UBICOMM 2024

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)

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,...

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

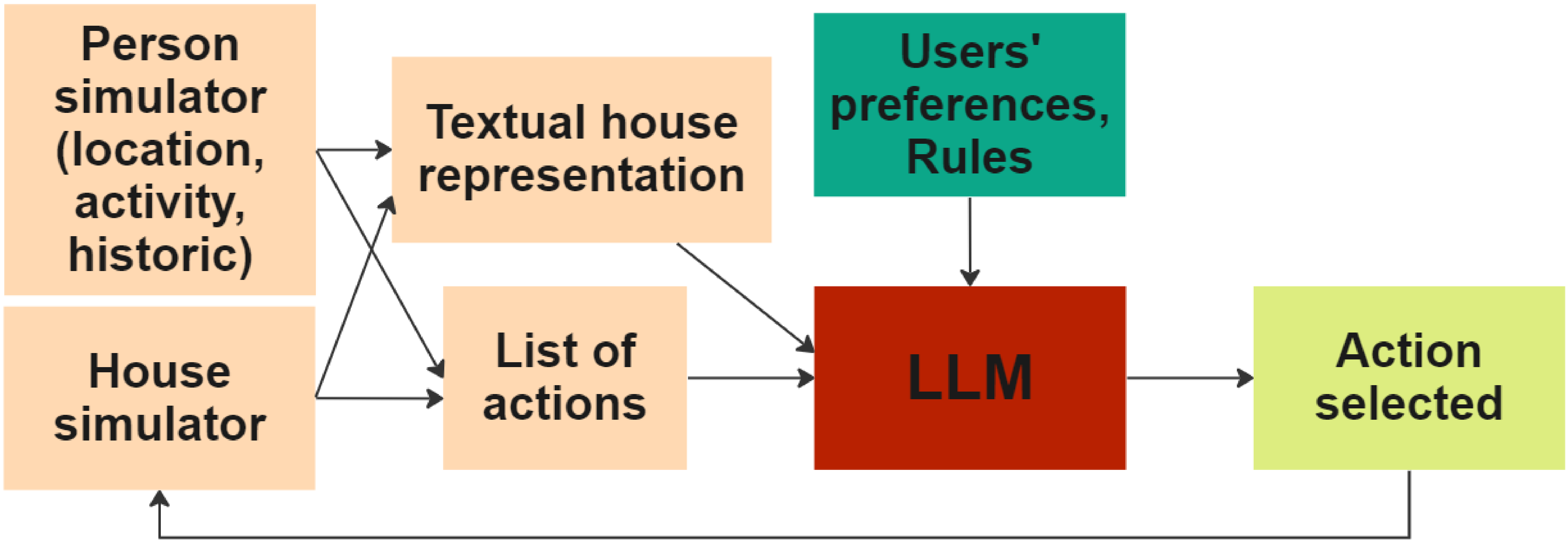
Benefits of LLMs

- 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

Proposed architecture



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"}

Textual

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

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.
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°

Prompting modelisation and methods

Injecting user preferences

Direct

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

No user preferences

DirectPref

A system prompt with the preferences, rules and generality from the database 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 main problems". For each, the LLM is asked to use the RAG to obtain the 3 closest preferences.
2. A prompt to ask for responses in the 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 response in the specified format.

Retrieval Augmented Generation

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 open-weight models according to the benchmarks.

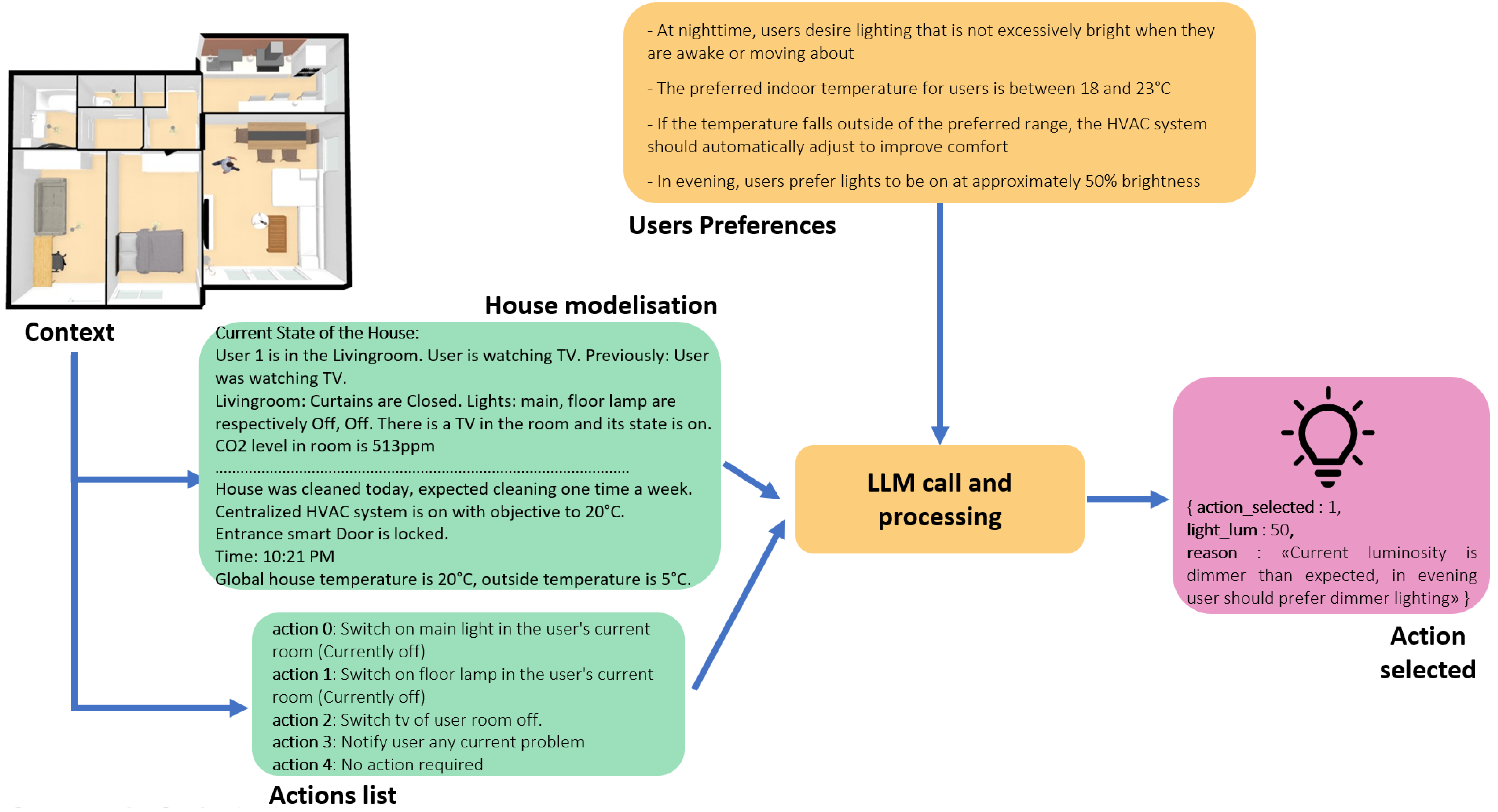
Qwen 1.5 14B
Alibaba Cloud

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.

Visual explanation of the architecture



Evaluation scenarios

Three metrics to evaluate the models, prompting and data representation

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

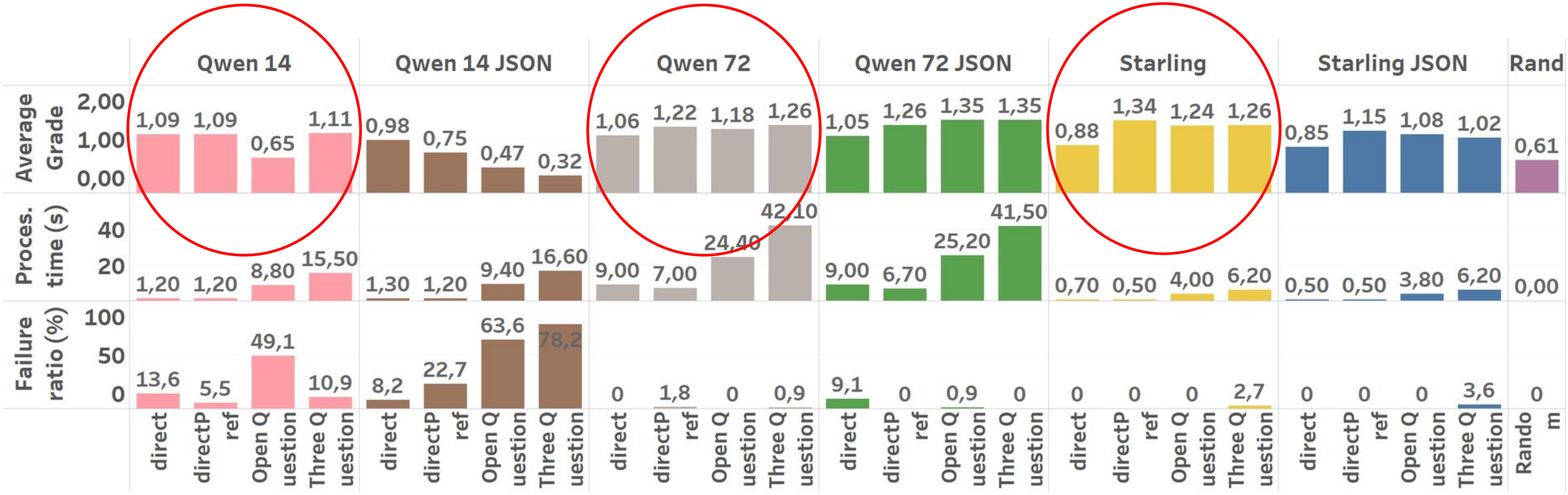
- **Processing time:**
Total runtime to obtain a response

- **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 Name	Grade	Associated Answer
Out of bed at night	2 1 0	Turn on auxiliary light or main light with reduced luminosity level Turn on main light Everything else
Watching TV: late evening	2 1 0	Turn on auxiliary light or main light with reduced luminosity level Turn on main light, open curtains, discuss Everything else
Out from bed issue with CO2	2 1 0	Inform user of risk Do an action and inform the user of risk Everything else
Going back to bed at night	2 1 0	Turn on auxiliary light or main light with reduced luminosity level Turn on main light Everything else
Evening sleeping: TV ON	2 1 0	Turn off TV Turn off anything on Everything else
At dinner watching TV	2 1 0	Turn on auxiliary light or main light with reduced luminosity level, open curtains Turn off the main light, do nothing Everything else
User out: TV is on	2 1 0	Turn off TV, turn off HVAC Turn off all lights Everything else
Too low temperature	2 1 0	Turn on HVAC Open Curtains Everything else
Low luminosity day	2 1 0	Open curtains Turn on any light in the room Everything else
Failed curtains	2 1 0	Turn on any light of the room Open curtains Everything else
Forgot to turn off lights	2 1 0	Turn off any lights, or HVAC Everything else

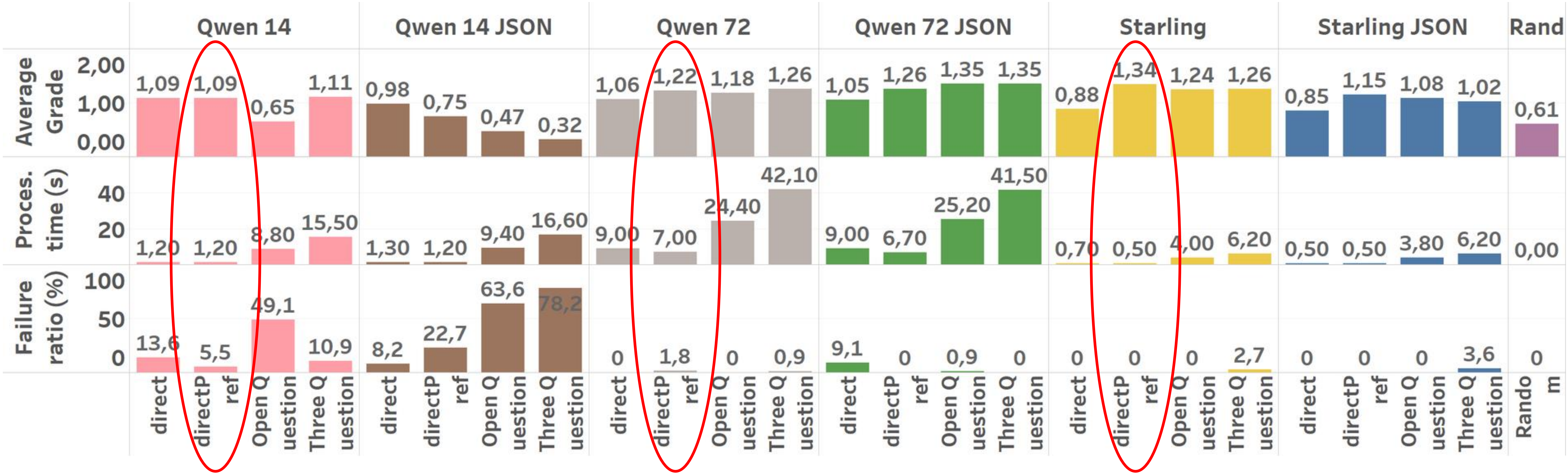
Results:
Data Representation



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

Minor impact on total response time

Results:
Users preferences

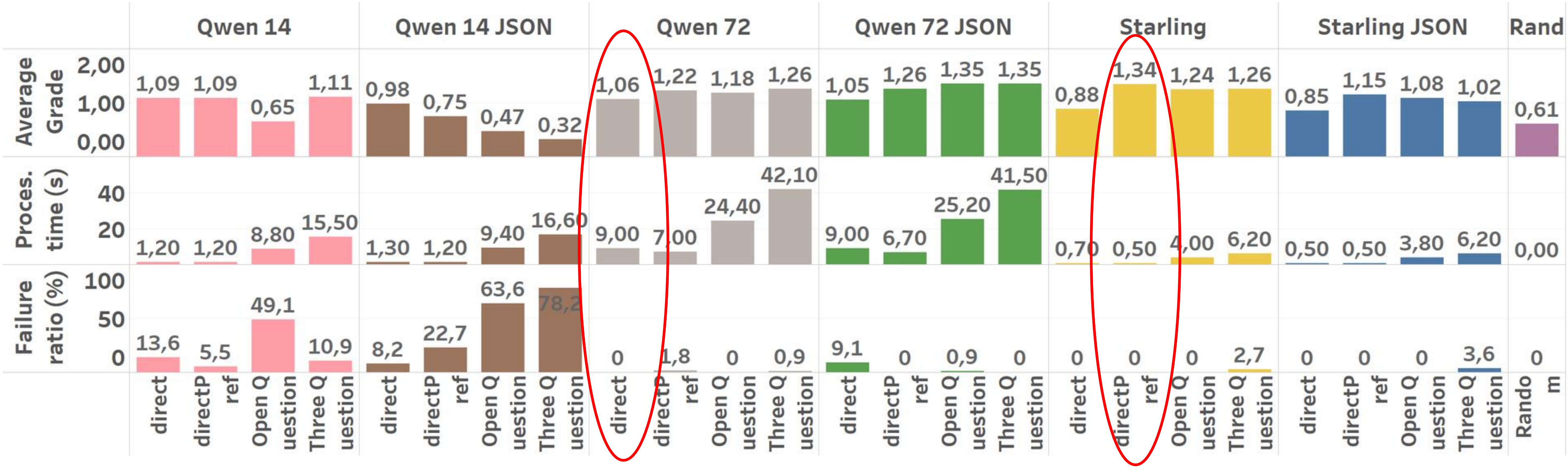


"directPref" prompt improved performance

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

Adding preferences brings significant benefits

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.

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

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.

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.

THANK YOU

Jordan Rey-Jouanchicot

jordan.reyjouanchicot@orange.com

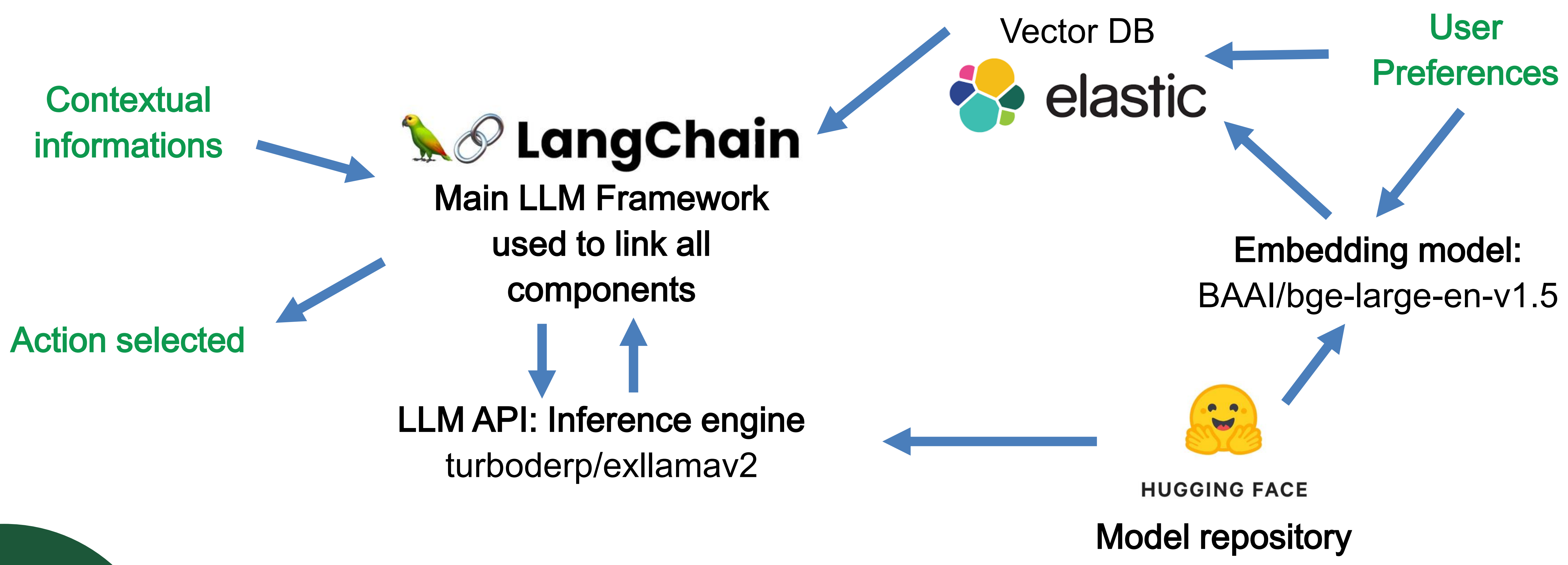
https://www.linkedin.com/in/jordanreyjouanchicot/

Appendix

Implementation

Hardware and software set-up

Software stack



Hardware stack

AMD NVIDIA

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