LLM Energy Lit Highlights

link to html version: https://tegorman13.github.io/ccl/llm_energy.html

Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models.

King, E., Yu, H., Lee, S., & Julien, C. (2024). **Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models.** Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(1), 1–38. https://doi.org/10.1145/3643505

Abstract

Smart home assistants function best when user commands are direct and well-specified—e.g., "turn on the kitchen light"—or when a hard-coded routine specifies the response. In more natural communication, however, human speech is unconstrained, often describing goals (e.g., "make it cozy in here" or "help me save energy") rather than indicating specific target devices and actions to take on those devices. Current systems fail to understand these under-specified commands since they cannot reason about devices and settings as they relate to human situations. We introduce large language models (LLMs) to this problem space, exploring their use for controlling devices and creating automation routines in response to under-specified user commands in smart homes. We empirically study the baseline quality and failure modes of LLM-created action plans with a survey of age-diverse users. We find that LLMs can reason creatively to achieve challenging goals, but they experience patterns of failure that diminish their usefulness. We address these gaps with Sasha, a smarter smart home assistant. Sasha responds to loosely-constrained commands like "make it cozy" or "help me sleep better" by executing plans to achieve user goals—e.g., setting a mood with available devices, or devising automation routines. We implement and evaluate Sasha in a hands-on user study, showing the capabilities and limitations of LLM-driven smart homes when faced with unconstrained user-generated scenarios.

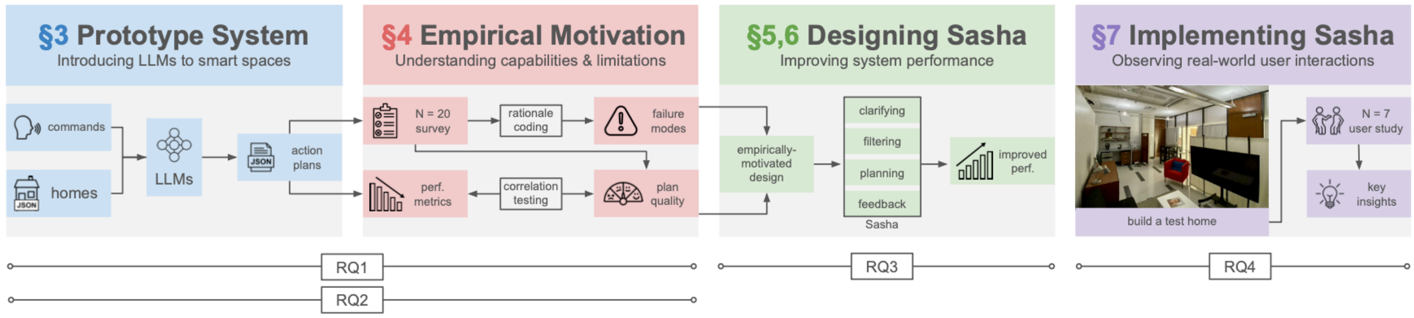


Fig. 1. Overview of the paper's four key components, their internal structure, and motivating research questions.

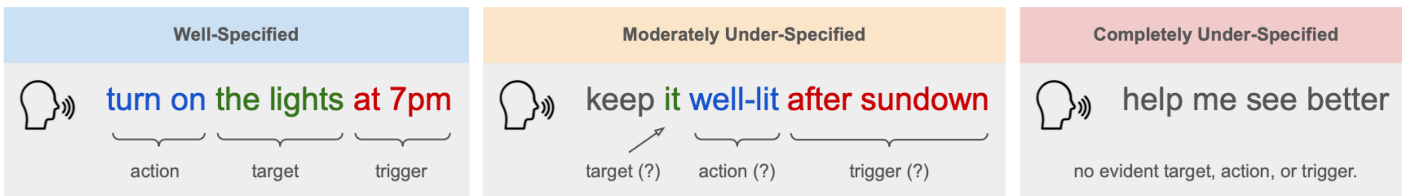


Fig. 2. User commands have varying degrees of specificity. Left: Well-specified commands define specific actions, specific target devices, and specific triggers. Middle: Moderately under-specified commands may allude to targets, actions, and triggers, but do not specify them. Right: Completely under-specified commands have no obvious targets, desired actions, or triggers. We focus on moderately and completely under-specified commands.

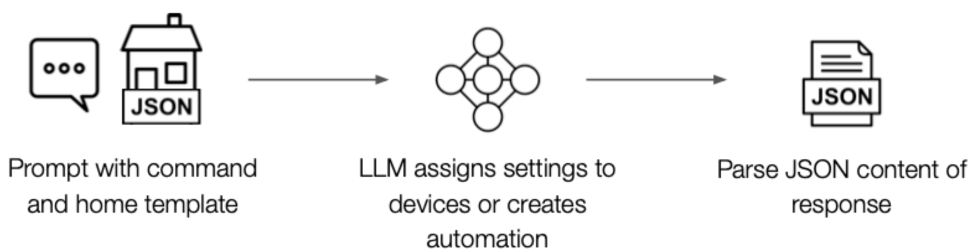


Fig. 3. We begin with an initial study of LLM behavior in smart home applications using this experimental setup.

```
You are an AI that controls a smart home. You receive user commands and
assign settings to devices in response.
User command: [ command ]
Devices: { JSON }
If there are devices relevant to the user command, respond with the device
JSON with settings assigned in the form: { "status": "success",
"devices": { }, "explanation": "" }
If there are no devices relevant to the user command, respond with JSON in the
form: { "status": "failure" }
```

```
You are an AI that controls a smart home. You receive user commands and
create automation routines in response.
User command: [ command ]
Devices: { JSON }
Sensors: { JSON }
If there are devices relevant to the user command, respond with JSON that
describes a sensor trigger and how you would change the devices (the action)
based on that trigger. Respond with JSON in this form: { "status":
"success", "trigger": { }, "action": { }, "explanation":
"" }
If there are no devices relevant to the user command, respond with JSON in
the form: { "status": "failure" }
```

Fig. 4. The text of immediate and persistent goal prompts (left and right, respectively). Immediate prompts produce action plans that can be executed immediately, while persistent prompts produce automation routines.

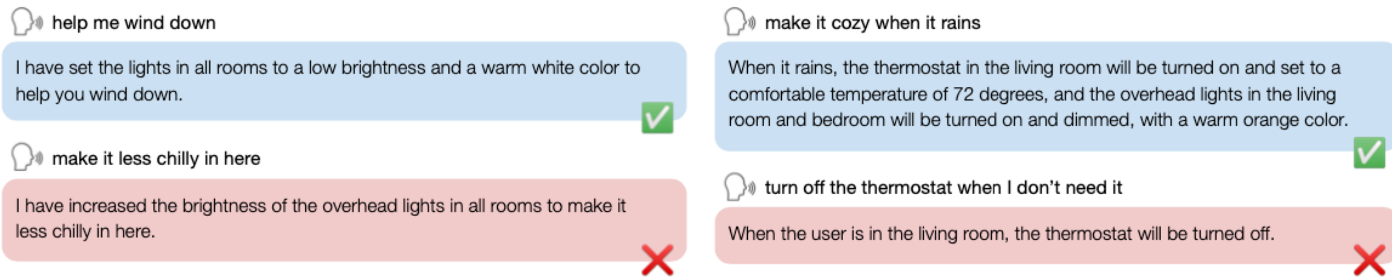


Fig. 5. GPT-3.5 responses to smart home commands. We prompt the model with a user command along with information about the devices and sensors available in the home. Responses are often remarkably creative, and often remarkably wrong.

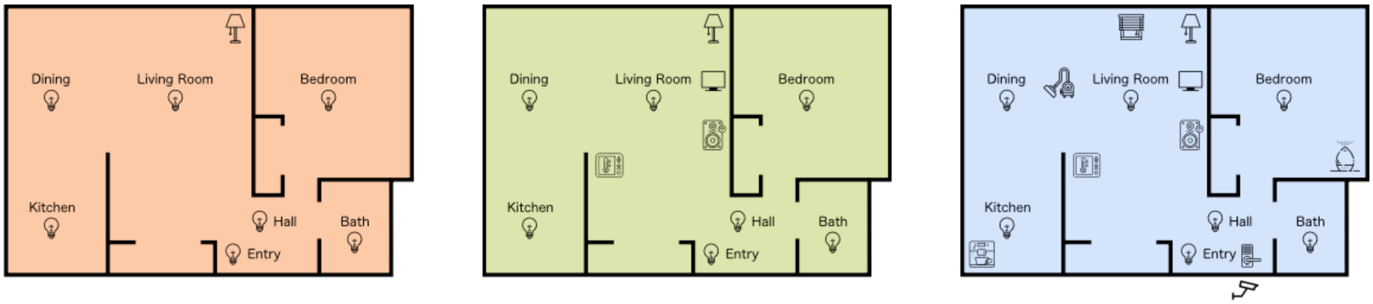


Fig. 6. We model three homes (h_1, h_2, and h_3 from left to right) with an increasingly diverse set of *devices*. All three homes have the same suite of *sensors*. The model's reasoning is based on a given command and the devices and sensors available.

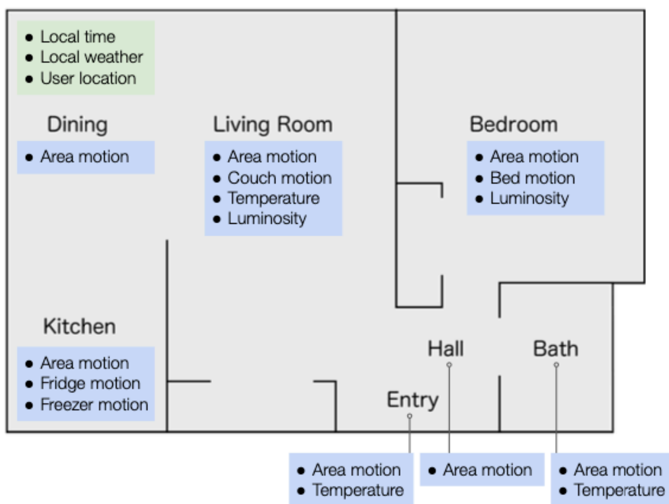


Fig. 7. Each home template in our set of homes h_1, h_2, and h_3 has the same suite of sensors, as shown.

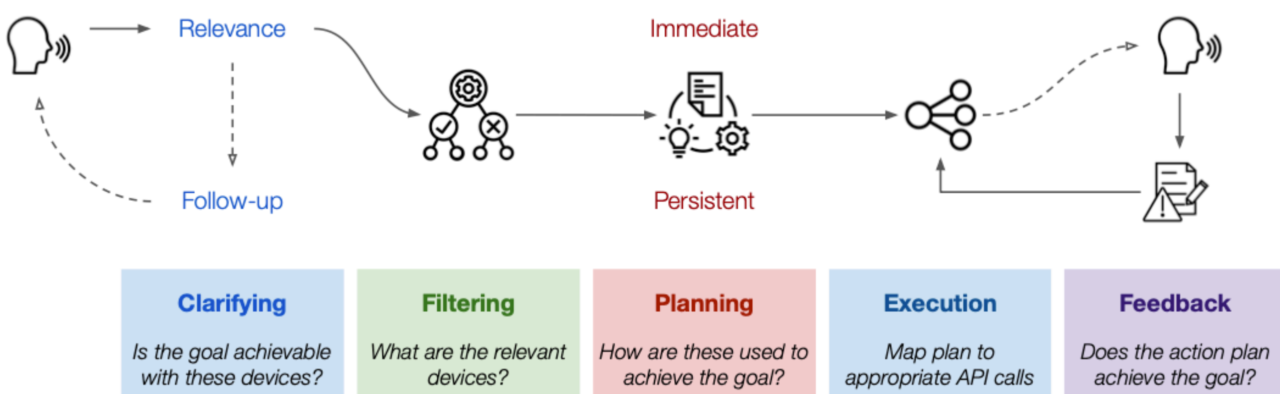


Fig. 11. Sasha uses iterative reasoning to generate consistently high-quality action plans that leverage whichever relevant devices are available in a given smart home.

Figure 1: Figure from King, Yu, Lee, et al. (2024)

Thoughtful Things: Building Human-Centric Smart Devices with Small Language Models

King, E., Yu, H., Vartak, S., Jacob, J., Lee, S., & Julien, C. (2024). **Thoughtful Things: Building Human-Centric Smart Devices with Small Language Models** (arXiv:2405.03821). arXiv. http://arxiv.org/abs/2405.03821

Abstract

Everyday devices like light bulbs and kitchen appliances are now embedded with so many features and automated behaviors that they have become complicated to actually use. While such "smart" capabilities can better support users' goals, the task of learning the "ins and outs" of different devices is daunting. Voice assistants aim to solve this problem by providing a natural language interface to devices, yet such assistants cannot understand loosely-constrained commands, they lack the ability to reason about and explain devices' behaviors to users, and they rely on connectivity to intrusive cloud infrastructure. Toward addressing these issues, we propose thoughtful things: devices that leverage lightweight, on-device language models to take actions and explain their behaviors in response to unconstrained user commands. We propose an end-to-end framework that leverages formal modeling, automated training data synthesis, and generative language models to create devices that are both capable and thoughtful in the presence of unconstrained user goals and inquiries. Our framework requires no labeled data and can be deployed on-device, with no cloud dependency. We implement two thoughtful things (a lamp and a thermostat) and deploy them on real hardware, evaluating their practical performance.

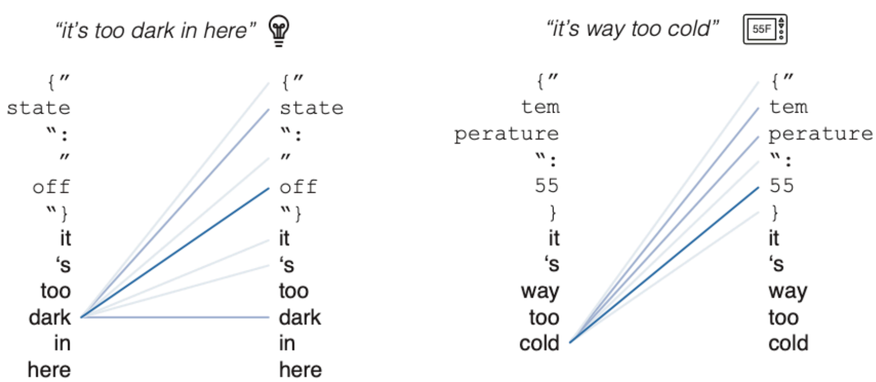


Fig. 2. Visualizations of attention in a transformer model (Phi-2) given input user commands and device states. Generalist base models are pre-trained on large amounts of code and unconstrained text, so they learn semantic relationships between commands and relevant machine-readable state (e.g., "dark" and "off"). In this paper, we fine-tune these models to generate responses with device-specific actions and explanations that adhere to a real device's state model.

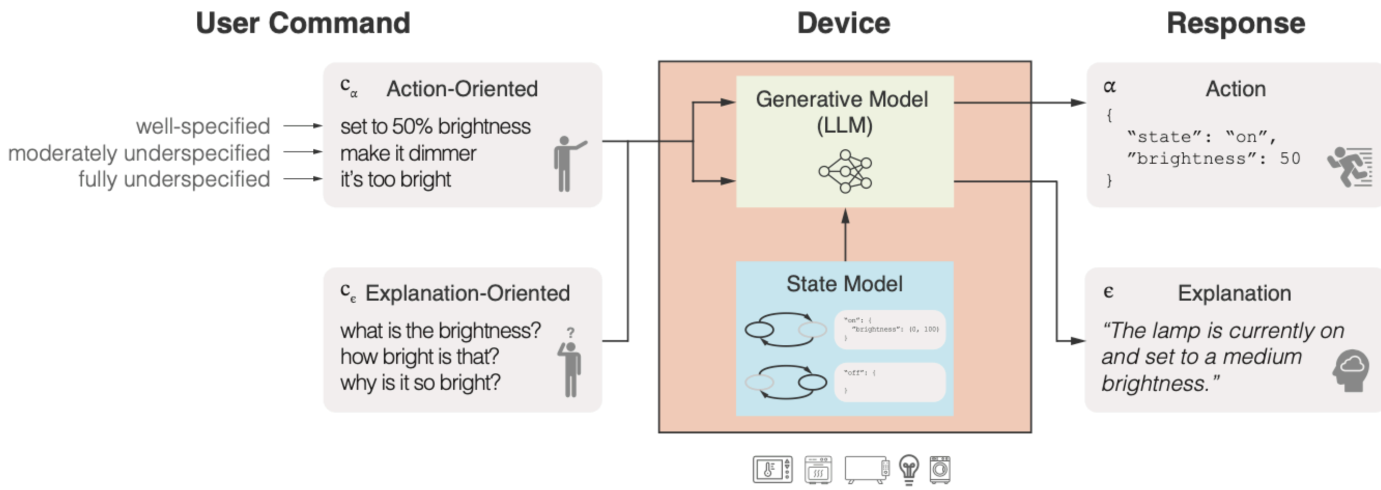


Fig. 3. Thoughtful things are devices that respond to unconstrained user commands with appropriate actions (i.e., state changes) or explanations (i.e., descriptions of current state and capabilities). We accomplish this by combining a small, fine-tuned generative language model with a formal system model. The LLM flexibly synthesizes new states and explanations in response to diverse user commands, while the system model grounds responses in a device's true capabilities.

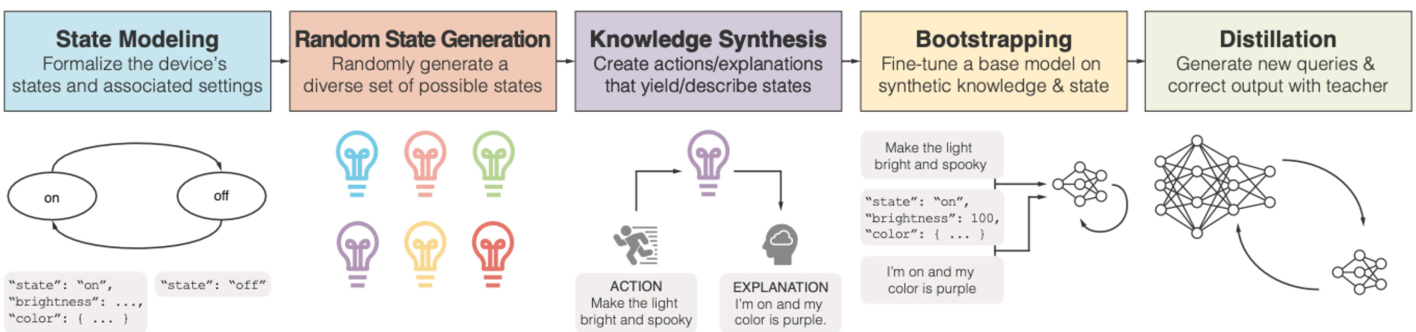


Fig. 4. Overview of our framework. Our five-step process leverages a combination of formal modeling, training data synthesis, and fine-tuning and distillation of large language models to train a lightweight model capable of generating appropriate settings and explanations for individual smart devices in response to unconstrained user commands.

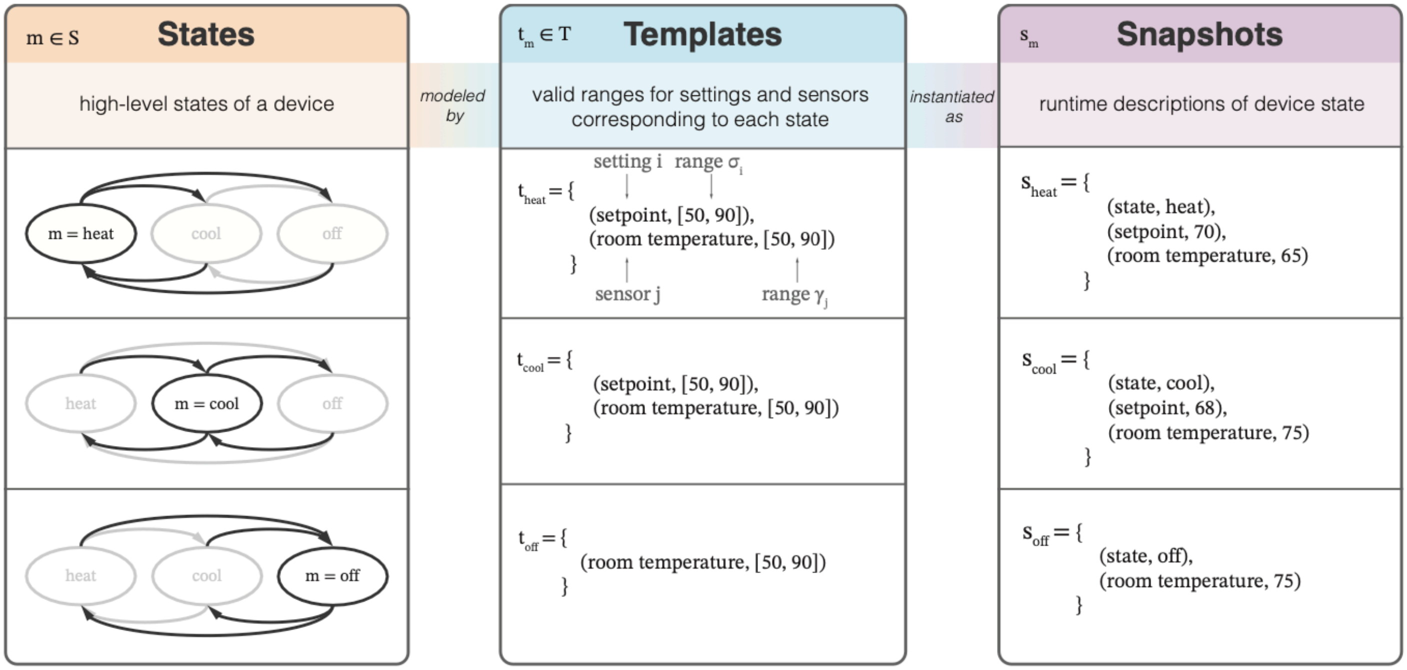


Fig. 5. Overview of device state models, with examples included for a thermostat. We model a device based on a high-level state machine S that describes valid transitions between states m (left). A lower-level template t_m associated with each state m captures each setting i and sensor j and their valid ranges σ_i, γ_j (center). At runtime, a snapshot s_m describes the current state of the device (right). A thoughtful thing leverages a fine-tuned small language model to *act* by generating valid snapshots for new states and to *explain* by describing the snapshot of the device's current state.

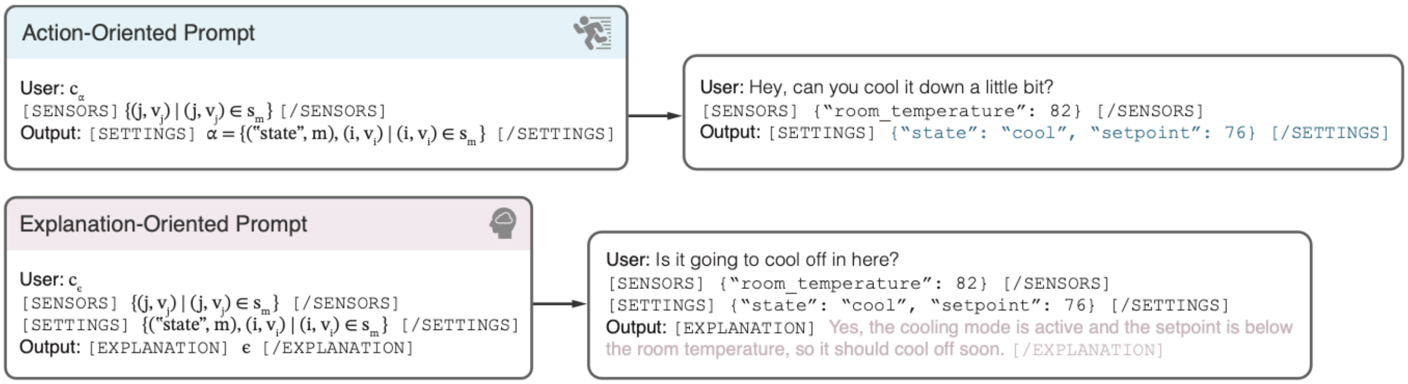


Fig. 7. Prompt formats that we use to fine-tune a small language model for device actions/explanations. Examples show prompts for a thermostat device, which has both "sensors" and "settings" in its state model. When fine-tuning, we substitute the full command, state snapshot (split into immutable sensors and mutable settings), and explanation (if applicable) into the prompt to create training instances. The fine-tuned model learns to complete the prompt with either actions or explanations (depicted as colored text) depending on the prompt.

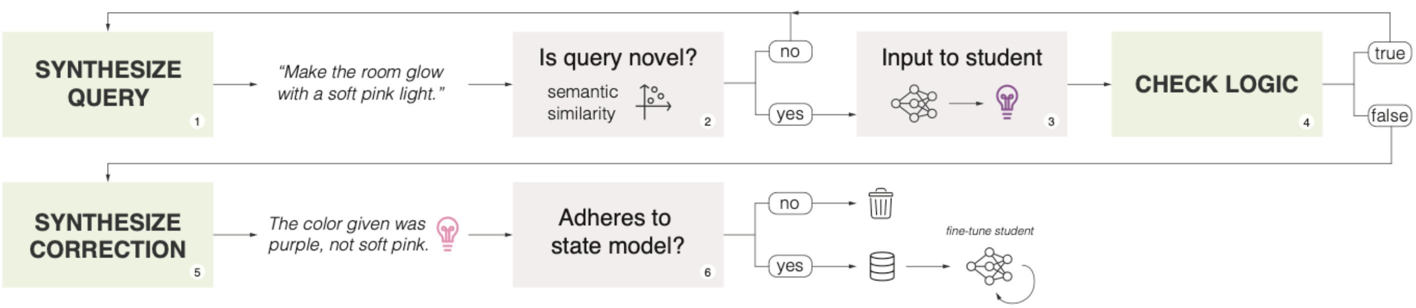


Fig. 8. We use an iterative distillation process to align the outputs of the small language model in a thoughtful thing (i.e., the "student" model) with the outputs of a higher-performance "teacher" LLM. Green boxes denote prompts to the teacher model. This process distills specialized knowledge about a given device's capabilities into the generative model that a thoughtful thing uses to act and explain in response to user commands.

Figure 2: Figure from King, Yu, Vartak, et al. (2024)

Designing Home Automation Routines Using an LLM-Based Chatbot.

Giudici, M., Padalino, L., Paolino, G., Paratici, I., Pascu, A. I., & Garzotto, F. (2024). **Designing Home Automation Routines Using an LLM-Based Chatbot.** Designs, 8(3), Article 3. https://doi.org/10.3390/designs8030043

Abstract

Without any more delay, individuals are urged to adopt more sustainable behaviors to fight climate change. New digital systems mixed with engaging and gamification mechanisms could play an important role in achieving such an objective. In particular, Conversational Agents, like Smart Home Assistants, are a promising tool that encourage sustainable behaviors within household settings. In recent years, large language models (LLMs) have shown great potential in enhancing the capabilities of such assistants, making them more effective in interacting with users. We present the design and implementation of GreenIFTTT, an application empowered by GPT4 to create and control home automation routines. The agent helps users understand which energy consumption optimization routines could be created and applied to make their home appliances more environmentally sustainable. We performed an exploratory study (Italy, December 2023) with N = 13 participants to test our application's usability and UX. The results suggest that GreenIFTTT is a usable, engaging, easy, and supportive tool, providing insight into new perspectives and usage of LLMs to create more environmentally sustainable home automation.

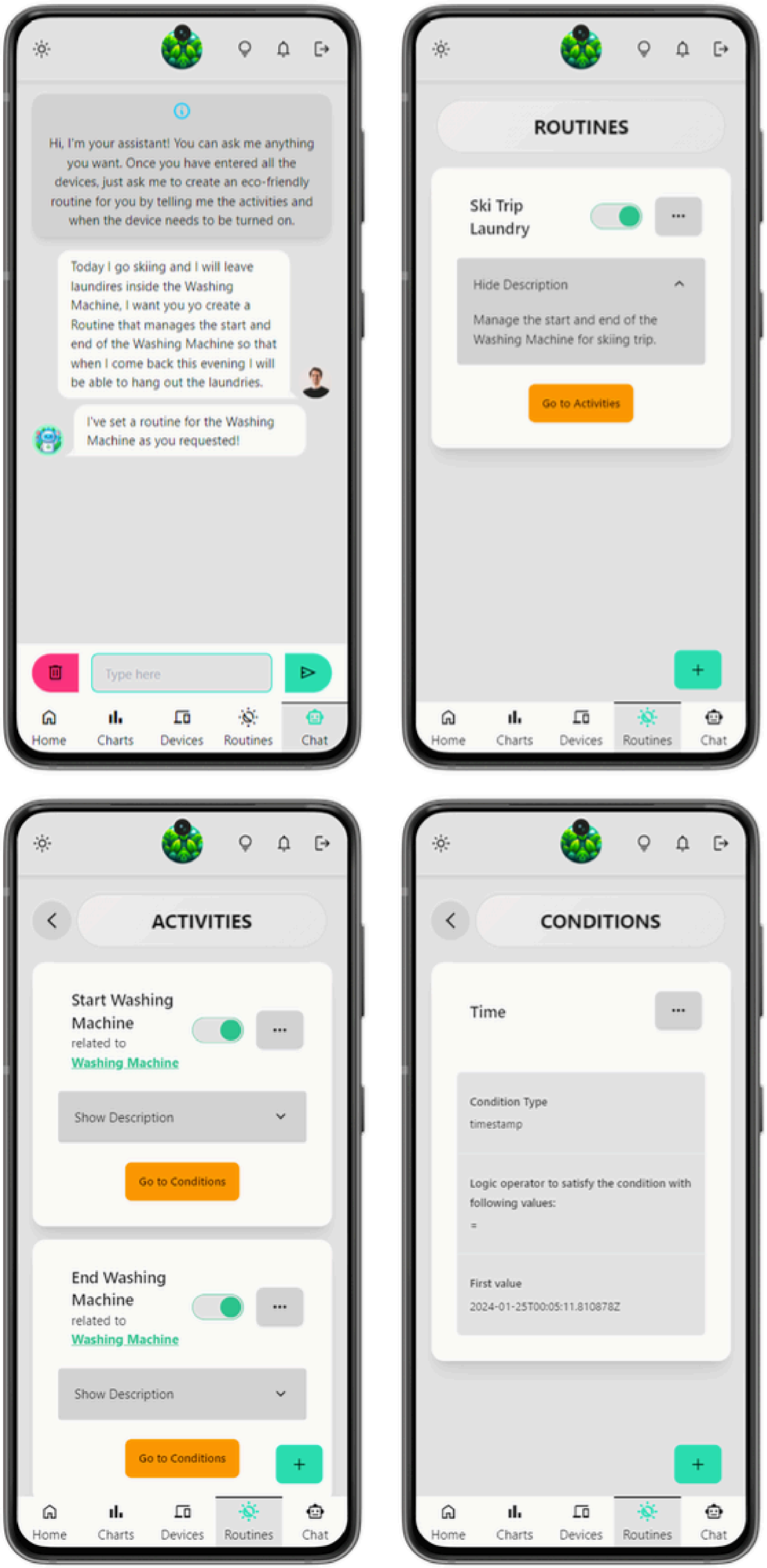


Figure 1. Overview of *GreenIFTTT* application.

Figure 3: Figure from Giudici et al. (2024)

AIoT Smart Home via Autonomous LLM Agents.

Rivkin, D., Hogan, F., Feriani, A., Konar, A., Sigal, A., Liu, X., & Dudek, G. (2024). **AIoT Smart Home via Autonomous LLM Agents.** IEEE Internet of Things Journal, 1–1. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2024.3471904

Abstract

The common-sense reasoning abilities and vast general knowledge of Large Language Models (LLMs) make them a natural fit for interpreting user requests in a smart home assistant context. LLMs, however, lack specific knowledge about the user and their home, which limits their potential impact. SAGE (Smart Home Agent with Grounded Execution), overcomes these and other limitations by using a scheme in which a user request triggers an LLM-controlled sequence of discrete actions. These actions can be used to retrieve information, interact with the user, or manipulate device states. SAGE controls this process through a dynamically constructed tree of LLM prompts, which help it decide which action to take next, whether an action was successful, and when to terminate the process. The SAGE action set augments an LLM's capabilities to support some of the most critical requirements for a smart home assistant. These include: flexible and scalable user preference management ("Is my team playing tonight?"), access to any smart device's full functionality without device-specific code via API reading ("Turn down the screen brightness on my dryer"), persistent device state monitoring ("Remind me to throw out the milk when I open the fridge"), natural device references using only a photo of the room ("Turn on the lamp on the dresser"), and more. We introduce a benchmark of 50 new and challenging smart home tasks where SAGE achieves a 76% success rate, significantly outperforming existing LLM-enabled baselines (30% success rate).

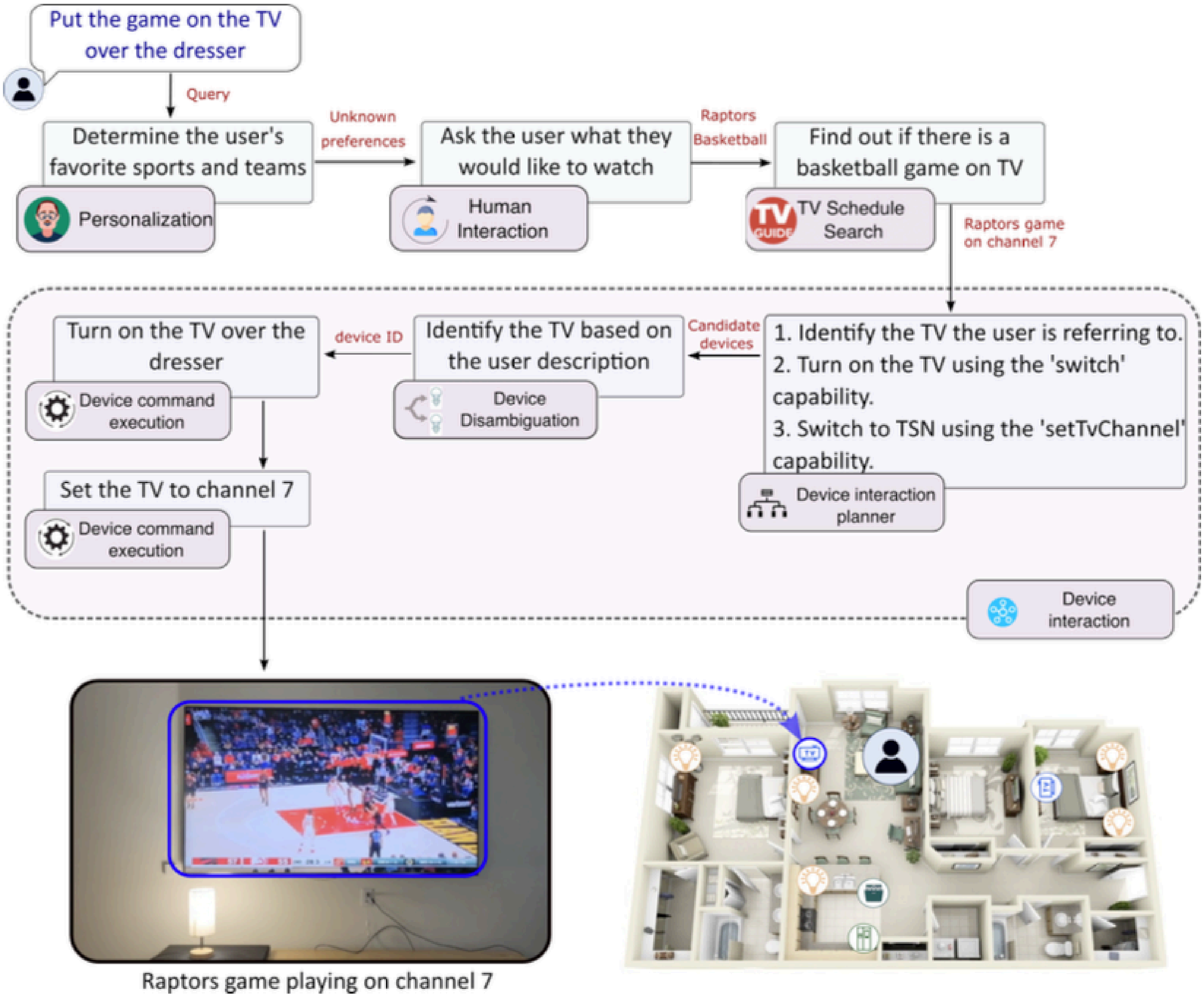


Fig. 1. **Demonstration of SAGE response process to user command: *"Put the game on by the dresser"*.** The figure illustrates the sequence of tools used by SAGE to complete the task. Note that each control flow decision was made by an LLM, not by hand-coded logic. The control flow is described step by step in Section V. This demo was executed using real SmartThings-enabled devices.

Figure 4: Figure from Rivkin et al. (2024)

Leveraging Large Language Models for enhanced personalised user experience in Smart Homes

Rey-Jouanchicot, J., Bottaro, A., Campo, E., Bouraoui, J.-L., Vigouroux, N., & Vella, F. (2024). **Leveraging Large Language Models for enhanced personalised user experience in Smart Homes** (No. arXiv:2407.12024). arXiv. http://arxiv.org/abs/2407.12024

Abstract

Smart home automation systems aim to improve the comfort and convenience of users in their living environment. However, adapting automation to user needs remains a challenge. Indeed, many systems still rely on hand-crafted routines for each smart object.This paper presents an original smart home architecture leveraging Large Language Models (LLMs) and user preferences to push the boundaries of personalisation and intuitiveness in the home environment.This article explores a human-centred approach that uses the general knowledge provided by LLMs to learn and facilitate interactions with the environment.The advantages of the proposed model are demonstrated on a set of scenarios, as well as a comparative analysis with various LLM implementations. Some metrics are assessed to determine the system's ability to maintain comfort, safety, and user preferences. The paper details the approach to real-world implementation and evaluation.The proposed approach of using preferences shows up to 52.3% increase in average grade, and with an average processing time reduced by 35.6% on Starling 7B Alpha LLM. In addition, performance is 26.4% better than the results of the larger models without preferences, with processing time almost 20 times faster.

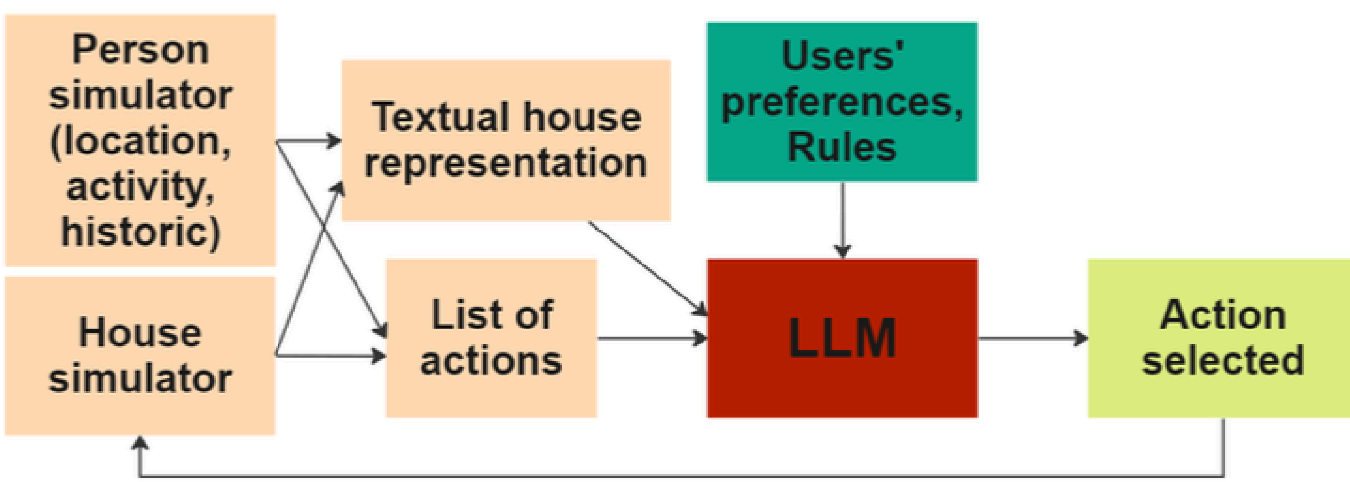


Figure 1: The implemented architecture

Figure 5: Figure from Rey-Jouanchicot et al. (2024)

Enhancing smart home interaction through multimodal command disambiguation.

Calò, T., & De Russis, L. (2024). **Enhancing smart home interaction through multimodal command disambiguation.** Personal and Ubiquitous Computing. https://doi.org/10.1007/s00779-024-01827-3

Abstract

Smart speakers are entering our homes and enriching the connected ecosystem already present in them. Home inhabitants can use those to execute relatively simple commands, e.g., turning a lamp on. Their capabilities to interpret more complex and ambiguous commands (e.g., make this room warmer) are limited, if not absent. Large language models (LLMs) can offer creative and viable solutions to enable a practical and user-acceptable interpretation of such ambiguous commands. This paper introduces an interactive disambiguation approach that integrates visual and textual cues with natural language commands. After contextualizing the approach with a use case, we test it in an experiment where users are prompted to select the appropriate cue (an image or a textual description) to clarify ambiguous commands, thereby refining the accuracy of the system's interpretations. Outcomes from the study indicate that the disambiguation system produces responses well-aligned with user intentions, and that participants found the textual descriptions slightly more effective. Finally, interviews reveal heightened satisfaction with the smart-home system when engaging with the proposed disambiguation approach.

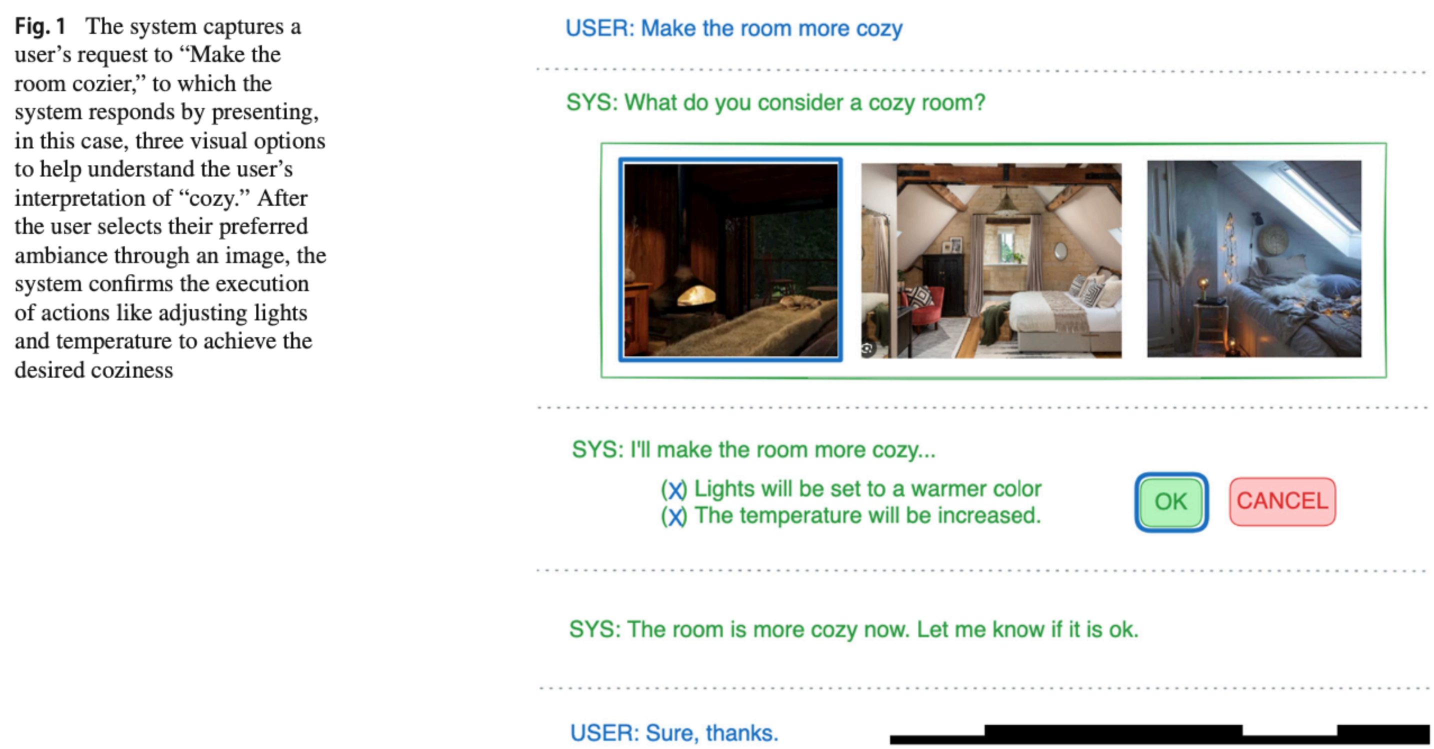


Figure 6: Figure from Calò & De Russis (2024)

A Human-on-the-Loop Optimization Autoformalism Approach for Sustainability

Jin, M., Sel, B., Hardeep, F., & Yin, W. (2023). **A Human-on-the-Loop Optimization Autoformalism Approach for Sustainability** (arXiv:2308.10380). arXiv. http://arxiv.org/abs/2308.10380

Abstract

This paper outlines a natural conversational approach to solving personalized energy-related problems using large language models (LLMs). We focus on customizable optimization problems that necessitate repeated solving with slight variations in modeling and are user-specific, hence posing a challenge to devising a one-size-fits-all model. We put forward a strategy that augments an LLM with an optimization solver, enhancing its proficiency in understanding and responding to user specifications and preferences while providing nonlinear reasoning capabilities. Our approach pioneers the novel concept of human-guided optimization autoformalism, translating a natural language task specification automatically into an optimization instance. This enables LLMs to analyze, explain, and tackle a variety of instance-specific energy-related problems, pushing beyond the limits of current prompt-based techniques. Our research encompasses various commonplace tasks in the energy sector, from electric vehicle charging and Heating, Ventilation, and Air Conditioning (HVAC) control to long-term planning problems such as cost-benefit evaluations for installing rooftop solar photovoltaics (PVs) or heat pumps. This pilot study marks an essential stride towards the context-based formulation of optimization using LLMs, with the potential to democratize optimization processes. As a result, stakeholders are empowered to optimize their energy consumption, promoting sustainable energy practices customized to personal needs and preferences.

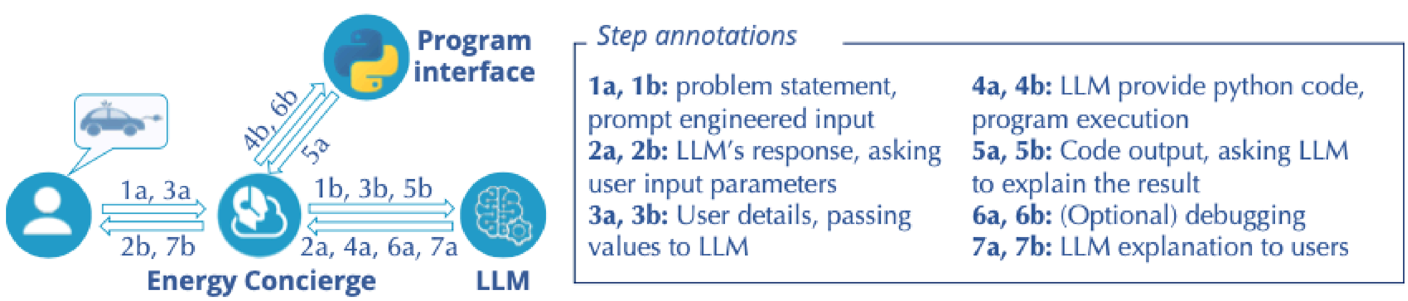


Figure 3: Energy Concierge framework. The user engages with an LLM through natural language queries and responses. The LLM identifies the necessary input parameters for optimization and generates Python code to address the problem. The program interface then executes the code and relays the solution back to the LLM, which subsequently provides a clear explanation to the user.

Figure 7: Figure from Jin et al. (2023)

Harmony: A Home Agent for Responsive Management and Action Optimization with a Locally Deployed Large Language Model

Yin, Z., Zhang, M., & Kawahara, D. (2024). **Harmony: A Home Agent for Responsive Management and Action Optimization with a Locally Deployed Large Language Model** (No. arXiv:2410.14252). arXiv. http://arxiv.org/abs/2410.14252

Abstract

Since the launch of GPT-3.5, intelligent home assistant technology based on large language models (LLMs) has made significant progress. These intelligent home assistant frameworks, such as those based on high-performance LLMs like GPT-4, have greatly expanded their functional range and application scenarios by computing on the cloud, enriching user experience and diversification. In order to optimize the privacy and economy of data processing while maintaining the powerful functions of LLMs, we propose Harmony, a smart home assistant framework that uses a locally deployable small-scale LLM. Based on Llama3-8b, an open LLM that can be easily deployed on a consumer-grade PC, Harmony does not send any data to the internet during operation, ensuring local computation and privacy secured. Harmony based on Llama3-8b achieved competitive performance on our benchmark tests with the framework used in related work with GPT-4. In addition to solving the issues mentioned above, Harmony can also take actions according to the user and home status, even if the user does not issue a command. For example, when the user wants to wake up later than normal on the weekend, Harmony would open the curtains only when the user gets up or prepare the room when the user comes home without requiring user commands.

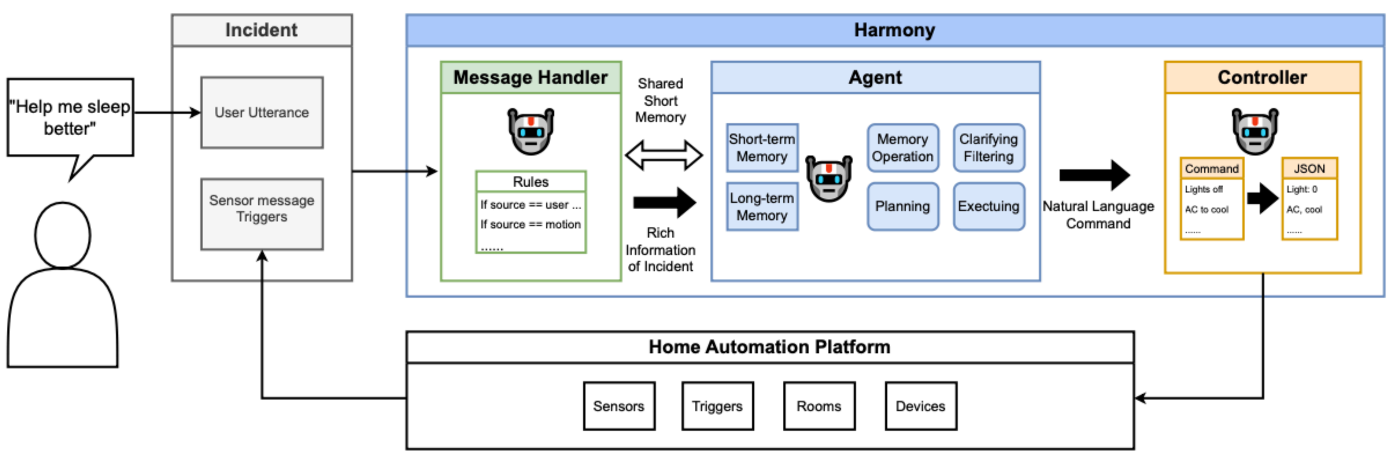


Figure 3: Illustration of Harmony

Figure 8: Figure from Yin et al. (2024)

Save It for the "Hot" Day: An LLM-Empowered Visual Analytics System for Heat Risk Management

Li, H., Kam-Kwai, W., Luo, Y., Chen, J., Liu, C., Zhang, Y., Lau, A. K. H., Qu, H., & Liu, D. (2024). **Save It for the "Hot" Day: An LLM-Empowered Visual Analytics System for Heat Risk Management** (arXiv:2406.03317). arXiv. http://arxiv.org/abs/2406.03317

Abstract

The escalating frequency and intensity of heat-related climate events, particularly heatwaves, emphasize the pressing need for advanced heat risk management strategies. Current approaches, primarily relying on numerical models, face challenges in spatial-temporal resolution and in capturing the dynamic interplay of environmental, social, and behavioral factors affecting heat risks. This has led to difficulties in translating risk assessments into effective mitigation actions. Recognizing these problems, we introduce a novel approach leveraging the burgeoning capabilities of Large Language Models (LLMs) to extract rich and contextual insights from news reports. We hence propose an LLM-empowered visual analytics system, Havior, that integrates the precise, data-driven insights of numerical models with nuanced news report information. This hybrid approach enables a more comprehensive assessment of heat risks and better identification, assessment, and mitigation of heat-related threats. The system incorporates novel visualization designs, such as "thermoglyph" and news glyph, enhancing intuitive understanding and analysis of heat risks. The integration of LLM-based techniques also enables advanced information retrieval and semantic knowledge extraction that can be guided by experts' analytics needs. Our case studies on two cities that faced significant heatwave events and interviews with five experts have demonstrated the usefulness of our system in providing in-depth and actionable insights for heat risk management.

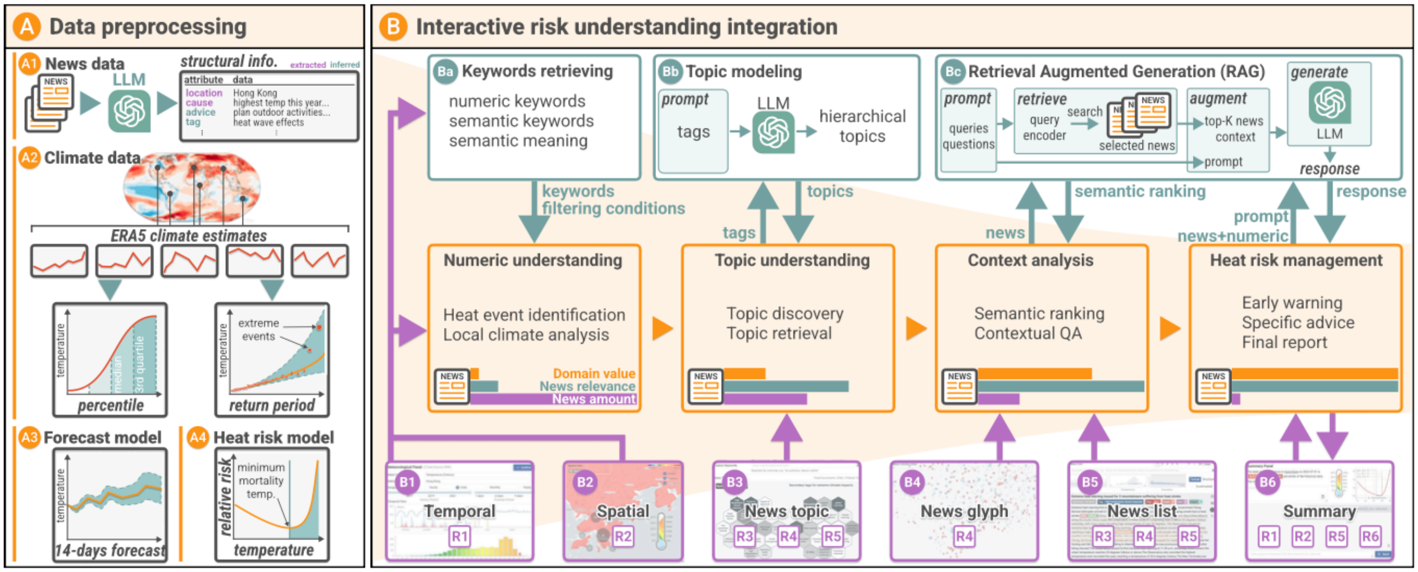


Fig. 2: The LLM-empowered pipeline contains two parts: data preprocessing (A) and interactive risk understanding integration (B). The data preprocessing involves extracting structural information using LLM (A1) and calculating climate indices (A2-4). In interactive risk understanding integration (B), heterogeneous understandings are integrated through keywords retrieving (Ba), topic modeling (Bb), and RAG (Bc). The interactive analysis process is supported by six views of *Havior* (B1-6) which fulfill the design requirements.

Figure 9: Figure from Li et al. (2024)

Follow-Me AI: Energy-Efficient User Interaction with Smart Environments

Saleh, A., Donta, P. K., Morabito, R., Motlagh, N. H., & Lovén, L. (2024). **Follow-Me AI: Energy-Efficient User Interaction with Smart Environments** (arXiv:2404.12486). arXiv. http://arxiv.org/abs/2404.12486

Abstract

This article introduces Follow-Me AI, a concept designed to enhance user interactions with smart environments, optimize energy use, and provide better control over data captured by these environments. Through AI agents that accompany users, Follow-Me AI negotiates data management based on user consent, aligns environmental controls as well as user communication and computes resources available in the environment with user preferences, and predicts user behavior to proactively adjust the smart environment. The manuscript illustrates this concept with a detailed example of Follow-Me AI in a smart campus setting, detailing the interactions with the building's management system for optimal comfort and efficiency. Finally, this article looks into the challenges and opportunities related to Follow-Me AI.

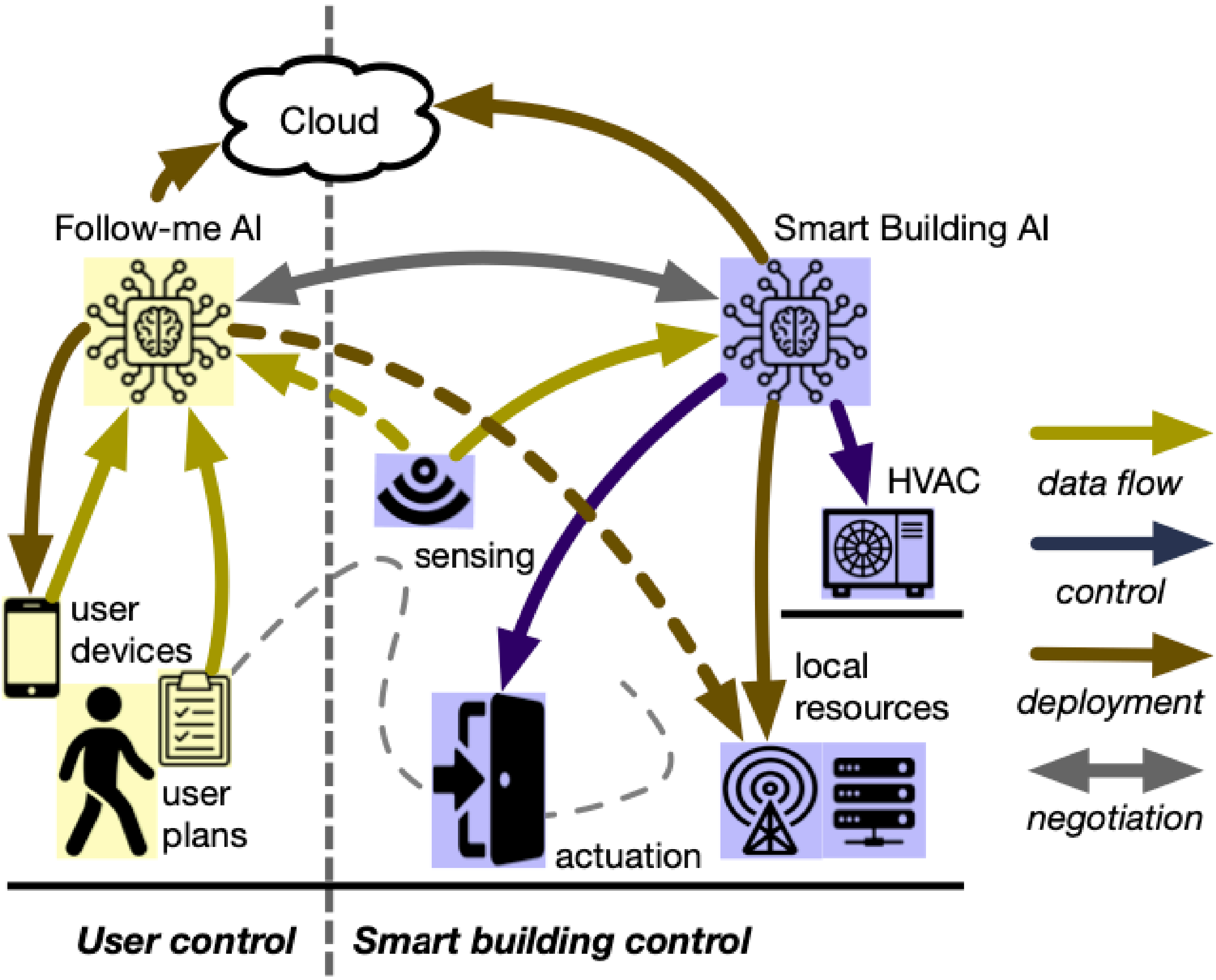
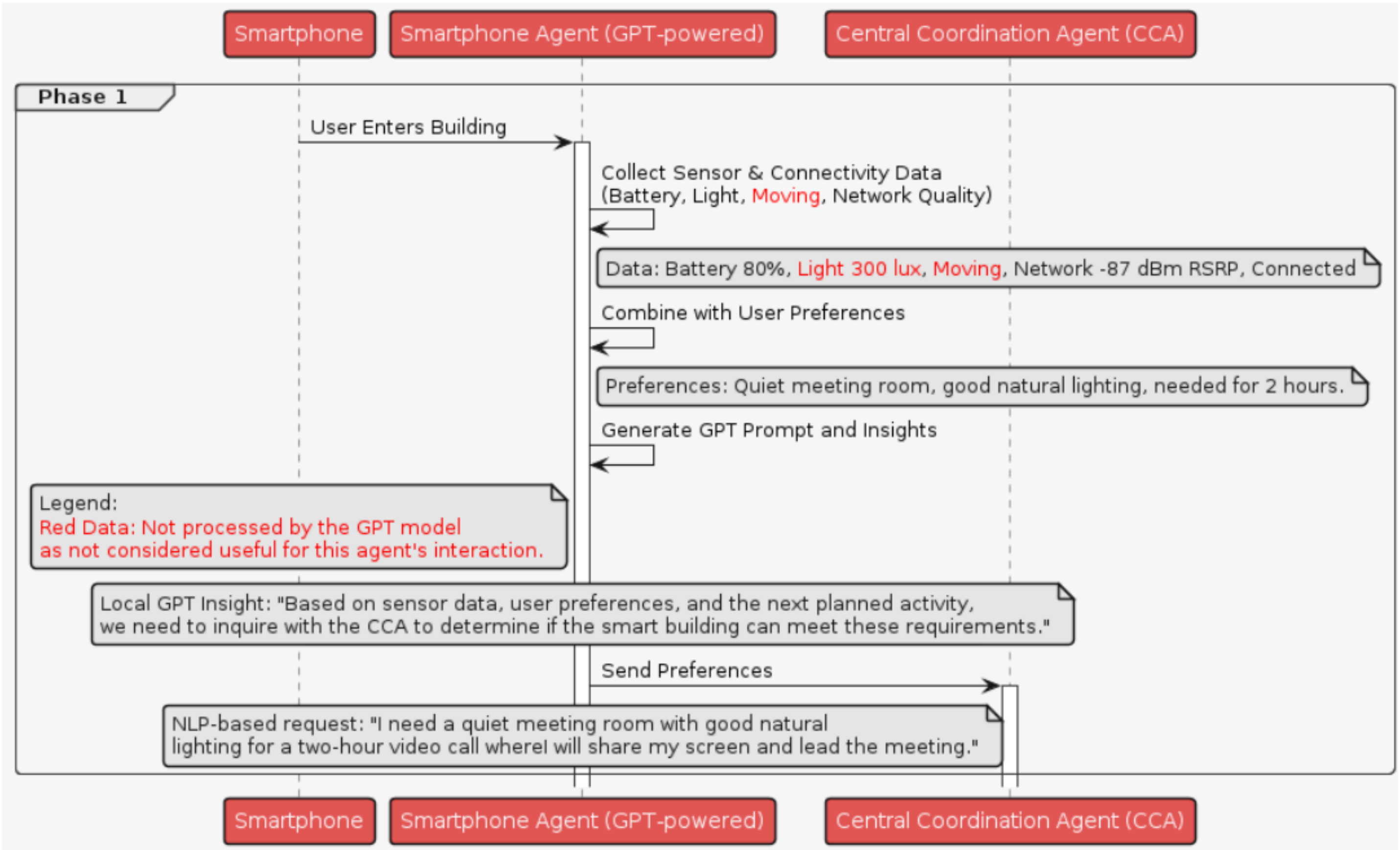


Fig. 1: Follow-Me AI for smart building interaction.


Smartphone
Smartphone Agent (GPT-powered)
Central Coordination Agent (CCA)
Phase 1
User Enters Building
Collect Sensor & Connectivity Data
(Battery, Light, Moving, Network Quality)
Data: Battery 80%, Light 300 lux, Moving, Network -87 dBm RSRP, Connected
Combine with User Preferences
Preferences: Quiet meeting room, good natural lighting, needed for 2 hours.
Generate GPT Prompt and Insights
Legend:
Red Data: Not processed by the GPT model
as not considered useful for this agent's interaction.
Local GPT Insight: "Based on sensor data, user preferences, and the next planned activity,
we need to inquire with the CCA to determine if the smart building can meet these requirements."
Send Preferences
NLP-based request: "I need a quiet meeting room with good natural
lighting for a two-hour video call whereI will share my screen and lead the meeting."
Smartphone
Smartphone Agent (GPT-powered)
Central Coordination Agent (CCA)

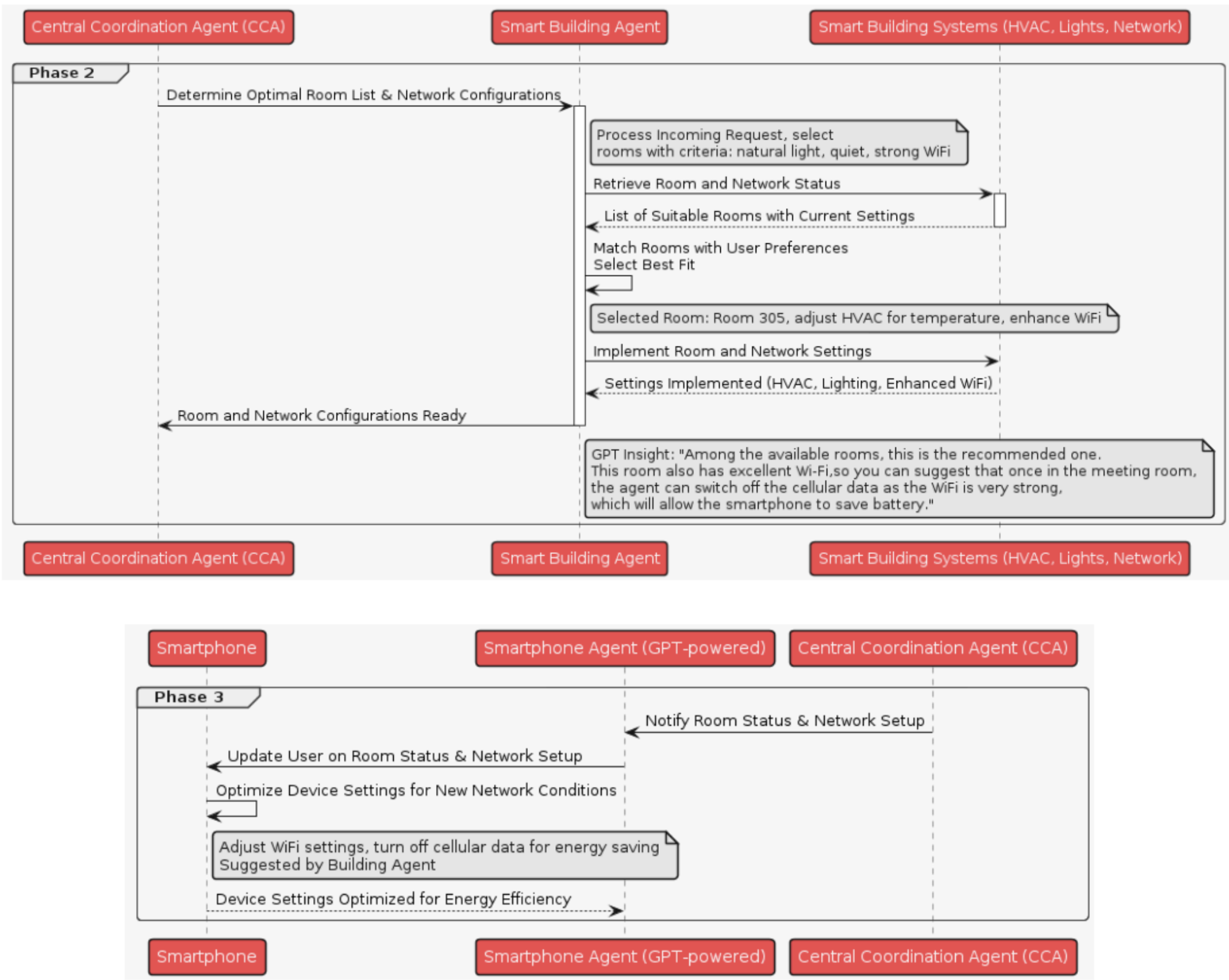


Fig. 3: Interaction diagram showing the data flow and agent collaboration within the smart campus environment, highlighting the roles of the Smartphone Agent, Central Coordination Agent, and Smart Building Agent in optimizing room conditions and device settings.

Figure 10: Figure from Saleh et al. (2024)

An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems

Yang, H., Siew, M., & Joe-Wong, C. (2024). **An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems** (arXiv:2403.16809). arXiv. http://arxiv.org/abs/2403.16809

Abstract

The increasing prevalence of Cyber-Physical Systems and the Internet of Things (CPS-IoT) applications and Foundation Models are enabling new applications that leverage real-time control of the environment. For example, real-time control of Heating, Ventilation and Air-Conditioning (HVAC) systems can reduce its usage when not needed for the comfort of human occupants, hence reducing energy consumption. Collecting realtime feedback on human preferences in such human-in-the-loop (HITL) systems, however, is difficult in practice. We propose the use of large language models (LLMs) to deal with the challenges of dynamic environments and difficult-to-obtain data in CPS optimization. In this paper, we present a case study that employs LLM agents to mimic the behaviors and thermal preferences of various population groups (e.g. young families, the elderly) in a shopping mall. The aggregated thermal preferences are integrated into an agent-in-the-loop based reinforcement learning algorithm AitL-RL, which employs the LLM as a dynamic simulation of the physical environment to learn how to balance between energy savings and occupant comfort. Our results show that LLMs are capable of simulating complex population movements within large open spaces. Besides, AitL-RLdemonstrates superior performance compared to the popular existing policy of set point control, suggesting that adaptive and personalized decision-making is critical for efficient optimization in CPS-IoT applications. Through this case study, we demonstrate the potential of integrating advanced Foundation Models like LLMs into CPS-IoT to enhance system adaptability and efficiency. The project's code can be found on our GitHub repository.

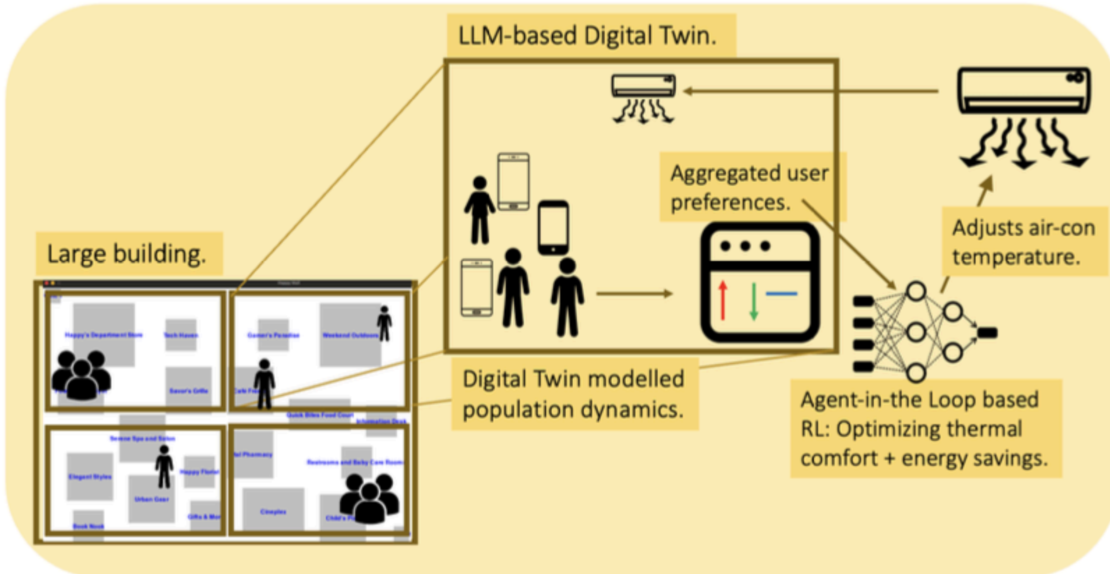


Fig. 1. The LLM-based Digital Twin Agent in the Loop Distributed Control (AitL-RL) Pipeline. The LLM-based digital twin simulates population behavior in the mall across the day, with multiple population groups such as "teen shoppers". Based on the simulation, user preferences are aggregated and input into the Agent-in-the-loop RL algorithm for offline training to optimize user comfort and energy savings.

Figure 11: Figure from Yang et al. (2024)

Can Private LLM Agents Synthesize Household Energy Consumption Data?

Almashor, M., & Miyashita, Y. (2024). **Can Private LLM Agents Synthesize Household Energy Consumption Data?** Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems, 664–668. https://doi.org/10.1145/3632775.3661993

Abstract

Reproducible science requires easy access to data, especially with the rise of data-driven and increasingly complex models used within energy research. Too often however, the data to reconstruct and verify purported solutions in publications is hidden due to some combination of commercial, legal, and sensitivity issues. This early work presents our initial efforts to leverage the recent advancements in Large Language Models (LLMs) to create usable and shareable energy datasets. In particular, we're utilising their mimicry of human behaviors, with the goal of extracting and exploring synthetic energy data through the simulation of LLM agents capable of interacting with and executing actions in controlled environments. We also analyse and visualise publicly available data in an attempt to create realistic but not quite exact copies of the originals. Our early results show some promise, with outputs that resemble the twin peak curves for household energy consumption. The hope is that our generalised approach can be used to easily replicate usable and realistic copies of otherwise secret or sensitive data.

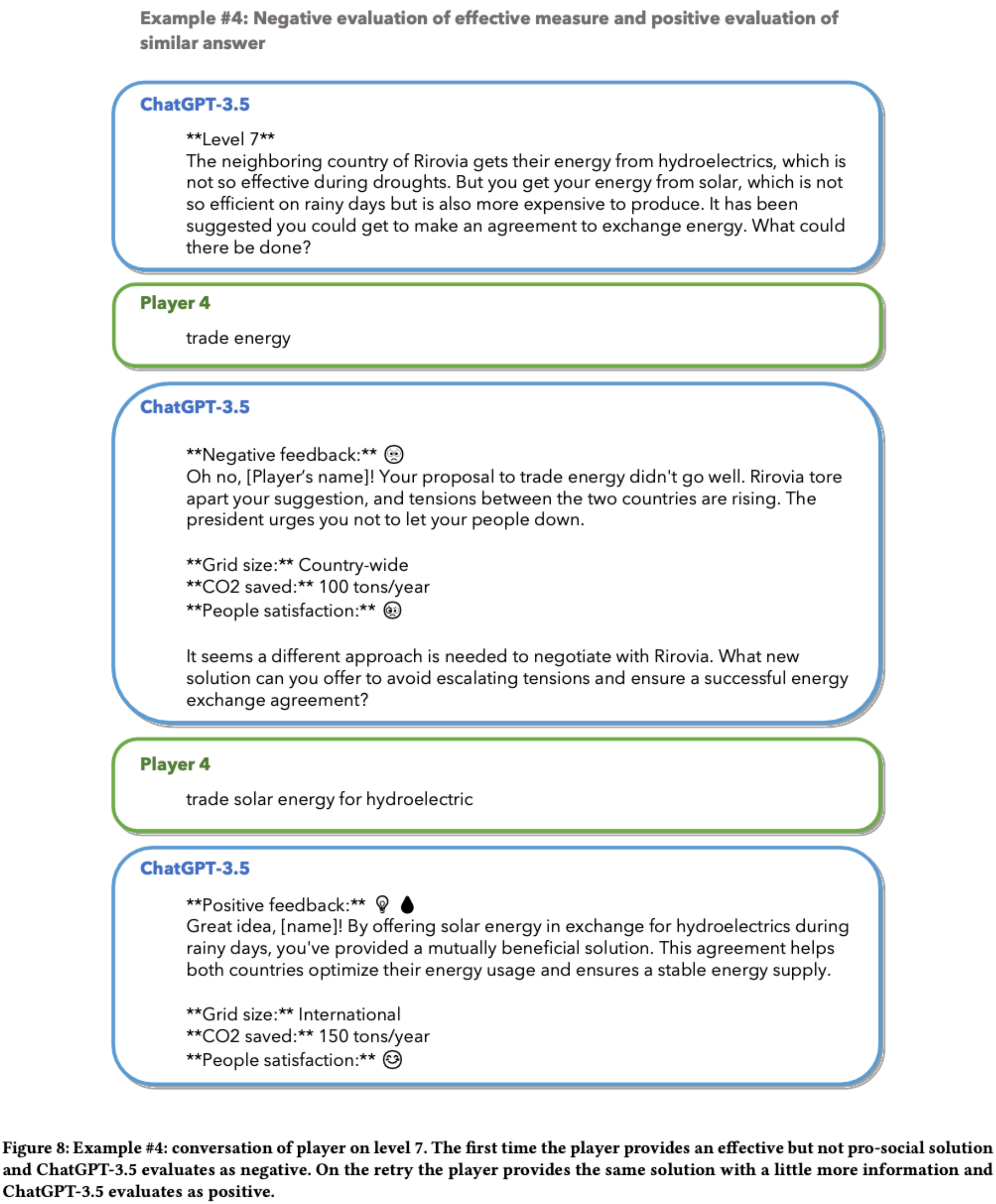
Almashor & Miyashita (2024)

Prompt-Gaming: A Pilot Study on LLM-Evaluating Agent in a Meaningful Energy Game.

Isaza-Giraldo, A., Bala, P., Campos, P. F., & Pereira, L. (2024). Prompt-Gaming: A Pilot Study on LLM-Evaluating Agent in a Meaningful Energy Game. Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3613905.3650774

Abstract

Building on previous work on incorporating large language models (LLM) in gaming, we investigate the possibility of implementing LLM as evaluating agents of open-ended challenges in serious games and its potential to facilitate a meaningful experience for the player. We contribute with a sustainability game prototype in a single natural language prompt about energy communities and we tested it with 13 participants inside ChatGPT-3.5. Two participants were already aware of energy communities before the game, and eight of the remaining 11 gained valuable knowledge about the specific topic. Comparing ChatGPT-3.5 evaluations of players' interaction with an expert's assessment, ChatGPT-3.5 correctly evaluated 81% of player's answers. Our results are encouraging and show the potential of using LLMs as mediating agents in educational games, while also allowing easy prototyping of games through natural language prompts.



Example #4: Negative evaluation of effective measure and positive evaluation of similar answer

ChatGPT-3.5

Level 7
The neighboring country of Rirovia gets their energy from hydroelectrics, which is not so effective during droughts. But you get your energy from solar, which is not so efficient on rainy days but is also more expensive to produce. It has been suggested you could get to make an agreement to exchange energy. What could there be done?

Player 4

trade energy

ChatGPT-3.5

Negative feedback:
Oh no, [Player's name]! Your proposal to trade energy didn't go well. Rirovia tore apart your suggestion, and tensions between the two countries are rising. The president urges you not to let your people down.

Grid size: Country-wide
CO2 saved: 100 tons/year
People satisfaction:

It seems a different approach is needed to negotiate with Rirovia. What new solution can you offer to avoid escalating tensions and ensure a successful energy exchange agreement?

Player 4

trade solar energy for hydroelectric

ChatGPT-3.5

Positive feedback:
Great idea, [name]! By offering solar energy in exchange for hydroelectrics during rainy days, you've provided a mutually beneficial solution. This agreement helps both countries optimize their energy usage and ensures a stable energy supply.

Grid size: International
CO2 saved: 150 tons/year
People satisfaction:

Figure 8: Example #4: conversation of player on level 7. The first time the player provides an effective but not pro-social solution and ChatGPT-3.5 evaluates as negative. On the retry the player provides the same solution with a little more information and ChatGPT-3.5 evaluates as positive.

Figure 12: Figure from Isaza-Giraldo et al. (2024)

Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry.

Lu, J., Tian, X., Zhang, C., Zhao, Y., Zhang, J., Zhang, W., Feng, C., He, J., Wang, J., & He, F. (2024). **Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry.** Energy and Built Environment. https://doi.org/10.1016/j.enbenv.2024.03.010

Abstract

Large language models (LLMs) have shown human-level capabilities in solving various complex tasks. However, it is still unknown whether state-of-the-art LLMs master sufficient knowledge related to heating, ventilation and air conditioning (HVAC) systems. It will be inspiring if LLMs can think and learn like professionals in the HVAC industry. Hence, this study investigates the performance of LLMs on mastering the knowledge and skills related to the HVAC industry by letting them take the ASHRAE Certified HVAC Designer examination, an authoritative examination in the HVAC industry. Three key knowledge capabilities are explored: recall, analysis and application. Twelve representative LLMs are tested such as GPT-3.5, GPT-4 and LLaMA. According to the results, GPT-4 passes the ASHRAE Certified HVAC Designer examination with scores from 74 to 78, which is higher than about half of human examinees. Besides, GPT-3.5 passes the examination twice out of five times. It demonstrates that some LLMs such as GPT-4 and GPT-3.5 have great potential to assist or replace humans in designing and operating HVAC systems. However, they still make some mistakes sometimes due to the lack of knowledge, poor reasoning capabilities and unsatisfactory equation calculation abilities. Accordingly, four future research directions are proposed to reveal how to utilize and improve LLMs in the HVAC industry: teaching LLMs to use design tools or software in the HVAC industry, enabling LLMs to read and analyze the operational data from HVAC systems, developing tailored corpuses for the HVAC industry, and assessing the performance of LLMs in real-world HVAC design and operation scenarios.

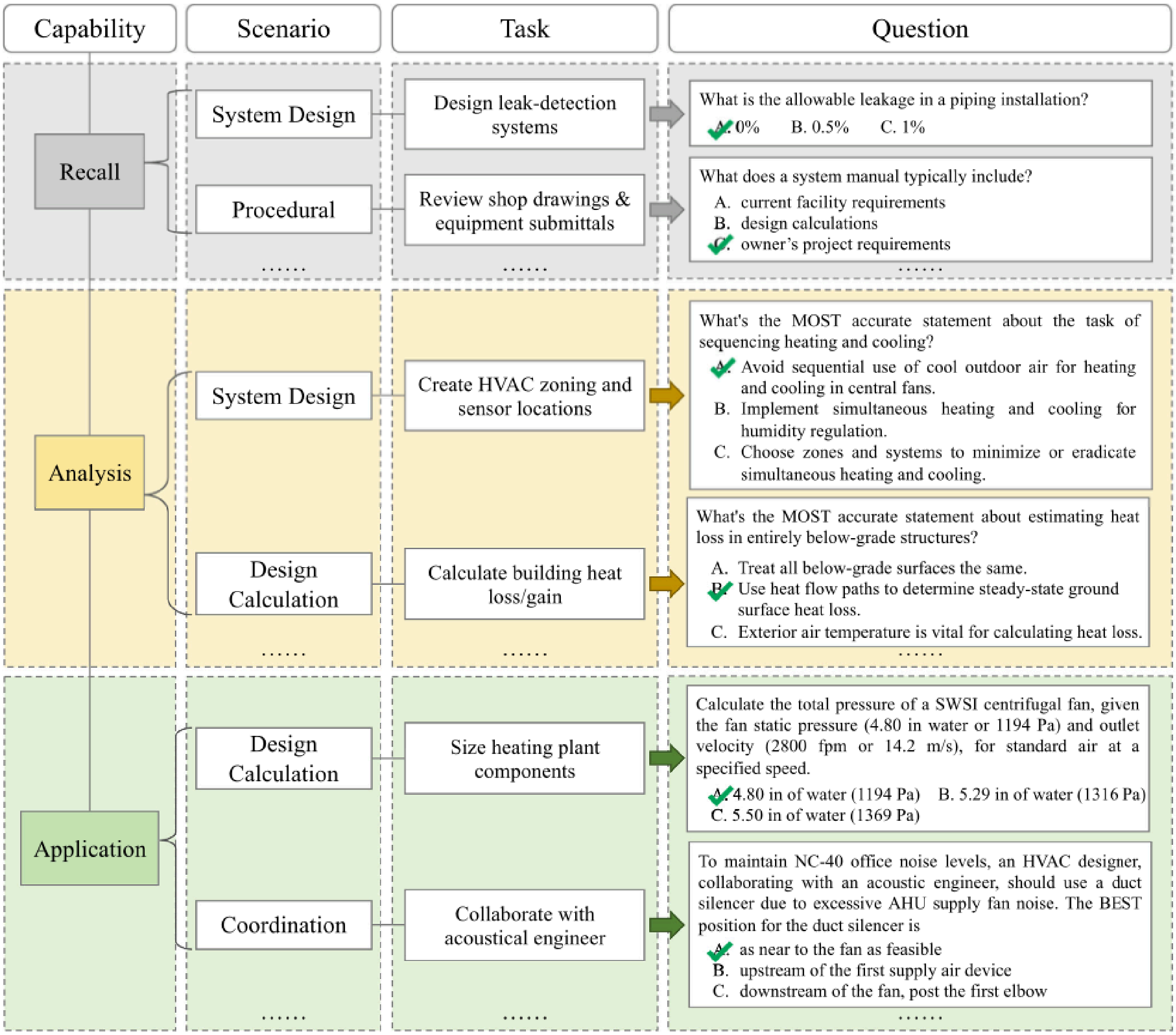


Fig. 3. Sample questions for professional capability evaluation of LLMs in the domain of HVAC systems.

Figure 13: Figure from Lu et al. (2024)

Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning.

Zhang, J., Zhang, C., Lu, J., & Zhao, Y. (2025). **Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning.** Applied Energy, 377, 124378. https://doi.org/10.1016/j.apenergy.2024.124378

Abstract

Large language models (LLMs) have exhibited great potential in fault diagnosis of heating, ventilation, and air conditioning systems. However, the fault diagnosis accuracy of LLMs is still unsatisfactory, due to the lack of effective diagnosis accuracy enhancement methods for LLMs. To fill this gap, this study proposes a LLM fine-tuning method supervised by data with fault and fault-free labels to enhance the fault diagnosis accuracy of LLMs. This method designs a LLM self-correction strategy to automatically generate a fine-tuning dataset based on the labeled data. The generated fine-tuning dataset is applied to fine-tune a LLM. Moreover, a data augmentation-based approach is put forward to adaptively update the fine-tuning dataset for iteratively developing a high-performance fine-tuned LLM. The proposed method is utilized to fine-tune the GPT-3.5 model using the air handling unit (AHU) fault dataset from the RP-1312 project. The results show that the diagnosis accuracy of the GPT-3.5 model is increased from 29.5 % to 100.0 % after model fine-tuning. Compared with the GPT-4 model, the fine-tuned GPT-3.5 model achieves a 31.1 % higher average diagnosis accuracy. The fine-tuned GPT-3.5 model is also applied to diagnose faults in two AHUs from another open-source dataset to verify the generalization ability of this model. The two AHUs have different system structures and sensor configurations compared to the AHU in the RP-1312 dataset, and this dataset is not utilized to fine-tune the GPT-3.5 model. The average diagnosis accuracy of the GPT-3.5 model is increased from 46.0 % to 99.1 % and from 38.8 % to 98.9 % for the faults in the two AHUs, respectively, after model fine-tuning. Furthermore, the proposed method is verified using two fault datasets from a variable air volume box and a chiller plant system. After fine-tuning the GPT-3.5 model using the two datasets, the average diagnosis accuracy of this model is increased from 33.0 % to 98.3 % for variable air volume box faults and from 36.0 % to 99.1 % for chiller plant system faults. This study provides an effective solution to the development of domain-specific LLMs for this domain.

Zhang et al. (2025)

Game of LLMs: Discovering Structural Constructs in Activities using Large Language Models.

Hiremath, S. K., & Plötz, T. (2024). **Game of LLMs: Discovering Structural Constructs in Activities using Large Language Models.** Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, 487–492. https://doi.org/10.1145/3675094.3678444

Abstract

Human Activity Recognition is a time-series analysis problem. A popular analysis procedure used by the community assumes an optimal window length to design recognition pipelines. However, in the scenario of smart homes, where activities are of varying duration and frequency, the assumption of a constant sized window does not hold. Additionally, previous works have shown these activities to be made up of building blocks. We focus on identifying these underlying building blocks–structural constructs, with the use of large language models. Identifying these constructs can be beneficial especially in recognizing short-duration and infrequent activities, which current systems cannot recognize. We also propose the development of an activity recognition procedure that uses these building blocks to model activities, thus helping the downstream task of activity monitoring in smart homes.

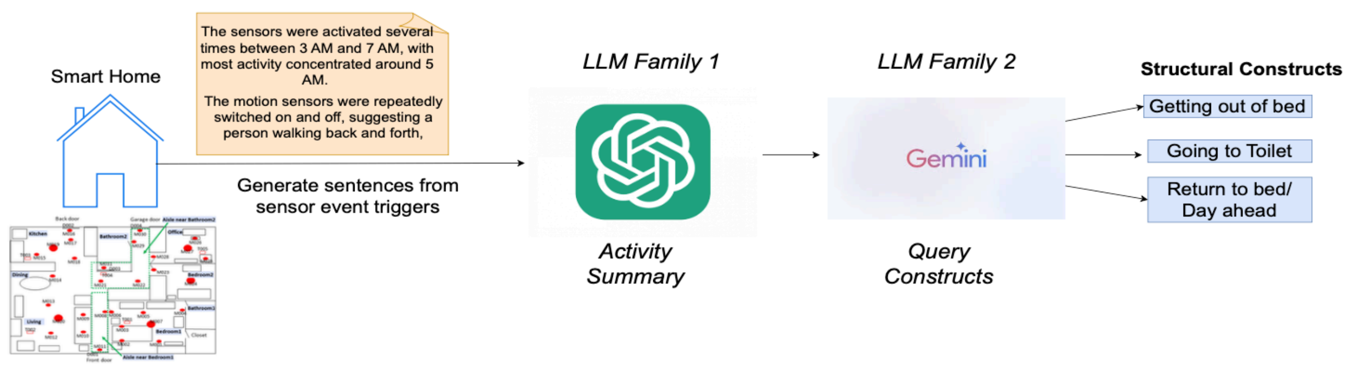


Figure 1: Overview of the proposed system. The proposed approach identifies the underlying structural concepts of activities observed in the smart home. First sentences detailing sensor event triggers are generated using information such as location and time of occurrence of activity [47]. Next a family of LLMs (GPT-4) is used to obtain a summarized version of varied instances of these activities. Subsequently, another family of LLMs (Gemini) is queried to identify the structural constructs.

Figure 14: Figure from Hiremath & Plötz (2024)

LLM-Based Open-Domain Integrated Task and Knowledge Assistants with Programmable Policies

Joshi, H., Liu, S., Chen, J., Weigle, R., & Lam, M. S. (2024). **LLM-Based Open-Domain Integrated Task and Knowledge Assistants with Programmable Policies** (arXiv:2407.05674). arXiv. http://arxiv.org/abs/2407.05674

Abstract

Programming LLM-based knowledge and task assistants that faithfully conform to developer-provided policies is challenging. These agents must retrieve and provide consistent, accurate, and relevant information to address user's queries and needs. Yet such agents generate unfounded responses ("hallucinate"). Traditional dialogue trees can only handle a limited number of conversation flows, making them inherently brittle. To this end, we present KITA - a programmable framework for creating task-oriented conversational agents that are designed to handle complex user interactions. Unlike LLMs, KITA provides reliable grounded responses, with controllable agent policies through its expressive specification, KITA Worksheet. In contrast to dialog trees, it is resilient to diverse user queries, helpful with knowledge sources, and offers ease of programming policies through its declarative paradigm. Through a real-user study involving 62 participants, we show that KITA beats the GPT-4 with function calling baseline by 26.1, 22.5, and 52.4 points on execution accuracy, dialogue act accuracy, and goal completion rate, respectively. We also release 22 real-user conversations with KITA manually corrected to ensure accuracy.

Joshi et al. (2024)

Large Language Models are Zero-Shot Recognizers for Activities of Daily Living

Civitarese, G., Fiori, M., Choudhary, P., & Bettini, C. (2024). **Large Language Models are Zero-Shot Recognizers for Activities of Daily Living** (arXiv:2407.01238). arXiv. http://arxiv.org/abs/2407.01238

Abstract

The sensor-based recognition of Activities of Daily Living (ADLs) in smart home environments enables several applications in the areas of energy management, safety, well-being, and healthcare. ADLs recognition is typically based on deep learning methods requiring large datasets to be trained. Recently, several studies proved that Large Language Models (LLMs) effectively capture common-sense knowledge about human activities. However, the effectiveness of LLMs for ADLs recognition in smart home environments still deserves to be investigated. In this work, we propose ADL-LLM, a novel LLM-based ADLs recognition system. ADLLLM transforms raw sensor data into textual representations, that are processed by an LLM to perform zero-shot ADLs recognition. Moreover, in the scenario where a small labeled dataset is available, ADL-LLM can also be empowered with few-shot prompting. We evaluated ADL-LLM on two public datasets, showing its effectiveness in this domain.

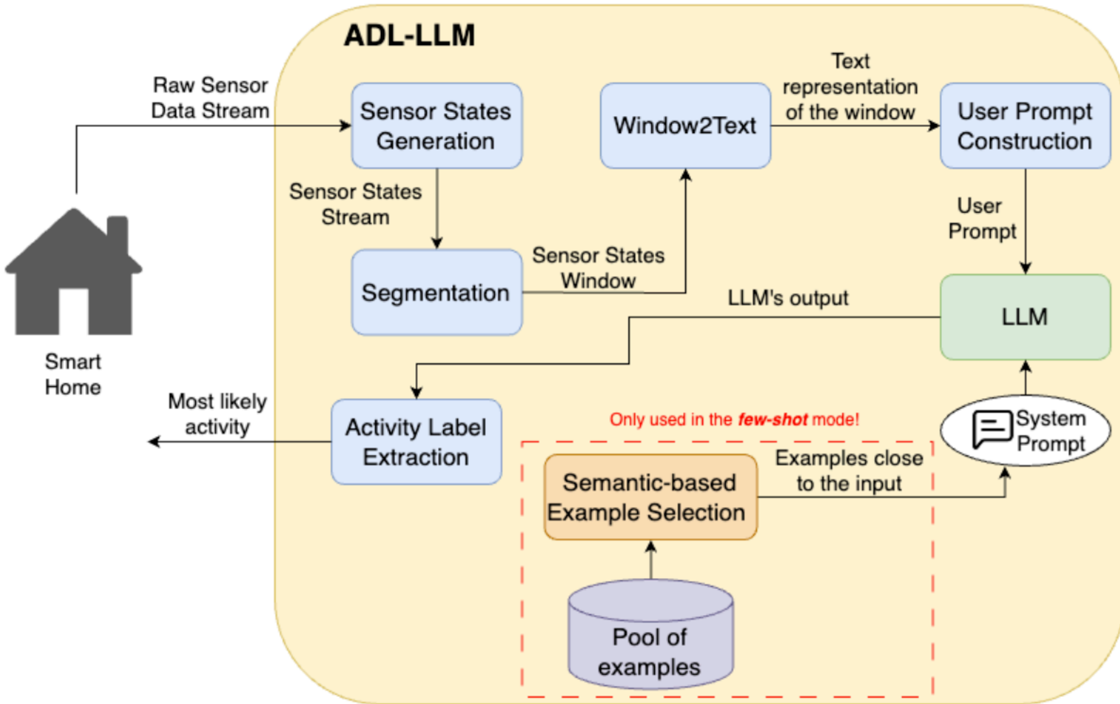


Fig. 1. Overall architecture of ADL-LLM. When the pool of examples is empty ADL-LLM, acts as a ***zero-shot*** ADLs recognition method. Otherwise, it is a ***few-shot*** approach.

State	t_s	t_e
OnCouch	2011-12-02 12:37:47	2011-12-02 14:46:29
FridgeON	2011-12-02 14:46:59	2011-12-02 14:47:08
NearStove	2011-12-02 14:47:18	2011-12-02 15:05:06
MicrowaveON	2011-12-02 14:47:22	2011-12-02 14:50:19

Fig. 2. An example of Window2Text in action on the UCI ADL dataset

Figure 15: Figures from Civitarese et al. (2024)

Large Language Models for Power Scheduling: A User-Centric Approach

Mongaillard, T., Lasaulce, S., Hicheur, O., Zhang, C., Bariah, L., Varma, V. S., Zou, H., Zhao, Q., & Debbah, M. (2024). **Large Language Models for Power Scheduling: A User-Centric Approach** (arXiv:2407.00476). arXiv. http://arxiv.org/abs/2407.00476

Abstract

While traditional optimization and scheduling schemes are designed to meet fixed, predefined system requirements, future systems are moving toward user-driven approaches and personalized services, aiming to achieve high quality-of-experience (QoE) and flexibility. This challenge is particularly pronounced in wireless and digitalized energy networks, where users' requirements have largely not been taken into consideration due to the lack of a common language between users and machines. The emergence of powerful large language models (LLMs) marks a radical departure from traditional system-centric methods into more advanced user-centric approaches by providing a natural communication interface between users and devices. In this paper, for the first time, we introduce a novel architecture for resource scheduling problems by constructing three LLM agents to convert an arbitrary user's voice request (VRQ) into a resource allocation vector. Specifically, we design an LLM intent recognition agent to translate the request into an optimization problem (OP), an LLM OP parameter identification agent, and an LLM OP solving agent. To evaluate system performance, we construct a database of typical VRQs in the context of electric vehicle (EV) charging. As a proof of concept, we primarily use Llama 3 8B. Through testing with different prompt engineering scenarios, the obtained results demonstrate the efficiency of the proposed architecture. The conducted performance analysis allows key insights to be extracted. For instance, having a larger set of candidate OPs to model the real-world problem might degrade the final performance because of a higher recognition/OP classification noise level. All results and codes are open source.

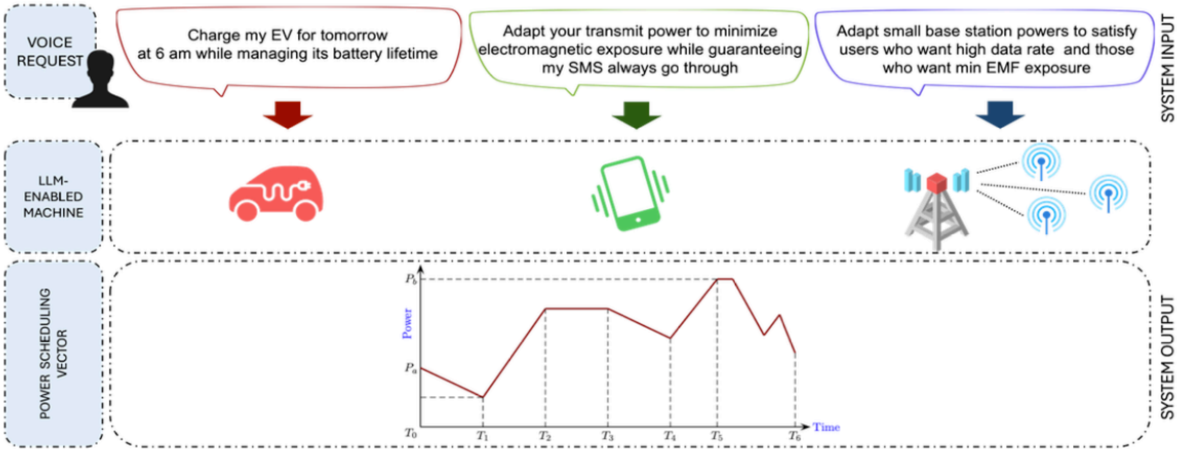


Fig. 1: Use-Cases of the Proposed Intelligent Power Scheduling System

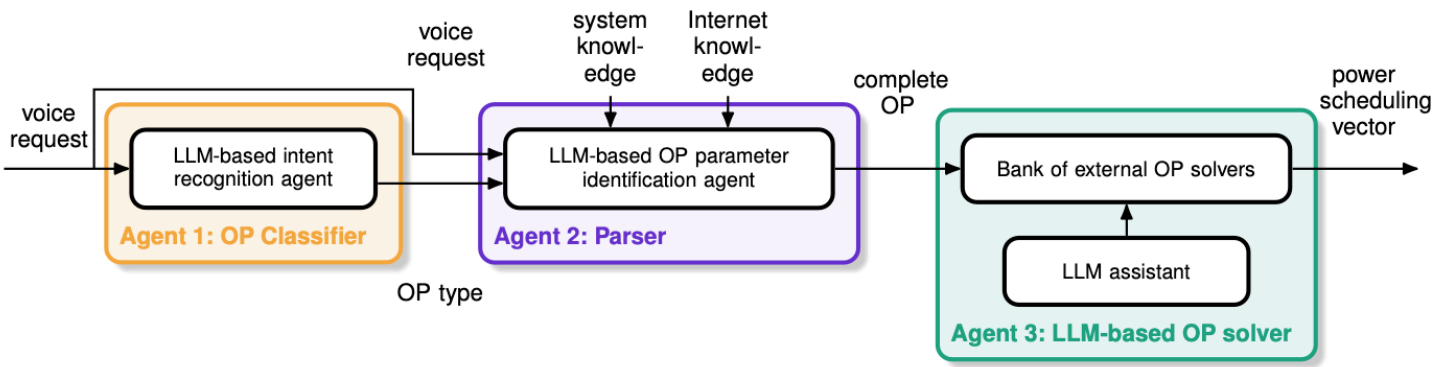


Fig. 2: Proposed multi-agent architecture for a voice request to power scheduling vector converter (VRQ2Vec)

You are an EXPERT in optimization problems in a smart home context. You have been trained to classify user requests in terms of EV charging into their corresponding optimization problem class. The FINAL GOAL is to provide the user with a power consumption vector that will satisfy the request.

Follow the different STEPS:
- Identify the performance metric required by the user using your knowledge. [...]
- Find the closest usual problem based on your knowledge.
- Select the corresponding optimization problem class.

When the user gives you a request to process, generate a FUNCTION CALL in the following format [...] Do not forget to generate the function call, it is really important [...] PRIORITIZE requests with common sense. Common sense and logics are crucial. For example [...]

Be very attentive to the KNOWLEDGE FILES. [...]

Your ANSWER has to contain [...]

A Recommendation System for Prosumers Based on Large Language Models.

Oprea, S.-V., & Bâra, A. (2024). **A Recommendation System for Prosumers Based on Large Language Models.** Sensors, 24(11), Article 11. https://doi.org/10.3390/s24113530

Abstract

As modern technologies, particularly home assistant devices and sensors, become more integrated into our daily lives, they are also making their way into the domain of energy management within our homes. Homeowners, now acting as prosumers, have access to detailed information at 15-min or even 5-min intervals, including weather forecasts, outputs from renewable energy source (RES)-based systems, appliance schedules and the current energy balance, which details any deficits or surpluses along with their quantities and the predicted prices on the local energy market (LEM). The goal for these prosumers is to reduce costs while ensuring their home's comfort levels are maintained. However, given the complexity and the rapid decision-making required in managing this information, the need for a supportive system is evident. This is particularly true given the routine nature of these decisions, highlighting the potential for a system that provides personalized recommendations to optimize energy consumption, whether that involves adjusting the load or engaging in transactions with the LEM. In this context, we propose a recommendation system powered by large language models (LLMs), Scikit-llm and zero-shot classifiers, designed to evaluate specific scenarios and offer tailored advice for prosumers based on the available data at any given moment. Two scenarios for a prosumer of 5.9 kW are assessed using candidate labels, such as Decrease, Increase, Sell and Buy. A comparison with a content-based filtering system is provided considering the performance metrics that are relevant for prosumers.

Oprea & Bâra (2024)

A conversational agent for creating automations exploiting large language models. Personal and Ubiquitous Computing.

Gallo, S., Paternò, F., & Malizia, A. (2024). **A conversational agent for creating automations exploiting large language models. Personal and Ubiquitous Computing.** https://doi.org/10.1007/s00779-024-01825-5

Abstract

The proliferation of sensors and smart Internet of Things (IoT) devices in our everyday environments is reshaping our interactions with everyday objects. This change underlines the need to empower non-expert users to easily configure the behaviour of these devices to align with their preferences and habits. At the same time, recent advances in generative transformers, such as ChatGPT, have opened up new possibilities in a variety of natural language processing tasks, enhancing reasoning capabilities and conversational interactions. This paper presents RuleBot + +, a conversational agent that exploits GPT-4 to assist the user in the creation and modification of trigger-action automations through natural language. After an introduction to motivations and related work, we present the design and implementation of RuleBot + + and report the results of the user test in which users interacted with our solution and Home Assistant, one of the most used open-source tools for managing smart environments.

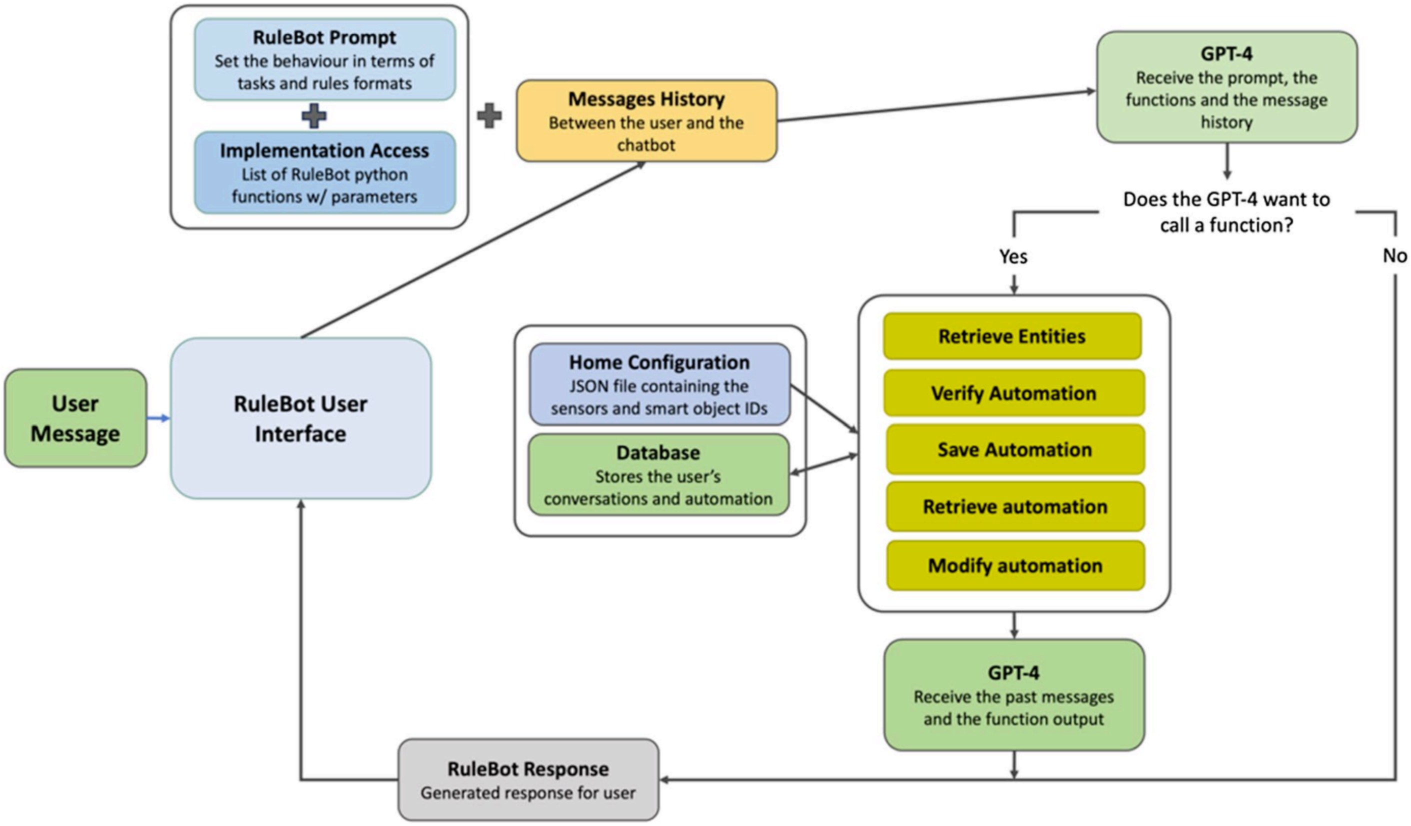


Fig. 1 The RuleBot++ architecture

Fig. 2 RuleBot++ web user interface

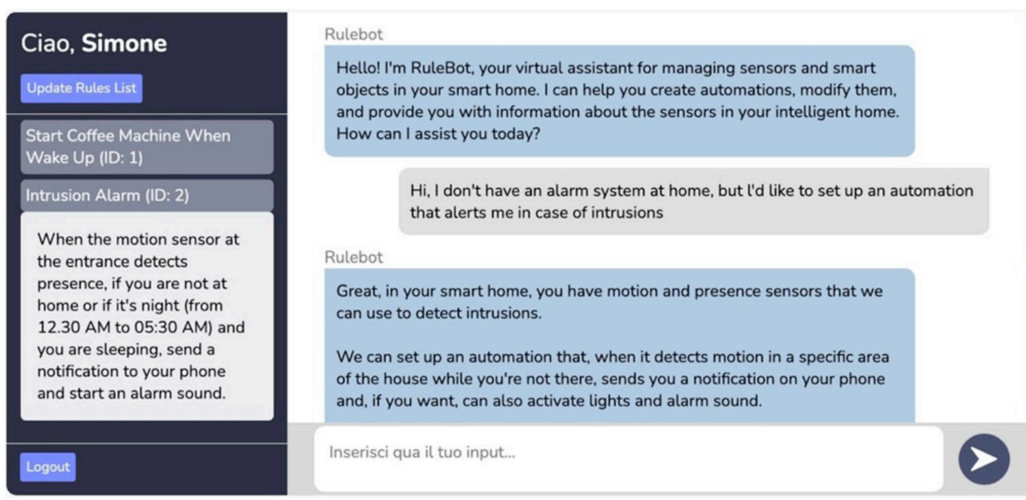


Figure 17: Figures from Gallo et al. (2024)

References

Almashor, M., & Miyashita, Y. (2024). Can Private LLM Agents Synthesize Household Energy Consumption Data? *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, 664–668. https://doi.org/10.1145/3632775.3661993

Calò, T., & De Russis, L. (2024). Enhancing smart home interaction through multimodal command disambiguation. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-024-01827-3

Civitarese, G., Fiori, M., Choudhary, P., & Bettini, C. (2024). *Large Language Models are Zero-Shot Recognizers for Activities of Daily Living* (arXiv:2407.01238). arXiv. https://arxiv.org/abs/2407.01238

Gallo, S., Paternò, F., & Malizia, A. (2024). A conversational agent for creating automations exploiting large language models. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-024-01825-5

Giudici, M., Padalino, L., Paolino, G., Paratici, I., Pascu, A. I., & Garzotto, F. (2024). Designing Home Automation Routines Using an LLM-Based Chatbot. *Designs*, *8*(3), 43. https://doi.org/10.3390/designs8030043

Hiremath, S. K., & Plötz, T. (2024). Game of LLMs: Discovering Structural Constructs in Activities using Large Language Models. *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 487–492. https://doi.org/10.1145/3675094.3678444

Isaza-Giraldo, A., Bala, P., Campos, P. F., & Pereira, L. (2024). Prompt-Gaming: A Pilot Study on LLM-Evaluating Agent in a Meaningful Energy Game. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3613905.3650774

Jin, M., Sel, B., Hardeep, F., & Yin, W. (2023). *A Human-on-the-Loop Optimization Autoformalism Approach for Sustainability* (arXiv:2308.10380). arXiv. https://arxiv.org/abs/2308.10380

Joshi, H., Liu, S., Chen, J., Weigle, R., & Lam, M. S. (2024). *LLM-Based Open-Domain Integrated Task and Knowledge Assistants with Programmable Policies* (arXiv:2407.05674). arXiv. https://arxiv.org/abs/2407.05674

King, E., Yu, H., Lee, S., & Julien, C. (2024). Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–38. https://doi.org/10.1145/3643505

King, E., Yu, H., Vartak, S., Jacob, J., Lee, S., & Julien, C. (2024). *Thoughtful Things: Building Human-Centric Smart Devices with Small Language Models* (arXiv:2405.03821). arXiv. https://arxiv.org/abs/2405.03821

Li, H., Kam-Kwai, W., Luo, Y., Chen, J., Liu, C., Zhang, Y., Lau, A. K. H., Qu, H., & Liu, D. (2024). *Save It for the "Hot" Day: An LLM-Empowered Visual Analytics System for Heat Risk Management* (arXiv:2406.03317). arXiv. https://arxiv.org/abs/2406.03317

Lu, J., Tian, X., Zhang, C., Zhao, Y., Zhang, J., Zhang, W., Feng, C., He, J., Wang, J., & He, F. (2024). Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry. *Energy and Built Environment*. https://doi.org/10.1016/j.enbenv.2024.03.010

Mongaillard, T., Lasaulce, S., Hicheur, O., Zhang, C., Bariah, L., Varma, V. S., Zou, H., Zhao, Q., & Debbah, M.

(2024). *Large Language Models for Power Scheduling: A User-Centric Approach* (arXiv:2407.00476). arXiv. https://arxiv.org/abs/2407.00476

Oprea, S.-V., & Bâra, A. (2024). A Recommendation System for Prosumers Based on Large Language Models. *Sensors*, *24*(11), 3530. https://doi.org/10.3390/s24113530

Rey-Jouanchicot, J., Bottaro, A., Campo, E., Bouraoui, J.-L., Vigouroux, N., & Vella, F. (2024). *Leveraging Large Language Models for enhanced personalised user experience in Smart Homes* (arXiv:2407.12024). arXiv. https://arxiv.org/abs/2407.12024

Rivkin, D., Hogan, F., Feriani, A., Konar, A., Sigal, A., Liu, X., & Dudek, G. (2024). AIoT Smart Home via Autonomous LLM Agents. *IEEE Internet of Things Journal*, 1–1. https://doi.org/10.1109/JIOT.2024.3471904

Saleh, A., Donta, P. K., Morabito, R., Motlagh, N. H., & Lovén, L. (2024). *Follow-Me AI: Energy-Efficient User Interaction with Smart Environments* (arXiv:2404.12486). arXiv. https://arxiv.org/abs/2404.12486

Yang, H., Siew, M., & Joe-Wong, C. (2024). *An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems* (arXiv:2403.16809). arXiv. https://arxiv.org/abs/2403.16809

Yin, Z., Zhang, M., & Kawahara, D. (2024). *Harmony: A Home Agent for Responsive Management and Action Optimization with a Locally Deployed Large Language Model* (arXiv:2410.14252). arXiv. https://arxiv.org/abs/2410.14252

Zhang, J., Zhang, C., Lu, J., & Zhao, Y. (2025). Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning. *Applied Energy*, *377*, 124378. https://doi.org/10.1016/j.apenergy.2024.124378