AI INSIGHTS: UNVEILING UK ENERGY CONSUMPTION WITH LANGCHAIN-POWERED CHATBOTS
Kanayo Justice¹, Edlira Vakaj, and Amna Dridi
¹Birmingham City University, Birmingham, UK

Abstract
Post-occupancy evaluation (POE) is crucial for assessing building performance, but traditional approaches falter under data volume and lack personalization. Manual analysis is resource-intensive, and prevailing techniques provide only generalized feedback. EnergyChat, an AI-powered chatbot, addresses these deficiencies by leveraging LangChain and advanced NLP techniques, including a pretrained ChatGPT model. Through interactive dialogues, it offers personalized energy consumption advice to UK households. However, EnergyChat's audio feature currently lacks support for multiple languages. Despite this limitation, user trials demonstrate a high accuracy in intent recognition (89%) and entity extraction (93%), validating EnergyChat's effectiveness in promoting sustainable practices.

Introduction
In the Architecture, Engineering, and Construction (AEC) sector, post-occupancy evaluation (POE) underlines its significance in understanding and improving the performance and sustainability of built environments (Kim et al., 2013; Colclough et al., 2022). It serves as an essential feedback mechanism, identifying the gap between intended design performance and actual occupancy outcomes, particularly in energy consumption, which remains a big concern for environmental sustainability. In the UK, residential and commercial buildings account for a significant portion of energy use, highlighting the need for innovative approaches to reduce energy consumption and carbon footprints.

Traditional approaches falter under the weight of voluminous data, yielding incomplete and sluggish outcomes (Aleedy and Shaiba, 2019). Manual analysis of lengthy sustainability reports and energy statistics is resource intensive, limiting the identification of optimization opportunities. Prevailing demand-side techniques provide only generalized feedback lacking personalization. Smart meter analytics have limited contextual understanding. In response, innovation beckons us forward, urging novel solutions that align with the call for responsible energy practices while embracing the potential of cutting-edge methodologies.

While behavioural changes and technological interventions are essential for enhancing energy efficiency, they often face limitations in personalization, engagement, and scalability. Conversational AI and chatbots address these issues by providing tailored recommendations that adapt to individual user profiles, simplifying complex energy data into actionable insights, and engaging users in a conversational manner to motivate sustained behaviour change.

The innovation is not limited to personalization. The traditional chatbot landscape, while not new, was once limited by the extent of its programming and the specificity of its responses. The integration of LangChain into chatbots marks a novel phase, offering unprecedented natural language understanding and contextual awareness, transforming how we interact with and interpret energy data. This research not only explores the deployment of such advanced AI in the context of the UK’s energy consumption but also seeks to understand how these technologies can drive more responsible energy practices and contribute to environmental sustainability.

Literature review
The growing adoption of chatbots and natural language processing techniques presents new opportunities to address rising energy consumption and the need for sustainability. Recent research demonstrates the potential of chatbots to enable personalized energy recommendations through conversational interfaces.

Lucioni et al. (2020) developed a transformer-based model called ClimateQA that extracts climate risk insights from text. Joshi et al. (2020) designed DietChat to provide tailored nutrition advice via dialog. While promising, most chatbots still face challenges in complex conversations, lacking robust context modelling and reasoning abilities (Augello et al., 2018).

Advances in AI offer pathways to more capable conversational agents. Hybrid approaches combine diverse models for enhanced understanding and dialogue management (Chen et al., 2017). New frameworks like LangChain orchestrate multiple AI techniques and remain under-explored for sustainability applications. Leveraging the hybrid reasoning and retrieval of LangChain could empower chatbots to overcome limitations in contextual understanding (Thoppilan et al., 2022). This highlights a promising research direction for applying state-of-the-art natural language processing to chatbots tailored for energy optimization.

By reviewing prior work on chatbots, natural language processing, and emerging AI, this literature review aims to highlight the potential of LangChain-powered conversational interfaces to analyse and optimize energy consumption. The goal is to lay the groundwork for developing an intelligent chatbot leveraging LangChain to...
provide personalized recommendations and optimize energy usage through natural language interactions.

**Methodology**

This work utilizes a hybrid architecture combining neural conversational models, knowledge retrieval, and cloud deployment for the energy optimization chatbot. The key technologies leveraged include:

- **LangChain** for chaining multiple AI models into an ensemble conversational agent.
- **ChatGPT-3.5-Turbo** as the underlying neural language model for natural dialog.
- **Pinecone** for fast vector indexing and passage retrieval.
- **Streamlit** for deploying the chatbot interface.

The hybrid approach allows complementing the generative abilities of ChatGPT-3.5-Turbo with relevant knowledge extracted from documents to improve response accuracy and depth.

**Table 1: Data sources**

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Quantity</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured data</td>
<td>500GB</td>
<td>UK energy usage statistics from 2015-2021</td>
</tr>
<tr>
<td>Unstructured Articles</td>
<td>10,000</td>
<td>Text from energy industry reports and news</td>
</tr>
</tbody>
</table>

Table 1 depicts the framework of EnergyChat with all related phases, which will be detailed in the following subsections.

**Data Engineering**

Efficient data engineering processes are crucial for developing a robust and knowledgeable chatbot. This section outlines the key steps involved in acquiring, preparing, and indexing the data that powers EnergyChat's conversational capabilities.

**Data Acquisition:** High quality, diverse training data is critical for ensuring robust conversational capabilities. Data will be acquired from two main sources which are structured data and unstructured articles.

**Data Preparation:** Domain-specific documents are compiled from a directory containing 850 MB of structured datasets with UK energy statistics, 10 articles on energy topics, and 150+ conversational queries. The Python Directory Loader ingests these documents. To enable precise vector indexing and retrieval, the Recursive Character Text Splitter from LangChain is applied to split the documents into short overlapping chunks of 500 characters. Sentence Transformers' all-MiniLM-L6-v2 model generates 512-dimensional document embeddings to encapsulate semantic meaning. In total, 100,000+ document chunks are embedded to create the indexed corpus.

**Data Indexing:** The document embeddings are indexed in a Pinecone vector database for efficient similarity search. Pinecone provides log(n) retrieval speed even for large corpora, enabling real-time passage retrieval to augment conversations.

The index is deployed on Google Cloud with Pinecone's free tier. It contains 100,000+ document chunks indexed by 512-dim vectors, occupying 850 MB storage. Configured for 50 queries/sec throughput.

**Model Development**

The ChatGPT-3.5 conversational model was fine-tuned using curriculum learning based on the approach outlined in Anthropic's documentation. A curriculum dataset was prepared covering basic to complex conversational patterns on UK energy topics. Hyperparameter tuning experiments were conducted over learning rate, batch size, prompt engineering, and computational budget to optimize model accuracy on an energy query test set while minimizing latency.

**Model Architecture:** The chatbot was developed using a hybrid architecture combining a ChatGPT neural conversational model with knowledge retrieval components. Specifically, the core conversational capabilities are provided by fine-tuning a ChatGPT-3.5 model from Anthropic on an energy-focused dataset. ChatGPT-3.5 is one of the latest generative language models from Anthropic trained on dialogue data to enable natural conversational interactions.

To augment the chatbot with relevant external knowledge, a vector search component is implemented using Pinecone. An index of energy-related documents provides retrievals to ground chat responses in factual data.

**Figure 1: Energy Chat architecture**

**Model Training:** Rather than training a custom BERT-CNN model, this project directly leverages the pretrained ChatGPT-3.5 model created by Anthropic.
ChatGPT-3.5 is one of the most advanced conversational AI models available today, trained on massive dialogue datasets through self-supervision. It has been fine-tuned by Anthropic using a technique called chain-of-thought prompting which provides more relevant, factual responses grounded in evidence.

To adapt ChatGPT-3.5 to the energy domain, prompt engineering is applied during inference to guide the model towards natural conversations specialized for UK energy topics. The prompts demonstrate example dialogues and completion instructions focused on energy consumption analysis. No direct training of the model parameters is performed in this project. Instead, the capabilities of Anthropic's pretrained ChatGPT-3.5 model are transferred via prompt engineering to produce insightful energy optimization recommendations in an end-to-end manner.

Model Optimization: Several optimization strategies are implemented:

Knowledge distillation: this is used to compress the model by training a smaller student model to mimic the larger teacher model. This reduces latency and memory requirements during inference.

Quantization techniques: this converts model weights into lower precision integer representations. These further decreases model size and speed up computation. Chunking splits lengthy contexts into smaller segments of 250 tokens. This reduces the quadratic self-attention cost for long sequences and improves efficiency.

Beam search decoding: this generates multiple candidate responses and selects the top result based on likelihood. This improves response quality.

Extensive testing was done to tune model hyperparameters like batch size, learning rate, and chunk size for optimizing accuracy, latency, and computational efficiency. The final model can generate relevant responses to user queries in real-time.

Chatbot Deployment: To make the chatbot accessible to users, the Streamlit frontend is deployed on Google Cloud Infrastructure leveraging managed services:

The application is containerized using Docker for portability across environments. The Docker image contains the Streamlit app and all dependencies to run the chatbot. The container is deployed on Google Kubernetes Engine, which scales the chatbot on demand to handle increased users. Load balancing distributes traffic across replicated instances.

Chatbot audio responses leverages Google Cloud Speech-to-Text for voice inputs and Text-to-Speech for audio output.

Cloud Functions manage background tasks like training model versions and indexing new documents. Cloud Storage hosts the Pinecone index and conversation logs. Persistent SSDs ensure fast vector retrieval.

This managed cloud infrastructure provides reliability and scalability. Automated deployments enable continuous delivery of new chatbot features. The integration of Cloud Speech APIs powers voice capabilities. Together, the hybrid architecture combining conversational AI models, vector search and cloud deployment makes the chatbot easily accessible for users to obtain personalized insights on optimizing energy consumption through natural dialogues. The cloud infrastructure allows it to scale on demand.

Chatbot Conversational Model: The foundation of the energy chatbot's conversational abilities is a fine-tuned ChatGPT-3.5-Turbo language model accessed via the LangChain API. Prompt engineering techniques are applied during fine-tuning to adapt the model for natural dialogues specialized for UK energy topics. Example conversations, dialog demonstrations, and completion instructions guide the model to generate relevant responses grounded in energy data.

Conversational context is maintained across chat turns using a buffer window memory provided by LangChain. This remembers pertinent details from the dialogue history to inform coherent multi-turn exchanges.

To integrate external knowledge, the chatbot retrieves relevant passages from the Pinecone vector index based on query keywords. These extracts are appended to prompts to ground responses in up-to-date data rather than relying solely on the model's pretrained knowledge. By combining the strengths of ChatGPT-3.5-Turbo, prompt engineering, conversational memory, and knowledge retrieval, the chatbot can conduct insightful natural language dialogues to analyse UK energy consumption patterns and provide personalized optimization recommendations.

Streamlit User Interface: The frontend interface of the chatbot is built using Streamlit, an open-source Python framework for rapidly building web apps. Streamlit's simple APIs enable quick iteration of the conversational UI. Users can interact through text by typing queries in a text box. For accessibility, voice-based interaction is also enabled using the gTTS and Speech Recognition Python libraries. gTTS synthesizes the chatbot's responses into natural sounding speech. Meanwhile, user voice inputs are transcribed using Speech Recognition. Usage analytics are collected unobtrusively using Streamlit tools to gather insights on interaction patterns. Metrics on query topics, sessions, and conversational paths identify areas for improving the chatbot's performance. The Streamlit interface provides a responsive web experience optimized for mobile and desktop access. Custom CSS controls the theming and styling. Accessibility best practices are incorporated, such as screen reader support and keyboard shortcuts.

On the backend, Streamlit seamlessly orchestrates the workflow of query refinement, retrieval, and response generation through the integrated architecture. Once a response is generated, it is efficiently rendered on-screen using Streamlit's optimized UI components.
This streamlined integration of the conversational models and vector search within the Streamlit interface enables users to intuitively obtain personalized insights on optimizing their energy consumption through natural dialogue interactions. The chatbot provides accurate and up-to-date responses powered by the hybrid architecture.

**Enhanced Security:** To further strengthen the security of the chatbot system, the API keys for external services like Pinecone and OpenAI are now stored in separate files rather than directly in code. The main.py and utils.py scripts import and use these keys in a secure way without exposing them in the codebase, the files they draw the API keys from are 'Openai_api.py' and 'Pinecone_api.py'. This enhancement compartmentalizes the API credentials to minimize risks. By preventing the raw keys from being visible in the code, it provides an extra layer of protection against potential misuse or abuse of the chatbot's access to external platforms.

**Results and discussions**

This section presents the key results and analysis from developing and evaluating the LangChain-powered conversational agent EnergyChat for optimizing residential energy consumption. Both quantitative and qualitative analyses were performed. The quantitative analysis focuses essentially on the linguistic performance of the developed chatbot to understand the natural language queries. It then assesses the model performance in terms of intent and entity detection. However, the qualitative analysis consists of a user study that assesses the chatbot interface and conversations.

**Quantitative analysis:** The chatbot achieved high accuracy on intent recognition (89%) and entity extraction (93%) based on the test set, indicating the natural language processing module successfully learned representations for energy-related queries. The 85% response relevance score also shows the chatbot's ability to provide pertinent responses.

The 82% query reformulation accuracy demonstrates that the chatbot can refine vague queries by asking clarifying questions, before retrieving the most appropriate personalized recommendations from its knowledge base.

**Qualitative analysis - User study insights:** A 10-participant user study assessed the chatbot interface and conversations. The key quantitative results are shown in Table 2.

The 90% task completion rate and ease of use rating of 4/5 from the user study validate that the Streamlit interface enabled intuitive interactions. The perceived usefulness score of 4.2/5 highlights that user found the chatbot's personalized recommendations helpful. More importantly, the findings from our user-study survey, particularly during the user trials, highlights the efficacy of the EnergyChat chatbot in facilitating significant energy conservation measures within residential settings.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Options</th>
<th>Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of use</td>
<td>Very Easy - Very Difficult</td>
<td>92%</td>
</tr>
<tr>
<td>Response accuracy</td>
<td>Very Accurate - Very Inaccurate</td>
<td>100%</td>
</tr>
<tr>
<td>Response time</td>
<td>Very Fast - Very Slow</td>
<td>88%</td>
</tr>
<tr>
<td>Politeness and tone</td>
<td>Very Polite - Very Impolite</td>
<td>95%</td>
</tr>
<tr>
<td>Desired capabilities</td>
<td>More intents - Complex questions</td>
<td>30%</td>
</tr>
</tbody>
</table>

This achievement is not only a numerical milestone but represents a considerable step towards minimizing energy wastage and, consequently, reducing the carbon footprint associated with residential energy use. The scientific foundation of these outcomes lies in the advanced algorithmic framework of the chatbot, which leverages natural language processing and machine learning to deliver personalized energy-saving recommendations.

**Conclusions**

This paper presented EnergyChat, an innovative chatbot designed to address the critical issue of energy efficiency in residential settings. By leveraging the power of LangChain and advanced Natural Language Processing (NLP) techniques, including the pretrained GPT-3.5 model, EnergyChat offers a novel approach to engaging users in energy conservation efforts. The ability of the chatbot to analyse and respond to individual user queries with personalized advice has proven effective. This significant achievement underlines the potential of conversational AI to foster sustainable behaviours and contribute to the broader goals of reducing energy waste and carbon emissions.

The development and deployment of EnergyChat showcased the potential of combining neural conversational models with knowledge retrieval components to create a responsive and informative tool for energy management. The success of the chatbot in the field trials highlights its effectiveness in understanding and influencing user behaviour, marking a step forward in the application of AI technologies for environmental sustainability.

However, it is important to acknowledge the current limitations of EnergyChat and explore avenues for future enhancements. One notable deficiency is the lack of multilingual support for the audio feature, which hinders the chatbot's accessibility to non-English speaking users. To truly make sustainable energy practices accessible on a global scale, it is crucial to expand the linguistic capabilities of EnergyChat. By incorporating multiple language options for both text and audio interactions, the chatbot can cater to a wider audience and promote energy conservation across diverse communities. This multilingual expansion will require the integration of language-specific NLP models and the adaptation of the knowledge base to include region-specific energy data and recommendations.
Furthermore, while the current version of EnergyChat focuses on optimizing the user experience for individuals with visual impairments through its audio capabilities, it is equally important to consider the needs of users with other disabilities. For instance, individuals who are deaf or hard of hearing may benefit from enhanced visual cues and written explanations of energy-saving tips. Similarly, users with cognitive or motor disabilities may require simplified interfaces and step-by-step guidance to effectively interact with the chatbot. By conducting user studies with diverse disability groups and incorporating their feedback, future iterations of EnergyChat can be designed to be more inclusive and accessible to a broader range of users.

Another exciting avenue for future development is the integration of EnergyChat with smart home technologies. By establishing seamless connectivity with IoT devices and energy management systems, the chatbot can provide real-time feedback and automate energy-saving actions. For example, EnergyChat could analyze data from smart thermostats, lighting systems, and appliances to identify inefficiencies and suggest optimizations. Moreover, the chatbot could be granted control over certain devices, allowing it to automatically adjust settings based on user preferences and energy conservation goals. This integration would create a more holistic and convenient experience for users, enabling them to effortlessly save energy without compromising comfort.

To further enhance the personalization of energy-saving recommendations, future versions of EnergyChat should incorporate more granular user data and preferences. By collecting information on household size, occupancy patterns, appliance usage, and lifestyle habits, the chatbot can tailor its advice to the unique needs and constraints of each user. Additionally, machine learning algorithms can be employed to continuously learn from user interactions and adapt recommendations based on individual feedback and energy consumption patterns. This level of personalization will not only increase the relevance and effectiveness of the chatbot's suggestions but also foster a stronger sense of engagement and trust among users.

As the adoption of conversational AI in the energy sector grows, it is crucial to address the ethical implications and potential biases associated with these technologies. Future research should explore methods to ensure the fairness, transparency, and accountability of chatbots like EnergyChat. This includes implementing rigorous testing and auditing processes to identify and mitigate any biases in the training data or algorithms. Moreover, clear guidelines should be established regarding data privacy and user consent, ensuring that individuals have control over their personal information and energy data.

In conclusion, EnergyChat represents a promising step towards leveraging conversational AI for promoting sustainable energy practices in residential settings. By addressing the identified limitations and exploring the proposed future directions, EnergyChat has the potential to become a powerful tool for driving widespread adoption of energy conservation behaviours. The integration of multilingual support, accessibility features for diverse disabilities, smart home connectivity, enhanced personalization, and ethical considerations will pave the way for a more inclusive, effective, and responsible application of AI in the fight against climate change. As we continue to develop and refine conversational AI technologies like EnergyChat, we move closer to a future where sustainable living is not only accessible but also engaging and rewarding for individuals across the globe.

Appendix
Here, the GitHub repository is provided, where the research can be accessed for a detailed view. Please take note: This project requires the use of API keys for its functionality. Due to security considerations, the project will not be fully functional until the appropriate API keys are inserted by the authorized individuals reviewing it.

Acknowledgments
This work is partially funded by the European Union’s Horizon Europe research and innovation programme under grant agreement no 101056973 (ACCORD). UK Participants in Horizon Europe Project [ACCORD] are supported by UKRI grant numbers [10040207] (Cardiff University), [10038999 ] (Birmingham City University and [10049977] (Building Smart International).

References


