International Journal of Innovative Computing 16(1–2) 161–169

INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Solar GenAI: An Offline LLM-Based Generative AI Agent for Solar Photovoltaics in Malaysia

Shafidah Shafian*
Solar Energy Research Institute (SERI)
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia
Email: norshafidah@ukm.edu.my

Submitted: 15/9/2025. Revised edition: 28/11/2025. Accepted: 5/1/2026. Published online: 28/7/2026
DOI: https://doi.org/10.11113/ijic.v16n1-2.689

***Abstract*—The adoption of solar photovoltaic (PV) systems in Malaysia is often limited by fragmented information and restricted access to reliable, localized knowledge. To address this challenge, Solar GenAI is introduced as an offline-capable, generative AI agent built on the Mistral-7B-Instruct model and deployed entirely on local hardware using the Ollama platform. Designed to function without internet connectivity, Solar GenAI ensures data privacy, institutional autonomy, and broad accessibility, particularly in bandwidth-limited or secure environments. The system integrates a Retrieval-Augmented Generation (RAG) pipeline grounded in a curated corpus of Malaysia-specific solar policies and technical documents. This architecture eliminates hallucinations and achieves 100% factual accuracy across representative queries, compared to a 60% hallucination rate in an LLM-only baseline. A caching mechanism further enhances responsiveness, reducing average query latency by 35%. The user interface was developed based on Nielsen's usability principles to support intuitive interaction across diverse user groups. By combining fully offline LLM deployment, grounded retrieval, and performance optimization, Solar GenAI provides a secure, accessible, and scalable platform to support solar PV awareness and education, institutional applications, and policy implementation in Malaysia.**

***Keywords*—Generative AI, Large Language Model, Local LLM, Solar Photovoltaic**

I. INTRODUCTION

Malaysia has embarked on a progressive renewable energy strategy as part of its national development agenda, aiming to diversify its energy mix, reduce carbon emissions, and ensure long-term energy security. Under the Malaysia Renewable Energy Roadmap (MyRER), the government targets achieving 31% (12.9 GW) of renewable energy capacity in the national electricity mix by 2025 and 40% (18.0 GW) by 2035 [1]. Solar energy, owing to Malaysia's favorable equatorial location, stands as the cornerstone of this transition. The country receives an average solar irradiance of 4.21- 5.56 kWh/m2/day, positioning it among the most solar-rich nations in Southeast Asia [2].

Despite initiatives such as Net Energy Metering (NEM) and Self-Consumption (SELCO) programs aimed at encouraging adoption, only a small percentage of Malaysian households and businesses have installed solar panels. This finding is supported by a study from Teoh, A. N (2020) with the result from 225 participant in Malaysia showed that the public was aware of the option of solar energy but was not ready to install solar photovoltaic (PV) [3]. This low uptake is influenced not only by cost but also by a pervasive knowledge gap and limited access to clear, reliable information. Essential resources ranging from government documents and technical manuals to supplier brochures and academic literature are fragmented and often too complex for the general public. Furthermore, the lack of a centralized, user-friendly platform makes it difficult to identify trustworthy data. This information overload and inconsistency hinder Malaysians' ability to make informed decisions about solar adoption, system design, or compliance with national energy policies.

To address these gaps, advancements in artificial intelligence (AI), particularly large language models (LLMs), offer a new pathway [4]. LLMs have demonstrated strong performance in natural language understanding and knowledge generation across specialized fields. Their capacity to simplify complex content and deliver interactive, tailored responses holds transformative potential for public engagement with solar technologies [5]. In the context of renewable energy, LLMs are increasingly explored to optimize energy consumption, develop sustainable solutions, promote environmental awareness, and simplify complex policy and regulatory content. These features position LLMs as powerful tools to bridge the solar knowledge gap in Malaysia by translating dense, fragmented content into accessible, conversational formats that can benefit a wide range of users across educational, professional, and community backgrounds.

However, mainstream LLMs typically depend on cloud-based infrastructure, which requires continuous internet connectivity for real-time interaction [6, 7]. In Malaysia, the need for an offline solution extends beyond rural connectivity challenges. Many schools, universities, government departments, and corporate organizations operate within secure networks that restrict access to external cloud services. This makes it difficult to integrate online AI tools into their workflows. Moreover, issues related to data privacy and national data sovereignty are increasingly important, as cloud platforms often transmit user information across international servers. This raises concerns when handling topics such as energy consumption, system planning, or financial investments in solar technology. The use of commercial cloud services may also involve recurring costs or create dependency on proprietary platforms, which can be impractical for public institutions or smaller organizations [8, 9]. These factors emphasize the need for an offline, locally deployed AI system that can provide verified solar information through a secure, centralized, and easy-to-use platform, suitable for users from various educational, professional, and community backgrounds.

To address these limitations, Solar GenAI was developed as an offline generative AI platform purpose-built to serve the Malaysian solar PV landscape. Solar GenAI, built on the Mistral-7B-Instruct model and deployed using the Ollama framework, operates fully on local hardware, independent of external APIs or internet connectivity. This design ensures data privacy, institutional autonomy, and broad accessibility even in bandwidth-limited or secure environments. The system is enhanced with a Retrieval-Augmented Generation (RAG) pipeline that dynamically grounds responses in a curated corpus of Malaysian solar content, including policies, incentives, and technical documents. Benchmark tests show that the RAG framework reduced hallucination rates from 60% to 0%, achieving 100% factual accuracy across representative solar-related queries. In addition, to improve system responsiveness and user experience, Solar GenAI integrates a caching mechanism that minimizes redundant computation. Performance tests demonstrated a 35% reduction in average query latency from approximately 12.0 s without caching to 7.8 s with caching. The platform also features a user-friendly graphical interface designed according to established usability principles, allowing individuals from diverse backgrounds to engage with the system effectively. By combining offline LLM deployment, context-aware retrieval, and performance-optimized architecture, Solar GenAI provides a secure, accurate, and accessible solution to bridge Malaysia's solar knowledge divide empowering public, academic, and institutional stakeholders to make informed decisions and accelerate clean energy adoption.

In summary, this paper presents Solar GenAI as the first offline generative AI system specifically designed for Malaysia's PV ecosystem. It addresses a critical research gap by providing a localized, retrieval-augmented, and privacy-preserving AI tool that overcomes the dependence on cloud-based LLMs. Unlike existing global AI platforms, Solar GenAI integrates a Malaysian solar knowledge base to deliver accurate, context-aware responses on national policies, incentives, and technical guidelines. This work contributes both a technical framework combining offline LLM deployment, RAG, and caching optimization and a practical demonstration of how AI can support education, research, and policy in advancing solar energy adoption.

II. METHODOLOGY

This study employed a structured methodological framework to design, develop, and evaluate Solar GenAI, an offline generative AI platform tailored to the Malaysian PV landscape. The process was conducted in four sequential phases: (A) system design and model configuration, (B) dataset development, (C) deployment and optimization, and (D) performance evaluation.

A. System Design and Model Configuration

The initial phase involved conceptualizing the overall architecture of Solar GenAI. The system was designed around the Mistral-7B-Instruct model, deployed locally via the Ollama framework to ensure full offline capability and data privacy. This stage established a baseline LLM-only configuration, allowing end-to-end interaction without reliance on external servers. However, limitations in factual grounding and localization prompted a transition to a RAG framework. The enhanced system architecture was structured to combine offline LLM reasoning with on-demand document retrieval, enabling context-aware and verifiable outputs aligned with Malaysia's solar policy environment.

The full architecture highlighting both the LLM-only (orange region) and RAG-enhanced (entire figure) workflows, is illustrated in Fig. 1. This diagram visualizes the transformation from a monolithic, prompt-based LLM to an intelligent, context-aware agent capable of high-fidelity, grounded question answering.

B. Dataset Development

A specialized knowledge base was developed to support RAG operations. The dataset consisted of open-access Malaysian policy documents, including the Malaysia Renewable Energy Roadmap (MyRER 2021–2035), the National Energy Policy (NEP 2022–2040), and the National Energy Transition Roadmap (NETR 2023) [1, 10]. It also comprised documentation on major government incentive programs like Net Energy Metering (NEM), the Net Offset Virtual Aggregation (NOVA) scheme, and the Self-Consumption (SELCO) initiative. Regulatory texts, including planning materials related to the LargeScale Solar (LSS) program, were incorporated to ensure policy completeness.

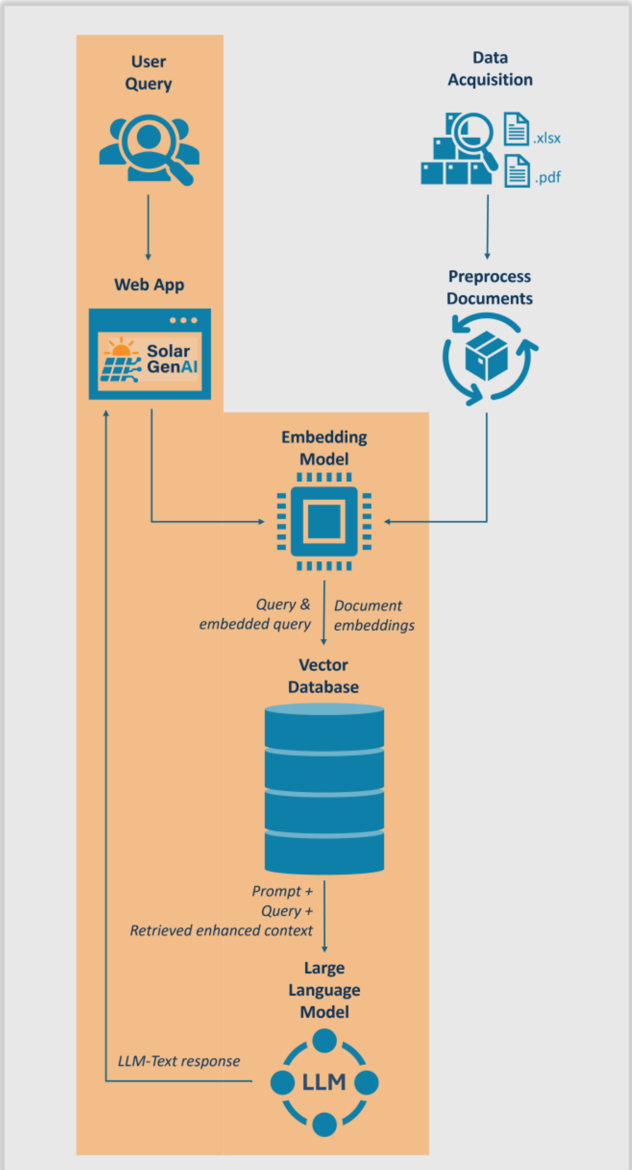

Fig. 1. System architecture of Solar GenAI showing the LLM-only configuration (orange region) to a RAG-enhanced framework (entire region).

In addition to government-issued content, selected scientific publications on renewable energy and solar PV systems were included to provide technical depth and evidence-based insights [11-24]. All documents used in this compilation were obtained in PDF format from publicly accessible Malaysian policy papers, institutional reports, and educational publications under open-access or public-domain licenses. Each document was processed using LangChain's unstructured document loader to extract clean, machine-readable text. To ensure ethical compliance, all texts were deliberately rephrased to avoid direct overlaps with copyrighted materials while maintaining factual accuracy and policy relevance. The resulting repository forms a high-quality, policy-aligned knowledge base optimized for embedding and retrieval.

After ingestion, documents were divided into coherent text segments of approximately 1,000 characters with slight overlaps to preserve context. Each segment was then transformed into a 384-dimensional embedding vector using the all-MiniLM-L6-v2 model and stored in a persistent vector database for cosine-similarity search. During inference, user queries are embedded within the same semantic space, matched to relevant segments, and combined with the prompt for grounded, traceable, and factually reliable responses.

C. Offline Deployment and Optimization

Following model integration, the system was deployed fully offline to evaluate its operational viability on consumer-grade hardware (NVIDIA RTX 4060 GPU, 8 GB VRAM; 32 GB RAM). A 4-bit quantized model version was employed to balance memory efficiency with response quality. Optimization included the implementation of a caching mechanism that stored vector embeddings and prior query responses, thereby reducing redundant computation. This technique, inspired by the principles of Least Recently Used (LRU) cache management, achieved a measurable reduction in latency while maintaining inference stability and accuracy.

D. Performance Evaluation

The final phase focused on assessing performance, reliability, and compliance. Quantitative evaluation compared the LLM-only and RAG-enhanced systems in terms of hallucination rate and response latency, forming the basis of the results presented in Section 3.

III. RESULTS AND DISCUSSION

A. Retrieval-Augmented Generation Hallucination and Accuracy Performance

LLMs have demonstrated impressive ability to generate text, but they are limited by the static knowledge encoded in their training data. As a result, they can produce confident-sounding answers that may be outdated or factually incorrect, a phenomenon often referred to as hallucination, where the model outputs information not grounded in reality [25]. RAG is an AI framework designed to address these limitations by grounding the LLM's responses in external knowledge sources. In essence, RAG allows an LLM to retrieve relevant information from a knowledge repository at query time, then use that information to generate more accurate and reliable answers. This approach ensures the model has access to up-to-date, domain-specific facts and even enables citing sources for transparency, thereby improving user trust [26]. In practical usage scenario, the evaluation showed that the LLM-only system often generated generic or outdated answers and even invented details. For example, when asked “Please describe yourself,” the baseline LLM hallucinated a false identity statement (“I am a Large Language Model trained by Mistra…”), whereas the RAG system answered in context (“I am Solar GenAI, an advanced and highly knowledgeable Solar PV Agent developed by Dr. Norshafidah…”) without misidentifying itself. Because the LLM had no domain data, it could not supply Malaysia-specific solar facts on demand (e.g. local climate, policies) and sometimes hallucinated environmental details.

Both setups were quantitatively evaluated on a representative set of solar-PV queries from basic definitions to Malaysia-specific questions and logged whether each answer contained hallucinations. Fig. 2(a) compares the average response time of the baseline LLM and RAG-enhanced systems. The RAG model demonstrates faster response time (6.64 s) as shown in blue bar compared to the LLM-only with 9.15 s shown in red. Although retrieval adds an extra step, the use of a compact vector store and prompt-only queries to the LLM kept latencies low. Overall, the RAG approach delivered a marked improvement in answer reliability with no penalty in responsiveness. Fig. 2(b) shows the LLM answers were marked as hallucinated in 60% of cases, meaning they contained incorrect or made-up information. In contrast, the RAG-enhanced Solar GenAI produced no hallucinations on any of the test queries and was factually correct in all cases. This result highlights the effectiveness of RAG in grounding responses using external context. Fig. 2(c) depicts the frequency of hallucinated (purple) and non-hallucinated (turquoise) responses generated by each model. Non-hallucinated or correct answer rate rose from 40% with LLM-only to 100% with RAG. These differences reflect the grounding effect of retrieval by supplying relevant context, making Solar GenAI able to answer niche questions correctly. The confusion matrix in Fig. 2(d) compares manually (true) labelled hallucination flags with automated (predicted) labelled based on predefined keywords. All 9 manually labelled hallucinated responses were missed by the keyword-based auto detection (false negatives), and none were correctly flagged (true positives = 0). Conversely, all 21 non-hallucinated responses were accurately identified as non-hallucinated (true negatives). This indicates the keyword-based auto-flagger has a 100% specificity but 0% sensitivity, failing to detect actual hallucinations.

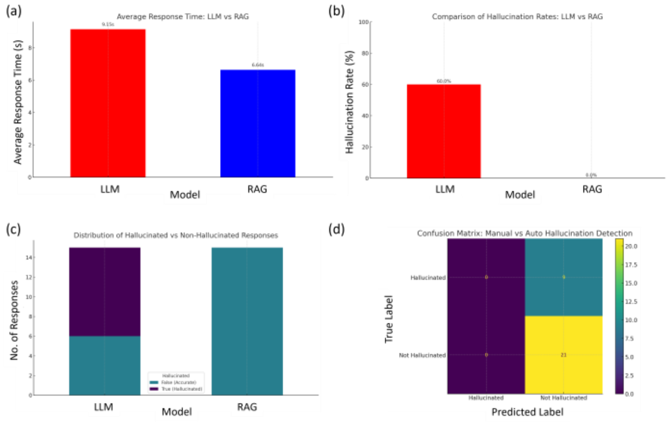

Fig. 2. (a) Average response time and (b) comparison of hallucination rates between LLM and RAG. (c) Distribution of responses marked as hallucinated. (d) Confusion matrix of manual versus automated hallucination detection.

A qualitative comparison of the actual outputs is presented in Table I below, which highlights three examples from the query evaluation. The LLM-only system hallucinated several critical facts. For instance, when asked “What is NEM?”, the LLM incorrectly identified it as a cryptocurrency called “New Economy Movement,” while the RAG-based system accurately explained that NEM refers to Net Energy Metering, a solar initiative introduced by the Malaysian government. Similarly, when comparing NEM with GoMEa and NOVA, the LLM mistakenly referred to GoMEa as a global meteorological system. In contrast, RAG correctly described both GoMEa and NOVA as solar-related initiatives aimed at reducing energy consumption in Malaysia. Lastly, in response to the query “What is the implication of NETR into Malaysia?”, the LLM misrepresented NETR as a vague principle about transparency, while the RAG model provided a detailed and locally accurate explanation of the National Energy Transition Roadmap, outlining its policy impact and significance.

Beyond retrieval, the system employs careful prompt design to further suppress hallucinations. The system prompt is structured to define Solar GenAI’s role and constraints; for instance, it explicitly states that the agent has no personal identity or opinions beyond the solar PV context. The system prepeds each answer with phrases like “Based on the context provided…” to signal reliance on retrieved facts. Crucially, the prompt instructs the model to refuse or hedge if a question goes beyond the available information. For example, if a query is outside the corpus scope, the model is guided to respond, “I do not have enough information to answer that,” rather than guessing. These hallucination-detection clauses (explicit refusal patterns) have been shown in prior work to improve factuality. In Solar GenAI, they prevent identity errors (e.g. masquerading as the model) and ensure factual consistency. In practice, this disciplined prompting meant that even ambiguous questions produced safe outputs grounded in context, rather than spurious or “imagined” content.

TABLE I. THE COMPARISON OF RESPONSE TO THE QUERY BASED ON LLM AND RAG.

Query	LLM	RAG
What is NEM?	NEM (New Economy Movement) is a peer-to-peer cryptocurrency and blockchain platform. It was originally called "Mijin" when it was launched in Japan in March 2015….	NEM, according to the provided context, stands for Net Energy Metering, which is a solar photovoltaic (PV) program….
Difference between NEM GoMEn and NOVA?	NEM (New Economy Movement) and GOMEn (Global Observing System for Meteorology, Oceanography, and Climate) are two completely different concepts….	The NEM GoMEn (Government Building Energy Management Programme) and the NOVA Programme serve similar purposes, both aiming to reduce energy expenses through solar PV systems….
What the implication NETR into Malaysia?	NETR, which stands for "Neutrality, Equality, Transparency, and Responsiveness," is a concept or set of principles rather than an acronym specific to any particular context. In the context of Malaysia, if we interpret NETR as a principle for governance or business….	The National Energy Transition Roadmap (NETR) has significant implications for Malaysia. Firstly, it serves as a blueprint for transitioning towards a net-zero future, emphasizing not just technological innovation but also a shift in mindset, strategies, and policies….

B. Cache Performance and Evaluation

Caching is a well-established method for enhancing computational efficiency by temporarily storing the results of expensive operations for reuse. In computer systems, caches exploit the principle of locality of reference, where recently or frequently accessed data are more likely to be needed again, thus reducing access time. The same concept applies to AI applications, where repeated computations or retrievals can be stored and reused to minimize redundant processing. Typical cache management follows algorithms such as Least Recently Used (LRU), time-to-live expiration, or response memorization [27, 28].

In Solar GenAI, caching is integrated at multiple layers to maintain responsiveness in offline and resource-limited environments. Upon initialization, the system loads the language model and vector database into memory and retains them across queries to avoid reloading overhead. The application also caches final responses for identical queries, enabling immediate output delivery without regenerating the result. Memory efficiency is managed using an LRU strategy to discard the least-accessed entries when capacity limits are reached.

To evaluate the impact of caching, a benchmark was conducted using identical queries under two conditions: with caching disabled and enabled. Each test measured (1) initialization time and (2) LLM response generation time. Timing was logged using high-resolution clocks and statistical analysis was performed using a paired t-test. As shown in Fig. 3(a), enabling the cache produced a substantial improvement in overall performance. Without caching, an average query in the test took around 12.0 s in total, of which roughly 4 seconds were spent on initialization tasks (orange portion) and ~8 seconds on the LLM's response generation (blue portion). With caching turned on, the initialization time per query dropped to effectively zero. The expensive setup steps are done once and reused leaving almost the entire query time attributable to LLM inference. The average total query time with caching was ≈7.8 s, essentially equal to the model's generation time since the cache eliminated the ~4 s overhead. This amounts to about a 35% reduction in latency on average as shown in the latency reduction in Eq. 1.

$$\text{Latency Reduction} = ((T_{\text{no\_cache}} - T_{\text{cache}})/T_{\text{no\_cahce}}) \times 100 \quad (1)$$

where Tno_cache is average response time without caching and Tcache is average response time with catching. Notably, the blue segments (LLM time) are nearly the same height in both scenarios, indicating that the intrinsic LLM computation time did not change (the LLM still needs to generate the answer token by token). It is the elimination of repeated initialization (orange segment) that yields the performance gain. These results confirm that caching significantly speeds up the application by removing redundant work. The improvement is not only practically large for user experience but also statistically significant – a paired t-test on the repeated queries showed the cached configuration was faster with $p \ll 0.01$ (on the order of 10^−5), rejecting the null hypothesis that caching makes no difference. In other words, there is a less than 0.001% probability that the observed ~4 s average improvement was due to random chance. Fig. 3(b) provides a query-level view of the performance differences. Each numbered position on the x-axis corresponds to one of the common queries (ordered from the slowest to fastest in the no-cache run). The red curve (no cache) is consistently above the green curve (cache enabled) for every query, underscoring that caching yielded faster responses across the board. The largest gap is observed at the leftmost point, which corresponds to an open-ended "self-introduction" query to the AI agent. In the no-cache run this query took about 22.5 s total, whereas with caching it completed in about 12.4 s, nearly a 10-s reduction. This dramatic speed-up was due to avoiding heavy initialization: without cache, the system likely had to load context or handle the query from scratch, while the cached run benefited from the model and data already being in memory. Most other queries show more modest yet significant gains of a few seconds each. For instance, a policy-related question that took ~12 s without cache was handled in ~8 s with cache, and a technical question that took ~15 s dropped to ~12 s with caching. In all cases, the cached execution either matched or outperformed the uncached one. There were no instances where caching made performance worse. It is noteworthy that the green line (cached) has small fluctuations, for example query 3 versus query 5, which reflect differences in LLM response time for different questions (some answers are longer or more complex). Crucially, those variations are minor compared to the consistent ~4 s savings of skipping re-initialization. This uniform benefit across diverse queries demonstrates the effectiveness of the caching strategy: regardless of query content, users experience faster answers mainly because the system isn't repeating setup work each time.

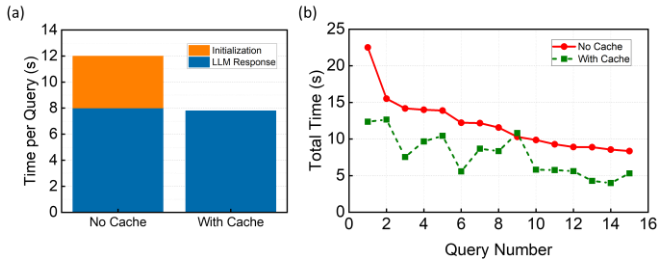

Fig. 3. (a) Average response time and (b) comparison of hallucination rates between LLM and RAG. (c) Distribution of responses marked as hallucinated. (d) Confusion matrix of manual versus automated hallucination detection.

C. Graphical User Interface

The Solar GenAI platform is equipped with a web-based graphical user interface (GUI) specifically designed to maximize accessibility, clarity, and ease of use across a wide spectrum of users in Malaysia. The interface design was guided by Nielsen's ten usability heuristics, which provide a robust framework for assessing system-user interaction and promoting intuitive design [29, 30]. The ten heuristics include visibility of system status, match between system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, help users recognize, diagnose, and recover from errors and lastly help and documentation.

As shown in Fig. 4, the main interface adopts a clean, minimalist layout composed of a header, a sidebar containing navigation and session controls, and a central content display panel. This structure ensures high visibility of system status, one of the key usability heuristics. During interaction, Solar GenAI provides continuous real-time feedback to keep users informed about ongoing processes.

Fig. 4. Main GUI of Solar GenAI deployed locally. The interface includes a header, sidebar with navigation and controls (left), and a central content display (middle). Key interface elements include intuitive labels, minimalist design, and real-time feedback, aligning with Nielsen's usability heuristics to support diverse users.

For instance, when a query is being processed, the application triggers a loading animation accompanied by a message such as "Running database()" or "Thinking…", indicating that the AI is retrieving relevant knowledge or generating a response. This feature is visualized in Fig. 5(a), which shows the actual system feedback during backend operations. Such status indicators assure users that their input has been received and is actively being handled, which is crucial for building trust, especially among users unfamiliar with AI behavior or those operating in environments with limited bandwidth.

The language used throughout the interface is deliberately aligned with domain-specific terminology relevant to Malaysia's solar PV ecosystem. As illustrated in Fig. 5(b), prompts such as "Type your question…" and clearly labeled sections like "About" and "Solar GenAI" reflect real-world phrasing familiar to users from educational, policy, or technical backgrounds. This alignment between system language and user expectations reduces the cognitive effort required to interact with the platform.

User control and freedom are supported via interface elements like the "Clear Chat" button, shown in Fig. 5(c), which allows users to reset conversations without restarting the application. This is particularly useful for users who are learning by exploration or who wish to reformulate their queries. Navigation between different sections such as moving from the About page to the main chat interface is seamless and non-disruptive, encouraging fluid user interaction.

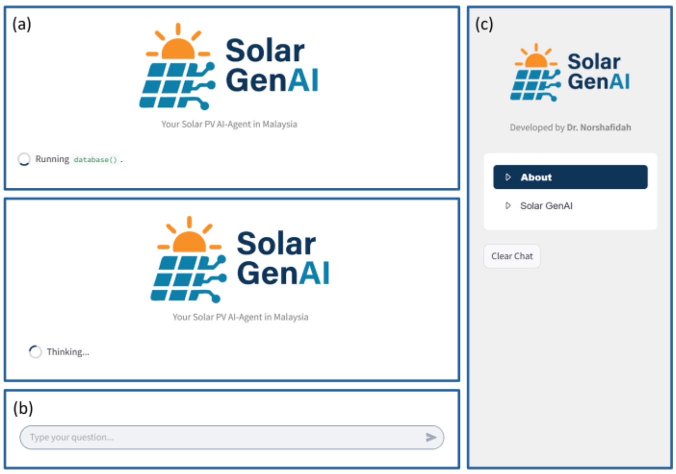

Fig. 5. Snapshots of the Solar GenAI GUI demonstrate usability features based on Nielsen's heuristics. (a) a real-time status indicator ("Running database ()") informs users that the system is setting up during initialization, (b) text input box ("Type your question…") guides user interaction, and (c) the sidebar enables easy navigation with labeled sections ("About", "Solar GenAI") and a "Clear Chat" button to reset the session.

Consistency in visual design is another key strength. As depicted in Fig. 6(a), the application uses uniform font choices, visual hierarchy, and color coding. User-generated messages are shown in dark blue chat bubbles aligned to the right, while AI-generated responses are displayed in light gray bubbles aligned to the left. This convention mirrors familiar chat applications and helps users quickly identify the source of each message. The layout remains consistent across sessions, promoting predictability and reducing the learning curve.

To prevent errors, the system provides input guidance via placeholder text and examples, steering users toward relevant queries. On the backend, Solar GenAI is programmed to handle unanswerable or out-of-scope questions by returning a polite refusal such as "I do not have enough information to answer that." This behavior minimizes hallucinations and misinterpretation, enhancing the reliability of responses. As highlighted in Fig. 6(b), the "About" section clearly outlines system capabilities and limitations, which helps users frame appropriate questions and avoid invalid inputs.

The design emphasizes recognition rather than recall. Interactive elements remain visible throughout the session, and previous messages are retained in a scrollable chat history, as seen in Fig. 6(c). This helps users reference earlier responses without needing to remember prior system outputs. All functional components are clearly labeled, further reducing the need for memorization and promoting accessibility for novice users.

Efficiency of use is maintained through streamlined workflows and caching. The interface supports both casual inquiries and technical queries from experienced users. System performance improves over time due to caching mechanisms that store vector embeddings and model states, resulting in faster interactions for repeated or similar queries. The aesthetic approach prioritizes minimalism and clarity. As shown earlier in Fig. 4, the interface avoids unnecessary elements or distractions. A restrained color scheme, clean typography, and spacious layout help users focus on the content of solar PV knowledge rather than on navigating the tool itself.

When errors do occur, users receive helpful and clear messages, such as "Error: Please try again." These messages are designed to be friendly and instructive, aiding users in diagnosing and correcting their inputs. This level of feedback supports a more forgiving and confidence-building user experience.

Lastly, the interface includes inline documentation to support independent use. The "About" section provides essential guidance on the tool's purpose, design, and how to use it effectively. Presented in a concise bullet-point format, this embedded help replaces the need for external manuals, making the system self-contained and easier to adopt.

Together, these usability-focused design decisions ensure that Solar GenAI is not only functionally robust but also highly approachable for a diverse Malaysian user base. From researchers and students to policy professionals and laypersons, the interface empowers all users to interact with the AI agent confidently, efficiently, and with a high degree of trust.

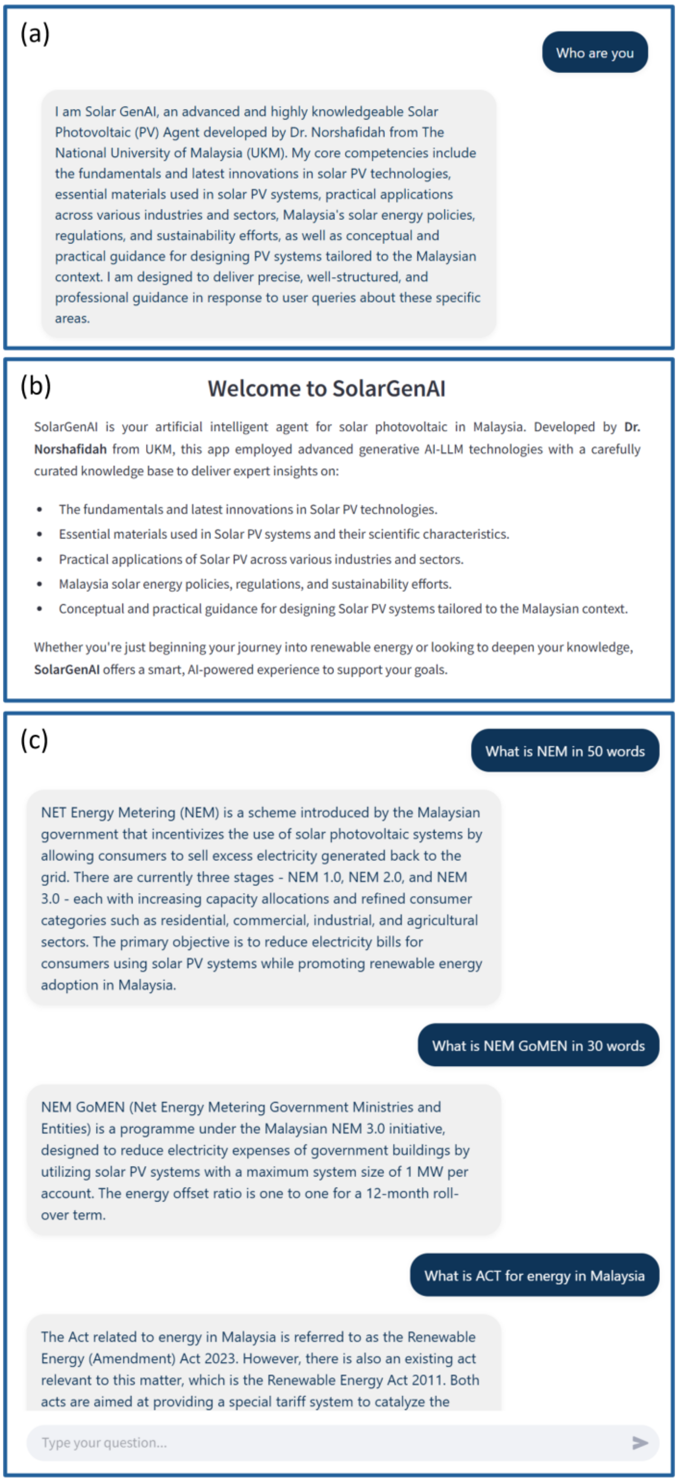

Fig. 6. Screenshots of the Solar GenAI interface showing key usability features. (a) Consistent design with clear fonts, layout structure, and chat bubble colors for user and AI messages. (b) The 'About' section gives a quick overview of Solar GenAI's purpose and features. (c) A scrollable chat history helps users easily review past responses without needing to remember them

IV. LIMITATIONS AND FUTURE WORKS

While Solar GenAI effectively addresses several key challenges in solar PV knowledge dissemination, it still has notable limitations. Currently, the system supports only English, which may hinder accessibility for user groups more comfortable with Malaysia's national and primary language, Bahasa Malaysia. To overcome this limitation, future development will involve fine-tuning the language model using Bahasa Malaysia corpora. This multilingual enhancement will improve contextual understanding, inclusivity, and outreach across diverse Malaysian user communities.

Secondly, the existing domain-specific retrieval dataset is static and does not reflect ongoing developments. The knowledge base must be manually updated to reflect new policy changes, incentives, or technical developments. In future versions, automated data updates will be introduced through scheduled batch ingestion or AI-assisted retrieval from verified energy databases, ensuring that Solar GenAI remains current and aligned with Malaysia's evolving solar PV landscape.

Another limitation lies in the model's lack of multimodal capability. While Solar GenAI performs well on text-based queries, it cannot yet process visual inputs such as solar panel layouts or technical schematics. Future iterations will explore integration of vision-language models (VLMs), such as CLIP or BLIP-2, to enable image-based diagnostics, solar component identification, and map-driven solar site planning. This multimodal enhancement will allow the system to analyze both textual and visual data for more comprehensive support in solar PV design and education.

Finally, while the caching mechanism effectively reduces latency, the system does not yet include per-user session memory or personalization. Adding long-term memory or preference-adaptive features while preserving offline operation could further enhance user engagement and relevance of responses, especially for repeat users. These upgrades will be designed to maintain user privacy and efficient local operation without dependence on external servers.

V. CONCLUSION

In conclusion, Solar GenAI demonstrates the practical viability of an offline, domain-specific generative AI system designed for Malaysia's PV landscape. By integrating a localized LLM, RA), optimized caching, and a user-centric interface, the system achieved substantial performance improvements where it reduces hallucination rates from 60 % to 0 % and enhancing response speed by 40 %. The interface, developed based on Nielsen's usability principles, effectively bridges technical capability with educational usability, fostering comprehension and engagement among diverse users. These outcomes confirm that localized AI frameworks can deliver accurate, efficient, and inclusive performance even in fully offline environments. Beyond its technical achievements, Solar GenAI serves as a replicable model for the deployment of generative AI in renewable-energy education and policy,
aligning with Malaysia's national vision for a sustainable energy transition and demonstrating how offline AI systems can deliver accurate and accessible solar PV knowledge to support broader clean-energy literacy and adoption.

ACKNOWLEDGMENT

The author would like to acknowledge the support of Geran Translasi UKM (TR-UKM), grant number UKM-TR2024-09, funded by Universiti Kebangsaan Malaysia (UKM), Malaysia.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

REFERENCE

[1] Sustainable Energy Development Authority Malaysia (SEDA). (2021). *Malaysia renewable energy roadmap: Pathway towards a low carbon energy system*. https://www.seda.gov.my/reportal/myrer/

[2] World Bank Group, Energy Sector Management Assistance Program (ESMAP), & Solargis. (2019). *Photovoltaic power potential: Malaysia*.

[3] Teoh, A. N., Go, Y. I., & Yap, T. C. (2020). Is Malaysia ready for sustainable energy? Exploring the attitudes toward solar energy and energy behaviors in Malaysia. *World, 1*(1), 90–103.

[4] Varshney, T. (2023). *Introduction to LLM agents*. NVIDIA Technical Blog. http://developer.nvidia.com/blog/introduction-to-llm-agents

[5] Aydin, O., Karaarslan, E., Erenay, F. S., & Bacanin, N. (2025). *Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma* (arXiv Preprint arXiv:2503.04765).

[6] Zhang, L., & Chen, Z. (2025). Opportunities of applying large language models in the building energy sector. *Renewable and Sustainable Energy Reviews, 214*, 115558.

[7] Liang, M., Hu, Y., Weng, H., Xi, J., & Yin, B. (2025). EnergyGPT: Fine-tuning large language model for multi-energy load forecasting. *Renewable Energy*, 123313.

[8] Cho, K., & Bahn, H. (2020). A cost estimation model for cloud services and applying to PC laboratory platforms. *Processes, 8*(1), 76.

[9] Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., Tang, L., Kang, Y., & Mars, J. (2024). Scaling down to scale up: A cost-benefit analysis of replacing OpenAI's LLM with open-source SLMs in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 280–291).

[10] Government of Malaysia. (2022). *National Energy Policy (DTN) 2022–2040*. https://myenergystats.st.gov.my/new-energy-tech1

[11] Mekhilef, S., Safari, A., Mustaffa, W., Saidur, R., Omar, R., & Younis, M. (2012). Solar energy in Malaysia: Current

state and prospects. *Renewable and Sustainable Energy Reviews, 16*(1), 386–396.

[12] Chua, S. C., & Oh, T. H. (2012). Solar energy outlook in Malaysia. *Renewable and Sustainable Energy Reviews, 16*(1), 564–574.

[13] Wahid, S. S. A., Ramli, M. S., Noorden, Z. A., Hassan, K. K., & Azli, S. A. (2017). A review on highlights and feasibility studies on solar energy utilization in Malaysia. In *AIP Conference Proceedings* (Vol. 1875, Article 030014).

[14] Vaka, M., Walvekar, R., Rasheed, A. K., & Khalid, M. (2020). A review on Malaysia's solar energy pathway towards carbon-neutral Malaysia beyond the COVID-19 pandemic. *Journal of Cleaner Production, 273*, 122834.

[15] Lau, L.-S., Choong, Y.-O., Ching, S.-L., Wei, C.-Y., Senadjki, A., Choong, C.-K., & Seow, A.-N. (2022). Expert insights on Malaysia's residential solar-energy policies: Shortcomings and recommendations. *Clean Energy, 6*(5), 619–631.

[16] Koerner, S. A., Siew, W. S., Salema, A. A., Balan, P., Mekhilef, S., & Thavamoney, N. (2022). Energy policies shaping the solar photovoltaics business models in Malaysia with some insights on COVID-19 pandemic effect. *Energy Policy, 164*, 112918.

[17] Kim, H., Kong, Y.-J., Kim, W.-S., Shafian, S., & Kim, K. (2024). Enhancing reproducibility in organic solar cell fabrication via static sequential deposition with cross-linked polymer donor and nonfullerene acceptor. *ACS Applied Polymer Materials, 6*(8), 5814–5821.

[18] Ibrahim, N. A., Alwi, S. R. W., Abd Manan, Z., Mustaffa, A. A., & Kidam, K. (2024). Climate change impact on solar systems in Malaysia: Techno-economic analysis. *Renewable and Sustainable Energy Reviews, 189*, 113901.

[19] Shafian, S., Mohd Salehin, F. N., Lee, S., Ismail, A., Mohamed Shuhidan, S., Xie, L., & Kim, K. (2025). Development of organic semiconductor materials for organic solar cells via the integration of computational quantum chemistry and AI-powered machine learning. *ACS Applied Energy Materials, 8*(2), 699–722.

[20] Zhang, X., Wang, L., Shafian, S., Wang, P., Zhao, Y., Wang, P., Wu, B., Zhai, J., Chen, J., & Sun, L. (2025). Crosslinking-driven chemical homogeneity enhances performance of pre-seeded perovskite solar cells. *Small, 21*, 2408362.

[21] Salehin, F. N. M., Chelvanathan, P., Goje, A. A., Ludin, N. A., Ibrahim, M. A., & Shafian, S. (2025). Design of blue, green and red colorful semitransparent films using $Ag/SnO_2/Ag$ color filter for integration into solar cells. *Results in Physics, 70*, 108172.

[22] Lee, S., Yoon, Y. S., Shafian, S., Kim, J. Y., & Kim, K. (2025). Sequential co-deposition of perovskite film: An effective way of tailoring bandgap in all-vacuum processed perovskite solar cells. *Small Methods, 9*, 2500104.

[23] Shafian, S., Husen, M. N., Xie, L., & Kim, K. (2025). Predicting high-performance perovskite solar cells using AI-based machine learning models. *Materials Today Sustainability, 31*, 101176.

[24] Zhai, J. X., Xie, L., & Shafian, S. (2025). Advancements in perovskite/CIGS tandem solar cells: Material synergies, device configurations, and economic viability for sustainable energy. *Nanotechnology Reviews, 14*, 20250196.

[25] Martino, A., Iannelli, M., & Truong, C. (2023). Knowledge injection to counter large language model (LLM) hallucination. In *Proceedings of the European Semantic Web Conference* (pp. 182–185).

[26] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

[27] Singh, S., Fore, M., Karatzas, A., Lee, C., Jian, Y., Shangguan, L., Yu, F., Anagnostopoulos, I., & Stamoulis, D. (2024). LLM-dCache: Improving tool-augmented LLMs with GPT-driven localized data caching. In *2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 1–4).

[28] Zou, L., Liu, T., Chen, K., Kong, J., & Deng, Y. (2024). *InstCache: A predictive cache for LLM serving* (arXiv Preprint arXiv:2411.13820).

[29] Nielsen, J. (2005). *Ten usability heuristics*.

[30] Martinez, W. L. (2011). Graphical user interfaces. *Wiley Interdisciplinary Reviews: Computational Statistics, 3*(2), 119–133.