
Benchmarking Multimodal Large Language Models on Electronic Structure Analysis and Interpretation

Izumi Takahara
Institute of Industrial Science
The University of Tokyo
kougen@iis.u-tokyo.ac.jp

Teruyasu Mizoguchi
Institute of Industrial Science
The University of Tokyo
teru@iis.u-tokyo.ac.jp

Abstract

Large language models (LLMs) are increasingly adopted in materials science, enabling automated literature mining, domain-specific scientific reasoning, and autonomous materials design. However, most existing systems are still limited to single-modality inputs, preventing the use of rich multimodal information inherent in the field. The electronic structure of materials is essential for predicting material properties, understanding their origins, and guiding new materials design, yet its integration into multimodal LLM (MLLM) frameworks has been rarely explored. Here, we present the first systematic benchmark of pre-trained MLLMs for density of states (DOS) interpretation. Using a high-fidelity dataset from first-principles calculations, we evaluate MLLMs on visual question answering and captioning tasks related to the interpretation of electronic structures, with captions scored by both human experts and MLLM-based evaluators. Our results reveal the capabilities and limitations of MLLMs in electronic structure analysis and provide a foundation for developing next-generation multimodal AI systems for materials design.

1 Introduction

Large language models (LLMs) are playing an increasingly pivotal role in materials science, offering new opportunities to accelerate materials discovery and deepen scientific understanding[1–3]. Recent studies have explored their use in automated literature mining from large corpora[4–7], domain-specific chatbots capable of responding to scientific queries[8, 9], and autonomous agents for materials discovery and design[10–13]. By leveraging their reasoning capabilities, LLM-based systems have the potential to assist researchers in hypothesis generation, data interpretation, and decision-making throughout the materials development process[14].

However, most existing LLM applications in materials science are limited to single-modality inputs, typically relying only on textual data. This limitation prevents them from exploiting the rich multimodal information available in the field, such as charge densities, microscopy images, and spectroscopic data, which provide essential insights for materials discovery and design. In recent years, progress has been made toward multimodal approaches, such as MatterChat[15], which integrates crystal graph representation with text. Nevertheless, one of the most critical modalities for understanding and designing materials, the electronic structure, remains largely unexplored for integration to multimodal LLM (MLLM) systems.

The electronic structure determines many fundamental properties of materials, such as electrical conductivity, optical absorption, and catalytic activity. The density of states (DOS) is a core representation of this structure, providing detailed insights into the distribution of the system’s electronic energy levels. Accurate interpretation of DOS requires both visual pattern recognition and domain-specific reasoning—capabilities that, if endowed to AI systems, could substantially enhance automated mate-

rials analysis, literature interpretation, and data-driven materials design. While several benchmarks for MLLMs in materials science have been developed in recent years [16, 17], there is no systematic benchmark dedicated to electronic structure interpretation that conducts both comprehensive quantitative and qualitative evaluations. Such evaluations are essential, as DOS analysis involves not only the quantitative extraction of values such as the band gap but also the qualitative interpretation of peak shapes and their relation to material properties.

In this work, we evaluate the capabilities of pre-trained MLLMs to perform inferential and critical analysis of electronic structures from DOS diagrams provided as image inputs, examining the applicability of their visual reasoning abilities to materials research. A key challenge in developing such a benchmark lies in the difficulty of preparing high-quality datasets that pair domain-specific visual representations with reliable ground-truth annotations, as data collected from literature often vary in format and consistency, making it challenging to construct a reliable dataset[18]. To address this, our study constructs an *ab initio* dataset based on high-accuracy first-principles calculations to enable rigorous and reproducible evaluation of MLLM performance. To conduct comprehensive evaluations, we introduce two complementary tasks: visual question answering (VQA) to test quantitative inference from DOS data, and captioning to assess qualitative interpretive ability. The captioning task is evaluated by both human experts and MLLM-based evaluators, allowing us to assess whether, beyond quantitative prediction accuracy, MLLMs align with human judgment in the qualitative interpretation of electronic structures. Our benchmark not only provides a snapshot of current MLLMs in materials science but also lays the foundation for future AI systems that natively integrate electronic structure information.

2 Experimental Setup

2.1 Task definition

In this study, the performance of MLLMs in interpreting DOS diagrams was evaluated through two complementary tasks: quantitative VQA task and qualitative captioning task, as illustrated in Figure 1. The VQA task involved predicting physical properties from DOS diagrams, including the valence band maximum (VBM), conduction band minimum (CBM), band gap, position of the Fermi level, and DOS at the Fermi level, which are key quantities that directly characterize the electronic structure and are essential for understanding material properties. The captioning task required generating a textual interpretation of the DOS, focusing on its features, the related electronic properties, and their correlations with material characteristics; this task was assessed by both human experts and MLLM-based evaluators. Each evaluator rated the generated captions on a five-point scale across three criteria: **accuracy** (the correctness of the description), **depth** (the depth of scientific insight provided), and **fluency** (appropriate use of the technical terminology and clarity of expression). Details of the scoring criteria and prompt templates for each task are provided in Appendix A.

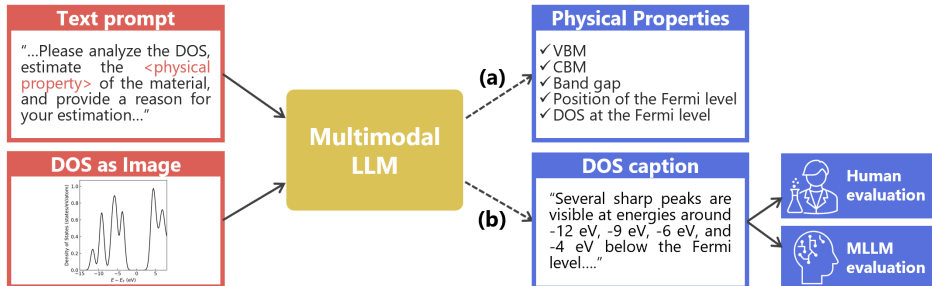


Figure 1: Overview of the evaluation framework for benchmarking MLLMs on DOS analysis and interpretation. (a) VQA task: given a prompt and DOS diagram, the model predicts key properties (VBM, CBM, band gap, position of the Fermi level, and DOS at the Fermi level). (b) Captioning task: the model generates a detailed textual interpretation of the DOS, evaluated by human experts and MLLMs.

2.2 Datasets and Models

All-electron density functional theory (DFT)[19, 20] calculations were performed with FHI-aims[21, 22] to construct an *ab initio* dataset linking DOS images with corresponding physical properties, spanning metallic, semiconductor, and insulating materials. For DOS image preparation, the Fermi level was aligned to 0 eV, and the x -axis covered -15 to 8 eV relative to the Fermi level. The y -axis range was determined from the maximum DOS value (DOS_{max} , in states/eV/atom) within this energy window and set to span from 0 to $1.1 \times DOS_{max}$. In the VQA and captioning tasks, 899 and 20 DOS samples were used, respectively. Detailed information on dataset construction is provided in Appendix B.

In this study, we evaluated the following MLLMs: o4-mini[23], GPT-4.1-mini[24], GPT-4o[25], and GPT-4o-mini[26], as well as publicly available open-weight models of moderate scale, including Llama-3.2-11B-Vision-Instruct[27], Pixtral-12B-2409[28], and Qwen2.5-VL-7B-Instruct[29]. The temperature parameter was set to 0.1.

3 Results

3.1 VQA-based DOS analysis

Table 1 shows the prediction results for key physical properties obtained from DOS plots, including the VBM, CBM, band gap, position of the Fermi level, and DOS at the Fermi level.

Table 1: Mean absolute error (MAE) for physical property predictions. Units for VBM, CBM, band gap, and E_F are eV; the unit for DOS at E_F is states/eV/atom. The smallest MAE in each column is shown in **bold**, and the second smallest is *underlined*.

Model	VBM	CBM	Band gap	E_F	DOS at E_F
o4-mini	0.842	0.742	1.131	<u>0.002</u>	0.049
GPT-4.1-mini	1.292	1.056	1.766	0.000	<u>0.085</u>
GPT-4o	<u>1.154</u>	<u>0.886</u>	<u>1.694</u>	0.000	0.217
GPT-4o-mini	5.549	1.674	3.195	1.498	0.941
Llama-3.2-11B-Vision-Instruct	2.385	2.734	1.778	0.953	0.840
Pixtral-12B-2409	6.991	1.480	2.401	1.056	0.540
Qwen2.5-VL-7B-Instruct	8.635	2.673	3.693	2.348	1.064

From Table 1, o4-mini shows the highest prediction accuracy across almost all physical properties, followed by GPT-4o and GPT-4.1-mini. In our dataset, the Fermi level is aligned to 0 eV, following the convention in electronic structure analysis. Consequently, correctly predicting its position requires recognizing the x -axis label " $E - E_F$ (eV)" accurately. The models o4-mini, GPT-4o, and GPT-4.1-mini answered this almost without error, and models that predicted the Fermi level position with high accuracy also tended to achieve relatively low MAE in predicting the DOS at E_F . In addition, across all models, VBM predictions tended to show larger MAE compared to those of the CBM.

Figure 2 shows scatter plots of predicted versus ground-truth values for (a) the VBM and (b) the CBM. From (a) and (b), we observe that the predictions by o4-mini follow the distribution of the ground-truth values, indicating that this model can relatively reliably extract VBM and CBM values from DOS diagrams. For VBM predictions, Pixtral 12B and Qwen2.5-VL 7B tended to output values on the lower-energy side of the valence band, whereas Llama-3.2 11B often produced values near the Fermi level. For CBM predictions, Pixtral 12B and Qwen2.5-VL7B frequently predicted values on the higher-energy side of the conduction band, while Llama-3.2 11B occasionally produced energies that actually fell within the valence band.

To examine how prediction accuracy varies among materials with different electronic structures, Figure 3 presents scatter plots showing the relationship between the ground-truth band gap and the prediction errors for each physical property. The region corresponding to materials with relatively small band gaps (≤ 2.5 eV), typically exhibiting semiconductor behavior, is shaded in gray, whereas the region with larger band gaps (> 2.5 eV) is highlighted in lavender.

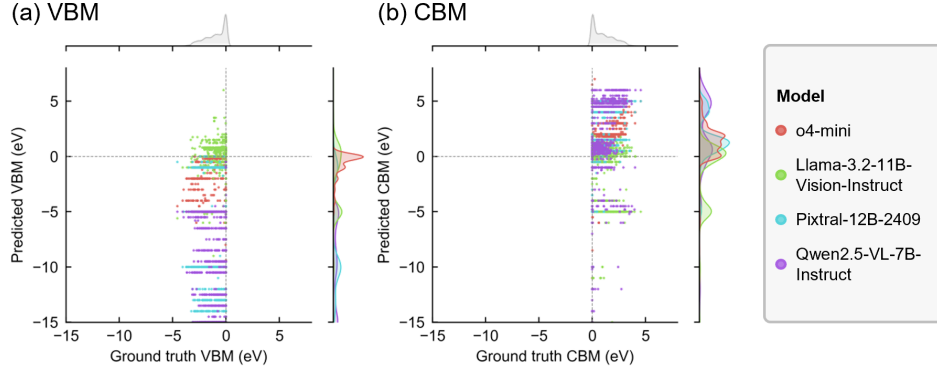


Figure 2: Scatter plots comparing ground-truth values and MLLM predictions for (a) the valence band maximum (VBM) energy and (b) the conduction band minimum (CBM) energy.

A nearly linear relationship is observed between the prediction errors and the ground-truth band gaps for (a) VBM, (b) CBM, and (c) band gap. This indicates that the LLMs do not fully capture the DOS features and tend to output several discrete preferred values regardless of the true band gap, producing multiple parallel trends. As a result, the models overestimate small band gaps and underestimate large ones, with predictions concentrated around 1.25–5 eV. This behavior may partly arise from pretraining bias. While o4-mini generates more diverse predictions that better reflect the underlying DOS features, a residual bias of this kind remains, and achieving more DOS-consistent predictions would likely require additional training or guidance through carefully designed prompts. For E_F predictions, Qwen2.5-VL 7B shows a similar tendency to output fixed values around -5 eV, and for DOS at E_F , large errors are observed for materials with vanishing band gaps, whereas these errors are substantially reduced in o4-mini.

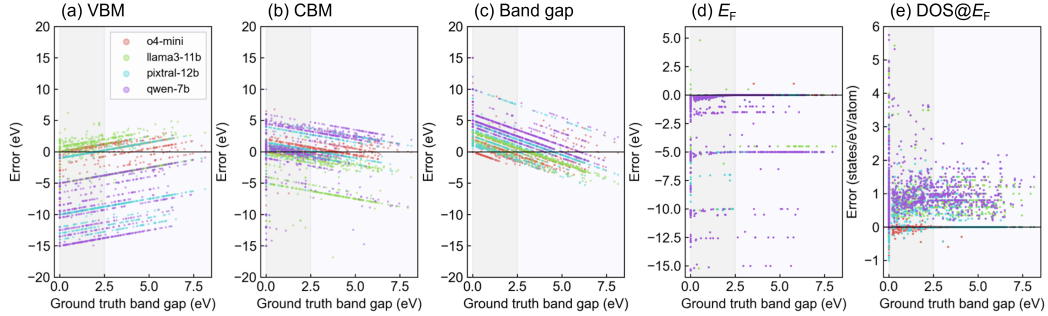


Figure 3: Scatter plots showing the relationship between the prediction errors for each physical property and the ground-truth band gap values. The region corresponding to band gaps between 0 and 2.5 eV are shaded in gray, while the region above 2.5 eV is highlighted in lavender.

3.2 Captioning-based DOS Interpretation

A captioning task was designed to assess DOS interpretation by MLLMs, focusing on interpretative aspects of scientific reasoning beyond quantitative prediction. The results are shown in Figure 4, which reports the mean and standard error of accuracy, depth, and fluency scores for captions generated by the MLLMs, as evaluated by human experts and MLLM-based evaluators. An example of generated captions is provided in Appendix C. The results indicate that MLLMs such as o4-mini, GPT-4.1-mini, and GPT-4o, which achieved high performance in the VQA task, also obtained high scores in accuracy, depth, and fluency in the captioning task under human evaluation. Notably, o4-mini obtained the highest scores in both accuracy and depth metrics, demonstrating reasoning that links DOS features to established principles in materials science (see Appendix C), suggesting the potential of MLLMs to provide deeper insights and support researchers beyond human limitations. Furthermore, when used as evaluators, these models showed scoring patterns broadly consistent with human assessments, highlighting their potential role as proxy evaluators for future model development and large-scale

benchmarking. A closer examination, however, reveals notable differences: GPT-4o tended to assign higher scores compared to both human experts and other MLLMs, GPT-4.1-mini produced evaluations more closely aligned with human judgments, and o4-mini also showed reasonable agreement but occasionally underestimated scores. For fluency, human evaluators consistently assigned lower scores, likely reflecting greater sensitivity to technical terminology and clarity of expression, suggesting the need for better alignment between human and MLLM evaluations.

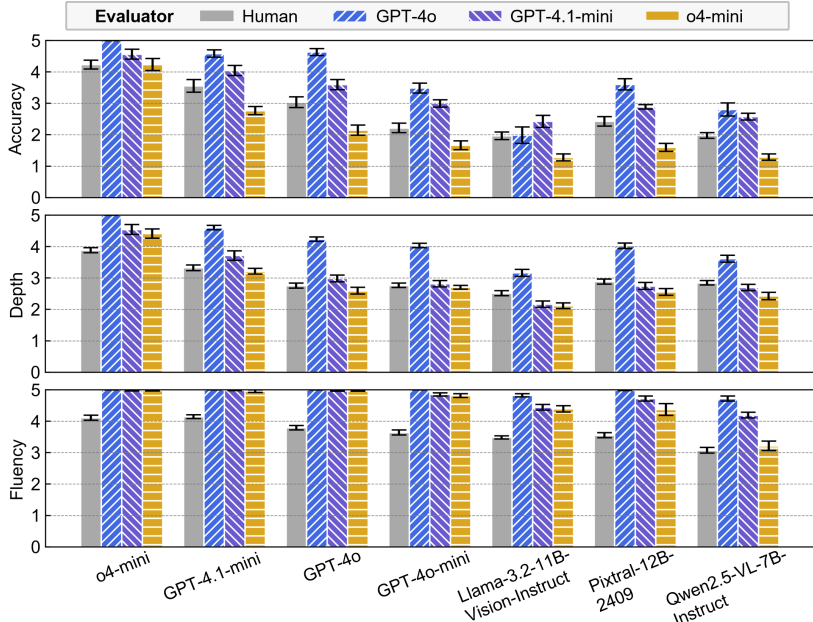


Figure 4: Scores of DOS captions generated by each LLM, evaluated by six human experts and LLM-based evaluators (GPT-4o, GPT-4.1-mini, and o4-mini). Accuracy, Depth, and Fluency scores were rated on a 1-5 scale. Bars show mean scores, with error bars denoting standard errors.

4 Conclusion

In this study, we conducted a systematic benchmark of pre-trained MLLMs for DOS interpretation, providing both quantitative and qualitative evaluations. In both VQA and captioning tasks, models such as GPT-4.1-mini and o4-mini demonstrated strong performance, indicating that current state-of-the-art MLLMs can capture certain structured aspects of electronic structure data. However, achieving more reliable predictions of quantitative physical properties such as band gaps would still require further domain adaptation. When used to evaluate DOS captions generated by MLLMs, their assessments generally aligned with human experts, though discrepancies remained in areas such as fluency, suggesting the need for better alignment between automated and human evaluations. This benchmark lays the groundwork for future research, including testing MLLM capabilities in more complex electronic structure analyses, such as orbital hybridization and spin polarization, and developing next-generation models that can reliably reason over these modalities.

Acknowledgement

This work was supported by JST ACT-X (Grant Number JPMJAX24DB) and JST BOOST (Grant Number JPMJBS2418), Japan. I. T. was supported by the MERIT-WINGS, The University of Tokyo.

References

- [1] Minh-Hao Van, Prateek Verma, Chen Zhao, and Xintao Wu. A survey of ai for materials science: Foundation models, llm agents, datasets, and tools, 2025. URL <https://arxiv.org/abs/2506.20743>.

- [2] Jinglan Zhang, Xinyi Chen, Xu Ye, Yulin Yang, and Bin Ai. Large language model in materials science: Roles, challenges, and strategic outlook. *Advanced Intelligent Discovery*, page 202500085, 2025.
- [3] Agada Joseph Oche and Arpan Biswas. Role of large language models and retrieval-augmented generation for accelerating crystalline material discovery: A systematic review, 2025. URL <https://arxiv.org/abs/2508.06691>.
- [4] Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 4(1):2356506, 2024.
- [5] Yifei Duan, Yixi Tian, Soumya Ghosh, Vineeth Venugopal, Jie Chen, and Elsa A. Olivetti. Llm-empowered literature mining for material substitution studies in sustainable concrete. *Resources, Conservation and Recycling*, 221:108379, 2025.
- [6] Youwan Na, Jeffrey J. Kim, Chanhyoung Park, Jaewon Hwang, Changgi Kim, Hokyung Lee, and Jehoon Lee. Advanced scientific information mining using llm-driven approaches in layered cathode materials for sodium-ion batteries. *Mater. Adv.*, 6:2543–2548, 2025.
- [7] Yanpeng Ye, Jie Ren, Shaozhou Wang, Yuwei Wan, Imran Razzak, Bram Hoex, Haofen Wang, Tong Xie, and Wenjie Zhang. Construction and application of materials knowledge graph in multidisciplinary materials science via large language model, 2025. URL <https://arxiv.org/abs/2404.03080>.
- [8] Xianjun Yang, Stephen D. Wilson, and Linda Petzold. Quokka: An open-source large language model chatbot for material science, 2024. URL <https://arxiv.org/abs/2401.01089>.
- [9] Yeonghun Kang and Jihan Kim. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature Communications*, 15(1):4705, 2024.
- [10] Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents, 2025. URL <https://arxiv.org/abs/2504.00741>.
- [11] Adib Bazgir, Rama chandra Praneeth Madugula, and Yuwen Zhang. Matagent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.
- [12] Alireza Ghafarollahi and Markus J. Buehler. Autonomous inorganic materials discovery via multi-agent physics-aware scientific reasoning, 2025. URL <https://arxiv.org/abs/2508.02956>.
- [13] Alireza Ghafarollahi and Markus J Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent ai. *Proceedings of the National Academy of Sciences*, 122(4):e2414074122, 2025.
- [14] Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7524–7555. Association for Computational Linguistics, 2025.
- [15] Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steve Farrell, Benjamin Erichson, Michael W. Mahoney, Andy Nonaka, and Zhi Yao. Matterchat: A multi-modal llm for material science, 2025. URL <https://arxiv.org/abs/2502.13107>.
- [16] Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 2025.
- [17] Sifan Wu, Huan Zhang, Yizhan Li, Farshid Effaty, Amirreza Ataei, and Bang Liu. Seeing beyond words: Matvqa for challenging visual-scientific reasoning in materials science, 2025. URL <https://arxiv.org/abs/2505.18319>.

- [18] Santiago Miret and N. M. Anoop Krishnan. Enabling large language models for real-world materials discovery. *Nature Machine Intelligence*, 7(7):991–998, 2025.
- [19] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, 1964.
- [20] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.
- [21] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175–2196, 2009.
- [22] Joseph W. Abbott et al. Roadmap on advancements of the fhi-aims software package, 2025. URL <https://arxiv.org/abs/2505.00125>.
- [23] OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, 2025. Accessed: 2025-08-17.
- [24] OpenAI. Introducing gpt-4.1. <https://openai.com/blog/gpt-4-1>, 2025. Accessed: 2025-08-17.
- [25] OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- [26] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. Accessed: 2025-08-17.
- [27] Meta-AI. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- [28] MistralAI. Pixtral 12b, 2024. URL <https://arxiv.org/abs/2410.07073>.
- [29] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- [30] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013. doi: 10.1063/1.4812323.
- [31] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996.

A Prompt Templates and Evaluation Criteria

Figures 4 and 5 show the prompt templates for the VQA task and the captioning task, respectively.

Prompt template for VQA task
You are given a density of states (DOS) plot. Please analyze the DOS and estimate the <property> of the material. Output exactly in JSON with keys 'reason', 'property', 'value', and 'unit'. Example: ```json <pre>{ "reason": "reason for your estimation", "property": "vbm", "value": -1.00, "unit": "eV" }</pre> ...

Figure 5: Prompt template for the VQA task on VBM energy inference. Values of "property," "value," and "unit" are varied depending on the target physical property.

Prompt template for captioning task
You are given a density of states (DOS) plot. Please analyze this DOS data and provide a concise, structured summary with the following points: 1. Key features in the DOS (peaks, gaps, Fermi level) 2. Electronic structure interpretation (metallic/semiconducting/insulating) 3. How this relates to material properties

Figure 6: Prompt template for captioning task.

Figure 6 presents the prompt template used in the scoring task for evaluating captions generated in the captioning task. In the scoring task, DOS captions generated by the LLMs are evaluated on three criteria—accuracy, depth, and fluency—using a five-point scale. Accuracy is scored higher when the caption correctly captures the features of the DOS; depth is scored higher when the caption provides deeper reasoning and stronger connections to material properties; and fluency is scored higher when technical terminology is used appropriately and the description is clear and concise. The same evaluation criteria were also applied in the assessments conducted by human experts.

Prompt template for scoring
You are given a DOS plot and a textual description of that plot. Please analyze the DOS plot, evaluate the textual description against it, and score it on three dimensions: accuracy, depth, and fluency. Scoring criteria (integers 1-5 only): - Accuracy: 1: Fails to capture the main peaks of the DOS or misinterprets the DOS. 3: Captures the major features but contains errors or omissions in the details. 5: Correctly identifies and describes peak positions, band gap width, and Fermi level. - Depth: 1: Limited to basic descriptions such as metallic/insulating behavior. 3: Provides simple correlations between DOS and physical properties. 5: Discusses the origins of peaks, links DOS shape to physical properties, and includes considerations for applications. - Fluency: 1: Misuses technical terms or uses unnatural expressions. 3: Generally readable but contains some ambiguous wording. 5: Clear, concise, and easy to read without causing confusion. Here is the description to evaluate: <LLM-generated DOS caption> Output exactly in JSON with `accuracy`, `depth`, and `fluency`. Example: ``json {"accuracy": 3, "depth": 4, "fluency": 2} ``

Figure 7: Prompt template for the scoring task. The placeholder **<LLM-generated DOS caption>** is replaced with a DOS caption generated by an LLM.

B Dataset preparation

The structural data obtained from the Materials Project[30] were used to calculate electronic structures with FHI-aims[21, 22], an all-electron DFT code employing numeric atom-centered orbitals. In this study, calculations were restricted to spin-unpolarized structures containing atoms with atomic numbers not exceeding 53 and fewer than 80 atoms in total. For the electronic structure relaxation, the *light* basis set was used, and the PBE[31] functional within the GGA approximation was employed as the exchange–correlation functional. As a result, a dataset of 8,991 converged calculations was obtained, each linking the computed DOS with the corresponding physical property values. DOS

data were stored as images with a resolution of 800×800 pixels. Here, the DOS was broadened using a Gaussian smearing width of 0.5 eV. Among all the data, 899 entries (10%) were sampled for the VQA task, and 20 entries were sampled for the captioning task.

C Example of DOS captions

Figure 8 presents an example of the input used in the captioning task, while Figures 9, 10, and 11 show the captions generated for this input by o4-mini, GPT-4o, and GPT-4.1-mini, respectively.

As shown in Figure 8, the DOS is provided in image form, and the prompt instructs the model to describe its key features, interpret the electronic structure, and discuss how these relate to material properties. In this example, the DOS illustrated in the figure represents the electronic structure of CaSO_4 , as obtained from DFT calculations.

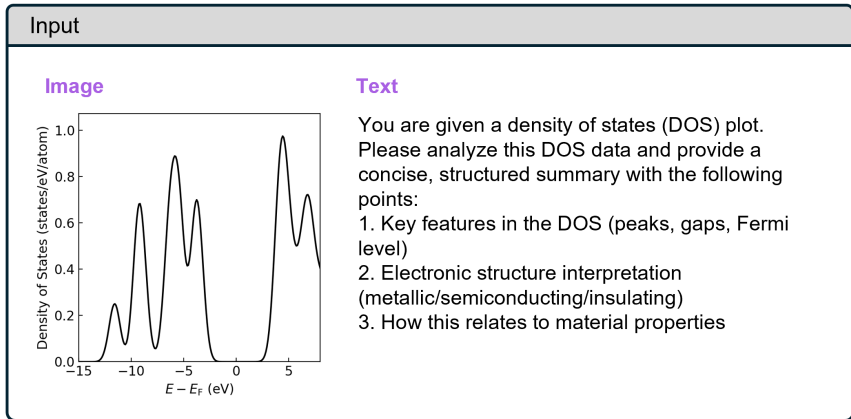


Figure 8: Example of inputs for the captioning task: a DOS diagram and the fixed text prompt used for all cases.

Based on this prompt, as shown in Figures 9, 10, and 11, the LLMs generate captions that describe peak features such as positions and intensities, provide interpretations of electronic properties from the electronic structure, and discuss their relation to material characteristics.

Figure 12 presents an example of captions generated by o4-mini for different DOS diagrams in the captioning task. Among the evaluated models, o4-mini achieved the highest scores in both accuracy and depth. The model was able not only to accurately extract multiple peak energies and DOS values but also, as shown in Figure 12, to include interpretations that relate DOS features to principles such as Hume–Rothery stabilization, offering insights that human experts also found insightful.

DOS caption by o4-mini
1.Key DOS features • Valence-band peaks: – ~-12 eV (0.25 states/eV/atom) – ~-9.5 eV (0.68 s/eV/atom) – ~-6.0 eV (0.90 s/eV/atom) – ~-4.0 eV (0.70 s/eV/atom) • Band-gap region: zero DOS from roughly -3 eV up to +3 eV (relative to EF) • Conduction-band peaks: – ~+4 eV (0.98 s/eV/atom) – ~+6 eV (0.50 s/eV/atom) – ~+7.5 eV (0.72 s/eV/atom) • Fermi level (EF) lies in the middle of this wide gap, with DOS(EF)=0 2.Electronic-structure classification • Zero DOS at EF → non-metallic • Large band gap (~7 eV, from top of valence near -4 eV to bottom of conduction near +3 eV) → wide-gap insulator 3.Implications for material properties • Very high electrical resistivity; negligible intrinsic conductivity • Transparency throughout the visible (gap exceeds visible photon energies) • Potential application as a dielectric (low leakage) or UV-transparent window • Likely good chemical/thermal stability typical of wide-gap oxides or nitrides

Figure 9: Example of a caption generated by o4-mini in the DOS captioning task.

DOS caption by GPT-4o
1.Key Features in the DOS: • Peaks: There are noticeable peaks around -12 eV, -8 eV, -4 eV, 2 eV, and 4 eV. These correspond to high density of electronic states at these energy levels. • Gaps: There is a significant gap around the Fermi level (0 eV), indicating a region with very low or no states. • Fermi Level: The Fermi level is positioned at 0 eV, which correlates with a relatively low DOS, suggesting minimal electronic states available at this energy. 2.Electronic Structure Interpretation: • The presence of a gap around the Fermi level suggests that the material is semiconducting. The energy gap divides the occupied and unoccupied states, typical of semiconductors. 3.Relation to Material Properties: • Semiconducting Behavior: The band gap implies that the material has semiconducting properties. At absolute zero, it behaves as an insulator, but with increasing temperature (or doping), electrons can be thermally excited across the gap, allowing conduction. • Applications: Materials with this kind of DOS are crucial in electronic and optoelectronic devices, as they can be used to control electron flow and are essential for transistors, diodes, and solar cells.

Figure 10: Example of a caption generated by GPT-4o in the DOS captioning task.

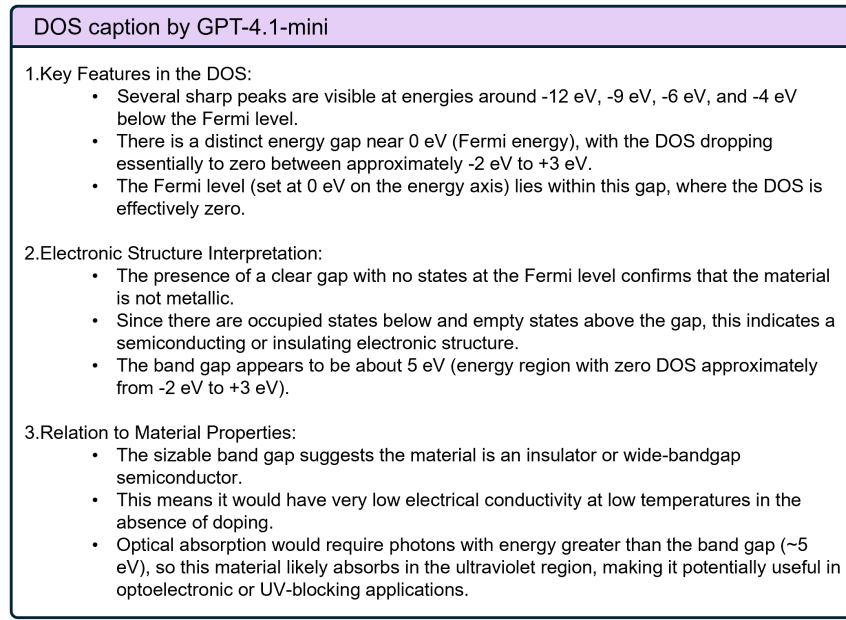


Figure 11: Example of a caption generated by GPT-4.1-mini in the DOS captioning task.

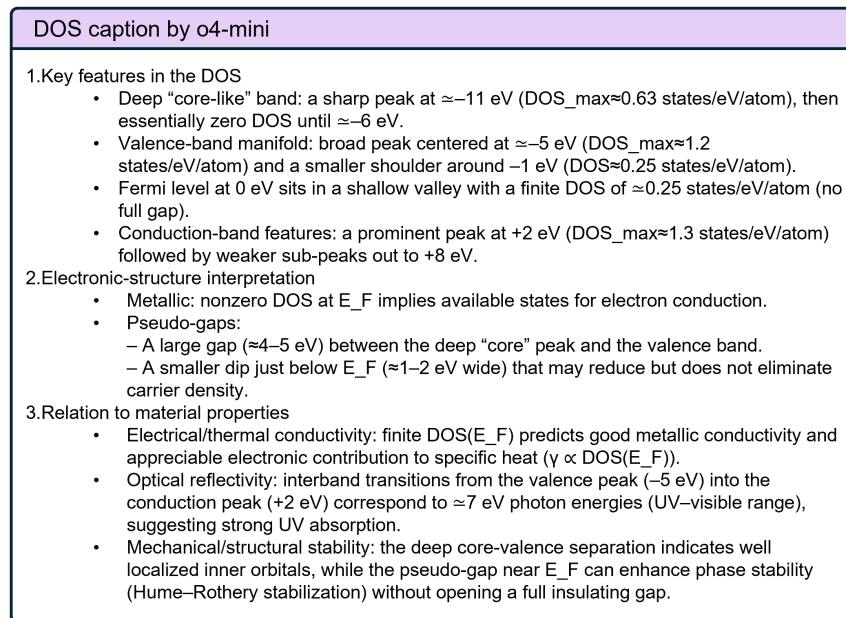


Figure 12: Example of a caption generated by o4-mini in the DOS captioning task.