

# SELLMA: Semantic Location through On-Device LLMs and WiFi Sensing

Martin Korelič  
Faculty of Computer and Information  
Science, University of Ljubljana  
Ljubljana, Slovenia

Octavian M. Machidon  
Faculty of Computer and Information  
Science, University of Ljubljana  
Ljubljana, Slovenia

Veljko Pejović  
Faculty of Computer and Information  
Science, University of Ljubljana  
Department of Computer Systems,  
Institute “Jožef Stefan”  
Ljubljana, Slovenia

## Abstract

Understanding a user’s semantic location is of critical importance in numerous areas of mobile computing, such as mobile healthcare, mobile advertising, and mobile personal assistance. Nevertheless, inferring semantic location remains challenging and often relies on translating raw geographical coordinates via third-party online services. In this paper we introduce SELLMA, an approach for semantic location inference that harnesses Wi-Fi SSID sensing and on-device querying of a specially crafted LLM. We implement SELLMA in Android and show that it can uncover a number of environmental and geographical descriptors of a users location in a privacy-preserving manner, without the need for GPS querying, and without reliance on Web-based services.

## CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computer systems organization** → *Embedded and cyber-physical systems*; • **Computing methodologies** → **Machine learning**.

## Keywords

Location sensing, Semantic location, Large Language Models, LLM fine-tuning, WiFi sensing, Ubiquitous computing

### ACM Reference Format:

Martin Korelič, Octavian M. Machidon, and Veljko Pejović. 2025. SELLMA: Semantic Location through On-Device LLMs and WiFi Sensing. In *The 8th International Workshop on Edge Systems, Analytics and Networking (EdgeSys '25)*, March 30–April 3, 2025, Rotterdam, Netherlands. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3721888.3722091>

## 1 Introduction

Since the early days of mobile sensing semantic location has represented staple information enabling context-aware computing and various applications thereof [3, 19]. For example, knowing that a tourist is in a particular part of town, the smartphone could provide notifications containing relevant sightseeing information, sensing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*EdgeSys '25, Rotterdam, Netherlands*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1559-4/25/03

<https://doi.org/10.1145/3721888.3722091>

### PROMPT:

User is nearby a Wi-Fi SSID named “[www.pattyfruits.cl](http://www.pattyfruits.cl)”. Based on the SSID provide location information and environmental context that the user might be in. Provide answers in sections and give reasoning.

### ANSWER:

#### 1. LOCATION INFERENCE FROM WI-FI SSID

- **COUNTRY:** Chile, Valparaiso
- The Wi-Fi SSID “[www.pattyfruits.cl](http://www.pattyfruits.cl)” suggests a Chilean origin, as the country code “.cl” is specific to Chile. The name “Pattyfruits” is also a Chilean brand, which further supports the inference of the location being in Chile.
- **COORDINATES:** (-32.9492448, -70.6656211)

#### 2. TYPE OF ENVIRONMENT

- **TYPE OF ENVIRONMENT:** Urban or commercial area
- The location indicates that the user is in a residential area, but the presence of a fruit juice bar suggests that the area is also commercial or tourist-oriented.

#### 3. OTHER ENVIRONMENTAL FACTORS

- **LIKELY FOOT TRAFFIC:** Moderate to high, depending on the time of day and day of the week
- **NEARBY AMENITIES:** Patty Frutas (a fruit juice bar) and possibly other nearby shops or restaurants

#### 4. LIKELY USER ACTIVITY

- **ACTIVITY:** Relaxing or socializing - The presence of a fruit juice bar suggests a leisurely or social activity, and the urban or commercial area type of environment indicates that the user is likely in a public place. The user is likely relaxing or socializing, possibly enjoying a drink or snack at the fruit juice bar.

### GROUND TRUTH:

**Location:** Arica, Chile

**Amenity:** Café

**Coordinates:** (-18.4449501, -70.2953886)

**Figure 1: Example SELLMA output when queried with a single sensed WiFi SSID “[www.pattyfruits.cl](http://www.pattyfruits.cl)”. The top text bubble (purple) depicted the prompt, here shortened for brevity. The bottom text bubble (yellow) shows the detailed SELLMA answer generated on a mobile device. Finally, the ground truth semantic information and actual geographic coordinates obtained through Open Street Maps are listed.**

that a user is in a library could be used to automatically switch the phone to the silent mode, while recognising that a user has just arrived to an amusement park could pop up ride reviews and advice on how to best spend the day there.

Traditionally, uncovering semantic location has relied first on obtaining a precise location, usually via a GNSS system, such as GPS, and then querying online services that perform translation from raw latitude and longitude information to contextual descriptors (e.g. Google Maps Platform [16], Google Geolocation API [15]). Such an approach, however, comes with serious drawbacks. GPS location is often unavailable in indoor environments and, from the energy consumption point of view, GPS querying can be prohibitively expensive for certain edge devices. Furthermore, the ability of

online services to perform accurate coordinate-to-semantics translation depends on the availability of (usually volunteered) descriptors related to particular coordinates, while the online querying implies that a user must trust that these services will not misuse the transmitted location information.

In this paper we present SELLMa (example of which is shown in Figure 1), a method for semantic location querying that does not require neither precise geographical coordinates sensing, nor remote database querying. Moreover, SELLMa can provide the semantics even for previously unseen locations. To accomplish this, our solution merges the reasoning capabilities and the world knowledge of large language models (LLMs) and the semantics hidden in Wi-Fi SSID descriptors. More specifically, SELLMa senses Wi-Fi SSIDs from a mobile’s surroundings and then uses these to populate a carefully crafted LLM prompt to uncover the location’s semantics, and potentially other information, such as the country/city/region, a user’s likely activity, as well as the absolute location coordinates. To avoid potentially sensitive data transfer, SELLMa hosts a specially pre-trained and optimised LLM directly on a mobile device. We implement a fully-functional SELLMa app on Android and through a series of experiments demonstrate its potential to uncover the semantic location of a user.

The specific contributions of our work include:

- We, to the best of our knowledge, for the first time integrate mobile sensing and LLM querying for the purpose of semantic location sensing, so that the two are both fully confined to an edge device;
- From publicly available sources we curate a comprehensive dataset of Wi-Fi SSID names and the corresponding semantic locations, and use this dataset in combination with a larger “teacher” model to fine-tune a smaller “student” LLM optimised for on-device mobile deployment;
- We implement a full-fledged Android application that enables semantic location querying, evaluate the quality of its outputs and assess the system-level overheads that sensing and on-device LLM querying bring.

With SELLMa we demonstrate that fine-tuned LLMs can infer semantic location and can indeed be optimised to perform such inference directly on contemporary edge devices. With this we open novel opportunities for semantic-aware services, while at the same time ensuring that a user’s privacy is preserved.

## 2 Related Work

Semantic location represents valuable information for a range of mobile applications. For instance, knowing the history of semantic locations visited by a person and their contacts can help us predict the next locations they will visit [17]. Semantic location can also guide dynamic adaptation of a mobile device’s user interface by automatically changing between alert modes, or controlling notification delivery [12, 22]. Finally, it is invaluable for understanding the behaviour of mobile users, such as their eating habits [2] or drug abuse [7].

Existing approaches for semantic location acquisition, such as Google Maps and Mapbox [18], provide an option for querying cloud-based proprietary databases with a user’s coordinates to obtain semantic location. Yet, sending raw coordinates to a third party

exposes the users to potential risks, including unauthorized tracking, data misuse, and personal information leakage. Moreover, GPS remains one of the most power-hungry aspects of mobile sensing and the one that may prohibit semantic location determination in certain scenarios [10]. Consequently, there is demand for solutions that are secure, energy-efficient and privacy-preserving [9].

Wi-Fi Service Set Identifiers (SSID), signal strength, and other meta-data are easily obtainable through wireless sensing and a number of studies explored semantic location inference from Wi-Fi data [1, 5, 8]. Indoor localisation research leveraged Wi-Fi signals [13, 29], Wi-Fi AP fingerprints [20] and temporal features in Wi-Fi connectivity data to classify a person’s indoor location to a semantically meaningful spatial extent [14]. For outdoor semantic localisation, publicly available Wi-Fi datasets were explored to extract the semantics of names and correlate them with known business venues [23]. None of the above approaches, however, can handle the inference over previously unseen locations.

Large language models’ (LLM) tremendous capabilities of generating content from the “world data” have been harnessed in ubiquitous sensing. Xu et al. [24] introduced the term “Penetrative AI”, emphasising the ability of making LLMs comprehend the physical world by interacting through the Internet of things (IoT) sensors and actuators. The same authors built on top of this study and proposed AutoLife [25], an automatic smartphone-based life journaling system. By analyzing the smartphone’s sensor data, the app synthesises comprehensive life journals based on the multimodal input to the vision and language models. Similar to our approach, the app senses Wi-Fi SSIDs and uses this information to infer something about the environment. However, AutoLife relies on a cloud-based LLM with no fine-tuning and additional online APIs, whereas we employed fine-tuned on-device LLM entirely offline and with no need to share the user’s data with other online services.

Recently, running LLMs on smartphones has been demonstrated [26, 27], and has soon found its way in mobile sensing applications. Thus, Zhang et al. [28] combine the on-device LLM with smartphone sensing for general context-aware and personalised services. However, their approach requires an Android terminal emulator inside a Linux environment app, which severely limits the solutions usability outside the demonstrated prototype. In our work, we engineer the LLM deployment and querying pipeline in a manner that enables its seamless use within an arbitrary Android application.

## 3 SELLMa

We now present the design and implementation of SELLMa, a solution for semantic location inference that combines Wi-Fi sensing and on-device LLM inference. Our approach comprises two main phases: (i) fine-tuning and optimising an LLM for semantic location inference from Wi-Fi SSID inputs, and (ii) deploying this model on a mobile device and integrating with Wi-Fi sensing to enable fully offline experience.

### 3.1 Fine-tuning LLM for semantic location

To construct a mobile-ready LLM, we first employ a larger model to create training dataset that connects Wi-Fi SSID information with well-structured semantic location information, and then fine-tune a smaller LLM using the above dataset.

**3.1.1 Dataset construction via teacher model.** To build a dataset for semantic location inference, we collect publicly available Wi-Fi SSIDs and their corresponding geolocation metadata through open-source APIs. Specifically, we use the Overpass API [21] to query OpenStreetMap entries with the tag `internet_access:ssid`, i.e. for locations whose both Wi-Fi SSID environment and geographic information are known. Each JSON-formatted record, thus contains Wi-Fi SSID name(s), the exact geographical coordinates (latitude and longitude), and relevant OpenStreetMap tags, such as the type of a venue/amenity.

In SELLMA, instead of using the input-output JSON entries, we use these entries to query a larger teacher LLM to generate structured explanations about each SSID and its environment. This allows us to, in a way, distill knowledge of the larger LLM, which is more expressive and could possibly infer the semantic location from certain clues in Wi-Fi SSIDs in situations where the smaller model would not. More specifically, we construct the prompt for the large LLM as shown in Figure 2 below.

**System:** You are a location reasoning model. You help infer the environment based on Wi-Fi SSIDs and geographical data.  
**User:** User is nearby a Wi-Fi SSID named "[SSID]". [LATITUDE, LONGITUDE] Using the Wi-Fi SSID and the geo-location information provided, answer the following questions in a structured format with reasoning for each answer:  
 Additional geolocation data: [GEOLOCATION DATA]

**Figure 2: Prompt template for semantic location reasoning.**

The placeholders [SSID], [LATITUDE, LONGITUDE], and [GEOLOCATION DATA] from Figure 2 are dynamically populated with context-specific information from the dataset. The teacher model is then instructed to answer the following queries related to semantic location:

- S1. Location inference from Wi-Fi SSID:** Based on the Wi-Fi SSID "[SSID]", give the estimation of the location (country, region, coordinates).
- S2. Type of environment:** Given the geographical location data, what type of environment is the user in? Are there any notable places nearby (e.g. landmarks, businesses etc.) based on the provided geo-information?
- S3. Other Environmental Factors:** What other factors can be inferred from the data (e.g., likely foot traffic, nearby amenities, noise levels)?
- S4. Likely User Activity:** Based on the SSID and the surrounding location information, what activity might the user be engaged in? (e.g., shopping, working, studying, relaxing, traveling). How does the environment affect what the user is likely doing at this moment?

These structured prompts encouraged the teacher model to provide detailed reasoning, explaining how it used both the SSID and the geo-information to arrive at each conclusion.

For the teacher model, we employ Meta-Llama-3.1-8B-Instruct [6], an 8-billion-parameter instruction-tuned LLM. Using the above procedure, we generate 3,000 unique samples spanning a wide variety of Wi-Fi SSIDs and locations around the globe.

**3.1.2 Mobile-ready LLM fine-tuning.** Deploying an 8-billion-parameter model on a contemporary smartphone is infeasible. Therefore, fine-tune a compact “student” model for on-device use. We choose Llama-3.2-3B-Instruct, a 3-billion-parameter, instruction-tuned LLM that still retains much of the broader “world knowledge” captured in the larger teacher model.

To make fine-tuning computationally feasible, we adopt a Low-Rank Adaptation (LoRA) approach [11], which significantly reduces training overhead by adding lightweight adapter layers. We fine-tune these adapters in the decoder layers with a LoRA rank ( $r$ ) of 32, a scaling factor ( $\frac{r}{d}$ ) of 1, and a learning rate of  $4 \cdot 10^{-5}$ . As the training data we use 2687 entries from the dataset constructed in Section 3.1.1. Upon the training completion, we merge the LoRA adapters back into the base LLM weights to avoid any additional overhead during inference. Even so, the resulting model weighs approximately 13 GB, which prevents deployment on a smartphone. Therefore, we apply dynamic quantisation to 8-bit integers, reducing the size to around 3.4 GB while preserving sufficient accuracy for semantic location inference (as demonstrated in Section 4).

## 3.2 LLM on-device deployment and integration with sensing

Following the offline fine-tuning phase, we implement the LLM on mobile devices to enable real-time semantic location inference without reliance on external services. This on-device process includes the deployment of the quantised LLM model, integration of Wi-Fi sensing, and the development of a user interface (UI) that facilitates inference.

The on-device inference process begins by scanning for Wi-Fi SSIDs using Android’s `WifiManager`, which retrieves nearby networks without requiring an active Internet connection. In Android, SSID scanning is treated as a sensitive action that may reveal a user’s location, thus the `ACCESS_FINE_LOCATION` permissions must be granted by the user. Once the SSIDs are obtained, this information is inserted into a prompt template similar to that used by the teacher model (i.e. Figure 2 followed by the queries S1-S4), yet omitting the explicit geolocation data. This prompt is then tokenised and passed to the LLM.

The LLM itself is the quantised version of the model described in Section 3.1.2 exported into the ONNX Runtime [4] format, encompassing weights, tokeniser, and other inference artifacts. To optimise computation during text generation, we harness the key-value (KV) caching mechanism preserves previously computed states, thereby reducing redundant calculations in the inference loop. Finally, the application presents the generated semantic insights to the user through a chat-based interface (example in Figure 1), offering real-time location reasoning in a privacy-preserving manner by keeping all data and inference locally on the device.

## 4 Experimental Evaluation

We now evaluate SELLMA’s ability to infer semantic location by comparing the responses of our student LLM (3B parameters) against ground truth derived from the teacher model (8B parameters) on 313 *Wi-Fi SSID – semantic location* instances that did not

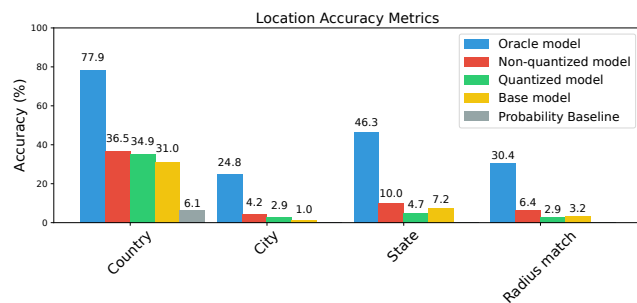
figure in the LLM fine-tuning set. Both normal and quantised versions of the student model are included in this evaluation, along with baseline models.

#### 4.1 Inference quality evaluation

We evaluate the models' performance on location reasoning from Wi-Fi SSIDs using the following main metrics: location accuracy, which includes both region accuracy (identifying the correct country, state, city, or county) and a 10 km radius match; cosine similarity, computed by comparing the text embeddings (from all-MiniLM-L6-v2 embedder model) of student and teacher responses; and environment type label accuracy, indicating how well the model's labeled environments align with the teacher's labels. Since location labels (e.g., country or region) can vary in format across different model responses, an external forward geocoding API was used to normalize these labels in both the ground truth and the predictions. For multi-label classification tasks such as *type of environment* and *likely user activity*, we employ the F1 score and Hamming loss with weighted averaging.

We compare the following small student models:

- (1) **Oracle model** – A trained student model given the exact geolocation data in the prompt (e.g., country or amenity type). This model is, thus, expected to provide very accurate information and answers that are as close to the original teacher model outputs as possible.
- (2) **Non-quantised model** – The primary trained student model running at full precision, with no explicit geolocation data in the prompt. This simulates real-world scenarios where only SSID information is available. However, this model is still prohibitively complex to be ran on an average commodity smartphone.
- (3) **Quantised model** – The student model quantised to lower precision. This model is readily deployable to resource-constrained Android smartphones.
- (4) **Base model** – The untrained version of the base student LLM, providing a baseline with no task-specific adaptation.
- (5) **Probability baseline** – A purely statistical model that samples locations and labels according to their observed frequency in the training set.



**Figure 3: Location accuracies of the responses generated by models compared against ground truth labels.**

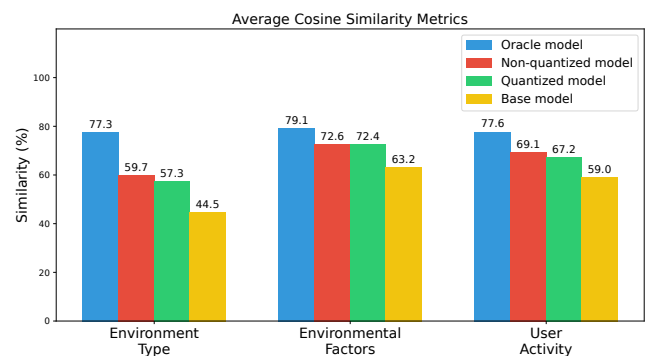
We first evaluate the ability of the models to infer geographical location information from Wi-Fi SSID data. Figure 3 illustrates that

	O	NQ	Q	B
USA	16.2	21.3	33.4	<b>54.8</b>
Germany	8	<b>19.7</b>	10.8	7.1
Hungary	9.9	12.1	10.5	1.6
France	3.8	5.4	4.5	5.8
Italy	2.9	<b>7.6</b>	7	4.5
Finland	1.1	1.9	1.3	1.3
Spain	1.2	2.2	2.2	2.2
Others	56.9	29.8	30.3	22.7

**Table 1: Distribution of the most frequent country prediction ratios across different variants of the student model (O - oracle version, NQ - non-quantised version, Q - quantised version and B - base version).**

the fine-tuned student model (non-quantised) achieves 36.5% accuracy in identifying the correct country which is significantly higher than 6.1% accuracy achieved by a probabilistic model. In addition, compared to the base model, we see that fine-tuning improves inference by about 5% (from 31% to 36.5%). Quantisation has only a slight negative impact on country-level predictions, implying that much of the location reasoning knowledge is preserved. This outcome is promising for real-world deployments, where memory and computational constraints necessitate lighter-weight models. All models fare much worse when finer granularity prediction is requested (i.e. city, state, or 10km radius match). Surprisingly, this is true for the Oracle model as well, although it remains the most accurate of all the tested models. The probability baseline performs poorly across the board, as simply sampling locations and labels from the training distribution yields only about 6% accuracy for country predictions, with performance on other location-related metrics dropping below 0.1% (hence not reported).

We also investigate where the misclassifications are made. Table 1 shows the distribution of country-level predictions. We observe that, compared to the oracle model, other LLM-based models tend to have predictions that are both less diverse and skewed toward a few countries. The base model, in particular, exhibits the highest skew (towards "USA"), reflecting a reliance on broad knowledge rather than task-specific cues.



**Figure 4: Cosine similarities of the responses generated by models computed against the teacher models responses.**

	Environment classification				
	O	NQ	Q	B	PB
F1 score	79.4	81.2	79	36.4	55.8

**Table 2: Environment and activity classification metrics of different model variants (O - oracle version, NQ - non-quantised version, Q - quantised version, B - base version, PB - probability baseline).**

We now focus on the semantics, and assess the models' ability to infer the environment type (e.g. restaurant, library, etc.), environment factors (e.g. noisy, busy, etc.), as well as the likely user activity in the environment. Figure 4 shows the cosine similarity between text embeddings of the teacher LLM-provided ground truth description and each of the models' output, while Table 2 shows the F1 score of the predicted environment type.

The oracle model, naturally, achieves highest match with the teacher model's responses (cosine similarity in Figure 4), however, there is a clear benefit in the student model fine-tuning, as the fine-tuned model (non-quantised) performs significantly better than the base model in each of the three – environment type, environmental factors, and user activity – cases. The F1 score does not take the answer wording in consideration, and here we see (Table 2) that fine-tuning indeed works very well. Without fine-tuning the base model (B) performs poorly, even worse than the probability baseline (PB). The fine-tuned student LLM (NQ) performs as well as the oracle solution (O). Quantisation, again, has very little effect on the inference quality with the quantised LLM (Q) losing only 2.2 F1 points compared to the non-quantised version (NQ).

#### PROMPT:

User is nearby a Wi-Fi SSID named "508 Муезерский - Петрозаводск, автовокзал". Based on the SSID provide location information and environmental context that the user might be in. Provide answers in sections and give reasoning.

#### ANSWER:

##### 1. LOCATION INFERENCE FROM WI-FI SSID

- **COUNTRY:** Russia
- **REGION:** Petrozavodsk (a city in the Republic of Karelia, Russia).
- **COORDINATES:** (61.6251111, 34.4441111)

The SSID "508 Муезерский - Петрозаводск, автовокзал" suggests that the user is in Petrozavodsk, a city in Russia. The presence of the word "автовокзал" (avtovokzal, meaning "bus station") further confirms this inference, as it is a common feature in Russian cities.

#### GROUND TRUTH:

Russia, Republic of Karelia, Petrozavodsk

**Figure 5: LLM response for SSID name input written in a non-Latin alphabet and not containing English language words.**

Finally, the results also demonstrate that LLMs exhibit capabilities in inferring location context by leveraging multilingual knowledge encoded during pretraining, which is depicted in an example shown in Figure 5. This includes the ability to translate and interpret Wi-Fi SSIDs in various languages, often connecting them to the correct geographical locations. This highlights the portability and global

	Prediction	Reasoning	Tokens/s	Cache
S1	20 s	35 s	2.25	107 MB
S2	23 s	31 s	2.11	126 MB
S3	27 s	47 s	1.98	151 MB
S4	14 s	54 s	1.92	174 MB
Avg	21 s	42 s	2.01	139.5 MB

**Table 3: Time taken for token generation in the observation (initial prediction) and reasoning (explanation) for each of the sections S, together with the tokens per second, and KV cache usage during inference.**

usability of SELLMA in diverse contexts worldwide with their wide variety of pretrained knowledge.

## 4.2 System evaluation

To assess SELLMA's usability on a real-world mobile device, we measure the following performance metrics on a Google Pixel 6 device: the time to load the model into memory, the time spent processing prompts before generation begins (prefill time), the rate at which tokens are generated, and the total amount of RAM required for model weights and runtime operations. The device features 8 GB of RAM (6.7 GB usable) and the initial prompts ranged from 360 to 430 tokens, influenced by the number and length of scanned Wi-Fi SSIDs; with a maximum of 650 tokens generated per prompt, total tokens occasionally approached 1,000.

Across a set of ten inference runs with the quantised (3B-parameter) model, loading the model into memory took approximately 2.14 s ( $\pm 0.32$  s), while prefill time averaged 25.87 s ( $\pm 2.25$  s). The model generated tokens at a rate of about 2.01 tokens/s and required around 2.3 GB of memory. Results of detailed testing across four representative tasks (S1–S4 from Section 3.1.1) listed in Table 3, show that the system took 21 s on average to generate the initial predictions, followed by 42 s for more detailed reasoning explanations per section. This corresponds to roughly 2 tokens generated per second, with approximately 140 MB of key-value (KV) cache used during inference. Predictions and reasoning times varied over tasks, but overall, the results indicate that SELLMA can handle moderately sized prompts and produce semantically rich answers in under a minute for each section S.

## 5 Discussion

SELLMA demonstrates the feasibility of running contextual reasoning about the location entirely on-device, using a locally deployed LLM. This approach offers several advantages, including (i) *privacy preservation*, as all the computation occurs locally, removing the need to transmit sensitive data to external servers, (ii) *global applicability*, as multilingual, context-rich knowledge embedded in modern LLMs allows them to interpret Wi-Fi SSIDs across various cultural and geographic contexts, supporting worldwide usability, (iii) *enhanced contextual awareness*, as by inferring environmental characteristics from SSIDs the model adds a novel layer of contextual intelligence to location-based services, potentially offering more fine-grained insights than traditional methods, and (iv) *tolerance to compression*, as the minimal drop in accuracy observed

between high- and low-precision models underscores how an over-parameterized LLM can retain much of its learned knowledge even after quantisation, minimising memory overhead without severely compromising performance.

Nevertheless, we also identify avenues for further improvements. Namely, we find that the model excels in settings with semantically rich SSIDs, while it can struggle when Wi-Fi labels are sparse or ambiguous, resulting in less precise location insights than those offered by GPS-based methods. Furthermore, we find that a balance needs to be found between assigning sufficient labels to the Wi-Fi SSIDs via the teacher model and preserving model ability to provide to-the-point answers – if the teacher output includes multiple relevant keywords this might promote further generalisation within the smaller model, reducing the precision of environment and activity classification. Finally, for certain practical applications, response generation needs to be accelerated. Part of the running time can be attributed to the prefill phase. This overhead may be reduced by restructuring the prompt to separate static and dynamic content, thus allowing the key-value (KV) cache for the static portion to be pre-computed and reused. Although this approach demands some storage for the precomputed cache, it remains practical given current smartphones' abundant storage capacity. Finally, future work could also investigate Retrieval-Augmented Generation (RAG), which integrates external knowledge about frequently encountered Wi-Fi SSIDs to improve inference quality.

## 6 Conclusion

This paper introduces SELLMA, a method for location-based contextual awareness powered by Wi-Fi sensing and edge-based LLM querying. To realise SELLMA, we curate a dataset of Wi-Fi SSIDs and semantic location descriptors and use it, with the help of a larger LLM, to fine-tune a smaller student LLM optimised for on-device deployment. The resulting system can infer rich contextual information solely from local Wi-Fi SSIDs, eliminating the need for GPS coordinates. With SELLMA we demonstrate that privacy-friendly, globally applicable semantic location sensing directly on mobile devices is indeed feasible.

## Acknowledgements

Our research was partly funded by Slovenian Research Agency projects "Context-Aware On-Device Approximate Computing" (J2-3047) and "approXimation for adaptable distributed artificial intelligence" (N2-0393), as well as core funding No. P2-0098.

## References

- [1] Mona A Aboelnaga, M Watheq El-Kharashi, and Ashraf Salem. 2018. Secure WiFi fingerprinting-based localization. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, 543–548.
- [2] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites 'n' bits: Inferring eating behavior from contextual mobile data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–33.
- [3] Diane J Cook, Michael Youngblood, Edwin O Heierman, Karthik Gopalratnam, Sira Rao, Andrey Litvin, and Farhan Khawaja. 2003. MavHome: An agent-based smart home. In *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003 (PerCom 2003)*. IEEE, 521–524.
- [4] ONNX Runtime developers. 2021. ONNX Runtime. <https://onnxruntime.ai/>. Version: 1.20.1.
- [5] Yinhan Dong, Guoxiong He, Tughrul Arslan, Yunjie Yang, and Yingda Ma. 2023. Crowdsourced indoor positioning with scalable WiFi augmentation. *Sensors* 23, 8 (2023), 4095.
- [6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [7] David H Epstein, Matthew et al. Tyburski, Ian M Craig, Karran A Phillips, Michelle L Jobs, Massoud Vahabzadeh, Mustapha Mezghanni, Jia-Ling Lin, C Debra M Furr-Holden, and Kenzie L Preston. 2014. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new method to study behavior in its geographical context. *Drug and alcohol dependence* 134 (2014), 22–29.
- [8] Yao-Chung Fan, Yu-Chi Chen, Kuan-Chieh Tung, Kuo-Chen Wu, and Arbee LP Chen. 2015. A framework for enabling user preference profiling through wi-fi logs. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2015), 592–603.
- [9] Jaap-Henk Hoepman. 2021. *Privacy is hard and seven other myths: Achieving privacy through careful design*. MIT Press.
- [10] Hagen Höpfner and Maximilian Schirmer. 2012. Energy efficient continuous location determination for pedestrian information systems. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*. 58–65.
- [11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [12] Andreas Komminos, Ioulia Simou, Antonis-Elton Frengkou, Nikolaos Gkorgkolis, and John Garofalakis. 2023. Where am I? Predicting user location semantics from engagement with smartphone notifications. *Journal of Ambient Intelligence and Humanized Computing* 14, 12 (2023), 15687–15703.
- [13] Phat-Vo Le Thanh, Quyen Ng Ph, Duy Nguyen, Bang Khuc, Aleksandr Gelgor, and Phuoc Nguyen TH. 2024. Public Dataset for Simultaneous Human Activity Recognition and Localization using WiFi Signals. In *2024 International Conference on Electrical Engineering and Photonics (EExPolytech)*. IEEE, 206–209.
- [14] Yiming Lin, Daokun Jiang, Roberto Yus, Georgios Bouloukakis, Andrew Chio, Sharad Mehrotra, and Nalini Venkatasubramanian. 2020. Locater: cleaning wifi connectivity datasets for semantic localization. *arXiv preprint arXiv:2004.09676* (2020).
- [15] Google LLC. 2025. Geolocation API Overview. <https://developers.google.com/maps/documentation/geolocation/overview>
- [16] Google LLC. 2025. Google Maps. <https://maps.google.com>
- [17] Eric Hsueh-Chan Lu and You-Ru Lin. 2023. A Self-Attention Model for Next Location Prediction Based on Semantic Mining. *ISPRS International Journal of Geo-Information* 12, 10 (2023), 420.
- [18] Mapbox. 2025. Location intelligence for business. <https://www.mapbox.com/>
- [19] Emiliano Miluzzo, Nicholas D Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B Eisenman, Xiao Zheng, and Andrew T Campbell. 2008. Sensing meets mobile social networks: the design, implementation and evaluation of the cenacle application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*. 337–350.
- [20] Jiazi Ni, Fusang Zhang, Jie Xiong, Qiang Huang, Zhaoxin Chang, Junqi Ma, BinBin Xie, Pengsen Wang, Guangyu Bian, Xin Li, et al. 2022. Experience: Pushing indoor localization from laboratory to the wild. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 147–157.
- [21] OpenStreetMap contributors. 2017. Planet dump retrieved from <https://planet.osm.org>. <https://www.openstreetmap.org>.
- [22] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 897–908.
- [23] Suranga Seneviratne, Fangzhou Jiang, Mathieu Cunche, and Aruna Seneviratne. 2015. SSIDs in the wild: Extracting semantic information from WiFi SSIDs. In *2015 IEEE 40th conference on local computer networks (LCN)*. IEEE, 494–497.
- [24] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.
- [25] Huatao Xu, Panron Tong, Mo Li, and Mani Srivastava. 2024. AutoLife: Automatic Life Journaling with Smartphones and LLMs. *arXiv preprint arXiv:2412.15714* (2024).
- [26] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088* (2024).
- [27] Zhenliang Xue, Yixin Song, Zeyu Mi, Le Chen, Yubin Xia, and Haibo Chen. 2024. PowerInfer-2: Fast Large Language Model Inference on a Smartphone. *arXiv preprint arXiv:2406.06282* (2024).
- [28] Shiquan Zhang, Ying Ma, Le Fang, Hong Jia, Simon D'Alfonso, and Vassilis Kostakos. 2024. Enabling on-device llms personalization with smartphone sensing. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 186–190.
- [29] Wei Zhang, Siwang Zhou, Liang Yang, Lu Ou, and Zhu Xiao. 2019. WiFiMap+: high-level indoor semantic inference with WiFi human activity and environment. *IEEE Transactions on Vehicular Technology* 68, 8 (2019), 7890–7903.