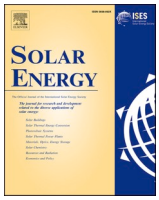
Solar Energy 315 (2026) 114801

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener



A unified large language model–based framework for heterogeneous PV image diagnosis

Baojie Li, Anubhav Jain *

Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ARTICLE INFO

Keywords:
PV image
Large language models
Fault diagnosis
Electroluminescence imaging
Infrared imaging
Photovoltaic systems

ABSTRACT

With advances in imaging technologies, modern photovoltaic (PV) systems generate large volumes of heterogeneous image data, including visible, electroluminescence (EL), and infrared (IR) images. Existing PV image analysis models, particularly deep learning approaches, are typically task-specific and lack cross-modality generalization. To address this limitation, this paper proposes an open-source large language model (LLM)–based unified framework for heterogeneous PV image diagnostics. Through task-aware diagnostic prompting, the framework enables analysis of visible, EL, and IR images within a single pipeline, supporting both zero-shot and few-shot inference and binary and multiclass classification. It is compatible with state-of-the-art multimodal LLMs, including ChatGPT, Gemini, Claude, Qwen, and CLIP. The framework is evaluated on PV module condition classification (clean, soiling, snow, hail, and bird droppings) using visible images, cell crack detection using EL images, and hotspot detection using IR images. GPT-5.1 in few-shot mode achieves the best performance, with classification accuracy exceeding 97.3%. Open-source models such as Qwen and CLIP also deliver competitive results on visible images (around 90% accuracy), though their performance is more limited on EL and IR modalities. On the full ELPV dataset, the framework achieves 83.5% zero-shot accuracy, within 2.8% of the supervised CNN baseline, confirming scalability to larger benchmarks. Practical aspects such as reproducibility, response latency, and confidence estimation are systematically analyzed. The framework operates across PV image modalities without modality- or task-specific training, making it well suited as a rapid pre-screening tool to support downstream detailed diagnostics. A benchmark dataset of diverse labeled PV images is also released.

1. Introduction

Large Language Models (LLMs) [1], like ChatGPT [2] and Gemini [3], have transformed how we interact with artificial intelligence. Initially focused on text-based tasks, these models excel at understanding and generating natural language. In recent years, LLMs have evolved into multimodal models [4,5], with outstanding capabilities not only in natural language processing tasks, but also remarkable potential in visual understanding and analysis [6,7]. These advanced multimodal LLMs offer innovative solutions for a wide range of tasks, including processing, understanding, and even generating images [8], enabling new possibilities for photovoltaic (PV) research [9,10].

The widespread deployment of PV imaging technologies such as drones and advanced cameras has resulted in the collection of extensive image datasets, including visible [11], electroluminescence (EL) [12], infrared (IR) [13], and photoluminescence (PL) images [14], etc. These images provide valuable insights into the health and performance of PV modules [15,16]. As a result, image-based inspection has become a routine component of operations and maintenance (O&M) workflows for both utility-scale and distributed PV installations [17]. As fleet sizes continue to grow, O&M teams face increasing pressure to inspect large numbers of modules efficiently, at low cost, and across heterogeneous data sources collected from drones, handheld cameras, and field thermographic or EL instruments [18].

Traditional PV image analysis methods commonly leverage artificial intelligence (AI) and machine learning (ML) techniques [19,20]. For visible image analysis, several approaches have been explored for different fault types. In the case of snow detection on PV modules, U-Net model has been applied in [21], while ResNet and VGG models are applied in [22]. For soiling analysis, Convolutional Neural Networks (CNN) −based approaches have been reported in [23], and BWR metric is applied in [24]. For bird dropping detection, YOLO models have been adopted in [25] and VGG model in [26]. As for the EL image analysis, machine learning models such as Artificial Neural Networks (ANN) and

* Corresponding author.
E-mail address: ajain@lbl.gov (A. Jain).

https://doi.org/10.1016/j.solener.2026.114801
Received 6 February 2026; Received in revised form 6 May 2026; Accepted 14 June 2026
Available online 25 June 2026

Support Vector Machines (SVM) are applied in [27,28]. Recently, deep learning methods have also been explored, including CNN [29,30], object detection models like YOLO in [31,32], and hybrid models in [15]. For IR image analysis, YOLO and ResNet architectures have been employed in [33,34], while CNN-based approaches in [35,36].

As can be seen from these studies, each model is designed for analyzing a specific type of PV image and cannot be generalized effectively across other image types. In addition, preparing a large, well-labeled dataset can be labor-intensive and time-consuming. Large public PV image datasets [11,12,13,14] and established transfer learning pipelines [37] have partly mitigated this burden, and recent advances in semi-supervised and self-supervised [38,39] learning further reduce the need for extensive labeling. Nevertheless, each new imaging modality or fault type generally still requires additional data curation and task-specific fine-tuning. This reliance on separate, task-specific models increases development cost, slows fault diagnosis, and limits scalability across diverse field conditions, motivating the need for a unified, training-free diagnostic framework that can flexibly handle heterogeneous PV imagery and serve as a rapid pre-screening tool in practical O&M deployment.

Some researchers have explored the application of traditional and multimodal LLMs in the PV field, with most efforts focused on tasks such as text mining and power prediction. For the literature or text mining, Liu et al. [40] use an optimized GloVe language model to analyze CdTe literature, accelerating material discovery for thin film CdTe solar cells. Sarahana et al. [41] apply NLP techniques, including AI-based sentiment analysis and topic modeling, to examine public perceptions and key themes related to solar energy deployment from public and social media texts. Kalle et al. [42] combine opinion mining, ChatGPT, and CA-GALT to assess global, regional, and local acceptance of solar power. For the power prediction, Lin et al. [43] proposed PV-VLM, a multimodal vision-language model that leverages sky images to improve intra-hour photovoltaic power forecasting. Zhang et al. introduced a GPT-based method for ultra-short-term distributed PV power forecasting, using virtual data for pre-training and limited real data for fine-tuning, which significantly improves prediction accuracy under data scarcity [44]. Yan et al. [45] proposed a probabilistic PV power forecasting method based on a multi-modal approach that uses a GPT-Agent to interpret weather conditions.

Research on PV image processing and analysis using LLMs remains quite limited. Deng et al. [46] develop CAP-PV, a language-aided few-shot learning framework based on CLIP model that leverages a CAP module to boost PV panel segmentation from remote sensing imagery. Boutana et al. [47] utilizes OpenAI's CLIP-based vision-language model to analyze visible images of PV panels and classify faults such as dust, snow, and physical damage. Both zero-shot and fine-tuned approaches are evaluated. Fine-tuning significantly improves fault detection performance, increasing accuracy from 38.5% to 92% for multi-class fault classification. However, this research is limited as it only evaluates the CLIP model on visible images, which cannot reflect the performance of the latest multimodal LLMs like GPT and Gemini or assess effectiveness on other important image types such as electroluminescence (EL) and infrared (IR) images. Guo et al. [48] proposed PVAL, a label-efficient LLM framework that detects, localizes, and quantifies rooftop PV panels in satellite imagery with 94% accuracy across six U.S. regions. By combining task decomposition, schema-guided outputs, and few-shot prompting, PVAL surpasses existing methods while enabling scalable, GIS-integrated PV assessment.

To date, most existing PV image analysis methods are task-specific and require separate models or retraining for each imaging modality. Beyond the PV domain, multimodal LLMs have demonstrated competitive performance on visual classification tasks in medical imaging [49], remote sensing [50], and industrial inspection [51]. This motivates exploring whether the same approach can be extended to the heterogeneous and technically specialized imagery encountered in PV diagnostics.

To complement these approaches, this paper proposes a unified LLM-based framework for heterogeneous PV image diagnostics that operates in zero-shot and few-shot regimes, reducing the overhead of domain-specific fine-tuning when new modalities or fault types are introduced. By leveraging task-aware diagnostic prompting, the framework enables the analysis of visible, electroluminescence (EL), and infrared (IR) images within a single pipeline. The contribution of this work lies in: (1) Design a unified diagnostic framework for heterogeneous PV image modalities (visible, EL, and IR) without modality-specific retraining; (2) Support both zero-shot and few-shot inference, as well as binary and multiclass classification tasks for different application scenarios; (3) Systematically accommodate state-of-the-art LLMs, while GPT-5.1 with few-shot mode achieves the best performance with classification accuracy over 97.3%; (4) Analyze the practical deployment considerations for the proposed framework, including reproducibility, prompt design, inference latency, and confidence levels; (5) Release a benchmark PV image dataset covering diverse image types to support evaluation of future image diagnostic models.

The remainder of this paper is organized as follows: Section 2 introduces the proposed framework, including the test image dataset, prompt design, inference mode, LLMs, and performance metrics. Section 3 presents PV image diagnosis results for visible, EL, and IR images. Section 4 further examines practical issues of the framework from reproducibility, prompt design, confidence level, and response time. Section 5 discusses the strengths, limitations, and future work direction of the proposed framework. Section 6 concludes the paper.

2. Methodology

The proposed unified LLM-based PV image diagnostic framework is illustrated in Fig. 1. This framework enables fault diagnosis of heterogeneous PV image modalities, including visible, electroluminescence (EL), and infrared (IR) images, within a single unified pipeline. It consists of task-aware diagnostic prompts supporting both binary and multiclass classification tasks, configurable inference modes (zero-shot or few-shot), and a multimodal LLM backbone, with representative models including GPT, Gemini, and Claude.

The model returns a structured JSON output containing the predicted class, a rubric-based confidence level, and a brief natural-language rationale, e.g., "the module surface is covered by a thick, uniform white layer with soft edges, consistent with snow rather than soiling", which explains why the LLM assigns the predicted fault type.

2.1. PV image test dataset

To evaluate the proposed framework, we build an open-source heterogeneous PV image test dataset, where three types of PV images are included: visible, electroluminescence (EL), and infrared (IR) images (https://github.com/DuraMAT/PV-LLM). To ensure variability and capture realistic conditions, these images were collected from different sources.

(i). PV visible image can reveal surface-level and environmental faults of PV modules. The dataset covers five common conditions—clean, snow, bird dropping, soiling, and hail, with images collected from the DuraMat Datahub [52] and open-source PV image datasets on Kaggle [53].

(ii). PV electroluminescence (EL) imaging detects internal defects and degradation invisible to the naked eye. The dataset includes EL images of healthy cells as well as cells with cracks at two severity levels: minor (small or mild cracks) and severe (large, multiple, or extensive cracks). Images are collected from [27] and public dataset on Kaggle [54]. These EL cell images are square in shape, with resolutions ranging from 100 × 100 px to 300 × 300 px.

(iii). PV infrared (IR) imaging can detect thermal anomalies that indicate underlying electrical or structural issues. The dataset

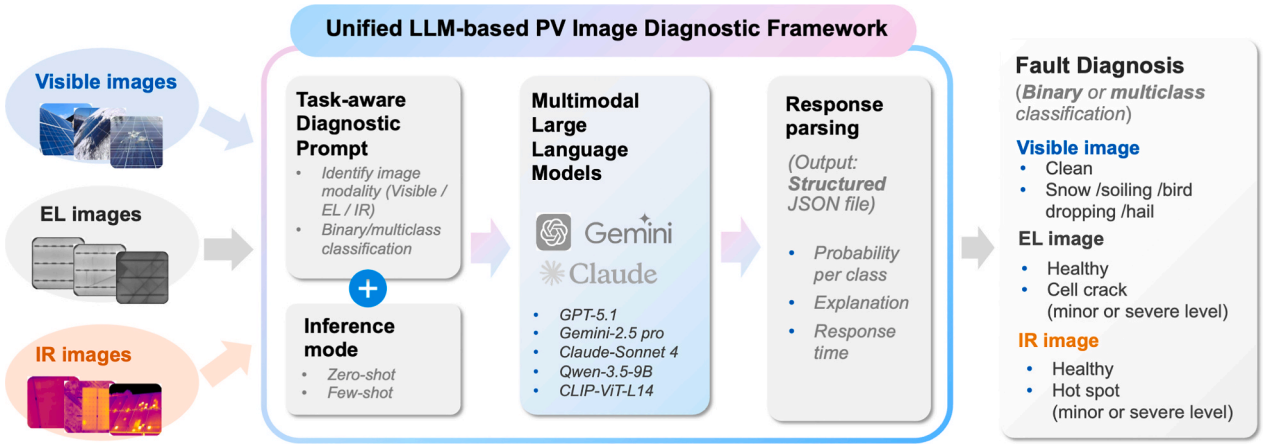


Fig. 1. Overview of the proposed unified LLM-based PV image diagnostic framework. The framework enables fault diagnosis of visible, electroluminescence (EL), and infrared (IR) PV images within a single pipeline, using task-aware diagnostic prompts (support binary and multiclass classification tasks), configurable inference modes (zero-shot or few-shot), a multimodal LLM backbone, and structured output parsing.

contains IR images of healthy PV modules as well as modules with hotspots at two severity levels: minor (a single localized thermal anomaly) and severe (multiple distinct hot regions within the module or array). Images are selected from [55,56].

The detailed configuration of the test dataset is presented in Table 1, where the classes for binary and multiclass classification are also listed. Binary classification focuses on detecting the presence of faults (e.g., healthy vs. faulty), serving as a fault alert mechanism. In contrast, multiclass classification provides fine fault diagnosis by distinguishing among specific fault types or severity. Examples of the test dataset are shown in Fig. 2.

Notably, these test images were selected to reflect a wide range of real-world scenarios. For example, for visible images, the test images include variations in shooting angle, scope (single or multiple PV modules), and presence of reflections from surrounding elements such as people, houses, trees, and the sun (like Fig. 2 (a) – Images 3, 12, 49). For EL images, the images exhibit variation in brightness and busbar number (three or four), adding to the complexity of the classification task. For IR images, the images feature wide variation in scales (from individual modules to large PV arrays), shooting angles, and thermal palettes or color mappings (e.g., Fig. 2 (c) – Images 11 and 34). This diversity makes the classification task more realistic and challenging, closely reflecting field conditions.

Table 1
Configuration of dataset.

Task	Visible images	EL images	IR images
Binary fault classification	• Clean ($50 \times 2 = 100$ images) • Faulty (contains images of snow, soiling, hail, bird dropping)	• Healthy ($50 \times 2 = 100$ images) • Crack (contains images of minor and severe crack)	• Healthy ($50 \times 2 = 100$ images) • Hotspot (contains images of minor and severe hotspot)
Multi-class fault classification	• Clean ($50 \times 5 = 250$ images) • Snow • Soiling • Hail • Bird dropping	• Healthy ($50 \times 3 = 150$ images) • Minor crack (small or mild cracks) • Severe crack (large, multiple, or severe cracks)	• Healthy ($50 \times 3 = 150$ images) • Minor hotspot (single hotspot) • Severe hotspot (multiple hotspot)

2.2. Task-aware diagnostic prompt design

Prompt design plays an important role in the proposed framework to perform reliable PV image diagnosis across heterogeneous PV images. To this end, we design a task-aware prompt that explicitly guides the LLMs through PV image identification, PV image type classification, and task-specific fault classification. The hierarchical design of the proposed prompt is illustrated in Fig. 3. Task routing, probabilistic constraints, and structured JSON output are enforced at each stage. For few-shot mode, example images (shown in Fig. 4) are encoded in *Base64* format and incorporated into the text-based prompt.

The designed prompt supports two application scenarios: binary and multiclass PV image classification (as defined in Table 1). The difference between the two scenarios lies in the definition of the target conditions. Full prompt specifications, including task-specific condition descriptions, are provided in the Supplementary Information (SI).

The LLM outputs an estimated confidence for each predicted class. To ensure these estimates are methodologically sound rather than free-form numeric values prone to hallucination, we adopt a structured four-level rubric: Low, Moderate, High, and Very High, mapped to numeric values of 0.25, 0.50, 0.75, and 1.00 respectively for downstream Brier Performance Index (BPI) [57] computation (presented in Section 4.1). This design follows the standardized severity-rating frameworks widely used in PV module reliability assessment (e.g., the IEA PVPS Task 13 Report [58]), in which discrete assessment levels are explicitly tied to observable evidence rather than free-form numeric scores.

The rubric is anchored to two complementary factors: (i) *class discrimination difficulty*, which differs between binary and multiclass tasks — for binary tasks, the rubric reflects the certainty of the clean/faulty distinction, whereas for multiclass tasks, it additionally requires the LLM to assess whether the selected class is clearly dominant relative to other candidate classes (e.g., distinguishing soiling from bird droppings, or minor from severe cracks); and (ii) *image quality and feature visibility*, with modality-specific evidence descriptors covering glare, sky reflections, viewing angle, and surface coverage for visible images; cell-structure sharpness, busbar contrast, noise, and vignetting for EL images; and thermal contrast, and palette ambiguity for IR images. The LLM is instructed to apply both factors jointly and to provide a 1–2 sentence rationale referencing specific visual evidence, ensuring that confidence claims are explicitly grounded in image content. The full rubric, modality-specific evidence criteria, and prompt template are provided in SI Section B.

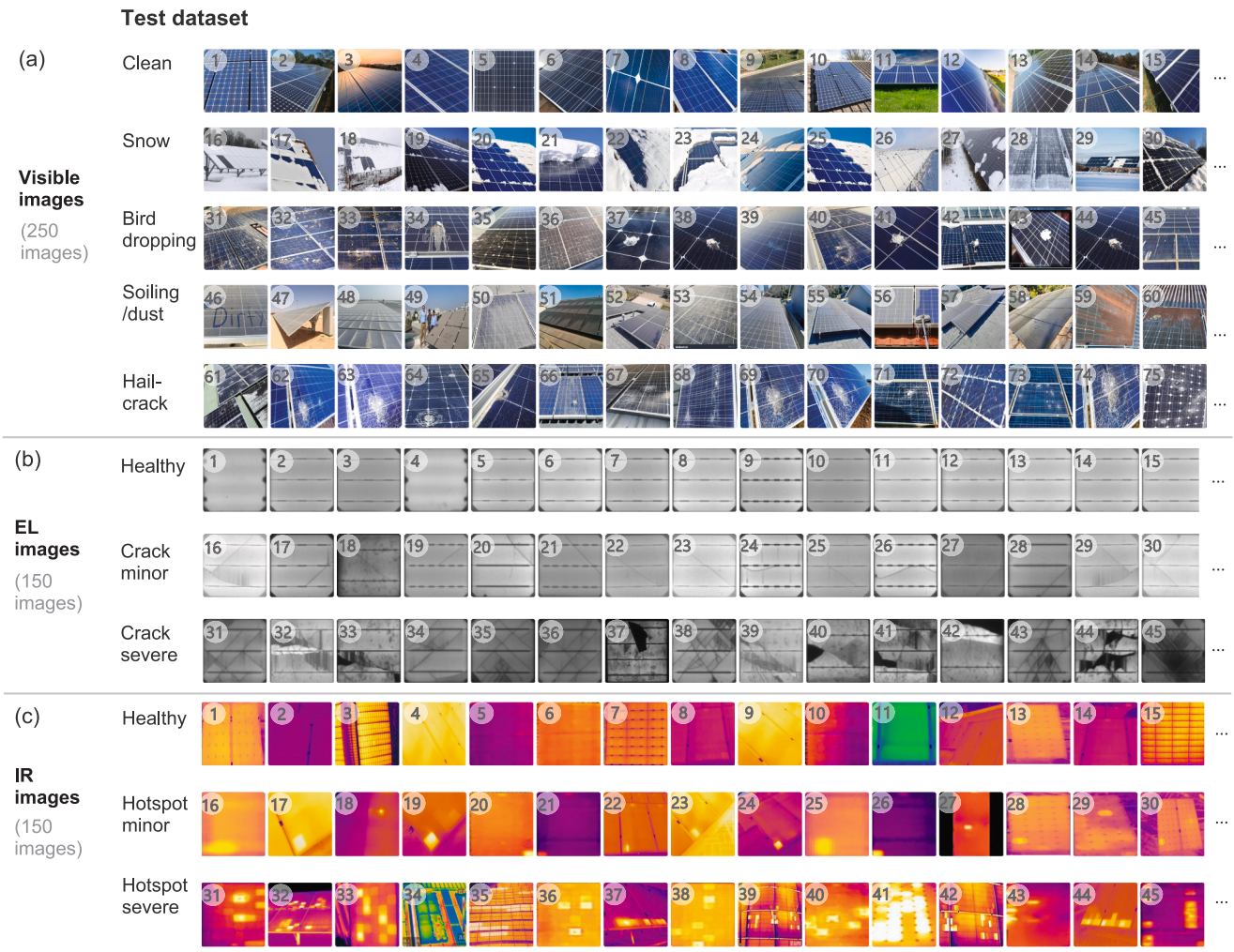


Fig. 2. Examples of test dataset. (a) visible, (b) EL, and (c) IR images. The dataset covers PV devices captured from diverse angles and perspectives, reflecting real-world inspection scenarios. Full dataset is available at https://github.com/DuraMAT/PV-LLM.

2.3. Multimodal large language models

To evaluate the proposed framework for PV image diagnostic, we select five recent multimodal vision-language models, covering three categories: (1) closed-source proprietary models: GPT-5.1 [59], Gemini 2.5 Pro [60], and Claude-Sonnet 4 [61]; (2) an open-weight multimodal LLM: Qwen3.5-9B [62], which supports offline deployment on consumer hardware via 4-bit quantization; (3) a contrastive vision-language model: CLIP ViT-L/14 [63], which can run offline and can provide reproducible results. This selection enables a comprehensive evaluation across different model architectures, access modes, and deployment constraints.

All models are integrated into the proposed unified framework through a standardized API-based interface, ensuring consistent input formatting, prompt structure, and output parsing across different vision-language models. The proposed framework is model-agnostic and extensible. New vision-language models can be readily incorporated by implementing the same interface without modifying the overall pipeline, prompts, or evaluation protocol.

2.4. Inference mode

The proposed unified framework supports two inference modes: (i) *zero-shot*: in which the LLM performs classification without access to labeled example images, relying solely on its pretrained knowledge and general visual reasoning capabilities; and (ii) *few-shot*: where a small number of labeled reference images are provided as in-context examples to guide the diagnostic process. The few-shot images are shown in Fig. 4.

For the CLIP baseline, the same two inference modes are implemented, but through embedding-space similarity rather than in-context prompting. In zero-shot mode, each class prototype is obtained by encoding a natural-language class definition with CLIP's text encoder, and a test image is assigned to the class with the highest cosine similarity to its visual embedding. In few-shot mode, the same labeled reference images used for the LLMs are encoded with CLIP's vision encoder and averaged per class to form visual prototypes, which are then combined with the text prototypes using equal weights (0.5/0.5). All embeddings are L2-normalized prior to similarity computation.

Under few-shot mode, for binary classification tasks, three representative PV images are selected for each category, while for multiclass tasks, two images are selected per category. This light-weight few-shot setting is adopted to provide sufficient diagnostic context while controlling prompt length.

2.5. Result parsing and error handling

The final step of the framework involves extracting the JSON-formatted text from the LLM output and converting it into a dictionary for subsequent analysis. The class associated with the highest predicted probability is selected as the diagnosed condition of the PV image.

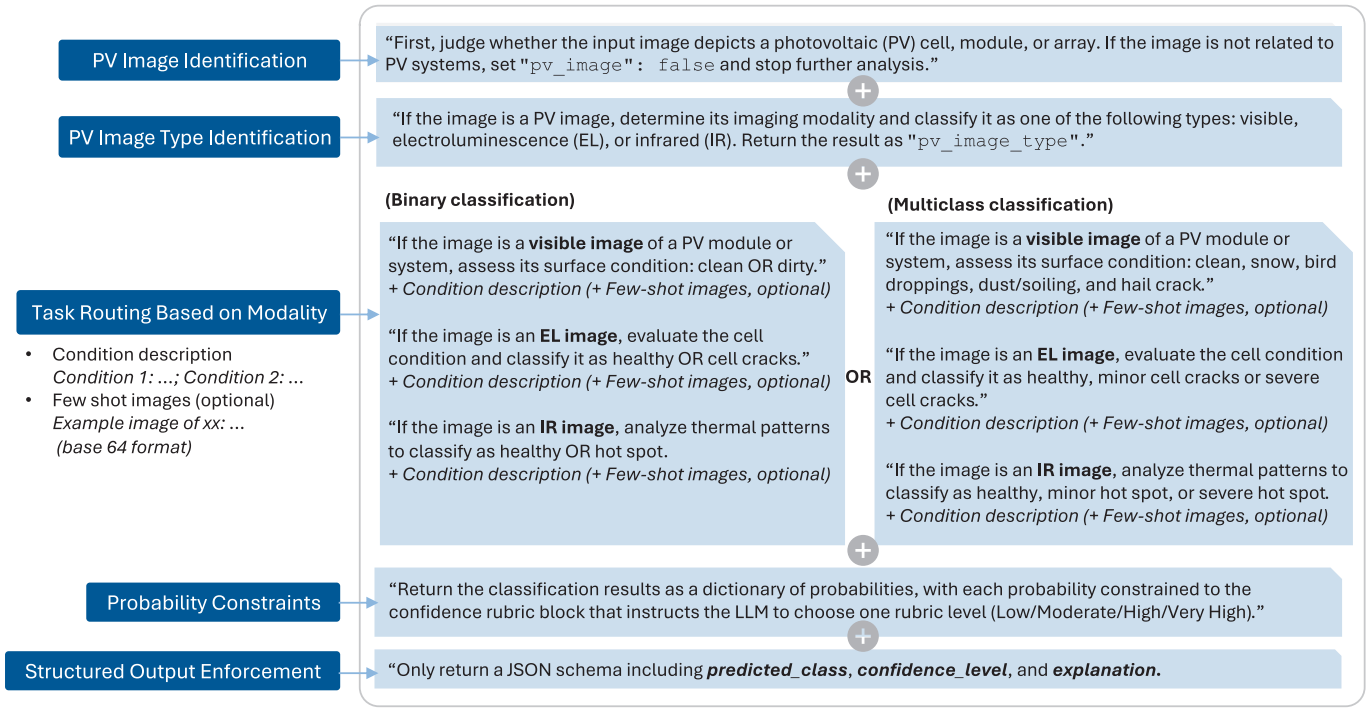


Fig. 3. Overview of the hierarchical design for unified PV image diagnostic prompt. The prompt decomposes the task into PV image identification, image type classification, and task-specific fault assessment with task routing, probabilistic constraints, and structured JSON output enforced at each stage. The designed prompt supports two application scenarios: binary and multiclass PV image classification.

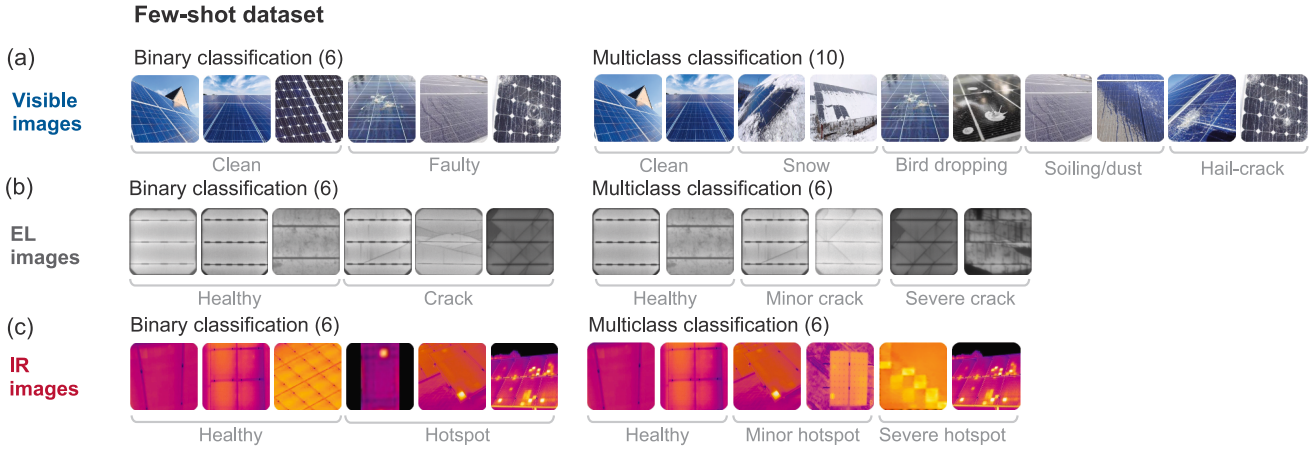


Fig. 4. Few-shot learning dataset. (a) visible, (b) EL, and (c) IR images. The dataset supports both binary and multiclass PV fault diagnosis tasks and is encoded into the prompt to guide the multimodal LLMs.

To ensure smooth and robust operation of the framework, several preventative functions are implemented to handle unexpected or anomalous outputs, such as malformed JSON, incorrect or missing class labels, inconsistent probability assignments, or incomplete responses. These safeguards improve the robustness of the pipeline, prevent runtime interruptions, and ensure reliable deployment of the proposed framework under a wide range of real-world conditions.

2.6. Evaluation metrics

We evaluate classification performance using three effective metrics: *Effective Accuracy*, *Effective Precision*, and the *Effective F1 Score* (Equations (1)–(5). These metrics are computed from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), and are scaled by the success rate ($R_{success}$, Equation (6). The success rate represents the proportion of valid and complete outputs in the correct JSON format. Since some LLMs may fail or generate invalid outputs for certain image analyses, scaling ensures that reported performance reflects both predictive quality and output validity.

$$EffectiveAccuracy = (TP + TN)/(TP + TN + FP + FN)*R_{success} \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$EffectivePrecision = Precision*R_{success} \quad (4)$$

$$EffectiveF1 = 2(Precision*Recall)/(Precision + Recall)*R_{success} \quad (5)$$

$$R_{success} = NumberofvalidJSONoutputs/Totaloutputs \quad (6)$$

3. PV image diagnostic results

This section evaluates the proposed framework using the PV image test dataset presented in Section 2.1. The framework with five LLMs under few-shot/zero-shot mode is tested for both binary and multiclass classification tasks. A summary of the results is shown in Fig. 5 with detailed classification metrics listed in Table 2. Detailed classification confusion matrices are listed in SI Section C.

It is shown that the proposed framework successfully handles the diagnostic of heterogeneous PV images within a unified pipeline. For the classification task, the binary classification task achieves higher accuracy than the multiclass task, since distinguishing between clean and faulty PV modules is inherently simpler than identifying specific fault types. Regarding the inference mode, few-shot learning generally outperforms zero-shot learning, as shown in Fig. 5, with an improvement of accuracy of 2.3–13.4%. Among the evaluated models, the proprietary online models (GPT-5.1, Gemini 2.5 Pro, and Claude-Sonnet 4) outperformed the open-source counterparts (Qwen3.5-9B and CLIP ViT-L/14). This performance gap is expected, as proprietary models benefit from substantially larger parameter counts, continuous updates, and access to broader-scale pretraining data. The open-source models achieved competitive accuracy on visible PV images, but their performance degraded notably on EL and IR modalities. In particular, Qwen3.5-9B exhibited the largest drop on EL imagery, while CLIP ViT-L/14 underperformed most on IR imagery. This modality-dependent gap likely reflects the underrepresentation of specialized PV imagery (EL grayscale and IR thermal) in the pretraining corpora of open-source vision-language models, which are predominantly trained on natural visible-light images. As for the success rate, Gemini occasionally fails to analyze the images, with a failure rate ranging from 6% to 23%, due to output errors or invalid responses, as shown in Table 2. Comparatively, GPT and Claude demonstrate a more stable and high success rate (near 100%).

From the perspective of PV image type, under the zero-shot mode, the classification accuracy for IR images is notably lower than that for visible and EL images (Table 2), with accuracy falling below 84%. This performance gap is likely attributable to the more complex and subtle visual patterns present in PV IR images, as well as their relatively limited representation in the pretraining data of current multimodal LLMs when compared to visible images.

Among the tested models, GPT with few-shot setting performs the best, achieving a high average classification accuracy of 97.3% under binary classification task, and 95% under multiclass classification task across the three types of PV images. Detailed classification result of GPT with few-shot setting is shown in Fig. 6.

From Fig. 6, overall, GPT-few shot achieves a good diagnosis performance across all the three types of PV images, with accuracy from 93 to 98%. For visible images, under binary classification task, 2 images were misclassified (faulty identified as clean), where only part of the PV array exhibits visible soiling, which likely led GPT to consider them as clean. As for the multiclass classification task, 19 out of 250 images were misclassified. Among them, 9 soiling images were similarly classified as clean, and 6 bird dropping cases were confused with soiling, which is reasonable given that the soiling or localized bird droppings can be visually ambiguous.

Regarding the EL images, the misclassifications of GPT-few shot fall into two main categories: healthy cells with background noise incorrectly identified as cracked, and minor cracks with faint dark lines or scratches misclassified as severe cracks.

As for the IR images, most misclassifications for GPT involve healthy modules being incorrectly labeled as hotspots. For the multiclass task, some images are classified as hotspots due to bright areas near the junction box, as seen in Fig. 6 – Images 5 and 6. Additionally, there is some confusion between minor and severe hotspots, due to ambiguous thermal patterns.

Overall, GPT-5.1 with few-shot is the best choice under the proposed unified PV image diagnostic framework, with an accuracy over 93% across visible, EL, and IR images under both binary and multiclass classification modes.

4. Reliability of the diagnostic framework

This section further examines the reliability and practical issues when implementing the proposed framework by analyzing result confidence (Section 4.1), impact of prompt design (Section 4.2), response time (Section 4.3), reproducibility (Section 4.4), and practical deployment in PV O&M (Section 4.5).

4.1. Confidence of results

The proposed framework converts the LLM outputs into a structured result, where each output is a dictionary containing the predicted class label, a confidence level selected from the four-tier rubric (Low / Moderate / High / Very High), and a short rationale referencing the visible evidence (see Section 2.2 and SI Section B). For analysis, the rubric levels are mapped post hoc to numeric values of 0.25, 0.50, 0.75, and 1.00 to enable quantitative comparison across models. We then compute the Brier Performance Index (BPI) [57] as in (7), which measures how well the model's rubric-derived confidence values match the true class outcomes while rewarding confident correct predictions and penalizing confident incorrect ones. A BPI close to 1 indicates more accurate and confident predictions, while a value close to 0 indicates

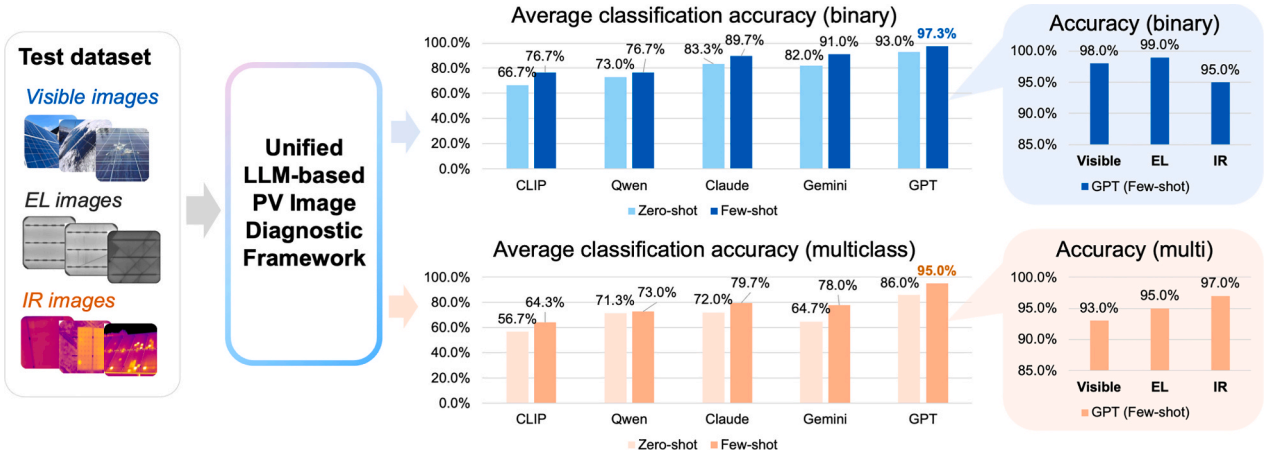


Fig. 5. Summary of results. The proposed framework successfully handles the diagnosis of heterogeneous PV images within a unified pipeline. GPT under few-shot mode achieves the best performance, with average classification accuracy of 97.3% (binary) and 95.0% (multiclass) across all the types of PV images.

Table 2
Classification results of PV image test dataset using 5 LLMs.

Image	Model	Binary classification				Multiclass classification			
		Accuracy	F1	Precision	Success	Accuracy	F1	Precision	Success
Visible	GPT (Zero-shot)	0.98	0.98	0.98	1.00	0.92	0.92	0.93	1.00
	Gemini (Zero-shot)	0.90	0.90	0.91	0.98	0.66	0.66	0.68	0.72
	Claude (Zero-shot)	0.91	0.91	0.91	0.95	0.76	0.75	0.82	0.98
	Qwen (Zero-shot)	0.90	0.90	0.92	1.00	0.72	0.71	0.77	1.00
	CLIP (Zero-shot)	0.85	0.85	0.88	1.00	0.69	0.66	0.76	1.00
	GPT (Few-shot)	**0.98**	**0.98**	**0.98**	1.00	**0.93**	**0.93**	**0.94**	1.00
	Gemini (Few-shot)	0.94	0.94	0.94	0.97	0.79	0.79	0.82	0.90
	Claude (Few-shot)	0.95	0.95	0.95	0.99	0.80	0.79	0.86	1.00
	Qwen (Few-shot)	0.92	0.92	0.93	1.00	0.74	0.74	0.78	1.00
	CLIP (Few-shot)	0.93	0.93	0.94	1.00	0.73	0.71	0.78	1.00
EL	GPT (Zero-shot)	0.98	0.98	0.98	1.00	0.82	0.81	0.86	1.00
	Gemini (Zero-shot)	0.79	0.79	0.79	0.80	0.72	0.68	0.79	0.91
	Claude (Zero-shot)	0.96	0.96	0.96	1.00	0.80	0.79	0.82	1.00
	Qwen (Zero-shot)	0.55	0.45	0.69	1.00	0.61	0.55	0.81	1.00
	CLIP (Zero-shot)	0.68	0.64	0.80	1.00	0.62	0.58	0.76	1.00
	GPT (Few-shot)	**0.99**	**0.99**	**0.99**	1.00	**0.95**	**0.95**	**0.96**	1.00
	Gemini (Few-shot)	0.95	0.95	0.95	0.95	0.59	0.59	0.59	0.59
	Claude (Few-shot)	0.96	0.96	0.96	1.00	0.91	0.91	0.92	1.00
	Qwen (Few-shot)	0.56	0.47	0.66	1.00	0.63	0.59	0.82	1.00
	CLIP (Few-shot)	0.86	0.86	0.89	1.00	0.75	0.72	0.82	1.00
IR	GPT (Zero-shot)	0.84	0.84	0.88	1.00	0.84	0.84	0.87	1.00
	Gemini (Zero-shot)	0.77	0.76	0.82	1.00	0.69	0.68	0.75	1.00
	Claude (Zero-shot)	0.63	0.58	0.73	1.00	0.60	0.58	0.66	1.00
	Qwen (Zero-shot)	0.74	0.72	0.83	1.00	0.81	0.79	0.85	1.00
	CLIP (Zero-shot)	0.47	0.45	0.47	1.00	0.39	0.37	0.38	1.00
	GPT (Few-shot)	**0.95**	**0.95**	**0.95**	1.00	**0.97**	**0.97**	**0.97**	1.00
	Gemini (Few-shot)	0.84	0.84	0.85	0.94	0.83	0.83	0.85	0.97
	Claude (Few-shot)	0.78	0.77	0.83	1.00	0.68	0.68	0.80	1.00
	Qwen (Few-shot)	0.82	0.81	0.87	1.00	0.82	0.82	0.85	1.00
	CLIP (Few-shot)	0.51	0.37	0.59	1.00	0.45	0.42	0.49	1.00

poor predictive performance. CLIP ViT-L/14 model is excluded from this analysis as it does not generate rubric-based confidence but relies on embedding-space similarity scores. The results for the four LLMs under the few-shot setting across three PV image classification tasks are presented in Fig. 7.

$$BPI = 1 - \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{7}$$

where, p_i is the model's rubric-derived confidence; y_i is the output status (1 if output is correct, 0 if wrong); N is the number of images.

From the model perspective, GPT consistently attains the highest BPI ($\geq$ 0.97) across all three tasks, reflecting both strong classification accuracy and well-calibrated confidence. Gemini performs comparably on visible (0.98) and EL (0.98) but declines on IR (0.93), likely due to the greater variability in thermal scale and color palettes within the IR test set, which makes interpretation more challenging for current LLMs. Claude follows a similar trend, remaining strong on visible (0.98) and EL (0.98) with a modest decline on IR (0.95). Qwen, the open-weight model evaluated, records consistently lower BPI across all tasks (0.89 on visible, 0.73 on EL, 0.72 on IR), with the largest gaps emerging on EL and IR imagery. This performance gap is consistent with the limited capacity of open-weight multimodal LLMs on specialized imagery such as EL grayscale and IR thermal images, which are underrepresented in their pretraining corpora.

From the task perspective, the framework yields higher-confidence outputs on visible and EL tasks across the proprietary models (GPT, Gemini, and Claude), likely because EL cell images exhibit clear, localized crack patterns and visible PV imagery is well represented in pretraining data. The IR task is the most difficult for all proprietary models, as reflected by the slight BPI drops in Gemini, Claude, and Qwen. In contrast, Qwen shows comparably low BPI on both EL and IR, further illustrating the modality-specific limitations of open-weight LLMs in PV diagnostics.

To verify that the reported BPI values reflect genuinely informative confidence rather than artifacts of high accuracy alone, we further benchmark each model's rubric-derived BPI against three constant-confidence baselines (c = 0.5, c = 1.0, and c = accuracy). The latter constitutes the theoretical ceiling achievable by any uniform confidence assignment under the Brier-based BPI formulation. Across all three modalities and both GPT and Claude, the rubric-based BPI exceeds the c = accuracy ceiling, with the most pronounced improvement observed on Claude's IR predictions (Δ = +11.7%). This confirms that the rubric mechanism produces per-sample confidence signal that cannot be replicated by any constant assignment, validating the use of BPI as a meaningful evaluation metric in this study. Detailed baseline comparisons are provided in SI Section D.

4.2. *Impact of prompt design*

To assess the value of the structured task-aware prompt, we conducted a targeted ablation study comparing three prompt variants across all models under zero-shot multiclass classification: (A) the full prompt (proposed), incorporating both task routing and condition instruction; (B) a partial prompt without condition instruction; and (C) a partial prompt without task routing. The average multiclass classification accuracies are reported in Table 3. CLIP is excluded from this ablation as its embedding-based classification mechanism does not support analogous prompt variants.

The results show that the proposed Prompt A consistently achieves the highest accuracy across all four models, outperforming Prompt B by 3–6% and Prompt C by 14–19%. Removing task routing (Prompt C) causes the largest degradation across all models, where GPT drops from 0.86 to 0.72, Gemini from 0.78 to 0.60, Claude from 0.80 to 0.61, and Qwen from 0.67 to 0.51. This indicates that task routing is the more critical component. Removing the condition instruction (Prompt B) yields a smaller but consistent decline, confirming that explicit condition definitions further sharpen class discrimination. Notably, Qwen exhibits the same degradation pattern as the proprietary models, suggesting that structured task-aware prompting benefits open-weight

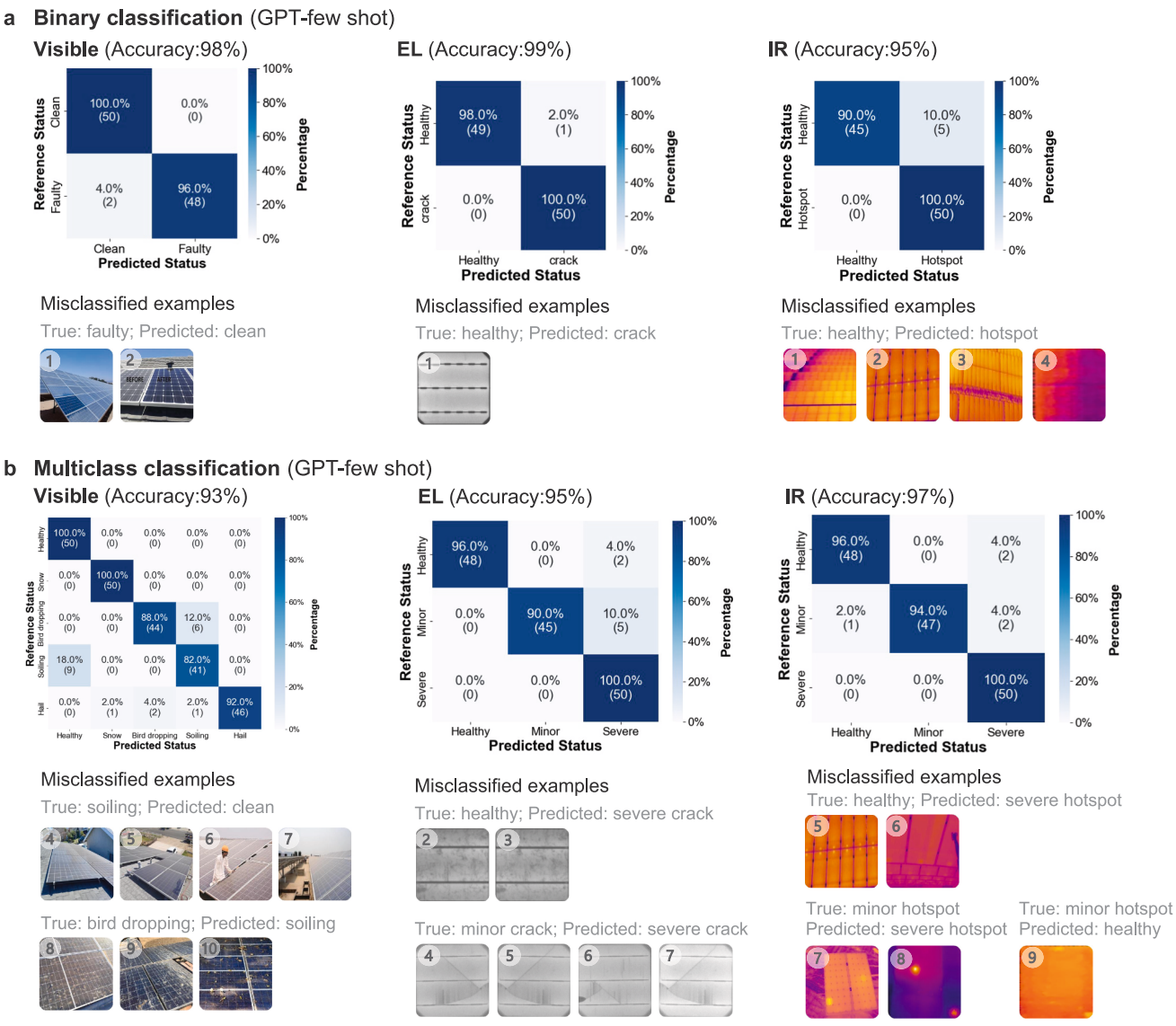


Fig. 6. Diagnosis results using GPT-few shot under the unified diagnosis framework (a) binary classification results with confusion matrix and examples of misclassified images; (b) multiclass classification with confusion matrix and examples of misclassified images. GPT under few-shot mode achieves a high accuracy (93–98%) across heterogeneous PV images.

multimodal LLMs as well, although its overall accuracy remains lower due to model scale. Together, these findings demonstrate the complementary benefit of structured task-aware prompting.

4.3. Response time

Response time is an important factor when implementing the LLM-based PV image diagnostic framework, especially in real-world applications where latency matters. Fig. 8 presents the distribution of inference times for five models: GPT, Gemini, Claude, Qwen, and CLIP, under zero-shot and few-shot settings across three binary classification tasks (visible, EL, and IR).

To ensure transparency and fair interpretation, we note the following context: (1) GPT-5.1 and Claude-Sonnet 4 were accessed via their standard developer API tiers; latency depends on API tier, geographic region, server load, and input/output token count. Our queries used approximately 1,500–2,200 input tokens (including Base64-encoded images) and 80–150 output tokens per request. (2) Gemini 2.5 Pro uses an extended reasoning (thinking) mode by default, which will increase response time. (3) Qwen3.5-9B and CLIP ViT-L/14 were evaluated locally as open-weight models on a MacBook Pro (2023, M3 Pro, 32 GB) using the integrated GPU. (4) Specialized inference providers (e.g., Groq [64], Together AI [65]) can deliver sub-second latency for compatible open-source models such as CLIP, and commercial API latency will vary across plans, providers, and time.

From the model aspect, CLIP, as a deterministic model, is much faster than all other models, with an average response time within 0.4 s per image. Among LLMs, Claude and GPT demonstrate the fastest performance in the zero-shot setting, taking around 2–3 s per image. Gemini, due to its default 'reasoning mode', takes an average inference time of 7–10 s per image with a notably wider distribution. Qwen exhibits high variability across tasks: while its zero-shot inference is comparable to GPT and Claude (~3–5 s), its few-shot inference is the slowest among all evaluated models, reaching 12–17 s per image, particularly on the IR task. This reflects the increased computational cost of processing multiple in-context image examples on the local hardware used for the open-weight model.

From the task perspective, the EL task requires the least inference time across most models. This is likely because the test EL images are simpler in nature, being grayscale, smaller in size (20–50 kB), and more

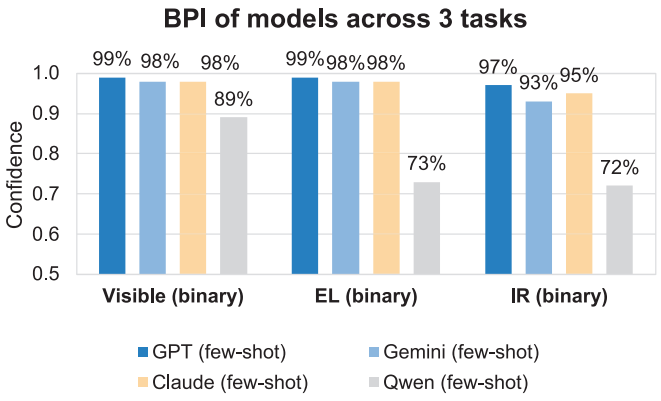


Fig. 7. Brier Performance Index (BPI) of four LLMs across three binary classification tasks. Proprietary models (GPT, Gemini, Claude) achieve high BPI on visible and EL imagery (≥0.98) but lower confidence on IR, indicating that thermal imagery is more challenging due to greater variability in color palettes and less distinctive defect patterns. GPT attains the highest BPI across all tasks (≥0.97), while the open-weight Qwen records consistently lower values, with pronounced drops on EL (0.73) and IR (0.72), reflecting the limited capacity of open-weight multimodal LLMs on specialized PV imagery underrepresented in their pretraining data.

Table 3
Comparison of average multiclass classification accuracy using 3 types of prompts.

Model	Prompt A (Full prompt, proposed)	Prompt B (No condition instruction)	Prompt C (No task routing)
GPT	0.86	0.83	0.72
Gemini	0.78	0.72	0.60
Claude	0.80	0.74	0.61
Qwen	0.67	0.62	0.51

uniform with fewer complex visual patterns, which allows for faster processing. The IR task tends to incur the longest inference time, especially in few-shot settings, due to its richer color information and the additional reasoning required to interpret thermal patterns.

As for the few-shot setting, which yields better classification performance (as presented in Section 3), the inference time increases by approximately 67% compared to zero-shot on average, reflecting the added computational overhead introduced by in-context examples. This trade-off is most pronounced for Qwen3.5-9B, where few-shot latency increases by 3–4 times over zero-shot. These observations emphasize the trade-offs between accuracy and latency when choosing between zero-shot and few-shot settings for time-sensitive applications and suggest that proprietary API-based models currently offer a more favorable balance for deployment scenarios where both accuracy and latency are critical.

Note that the latencies reported here should be interpreted as indicative benchmarks for the specific API configurations and hardware used, not as absolute performance figures.

4.4. Reproducibility

For the LLM-based PV image diagnostic framework, considering the non-deterministic nature of LLMs, the reproducibility becomes one important factor. To assess the reproducibility, we conducted three repeated runs for each binary classification task, Visible, EL, and IR, using the same set of 100 images. In each prompt, we explicitly set the '*temperature*' parameter to 0. For LLMs, '*temperature*' controls output randomness, where 0 makes responses deterministic, while higher values increase variability. To maximize reproducibility, we document the following for each model: exact model version identifier (GPT-5.1,

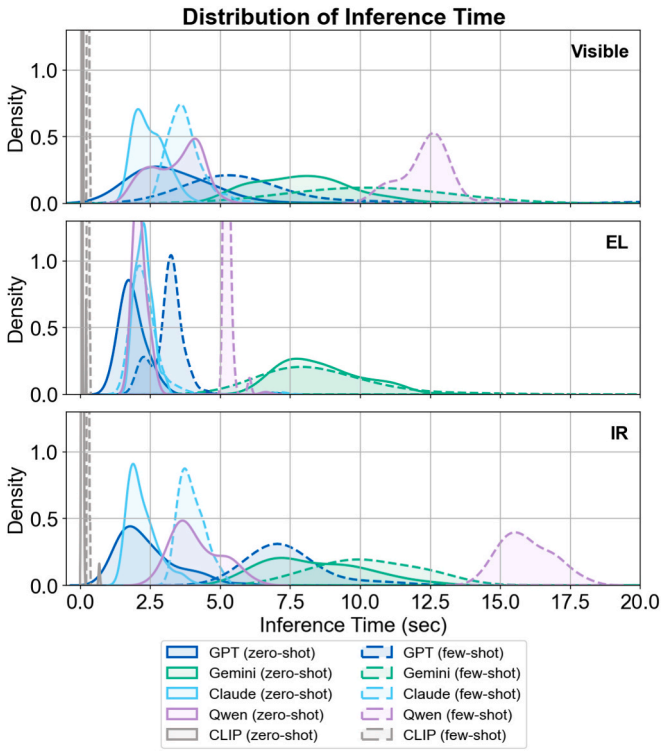


Fig. 8. Distribution of inference time for five models (GPT, Gemini, Claude, Qwen, and CLIP) under the proposed framework with zero-shot and few-shot settings across three binary classification tasks (visible, EL, IR). Overall, the few-shot setting increases inference time. CLIP, as a local deterministic model, is the fastest, with an average response time under 0.4 s per image. Among LLMs, Claude and GPT in zero-shot mode are the fastest, with a mean inference time of approximately 2–3 s per image, while Gemini exhibits a slower and wider response distribution (~7–10 s) due to its default reasoning mode. Qwen, evaluated locally as an open-weight model, shows the largest few-shot overhead, reaching 12–17 s per image on the IR task.

Gemini-2.5-Pro, Claude-Sonnet-4), API access date range, and approximate input/output token counts per query. This will allow the evaluation conditions to be reproduced as closely as possible.

We visualize these results in Fig. 9, where the reproducibility patterns of each model are shown across the 100 test samples for all three tasks. We define a '*Match*' as a case where all three outputs are identical, indicating perfect reproducibility. A '*Partial match*' occurs when at least one of the three outputs differs. An '*Invalid*' result is recorded when at least one of the outputs is unrecognizable or cannot be interpreted.

As shown in Fig. 9, the framework with GPT or Claude exhibits high reproducibility, with over 96% of samples yielding consistent predictions across all three runs. In contrast, Gemini displays a substantially higher rate of invalid outputs (18–29%) and partial matches (4–6%). Notably, most of Gemini's invalid results occur when classifying clean or healthy PV images, corresponding to the first 50 samples in Fig. 9 (a–c).

Despite the deterministic setting (*temperature* = 0), all tested frameworks with different multimodal LLMs still exhibit a degree of variability in their outputs. Nevertheless, GPT and Claude demonstrate considerably greater stability compared to Gemini in the proposed diagnostic framework.

4.5. Practical deployment in PV O&M

From a practical O&M standpoint, the proposed framework is well suited as a rapid first-pass screening tool within existing drone-based inspection pipelines. Its structured JSON output — containing the class label, per-class confidence scores, and a natural-language rationale — can be directly ingested by asset management systems (e.g., to flag

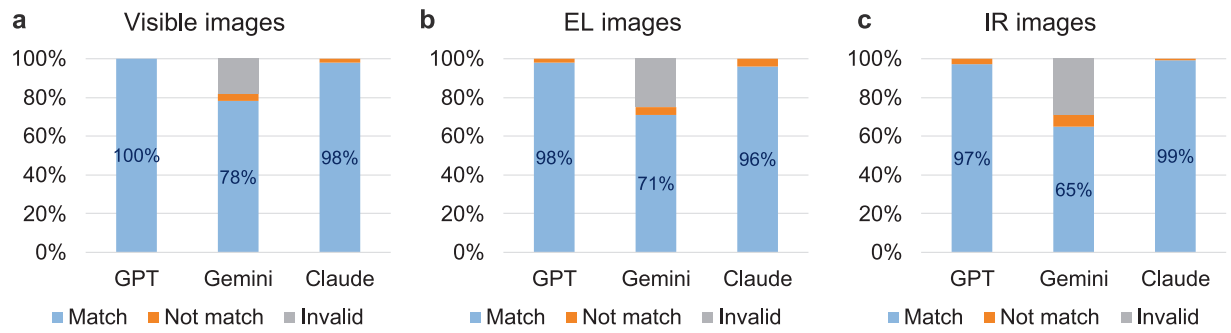


Fig. 9. Reproducibility results of three LLMs (zero-shot) across three binary classification tasks: (a) Visible, (b) EL, and (c) IR. Each model was evaluated three times on the same 100 images. Both GPT and Claude demonstrate high reproducibility, with the percentage of all-three-test matches exceeding 96%, highlighting their stability of output under the proposed framework.

modules for follow-up inspection) without additional post-processing. The zero-shot capability is particularly valuable for new plant types or emerging fault signatures for which labeled training data are not yet available. The main constraints for field deployment are API latency (2–8 s per image for online models, see Section 4.3) and dependence on internet connectivity; for sites with limited connectivity, the local CLIP model provides offline operation at the cost of lower accuracy.

A promising direction for future investigation is a hybrid architecture, in which the LLM-based framework performs initial screening and confidence-weighted routing, while specialized deep learning models (e.g., SegFormer [66] for hotspot localization, YOLO for crack extraction [31]) handle precision-critical downstream tasks.

5. Discussion

Multimodal LLMs serve as the core of the proposed framework, which brings a major advantage compared to image-trained deep learning models (e.g., CNNs, YOLO, U-Net), including free of requiring large-scale data, no tedious and task-specific training, and strong generalization across PV image types. As evaluated in Section 3, the test dataset reflects a wide range of real-world scenarios, including large variations in image size, image format, shooting angle, brightness, scope (from single PV module to entire array), presence of reflections from surrounding, number of busbars (for EL cell image), and even thermal palettes or color mappings (for IR images). Despite this high degree of heterogeneity, the great majority of cases were correctly handled by the LLM-based framework, achieving an overall accuracy of 97.3%. These results highlight the robustness and practical applicability of multimodal LLMs applied in the framework.

Reproducibility is an inherent limitation when proprietary multimodal LLMs are used, since these models are continuously updated by their providers and their internal weights are not publicly accessible. For applications requiring fully reproducible results, open-source models such as Qwen3.5-9B and pretrained deterministic models such as CLIP can be employed. However, the performance of open-source models remains inferior to that of state-of-the-art proprietary multimodal LLMs. In this study, CLIP ViT-L/14 achieved an overall binary classification accuracy of 66.7% in the zero-shot setting and 76.7% in the few-shot setting, while Qwen3.5-9B reached 73.0% (zero-shot) and 76.7% (few-shot), compared to 97.3% achieved by GPT-5.1 under few-shot conditions. These results indicate that, while open-source and pretrained ViT-based models offer reproducibility and the advantage of offline deployment, state-of-the-art proprietary MLLMs demonstrate substantially stronger performance on heterogeneous PV images.

To enhance reproducibility for LLM-based applications, as adopted in this study, several practical strategies are recommended: (i) explicitly reporting the exact model versions and API access dates, and fixing decoding parameters (e.g., temperature = 0) to minimize stochastic variation; (ii) including open-source offline models (such as CLIP and Qwen) as a reproducible baseline alongside proprietary models; (iii) applying ensemble-based strategies such as repeated inference with majority voting or confidence-based aggregation to stabilize outputs; and (iv) releasing benchmark datasets, prompts, and inference scripts so that the pipeline can be re-evaluated with any current or future LLM.

Although no traditional CNN model can operate across all three PV image modalities, we compared the proposed framework against a task-specific supervised CNN baseline [27] on the ELPV dataset as a representative benchmark. The framework with GPT-5.1 achieves 83.5% accuracy zero-shot, 2.8% lower than the proposed supervised CNN model (86.3%), despite requiring no task-specific training. Moreover, unlike the CNN baseline, which is restricted to a single modality and task, the proposed framework operates uniformly across visible, EL, and IR imagery, offering cross-modal generalization that single-modality supervised models cannot provide.

The proposed framework is designed for classification of PV images, where multimodal LLMs under this framework demonstrate strong performance (like GPT-5.1). However, for more advanced and precision-critical tasks, such as hotspot area localization, cell crack extraction, and power loss quantification, their performance remains limited (see SI Section E for details). Some pretrained computer vision models available on Hugging Face, such as SegFormer [66] and RMBG [67], show potential for specific image preprocessing tasks, including image segmentation and object detection. These models could be further explored or integrated with the proposed framework for future applications in the PV domain. While the LLMs-based framework cannot yet fully replace image-trained deep learning models, they are promising for pre-analysis and classification, supporting downstream deep learning models in delivering detailed diagnostics.

A potential limitation of evaluating commercial multimodal LLMs is that the publicly available PV image datasets used in this study may overlap with the models' training set, which cannot be fully verified for closed-source models. Several observations, however, suggest that this concern has limited impact on our findings. To directly test for potential memorization, we evaluated the framework on a held-out subset of EL images from the DuraMAT project (n = 32) [52], publicly released in August 2025 — after the September 2024 knowledge cutoff of GPT-5.1. The framework achieved 97% accuracy on this held-out set, consistent with the 98% overall benchmark accuracy, providing direct evidence that performance reflects genuine visual reasoning rather than memorization. In addition, the relative performance trends across models, modalities, and inference modes are consistent and interpretable, and few-shot prompting yields a consistent 2.3–13.4% improvement over zero-shot. Both observations are inconsistent with memorization-driven results. Future evaluations on additional newly collected, non-public PV image datasets will further help isolate any potential training-data overlap.

The major contribution of this work lies in the systematic evaluation and unified deployment of multimodal LLMs for heterogeneous PV

image analysis, rather than in proposing a new model architecture or learning algorithm. The underlying LLMs are used unmodified, with the value coming from the unified prompting design, structured-output handling, robustness safeguards, and benchmarking under realistic diagnostic scenarios. Building on this foundation, a particularly promising direction enabled by this framework is to move beyond reliance on a single PV image modality for diagnostics and instead integrate multiple complementary data sources to achieve a more comprehensive assessment of PV module health. Such cross-modality inputs may include combinations of visible, EL, and IR images, as well as the fusion of 2-D images with 1-D production data such as power output and I–V curves. This multimodal diagnostic setting is well suited to multimodal LLMs, which naturally leverage cross-modal reasoning, enabling flexible integration of heterogeneous inputs. Such capabilities of LLMs position the proposed framework as a promising foundation for future AI agent–based, expert-level diagnostic systems for large-scale PV installations.

While the present study evaluates five representative models spanning closed-source frontier multimodal LLMs and open-source baselines (Qwen3.5-9B and CLIP ViT-L/14), the proposed framework is model-agnostic and can be readily extended to other multimodal LLMs as the field evolves. In addition, future online proprietary models, such as upcoming versions of the GPT, Gemini, and Claude series, can be seamlessly integrated through standard API interfaces to deliver improved performance.

Given the rapid release cadence and frequent deprecation of MLLMs, we have released the benchmark PV image dataset publicly at https://github.com/DuraMAT/PV-LLM to support continuous evaluation of emerging models within the proposed framework. An interactive web interface is also available for intuitive exploration of this LLM-based framework for PV image diagnosis at https://pvtools.lbl.gov/pv-image, where users can upload images and view both the predicted class probabilities and the accompanying natural-language explanations generated by the LLM.

Key directions for future work include: (i) extending the framework to fine-grained tasks such as defect localization, severity grading, and degradation tracking by integrating specialized deep learning models, where the LLM performs cross-modality screening and the specialized models handle precision-critical tasks; (ii) integrating 2D PV images with 1D operational data (e.g., I–V curves, power output, weather conditions) within a single LLM-based reasoning pipeline to enable advanced fault diagnosis and root-cause analysis at the system level; (iii) validating the framework on larger, geographically diverse field datasets covering a wider range of module technologies, mounting configurations, and environmental conditions; and (iv) deploying the framework as part of an AI-agent–level health monitoring system with closed-loop feedback to inspection scheduling and asset management.

6. Conclusion

This work proposes a unified framework for heterogeneous PV image analysis based on multimodal large language models (LLMs). By leveraging task-aware diagnostic prompting, the framework enables diagnosis of PV visible, electroluminescence (EL), and infrared (IR) images within a single pipeline. Both binary and multiclass classification and inference modes are supported. The proposed framework can effectively accommodate diverse model types, including online LLMs (GPT-5.1, Gemini 2.5 Pro, and Claude Sonnet 4) and local open-source model (CLIP ViT-L14). Among these models, GPT-5.1 with few-shot learning demonstrates outstanding performance, achieving average accuracy of 97.3% across all image types. This framework leverages the strengths of multimodal LLMs, including strong adaptability to diverse input PV images in terms of image size, spatial scope, and viewing angle, while avoiding modality-specific and task-specific training processes. It is well suited as a rapid pre-screening tool for PV image analysis to support detailed diagnostics. Moreover, this framework can seamlessly accommodate future LLMs and provide a promising foundation for cross-modality PV diagnostics—such as the joint integration of 2D PV images with 1D PV data (e.g., power data and I–V curves), toward AI agent–level comprehensive health monitoring of PV systems.

CRediT authorship contribution statement

Baojie Li: Visualization, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Anubhav Jain:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Lawrence Berkeley National Laboratory is funded by the DOE under award DE-AC02-05CH11231. Funding provided by U.S. DOE Office of Critical Minerals and Energy Innovation (CMEI) Solar Energy Technologies Office (SETO) and as part of the Durable Module Materials Consortium 2 (DuraMAT 2) funded by the U.S. DOE Office CMEI SETO, Agreement 38259. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.solener.2026.114801.

Data availability

The benchmark test dataset is available at https://github.com/DuraMAT/PV-LLM. An interactive web interface for intuitive exploration of multimodal LLM performance on PV image analysis is also provided at https://pvtools.lbl.gov/pv-image.

References

[1] A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models, Nat. Rev. Phys. 5 (5) (2023) 277–280, https://doi.org/10.1038/S42254-023-00581-4.
[2] T.B. Brown, et al., language models are few-shot learners, Adv. Neural Inf. Process. Syst. Vol. 2020-December (2020). Accessed: Jul. 28, 2025. [Online]. Available: https://arxiv.org/pdf/2005.14165.
[3] G. Team *et al.*, "Gemini: A family of highly capable multimodal models," Dec. 2023, Accessed: Jul. 28, 2025. [Online]. Available: https://arxiv.org/pdf/2312.11805.
[4] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2023, Accessed: Jul. 27, 2025. [Online]. Available: https://arxiv.org/pdf/2303.08774.
[5] Google DeepMind, "Gemini: A family of highly capable multimodal models." Accessed: Jul. 27, 2025. [Online]. Available: https://deepmind.google/technologies/gemini/.
[6] D. Truhn, J.N. Eckardt, D. Ferber, J.N. Kather, Large language models and multimodal foundation models for precision oncology, npj Precis. Oncol. 8 (1) (Dec. 2024) 1–4, https://doi.org/10.1038/S41698-024-00573-2.
[7] D. Ferber, et al., In-context learning enables multimodal large language models to classify cancer pathology images, Nat. Commun. 15 (1) (2024) 1–12, https://doi.org/10.1038/s41467-024-51465-9.
[8] H. Tan, et al., General generative AI-based image augmentation method for robust rooftop PV segmentation, Appl. Energy 368 (2024) 123554, https://doi.org/10.1016/J.APENERGY.2024.123554.
[9] S. Majumder, et al., Exploring the capabilities and limitations of large language models in the electric energy sector, Joule 8 (6) (2024) 1544–1549, https://doi.org/10.1016/j.joule.2024.05.009.

[10] S.K. Choudhary, A. Mondal, Impact of large language models (LLMs) and artificial intelligence (AI) on renewable and sustainable energy, in: Large Language Models for Sustainable Urban Development, Springer, 2025, pp. 3–43, https://doi.org/10.1007/978-3-031-86039-3_1.
[11] J.M. Malof, K. Bradbury, L.M. Collins, R.G. Newell, Automatic detection of solar photovoltaic arrays in high resolution aerial imagery, Appl. Energy 183 (Dec. 2016) 229–240, https://doi.org/10.1016/J.APENERGY.2016.08.191.
[12] V.E. Puranik, R. Kumar, R. Gupta, Progress in module level quantitative electroluminescence imaging of crystalline silicon PV module: a review, Sol. Energy 264 (Nov. 2023) 111994, https://doi.org/10.1016/J.SOLENER.2023.111994.
[13] J.A. Tsanakas, L. Ha, C. Buerhop, Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: a review of research and future challenges, Renew. Sustain. Energy Rev. 62 (2016) 695–709, https://doi.org/10.1016/J.RSER.2016.04.079.
[14] B. Doll, et al., Photoluminescence for defect detection on full-sized photovoltaic modules, IEEE J. Photovolt. 11 (6) (2021) 1419–1429, https://doi.org/10.1109/JPHOTOV.2021.3099739.
[15] H. Yousif, Z. Al-Milaji, Fault detection from PV images using hybrid deep learning model, Sol. Energy 267 (2024) 112207, https://doi.org/10.1016/J.SOLENER.2023.112207.
[16] B. Li, C. Delpha, D. Diallo, A. Migan-Dubois, Application of artificial neural networks to photovoltaic fault detection and diagnosis: a review, Renew. Sustain. Energy Rev. (2020), https://doi.org/10.1016/j.rser.2020.110512. p. 110512.
[17] V. Lopes, et al., Comprehensive overview of photovoltaic plants operations and maintenance using UAV-based IR imaging and advanced image analysis techniques, Renewable Energy Focus 57 (2026) 100825, https://doi.org/10.1016/J.REF.2026.100825.
[18] L. Hernández-Callejo, S. Gallardo-Saavedra, V. Alonso-Gómez, A review of photovoltaic systems: design, operation and maintenance, Sol. Energy 188 (2019) 426–440, https://doi.org/10.1016/J.SOLENER.2019.06.017.
[19] M. Waqar Akram, G. Li, Y. Jin, X. Chen, Failures of photovoltaic modules and their detection: a review, Appl. Energy 313 (2022), https://doi.org/10.1016/J.APENERGY.2022.118822.
[20] J.F. Gaviria, G. Narváez, C. Guillen, L.F. Giraldo, M. Bressan, Machine learning in photovoltaic systems: a review, Renew. Energy 196 (2022) 298–318, https://doi.org/10.1016/J.RENENE.2022.06.105.
[21] M.T. Araji, A. Waqas, R. Ali, Utilizing deep learning towards real-time snow cover detection and energy loss estimation for solar modules, Appl. Energy 375 (2024) 124201, https://doi.org/10.1016/J.APENERGY.2024.124201.
[22] O. Ozturk, B. Hangun, O. Eyecioglu, Detecting snow layer on solar panels using deep learning, in: 10th IEEE International Conference on Renewable Energy Research and Applications, ICRERA 2021, 2021, pp. 434–438, doi: 10.1109/ICRERA52334.2021.9598700.
[23] R. Cavieres, R. Barraza, D. Estay, J. Bilbao, P. Valdivia-Lefort, Automatic soiling and partial shading assessment on PV modules through RGB images analysis, Appl. Energy 306 (2022) 117964, https://doi.org/10.1016/J.APENERGY.2021.117964.
[24] M. Yang, J. Ji, B. Guo, Soiling quantification using an image-based method: effects of imaging conditions, IEEE J. Photovolt. 10 (6) (2020) 1780–1787, https://doi.org/10.1109/JPHOTOV.2020.3018257.
[25] S. Kshetrimayum, J.J.H. Liou, Y.P. Huang, A deep learning based detection of bird droppings and cleaning method for photovoltaic solar panels, Conf. Proc. IEEE Int. Conf. Syst. Man Cybern. (2023) 3398–3403, https://doi.org/10.1109/SMC53992.2023.10394560.
[26] A. Moradi Sizkouhi, M. Aghaei, S.M. Esmailifar, A deep convolutional encoder-decoder architecture for autonomous fault detection of PV plants using multi-copters, Solar Energy 223 (2021) 217–228, https://doi.org/10.1016/J.SOLENER.2021.05.029.
[27] S. Deitsch, et al., Automatic classification of defective photovoltaic module cells in electroluminescence images, Sol. Energy 185 (Jun. 2019) 455–468, https://doi.org/10.1016/J.SOLENER.2019.02.067.
[28] A.M. Karimi, et al., Automated pipeline for photovoltaic module electroluminescence image processing and degradation feature classification, IEEE J. Photovolt. 9 (5) (2019) 1324–1335, https://doi.org/10.1109/JPHOTOV.2019.2920732.
[29] W. Tang, Q. Yang, K. Xiong, W. Yan, Deep learning based automatic defect identification of photovoltaic module using electroluminescence images, Sol. Energy 201 (2020) 453–460, https://doi.org/10.1016/J.SOLENER.2020.03.049.
[30] J. Zhang, X. Chen, H. Wei, K. Zhang, A lightweight network for photovoltaic cell defect detection in electroluminescence images based on neural architecture search and knowledge distillation, Appl. Energy 355 (2024) 122184, https://doi.org/10.1016/J.APENERGY.2023.122184.
[31] X. Chen, T. Karin, A. Jain, Automated defect identification in electroluminescence images of solar modules, Sol. Energy 242 (2022) 20–29, https://doi.org/10.1016/J.SOLENER.2022.06.031.
[32] F.M.A. Mazen, R.A.A. Seoud, Y.O. Shaker, Deep learning for automatic defect detection in PV modules using electroluminescence images, IEEE Access 11 (2023) 57783–57795, https://doi.org/10.1109/ACCESS.2023.3284043.
[33] F. Hong, J. Song, H. Meng, W. Rui, F. Fang, Z. Guangming, A novel framework on intelligent detection for module defects of PV plant combining the visible and infrared images, Sol. Energy 236 (2022) 406–416, https://doi.org/10.1016/J.SOLENER.2022.03.018.
[34] A. Di Tommaso, A. Betti, G. Fontanelli, B. Michelozzi, A multi-stage model based on YOLOv3 for defect detection in PV panels based on IR and visible imaging by unmanned aerial vehicle, Renew. Energy 193 (2022) 941–962, https://doi.org/10.1016/J.RENENE.2022.04.046.
[35] R. Tang, Z. Ren, S. Ning, Y. Zhang, Fault classification of photovoltaic module infrared images based on transfer learning and interpretable convolutional neural network, Sol. Energy 276 (2024) 112703, https://doi.org/10.1016/J.SOLENER.2024.112703.
[36] N. Kellil, A. Aissat, A. Mellit, Fault diagnosis of photovoltaic modules using deep neural networks and infrared images under Algerian climatic conditions, Energy 263 (2023) 125902, https://doi.org/10.1016/J.ENERGY.2022.125902.
[37] X. Xie, G. Lai, M. You, J. Liang, B. Leng, Effective transfer learning of defect detection for photovoltaic module cells in electroluminescence images, Sol. Energy 250 (2023) 312–323, https://doi.org/10.1016/J.SOLENER.2022.10.055.
[38] L. Bommes, et al., Anomaly detection in IR images of PV modules using supervised contrastive learning, Prog. Photovolt. Res. Appl. 30 (6) (2022) 597–614, https://doi.org/10.1002/PIP.3518.
[39] R. Yang, et al., Weakly-semi supervised extraction of rooftop photovoltaics from high-resolution images based on segment anything model and class activation map, Appl. Energy 361 (2024) 122964, https://doi.org/10.1016/J.APENERGY.2024.122964.
[40] X. Liu, K. Barth, D. Windridge, K. Xu, Accelerating material discovery for CdTe solar cells using knowledge intense word embeddings, in: Conference Record of the IEEE Photovoltaic Specialists Conference, 2024, pp. 281–283, https://doi.org/10.1109/PVSC57443.2024.10749288.
[41] S. Shrestha, I. Bittencourt, A.S. Varde, P. Lal, AI-based modeling for textual data on solar policies in smart energy applications, in: 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA), 2024, pp. 1–8, 10.1109/IISA62523.2024.10786713.
[42] K. Nuortimo, J. Harkonen, K. Breznik, Global, regional, and local acceptance of solar power, Renew. Sustain. Energy Rev. 193 (2024) 114296, https://doi.org/10.1016/J.RSER.2024.114296.
[43] H. Lin and M. Yu, "PV-VLM: A Multimodal Vision-Language Approach Incorporating Sky Images for Intra-Hour Photovoltaic Power Forecasting," Apr. 2025, Accessed: Jul. 28, 2025. [Online]. Available: https://arxiv.org/pdf/2504.13624.
[44] H. Zhang, J. Yang, S. Fan, H. Geng, C. Shao, An ultra-short-term distributed photovoltaic power forecasting method based on GPT, IEEE Trans. Sustain. Energy (2025), https://doi.org/10.1109/TSTE.2025.3564975.
[45] Z. Yan, D. Zhenyuan, Y. Xu, and Z. Zhou, "Probabilistic PV Power Forecasting by a Multi-Modal Method using GPT-Agent to Interpret Weather Conditions," pp. 1–6, Sep. 2024, doi: 10.1109/ICIEA61579.2024.10665072.
[46] R. Deng *et al.*, "Cap-Pv: A Language-Aided Few-Shot Learning Framework for Enhancing Pv Panel Segmentation Generalizability Via Clip," 2025, doi: 10.2139/SSRN.5253569.
[47] Y. Boutana, A. Soukkou, S. Haddad, Enhancing fault detection in solar photovoltaic systems using multi-modal models. 2024 International Conference on Advances in Electrical and Communication Technologies, ICAECOT 2024, 2024 doi: 10.1109/ICAECOT62402.2024.10828814.
[48] M. Guo, Y. Weng, Solar photovoltaic assessment with large language model, Appl. Energy 402 (2025) 126835, https://doi.org/10.1016/J.APENERGY.2025.126835.
[49] D. Ferber, et al., In-context learning enables multimodal large language models to classify cancer pathology images, Nat. Commun. 15 (1) (2024) 10104, https://doi.org/10.1038/s41467-024-51465-9.
[50] X. Chen, M. Zheng, B. Liu, Y. Zhang, C. Chen, P. Zhang, Large multimodal model for zero-shot scene classification of high spatial resolution remote sensing images, Remote Sens. Lett. 16 (4) (2025) 449–459, https://doi.org/10.1080/2150704X.2025.2470879.
[51] M. Zhao, T. Fu, H. Yu, K. Niu, and B. Li, "IADGPT: Unified LVLM for Few-Shot Industrial Anomaly Detection, Localization, and Reasoning via In-Context Learning," Aug. 2025, Accessed: Apr. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2508.10681.
[52] DuraMat, "DuraMat Data Hub." Accessed: Jan. 21, 2025. [Online]. Available: https://datahub.duramat.org/.
[53] Afroz, "Solar Panel Images Clean and Faulty Images." Accessed: Jan. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/pythonafroz/solar-panel-images.
[54] Y. ZHANG, "Dataset of solar cells defect segmentation." Accessed: Jan. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/yaozhang01182010/dataset-of-solar-cells-defect-segmentation.
[55] E. L. C. H. F.-M. É. R.-G. A.-D. N. S. Alfaro Mejia, "Photovoltaic System Thermal Images." Accessed: Jan. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/photo-voltaic-cell-images.
[56] UITM, "Hotspot Computer Vision Project." Accessed: Jan. 21, 2025. [Online]. Available: https://universe.roboflow.com/uitm-eyqpa/hotspot-ahbdm/dataset/1.
[57] M.S. Roulston, Performance targets and the Brier score, Meteorol. Appl. 14 (2) (2007) 185–194, https://doi.org/10.1002/MET.21.
[58] IEA, "IEA PVPS Report TASK -13: Review of Failures of Photovoltaic Modules Final," May 28, 2014, *IEA-PVPS*. Accessed: Apr. 25, 2026. [Online]. Available: https://iea-pvps.org/key-topics/review-of-failures-of-photovoltaic-modules-final/.
[59] OPENAI, "GPT-5." Accessed: Oct. 08, 2025. [Online]. Available: https://openai.com/gpt-5/.
[60] Google AI., "Gemini." Accessed: Jan. 15, 2025. [Online]. Available: https://ai.google/gemini/.
[61] Anthropic, "Introducing Claude 4." Accessed: Jul. 27, 2025. [Online]. Available: https://www.anthropic.com/news/claude-4.
[62] Qwen Team, "Qwen3.5: Towards Native Multimodal Agents," Feb. 2026. [Online]. Available: https://qwen.ai/blog?id=qwen3.5.
[63] OPENAI, "CLIP ViT-L/14 model." Accessed: Oct. 06, 2025. [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14.

[64] I. Groq, "Groq LPU Inference Engine." Accessed: Apr. 26, 2026. [Online]. Available: https://groq.com.
[65] Together AI, "Together Inference Platform." Accessed: Apr. 26, 2026. [Online]. Available: https://www.together.ai.
[66] NVDIA, "SegFormer." Accessed: Oct. 08, 2025. [Online]. Available: https://huggingface.co/nvidia/segformer-b1-finetuned-cityscapes-1024-1024.
[67] BRIAAI, "RMBG." Accessed: Oct. 08, 2025. [Online]. Available: https://huggingface.co/briaai/RMBG-1.4.