The City as a Personal Assistant: Turning Urban Landmarks into Conversational Agents for Serving Hyper Local Information

UTKU GÜNAY ACER, Nokia Bell Labs, Belgium
MARC VAN DEN BROECK, Nokia Bell Labs, Belgium
CHULHONG MIN, Nokia Bell Labs, United Kingdom
MALLESHAM DASARI[*], Carnegie Mellon University, USA
FAHIM KAWSAR, Nokia Bell Labs, United Kingdom

Conversational agents are increasingly becoming digital partners in our everyday computational experiences. Although rich and fresh in content, they are oblivious to users' locality beyond geospatial weather and traffic conditions. We introduce Lingo, a hyper-local conversational agent embedded deeply into the urban infrastructure that provides rich, purposeful, detailed, and in some cases, playful information relevant to a neighbourhood. Drawing lessons from a mixed-method contextual study (online survey, $n = 1992$ and semi-structured interviews, $n = 21$), we identify requirements for such a hyper-local conversational agent and a sample set of questions serving urban neighbourhoods of Belgium. Our agent design is manifested into a two-part system. First, a multi-modal reasoning engine serves as a hyper-local information source using automated machine-learning models operating on camera, microphone, and environmental sensor data. Second, a smart conversational speaker and a smartphone application serve as hyper-local information access points. Finally, we introduce a covert communication mechanism over Wi-Fi management frames that bridges the two parts of our Lingo system and enables the privacy-preserving proxemic interactions. We describe the design, implementation, and technical assessment of Lingo together with usability ($n = 20$) and real-world deployment ($n = 5$) studies. We reflect on information quality, accessibility benefits, and interaction dynamics and demonstrate the efficacy of Lingo in offering hyper-local information at the finest granularity in urban neighbourhoods while reducing access time up to a factor of 25.

CCS Concepts: • **Human-centered computing → Sound-based input / output**; **Ubiquitous and mobile computing systems and tools**; *Ubiquitous computing*; *Mobile computing*.

Additional Key Words and Phrases: Conversational Agent, Citizen Engagement, Edge AI, Spontaneous Interaction

ACM Reference Format:
Utku Günay Acer, Marc van den Broeck, Chulhong Min, Mallesham Dasari, and Fahim Kawsar. 2022. The City as a Personal Assistant: Turning Urban Landmarks into Conversational Agents for Serving Hyper Local Information . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 40 (June 2022), 31 pages. https://doi.org/10.1145/3534573

[*] Part of the work reported in this paper has been conducted while the author was an intern at Nokia Bell Labs, Cambridge, UK

Authors' addresses: Utku Günay Acer, Nokia Bell Labs, Belgium, utku_gunay.acer@nokia-bell-labs.com; Marc van den Broeck, Nokia Bell Labs, Belgium, marc.van_den_broeck@nokia-bell-labs.com; Chulhong Min, Nokia Bell Labs, United Kingdom, chulhong.min@nokia-bell-labs.com; Mallesham Dasari, Carnegie Mellon University, USA, malleshd@andrew.cmu.edu; Fahim Kawsar, Nokia Bell Labs, United Kingdom, fahim.kawsar@nokia-bell-labs.com.

2474-9567/2022/6-ART40 $15.00
https://doi.org/10.1145/3534573

1 INTRODUCTION

Conversational agents[1] are now pervasive, integrated into mobile phones, smart speakers, and even in cars. The remarkable advancement of machine learning is causing a seismic shift, in that conversational agents are now able to understand human speech and transform the text into speech in a similar way to humans [44] in everyday living spaces (even on-the-go). Naturally, this created interminable possibilities, uncovering novel, productive and useful experiences with conversational agents for accessing and interacting with digital services in many and diverse applications including HCI [27], customer experience [33], conversational commerce [36], medicine [47, 48], entertainment [23], education [32], and social work [10]. However, unfortunately, beyond weather and traffic conditions, the services provided by today's conversational agents do not consider the locality of users.

For long, ubiquitous computing research has attempted to understand location awareness to serve location-based services [15, 24]. Many applications have emerged offering a variety of experiences, including navigation support, recommendation for venues, augmentation of a search for people, places, things, etc. Unfortunately, these applications too often lack a locality view both temporally and spatially, i.e., information that is only available locally to local citizens. For example, consider a citizen would like to learn the spatiotemporal events such as the presence of garbage collector on the street, allergen concentration in the nearby street, or events of recent past, such as whether the postman visited the street already, or qualitative aspects of a neighbourhood such the ambience or safety. This extreme local information is not available in today's location-based services. With the emergence of the Internet of Things (IoT), local authorities increasingly use connected cameras and sensors to understand their cities and ensure their citizens' societal, economic, and environmental well-being. Naturally, in the past few years, we have seen many work that leverage diverse Internet of Things (IoT) systems in urban spaces that offer a quantitative view of the urban landscape e.g., noise, air pollution or mobility [4, 6, 17, 20].

We see a unique opportunity to bring these various systems together and offer citizens a conversational experience to access hyper-local information of their neighbourhood. To this end, in this paper, we present Lingo, a hyper-local conversational agent placed in urban landmarks and provide access to rich, purposeful, detail, and in some cases, playful information relevant to a neighbourhood. We first report a mixed-method contextual study (online survey, $n = 1992$ and semi-structured interviews, $n = 21$) that informed us of such an agent's information affinity and interaction modalities. In particular, we identified a set of questions as representative information for this type of service in the context of urban neighbourhood of Belgium. Then, grounded on these results, we developed a two-part system as illustrated in Figure 1. The first part of our system is a multi-modal reasoning engine called Lingo Observer that serves as a hyper-local information source and uses automated machine-learning models on a camera, a microphone and environment sensor data. We envision these observers to be embedded in urban landmarks, e.g., street lampposts or wall poles. The second part of our system, Lingo Agent, is manifested in a smart conversational speaker and a smartphone application that serve as hyper-local information access points over voice and text-based interactions. These two components are connected using a covert communication mechanism over Wi-Fi management frames that ensure privacy-preserving proxemic interactions, i.e., a user needs to be physically co-located (in this case within Wi-Fi range) to access the information.

We evaluated Lingo through a usability study ($n = 20$) followed by a real-world deployment of Lingo in two neighbourhoods and five households. Quantitative and qualitative insights gained from these studies uncover various facets, including information quality and diversity, topical and spatiotemporal coverage, access control and extension, etc. For instance, our participant found spatiotemporal events of the recent past (e.g., visit of a postman) or invisible attributes (e.g., pollen concentration) offer maximum utility. However, they prefer to have wider coverage both topically and spatially. The representative questions received good acceptance from the participants, and the access time to information has found to be significantly reduced up to a factor of 25.

[1] Here, conversational agents refer to voice assistants such as Alexa, Siri, Cortana, etc.

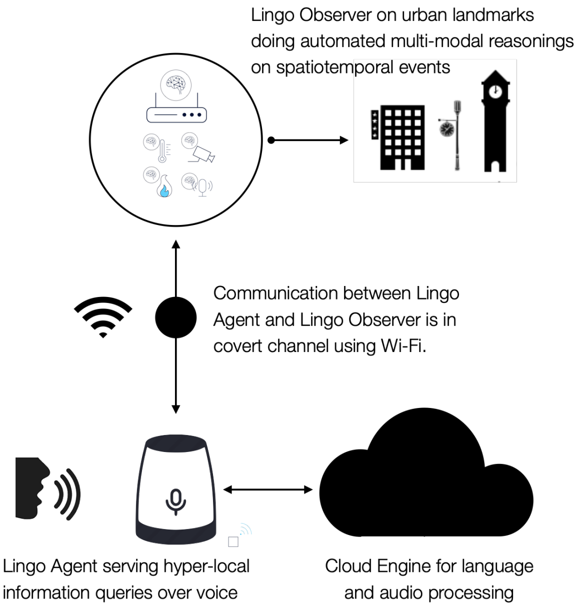

Lingo Observer on urban landmarks doing automated multi-modal reasonings on spatiotemporal events
Communication between Lingo Agent and Lingo Observer is in covert channel using Wi-Fi.
Lingo Agent serving hyper-local information queries over voice
Cloud Engine for language and audio processing

Fig. 1. A hyper-local conversational agent offers spatiotemporal events of a neighbourhood through a conversational user interface and multi-modal reasoning engine embedded on an urban landmark.

Taken together these and the rest of our findings demonstrate the many ways hyper-local conversation agents can bring substantial benefits to urban citizens. The main contributions of this work are as follows:

- We report the findings of a mixed-method contextual study to uncover the information affinity and interaction modality of hyper-local information that led to twenty representative questions.
- We present a concrete manifestation of our first-of-its-kind solution in designing a hyper-local conversational agent and describe its various design and technical facets in detail.
- We offer insights from real-world deployment coupled with controlled usability evaluation of our hyper-local conversational agent to put forward exciting research directions in urban computing.

In what follows, we first present the contextual study that informs the design of Lingo. Next, we provide an in-depth technical view of our solution. Then we present an evaluation of Lingo from three perspectives – system, usability and real-world deployment. We then reflect on several implications that emerged from our evaluation. We revisit related past research before concluding the paper.

2 RELATED WORK

Lingo is a hyper-local conversational agent embedded deeply into the urban infrastructure providing spatiotemporal information relevant to a neighbourhood. Three research areas are relevant to Lingo, namely, conversational agents, location-based systems, sensory and crowdsourcing systems and citizen science. In this section, we review related research in these areas.

2.1 Conversational Agents

Conversational agents have become a significant part of the computational experience in the last decade. These assistants understand spoken commands from users to perform various tasks that are constrained by the applications installed on user devices. With the advancement in machine learning, the agents are able to understand the user speech in audio signals and convert text into speech. Once the user makes a vocal query, the agent sends raw audio signals to a remote server for natural language processing to recognise the intent of the user, and carries out the requested service. In addition to commercial-grade conversational agents (Alexa, Siri, Google, Cortana, etc.), specialised agents are used for accessing and interacting with digital services in many and diverse applications including education [32], customer experience [33], conversational commerce [36] and medicine [48].

Prior work has extensively studied conversational agents as a communication modality to receive digital services and explores user practices and expectations in personal devices and/or settings such as homes [7, 11, 28, 37]. The way these agents affect the social dynamics in home environments has been presented in [50]. On the other hand, we investigate the use of such devices in urban spaces to extract extremely local information. Pearson *et. al* study the use of conversational agents in urban settings and how they can improve the citizens' daily lives [35]. Lingo builds upon such urban conversational agents to provide an end-to-end system pipeline that not only accepts queries, but also computes the responses to these queries with an ML engine. This engine executes multiple models locally to compute responses to hyper-local queries rather than relying completely on 3rd party service providers.

2.2 Location-Based Systems

Location-based systems have been widely investigated to provide users with the local information. They obtain individual's location data, typically from a mobile device, and provide the requested information that is often retrieved from a remote server [15, 24, 41]. A large number of such applications have emerged offering a variety of experiences including navigation support, recommendation for venues, augmentation of a search for people, places, things, etc. While these applications enable users to easily spot local information of interest, they often lack a locality view both temporally and spatially, i.e., information only available locally to local citizens. Consider a citizen who would like to learn when the local events take place such as the postman passing by, qualitative properties of a neighbourhood such as friendliness of the inhabitants, nearby shops that serve a particular product, the tourism related facts about a location that relate its historical and cultural background, etc. This extreme local information, unfortunately, is not available on today's location-based services. In contrast to those, Lingo is designed to offer hyper-local information which is usually available to obtain only on the site. To this end, we systematically explore the variety of spatio-temporal information that a citizen wants to have from the neighbourhood and develop a self-contained sensory box that can be embedded deeply into the urban infrastructure.

2.3 Sensory and Crowdsourcing Systems

Sensory and crowdsourcing systems have been proposed to accommodate the fine-grained sensory view of an urban setting including both quantitative and qualitative fronts [4, 6, 17, 20]. In the context of participatory and opportunity sensing, a series of projects have been developed to maximise the sensory coverage of a wide-city by leveraging mobile users and vehicles. For example, PotholePatrol [17], Nericell [31], and CommuniSense [43] assess the state of roads and traffic conditions by leveraging sensor-equipped vehicles. Ear-Phone [40] and NoiseTube [29] use mobile phones to measure noise levels and create a noise map of urban areas. Another direction of extending the information coverage is to leverage crowdsourcing, i.e., by allowing a group of citizen to manually provide the requested information of a city. Alt et al. presented a location-based crowdsourcing platform that distributes tasks to workers by integrating location as a parameter [6]. CrowdOut [9] is a tool to report traffic related infringements and problems to local authorities, such as illegally parked cars, broken signs, and signals, or road quality in general. In [4], the authors presented a wearable crowd-sourcing system that embeds crowdsourcing tasks to the daily routines of a mobile workforce. While these sensory and crowdsourcing systems mainly leverage mobile users and vehicles in order to maximise the coverage of the information in an urban setting, Lingo is designed to be embedded deeply into the urban infrastructure and provide citizens with a handy access to the information of a city.

2.4 Citizen Science

To understand citizen experiences at a level of spatio-temporal granularity, several studies have been made to devise computational methods that automatically profile urban areas and quantify citizen experiences such as recognisability, walkability, and happiness. In [38], the authors examined the urban features that contribute to the walkability of a city based on the social media data of Flickr and Foursquare. Venerandi et al. [49] also leveraged user-generated content (Open Street Map and Foursquare) to profile urban neighbourhoods in terms of functional advantages, which was then used to automatically uncover socio-economic deprivation of urban areas. In [39, 54] explored the recognisability of a city by proposing a new image ranking technique that identifies memorable city pictures based on the prediction of whether a neighbourhood makes people happy. In [39], the authors used classical text mining techniques to infer citizen happiness from Twitter conversations, and embedded sentiments. While these works attempted to quantify subjective citizen experiences using online data, we focus on hyper-local purposeful information of a neighbourhood.

3 CONTEXTUAL STUDY

We begin by reporting a mixed-method contextual study to uncover the information affinity and interaction modality concerning hyper-local information services. In particular, we are interested in identifying a set of questions that can serve as a representative sample for Lingo. We first describe an online survey followed by a semi-structured interview study.

3.1 Online Survey

3.1.1 Participants and Methods. In order to gain a broad understanding of topical coverage of formative preferences of potential users of a hyper-local information system, we have conducted an online survey over Amazon Turk. We have not placed any constraints on the participant their gender, age, where they live, etc.

After carefully reviewing related work, we have come up with a series of multiple choice and multiple answer questions to determine the set of information types users are interested in and the modality of communication they prefer to retrieve hyper-local information. In addition, the survey questions retrieve participants' behaviour with various modalities of smart device interaction, whether they are accustomed to conversational agents and if/how they currently acquire information about their neighbourhoods. We compensated each participant for completing the survey with an amount of $ 1.00.

We have collected responses from 1992 participants (51.9% male, 48.1% female). The ages of the participants range from 19 to 70 and 58.1% of them are between 25 and 40. The participants are distributed across five continents with 80% of them living in North America and Europe. When we asked them to rank their digital mindset on a scale between 1 and 10 (1 lowest, 10 highest), 87% responded that it was above 5. We were able to collect all the responses within 24 hours.

3.1.2 Results. 90% of the participants indicate that they would like to be more informed about their surroundings in the city they live. 88.9% of the people said they would be interested in effective approaches to access information about their neighbourhood that do not require lengthy search sessions. There are two specific aspects on which we delved deeper:

(1) *Information Affinity*: 45% of the participants indicated that they would make health related queries to gather information including air quality, allergens, noise level, street cleanliness, etc. For 50.9%, community-related questions to provide information about new shops, social events, places with the best overall mood, crowded areas, etc. 59.2% have said that they like to inquire infrastructure related contents such as public transportation, planned street works in an area. Safety related requests such as contacting the police, receiving and submitting warnings are important for 48% of the participants.

(2) *Interaction Modality*: We found that most people mentioned that they would use a conversational agent either using voice or text (chat bot). Voice was the preferred medium for 60.5% of the participants while text was the next preferred medium accumulating 43.8%.

This survey essentially allowed us to identify four critical topical areas of hyper-local information, namely - *health, infrastructure, community, and safety*. In addition, we also have identified that voice and text are two preferred modalities to obtain such information. Grounded on this initial finding, we then move to a more formal contextual study with semi-structured interviews to further understand information affinity and interaction modalities.

3.2 Semi-Structured Interviews

3.2.1 Participants and Methods. We follow the online survey with a two phase study. In order to see whether the responses from a global audience obtained from the online survey holds in a more culturally restrained setting, i.e. one particular country, we recruited 21 individuals (13 men, 8 women, age range 25 - 72) from our research facility in Belgium following stratified sampling with snowball sampling within each stratum. Each participant is compensated with a 10 € gift card.

We interviewed the participants following a technique called laddering. We asked them to answer the same set of questions as those in the online survey. Each interview took about 45 minutes. We completed all the interviews in 3 days. Unlike the survey, we allowed participants to provide open-ended answers and then asked them to elaborate their responses to understand the type of information they are interested in, what communication modality they prefer and whether they are comfortable with other modalities, their expectations and concerns about a possible system, etc. We recorded and transcribed these interviews and coded the individual responses using an affinity diagram to extract keywords and analysed them with thematic analysis. Through this analysis, we came up with 20 representative questions where a hyper-local information delivery system can be used to answer in the same cultural context.

In Phase 2, we sent out the list of potential questions questions to these participants over email and asked them to rank their importance concerning their household and lifestyle necessities.

3.2.2 Phase One Results. We report the interview results from two perspectives.

(1) *Information Affinity*: The participants essentially echoed our survey responses highlighting their desire for neighbourhood information concerning health outbreaks (e.g., pollen, flu, etc.) (n = 11), community events (e.g., street party, local school events, etc.) (n = 9, infrastructure related (e.g., construction schedule, noise management, etc.) (n = 12) and safety related (e.g., petty crime, etc.) (n = 10). Most participants have acknowledged that this system can help them engage with their neighbourhoods in a more informed manner. Some participants (n = 15) mentioned that they are not as aware of the city they live in as they would have desired. They typically use social media or monthly city magazines to learn about events where they live; however, filtering information generally is a problem.

(2) *Interaction Modality*, 8 participants have expressed that they prefer voice-based interaction for such information access. Multiple (n = 13) of them mentioned conversational agents available in their homes today, e.g., Siri, Alexa, etc. 11 participants have said they would prefer mobile applications to access such

Table 1. List of questions identified from the survey and interview studies sorted according to the ranked score received.

ID	Question	Category	Type	Property	Score
1	What is the pollen count in the street?	Health	Dynamic	Realtime	4.84
2	How crowded is the nearby street now?	Infrastructure	Dynamic	Realtime	4.81
3	Has the garbage collector come already today?	Community	Dynamic	Historical	4.79
4	Is there a garbage collector on the street now?	Community	Dynamic	Realtime	4.73
5	Has the postman come already today?	Community	Dynamic	Historical	4.73
6	Is there any postman now on the street?	Community	Dynamic	Realtime	4.71
7	Are there many dogs on the street in the area right now?	Safety	Dynamic	Realtime	4.43
8	Is the street quite now?	Safety	Dynamic	Realtime	4.29
9	Is the bench on the street empty?	Community	Dynamic	Realtime	4.17
10	Is there a snack van on the street?	Infrastructure	Dynamic	Realtime	3.91
11	Is there a free table in the cafe?	Infrastructure	Dynamic	Realtime	3.89
12	How is the humidity on the street now?	Health	Dynamic	Realtime	3.67
13	Is the street warmer than yesterday?	Health	Dynamic	Historical	3.61
14	Is it safe to walk alone here?	Safety	Dynamic	Historical	3.46
15	I smell gas, whom do I alert?	Safety	Static	Historical	3.32
16	Can I take my dog inside the town hall?	Infrastructure	Static	Historical	3.19
17	Does the street have security camera?	Safety	Static	Historical	3.01
18	Is there a pharmacy in the street?	Health	Static	Historical	2.93
19	What is the number for calling an ambulance?	Health	Static	Historical	2.86
20	Where is the nearest garbage bin?	Infrastructure	Static	Historical	2.77

information due to their familiarity with chat agents and messaging apps; however, the majority (n=20) said they are comfortable with using voice as a communication modality.

Overall, all participants have expressed that they would like to use hyper-local information service if available. However, one general concern was the possibility of abuse by a provider to flood users with advertisements leading to information overload. One participant said

> *"I don't want yet another way of receiving advertisements. If it is only the information I request, then it is fine.."*

As a remedy, some participants noted that the system must not just push information, but it should only respond to specific queries. Another point of discussion was privacy. While the participants are willing to share limited data to receive, they are most interested in public information inquiries such as an event in the area, housing, health conditions, public transport, etc. Because the system provides spatiotemporal public information rather than a personal service that requires user data, it was perceived positively.

The analysis of interview sessions helped us identify specific questions of interest in the previously determined topics, i.e., health, infrastructure, community, and safety. Using affinity diagrams, we coded participants' comments to extract keywords, cluster and then generate a list of 20 questions in these four categories. We further characterised these questions based on their content type (static or dynamic) and temporal nature (real-time or historical). Dynamic questions are highly dependent on spatiotemporal data and challenging to obtain from online services. For example, *"Are there many dogs in the street in this area?"* can not be obtained using a web search. On the other hand, a static question can be answered using already available tools and data on the web, e.g., *"Can I take my dog inside the town hall?"* The final list of these questions following this session is depicted in Table 1.

3.2.3 Phase Two Results. Once we have retrieved the set of questions through the thematic analysis over the transcription of the group discussion, we sent out the list of questions to these participants over email. We asked our participants to rank all questions on a Likert scale from 1 (not important) to 5 (very important), reflecting their importance based on their household and lifestyle necessities. Our objective here was to validate to what extent we have accurately captured their collective feedback from phase one. Table 1 questions are sorted according to the score received for each question. We can observe that 17 out of 20 questions (both static and dynamic) received a score of above 3 (midpoint) which we consider a good indication of their agreement. Please do note that we did not apply a more established metric on information quality in this phase. However, we revisit this assessment in our usability evaluation with more established metrics.

4 LINGO: DESIGN DECISIONS

The previous section described the contextual study that informed us on the information affinity and interaction modality. We also extracted 20 questions (14 Dynamic and 6 Static) that we consider is a good starting set to evaluate the notion of hyper-local information services. In this section, we discuss four design decisions that we followed to develop our Lingo solution.

Automated Multimodal Reasoning to Offer Dynamic Information: Many location-based services today use location as a critical context to serve location-based information [15, 24, 41]. We have also seen remarkable growth in video analytics systems to offer human-like reasoning with machine imaging [8, 22, 51]. Similarly, recent research on audio analytics has shown the ability of machine-learning models to understand ambient sound beyond human speech [25, 26, 34]. Naturally, grounded on this research, our first design decision is to develop an automated multimodal reasoning engine using a camera, a microphone and environment sensors. We posit that by combining these sensors with lightweight machine learning models, we would characterise neighbourhood situations automatically and dynamically to serve user queries. To this end, in Lingo, we have developed a brand-new edge device, called Lingo Observer, that uses multimodal machine learning models operating on camera, microphone and environment sensor to answer 14 dynamic questions. In addition, this component maintains a set of fixed information to serve static questions. We discuss the specific models and their usage in Lingo Observer in the later section.

Multimodal Question-Answer as Interaction Primitive: Our next design issue is the access mechanism to our multimodal reasoning engine. Our contextual study informed users desire to interact with neighbourhood information services with voice and text. With an incredible proliferation of conversation agents in recent years, it is natural that voice interactions need to follow a question-answer paradigm established in the current conversational user interface domain. Text-based interaction can also follow this pattern, although it can be automated with predefined questions. Accordingly, in Lingo, we have developed Lingo Agent that serves as user interface for the users in two different forms. First, a smart speaker offers a conversational experience, and second, a smartphone application offers a text-based experience to interact with neighbourhood information services. We have used state-of-the-art NLP engines to accommodate these interactions and extract the intent of an interaction that is then mapped to our multimodal reasoning engine capability. To support this mapping, we devise a new service descriptor abstraction that we call *codelet*, and implemented in our solution to serve both real-time and historical questions. These details are further discussed in the following section.

Hyper-local Access with Disconnected Covert Communication Channel: One of the fundamental issues reported in urban computing literature is the lack of awareness of service availability in a neighbourhood. Our objective here is to serve hyper-local information of the neighbourhood to its citizens. So, the challenge is to bridge this awareness gap with easy accessibility and constant availability. One obvious choice is to connect our reasoning engine to a global service point over the Internet, which demands dedicated stakeholders managing the service. We argue such service management can not be achieved easily at fine granularity, for instance, at a street level through a centralised system. Besides, such a system begs privacy and data protection, which further complicates the practical management issues. We argue one way to address this is to bring the principle of proxemic interaction [18, 30] which has shown success for interacting with urban landmarks, such as a public display. These interactions do not demand a global service and communication infrastructure, rather a co-location of the user and the service provider. In Lingo, we borrowed this principle and developed a proxemic interaction channel using covert communication with Wi-Fi. Here we use Wi-Fi instead of near-field communication protocols such as Bluetooth, Ultra-Wide Band or ZigBee for two reasons - the ubiquity of Wi-Fi and the range of Wi-Fi. However, we also wanted to minimise the access control challenges typical to a Wi-Fi network. As such, we developed a covert communication protocol using Wi-Fi management frames that do not require users to authenticate themselves to receive the Wi-Fi service, in this case, interaction with the Lingo reasoning engine. As long as a user is within the range of Wi-Fi coverage of Lingo, i.e., proxemic, a user can freely interact in a disconnected fashion. This design allows us to offer access to hyper-local information without any global service provider, instead of just being co-located with the service. We describe the detail of this covert communication protocol in the following section.

Privacy-Preserving Data Management: We integrate Lingo in urban landmarks for gathering information and events related to a neighbourhood to serve its citizens. Naturally, such a life-log of neighbourhood demands discussions concerning data privacy, ownership and protection. We consider the breadth and depth of this issue - data protection of public information - is out of the scope of this paper. However, we have made two design decisions to conform to General Data Protection Regulation (GDPR). First, we do not store raw sensor data in local storage. Instead, only model inferences are stored to serve historical queries. Second, we only serve processed information, e.g., answer to questions, ensuring zero access to the raw data stream. We argue that these two design decisions provide minimal but adequate support to meet the GDPR concerning public surveillance.

The following section describes the Lingo system that implements these design principles in its constituent components.

5 LINGO: SYSTEM DESCRIPTION

The Lingo system consists of two components, *Lingo Observer* and *Lingo Agent*. A Lingo Observer is a self-contained sensing box that obtains and offers hyper-local information. We envision the Lingo Observers will be embedded on local landmarks in an urban landscape such as a light post, a building, a tower, each responsible for its vicinity to provide spatiotemporal information. Lingo Agents are end-user devices that have the capability to issue queries by interacting with Lingo Observer. These components share a proxemic communication protocol that allows them to exchange query and response messages through a covert channel that does not require a persistent connection.

5.1 Lingo Observer

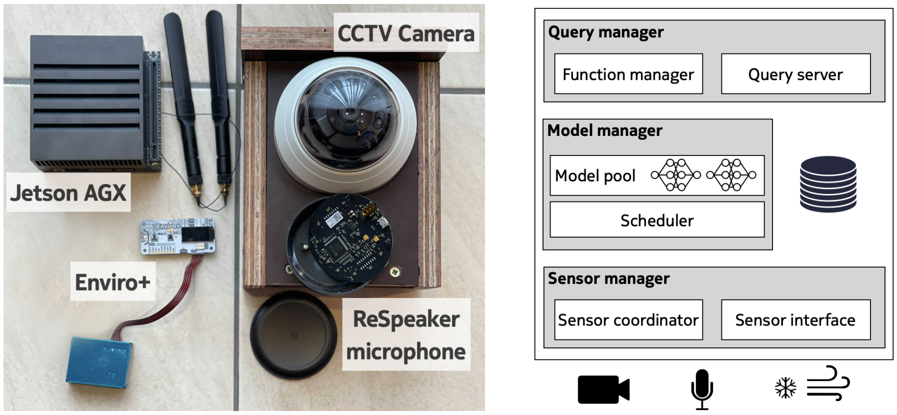
We prototyped the Lingo Observer using Nvidia Jetson AGX [2] (a GPU-powered embedded board released by Nvidia), and three types of sensors (a camera, a microphone, and an environment sensor) as shown in Figure 2a. The Nvidia Jetson AGX board hosts a 8-code Nvidia Carmel Arm, a 512-core Nvidia VoltaTM GPU with 64 Tensor Cores able to deliver up to 32 TOPs, and 32 GB of LPDDR4 RAM, having the size of 105 mm × 105 mm × 65 mm. For sensors, we use a) a Samsung QND-6030RP CCTV camera [3], b) ReSpeaker Mic array [4], and c) an Enviro+ module [5] for environment sensing. The QND-6030RP camera has build-in 6mm fixed lens and provides

[2] https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit

[3] https://www.hanwha-security.com/en/products/camera/network/dome/QND-6030R/overview/

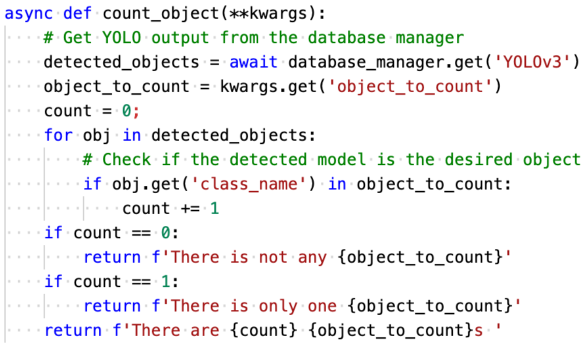
[4] https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

[5] https://shop.pimoroni.com/products/enviro?variant=31155658457171



(a) Hardware setup (b) System architecture

Fig. 2. Lingo Observer.



```
async def count_object(**kwargs):
    # Get YOLO output from the database manager
    detected_objects = await database_manager.get('YOLOv3')
    object_to_count = kwargs.get('object_to_count')
    count = 0;
    for obj in detected_objects:
        # Check if the detected model is the desired object
        if obj.get('class_name') in object_to_count:
            count += 1
    if count == 0:
        return f'There is not any {object_to_count}'
    if count == 1:
        return f'There is only one {object_to_count}'
    return f'There are {count} {object_to_count}s '
```

Fig. 3. A snippet for a codelet function for Question 7.

the maximum 2M resolution (1920×1080). The ReSpeaker Mic array is an array of 4 high performance digital microphones, specially designed to have the improved voice quality. The Enviro+ module has a set of environment sensors: BME280 temperature, pressure, humidity sensor, LTR-559 light and proximity sensor, MICS6814 analog gas sensor. We also connect particulate matter (PM) sensor to monitor the air-quality.

Figure 2b shows the system architecture of the Lingo Observer composed of three main components.

5.1.1 *Query Manager.*

The query manager is responsible for computing responses to Lingo Agent queries. The query retrieved through the proxemic communication protocol (§5.3) includes a function identifier `function_id` along with a number named arguments. The query manager provides a light-weight interface *get_response(<function_id>, **kwargs)* that prompts the execution of the corresponding codelet function. Figure 3 illustrates a code snippet for a codelet function. A codelet function typically reads model outputs from the local database that the model manager writes into. The codelet interprets output from at least one model and returns an answer in text. Each function answers one or more questions described in Table 1.

Table 2. List of models in Lingo Observer.

Task	Sensor	Model	Supporting questions (labels of interest)
Crowd counting	Camera	CrowdNet [12]	2 (number of people), 9 (number of people), 11 (number of people)
Object detection	Camera	YOLOv3 [42]	3 (garbage van), 4 (garbage van), 5 (postman), 6 (postman) 7 (dog), 9 (bench), 10 (snack van), 11 (table)
Sound event classification	Microphone	YAMNet [3]	8 (silence), 14 (car passing by, screaming,)
Environment monitoring	Enviro+	EnviroMon	1 (PPM), 12 (humidity), 13 (temperature)

5.1.2 Model Manager. The model manager is responsible for the execution of sensing models to offer the inference output to the answer functions. To support the *dynamic* questions in Table 1, the Lingo Observer adopts four sensing models as follows:

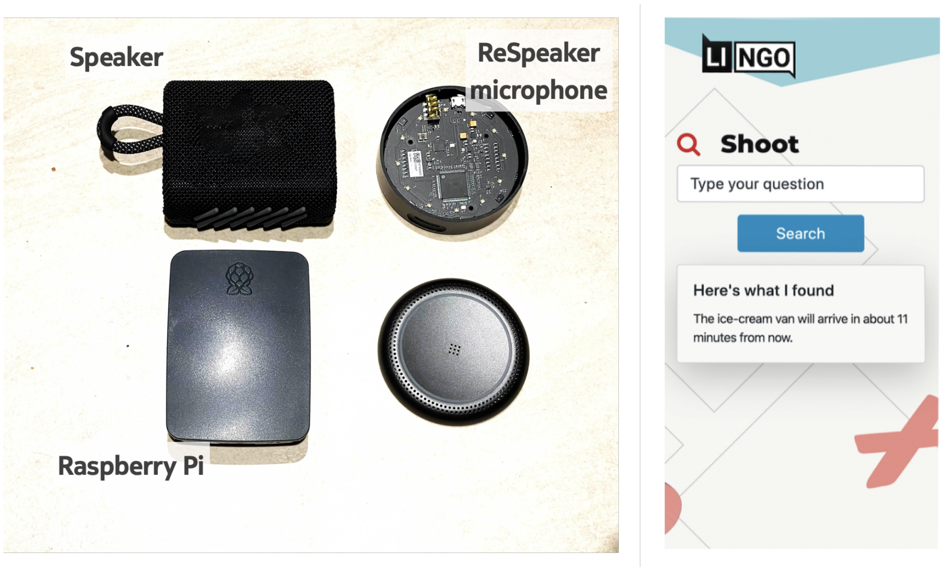
- **CrowdNet [12]:** CrowdNet is a deep convolution network specially designed for dense crowd counting. It takes an image with the size of 400×400 and outputs the estimated count of people in the input image.
- **YOLOv3 [42]:** YOLOv3 is an object detection model which takes an image (416×416) as input and outputs a list of detected objects with a bounding box. Since the pre-trained YOLOv3 does not provide the full set of labels required for our questions, we collected the images of the objects we are interested in and re-trained the YOLOv3 model.
- **YAMNet [3]:** YAMNet is a pre-trained audio model for the sound event classification, released by Google in 2019. It is trained with Google's AudioSet that contains 5.8 thousands of hours of audio data (16 kHz mono) and predicts 521 audio event classes.
- **EnviroMon:** We build our custom model, called EnviroMon, for environment monitoring. It processes the raw data of environment sensors by applying post processing operations such as smoothing and calibration.

Note that, for static questions, the Lingo Observer stores the relevant information to the local database prior to the deployment. When the sensor data is sampled, the scheduler triggers the execution of the corresponding model managed in the model pool and stores the inference output to the local storage. It is important to note that the inference of four models are performed continuously in the background in order to support the historical queries and minimise the Q&A latency.

5.1.3 Sensor Manager. The sensor manager manages the acquisition of sensor data. The sensor interface is responsible for communicating with the sensors installed in the Lingo Observer. It is implemented using the real-time streaming protocol (RTSP) to communicate with the CCTV camera, USB interface for the ReSpeaker Mic array, and I2S interface on J21 header for the Enviro+ sensor. The sensor coordinator reads the sensor data stream with the following configurations: video (640 × 480, 30 Hz), audio (16 kHz, mono), and Enviro+ (1 Hz), and resizes the image to 400×400 and 416×416 for CrowdNet and YOLOv3, respectively. Then, it forwards the data streams to the model manager to trigger the execution of model inferences. Note that the scheduler in the model manager discards the old samples if the execution time of the model inference exceeds the sampling interval of the corresponding sensor.

5.2 Lingo Agent

The devices that have the capability to issue queries are called *Lingo Agents*. These devices capture the user request, either in text or speech, infer the structured query using Natural Language Processing tools, send the query to the Lingo observer, retrieve the response and present it to the user in the same modality that it received the query. We have developed two classes of agents. In the first class, a smart speaker with a fixed location is equipped with a microphone and a speaker to serve audio queries. For mobile users, we provide an Android application that accepts both textual and audio queries.



Here's what I found

The ice-cream van will arrive in about 11 minutes from now.

Fig. 4. (a) Static Lingo Agent - Smart Speaker (left), (b) Mobile Lingo Agent - Android App (right)

5.2.1 Static Lingo Agent - Smart Speaker. A static Lingo Agent (Figure 4a) consists of a computation unit, a microphone and a speaker. In our implementation, the computation unit is a Raspberry Pi[6] and the microphone is a ReSpeaker Mic Array (the same microphone we used for the Lingo Observer). We envision a static Agent to act as an urban voice assistant that citizens can use to fetch information about their proximity. In addition, these Internet connected devices can be used to provide other services including other voice assistants.

Similar to commercial voice assistants such as Siri, Alexa, Google Assistant, Lingo Agent expects an explicit cue or a wake up command, e.g., 'Hey Lingo', to capture and handle audio from users. To do this, Lingo runs an off-the-shelf, lightweight Keyword Spotting model[7]. Upon hearing the utterance of this cue, the Agent records the audio using microphone, capturing the user's question in audio. This audio clip needs to be analyzed to understand what actual phenomenon the user is interested in through Natural Language Processing (NLP) and construct a formal query for the observer.

Once the question is answered by the Lingo Observer, the response in text is played back to the user. To achieve this, we a use commercial text to speech (TTS) engine[8]. This engine accepts textual input and returns an audio content.

5.2.2 Mobile Lingo Agent - Android application. In order to serve users directly from their personal devices, we also developed an Android application as shown in Figure 4b.

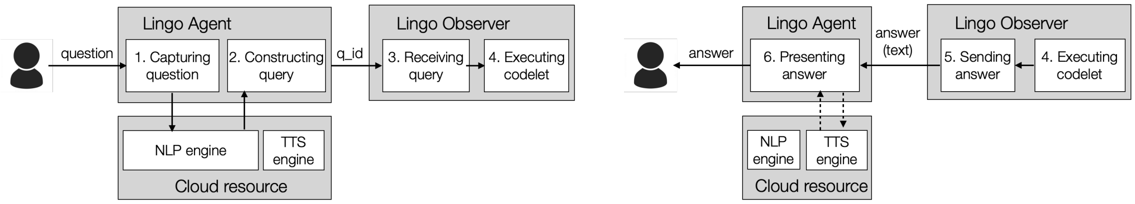
This application provides an interface for the user to enter the question in text. Similar to the audio interface of the static Agent, this textual question is processed to understand the intention of the user and construct a query for the observer. Once a response arrives from the observer, the application presents it to the user in text as well.

5.2.3 Lingo Agent Software. Though static and mobile Agents differ in how they capture the user questions, they use the common components to interpret user questions to understand their intentions, extracting the necessary

[6] https://www.raspberrypi.org/products/raspberry-pi-3-model-b/

[7] https://github.com/Picovoice/porcupine

[8] https://cloud.google.com/speech-to-text



(a) Pipeline for processing a question

(b) Pipeline for processing an answer

Fig. 5. End-to-end Q/A pipeline

arguments for the executions of the functions to prepare an answer at the observer and making a query to the observer.

In order to produce a query from a question in natural language, we use Dialogflow[9], a commercial off-the-shelf Natural Language Processing (NLP) engine. While open source NLP engines exist for local deployment, such models require a high amount processing and energy to perform intent matching and automatic speech recognition tasks. While Agent devices are not considered powerful enough to carry out these tasks, our goal in the observer is to allocate as much resources as possible to model deployment and inferences to capture local information. Instead, we use the Internet connectivity of Agents to offload these tasks to a cloud deployment.

Dialogflow allows developers to match expressions in natural language to a number of intents. It provides a set of APIs for applications to post audio recordings or texts for intent matching. As explained in §3, Lingo observers provide answers for a set of pre-selected questions each answered by a codelet function. Each intent prompts execution of a function that may use a number of parameters extracted from the user question. These parameters include date and time that are used to answer questions regarding information from the past. The response from Dialogflow may include other parameters that are necessary for the execution of the functions. For example, a function associated with object detection may accept an argument that filter the detected objects to a certain type.

The response received from Dialogflow includes a function identifier and a list of arguments that the Agent uses to construct a query. This query is then sent to the Lingo Observer a response is received using the covert communication mechanism explained in §5.3. The Agent then presents the response to the user either through audio or text.

The end-to-end question handling pipeline is summarized in Figure 5.

5.2.4 Observer Selection for Agents. Agents and observers use a Wi-Fi based covert communication mechanism to communicate and the agents rely on Wi-Fi scanning for the discovery of observers (§5.3). While WiFi scan returns all the access points nearby, our discovery mechanism can filter those can provide Lingo observer functionality. It also returns the strength of the signal received from each observer. This intrinsically limits the geographical scope as the discovered observers are within the Wi-Fi range of the agent.

When sending a question, the Agent selects the nearest observer that can answer the question. To do this, we follow a simple heuristic and select the observer with highest signal strength. In fact, the signal strength can greatly fluctuates and the Agent may momentarily receive a weaker signal from a closer Observer due to scattering and reflection. However, we disregard these effects to avoid running complex localization methods in the Agents.

[9] https://cloud.google.com/dialogflow

5.3 Proxemic Communication

In this section, we describe the proxemic communication mechanism between Lingo Agents and Observers to send user questions and answers. Rather than transport layer connection, Lingo uses a MAC layer covert communication, leveraging the ubiquity of Wi-Fi access points. This mechanism does not require authentication or association with an access point. Instead, it relies on IEEE 802.11 management frames to exchange messages. It also allows Lingo Observers to serve Agents directly without the need for service discovery, persistent channels, cloud based proxies, etc.

5.3.1 Background on IEEE 802.11 Standard. Wi-Fi is a very popular wireless networking technology with a high density of Access Points (AP), and provides a pervasive way to connect to the Internet all over the world. It is expected that by 2023 the total number of public APs will be nearly 628 million globally, up from 169 millions in 2018. Within the same time frame, the number of Wi-Fi capable handheld and personal devices will increase from 4.9 billion to 6.7 billion [1]. In addition to their density, APs intrinsically provide a fine-grained location information.

Wi-Fi networks are governed by the IEEE 802.11 standard [2]. The operation of the network, the communication between the AP and the devices is maintained through three types of layer-2 frames. *Data* frames are used to send data from the AP to a device, and vice versa. *Control* frames police the devices' access to the wireless medium without causing a collusion, i.e. by preventing two devices transmitting at the same time. *Management* frames on the other hand, are used to provide and maintain connectivity. Their functionality includes authenticating user devices and associating them with the AP, broadcasting AP presence and its service set identifier (SSID), probe requests for devices to scan surrounding APs, probe responses for APs to informing the scanning devices, etc. A device does not need to be associated with an AP to receive a management from it. In fact, even if a device is associated with an AP, it can exchange management frames with another AP. A particular type of management frame, *Action* frame, can extend the functionality of the Wi-Fi devices. General Advertisement Service (GAS) for example uses such frames to query APs to find higher layer advertisements including roaming associations that can be used to make decisions on associating with APs [19].

Wi-Fi APs serve the devices that are associated with them on a channel, i.e. a range of bandwidth spectrum, that depends on the physical layer standard. In order to send/receive frames to/from an AP, the devices need to adjust their frequency to the AP channel.

5.3.2 Lingo Covert Communication. We deploy our Lingo Observers to have local connectivity. In addition to computation, they provide a Wi-Fi AP functionality. Hence, in the rest of this section, we use AP and Observer terms interchangeably.

In order to encapsulate the queries, and the responses to these queries, we use action frames. This way, we can send up to 2.3 KBytes of data in a frame. While this is not enough to carry out general purpose communication, it is enough to support hyper-local queries and their responses. We follow a simple protocol to send queries and receive responses as illustrated in Figure 6, an explained below.

Once a Lingo Agent has query, the device first begins an active Wi-Fi scan to discover nearby Observers (1). To do that, it broadcasts probe request frames. Once an AP receives these frames, it responds with a probe response frame to announce its presence along with a number of *information elements* that define the service it provides. In Lingo, we add a new element to the probe responses that indicates the Observer can be used to serve hyper-local queries (2). If multiple Observers are discovered by the Agent, one Observer is selected using the mechanism explained in §5.2. Once the Observer is decided, the device then switches to its channel and sends the structured query encapsulated in an action frame (3). Since serving query involves execution of a codelet function, the Observer cannot immediately return the answer; however small it is. In order to save energy, the Agent, by default, turns off the power the system components including the Wi-Fi chip as much as possible. In addition, the

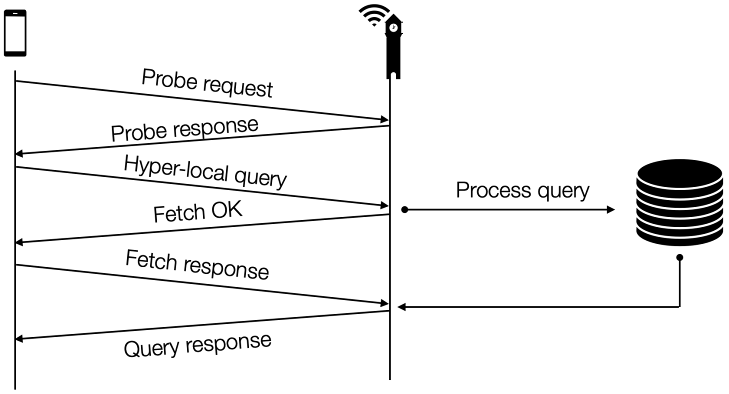


Fig. 6. Proxemic Communication Protocol

Agent may switch to another channel to for Internet connectivity. In order to receive a frame from the Observer, on the other hand, the Agent has to turn its antenna on and be on the same channel. To achieve this, the Observer sends the Agent a *Fetch OK* message (4) that prompts the Agent to send a *Fetch response* message (5). As we show in §6.1, the codelet execution occurs very fast and within the one round-trip time of management frames. Note that the handling of management frames takes longer than data frames as the first is processed in the user space of the device where as the latter is handled in the kernel. Finally, the Observer sends a *Response* message to the Agent (6).

In each interaction to serve a query, the query is initiated with a query identifier by the Agent. All the exchanged messages regarding this query include this identifier. Once the Observer computes the response to the query, and it arrives the communication unit the response is inserted into a buffer where the identifier points to this response. Once *Fetch Response* message arrives, the identifier it includes is used to retrieve the response from the buffer, which in turn is sent to the Agent.

Our covert communication mechanism is heavily influenced by WiPush[5]. One major difference is that the interaction is initiated by the Agent to query information rather that the access point pushing the information to the discovered devices without any prompt.

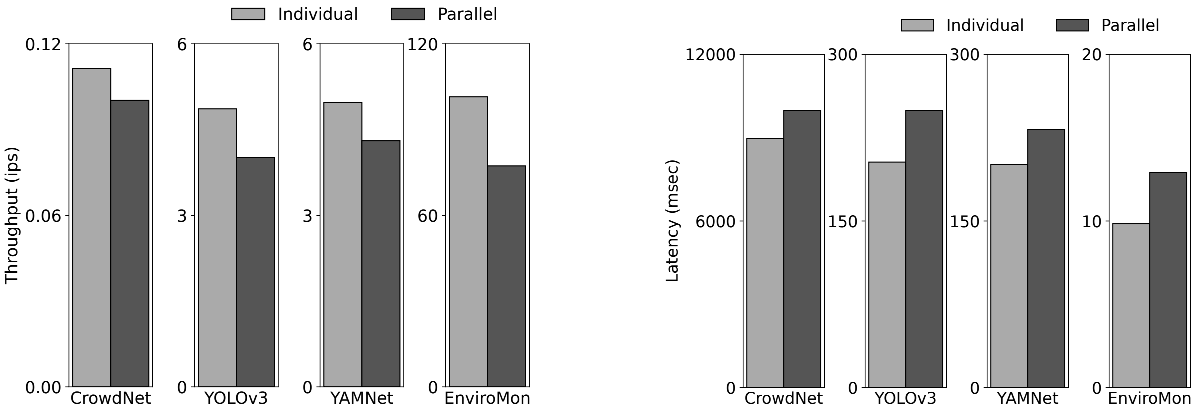
5.3.3 *Implementation.*

We implement Lingo by modifying the wpa_supplicant daemon in Agents. As discussed before, in our design Observers also provide AP functionality to communicate with Agents. In Observers, we have introduced the Lingo functionality in the hostapd daemon. These daemons handle management of Wi-Fi networks including discovery, authentication, association, etc.

While it is easier to modify the system daemons in a hardware platform such as Raspberry Pi and Nvidia AGX, it is less straightforward in Android devices. To achieve this, we use a Google Pixel 3 device and recreate its firmware using Android Open Source Project[10]. The applications in Lingo Agents and the query server in Lingo Observer, communicate with these daemons through system sockets to acquire messages that are sent to issue or answer queries.

Development for both daemons is based on version v2.10-devel. In total, it took 3726 lines of C code to add functionality both daemons, whereas the entire code base consists of roughly 650K lines of code.

[10] https://source.android.com/



(a) Inference throughput (inferences per second)

(b) Inference latency

Fig. 7. Computational footprint of Lingo Observer

6 LINGO: EVALUATION

In this section, we report a three-phase evaluation of Lingo. First, we present the benchmark of various system components concerning computation footprint and performance metrics. Next, we present a small-scale user study that reflects on the usability and utility of Lingo. Finally, we describe an in-the-wild evaluation of Lingo with real-world deployment among five households in two neighbourhood for one week.

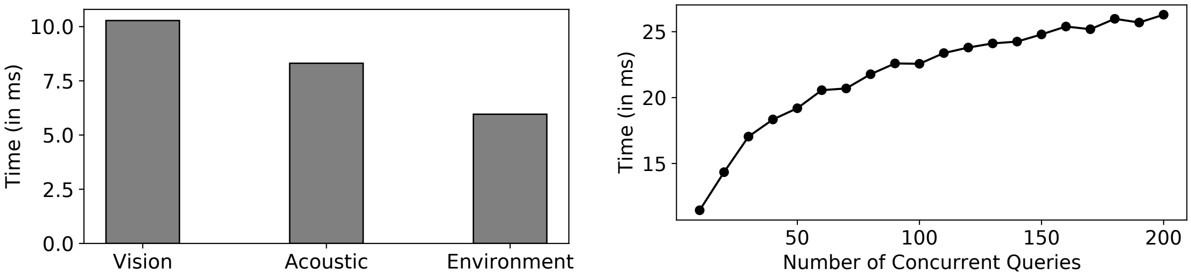
6.1 System Evaluation

We begin by reporting the computational footprint of various system component of Lingo.

6.1.1 Computational Footprint of Lingo Observer. We described in §5.1 that Lingo observer is composed of three components managing sensor data, model, and queries running on an NVidia Jetson AGX board. We report here the performance of two components, model manager and query manager, as they contribute to the overall experience of Lingo.

Model execution performance: On model management, we are primarily interested in the throughput (the number of inferences per unit time) and latency (execution time required for each inference) while running multiple models simultaneously. We have reported that for the current implementation, we have used four models, two for vision tasks (YOLOv3 and CrowdNet), one for the acoustic tasks (YAMNet) and one for environment tasks (EnviroMon). For this experiment, we run these models concurrently, i.e., the end-to-end inference pipeline of each model is loaded in the memory and fed with data with the maximum sampling rate and resolution of each underlying sensors - a camera, a microphone, and environment sensors. Furthermore, we use the batch size of 1, i.e., every data sample is fed to the model as soon as it is available for inference. Although recent literature has shown the runtime optimisation of multiple model execution, e.g., with selective batch size, dynamic data sharing, or GPU scheduling [14, 45], in this experiment, we did not apply these techniques.

Figure 7a shows the overall throughput of Lingo Observer while running these models in parallel comparing against the situation where each model is running individually; we consider such a situation as a baseline to understand the upper bound of the model performance. The results show that, even with parallel execution of four models, the Lingo Observer achieves 0.1, 4.0, 4.3, and 77.3 inferences per second for CrowdNet, YOLOv3,



(a) Codelet execution latency in isolation

(b) Codelet execution latency with concurrent queries

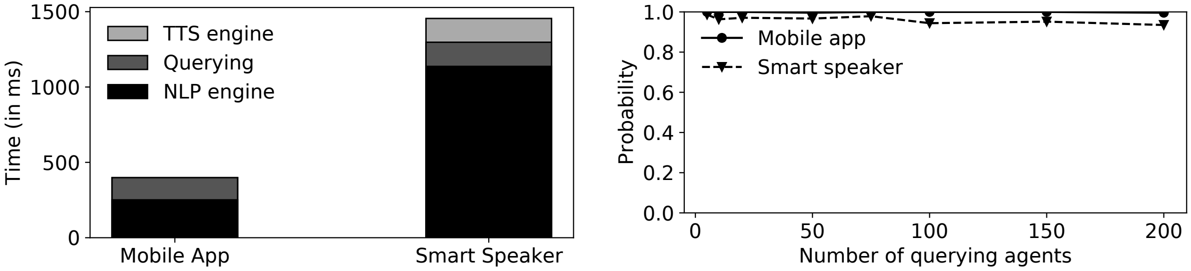
Fig. 8. Codelet execution performance

YAMNet, and EnvirMon, respectively. This result highlights that Lingo Observer can capture multiple events without compromising coverage to support real-time and historical queries. Interestingly, we can also observe that the parallel execution of four models do not sacrifice much compared to *individual* where a single model is executed; the throughput decreases 10% to 23%. Figure 7b shows the overall latency of Lingo Observer while running these models in parallel compared to the *individual* situation. We also observe that, even with the parallel case, four models achieve the latency of 9.97, 0.25, 0.23, and 0.01 seconds, respectively. This latency indicates that our current implementation, Lingo Observer, can adequately process various events within the tight latency target required to serve real-time queries. It is important note that, although CrowdNet takes 9.97 seconds for one inference on average, it does not mean that the user needs to have such a long latency. Since the Lingo Observer generates the answer based on the latest model output stored in the local database, this inference latency is not included in the Q&A latency. We also believe that, the interval of 10 seconds for crowd counting is reasonable considering human mobility and crowd dynamics.

These results also suggest that for the chosen set of models, our selected hardware board is adequate. However, we acknowledge that this performance strongly depends on the model size, architecture type and complexities, and the result reported here should be considered only for the specific set of models used in this experiment.

Query serving performance: We are primarily interested in the latency of the query manager in serving responses to user queries. Three different operations contribute to this latency - reading model output, executing a codelet function to prepare a query response and assembling the Lingo communication packet to send over Wi-Fi management frames. Figure 8a shows the latency of serving a single query for the three different types of questions concerning vision, acoustic and environment models. We can observe that Lingo Observer prepares a response to user queries within 15 milliseconds across these varieties of tasks. Next, we look at the scale of this serving capability. Figure 8b illustrates the average latency with an increasing number of concurrent queries concerning vision tasks. We notice that the execution latency increases up to 25 milliseconds when the Lingo Observer has 200 concurrent queries to answer.

6.1.2 Computational footprint of Lingo Agent. For Lingo Agent, we are interested in the end-to-end latency of serving a user query. There are a number of operations involved in this, including capturing the audio, processing the audio with a cloud-based NLP engine to extract the intent and respective question, broadcasting the question to Lingo Observer, receiving the response from Lingo Observer, generating the audio response using a cloud-based TTS engine and finally playing the audio. For the text-based agent, the NLP engine works on the textual input the user has provided and no speech generation is necessary. Figure 9a illustrates the breakdown of these operations aggregated over 1K queries, and we observe that the end-to-end latency for serving a single



(a) End-to-End query serving latency

(b) Query serving probability with increasing agents

Fig. 9. Query serving performance

query is within 1500 milliseconds for audio and 400 milliseconds for text. For this experiment, the Lingo Agents (smartphone application and smart speaker) were connected to a 100 Mbps Wi-Fi network. We notice that two sub-tasks contribute to the maximum latency incurred by this workflow - the cloud operations and the Lingo communication. The former is attributed to the respective cloud provider and beyond our control. However, the latter latency is yielded due to the best-effort mechanism of our Wi-Fi management frame-based communication protocol. Our experiments showed, the agent receives a response to its question once it is sent to the observer under 200 ms with 90% probability. While the end-to-end latency for query serving for audio takes considerably higher than text since automatic speech recognition is also applied, we consider this performance is acceptable, valuable and usable given it is still with a time that is comparable with commercial voice assistants.

As we described earlier, this protocol enables us to achieve our hyper-locality objective in a disconnected fashion; however, it comes at the expense of a best-effort communication that does not guarantee query serving. Figure 9b shows how the serving probability changes with the increasing number of agents. Instead of using actual users, we have programmed another device, Raspberry Pi, to artificially inject hyper-local queries and fetch response messages across the wireless medium with randomly generated MAC addresses. For each of these addresses, the messages are injected with a period of 10 seconds for 15 minutes. Our agents programmatically make queries at the same rate. Even though the mechanism is best-effort, a concentration of up to 200 agents does not diminish the delivery performance for either type of agent. The slight difference is due to device specifications. Still, consistently, our agents receive responses to more than 90% of their queries.

6.1.3 Lingo Accuracy. We investigate the accuracy of two system components of the Lingo system, model inference on Lingo Observer and query construction on Lingo Agent. It is important to note that these operations are built based on off-the-shelf models and libraries, and our goal is to demonstrate their performance on the target situation.

Model inference accuracy: To measure the inference accuracy of the models used in Lingo, we deployed the Lingo Observer at two places facing the nearby street of our research facility for one hour and randomly sampled 500 images and 200 audio clips in total. Then, the researchers manually tagged the ground truth information and measured the model accuracy. We did not include EnviroMon for the study because environment monitoring does not require a complex inference logic and its performance is highly tied to the sensor specification.

Table 3 shows the inference accuracy of the models deployed on Lingo Observer. For YoloV3 and YAMNet, we report the average F_1 score of all target labels. Overall, the models in Lingo show reasonable accuracy to capture the target labels required for serving hyper-local information. For CrowdNet, it shows unsatisfactory accuracy for precisely monitoring the number of people (the error ranging around ± 10), but its performance is

Table 3. Model inference accuracy.

Model	Target labels	Metric	Result
CrowdNet	Number of people	Mean absolute error	10.2
YoloV3	garbage van, postman, dog, bench, snack van, table, others	F_1 score	0.81
YAMNet	silence, car passing by, screaming, others	F_1 score	0.86

sufficient to support the relevant questions (2, 9, 11), i.e., the level of crowdness and whether place is empty or not. Surprisingly, the accuracy of the audio model (YAMNet) is lower than expected with a precision of 0.92 and a recall of 0.81. According to our investigation, the errors mostly occur when the audio event is made far from the microphone and the relevant audio signal is weakly captured. YoloV3 too performs worse than originally reported in [42] (precision 0.88, recall 0.75), since the model is trained on a more limited data that we could retrieve through 3rd party tools such as search engines.

Query construction accuracy: We investigate the query construction accuracy of Lingo Agent and measure the intent matching accuracy using both the text and speech modalities. We consider that intent matching is correct if the Dialogflow module outputs the intended question identifier. A query is correctly constructed if the user questions are matched to the correct intent by the Dialog agent. In order to evaluate the accuracy of static Lingo Agent, we recruited 10 people from our research facility (6 men, 4 women, age range 22 – 56) and asked them to speak out 20 Lingo questions. For mobile Lingo Agent, we made queries through text through the custom app.

The results show that Lingo Agent achieves 100% of intent matching accuracy with text input. Surprisingly, Lingo Agent also shows reasonable accuracy, 94% with speech. This is because we have a relatively low number of questions, i.e., 20, and the keywords of the questions are mostly well kept, e.g., postman, gas, and pharmacy. Thus, even with some words are incorrectly identified, the Lingo Agent rules in DialogFlow can correctly match the intended question. Moreover, the error rate varies by the individual participant varying between 85% and 100%.

6.2 Usability Evaluation

In the second phase of our evaluation, we conducted a small-scale usability study to understand 1) the utility of the questions served by Lingo, 2) accessibility benefits, and 3) the dynamics of two interaction modalities – text and voice.

6.2.1 Participants, Apparatus and Methods. We recruited 20 citizens (12 men, 8 women, age range 22 – 62) who live in a *smart zone* in Antwerp, Belgium[11]. For recruiting, we used stratified sampling with snowball sampling within each stratum. All participants owned a mobile device (a smartphone or a tablet or both) and a smart speaker and considered themselves digitally savvy with a score of minimum 7 out of 10. As they live in an area where a number of smart-city solutions are deployed, they are familiar with various state-of-the-art and cutting edge technologies. Our experiment followed a 2 × 2 factorial within-subject design with the different interaction modalities (voice and text) and the question categories (dynamic, static) as the factors. Seven dynamic questions and three static questions were randomly assigned to the conditions. The order of modality condition was counterbalanced and presented in blocks. For example, both question categories (dynamic, static or vice versa) with voice happened in one block without allowing the user to switch to text. With 20 subjects, this study recorded 80 trials.

[11] https://antwerpsmartzone.be/en/

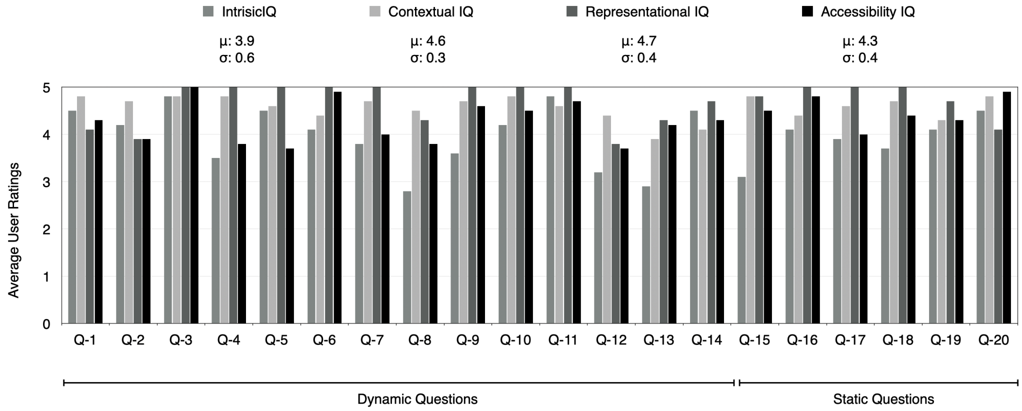


Fig. 10. Average ratings of users across different dimensions on the information quality of the queries served by Lingo.

For this controlled study, our setup included a Lingo Observer placed on a tripod facing a street in the smart zone, and the study room included Lingo Agents – both the Android application running on a Google Pixel 3 smartphone and the smart speaker as described in §5.2.

Initially, participants were given a written task description and were asked about their demographics and device experiences. Next, the participants tested each modality condition with a sample question, *"How is the weather today?"* in the training phase, first asking the Lingo Agent running on smartphone using text, and then asking the same question to the smart speaker over voice. Then for each condition, participants were given the set of questions (7 dynamic, 3 static) displayed on a tablet screen, first to find the answer without using Lingo Agent (with their own way such as googling, going outside), and then with the specific Lingo Agent (smartphone application in the first condition, and smart speaker in the second condition). For each question, we stipulated a maximum time limit of 180 seconds both without and with Lingo. We recorded the time required to acquire the answer by the participants. After each condition, the participants were requested to complete a SUS questionnaire[12]. After completing all conditions, we conducted semi-structured interviews with the participants and asked them to rate the quality of the questions. We borrowed metrics from the seminal work of Wang and Strong [52] on Information Quality (IQ) for this rating. Participants were asked to rate each question on four dimensions.

- Intrinsic IQ – covering accuracy, objectivity, believability and reputation.
- Contextual IQ – covering relevance, value-addition, timeliness and completeness.
- Representational IQ – covering interpretability, format, coherence, and compatibility.
- Accessibility IQ – covering accessibility and security

Each interview was audio-recorded for later analysis. We analysed this data by coding participants' responses using affinity diagrams. The total time for each session was about 90 minutes and each participant is provided with a 20 € gift certificate for the reimbursement of their time.

6.2.2 Results. We discuss the results from three perspectives, information quality, accessibility benefit and interaction dynamics

[12] The System Usability Scale (SUS) is a technology agnostic survey that is used to assess the usability of a variety of products or services. It is composed of ten statements contributing to a single score ranging from 0 to 100 [13].

Information Quality: We begin by reflecting on the utility of the information served by Lingo. Each question was ranked against four dimensions using a 5-point Likert scale. Figure 10 illustrates the average ratings of all participants of all twenty questions with dynamic and static characteristics. Interestingly, all questions were rated relatively high across all the users in all four dimensions. On intrinsic quality ($\mu = 3.9, \sigma = 0.6$), both dynamic and static answers were conceived to be accurate and objective. Follow up interviews with the participants shed some interesting insights, however. Participants found the objective nature of the responses (e.g., Yes or No) on highly granular aspects of a neighbourhood, such as *"Has the postman come already today?"* was very useful. Besides, they commented that they could not find such information on the web today for their neighbourhood. One particular comment by **P11** was:

> *"If I expect critical documents, I can't leave home before the post arrives and every day they arrive at a different time. It is not practically to keep checking the mailbox or ask a neighbour."*

Similarly, **P3** said:

> *"I thought the answers were very objective on issues that I could not know otherwise, like postman visit or pollen concentration, even by asking my neighbours..."*

However, Lingo's responses yielded a relatively lower score on this dimension compared to the other three. Our interviews revealed that the relative importance of the question is highly subjective and depends on individuals lifestyle demands and preferences. For instance, participants with pollen allergy found Q1 very useful, while the others did not. We observe a similar pattern for questions that involve pets or outdoor activities.

On contextual quality ($\mu = 4.6, \sigma = 0.3$), Lingo's responses were highly valued. In particular, the temporal nature concerning real-time delivery of current and immediate past events was considered very useful. One comment from **P18** was:

> *"What I liked about this system is that it can tell me about the recent past as well as what is happening now..."*

On representational quality ($\mu = 4.7, \sigma = 0.4$) and accessibility quality ($\mu = 4.3, \sigma = 0.4$), the responses were equally positive. On representation, the simplicity of the answers (such as "yes or no", "high or low") was highlighted by the participants as critical for their interpretation. Similarly, on accessibility, the fine-granularity of the responses was key for participants positive ratings.

> *"It is good that the responses don't leave anything for interpretation. I don't need to know exact number for humidity, only if it is too humid for comfort..."*

Although one of the unique aspects of Lingo is its ability to answer questions that are dynamic in nature, requiring access to real-time data in finest granularity at the spatial scale, we did not observer any significant difference in the overall information quality score between dynamic and static questions. We run a Chi-square test to confirm this aspect ($p > 0.5$). We delved deeper into this aspect during our interviews, and it was revealed that most participants did not differentiate the actual retrieval mechanism, rather rated on the utility of the information delivered concerning timeliness and accuracy. According to **P6**

> *"Text or speech doesn't make any difference as long as I receive correct information without latency"*

Accessibility: Next, we look at the accessibility aspect of Lingo, and in particular, we are interested in understanding whether Lingo offers faster access to accurate information than traditional sources. e.g., a web search engine. As illustrated in Table 4, Lingo can significantly reduce the time to access such hyper-local information up to 18.7× and 25.4× for static and dynamic questions respectively. However, please do note that, out of 14 dynamic questions, all participants did not manage to find any answers for several questions (Q3, Q5, Q9, Q13 – that look at recent past), and for some questions, their answers were not accurate or granular enough (e.g., Q1 – pollen count, Q12 – street humidity). Finally, most participants physically went outside the room for the

Table 4. Average time required to find the answers of the question without and with Lingo.

	Static Questions (Average Time to Complete)	Dynamic Questions (Average Time to Complete)
Without Lingo Agent	22.64 seconds	96.5 seconds
With Lingo Agent	1.2 seconds	3.8 seconds
Time Reduction	**18.7x**	**25.4x**

rest of the questions to check the street. This resulted in 14% of questions not answered within the 180 second timeout. Moreover, every participant experienced timeout at least once. On the other hand, Lingo answered all the questions before timeout. Participants highlighted these aspects as the key value of a system like Lingo. One particular remark from **P11** was:

> *"The fast access to this detailed information is impressive, and I can see myself using such systems regularly, hopefully with more information..."*

While information is provided quickly with Lingo, 90% confidence interval for query completion time is between 0.7 and 1.9 seconds for static questions, and 0.9 and 7.5 seconds for dynamic questions. In particular, computing answers to historical queries take shorter than real time queries ($\mu = 4.1, \sigma = 3.7$, e.g. Q3 vs Q4. While real time queries prompt a model execution; historical queries, just like static questions, are answered by retrieving values from the local data store.

Interaction Dynamics: Finally, we shift our focus to interaction dynamics for Lingo Agents, and in particular, we wanted to understand the overall usability of Lingo Agents concerning interaction modalities. The SUS scores for text was $\mu = 79.83, \sigma = 8.61$ and voice was $\mu = 78.13, \sigma = 2.13$. We did not expect any conclusive results here, given the proliferation of voice-based interfaces in recent years. Regardless of this, we conducted a one-way repeated measure ANOVA on SUS scores and did not observe any statistical significance ($p > 0.5$). We reflected on this in our follow-up interviews. Most of our participants suggested that they do not have any preference over the two mechanisms, given at home, they are comfortable talking to conversational agents. In an outdoor setting, participants also commended the text-only interface as the right design, which was, of course, informed by our contextual studies as discussed in §4. One particular remark from **P2** was:

> *I like the fact you have two options, as I will not be talking outside with my mobile agent, but would do comfortably at home, it has turned into a habit now...*

In recent literature, we have seen several studies addressing the social acceptance of conversational agents and interaction dynamics across different modalities [10, 16, 21, 46, 53, 55]. This result adds support to those research. For Lingo, we consider both interfaces are helpful and practical.

Besides these three key aspects, our interviews also revealed interesting aspects concerning spatial and topical coverage of the questions and privacy issues. We will discuss these aspects in a later section.

6.3 In-the-Wild Evaluation

In our final phase of the evaluation, we have deployed our Lingo solution to two streets in Belgium (Sint-Katelijne-Waver and Hoboken) and engaged five households for one week. This section reports on the quantitative and qualitative assessment of Lingo during this in-the-wild evaluation phase.

6.3.1 Participants, Apparatus, Methods. We recruited five households in two neighbourhoods of Belgium using the mailing list of our research facility. Four of the households had two members (couple), and one had four

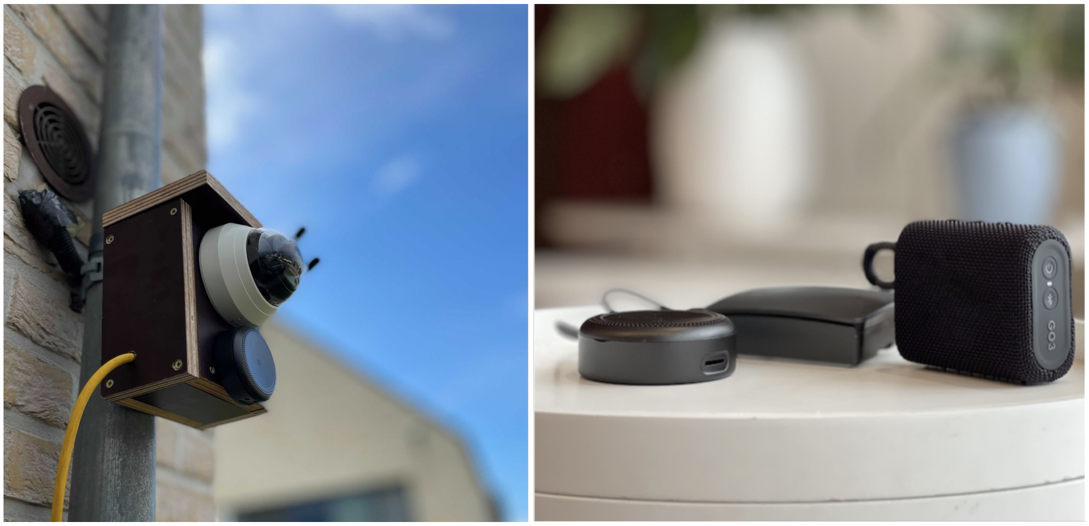


Fig. 11. Lingo Observer installed on the wall pole of a household (left) and Lingo Agent placed in the living room of a household (right).

members (parents and two children). All households had smart speaker-based conversational agents and several other connected devices and should be considered digitally savvy.

We installed Lingo observer in the wall pole of two households, one in Sint-Katelijne-Waver and one in Hoboken. We modified the Lingo Observer and put bigger storage as the unit recorded the raw video and audio throughout the deployment. We powered the units using extension cables connected to the external power units of the two households that participated in the study. We provided a Lingo Agent smart speaker to each home and installed the Lingo agent mobile applications to at least one phone in each household. Figure 11 illustrates the deployment of Lingo Observer in a street lamp post and Lingo agent in one of the households. In addition, we placed a post with a camera sign to notify citizens about the camera recording.

After the installation we gave a short demonstration of Lingo to each household and then requested one of the members to perform two queries using Lingo agents - smartphone application and smart speaker. After that, we provided them with a printed list of twenty questions and requested them to ask these questions to Lingo Agents at various points based on their need for one week. Finally, after one week, we revisited these households and collected the units. At this point, we conducted an in-depth interview with all household members who used the Lingo solution, and discussed various aspects of Lingo and their experiences. The interviews were audio recorded and afterwards transcribed for coding using an affinity diagram to extract keywords and analysed them with thematic analysis. Each interview took about 90 minutes. Each household received a 100 € gift card for the their participation in our evaluation.

During the deployment period, the Lingo observer recorded all traces, both sensory input and model output, in local storage, and every single question asked during the logged both in the Lingo observer and in Lingo agent. The raw sensor data is stored only for the sake of performance evaluation and not necessary for the operation.

6.3.2 Quantitative Analysis. We report the results of our deployment study from two aspects. We first report on various quantitative aspects of the deployment phase. Table 5 summarises the usage of the Lingo during the deployment phase among the five households over one week. As we can see, on average, 27.2 questions were asked per day, which we consider relatively high for conversational agents. Later interviews revealed that several

Table 5. Various quantitative metrics reflecting the results of the Lingo Deployment on five households for one week.

Aspects	**Quantification**
Total questions	191
Average questions per household	38.2 (σ = 7.1)
Average questions per day	27.2 (σ = 3.8)
Total questions over voice	130
Total questions over text	61
Static questions percentage	17.8% (n = 34)
Static questions accuracy	100%
Dynamic questions percentage	82.2% (n = 157)
Dynamic questions accuracy	73% (n = 114)
Number of time each question asked	Max : 21, Min : 3, μ = 9.55, σ = 5.09
3 most popular questions	Q5. Has the postman come already today? (n = 21) Q1. What is the pollen count in the street? (n = 19) Q13. Is the street warmer than yesterday? (n = 15)

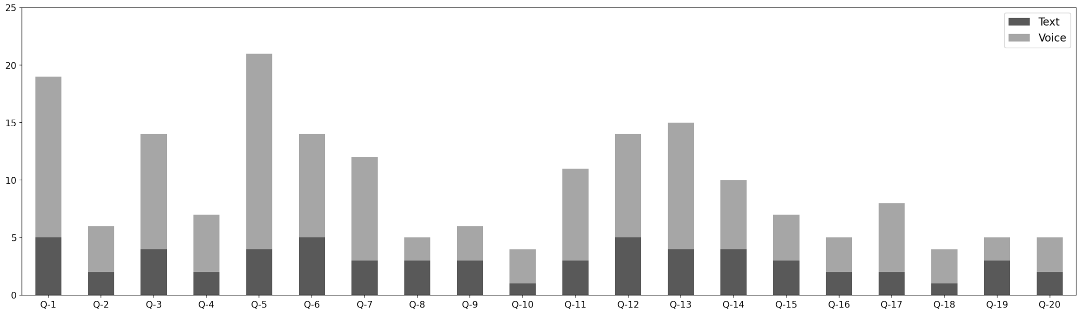


Fig. 12. Distribution of questions asked by users over the course in-the-wild evaluation

times the same questions were asked multiple times, or multiple questions were asked in short succession to test the system's boundaries.

As mentioned earlier, the timing of each question was logged, including start time, delivery time and the question itself. Thus, in total, 157 times dynamic questions were asked by our subjects. After the deployment, we inspected these 157 instances and compared the model and codelet output in response to query against the raw data through manual inspection (e.g., labelling the video and audio stream and checking environment sensors calibrated output). This inspection yielded an accuracy of 73%, i.e., 114 questions were accurately answered by the model and codelets of Lingo Observer. However, three aspects caused the failure of Lingo to answer the questions accurately. First, the model inaccurately classified wrong targets (e.g., postman or garbage van) either due to poor visibility or inadequate coverage (n = 26). Next, the questions arrived at the Lingo Observer incurred a delay due to the broadcasting latency, causing the live view to change and missing an object of interest, e.g., presence of a dog or the arrival of an ice cream van (n = 5). Finally, the mapping of the question from user voice input was wrong, leading to the default response from the Lingo observer (n = 12).

Concerning questions popularity, we noticed that questions related to the recent past were generally popular than real-time events (e.g., Q3, Q5, Q13). Furthermore, environmental conditions that can not be perceived naturally received attention from the subjects (e.g., Q1, Q12, Q13). Moreover, dynamic questions were more popular than static questions. Figure 12 shows the distribution of questions directed by users to Lingo Observers both using text and voice as a modality.

We have not engaged the subjects of our deployment study with a questionnaire on information quality. Instead, during our in-depth interview with each household, we discussed the overall utility and various other aspects of Lingo that we present next.

6.3.3 Qualitative Analysis. The final interviews with our participant households uncovered several interesting aspects concerning hyper-local information services. In this section, we reflect on these aspects. In the next section, we will further discuss the implication of these subjective experiences and the rest of our evaluation findings.

Information Quality: All of our participants collectively experienced Lingo very positively. The first highlight was Lingo's ability to capture ephemeral information to explain the recent past as a street-level granularity with immediate access. One participant mentioned that it is impossible today to access such hyper-local information without using in-home security camera footage for manual inspection. Information such as the historical presence of the postman or a specific object of interest on the street seems to be extremely useful for our participants. One particular remark from **H2** was:

> *The true benefit of your system is that we can revisit the past very quickly, I think this can solve a lot of critical everyday problems...*

The second highlight was the hyper-locality of information. Our subjects were very optimistic concerning the fine-granular coverage of Lingo together with timeliness. We received multiple remarks suggesting the hyper-locality concerning neighbourhood specific information as key benefits of Lingo as web-scale service providers do not offer at such scale. **H3** said:

> *Apps can't tell me what is happening near me. This feels like the while system works for only me*

On the other hand, they also noted some particular questions may be more suitable in an urban setting rather than a suburban street. **H2** commented:

> It is never very crowded here but this may be useful in the train station for example. I already know answers to some questions but they might be useful somewhere I am not familiar with...

The final aspect was the diversity of Lingo concerning information types. Our participants remarked that providing invisible metrics such as environment attributes next to audio/visual-based information was refreshing for them, especially when served in a single system. One comment from **H5** was:

> *I liked that you provided various types of information, including ones that it impossible for me to perceive, like pollen concentration. This information is important to me...*

Another remark that captured this aspect well was from **H1**:

> *The power of this system is that you have combined various helpful information necessary to my home in one unit, and the questions-answers were straightforward to make a quick decision which is vital for household chores...*

Accessibility: In general, participants were reasonably positive concerning the accessibility of Lingo, especially for the variety of hyper-local events. As we have observed during our usability study, our deployment participants also mentioned that it would have been impossible for them to know some of the hyper-local events without manually inspecting their streets. It was clearly highlighted that Lingo's ability to answer questions in real-time on things that are happening right now in a nearby place or recent past was considered vital benefits.

However, participants questioned the spatial coverage both concerning the event capture and communication. While they understood the capability of Lingo is spatially constrained, they wondered how to scale the system and what it would entail for their community. One of the critical aspects of Lingo was its ability to operate using covert communication channel in a disconnected fashion, allowing easy placement of Lingo observer on civic landmarks and dynamic scaling without massive infrastructure support. We brought this aspect during the interviews while discussing access and scale of Lingo services. Our participants were unaware of the technical details. However, it was a pleasant surprise that the communication between Lingo Observer and Agent is entirely local and has significant benefits concerning privacy.

While they appreciated this design and acknowledged the current limitation in building a distributed network of Lingo Observer, they wanted to extend the range of the coverage. We discuss this more in the next section.

Interaction Dynamics: The participants used voice to interact with Lingo 68% of the time. Interviews revealed that a smart speaker is a common feature in household routines among the participants, and it was most convenient for them to speak instead of typing while at home. Besides, they mentioned that mobile application would be helpful while they are away from home. This subjective experience echoes with our findings in the second phase of our evaluation. Overall, our participants found that the interaction with Lingo is positive. However, they highlighted two negative aspects. First, when Lingo fails to understand the question or to find an answer, it only replies with a default answer without any further feedback, which was frustrating for the users. Second, there is often a variable delay in responding to the question, and it was difficult for them to understand whether to wait or repeat the questions. These aspects were mentioned for both text-based and voice-based interactions. We acknowledge that this is a design fault of the current Lingo Agent manifestations. We plan to rectify them in the subsequent iterations.

In the next section, we further delve into few implications emerging from our three-phase evaluations.

7 DISCUSSION

In the previous section, we presented three different evaluations of Lingo. In this section, we reflect on the implications of the key findings of these results.

Topical coverage of the information: The contextual study conducted in §3 informed the current design and capabilities of Lingo. While the system design aspects (e.g., dynamic information through automated ML-based reasoning or covert communication for hyper-local proxemic interaction) manage to meet user expectations, we observed participants' desire to access more diverse information. Besides, some of the information we presented, for instance, allergen concentration or presence of an ice cream van, satisfy specific user groups. Furthermore, our participants' geographical, socio-economic and cultural context also shape the coverage of information for our prototype. We seek to draw attention to the broader UbiComp community on this aspect and expect future contextual studies to address various communities to bring more diverse information types and category to increase the utility of a system such as Lingo.

Spatio-temporal coverage of the information: One of the design principles of Lingo is to bind the observer to a local region. However, this hyper-local placement has limitations concerning information coverage and its accuracy and robustness. For instance, our current system can only infer events within the viewpoint of the camera or the microphone range. Besides, the Wi-Fi range, which is typically 30-50 meters, also limits the accessibility of the system. Our participants, in particular, who took part in the deployment study, mentioned that in many instances, they want to know the status of an event in the next street but immediately realised Lingo could not serve that. Similarly, they also realised that they need to be within the range of Lingo observer to interact with it, and pointed out that - in the outdoor setting, they would expect several Lingo Observers are present to avoid moving to a specific place for receiving the information. These issues can be addressed by scaling the Lingo Observer nodes and creating a mesh network among the observers. We plan to work on this in the future avenue of our work to extend the spatio-temporal coverage of the Lingo Observer. An alternative app-based approach where user location is used to send relevant information may on the other hand comes at the cost compromised privacy where neighbourhood-level data regarding citizens' daily lives are stored remotely.

Extensibility of the system: Extending topical and spatio-temporal coverage of Lingo essentially demands the ability to push new capabilities to Lingo Observer. In our current proof-of-concept, the capabilities are tight and hard-coded (e.g., a specific set of ML models). A more traditional app-based information delivery approach, where the user location is tracked and a remote server delivers the relevant information, can provide a larger set of capabilities. This requires neighbourhood-level information is stored remotely, introducing privacy concerns. With Lingo, we need to rethink this system aspect holistically as this challenge demands bringing agile and flexible Machine Learning Operations (MLOPs) with a multi-tenant model serving capabilities to edge devices like Lingo Observer. Although we have seen remarkable advancement in model execution on constrained devices, such cloud-scale continuous deployment and integration of ML-based services on edge devices require brand new architectural thinking. Some of the Lingo Observer design decisions, e.g., codelet and separation of model and query server, already provide some exciting options towards this objective. However, we would like to address this issue in our future work, taking a more systematic approach.

Privacy and anonymity: Another critical aspect of Lingo is the privacy-preserving manifestation of spatio-temporal events in Lingo Observer. During this study, we carefully applied the principles of General Data Protection Regulation (GDPR) in our solution, particularly for the deployment study. However, given the nature of data that such a solution can accumulate, it is essential to scrutinise the collection, storage and usage of information acquired using such a solution through a GDPR lens. Our participants brought this aspect multiple times to raise their concern on who gets to access the raw footage of these recordings. We are increasingly observing CCTV cameras in our street and security cameras on peoples home. The ownership of the data of these systems is relatively well understood or at least has a good service level agreement. This is not the case for Lingo, and common questions came to our discussion during the interview phase: who owns the data, who has access to the raw data, and manages the data, e.g., deletion or copy. These are essential questions, and we acknowledge that Lingo design does not answer many of these aspects that demand deep discussions across various stakeholders, including community and city corporations. In the current Lingo design, we have limited access to data to Lingo with serving queries anonymously. However, we anticipate that the materialisation and scaling of such a solution in the real world would require deep and thoughtful discussions. We call attention to the community working on privacy and data protection for the urban environment to consider systems such Lingo and respective implications. However, we expect system-level design to accommodate various principles should be straightforward.

We discussed an ethic review with our institution regarding the Lingo design and its deployment for evaluation, but they confirmed that IRB was not necessary per our institution's policy.

Community interaction: One of the recommendations we have received from our participants is to offer the capability to broadcast information by themselves, either using notification or current query based interactions. We did consider this aspect during the Lingo design, as it was mentioned in our contextual study. However, we did not accommodate this capability in the prototype to avoid misuse and advert publishing. During the interviews, our participants mentioned that currently, many neighbourhoods use dedicated channels in popular group messaging apps to inform and learn about community events. We received suggestions that a system like Lingo should be able to participate or access such information. Of course, such integration would break the fundamental design principles of Lingo, e.g., privacy-preserving, hyper-local proxemic interaction with automated reasoning. However, moving forward, we need to rethink whether Lingo could have dedicated group communication capabilities, for instance, by extending the current Lingo Agent application to facilitate community interaction.

Limitations: This research was conducted in Belgium. Certainly, the results presented here must be interpreted in the context of the culture and infrastructure in which they were performed. We expect our results are most appropriate for designers of urban applications for citizens in Europe or countries with similar cultures and levels of technology adoption.

Next, our set of questions was minimal and designed based on the outcome of the contextual study. This research offered both quantitative and qualitative assessment of this minimal set, but we do not consider them either complete or adequate. We sincerely acknowledge this as a limitation of our research. However, we hope that this research has offered an exciting foundation for further research to address diverse communities to derive a richer, more extensive and meaningful information set.

Message exchange through management frames with association means that anyone sniffing the wireless medium can inspect the frame payloads revealing the content of messages. Even though Lingo offers anonymous query serving, this may potentially lead to security risks. A large scale, real world deployment of Lingo needs to devise encryption mechanisms so that only the Agent and the Observer can extract the contents of the message exchange.

Finally, the scale of both our usability study and in-the-wild deployment was small. We want to mention that both studies were conducted during the ongoing COVID-19 pandemic, which significantly reduced our ability to recruit and deploy Lingo. Hence, the results reported here should not be considered as general, instead, interpreted in the context of the study setup as they may be subject to novelty effect. As such, further validation studies in a variety of different communities are necessary to assess and widely apply the implications of our system. We sincerely acknowledge these limitations. However, we hope that our faithful observations and data-driven insights uncovered interesting implications as reported here and inform the design and development of future solutions that offer hyper-local information to urban citizens.

8 CONCLUSION

In this paper, we present Lingo, a hyper-local conversational agent placed in urban landmarks that offers rich and purposeful spatiotemporal information relevant to a neighbourhood. We informed the design on Lingo based on a mixed method contextual study. A set of twenty questions served as the representative information for Lingo. We described various technical components of Lingo, namely – an observer serving as a hyper-local information source and an agent serving as a hyper-local information access point. We reflect on several technical aspects of these components including multi-modal reasoning engine, codelet, and covert communication that provide the backbone for Lingo. Multi-phased evaluation of Lingo including system benchmark, usability and real-world deployment highlighted its efficacy and limitation on information quality, access mechanism and interaction modality. We hope the technical vision, our prototype implementation and multi-faceted findings from the studies present a solid foundation for future research in this direction, in particular those addressing hyper-local information service to urban citizens.

ACKNOWLEDGMENTS

This work is supported by the ESSENCE project. ESSENCE is an icon project realised in collaboration with imec and with project support from VLAIO (Flanders Innovation & Entrepreneurship).

REFERENCES

[1] 2020. Cisco Annual Internet Report (2018–2023). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf. [Accessed: 2021-05-15].

[2] 2021. IEEE Standard for Information Technology–Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks–Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)* (2021), 1–4379. https://doi.org/10.1109/IEEESTD.2021.9363693

[3] 2021. YAMNet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. [Accessed: 2021-05-15].

[4] Utku Günay Acer, Marc van den Broeck, Claudio Forlivesi, Florian Heller, and Fahim Kawsar. 2019. Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 35 (June 2019), 32 pages. https://doi.org/10.1145/3328906

[5] Utku Günay Acer and Otto Waltari. 2017. WiPush: Opportunistic Notifications over WiFi without Association. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (Melbourne, VIC, Australia) *(MobiQuitous 2017)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3144457.3144492

[6] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. 2010. Location-Based Crowdsourcing: Extending Crowdsourcing to the Real World. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (Reykjavik, Iceland) *(NordiCHI '10)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/1868914.1868921

[7] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (apr 2019), 28 pages. https://doi.org/10.1145/3311956

[8] Ganesh Ananthanarayanan, Paramvir Bahl, Peter Bodík, Krishna Chintalapudi, Matthai Philipose, Lenin Ravindranath, and Sudipta Sinha. 2017. Real-time video analytics: The killer app for edge computing. *computer* 50, 10 (2017), 58–67. https://doi.org/10.1109/MC.2017.3641638

[9] Elian Aubry, Thomas Silverston, Abdelkader Lahmadi, and Olivier Festor. 2014. CrowdOut: A mobile crowdsourcing service for road safety in digital cities. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. 86–91. https://doi.org/10.1109/PerComW.2014.6815170

[10] Matthias Baldauf, Stefan Ribler, and Peter Fröhlich. 2019. Alexa, I'm in Need! Investigating the Potential and Barriers of Voice Assistance Services for Social Work. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI 2019)*. Association for Computing Machinery, New York, NY, USA, Article 50, 6 pages. https://doi.org/10.1145/3338286.3344397

[11] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (sep 2018), 24 pages. https://doi.org/10.1145/3264901

[12] Lokesh Boominathan, Srinivas S S Kruthiventi, and R. Venkatesh Babu. 2016. CrowdNet: A Deep Convolutional Network for Dense Crowd Counting. arXiv:1608.06197 [cs.CV]

[13] John Brooke. 1996. *"SUS-A quick and dirty usability scale." Usability evaluation in industry.* CRC Press. https://www.crcpress.com/product/isbn/9780748404605 ISBN: 9780748404605.

[14] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J Franklin, Joseph E Gonzalez, and Ion Stoica. 2017. Clipper: A low-latency online prediction serving system. In *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*. 613–627.

[15] Subhankar Dhar and Upkar Varshney. 2011. Challenges and Business Models for Mobile Location-Based Services and Advertising. *Commun. ACM* 54, 5 (May 2011), 121–128. https://doi.org/10.1145/1941487.1941515

[16] Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human–Computer Interaction* 31, 4 (2015), 307–335. https://doi.org/10.1080/10447318.2014.986642

[17] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. 2008. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring. In *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services* (Breckenridge, CO, USA) *(MobiSys '08)*. ACM, New York, NY, USA, 29–39. https://doi.org/10.1145/1378600.1378605

[18] Saul Greenberg and Nicolai Marquardt. 2016. Using Social Science Theory to Inspire Surface Design: A Case Study of Proxemic Interactions. In *Designing Digital Surface Applications*, Frank Maurer (Ed.). SurfNet, University of Calgary, Calgary, Canada, 26–38.

[19] Vishal Gupta and Mukesh Rohil. 2012. Enhancing Wi-Fi with IEEE 802.11u for Mobile Data Offloading. *International Journal of Mobile Network Communications & Telematics* 2 (08 2012). https://doi.org/10.5121/ijmnct.2012.2403

[20] Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2012. Mapping Community Engagement with Urban Crowd-Sourcing. In *Proc. When the City Meets the Citizen Workshop (WCMCW)*. AAAI, Palo Alto, CA, USA, 14–19. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4749/5102

[21] Inseok Hwang, Youngki Lee, Chungkuk Yoo, Chulhong Min, Dongsun Yim, and John Kim. 2019. Towards Interpersonal Assistants: Next-Generation Conversational Agents. *IEEE Pervasive Comput.* 18, 2 (2019), 21–31. https://doi.org/10.1109/MPRV.2019.2922907

[22] Samvit Jain, Xun Zhang, Yuhao Zhou, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, Paramvir Bahl, and Joseph Gonzalez. 2020. Spatula: Efficient cross-camera video analytics on large camera networks. In *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 110–124. https://doi.org/10.1109/SEC50012.2020.00016

[23] Michael J Kuhn. 2015. Virtual game assistant based on artificial intelligence. US Patent 9,202,171.

[24] Axel Küpper. 2005. *Location-based services: fundamentals and operation.* John Wiley & Sons. https://doi.org/10.1002/0470092335.ch2

[25] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for

Computing Machinery, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[26] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (March 2019), 18 pages. https://doi.org/10.1145/3314404

[27] P. Luff, D. Frohlich, and N.G. Gilbert. 2014. *Computers and Conversation.* Elsevier Science. https://books.google.be/books?id=UmniBQAAQBAJ

[28] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[29] Nicolas Maisonneuve, Matthias Stevens, Maria Niessen, and Luc Steels. 2009. NoiseTube: Measuring and mapping noise pollution with mobile phones. 215–228. https://doi.org/10.1007/978-3-540-88351-7_16

[30] Nicolai Marquardt and Saul Greenberg. 2015. Proxemic Interactions: From Theory to Practice. *Synthesis Lectures on Human-Centered Informatics* 8 (02 2015), 1–199. https://doi.org/10.2200/S00619ED1V01Y201502HCI025

[31] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. 2008. Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems* (Raleigh, NC, USA) *(SenSys '08)*. Association for Computing Machinery, New York, NY, USA, 323–336. https://doi.org/10.1145/1460412.1460444

[32] Lindsay C. Page and Hunter Gehlbach. 2017. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. *AERA Open* 3, 4 (2017). https://doi.org/10.1177/2332858417749220 arXiv:https://doi.org/10.1177/2332858417749220

[33] Salvatore Parise, Patricia J Guinan, and Ron Kafka. 2016. Solving the crisis of immediacy: How digital technology can transform the customer experience. *Business Horizons* 59, 4 (2016), 411–420. https://doi.org/10.1016/j.bushor.2016.03.004

[34] Chunjong Park, Chulhong Min, Sourav Bhattacharya, and Fahim Kawsar. 2020. Augmenting Conversational Agents with Ambient Acoustic Contexts. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 33, 9 pages. https://doi.org/10.1145/3379503.3403535

[35] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. *StreetWise: Smart Speakers vs Human Help in Public Slum Settings.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300326

[36] Nishant Piyush, Tanupriya Choudhury, and Praveen Kumar. 2016. Conversational commerce a new era of e-business. In *2016 International Conference System Modeling Advancement in Research Trends (SMART)*. 322–327. https://doi.org/10.1109/SYSMART.2016.7894543

[37] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. *Voice Interfaces in Everyday Life.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[38] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. 2015. The Digital Life of Walkable Streets. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 875–884. https://doi.org/10.1145/2736277.2741631

[39] Daniele Quercia, Diarmuid Ò Séaghdha, and Jon Crowcroft. 2012. Talk of the city: Our tweets, our community happiness. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6.

[40] Rajib Kumar Rana, Chun Tung Chou, Salil S. Kanhere, Nirupama Bulusu, and Wen Hu. 2010. Ear-Phone: An End-to-End Participatory Urban Noise Mapping System. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks* (Stockholm, Sweden) *(IPSN '10)*. Association for Computing Machinery, New York, NY, USA, 105–116. https://doi.org/10.1145/1791212.1791226

[41] Jonathan Raper, Georg Gartner, Hassan Karimi, and Chris Rizos. 2007. Applications of Location-Based Services: A Selected Review. *J. Locat. Based Serv.* 1, 2 (June 2007), 89–111. https://doi.org/10.1080/17489720701862184

[42] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs.CV]

[43] Darshan Santani, Jidraph Njuguna, Tierra Bills, Aisha W. Bryant, Reginald Bryant, Jonathan Ledgard, and Daniel Gatica-Perez. 2015. CommuniSense: Crowdsourcing Road Hazards in Nairobi. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 445–456. https://doi.org/10.1145/2785830.2785837

[44] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. *arXiv:1503.02364 [cs]* (March 2015). http://arxiv.org/abs/1503.02364 arXiv: 1503.02364.

[45] Haichen Shen, Lequn Chen, Yuchen Jin, Liangyu Zhao, Bingyu Kong, Matthai Philipose, Arvind Krishnamurthy, and Ravi Sundaram. 2019. Nexus: A GPU Cluster Engine for Accelerating DNN-Based Video Analysis. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles* (Huntsville, Ontario, Canada) *(SOSP '19)*. Association for Computing Machinery, New York, NY, USA, 322–337. https://doi.org/10.1145/3341301.3359658

[46] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26. https://doi.org/10.1631/FITEE.1700826

[47] Jaisie Sin and Cosmin Munteanu. 2019. An Information Behaviour-Based Approach to Virtual Doctor Design. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI 2019)*. Association for Computing Machinery, New York, NY, USA, Article 44, 6 pages. https://doi.org/10.1145/3338286.3344391

[48] Myrthe L. Tielman, Mark A. Neerincx, Rafael Bidarra, Ben Kybartas, and Willem-Paul Brinkman. 2017. A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories. *Journal of Medical Systems* 41, 8 (Aug. 2017), 125. https://doi.org/10.1007/s10916-017-0771-y

[49] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 254–264. https://doi.org/10.1145/2675133.2675233

[50] Alexandra Voit, Jasmin Niess, Caroline Eckerth, Maike Ernst, Henrike Weingärtner, and Paweł W. Woźniak. 2020. 'It's Not a Romantic Relationship': Stories of Adoption and Abandonment of Smart Speakers at Home. In *19th International Conference on Mobile and Ubiquitous Multimedia* (Essen, Germany) *(MUM 2020)*. Association for Computing Machinery, New York, NY, USA, 71–82. https://doi.org/10.1145/3428361.3428469

[51] Li Wang and Dennis Sng. 2015. Deep learning algorithms with applications to video analytics for a smart city: A survey. *arXiv preprint arXiv:1512.03131* (2015).

[52] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manage. Inf. Syst.* 12, 4 (March 1996), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[53] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 542, 12 pages. https://doi.org/10.1145/3290605.3300772

[54] Vinicius Zambaldi, Joao Pesce, Daniele Quercia, and Virgilio Almeida. 2014. Lightweight contextual ranking of city pictures: urban sociology to the rescue. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[55] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93. https://doi.org/10.1162/coli_a_00368