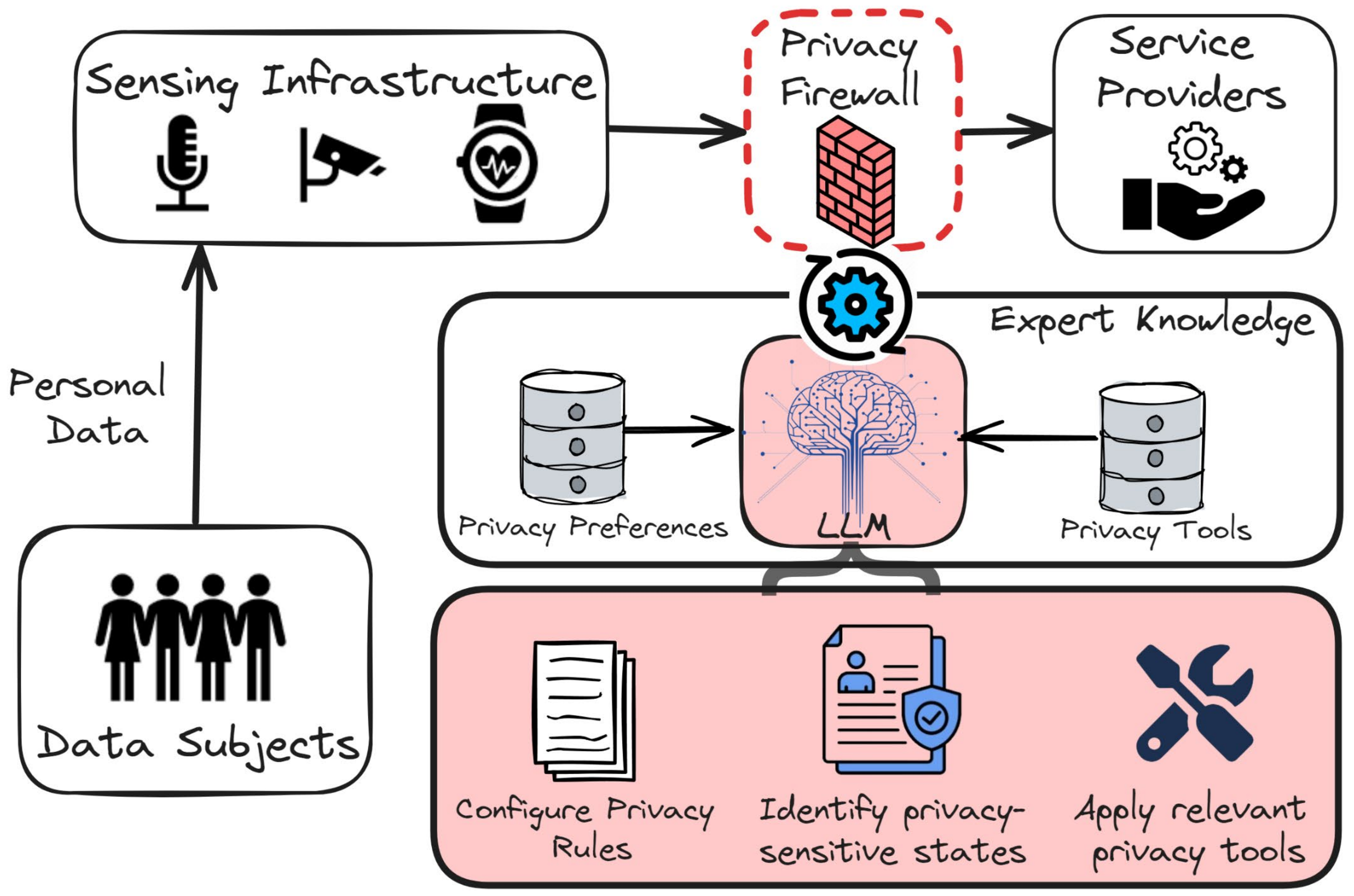
PrivacyOracle : Configuring Sensor Privacy Firewalls with Large Language Models in Smart Built Environments

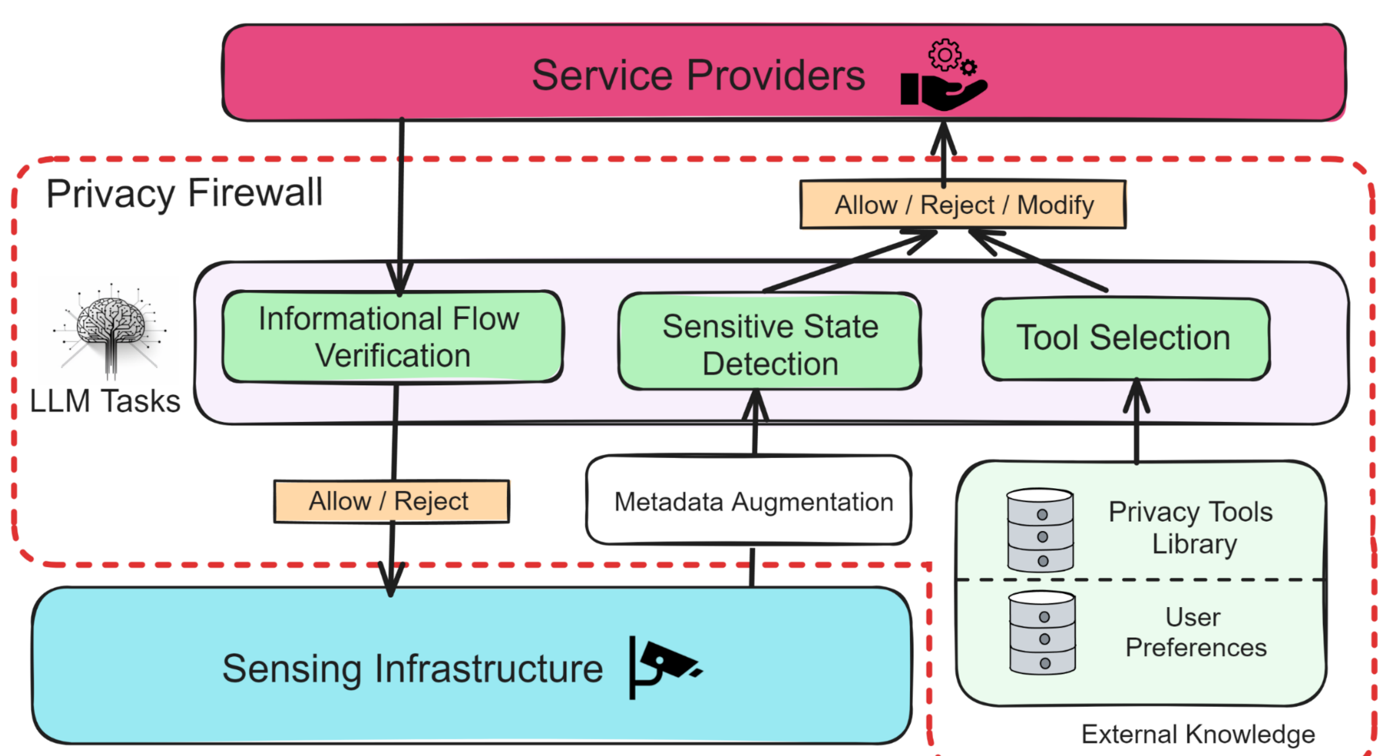
Brian Wang (UCLA), Luis Antonio Garcia (University of Utah), Mani Srivastava (UCLA)

Objectives

We propose a system for automatically managing privacy decisions of sensory data on behalf of data subjects in smart built environments,



LLM-based Privacy Firewalls



Our system, **PrivacyOracle**, accomplishes 3 tasks:

- *Informational flow verification* uses world knowledge of legal rules to identify violations of privacy
- *Sensitive state detection* uses lower level sensor information and preferences to identify privacy-sensitive states
- *Tool selection* identifies relevant privacy tools for a privacy preference and builds appropriate data processing pipelines

Methods

We utilize GPT-3.5 and GPT-4 to generate a variety of privacy decisions, with examples shown below.

Privacy Preferences

- Don't reveal information about *hygiene routines*
- Don't reveal information about *sedentary lifestyles*
- Don't reveal information about *house occupancy*

Objective:
Given a sequence of sensor events in a smart home, your task is to identify privacy-sensitive segments of time relevant to the user's privacy preference.

Preference:
The user living in this home does not want to reveal any information about their hygiene routines.

Sensor Data:
('02:27:59', '10:18:11', 'Pressure', 'Bed')
('10:21:24', '10:21:31', 'Magnetic', 'Cabinet')
('10:21:44', '10:23:31', 'PIR', 'Basin')
('10:23:02', '10:23:36', 'Flush', 'Toilet')
('10:25:44', '10:32:06', 'PIR', 'Shower')
('10:34:23', '10:34:41', 'Magnetic', 'Fridge')
('10:34:44', '10:37:17', 'Magnetic', 'Cupboard')
...

LLM Output:
Based on the user preference, these are times when data should not be revealed:

('10:21:44', '10:23:31', 'PIR', 'Basin')
('10:23:02', '10:23:36', 'Flush', 'Toilet')
('10:25:44', '10:32:06', 'PIR', 'Shower')
('10:49:48', '10:51:13', 'PIR', 'Basin')

Objective:
Please verify if the following informational flow is permitted under HIPAA.

Information Flow:
Sender: A voice assistant. Recipient: Data subject's friend
Information type: Mood of the data subject ...

LLM Output:
Permitted under HIPAA.

Top right (green) describes the sensitive state detection task. Left (purple) describes the informational flow task. Below (blue) describes the tool selection task.

Objective:
Build a simple dataflow program to perform image transformations according to a set of preferences.

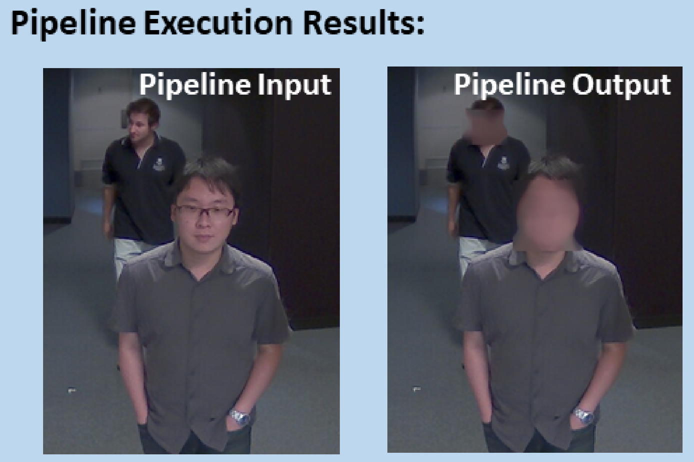
Available Tools:
- 'detect_face', which obtains an image, detects faces in an image, and returns the image.
- 'blur_face' which distorts the appearance of one's face given a detected face and makes it incredibly blurry.
- 'block_image' which returns a blank image given a detected face.
- 'facefusion' which takes in a video and automatically swaps appearances of faces with a predefined face, and outputs a video with the swapped appearances, but information such as expressions and emotions are retained.

...

Preferences:
Users in the images wish to hide their facial appearance, but are fine with the images revealing other information, such as the color of their shirt. Please create a pipeline based on this preference.

LLM Output:
Based on the user's preference to hide their facial appearance but not other information, the pipeline would look like this:
['detect_face', 'blur_face', 'save_image_to_directory']

Pipeline Execution Results:



Privacy Preferences:

A. Users in the image wish to hide any trace of them being part of the video
B. Users in the image wish to hide their facial appearance but *are fine with images revealing other information, such as the color of their shirt*
C. Users in the image wish to hide their facial appearance but *would like to enable emotion classification services on their facial data*

Results

Privacy-Sensitive State	IoU	F1
Hygenic Activities	0.684	0.831
Sedentary Lifestyle	0.844	0.983
House Occupancy	0.961	0.701

Measuring agreement between LLM-inferred privacy sensitive states and ground truth sensitive states

For *informational flow verification*, we use a manually curated dataset of 16 HIPAA scenarios with various contextual integrity parameters, and achieve a false positive rate and false negative rate of **6.25%.**

Requirement	Age F1	Gender F1	Race F1	Emotion F1
A	0.0	0.0	0.0	0.0
B	0.208	0.373	0.151	0.081
C	0.187	0.371	0.288	0.407

Privacy (age, gender, race recognition) vs. Utility (emotion) on LLM-generated processing pipelines for each privacy preference

In the *Tool Selection* task, each tool has different privacy/utility costs and the LLM selects the appropriate tool for each preference.

Future Work

- Examine enforcement of privacy decisions in non-cooperative sensing environments
- Evaluation of LLM results on different prompting strategies and validation mechanisms
- Managing conflicting privacy requires among users (democratization/negotiation)

Acknowledgements

This research was sponsored in part by the National Science Foundation (NSF) under awards # 1705135, 2124252 and 2211301, the NIH mDOT Center under Award #1P41EB028242, the DEVCOM ARL under Cooperative Agreement #W911NF-17-2-0196, and the DARPA ANSR Program under Contract #FA875023C0519.