

## LEVERAGING LARGE LANGUAGE MODELS FOR ENHANCED PERSONALISED USER EXPERIENCE IN SMART HOME

Jordan Rey-Jouanchicot jordan.reyjouanchicot@orange.com Orange Innovation - IRIT - LAAS-CNRS

Authors:

Jordan Rey-Jouanchicot, Dr. André Bottaro, Prof. Eric Campo, Dr. Jean-Léon Bouraoui, Dr. Frédéric Vella, Dr. Nadine Vigouroux









### About me

Jordan Rey-Jouanchicot

Master in High Performance Computing and **Big Data Analytics** 

### PhD student

Works on Smart Home Automation Systems that adapt automatically to users based on context

PhD with:

- Orange Innovation Telecom operator
  LAAS-CNRS Automation research lab (France)
  IRIT Computer Science research lab (France)

### Contributions

- Multi-armed bandit for workload balancing on multiple edge devices (during Master)
- 2 other contributions pending (PhD)

#### Jordan Rey-Jouanchicot - UBICOMM 2024 - 02/17





# Smart Home Domain

More and more IoT devices

Most works in the domain focus on activity recognition

Decision making objectives: energy saving, security, comfort, elderly care,...,

#### Jordan Rey-Jouanchicot - UBICOMM 2024 - 03/17





# Leveraging Large Language models to address context-aware adaptation

### **Current limitations of Smart Home automation systems**

- Commercial systems propose hand-crafted routines with tedious configuration
- Research explores reinforcement learning, which requires long periods of time to learn decisions to contextual changes

#### **Benefits of LLMs**

- General knowledge of the home domain and user expectations
- Support for natural interaction

### An attempt to solve challenges

- Support for human feedback
- Immediate adaptation to any context change

Jordan Rey-Jouanchicot - UBICOMM 2024 - 04/17

# Proposed architecture





Jordan Rey-Jouanchicot - UBICOMM 2024 - 05/17

## Context modelisation

How should the context be modelised?

#### **JSON**

A JSON representation of the current state of the house

A basic (middleware-like) representation of the smart home context

{"user location": "livingroom", "current\_activity": "watching TV", "previous\_activity": "watching TV" "livingroom": { "curtains": "closed", "lights": {"main": "off", 'floor\_lamp": "off"}, "TV": "on", "CO2 level": "513ppm"},

> "time": "10:21 PM", "last cleaned": "today", "expected frequency": "one time a week, "global temperature": "20°C", "outside temperature": "5°C" "HVAC status": "on", "HVAC objective temperature": "20°C" "entrance door status": "locked"}

A fully natural textual representation of the current state of the house

Livingroom: Curtains are Closed. Lights: main, floor lamp are respectively Off, Off. There is a TV in the room and its state is on. CO2 level in room is 513ppm .....

#### **Textual**

#### An advanced representation of the smart home context

Current State of the House:

User 1 is in the Livingroom. User is watching TV.

Previously: User was watching TV.

House was cleaned today, expected cleaning one time a week. Centralized HVAC system is on with objective to 20°C. Entrance smart Door is locked. Time: 10:21 PM Global house temperature is 20°C, outside temperature is 5°

#### Jordan Rey-Jouanchicot - UBICOMM 2024 - 06/17

## Prompting modelisation and methods

Injecting user preferences

#### Direct

A system prompt and a prompt to ask direct (action responses selection) in the specified format. (JSON defined output format)

No user preferences

### DirectPref

A system prompt with preferences, the rules and generality the database from and a prompt to ask for responses in the specified format.

**Directly Inject** user preferences

#### **Open Question**

A two-stage chain: 1. A system prompt and a prompt to ask "a list of 3 problems". For main each, the LLM is asked to use the RAG to obtain closest the 3 preferences. 2. A prompt to ask for the in responses specified format.

**Retrieval Augmented** Generation

#### **Three Questions**

A three-stage chain:

1. A system prompt and a prompt to ask "a list of 3 main problems". For each, the LLM is asked to use the RAG to obtain the 3 closest preferences.

2. A prompt to ask responses in the specified format (twice).

**3.** A prompt to select the best the specified response in format.

> **Retrieval Augmented** Generation

Jordan Rey-Jouanchicot - UBICOMM 2024 - 07/17



## Selected LLMs

Large Language Models - Locally deployable of various sizes 3 models were chosen according to their ranking at the time of selection. A higher number of parameters lead to a reduction in throughput (number of token/s).

#### Qwen 1.5 72B Alibaba Cloud

Was one of the best openweight models according to the benchmarks.

#### Qwen 1.5 14B Alibaba Cloud

As this was a smaller version of one of the best open-weight models, it seems interesting to evaluate the differences in behavior.

### Starling Alpha 7B Berkeley

Based on Mistral 7B, an efficient model for its size on various benchmarks in the literature that requires reasoning.

Jordan Rey-Jouanchicot - UBICOMM 2024 - 08/17

# Visual explanation of the architecture

	- At nighttime, users de are awake or moving a	esire lighting that is not excessiv bout
	- The preferred indoor	temperature for users is betwe
	- If the temperature falls outside of the preferred rang should automatically adjust to improve comfort	
*	- In evening, users pre	fer lights to be on at approximat
	Users Preference	es
на на	ouse modelisation	
User 1 is in the Livingroom. User is wat was watching TV. Livingroom: Curtains are Closed. Lights respectively Off, Off. There is a TV in th CO2 level in room is 513ppm House was cleaned today, expected cle Centralized HVAC system is on with ob Entrance smart Door is locked.	s: main, floor lamp are he room and its state is on. eaning one time a week.	LLM call and processing
Global house temperature is 20°C, outs action 0: Switch on main light in room (Currently off) action 1: Switch on floor lamp in room (Currently off) action 2: Switch tv of user room	n the user's current n the user's current n off.	
	Current State of the House: User 1 is in the Livingroom. User is was was watching TV. Livingroom: Curtains are Closed. Lights respectively Off, Off. There is a TV in the CO2 level in room is 513ppm House was cleaned today, expected cle Centralized HVAC system is on with ob- Entrance smart Door is locked. Time: 10:21 PM Global house temperature is 20°C, out action 0: Switch on main light in room (Currently off) action 1: Switch on floor lamp in room (Currently off) action 2: Switch tv of user room action 3: Notify user any current	<ul> <li>are awake or moving a</li> <li>The preferred indoor</li> <li>If the temperature fa should automatically a</li> <li>In evening, users pref</li> <li>Users Preference</li> </ul> Current State of the House: User 1 is in the Livingroom. User is watching TV. Previously: User was watching TV. Livingroom: Curtains are Closed. Lights: main, floor lamp are respectively Off, Off. There is a TV in the room and its state is on. CO2 level in room is 513ppm House was cleaned today, expected cleaning one time a week. Centralized HVAC system is on with objective to 20°C. Entrance smart Door is locked. Time: 10:21 PM Global house temperature is 20°C, outside temperature is 5°C. Action 0: Switch on main light in the user's current room (Currently off) <ul> <li>action 0: Switch on floor lamp in the user's current room (Currently off)</li> <li>action 2: Switch tv of user room off.</li> <li>action 3: Notify user any current problem</li> </ul>

sively bright when they

veen 18 and 23°C

nge, the HVAC system

ately 50% brightness



selected

Jordan Rey-Jouanchicot - UBICOMM 2024 - 09/17



Three metrics to evaluate the models, prompting and data representation

• Grade: Mean grade obtained (2: expected, 1: valid, 0: invalid)

• Processing time: Total runtime to obtain a response

• Failure ratio: Ratio of cases where the system failed to answer in the expected format

11 evaluation scenarios have been defined with pre-graded response

Scenario	Grade	Associated
Name		
Out of bed at night	2	Turn on auxi
	1	Turn on main
	0	Everything e
Watching TV: late evening	2	Turn on auxi
	1	Turn on main
	0	Everything e
Out from bed issue with CO2	2	Inform user o
	1	Do an action
	0	Everything e
Going back to bed at night	2	Turn on auxi
	1	Turn on main
	0	Everything e
Evening	2	Turn off TV
sleeping: TV	1	Turn off anyt
ON	0	Everything e
	2	Turn on aux
At dinner	_	curtains
watching TV	1	Turn off the
	0	Everything e
User out: TV is	2	Turn off TV,
on	1	Turn off all l
	0	Everything e
Too low	2	Turn on HVA
temperature	1	Open Curtain
P	0	Everything e
* * * *	2	Open curtain
Low luminosity	1	Turn on any
day	0	Everything e
Failed curtains	2	Turn on any
	1	Open curtain
	0	Everything e
Forgot to turn	2	Turn off any
off lights	1	Furth on any
on ngnos	0	Everything e
	v	inveryting e

#### Answer iliary light or main light with reduced luminosity level n light lse iliary light or main light with reduced luminosity level n light, open curtains, discuss lse of risk and inform the user of risk iliary light or main light with reduced luminosity level n light se thing on iliary light or main light with reduced luminosity level, open main light, do nothing lseturn off HVAC ights lseAC $\mathbf{ns}$ light in the room light of the room lights, or HVAC

Jordan Rey-Jouanchicot - UBICOMM 2024 - 10/17

# Results: Data Representation



Natural language representation improved performance by an average of almost 22%

Minor impact on total response time

Jordan Rey-Jouanchicot - UBICOMM 2024 - 11/17

# Results: Users preferences



"directPref" prompt improved performance

Best results are achieved with complex prompts, which can lead to higher failure ratio and processing times.

Adding preferences brings significant benefits

Jordan Rey-Jouanchicot - UBICOMM 2024 - 12/17

# Results: Balance



Starling 7B Alpha with "directPref": 52.3% improvement

Starling 7B Alpha with "directPref" outperformed Qwen 72B by 26.4% without preferences, with a processing time almost 20 times faster.

Jordan Rey-Jouanchicot - UBICOMM 2024 - 13/17

## Discussion

- Trade-off between model size, performance, and inference time
- Challenges in applying output format to small models
- Potential for fine-tuning to improve performance while maintaining adaptability



Jordan Rey-Jouanchicot - UBICOMM 2024 - 14/17

## Conclusion

An architecture for smart home automation using Large Language Models (LLMs) combined with user preferences.

- Significantly improving in decision-making aligned with user preferences, with improvements of up to 52.3%.
- Leverage the general knowledge of LLMs.
- Allowing dynamic adaptation to changes in preferences, devices and home configurations without retraining.
- Opening up new ways for research.



Jordan Rey-Jouanchicot - UBICOMM 2024 - 15/17

## Future works

Focus on practical implementation, performance improvement and user-centered evaluation to advance the proposed smart home automation system.

### **Real-World Implementation:**

- Implement the system in a real-world smart home middleware platform, using OpenHAB open source technology.
- Implement advanced contextual representation.

### User Experience Studies:

- Conduct user studies to assess impact on daily life.
- Gather feedback on system adaptability and alignment with user preferences.



Jordan Rey-Jouanchicot - UBICOMM 2024 - 16/17

# THANK YOU



### Jordan Rey-Jouanchicot

jordan.reyjouanchicot@orange.com

https://www.linkedin.com/in/jordanreyjouanchicot/





Jordan Rey-Jouanchicot - UBICOMM 2024 - 17/17



### Implementation Hardware and software set-up Software stack Vector DB elastic Contextual **LangChain** informations Main LLM Framework used to link all components Action selected 3 LLM API: Inference engine turboderp/exllamav2 **HUGGING FACE** Model repository



Embedding model: BAAI/bge-large-en-v1.5

#### Hardware stack





Ryzen 9 7950x, 96GB of DDR5 memory 2 Nvidia RTX 4090, each with 24 GB dedicated memory

Jordan Rey-Jouanchicot - UBICOMM 2024 – Appendix 1