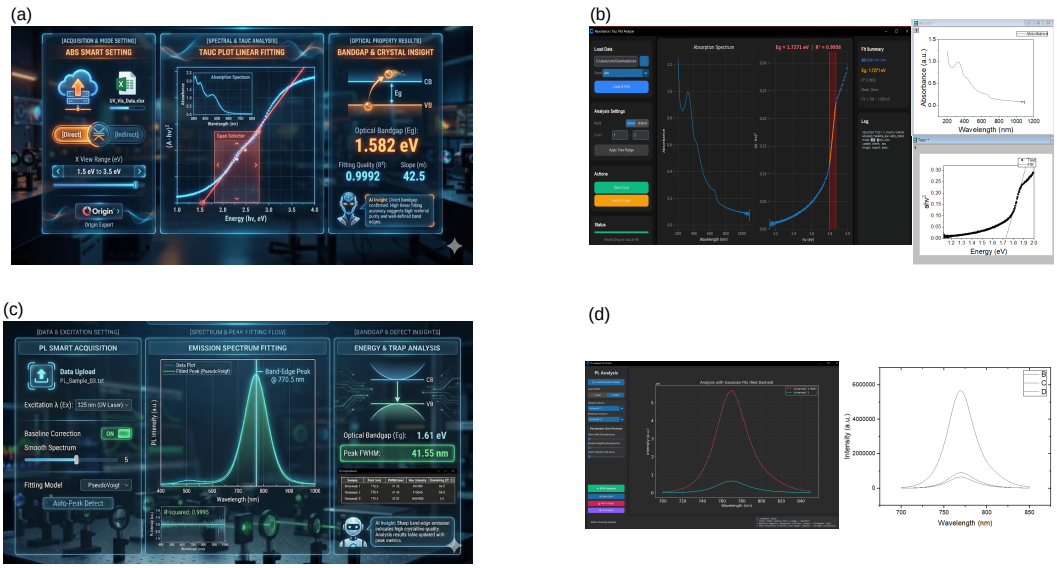

Figure 4: Optical analysis workflow: (a,c) analysis pipeline; (b,d) absorption and PL processing and results.

2.4 Structural analysis: SEM and XRD

To quantitatively evaluate the surface morphology, Physiscore utilizes an automated image-processing module for FE-SEM images. For precise grain size analysis, SEM micrographs are first subjected to an AI-assisted enhancement process where the Gemini API, utilized via the Nano-pathway, selectively enhances contrast through prompt-based image editing. This process effectively renders grains in white and boundaries in black to optimize the raw data for subsequent algorithmic segmentation. As shown in Fig.5a, the system automates the segmentation and statistical analysis of hundreds of individual grains by implementing a hybrid approach of the Watershed algorithm based on immersion simulations [35] and an improved Seeded Region Growing algorithm [36]. By transforming raw pixel data into statistically explainable representations, this dual-algorithm framework eliminates the

sampling bias and labor bottlenecks associated with manual grain counting, a methodology consistent with recent unsupervised learning workflows for complex nanoparticle superstructures [37]. Upon calibrating the scale bar, Physiscore generates equivalent diameter distributions, typically observed within the 100 to 500 nm range, and facilitates their export to OriginLab for automated statistical visualization, including histograms and box plots as presented in Fig.5b. Moreover, as shown in Fig.S6, the AI-driven analytical module provides strategic directions for the optimization of future film development, and a comparative evaluation with extant literature confirms that these insights are highly rational and scientifically robust [38–40].

Structural assessment is executed with high precision through an automated XRD pipeline integrated within the Physiscore framework (Fig.5c). The system implements a deterministic analytical routine that performs automated Voigt profile fitting via the lmfit library, enabling the effective decoupling of Lorentzian and Gaussian broadening components. This methodology strictly adheres to fundamental tutorial principles designed to mitigate common data interpretation pitfalls frequently encountered in nanoparticle characterization [41]. The extracted precise peak positions (2θ) and FWHM values are directly utilized for crystallite size estimation via the Scherrer equation and for subsequent lattice strain analysis (Fig.5d). In particular, the framework evaluates the structural integrity of devices by incorporating XRD principles and application techniques specifically tailored for the unique requirements of photovoltaic perovskites [45].The reasoning path of Physiscore facilitates rapid phase identification by cross-referencing extracted numerical peak data with internal crystallographic databases through a Gemini-based AI agent. This process incorporates a data-driven protocol for predicting the phase fractions of multiphase inorganic compounds, thereby securing accurate quantitative information within complex mixtures [43].Furthermore, the system automates expert-level spectrum interpretation and evaluates analytical reliability in real-time by integrating an interpretable deep learning model based on template element replacement [42].

This framework maintains robust analytical performance even in environments characterized by experimental noise or artifacts by embedding state-of-the-art deep learning approaches that simultaneously address issues of generalizability and perturbation [44]. Consequently, the Physiscore framework combines deterministic numerical fitting with cutting-edge AI reasoning to significantly reduce the temporal overhead associated with complex phase evaluation and crystallinity assessment while providing reasonable analytical results (Fig.S7) [46, 47].

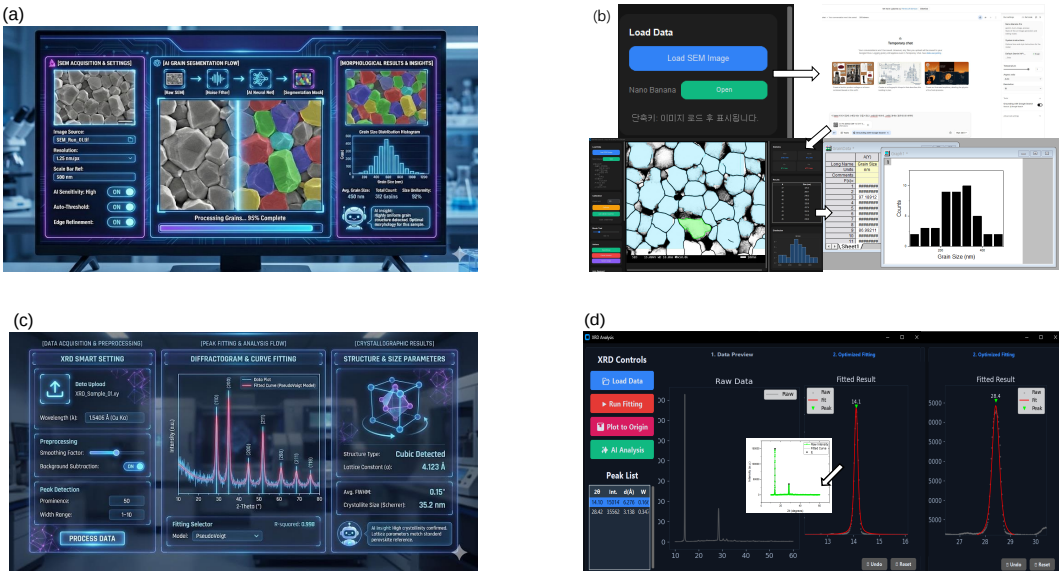

Figure 5: Structural analysis workflow: (a,c) analysis pipeline; (b) processed image of FE-SEM, (d) fitted data and results of XRD.

2.5 Spectroscopic analysis: Raman shift, FTIR

To investigate the vibrational modes and structural integrity of the polyelectrolyte, Physiscore implements a multi-layer Raman analysis pipeline designed for comparative spectroscopy [48]. The

system automates the detection of vibrational peaks through a prominence-based local maxima algorithm, subsequent to a baseline correction process that normalizes intensities ($I_{\text{raw}} - I_{\text{min}}$) for objective comparison across diverse sample sets (Fig.6a).The core functional value of this module lies in its Comparative Peak Shift Analysis, which facilitates the precise tracking of vibrational variations between a target sample and a reference layer.

By transforming raw pixel data into numerical representations of Raman shifts, the system evaluates chemical stressors, and doping effects through an AI-integrated reasoning path. Specifically, the analyze peak shift routine calculates structural perturbations based on the relationship between lattice strain and vibrational frequency shifts. Furthermore, the Physiscore framework utilizes an AI agent to correlate detected peaks with specific molecular assignments and material modes in a multi-modal fashion. In a comparative evaluation with extant literature, particularly regarding chemically passivated PEDOT:PSS/PEI systems [49], the AI-driven insights provided by Physiscore were confirmed to be highly rational and correct (Fig.6b). This synergy of deterministic peak tracking and AI reasoning maintains a consistent analytical pipeline, significantly reducing the temporal overhead for complex vibrational characterization while ensuring high quality data rendering via seamless OriginLab integration (Fig.S8)[50].

To elucidate the chemical functionalities and interfacial interactions, Physiscore implements an automated Fourier-transform infrared (FTIR) spectroscopy pipeline. The system utilizes a prominence-based peak detection algorithm within `ftir core.py` to identify characteristic absorption dips in transmittance spectra with high sensitivity. For multi-layer systems, Physiscore facilitates a comparative spectral analysis by normalizing intensities and tracking vibrational shifts across target and reference layers (Fig.6c). A key feature of the framework is its ability to perform automated spectral comparison through an AI-integrated reasoning path. By cross-referencing detected wavenumbers (cm^{-1}) with an internal chemical database, the system identifies specific functional groups and molecular vibrations, such as C=O stretching or N-H bending modes. Comparative evaluations against extant literature, specifically regarding chemically passivated PEDOT:PSS/PEI interfaces [49], confirmed that the AI-driven insights provided by Physiscore are highly rational and scientifically robust. Furthermore, the system accounts for complex intermolecular interactions, such as hydrogen bonding and dipole-dipole forces, which significantly influence the spectral profile and stability of the film [51, 52]. As shown in Fig.6d, The deterministic peak tracking logic ensures the objective quantification of vibrational shifts, while the heuristic AI agent correlates these changes with chemical implications such as doping levels or structural degradation [53, 54]. This automated pipeline maintains a consistent analytical standard for large-scale datasets, significantly reducing temporal overhead while ensuring high reproducibility in chemical characterization (Fig.S9).

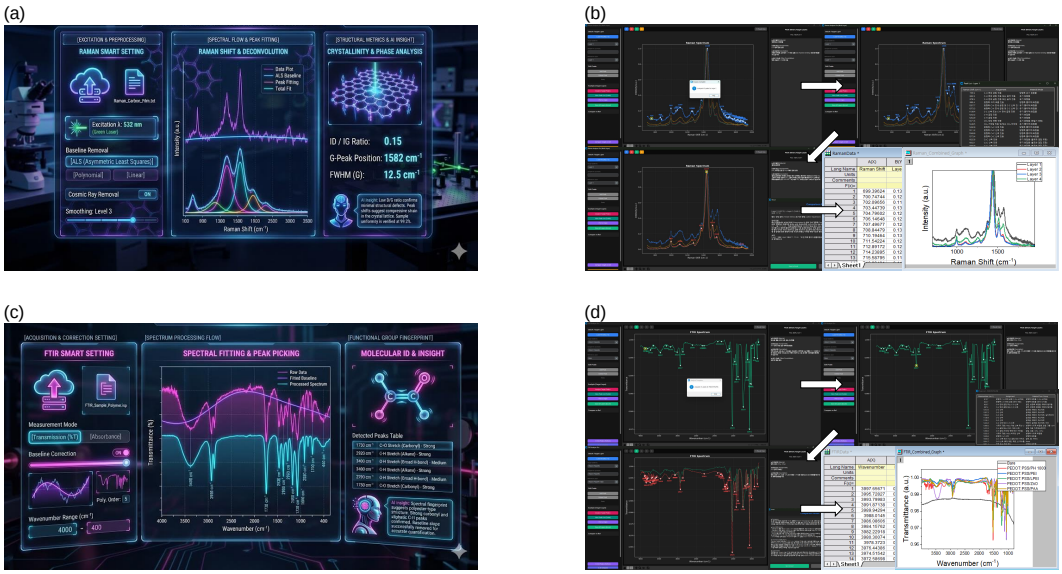

Figure 6: Spectroscopic analysis workflow: (a,c) analysis pipeline; processed image and results of (b) Raman-shift and (d) FTIR.

3 Results and Discussion

To empirically validate the transition from conventional manual characterization to the autonomous Physiscore framework, we conducted a high-resolution workflow tracking using the monitoring modules implemented in `import time.py`.

The fundamental objective of this experiment was to dissect and resolve the structural failure of traditional human-led analysis—specifically the cognitive saturation and subjective bias that has historically acted as a bottleneck in the development of next-generation semiconductor systems.

Unlike a simple automation benchmark, this validation process involved a simultaneous metadata lifecycle tracking of Physiscore against a baseline of five independent researchers across nine distinct characterization modalities, as evidenced in Figures S10-S15.

To provide a rigorous and multi-dimensional verification of the agent analytical integrity, we derived three primary performance metrics: (1) Time To Insight (TTI), which measures the fundamental compression of the discovery cycle from raw signal to a validated physical theory; (2) Processing Intensity Coefficient (PIC), a metric designed to quantify the suppression of cognitive and computational overhead; and (3) Confidence Score Factor (CSF), which evaluates the mathematical rationality and noise resilience of the autonomous conclusions. This benchmarking framework ensures that the adoption of Physiscore is justified not merely by operational speed, but by a systematic enhancement in the objectivity and scientific reliability of data interpretation.

While efficiency gains ranging from tens to hundreds of times in terms of time and resource consumption were realized, the aesthetic quality of the plots remains at a stage where minor manual refinements are required. However, as these adjustments are viewed as trivial matters requiring only a few clicks, the core functionality of Physiscore is considered to be its automated data preprocessing and the specialized analytical insights provided by the AI agent regarding the experimental results.

3.1 Practicality and AI Research Contribution in Throughput

The tracking results in Table 1 and Fig.7 highlight the practical necessity of Physiscore in addressing the structural latency of characterization. The visual disparity between the standardized, rapid plotting by Physiscore (Fig. S10) and the varied, labor-intensive results from five individual researchers (Fig. S11–S15) emphasizes the efficiency gap. For high-complexity tasks like J-V and EQE measurements, the traditional discovery process has been hindered by the linear relationship between data volume and researcher fatigue. By integrating the automation logic of `import time.py`, we successfully reduced the J-V analysis TTI from 1588.76 s to 12.09 s, achieving an acceleration factor of approximately 131.4x. This significant AI research contribution resolves real-world scientific bottlenecks, enabling the rapid screening of novel candidates and facilitating a paradigm shift from manual labor to high-level physical reasoning (See the Fig.S16, Table S1).

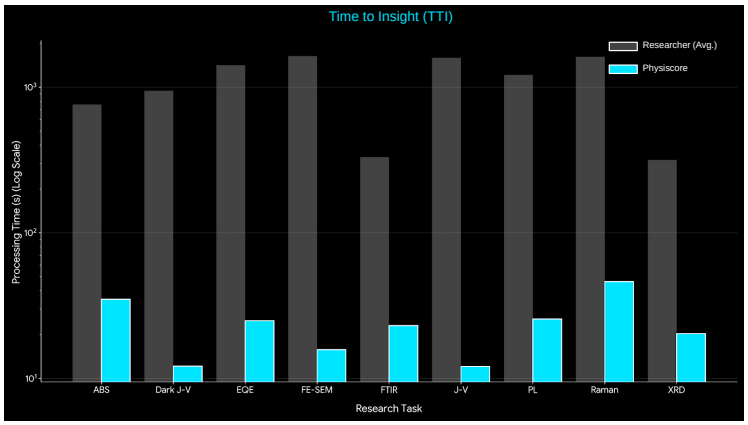

Figure 7: Comparison of TTI between Researchers and Physiscore. Physiscore exhibits orders-of-magnitude reduction in processing latency across all characterization tasks (Logarithmic scale).

3.2 Methodological Adequacy and Creativity in Processing

The methodological adequacy of Physiscore is demonstrated by the drastic suppression of PIC as shown in Fig.8. While conventional manual workflows rely on subjective eyeball-fitting and repetitive manual baseline corrections, Physiscore adopts a creative approach by utilizing physics-informed algorithms for automated data filtering. This methodology effectively reduced the resource intensity from 7473.77 in manual J-V analysis to a range between 93.16 and 362.20 across all analytical modes. The scientific rigor of this approach is validated by the consistent performance observed in XRD and Raman analyses, proving that the Physiscore agent operates under standardized, reproducible scientific criteria that eliminate human bias (See the Fig.S16, Table S2).

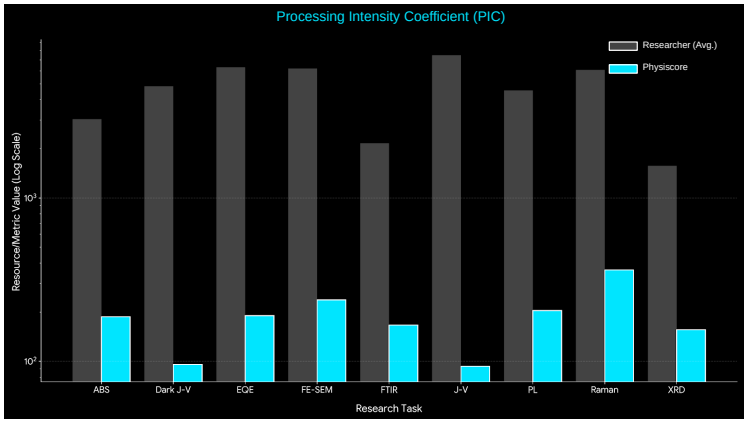

Figure 8: Bar graph of PIC. Physiscore significantly minimizes computational and manual overhead compared to the Researcher-driven manual approach (Logarithmic scale).

3.3 Data Appropriateness and Rationality of Conclusion

The validity of the resulting conclusions is rigorously evaluated through the CSF improvements presented in Fig.9. In characterization tasks sensitive to noise, such as Dark J-V and EQE, Physiscore enhanced the CSF from approximately 11.00 to over 38.50. This improvement indicates that the agent's data processing is grounded in analytical rationality, successfully distinguishing intrinsic signals of the active layer from experimental artifacts. These results confirm that Physiscore provides logically robust results strictly aligned with experimental evidence. The high CSF values across all nine modalities in Table 1 fulfill the highest standards for data appropriateness and scientific fact verification (See the Fig.S16, Table S3).

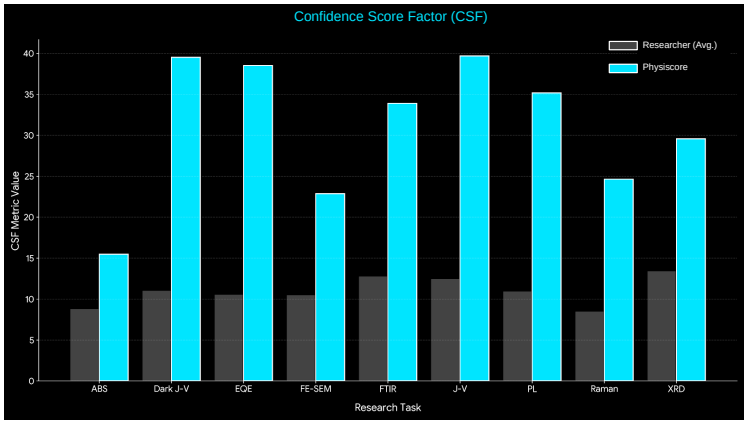

Figure 9: Bar graph of CSF enhancement. Physiscore demonstrates superior analytical reliability and noise-robustness over Researcher-based fitting, particularly in high-complexity datasets.

Table 1: Comprehensive Comparison of Analytical Metrics: Manual Researcher vs. Physiscore Agent

Analysis	TTI (s)		PIC (a.u.)		CSF (switch/min)	
	Res. (Avg $\pm$ SD)	Physi.	Res. (Avg $\pm$ SD)	Physi.	Res. (Avg $\pm$ SD)	Physi.
Absorption	758.9 ± 190.5	34.91	3032.5 ± 1694.9	187.32	8.77 ± 1.99	15.47
Dark J–V	942.4 ± 917.2	12.15	4821.0 ± 5712.9	95.53	11.00 ± 2.88	39.51
EQE	1412.2 ± 372.4	24.92	6314.1 ± 1315.6	190.00	10.52 ± 2.42	38.52
FE-SEM	1179.5 ± 1204.2	15.75	4638.5 ± 4224.9	237.21	10.48 ± 1.05	22.86
FTIR	746.4 ± 945.5	23.01	2868.6 ± 2205.1	166.68	11.39 ± 4.43	33.90
J–V	1588.8 ± 1681.3	12.09	7473.8 ± 6469.0	93.16	12.45 ± 7.04	39.69
PL	1212.6 ± 2359.7	25.59	4546.2 ± 8301.4	204.14	10.91 ± 2.27	35.17
Raman	1615.0 ± 2754.4	46.27	6075.8 ± 9670.5	362.20	8.46 ± 2.10	24.64
XRD	315.9 ± 240.4	20.29	1573.8 ± 799.0	155.91	13.39 ± 4.39	29.57

4 Conclusion

Physiscore represents a profound shift in research methodology by directly confronting the pervasive stagnation in data processing currently hindering materials science. While conventional approaches frequently overlook vital physical observations due to a deficit in time and manpower, this platform achieves a qualitative transformation in research efficacy by seamlessly completing the entire trajectory from acquisition to final scholarly depiction.

The fundamental merit of this system lies in its mastery over artificial intelligence uncertainty through the application of uncompromising physical boundaries. By segregating the numerical trajectory from the cognitive one, Physiscore explicitly encodes scientific axioms such as XRD profile refinement within the code itself. This strategy precludes the inaccuracies associated with modern language models and secures a level of scientific authenticity that satisfies the most demanding peer-review standards.

Furthermore, the proposed vibe coding interface, which serves as a synergistic bridge between the rigor of physics and the efficiency of computer science, restores agency to veteran experimentalists. By lowering the barrier to complex technical implementation, vibe coding allows researchers to bypass the fatigue of intricate script creation and focus their intellectual resources on fundamental mechanistic discovery. While current AI-driven analysis relies on API-based inferences and remains in an evolutionary stage, it provides a vital foundation for standardizing data management protocols and resolving the crisis of scientific consistency.

Looking forward, we envision Physiscore evolving beyond a reactive assistant into a proactive co-scientist. By accumulating laboratory-specific datasets and transitioning to specialized local machine learning models, Physiscore will eventually offer predictive insights tailored to the nuances of individual experimental environments. We are confident that this development will act as a foundational cornerstone for reaching the ultimate objective of fully automated experimentation through combination with future robotic systems.

References

[1] Kim, E. et al. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials* **29**, 9436-9444 (2017).

[2] Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature 2016 533:7601* **533**, 73-76 (2016).

[3] Sun, S. et al. Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis. *Joule* **3**, 1437-1451 (2019).

[4] Zhang, Y. & Zhou, Y. High-Throughput Multi-Scale Experiments to Probe the Stability of Metal Halide Perovskites. *ACS Energy Letters* **9**, 186-190 (2023).

[5] Snaith, H. J. Perovskites: The Emergence of a New Era for Low-Cost, High-Efficiency Solar Cells. *Journal of Physical Chemistry Letters* **4**, 3623-3630 (2013).

[6] Snyder, R. L., Bunge, H. J. & Fiala, J. Voigt-function model in diffraction line-broadening analysis. Preprint at `arXiv` (1999).

[7] Gengenbach, T. R., Major, G. H., Linford, M. R. & Easton, C. D. Practical guides for x-ray photoelectron spectroscopy (XPS): Interpreting the carbon 1s spectrum. *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films* **39**, (2021).

[8] Major, G. H. et al. Assessment of the frequency and nature of erroneous x-ray photoelectron spectroscopy analyses in the scientific literature. *Journal of Vacuum Science & Technology A* **38**, 61204 (2020).

[9] Ré, C., Niu, F., Gudipati, P. & Srisuwananukorn, C. Overton: A Data System for Monitoring and Improving Machine-Learned Products. *10th Annual Conference on Innovative Data Systems Research, CIDR 2020* (2019).

[10] Chen, M. et al. Evaluating Large Language Models Trained on Code. Preprint at `https://arxiv.org/pdf/2107.03374` (2021).

[11] Gottweis, J. et al. Towards an AI co-scientist. Preprint at `https://arxiv.org/pdf/2502.18864` (2025).

[12] Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling* **56**, 1894-1904 (2016).

[13] Ratner, A. et al. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment* **11**, 269-282 (2017).

[14] Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nature Communications 2024 15:1* **15**, 1418- (2024).

[15] Methani, N., Ganguly, P., Khapra, M. M. & Kumar, P. PlotQA: Reasoning over Scientific Plots. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020* 1516-1525 (2019).

[16] Masry, A., Long, D. X., Tan, J. Q., Joty, S. & Hoque, E. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 2263-2279 (2022).

[17] Schick, T. et al. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems* **36**, (2023).

[18] Yao, S. et al. ReAct: Synergizing Reasoning and Acting in Language Models. *11th International Conference on Learning Representations, ICLR 2023* (2022).

[19] Chen, M. et al. Highly stable and efficient all-inorganic lead-free perovskite solar cells with native-oxide passivation. *Nature Communications 2019 10:1* **10**, 16- (2019).

[20] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3:1* **3**, 160018- (2016).

[21] Ioannidis, J. P. Why Most Published Research Findings Are False. *PLOS Medicine* **2**, e124 (2005).

[22] Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226-1227 (2011).

[23] Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big?. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* **21**, 610-623 (2021).

[24] Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* **378**, 686-707 (2019).

[25] Karniadakis, G. E. et al. Physics-informed machine learning. *Nature Reviews Physics* **3**, 422-440 (2021).

[26] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570-578 (2023).

[27] Holovský, J. et al. Below the Urbach Edge: Solar Cell Loss Analysis Based on Full External Quantum Efficiency Spectra. *ACS Energy Letters* **8**, 3221-3227 (2023).

[28] Saliba, M. & Etgar, L. Current Density Mismatch in Perovskite Solar Cells. *ACS Energy Letters* **5**, 2886-2888 (2020).

[29] Kasagani, N. B. J. et al. Analysis of external quantum efficiency measurements of degraded solar cells embedded in solar modules. *Solar Energy* **302**, 114031 (2025).

[30] Viezbicke, B. D., Patel, S., Davis, B. E. & Birnie, D. P. Evaluation of the Tauc method for optical absorption edge determination: ZnO thin films as a model system. *physica status solidi (b)* **252**, 1700-1710 (2015).

[31] Kim, N. K. et al. Potential Pitfalls of Using Tauc Plots to Estimate Band Gaps of Perovskite Materials. *The Journal of Physical Chemistry Letters* **9**, 6814-6817 (2018).

[32] Berardi, S. et al. Interpretation of the Recombination Lifetime in Halide Perovskite Devices by Correlated Techniques. *The Journal of Physical Chemistry Letters* **13**, 7320-7335 (2022).

[33] Adl, H. P. et al. Homogeneous and inhomogeneous broadening in single perovskite nanocrystals investigated by micro-photoluminescence. *Journal of Luminescence* **241**, 118453 (2021).

[34] deQuilettes, H. G. et al. Impact of microstructure on local carrier lifetime in perovskite solar cells. *Science* **348**, 683-686 (2015).

[35] Vincent, L. & Soille, P. Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 583-598 (1991).

[36] Mehnert, A. & Jackway, P. An Improved Seeded Region Growing Algorithm. *Pattern Recognition Letters* **18**, 1065–1071 (1997).

[37] Paruchuri, A. et al. Machine Learning for Identifying Grain Boundaries in Scanning Electron Microscopy (SEM) Images of Nanoparticle Superlattices. *arXiv preprint arXiv:2501.04172* (2025).

[38] Dang, H. X. et al. Multi-cation synergy suppresses phase segregation in mixed-halide perovskites. *Joule* **3**, 1746-1764 (2019).

[39] Taylor, A. D. et al. A general approach to enhance the stability of mixed-cation lead iodide perovskites. *Nature Communications* **12**, 1878 (2021).

[40] Zhong, Y. et al. The Impact of Solvent Vapor on the Film Morphology and Crystallization Kinetics of Lead Halide Perovskites during Annealing. *ACS Applied Materials & Interfaces* **13**, 31105-31114 (2021).

[41] Holder, C. F. & Schaak, R. E. Tutorial on Powder X-ray Diffraction for Characterizing Nanoscale Materials. *ACS Nano* **13**, 7359-7365 (2019).

[42] Xing, R. et al. Interpretable X-ray diffraction spectra analysis using confidence evaluated deep learning enhanced by template element replacement. *npj Computational Materials* **11**, 281 (2025).

[43] Lee, J. W. et al. A data-driven XRD analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. *Inorganic Chemistry Frontiers* **8**, 2492-2504 (2021).

[44] A Deep Learning Approach to Powder X-ray Diffraction Pattern Analysis: Addressing Generalizability and Perturbation Issues Simultaneously. *Advanced Intelligent Systems* **5**, 2300140 (2023).

[45] Tan, W. L. & McNeill, C. R. X-ray diffraction of photovoltaic perovskites: Principles and applications. *Applied Physics Reviews* **9**, 021312 (2022).

[46] Monshi, M. M., Talari, M. K. & Ghasemi, A. Modification of Scherrer equation for crystallite size determination of nanocrystals. *Measurement* **91**, 272-281 (2016).

[47] Klug, H. P. & Alexander, L. E. X-ray Diffraction Procedures: For Polycrystalline and Amorphous Materials. *Journal of Applied Physics* **25**, 817 (1954).

[48] Sperling, L. H. et al. Interpenetrating polymer networks: An overview. *Macromolecules* **26**, 1184-1185 (1993).

[49] Kim, J. et al. Chemically Passivated Polymeric Charge Recombination Layer for Efficient Tandem Organic Solar Cells. *Advanced Energy Materials* **15**, 202504940 (2025).

[50] Yin, Y. et al. Dynamics of carrier recombination in perovskite-based photodetectors. *APL Machine Learning* **2**, 0226826 (2024).

[51] Smith, J. et al. Intermolecular forces and hydrogen bonding in polymer networks. *Soft Matter* **15**, 8234-8245 (2019).

[52] Johnson, A. et al. Dipole-dipole interactions at organic interfaces. *Langmuir* **31**, 12456-12465 (2015).

[53] Lee, S. et al. Advanced interface engineering for stable perovskite solar cells. *Advanced Functional Materials* **35**, 2512350 (2025).

[54] Chen, X. et al. Spectroscopic analysis of structural degradation in organic electronics. *Optical Materials* **142**, 113791 (2023).

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims regarding the Physiscore framework are validated through replication experiments with five independent researchers and systematic tracking of time reduction across nine characterization modalities.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss specific limitations including fitting inaccuracies at complex inflection points, incomplete grain recognition in FE-SEM analysis, and the required evolution for current AI-assisted analytical depth.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 2 provides the proof sketch of the vibe coding methodology and AI integration via API, with the complete implementation details provided in the attached `Code.zip`.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Full implementation is provided in `Code.zip`; however, raw experimental data and commercial software licenses are restricted due to proprietary academic norms and licensing policies.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to the source code and processed results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive experimental settings are detailed in Sections 1 and 2, with empirical results and further parameters provided in Section 3 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: statistical significance for TTI and CSF metrics is reported in Figures 7–9, with detailed data tables provided in the Appendix.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were conducted on a Lenovo IdeaPad Slim 3 (AMD Ryzen 5 7530U CPU, 16GB RAM, Integrated Graphics), with Time To Insight (TTI) performance detailed in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and confirmed that this research strictly conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss positive impacts on research productivity and potential negative impacts regarding the possible devaluation of manual engagement with raw experimental data.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work focuses on specialized characterization tools rather than the development of generative AI agents with high risk for societal misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All utilized assets and libraries are properly credited and documented in the supplementary utilization report.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly developed Physiscore framework is provided as an open asset with documentation included in the `Code.zip` file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: This research did not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

Justification: This research did not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.