Sim911: Towards Effective and Equitable 9-1-1 Dispatcher Training with an LLM-Enabled Simulation

Zirong Chen[1], Elizabeth Chason[1], Noah Mladenovski[2], Erin Wilson[2], Kristin Mullen[2], Stephen Martini[2], Meiyi Ma[1]

[1]Department of Computer Science, Vanderbilt University, Nashville, Tennessee 37235, USA
[2]Metro Department of Emergency Communications, Nashville, Tennessee 37211, USA
{zirong.chen, elizabeth.r.chason, meiyi.ma}@vanderbilt.edu
{noah.mladenovski, erin.wilson, kristin.mullen, stephen.martini}@nashville.gov

Abstract

Emergency response services are vital for enhancing public safety by safeguarding the environment, property, and human lives. As frontline members of these services, 9-1-1 dispatchers have a direct impact on response times and the overall effectiveness of emergency operations. However, traditional dispatcher training methods, which rely on role-playing by experienced personnel, are labor-intensive, time-consuming, and often neglect the specific needs of underserved communities. To address these challenges, we introduce Sim911[1], the first training simulation for 9-1-1 dispatchers powered by Large Language Models (LLMs). Sim911 enhances training through three key technical innovations: (1) knowledge construction, which utilizes archived 9-1-1 call data to generate simulations that closely mirror real-world scenarios; (2) context-aware controlled generation, which employs dynamic prompts and vector bases to ensure that LLM behavior aligns with training objectives; and (3) validation with looped correction, which filters out low-quality responses and refines the system performance. Beyond its technical advancements, Sim911 delivers significant social impacts. Successfully deployed in the Metro Nashville Department of Emergency Communications (MNDEC), Sim911 has been integrated into multiple training sessions, saving time for dispatchers. By supporting a diverse range of incident types and caller tags, Sim911 provides more realistic and inclusive training experiences. In our conducted user study, 90.00 percent of participants found Sim911 to be as effective or even superior to traditional human-led training, making it a valuable tool for emergency communications centers nationwide, particularly those facing staffing challenges.

1 Introduction

Emergency response services are essential for public safety, managing approximately 240 million 911 calls annually in NYC, according to city-wide statistics all year (NYC-911 2022). However, there is a critical staffing shortage, with a third of centers reporting more vacancies in 2023 compared to 2019, resulting in approximately 25,000 unfilled positions nationwide. This staffing crisis increases the workload on current staff (Chen et al. 2022, 2023), leading to dispatcher burnout and impacting emergency service quality (NICE 2023). As urban areas in the US grow, the strain on emergency response systems intensifies. Rapid urbanization and population growth demand effective solutions to adapt and manage these increasing pressures (Ma et al. 2019).

Traditional training environments prepare trainees for real-world dispatcher roles by employing role-playing scenarios where experienced dispatchers coach trainees through simulated calls. The need for experienced dispatchers to participate in training diverts essential personnel from actual emergency duties, causing inconsistencies in training quality and reducing the availability of skilled staff, particularly in underserved areas (Saxon et al. 2022; Afonso 2021). However, traditional training methods, which rely on role-playing by experienced personnel, are labor intensive, time consuming, and often overlook the specific needs of underserved communities.

In light of these demands, exploring innovative technological solutions is critical. Advancements in artificial intelligence (Ma, Stankovic, and Feng 2021), especially Large Language Models (LLMs), offer promising methods for enhancing training environments. Employing LLMs to simulate caller interactions can reduce reliance on human resources, improving training efficiency and consistency (Naveed et al. 2023; Wang et al. 2023; Carta et al. 2023). However, directly applying plain LLM agents is not ideal. In our preliminary investigations, we identified the following **challenges**: (1) Achieving consistent *realistic* simulations is difficult without detailed factual databases, despite meticulous prompt engineering across different LLMs. This lack of realism results in simulations that do not fit the local context, making the training less effective and potentially confusing for trainees. (2) LLMs excel at generating coherent content, but tend to fabricate details, undermining *authenticity*. Simulations with fabricated geographic information lead dispatchers to make decisions based on incorrect data, compromising the effectiveness of the emergency response. (3) The needs of *vulnerable populations* in metropolitan areas are often understudied during conventional training, leaving practitioners unprepared. This lack of inclusiveness results in biased training, inadequate preparation of dispatchers to handle calls from vulnerable groups, and causes disparities in emergency response. (4) The inherently complex nature of 9-1-1 calls presents significant challenges, even for human trainers. Human-led training may also fail to capture

[1]Code and Demo: https://meiyima.github.io/angie.html

these complexities, as discussed in Section 2.

In this paper, we introduce Sim911, the first system that leverages LLMs to simulate realistic 9-1-1 calls, specifically designed to enhance dispatcher training. Sim911 focuses on creating effective and equitable simulation experiences tailored to the local metro area. Sim911 comprises three key components: *knowledge construction*, *context-aware controlled generation*, and *validation with looped feedback*. Knowledge construction organizes real-world information into retrieval knowledge bases, while context-aware controlled generation fine-tunes the LLM's behavior through human-designed instructions. Validation with looped feedback ensures high-quality outputs by filtering out low-quality responses.

We summarize our **technical innovations** and **contributions** as follows: (1) *Innovative Knowledge Construction from 9-1-1 Calls*: Sim911 organizes real-world call data into detailed knowledge bases, allowing for the generation of contextually accurate and realistic training simulations, supporting 57 different incident types. (2) *Context-Aware Controlled Generation*: Sim911 strategically and dynamically uses advanced techniques, such as Chain-of-Thought (CoT) and Retrieval-Augmented Generation (RAG), to tailor LLM behavior during training sessions. (3) *Validation with Looped Correction*: Sim911 includes a unique validation process that filters out low-quality responses, ensuring high-quality and scenario-appropriate outputs. (4) *Focus on Social Equity*: Sim911 emphasizes training that addresses the needs of underserved and vulnerable communities, incorporating relevant data to better prepare dispatchers for diverse real-world scenarios. (5) *Real-World Deployment and Evaluation*: Successfully deployed in DEC, Sim911 has proven an effective tool in enhancing 9-1-1 dispatcher training from experimental results on real-world data.

Beyond technical advancements, Sim911 delivers significant **social impacts**: (1) Sim911 has been successfully deployed in DEC's training programs, seamlessly integrated into 4 training classes across different service sites. (2) To the date of this paper, Sim911's system logs reveal a total active simulation time of 26.55 hours, effectively saving this time for MNDEC dispatchers. (3) Sim911 supports 57 different real-world incident types and covers 14 caller tags, such as "unhoused" and "non-English speaking," to enrich caller profiles. (4) In a user study conducted with DEC, 90.00% of participants found Sim911 to be as effective or even superior to traditional human-led training. Additionally, Sim911 received an average helpfulness score of 4.89 for its assistance in call-taking training. (5) Sim911 has the potential to assist emergency communications centers throughout the US with limited staffing by allowing trainees to engage individually with the training program.

2 Motivating Study

We analyzed 11,841 real-world recordings (from Nov. 2022 to May 2024) and reviewed 33 randomly selected conventional training pieces, leading to the following observations.

Traditional training is laborious and time-consuming. In traditional training setups, each trainee engages in call simulations, assuming three roles: the call-taker, the caller, and the instructor. The trainee, as the call-taker, manages calls directed by the instructor and an experienced dispatcher. Each trainee typically participates in 60 independent simulated calls, and each call requires the participation of three participants. Based on past recordings, the average call duration is 3.5 minutes; with an average of 12 trainees per session, the total time commitment for experienced dispatchers amounts to at least 84 working hours per session.

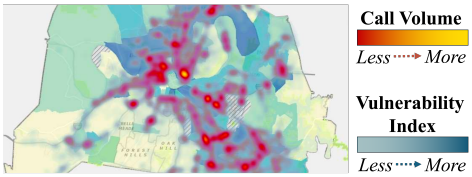

Figure 1: Year-round Distribution of 9-1-1 Calls and Vulnerability Index in Nashville, 2023.

Real-world 9-1-1 calls cover a wide spectrum of incident types and contextual scenarios. From our analysis of past phone call recordings, we identified over 200 distinct incident specifications. However, during initial training, each trainee is exposed to only 40 incident types and 15 call templates. Our review reveals that, on average, trainees cover only 48.00% of the incident types in the first 3-day program, and only 61.54% of special contexts or requests are adequately addressed. This limited exposure fails to prepare trainees for the variety of incidents they will encounter.

Caller images are critical for call-taking training but rarely considered. Equity and inclusiveness are often overlooked in conventional role-playing simulations during dispatcher training. Even with guidance from experienced dispatchers, these simulations frequently struggle to empathetically and accurately capture the nuanced experiences of vulnerable groups. Among the 33 training scenarios we reviewed, only 4 focused on vulnerable populations (such as non-native English speakers, who may use different language patterns; and callers from lower-income housing areas, who might have limited access to personal vehicles) representing just 12.12%. However, government statistics, see Figure 1, indicate that the needs of various vulnerable groups are significantly reflected in real-world 9-1-1 calls. This discrepancy highlights the importance of incorporating diverse caller images into training, as different scenarios might require distinct call-taking skill sets to effectively handle real-life situations.

3 Methodology

This section first provides an overview of Sim911. Then we introduce the technical aspects of how Sim911 works in Sections Knowledge Construction, Context-aware Controlled Generation, and Validation with Looped Correction.

Sim911 simulates calls by playing the role of 9-1-1 callers and interacting directly with the trainees. It comprises three main components, depicted in Figure 2: *knowledge construction*, *context-aware controlled generation*, and *valida-*

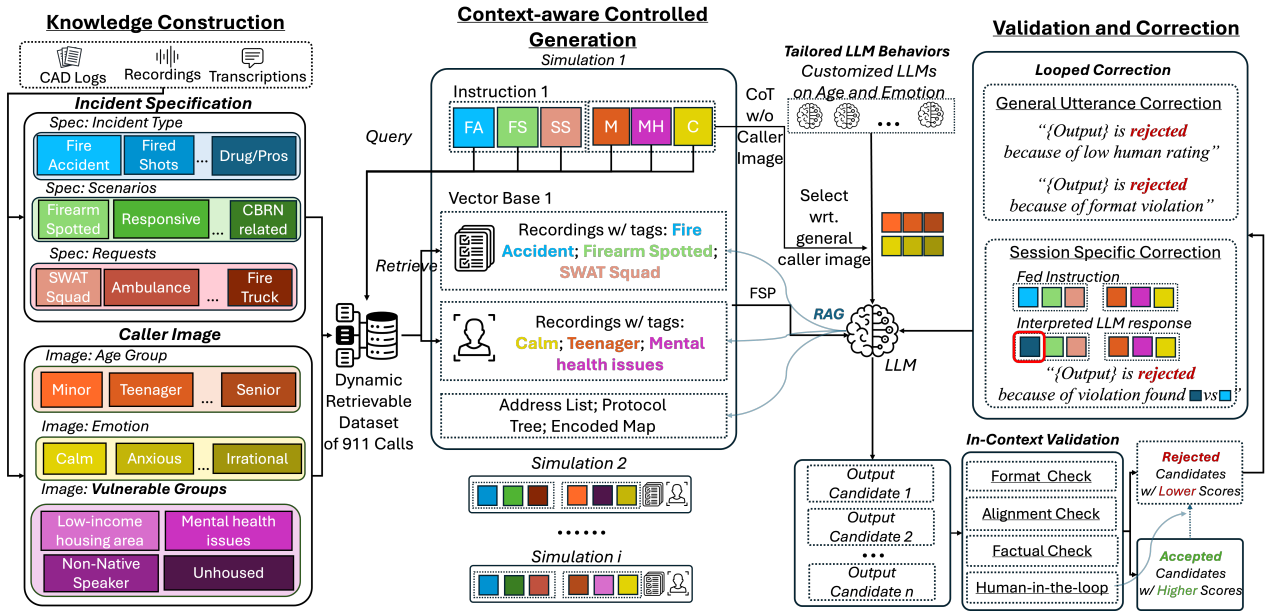

Figure 2: An Overview of Sim911's Key Components: *knowledge construction* (KC), *context-aware controlled generation* (CaCG), and *validation with looped correction* (VLC). KC integrates real-world data into knowledge bases before runtime. CaCG tailors LLM behaviors during runtime. VLC involves in-context validation during runtime and periodic correction after runtime.

tion with looped correction. During the *Knowledge Construction* phase, 11,841 calls are analyzed to develop knowledge bases containing tags for incident specifics and caller images. This ensures that pertinent information is readily available during simulations. At runtime, Sim911 utilizes these tags to select the most suitable LLM backends, query the knowledge bases, and generate prompts in the *context-aware controlled generation* phase. This process tailors prompts to include incident details and caller profiles, ensuring that LLM responses align with simulation requirements. The *Validation with Looped Correction* component filters out low-quality responses in as the simulation goes.

3.1 Dynamic Knowledge Construction

As a first step, we build an in-depth review and sophisticated reconstruction of the existing dataset, which has three key data sources: Computer-Aided Dispatching(CAD) logs, archived 9-1-1 call recordings, and their corresponding transcriptions.

Detailing Two Components in 9-1-1 Calls We integrate insights from dispatcher teams at MNDEC to identify two key components in 9-1-1 call handling: *Incident Specifications* and *Caller Images*. We use finely-grained tags for each data entry to create more accurate simulations. In our annotation work, we manually review each call and apply all relevant tags. A single call may be annotated with multiple tags to ensure comprehensive coverage.

Incident Specifications (*IS*) capture critical details of incidents, including: (1) *Incident Type*: Categorizes the incident, from routine (e.g., illegal parking) to critical (e.g., severe medical emergencies). (2) *Scenario Context*: Adds situational context, such as environmental conditions (e.g., severe weather), potential threats (e.g., sightings of firearms), or specific events (e.g., large public gatherings). (3) *Special Requests*: Identifies specific instructions, like the need for specialized units (e.g., bomb squads) or coordination with other agencies (e.g., fire departments).

Caller Images (*CI*) create a comprehensive caller profile, enhancing the LLM's understanding of the caller's perspective, especially for vulnerable groups. This includes: (1) *General Tags*: Profiles the caller by age (e.g., minor, adult) and emotion (e.g., "neutral", "anxious"). Dispatchers assign these tags based on conversation clues (e.g., "My mom is mid-70s and living alone") or voice analysis. These tags are less sensitive and linked to pre-customized LLM agents to avoid identification confusion (Wei, Haghtalab, and Steinhardt 2024). (2) *Vulnerable Groups*: These government-introduced tags include descriptors such as "low-income housing area" (if the call originates from a lower-income area, according to year-round statistics), "mental health" (if the caller exhibits potential mental health issues, such as bipolar disorder or depression; inferred from the conversation), "non-native speaker" (if the caller uses limited English), and "unhoused" (if the caller indicates lack of stable, permanent housing, inferred from conversation clues). These tags are considered highly sensitive and remain hidden during runtime due to ethical concerns.

Specializing Knowledge Bases for Contextual Control We leverage the Retrieval Augmented Generation (RAG) approach, which enhances LLMs for tasks requiring deep

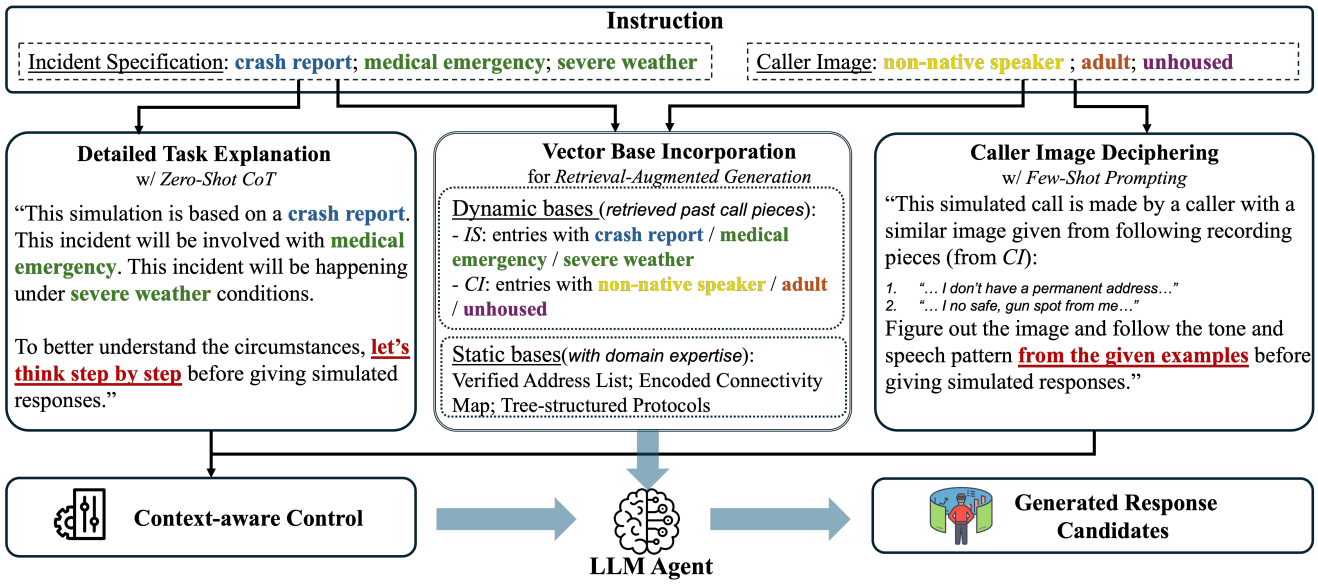

Figure 3: An Example of the 3-step *Context-aware Controlled Generation* with Vector Base Incorporation, Detailed Task Explanation, and Caller Image Deciphering. The incident type specification (*IS*) comes with tags **crash report** (incident type), **medical emergency**(special request), and **severe weather** (scenario contexts), and the caller image (*CI*) is set to be an **unhoused non-native speaker adult**.

knowledge by incorporating external databases as reference points during content generation. This methodology, as discussed by (Lewis et al. 2020), improves the LLMs' ability to provide accurate and relevant outputs. Here, we introduce the two major knowledge bases for runtime use.

Factual Bases. First, we build a static base, which contains factual knowledge: (1) *Validated Address List*: a comprehensive list of real addresses within the local area; (2) *Encoded Map with Connectivity Information*: beyond simple address listings, this map provides detailed information about the connectivity between locations; (3) *Tree-Structured Protocols*: a collection of protocols for various types of emergency incidents, organized in a tree structure. These protocols detail the question sequence dispatchers should follow, ensuring Sim911's simulations adhere to procedural standards of emergency response.

Retrievable Bases. This base includes data entries tagged according to Incident Specifications (*IS*) and Caller Images (*CI*). The retrievable base allows Sim911 to query and retrieve necessary data samples that enhance the simulation experience during runtime.

3.2 Context-aware Controlled Generation

Each simulation runtime begins with predefined tags (referred to as 'instructions') that detail the desired scenarios. These tags delineate Incident Specifications (*IS*) and Caller Images (*CI*), guiding the setup for each simulation. We select the most appropriate preset backend for each simulation based on less sensitive *CI* attributes (emotion, age) and gather data associated with these tags from knowledge bases. This ensures a well-informed simulation environment tailored to the given simulation instructions. Context-aware Controlled Generation employs advanced prompting techniques, including Chain-of-Thought (CoT), Retrieval-Augmented Generation (RAG), and Few-shot Prompting (FSP), to enhance LLM performance (Touvron et al. 2023; Kaplan et al. 2020; Wei et al. 2022). Unlike the direct and static application of these techniques, our approach dynamically adapts to the emergency response context, with a focus on both *IS* and *CI*. The context-aware controlled generation process consists of three major steps, illustrated in the running example in Figure 3: (1) *Vector Base Incorporation*: We statically mount the fact bases and dynamically retrieve all past call pieces associated with given tags in both *IS* (e.g., crash report, medical emergency, severe weather) and *CI* (e.g., non-English speaking, adult, unhoused). The LLM backend is granted access to both vector bases; (2) *Detailed Task Explanation*: We elaborate on the instruction through step-by-step explanations, setting the stage for how the simulation should proceed. This preparation allows the LLM to conceptualize the simulation's context and objectives before initiation; (3) *Caller Image Deciphering*: By including examples of past utterances and interactions associated with both similar tags, we provide the LLM with contextually relevant examples to draw from. This repository of past interactions enriches the LLM's understanding and ability to generate responses that are both consistent with the user's profile and grounded in real-world examples.

3.3 Validating with Looped Correction

During runtime, Sim911 employs an in-context validation process with a co-pilot design (Chen et al. 2024) to prevent negative examples and iteratively loops back to the LLM backend until a validated response is obtained. We adjust

this threshold in practice to balance latency and accuracy. The *In-Context Validation* process includes four key checks: (1) *Format check*: This check ensures that the generated response adheres to the expected format. Any response that violates the format requirements is discarded to prevent system errors. (2) *Alignment check*: This step utilizes a BERT-based classifier (Devlin et al. 2019) to extract incident specifications from the response. The extracted specifications are then compared with the expected instructions, and any misalignment results in the response being discarded. (3) *Factual check*: A RoBERTa-based question-answering framework (Liu et al. 2019) is used to query key details, such as location information, by asking preset questions (e.g., "What is the address?"). If the extracted address does not exist in real life, the response is discarded. (4) *Human-in-the-Loop check*: This step allows users to provide immediate feedback on the generated response, supporting both written comments and scaled (1-5) ratings. Users can reject any response that does not meet their standards, and this feedback is systematically collected for further analysis.

4 Evaluation of Sim911

Sim911 introduces a pioneering AI-driven system to enhance call-taker training for emergency response scenarios. Due to its novelty, there is limited existing literature to guide its evaluation. To provide a comprehensive assessment, we not only report system-level performance but also conduct a study on the component-wise impacts using pre-configured runtimes. This approach allows us to evaluate Sim911 component by component without disrupting its ongoing deployment at DEC. Therefore, our evaluation of Sim911's performance, focusing on ***effectiveness*** and ***equity***, consists of two components: (1) component-wise analysis using pre-configured runtimes, and (2) system-level assessment during real-world deployment. The framework of Sim911 is generalizable with any LLM backends. GPT-4o is tested to be the optimal LLM backend for Sim911 by the date of this submission. We fetched GPT-4o responses using OpenAI API and tested the workflow on a machine with a 2.50GHz CPU, 32GB RAM, and Nvidia RTX 3080Ti GPU.

Component-wise analysis with pre-configured runtimes: We begin by extracting instructions from 2,641 past calls in the MNDEC database, spanning 13 incident types (e.g., Motor Vehicle Accidents 11%, Lost and Stolen 10%, Aggressive Drivers 10%) from Nov. 2022 to May 2024, based on Incident Specifications (*IS*) and Caller Images (*CI*) provided by expert annotations. We then replicate the dialogue flows using rule-based scripts that mimic the questions typically asked by call-takers. These instructions and replicated dialogue flows are used to simulate calls with Sim911. Sim911 operates without access to highly granular details. For example, if a past call involved an abandoned vehicle with a specific license plate and tinted windows, Sim911 would only be informed that the simulation involves an abandoned vehicle, without further specifics like the license plate or tinted windows. To ensure a fair evaluation, we exclude these granular discrepancies from our analysis. Effectiveness and equity scores are obtained through call-wise comparisons under control experiments. We record average scores with standard deviations to ensure robust evaluation.

System-level assessment during real-world deployment: During Sim911's deployment, we collected data from 3,416 system interactions and 3,409 user interactions across both complete (228) and incomplete simulations, each guided by its own set of instructions (*IS* and *CI*). These data are utilized in assessing Sim911's authenticity. Additionally, we conducted a user study in collaboration with MNDEC to evaluate Sim911 at a system level. This study involved trainees and personnel from DEC, including those from training management and quality assurance. The user study gathered scaled feedback (1-5) from MNDEC personnel on several key aspects, e.g., *realism* ('How similar or vivid are the calls generated by Sim911 compared to real-world calls?'), *authenticity* ('Are Sim911's responses valid and true to real-life situations?'), *equity* ('How well does Sim911 simulate the experiences of vulnerable callers?'), and *helpfulness* ('How helpful is Sim911 in assisting with call-taking training?'). Written comments were also collected to provide additional insights. We review system logs and user feedback to assess effectiveness and equity, with further details discussed later.

4.1 Effectiveness of Sim911

We assess Sim911's effectiveness by evaluating its realism and authenticity.

Realism: *"How closely do Sim911's simulations mirror real-world calls?"* We use the following metrics to evaluate Sim911's performance on pre-configured runtimes: *Perplexity* (a measure of distributional similarity commonly used in language model training; it assesses how reasonable the generated texts are compared to a reference set), *METEOR* (Banerjee and Lavie 2005) (text generation metric that balances precision and recall, considering word stems, synonyms, and word order to determine how closely a generated text mirrors a reference text), and *TTR* (Type-Token Ratio; measures lexical diversity by comparing the number of unique words to the total number of words in the text).

Authenticity: *"Does Sim911 provide accurate, true-to-life information without fabricating given instructions?"* For evaluation, we break authenticity down into two aspects: "matter of facts" and "simulation alignment." For the first, we focus on the accuracy of the given location in a simulation, as recommended by MNDEC experts. We use the Google Maps API with Geocoding (Google Maps Platform 2024) to verify the geographic information provided in the simulation and report the *locating success rate*. To measure *simulation alignment*, we use the copilot's results to determine if the indicated Incident Specification (*IS*) aligns with the one provided in the simulation instructions. System-level performance during real-world deployment is assessed through quantitative analysis of system logs.

From the statistics in Table 1, we observe the following key points. When all components are enabled, Sim911 achieves optimal results in both realism (PPL=11.07, METEOR=0.85) and authenticity (GMap=99.19%, SAR=98.42%). Disabling knowledge construction (KC) and the RAG sub-component of CaCG leads to significant drops in realism (PPL=31.22 and PPL=57.19, respectively). Similarly, turning off validation with looped

		REALISM			AUTHENTICITY	
		PPL↓	METEOR↑	TTR↑	GMap(%)↑	SAR(%)↑
Sim911¬KC		31.22±13.38	0.22±0.05	0.88±0.01	83.00±5.11	94.11±1.66
Sim911¬CaCG	¬CoT	21.98±6.64	0.67±0.18	0.85±0.02	98.13±1.06	90.83±2.71
	¬FSP	12.55±5.71	0.75±0.21	0.94±0.01	98.11±1.08	98.14±1.06
	¬RAG	57.19±12.22	0.19±0.12	0.92±0.02	61.47±6.89	96.44±2.60
Sim911¬VLC		18.89±8.19	0.77±0.08	0.88±0.02	89.11±5.15	89.48±3.12
Sim911¬All(*GPT-4o*)		61.99±13.91	0.12±0.02	0.88±0.02	61.01±11.31	81.63±3.97
Sim911		**11.07±5.49**	**0.85±0.03**	**0.94±0.01**	**99.19±0.81**	**98.42±1.58**

Table 1: ***Effectiveness*** of Sim911 in terms of **REALISM** and **AUTHENTICITY**. The metrics include ***PPL*** (Perplexity), ***TTR*** (Type-Token Ratio), ***GMap*** (Google Maps API locating success rate), and ***SAR*** (Simulation Alignment Rate from copilots).

correction (VLC) reduces both realism and authenticity, though the system remains moderately effective. When all components are disabled, the system's performance declines significantly, particularly in realism (PPL=61.99) and authenticity (SAR=81.63%). In conclusion, *Sim911 demonstrates high effectiveness in terms of realism and authenticity in real-world deployment when all components are active. Disabling components harms Sim911's overall effectiveness.*

4.2 Equity of Sim911

We assess Sim911's equity features by evaluating *"how effectively it provides simulation experiences for different caller groups"*, represented by each supported tag in the caller image (*CI*). Recognizing that some tags are subjective and challenging to quantify, we adopt two general approaches to study these equity features. We employ fine-tuned BART (Lewis et al. 2019), a state-of-the-art model for zero-shot text classification, to evaluate Sim911-generated emergency call texts against a predefined set of image tags. For each generated call x_i, associated with ground truth tags $\mathcal{T}(x_i)$, BART predicts the presence or absence of each tag T_j using a binary classifier $C_j(x_i)$, which outputs 1 if x_i is associated with T_j, and 0 otherwise. The predicted tags form a binary vector $\hat{\mathcal{T}}(x_i) = \{C_1(x_i), C_2(x_i), \ldots, C_k(x_i)\}$. Accuracy for each call is calculated by comparing $\hat{\mathcal{T}}(x_i)$ with $\mathcal{T}(x_i)$ using the formula $\text{Acc}(x_i) = \frac{1}{k}\sum_{j=1}^{k} \mathbb{I}\left(C_j(x_i) = \mathbb{I}(t_{ij} \in \mathcal{T}(x_i))\right)$. This classification is iteratively performed for each tag, and the overall accuracy is determined by averaging the individual accuracies across all generated calls as *BART Score* $= \frac{1}{n}\sum_{i=1}^{n} \text{Acc}(x_i)$. Second, we perform a textual similarity analysis based on syntax (Context-Free Grammar Parser), lexicon (TF-iDF), and sentiment (Loria 2018). We compare the generated outputs tagged as A with both the ground truth tagged as A and not-A. To quantify the strength of classification for each tag, we calculate the *Margin Score* $(\text{Similarity}(A) - \text{Similarity}(\neg A))/(\text{Similarity}(A) + \text{Similarity}(\neg A))$, where Similarity(.) is the overall syntactic similarity of the output to reference texts with a given tag.

Besides these two approaches, we additionally introduce the following tag-specific evaluation methods: (1) *NRCLex* (Mohammad and Turney 2013) for unsupervised textual ***emotion*** detection, where we analyze the accuracy similarly to the *BART Score*. (2) *Gunning Fog Index*, a well-known method in linguistics of text readability analysis, is used to assess the readability of the text for *non-native English speakers*. Gunning Fog Index outputs a readability level and we analyze this score similarly to the *Margin Score*.

From the statistics in Table 2, we derive the following findings. Sim911 achieves strong performance across all caller image tags when all components are enabled, including age groups (BART=83.11%, Margin=0.34), emotion ranges (BART=85.66%, Margin=0.36), and unhoused populations (BART=73.94%, Margin=0.21). Disabling the FSP sub-component of CaCG results in notable declines for age groups (BART=62.90%, Margin=0.09) and mental health tags (BART=64.55%, Margin=0.11). Turning off knowledge construction (KC) significantly reduces performance, especially for low-income housing (BART=51.13%, Margin=0.04) and mental health (BART=67.67%, Margin=0.17). Similarly, disabling validation with looped correction (VLC) leads to a decrease in metrics for mental health (BART=83.22%, Margin=0.21). When all components are disabled, the system's performance deteriorates significantly, particularly for the low-income housing tags (BART=44.44%, Margin=0.04) and non-native speakers (BART=69.13%, Margin=0.21). *In conclusion, Sim911 delivers equitable and inclusive simulations in real-world deployment when all components are enabled. Disabling components negatively impacts Sim911's equity features.*

4.3 Insights from User Study

We collected 10 anonymous feedback from trainees (x2), active call-takers/dispatchers (x2), and training officers (x6) at DEC. Surveys are contributed by MNDEC based on the availability. Responses included yes/no questions, written comments, and a scaled rating system: Not at all (1), Neutral (2), Somewhat (3), Very much (4), and Perfectly (5). We find following insights.

Effectiveness and Equity: Sim911 received scores of 4.50 for realism and 4.70 for authenticity. In terms of equity, it performed well across various caller image tags, with average scores as follows: Age Groups (4.25), Emotion Ranges (4.20), Unhoused (4.10), Mental Health (4.25), Non-Native Speakers (4.25), and Low-Income Housing (4.10). Additionally, Sim911 earned an average score of 4.89 for *"How effectively does Sim911 support call-taker training in real-life scenarios?"*. One participant commented: *"I was surprised by how well it handled a call as a pregnant woman. I even*

		AGE GROUPS		EMOTION RANGES			UNHOUSED	
		BART(%)↑	Margin↑	BART(%)↑	Margin↑	NRCLex↑	BART(%)↑	Margin↑
Sim911¬KC		66.46±3.89	0.13±0.11	71.68±3.55	0.19±0.14	69.85±2.61	59.87±4.45	0.11±0.21
Sim911¬CaCG	¬CoT	78.65±2.95	0.29±0.15	86.12±3.56	0.34±0.23	**77.11±3.44**	72.44±3.19	0.20±0.10
	¬FSP	62.90±3.51	0.09±0.06	63.41±4.71	0.08±0.06	65.97±3.12	60.11±3.51	0.01±0.20
	¬RAG	64.63±4.13	0.13±0.12	68.11±3.78	0.22±0.19	61.00±3.10	61.84±5.76	0.05±0.16
Sim911¬VLC		78.71±4.29	0.22±0.17	76.12±4.69	0.25±0.18	64.11±3.90	69.41±4.17	0.15±0.14
Sim911¬All(*GPT-4o*)		59.11±5.55	0.07±0.05	57.87±5.86	0.21±0.19	51.19±4.51	51.12±5.46	0.17±0.22
Sim911		**83.11±2.82**	**0.34±0.26**	**85.66±3.17**	**0.36±0.22**	73.31±2.16	**73.94±4.31**	**0.21±0.13**
		MENTAL HEALTH		**NON-NATIVE SPEAKERS**			**LOW-INCOME HOUSING**	
		BART(%)↑	Margin↑	BART(%)↑	Margin↑	Gunning Fog↑	BART(%)↑	Margin↑
Sim911¬KC		67.67±4.41	0.17±0.11	70.19±4.56	0.22±0.21	0.02±0.17	51.13±5.17	0.04±0.11
Sim911¬CaCG	¬CoT	81.67±5.40	0.31±0.11	82.13±4.87	0.45±0.10	0.17±0.11	74.49±4.11	0.13±0.18
	¬FSP	64.55±3.42	0.11±0.14	72.13±4.41	0.39±0.11	0.09±0.14	57.78±4.14	0.07±0.11
	¬RAG	76.71±4.45	0.01±0.09	75.33±5.19	0.23±0.18	0.11±0.17	64.65±5.77	0.01±0.14
Sim911¬VLC		83.22±3.77	0.21±0.20	81.19±3.11	0.24±0.26	0.07±0.09	63.18±5.13	0.16±0.13
Sim911¬All(*GPT-4o*)		61.62±4.15	0.04±0.11	69.13±4.36	0.21±0.22	0.03±0.12	44.44±3.72	0.04±0.02
Sim911		**86.16±3.37**	**0.33±0.14**	**84.41±4.80**	**0.48±0.11**	**0.17±0.13**	**77.98±3.37**	**0.19±0.21**

Table 2: Caller Image Tag-wise ***Equity*** Features Analysis. The metrics include ***BART*** (BART Score), ***Margin*** (Margin Score), ***NRCLex*** (accuracy on NRCLex results across tags), and ***Gunning Fog*** (margin score on Gunning Fog Index across tags).

managed to successfully deliver a baby on the phone!" Another shared: *"When it played the role of a kid caller, it acted just like a real child—refusing to do anything until his mom arrived on the scene."* These results emphasize Sim911's effectiveness in preparing call-takers by simulating diverse caller profiles and challenging real-life situations.

Comparison to Human-led Training: 9 out of 10 participants found Sim911 to be on par with or better than traditional human-led training. One participant remarked: *"Sim911 is a great starting point because it comes up more incident types than what we do right now. It's a valuable tool for enhancing our training."* Another said: *"It's impressive how Sim911 can simulate different callers (images). Trainees can be exposed to rare but useful calls that we could not (simulate) in the past."* These findings highlight that Sim911 not only complements human-led training but also enhances it by providing a broader range of incident types and scenarios that are difficult to replicate manually.

5 Related Work

Simulation-based training is a key component in various fields such as healthcare, aviation, and emergency services, where it provides a controlled environment for skill development without real-world risks (Suresh et al. 2023; Preiksaitis and Rose 2023; Daun et al. 2023). This method enhances critical thinking, decision-making, and practical skills by allowing repeated exposure to diverse and sometimes hazardous scenarios (Ibrahim et al. 2023; Flores, Ziakkas, and Dillman 2023; Rahman et al. 2023). Recent technological advancements, including Augmented Reality (AR), have begun to enhance traditional training setups, offering more immersive training experiences (Fitria 2023; Pfaff et al. 2020; Li et al. 2018; Ummenhofer et al. 2019), especially those for emergency responses (Parry et al. 2022; Mehta et al. 2022). Despite these innovations, most training simulations rely heavily on human-scripted scenarios and instructor feedback, limiting scalability and adaptability (Violato et al. 2023; Salvato et al. 2021; de Paula Ferreira, Armellini, and De Santa-Eulalia 2020). Large Language Models (LLMs) are emerging as a transformative tool for dialogue-focused simulations, able to generate dynamic interactions (Webb 2023; Thoppilan et al. 2022; Gong et al. 2023). But their integration into training programs must carefully address accuracy, ethical concerns, and potential biases to ensure effectiveness (Shanahan, McDonell, and Reynolds 2023; Shayegani et al. 2023; Yao et al. 2024; Salewski et al. 2024).

6 Summary

In this paper, we introduce Sim911, the first AI-driven simulation environment designed to assist 9-1-1 dispatcher training under emergency response scenarios. Sim911 aims to enhance the preparedness of emergency dispatchers, contributing to the resilience and safety of urban populations. Evaluation results on pre-configured runtimes and real-world deployment show that Sim911 effectively delivers realistic, authentic, and equitable simulations, to assist dispatcher training with the integration of knowledge construction, context-aware controlled generation, and validation with looped correction. More details regarding human-led training studies, running examples, user study settings, and extended related work are available in the Appendix (SimER-Project 2024).

This work can help emergency communications centers with limited staffing by allowing trainees to interact individually with the training program. Nearly 6,000 emergency communications centers could benefit from this training opportunity. Additionally, the GenAI-enabled solution can be extended to other training spaces, such as education and healthcare. In future work, we will extend Sim911 to other application domains for a broader social impact.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under Grants 2427711, the Google Academic Research Award, OpenAI Researcher Access Program, and the U.S. Department of Education under Grant R305C240010. The opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

References

Afonso, W. 2021. Planning for the unknown: Local government strategies from the fiscal year 2021 budget season in response to the COVID-19 pandemic. *State and Local Government Review*, 53(2): 159–171.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Carta, T.; Romac, C.; Wolf, T.; Lamprier, S.; Sigaud, O.; and Oudeyer, P.-Y. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, 3676–3713. PMLR.

Chen, Z.; Li, I.; Zhang, H.; Preum, S.; Stankovic, J. A.; and Ma, M. 2022. Cityspec: An intelligent assistant system for requirement specification in smart cities. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, 32–39. IEEE.

Chen, Z.; Li, I.; Zhang, H.; Preum, S.; Stankovic, J. A.; and Ma, M. 2023. CitySpec with shield: A secure intelligent assistant for requirement formalization. *Pervasive and Mobile Computing*, 92: 101802.

Chen, Z.; Sun, X.; Li, Y.; and Ma, M. 2024. Auto311: A Confidence-Guided Automated System for Non-emergency Calls. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 21967–21975.

Daun, M.; Grubb, A. M.; Stenkova, V.; and Tenbergen, B. 2023. A systematic literature review of requirements engineering education. *Requirements Engineering*, 28(2): 145–175.

de Paula Ferreira, W.; Armellini, F.; and De Santa-Eulalia, L. A. 2020. Simulation in industry 4.0: A state-of-the-art review. *Computers & Industrial Engineering*, 149: 106868.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Fitria, T. N. 2023. Augmented reality (AR) and virtual reality (VR) technology in education: Media of teaching and learning: A review. *International Journal of Computer and Information System (IJCIS)*, 4(1): 14–25.

Flores, A. D. C.; Ziakkas, D.; and Dillman, B. G. 2023. Artificial Cognitive Systems and Aviation training. *Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems*, 69(69).

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Google Maps Platform. 2024. Google Maps Platform. https://mapsplatform.google.com/. Accessed: 2024-05-15.

Ibrahim, S.; Lok, J.; Mitchell, M.; Stoiljkovic, B.; Tarulli, N.; and Hubley, P. 2023. Equity, diversity and inclusion in clinical simulation healthcare education and training: An integrative review. *International Journal of Healthcare Simulation*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, Y.; Wu, J.; Tedrake, R.; Tenenbaum, J. B.; and Torralba, A. 2018. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loria, S. 2018. textblob Documentation. *Release 0.15*, 2.

Ma, M.; Preum, S. M.; Ahmed, M. Y.; Tärneberg, W.; Hendawi, A.; and Stankovic, J. A. 2019. Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems*, 4(2): 1–28.

Ma, M.; Stankovic, J. A.; and Feng, L. 2021. Toward formal methods for smart cities. *Computer*, 54(9): 39–48.

Mehta, R.; Moats, J.; Karthikeyan, R.; Gabbard, J.; Srinivasan, D.; Du, E.; Leonessa, A.; Burks, G.; Stephenson, A.; and Fernandes, R. 2022. Human-centered intelligent training for emergency responders. *AI Magazine*, 43(1): 83–92.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3): 436–465.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

NICE. 2023. State of 911 Staff Performance and Retention Benchmark Study Report. https://www.nicepublicsafety.com/resources/state-of-911-staff-performance-and-retention-benchmark-study-report?UTM_source=NENAad.

NYC-911. 2022. Next-Gen 911 on Target for 2024 Completion. https://www.nyc.gov/content/oti/pages/press-releases/next-gen-911-on-target-2024-completion. Accessed: 08-15-2023.

Parry, A. E.; Kirk, M. D.; Colquhoun, S.; Durrheim, D. N.; and Housen, T. 2022. Leadership, politics, and communication: challenges of the epidemiology workforce during emergency response. *Human Resources for Health*, 20(1): 33.

Pfaff, T.; Fortunato, M.; Sanchez-Gonzalez, A.; and Battaglia, P. W. 2020. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*.

Preiksaitis, C.; and Rose, C. 2023. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR medical education*, 9: e48785.

Rahman, M. A.; Jia, L.; Mirza, E.; Preum, S. M.; Alemzadeh, H.; Williams, R. D.; and Stankovic, J. A. 2023. emsReACT: A Real-Time Interactive Cognitive Assistant for Cardiac Arrest Training in Emergency Medical Services. In *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 120–128. IEEE.

Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; and Akata, Z. 2024. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. *Advances in Neural Information Processing Systems*, 36.

Salvato, E.; Fenu, G.; Medvet, E.; and Pellegrino, F. A. 2021. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 9: 153171–153187.

Saxon, N.; Villena, P.; Wilburn, S.; Andersen, S.; Maloney, D.; and Jacobson, R. 2022. Annual Survey of Public Employment & Payroll Summary Report: 2021. *US Census Bureau*.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Shayegani, E.; Mamun, M. A. A.; Fu, Y.; Zaree, P.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

SimER-Project. 2024. SimER: AI-driven Simulation for Call-Taker Training in Emergency Responses. https://meiyima.github.io/simer.html#resources. Accessed: 2024-12-19.

Suresh, D.; Aydin, A.; James, S.; Ahmed, K.; and Dasgupta, P. 2023. The role of augmented reality in surgical training: a systematic review. *Surgical Innovation*, 30(3): 366–382.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ummenhofer, B.; Prantl, L.; Thuerey, N.; and Koltun, V. 2019. Lagrangian fluid simulation with continuous convolutions. In *International Conference on Learning Representations*.

Violato, E.; MacPherson, J.; Edwards, M.; MacPherson, C.; and Renaud, M. 2023. The use of simulation best practices when investigating virtual simulation in health care: A scoping review. *Clinical Simulation in Nursing*, 79: 28–39.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Webb, J. J. 2023. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. *Cureus*, 15(5).

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.