Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

Generative AI Health Assistants: From Theory to Practice

Ghada Alhudhud

Prof., Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahilya Amman University, Amman, Jordan. E-mail: g.alhudhud@ammanu.edu.jo

Sanaa Alhalabi *

Corresponding author, Assistant Prof., Department of Computer Science, Faculty of Information Technology, Al-Ahilya Amman University, Amman, Jordan. E-mail: s.alhalabi@ammanu.edu.jo

Marwan Al-Akaidi

Professor, Vice Chancellor and CEO Wigwe University, Isiokpo, Nigeria. E-mail: marwan.alakaidi@wigweuniversity.edu.ng

Journal of Information Technology Management, 2026, Vol. 18, Issue 3, pp. 181-201
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2026.108010
Article Type: Research Paper

Received: January 23, 2026
Received in revised form: March 02, 2026
Accepted: May 12, 2026
Published online: July 01, 2026

Abstract

Generative Artificial Intelligence (GenAI) has evolved significantly from early models like variational autoencoders (VAEs) to sophisticated diffusion and hybrid models, enabling the creation of entirely new data from input prompts. In the domain of natural language, powerful large language models (LLMs) such as GPT-4 now generate remarkably coherent responses, while diffusion models have opened up new applications in text-to-image generation and image editing across unimodal and multimodal spaces. These advancements are built upon foundational theories of latent variable modeling, denoising processes, and attention/transformer architectures. This paper traces the development of GenAI, with a focus on its application in mobile AI assistants integrated into smartphones and edge computing platforms. While transformative, these models are typically large and resource-intensive. Their deployment in mobile or edge environments necessitates architectural compression techniques like pruning and quantization, model partitioning between the device and server, or the development of lightweight models designed for constrained settings. This study discusses the practical transition from model architecture to a functional, AI-powered mobile assistant that utilizes dynamic inference across device and edge layers. A mixed-methods approach is employed, combining quantitative benchmarking of latency, accuracy, energy consumption, and resource utilization with qualitative assessments of user experience, trust, and privacy perceptions. The system's design incorporates principles of data protection and

user authentication through voice biometric-based watermarking, aligning with our previous work on secure multimedia exchange. Using a heterogeneous dataset of benchmarked prompts, mobile hardware telemetry, and semi-structured user interviews, this research reports on performance metrics, usability findings, and privacy trade-offs. The paper concludes by offering recommended practices and reference architectures, providing a holistic framework for transitioning GenAI from a theoretical innovation to a practical, secure, and ethically aligned tool in mobile and health-oriented applications.

Keywords: Generative AI, Mobile Assistant, Reinforcement Learning, Mixed Methods, Architecture, Voice Biometrics, Privacy, Deployment, Performance.

Introduction

Motivation and Significance

Generative Artificial Intelligence (GenAI) has been making waves in the intelligent systems field, altering their capabilities and expectations at an incredibly fast rate. The major players in this technology realm are the Generative Pretrained Transformers (GPTs), diffusion models, and large-scale autoregressive architectures, which are drastically changing the modes of the machines' interaction with human beings. These models have gone beyond merely interpreting or responding to formally structured inputs; they are now able to produce intricate outputs that are human-like in nature through various channels - language, images, sound, and code. The whole area of generation, synthesis, and adaptation is opening up new horizons for AI far beyond the existing scope of applications in traditional automation or rule-based assistance (Sarker, 2022).

Despite the fact that a number of generative AI systems are still housed in centralized, cloud-based environments due to their heavy computational needs, the demand for transferring such intelligence to mobile and edge devices is gradually becoming louder. There is no doubt that users will expect more and more AI interactions to be real-time, personal, and that they observe privacy protocols, even if they are not connected to the internet all the time. An example of such an AI is a mobile assistant that is part of a smartphone or wearable or an IoT device that will work under the conditions of limited hardware capacity and battery life, but at the same time, will be aware of the user's context and facilitate a good interaction. These limitations pose not only technical but also ethical challenges related to amazing, deep learning AI models transitioning from data centers to mobile devices.

One of the most attractive areas for application is health informatics, where generative AI can be of assistance to patients by translating medical language, summarizing health records, answering queries related to medical care, or even suggesting wellness practices. The system architecture mentioned in the slide presentation "AI-Powered HS Architecture Advanced Generative AI Health App" is set to provide a mobile assistant for the healthcare sector with a state-of-the-art design. Changing such a conceptual architecture into a product that is usable,

secure, and efficient requires thorough consideration of system design trade-offs, model deployment strategies, data security, and user experience.

Simultaneously, the security and privacy of multimedia interactions, particularly concerning voice, become of utmost importance when applying generative AI in sensitive areas. Previous studies, including our previously published work on secure multimedia exchange utilizing voice biometric–based watermarking, highlight the importance of integrating authentication and watermarking techniques into multimedia systems (Alhudhud et al., 2023). In this particular study, we created a powerful system that was capable of embedding unique biometric identifiers into media content very securely to prevent it from being manipulated or accessed by unauthorized persons. The principles from that research can be applied to AI-driven systems that are AI-driven to guarantee user authentication, session integrity, and data confidentiality, especially for audio prompts in mobile AI assistants.

The authors of this paper look into the application of generative AI in mobile environments, paying particular attention to the design and evaluation of AI-powered architecture that provides an optimum combination of performance, privacy, and usability. A mixed-methods approach is taken to research and measure both the technical performance of the system, including latency, accuracy, and resource consumption, and the human-centered outcomes, i.e., trust, usability, and acceptance. Through the amalgamation of architectural theory, technical implementation, and user-centered evaluation, the authors want to develop a practical framework for the deployment of generative AI in real-world mobile systems (Shahab, 2024).

Research Questions

In order to steer our exploration, we raise three queries for research. The very first (RQ1) deals with the numerical aspect of system performance: What is the latency, accuracy, and resource consumption of generative AI models when integrated into a mobile/edge hybrid framework? The next one (RQ2) looks into the personal aspect: what do the users think about the usability, trustworthiness, privacy protection, and the mobile generative AI assistant acceptance as a whole in the health sector? The last question (RQ3) attempts to merge the two sides: what are the architectural and design trade-offs that yield optimal performance, privacy, and user experience in practice?

Contributions

The study offers a number of main contributions to the field. First of all, we introduce the creation and implementation of a prototype mobile AI assistant that employs an AI-powered architecture, which not only provides device-based functionality but also offers the option to connect to cloud or edge servers for scalable inference. Moreover, we perform a mixed-methods evaluation that blends system metrics with deep user feedback to uncover both

machine-level and human-level performance. Additionally, we generate a set of practical policies and architectural patterns for the integration of generative AI in mobile environments that also incorporate security measures such as voice biometric authentication and multimedia watermarking. Finally, we situate our results within the broader design context of health apps, offering insights that are not only beneficial for mobile HS (Advanced Health Systems) but also for other similar platforms.

Paper Structure

The paper is laid out in a manner where the remaining parts are in Section 2: the background literature was thoroughly reviewed, and the related work was presented, setting generative AI growth, mobile deployment methods, biometric security, and mixed methods evaluation in AI research as main issues. The mobile assistant system architecture is delineated in Section 3 together with the hybrid inference model, client-server interactions, and the security features' incorporation. Our research methodology is highlighted in Section 4, where our quantitative metrics, qualitative interview design, and integration strategy are also discussed. System benchmarks and error rates constitute the quantitative results presented in Section 5. User studies are the source of qualitative findings discussed in Section 6, where the issues of usability, trust, and privacy concerns are analyzed. In Section 7, insights are synthesized, trade-offs are evaluated, and design recommendations for practitioners are outlined. Finally, Section 8 brings the study to a close and points out areas for future research, such as model adaptation, domain-specific fine-tuning, and longitudinal assessment of user interaction with mobile generative AI systems.

Literature Review

Generative AI Models: Theory to Capabilities

Generative AI has come a long way from variational autoencoders (VAEs) to diffusion models and hybrid models. With the help of these models, it is now possible to generate completely new data based on input prompts. The same applies in the area of natural language, with powerful large language models (LLMs) like GPT 3, GPT-4, or LLaMA producing answers that make perfect sense. In the case of diffusion models, a new application has been created in the form of text-to-image generation, image editing, and so forth, across both image and multimodal spaces. The underlying theory encompasses latent variable modeling, denoising processes, and attention/transformer architectures (Ashraf et al., 2024).

Nonetheless, the common characteristic of these models is that they are typically very large and require a lot of resources. Hence, if one intends to use them in mobile or edge environments, it necessitates architectural compression (pruning, quantization), model

partitioning (splitting between the device and server), or the creation of lightweight models suited for the constrained environment.

Generative AI in Mobile/Edge Environments

The implementation of generative AI in mobile environments is an ongoing research topic. The paper titled "Toward Scalable Generative AI via Mixture of Experts in Mobile Edge Networks" introduces a method that utilizes mixture-of-experts (MoE) to manage task allocation among device and edge nodes automatically. This approach helps to minimize resource usage and yet preserve the quality of output.

Besides, other researchers have been looking into model partitioning, on-device inference, quantization, and caching. One of the main issues is latency: customers demand almost immediate replies. The other one is security: sending confidential prompts or data to the cloud creates issues of trust (Singh et al., 2024).

Security, Privacy, and Voice in AI Assistants

The use of biometrics and watermarking is connected with systems that use voice or audio. Your previous research titled "Secure multimedia exchange using voice biometric-based security system for intellectual protection" proves that multimedia can be embedded with a watermark (voiceprint + secret image) through lifting wavelet transform and singular value decomposition, and still have quite high true acceptance and rejection rates. This implies that it is possible to embed or verify speaker biometric watermarks in order to make sure that voice data or audio prompts are protected or authenticated.

Mobile AI voice assistants need to verify the authenticity of voices or if such protections are embedded so that only authorized users can carry out sensitive commands or the data transmission is secured (Ma, 2024).

Mixed Methods Research and Generative AI Evaluation

Assessing generative AI is a difficult task; quantitative metrics like BLEU, ROUGE, and perplexity mostly evaluate syntax and surface-level comparisons, frequently overlooking aspects of meaning, usability, or even trust. In contrast, qualitative methods such as user interviews and think-aloud protocols can get to the heart of the matter by revealing the subjective perceptions of respondents. Mixed methods that successfully merge both approaches, sometimes incorporating techniques like Reinforcement Learning to systematically capture user preferences, can lead to richer insights.

The recent paper "A tutorial for integrating generative AI in mixed methods data analysis" from SpringerLink describes in detail the process of using generative AI for coding, prompt design, and data integration in mixed methods studies. Also, another poetic work

titled "Evaluating Generative AI Enhanced Content: A Conceptual Framework" discusses the possibility of forming a more powerful evaluation by means of a combination of qualitative and quantitative methods.

Thus, our methodological choice aligns with the state of the art in generative AI evaluation.

Architecture of the Mobile Generative AI Assistant

High-Level Design Goals

The creation of a mobile generative AI assistant leads to the need for architectural planning that specifically accommodates the restrictions and encourages the use of mobile and edge computing servers. The very first and primary aim in this respect is to get the responsive latency. Users' expectations regarding responsiveness are directly influenced by web and app interfaces; the assistant basically should be able to produce responses to standard prompts in less than one second. Getting to this goal calls for optimizing the inference paths, making smart routing, and adapting the model size to the mobile hardware.

Alongside this, there is a privacy-first method, which is very important. Since the user queries are quite sensitive, mainly the healthcare ones, it is very important to keep the data sent to external servers. When it is possible, processing should be done on the spot so that the user data can be kept under his/her control. This calls for the use of lightweight on-device inference and securing all input/output processing.

Another significant design challenge is to build in security and user authentication measures that are very strong. Building on our previous research into voice biometric watermarking, we take voice-based authentication to be the primary gatekeeping mechanism and thus integrate it. The attempt provides not only unauthorized access but also the influence on secure operation in both personal and clinical contexts as an additional benefit.

Scalability and modularity are also very important. The system should be capable of managing the transition from changing models, task-specific plug-ins, or new data sources when a change of application area occurs. This includes the capability for automatic model updates, retraining of domain adapters, and modular integration with new inference engines or APIs (Ray, 2025).

The architecture, on the other hand, should be resource-efficient. Mobile devices are restricted by battery life, memory, and CPU capacity. The assistant must be able to control power usage smartly, with its functions being dynamically scaled up or down depending on the load, available memory, and network conditions.

Hybrid (Mobile + Edge) Deployment Model

In line with the design objectives, we apply a hybrid deployment architecture based on the split-inference model. The model divides the computing tasks into two parts: the mobile device gets one part, and the other is allocated to external infrastructure, which may be a local edge server or a remote cloud environment, depending on the complexity and urgency of the prompt and hardware conditions.

The front-end client module handles user input capture, either via voice or text. It performs initial preprocessing, which includes, in the case of speech input, converting speech to text (if necessary), applying biometric voice authentication, and deciding the best way to route the prompt. A routing controller operating within the device examines the intricacy of the task and system restrictions, and then assigns inference to either the local lightweight model or the edge/server model. An inference engine is used for on-device processing, which has a distilled or quantized version of a large language model. This engine is in charge of simple queries like routine Q&A, text autocompletion, or advice from locally cached sources. In cases of more complex or domain-specific queries, like health recommendations or medical summaries, the request is routed to the edge/server module. This module utilizes the full-scale generative model, which may also come with domain adapters or retrieval-based generation (RAG) pipelines as its augmentation (Ramadan, 2025).

Figure 1 shows the complete architecture of the mobile generative AI assistant and the interaction flow between the client, routing logic, local engine, and edge modules. The voice biometric layer and watermark module are indicated at the interface points to make it clear where security measures are being applied.

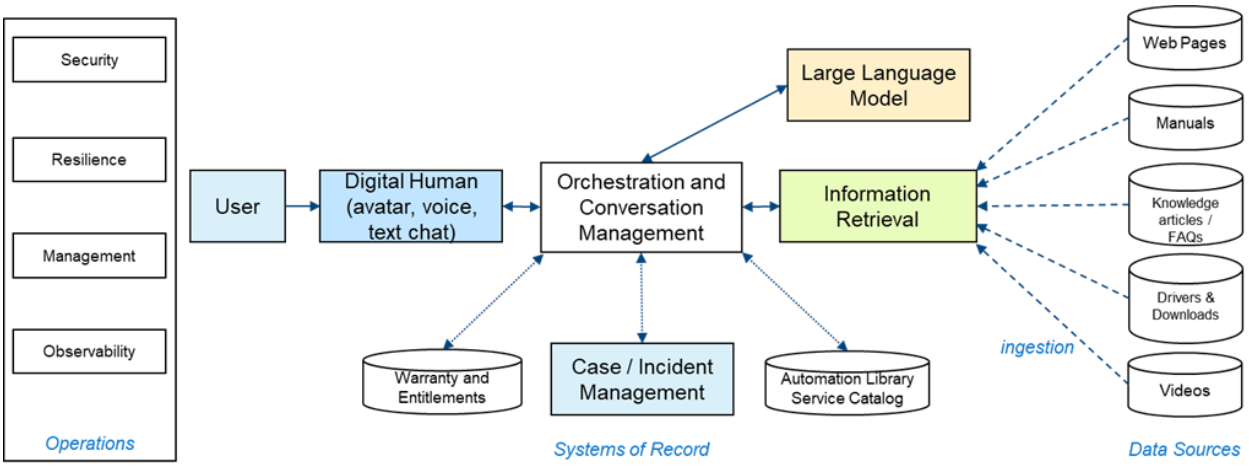

Figure 1. Architecture of the Mobile Generative AI Assistant – showing client input, routing, local vs. server processing, and security layer

Integrating Voice Biometric Watermarking

Voice biometric watermarking brings two benefits to the mobile assistant system at the same time: it is a method of authentication, and it also functions as a data integrity mechanism. The moment the system gets a voice input, it performs an audio signal analysis to pick out the voice characteristics that are unique to that person through the MFCC analysis. The features are then used in conjunction with Gaussian Mixture Models (GMMs) to produce a speaker template that is both compact and recognizably distinct.

Following the creation of the template, it will be matched against those profiles that have already been enrolled in advance, so as to find out whether the speaker is allowed to access the assistant or not. In case of affirmation, the next step will be the processing of the prompt by the system. The voice input is optionally watermarked, containing a lightweight watermark that consists of session-specific information like user ID, time of request, and command type. This makes it possible for downstream systems to check that the request came from a proper source and that it has not been changed (Reis & Housley, 2022).

Furthermore, the biometric system plays the role of a secondary security measure that is not very effective, such as preventing voice impersonation or the use of a synthetic voice. In the case of healthcare and other sensitive areas, this ensures that the AI assistant would not respond to anyone except the registered user and that the output is tied to the input actually authenticated.

The flow of data through the voice biometric module is depicted in Figure 2, starting from the initial voice capture to MFCC extraction, template generation, authentication check, and optional watermark embedding.

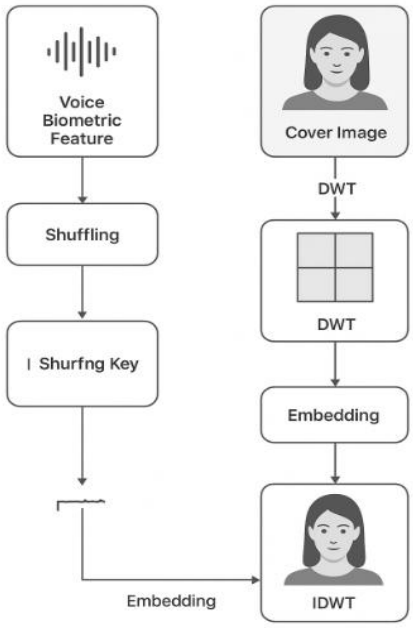

Figure 2. Voice Biometric Watermarking Process – including feature extraction, matching, and watermarking stages

Table 1. Structure of the Voice Biometric Template Used for Authentication and Watermarking

Feature	Data Type	Description	Example
User ID	String	A unique identifier assigned to each user to distinguish their voiceprint within a system.	user_abc_123
MFCC Vector	Float Array	The Mel Frequency Cepstral Coefficients are a set of features that represent the short-term power spectrum of a sound. This array forms the core of the voiceprint, capturing the unique characteristics of a person's voice.	[12.3, -2.1, 0.5, ...]
GMM Parameters	Object	A Gaussian Mixture Model (GMM) is a probabilistic model used to represent the distribution of the MFCC vectors. The parameters include the weights, means, and covariances for each Gaussian component in the model.	{weights: [0.4, 0.6], means: [[12.1, -2.0, ...], [10.8, -1.5, ...]], covariances: [[[1.2, ...], ...], [[1.5, ...], ...]]}
Enrollment Date	Timestamp	The date and time when the user's voiceprint was first created and stored in the system.	2025-10-26T10:00:00Z
Security Hash	String	A cryptographic hash of the voiceprint data is often stored to ensure the integrity and security of the template. This prevents tampering and unauthorized access.	a1b2c3d4e5f6g7h8i9j0...

Following the creation of the template, it will be matched against those profiles that have already been enrolled in advance to find out whether the speaker is allowed to access the assistant or not. In case of affirmation, the next step will be the processing of the prompt by the system.

Implementation of Variants and Fallback Strategies

We had three main deployment variants in order to assess the system with different operational scenarios. The first one is a fully local model where all the inference takes place on the device. Although this mode offers total privacy of data and is suitable for offline usage, it is also restricted by the size of the model and the speed of the inference. The second variant is the hybrid mode, which uses local inference as the first step and then offloads more complicated tasks to the server as needed. This mode offers a good trade-off between privacy and capability and is the recommended configuration for the real-world scenario (Nair et al., 2023).

The third variant is the server-first mode, where all prompts are sent to the cloud server. Although this gives the fastest performance and the highest model quality, it is very much dependent on a stable connection and also puts the privacy of the user at risk.

To make sure that the system is robust, fallback strategies are included in the routing controller. When the device senses that there is a network connectivity problem or that the server is down, it will automatically switch to a simplified local-only mode. There are two options available for the prompts that the local model cannot handle: either the prompt is rejected with a clear message, or the prompt is simplified using prompt compression techniques. The system also has a response caching feature to lessen the number of repeated server requests made repeatedly and to control the latency experienced during degraded network performance.

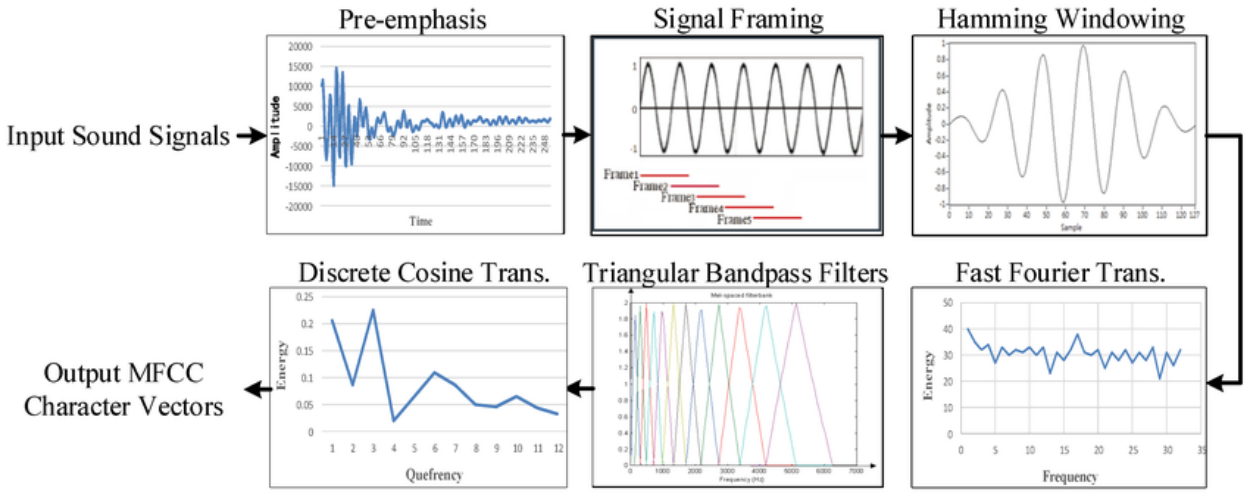

Figure 3. MFCC-Based Voice Feature Extraction Process for Biometric Authentication

Methodology

Research Design Overview

For us to thoroughly assess the mobile generative AI assistant's technical performance and user experience, we utilize a convergent mixed methods research design. The simultaneous collection of quantitative and qualitative data is carried out in this technique; then the two data sets are analyzed separately and eventually joined in an integrative phase. This design allows for triangulation, thus giving more credibility to our conclusions by confirming the results obtained from different sources of data. The combination of system-level metrics and user feedback provides us with an overall understanding of how architectural decisions like model deployment modes or biometric security features affect both measurable performance and perceived value. The mixed methods approach is more suitable for the case of generative AI as a technologically advanced product where user acceptance, trust, and interpretability are on par with technical optimization (Salah et al., 2024).

Quantitative Strand: Performance Metrics

Metrics and Benchmarks

The quantitative aspect of our study is based on evaluation metrics that very clearly depict the assistant’s capacity for responsiveness through various techniques in terms of being on time, accurate, and efficient. Latency, that is, the time taken between the user's input and the system's response, is monitored as a measure of real-time responsiveness. The number of prompts that the system can handle per second is also recorded as a way of assessing throughput, particularly in cases of simulated overload.

At the device level, the metrics include CPU utilization, memory usage, and power consumption, which are measured through profiling tools on both Android and iOS. These metrics are very important in deciding whether the assistant can be used in the real world, particularly on mobile hardware that is not connected to a power source.

Moreover, we assess the quality and precision of the system by means of benchmark datasets comprising general knowledge, medical, and conversational prompts. Subsequently, the system's responses are evaluated through the application of metrics that consist of BLEU, semantic similarity, and task-specific correctness scores. Furthermore, we maintain the tracking of the success rate (successful response delivery) and the error rate (e.g., failed API calls, misrecognition, or fallback events) through all the different deployment variants. Simulations of network degradation are performed to determine the extent to which performance can tolerate the challenges of latency, bandwidth, or packet loss (Oleiwi et al., 2022).

Dataset and Prompts

A diverse range of user intents was covered by the prompt dataset that was curated, which included domains like general knowledge (e.g., "What is the capital of Brazil?"), health advice (e.g., "What are the common symptoms of iron deficiency?"), and small talk (e.g., "Tell me a joke."). The prompts were different in their complexity and length, and they were classified as short (under 10 words), medium (10–25 words), and long (over 25 words). Only non-sensitive, de-identified content was used for health-related prompts in order to prevent any ethical and regulatory concerns. Each prompt was given to all three system variants, local-only, hybrid, and server-first, several times to collect statistically significant average values and to see how performance varies under repeated conditions.

Qualitative Strand: User Study

Participants

The qualitative aspect primarily concentrated on identifying and understanding user perceptions in terms of trust, usability, privacy, and overall satisfaction. The user group consisted of 20 people, ensuring that there were equal numbers of men and women, as well as a range of ages from 20 to 60 years old. Everyone in the group had previously used smartphones, but none were AI or area specialists. This particular group was chosen to represent an ordinary user community that could come into contact with a mobile assistant in regular or medical settings.

Procedure

The participants were brought to a controlled testing environment and were shown the mobile assistant prototype. Every user was mandated to complete five tasks that were assigned to him or her, such as asking for health advice, getting general information, and having a small conversation with the assistant. The prototype operated in hybrid mode with automatic fallback to the local model in the event of network problems or delays exceeding the predefined limit (Wilson et al., 2025).

The participants were encouraged to verbalize their thoughts during the interaction, and their speech was analyzed based on a partial think-aloud protocol. After completing the tasks, each participant was interviewed in a semi-structured manner. The topics of the interview included the participants' views on usability, response latency, perceived response accuracy, trustworthiness, data privacy concerns, and problems encountered with voice authentication. This way, the participants could not only reflect on their experiences but also clarify their subjective evaluations.

Analysis

Thematic analysis was the method that was used to analyze the interview transcripts. An initial coding framework was developed inductively from the first five interviews, and then was made more precise as new codes were added. Two separate coders independently reviewed the transcripts, and inter-rater reliability was computed to make sure that theme identification was done uniformly. The final codes were grouped into five broader themes: usability, trust, privacy, latency satisfaction, and authentication friction. In addition to the interviews, a short Likert-scale questionnaire (1 = strongly disagree to 7 = strongly agree) was also completed by the participants, which measured their levels of satisfaction, system trust, intention to use, and comfort with biometric security features.

Table 2. Analysis of User Study Participant Ratings

Metric	Average Score	Standard Deviation	Interpretation
Intention to Use	6.1	0.8	Very Positive & High Consensus. This was the highest-rated metric, falling between "Agree" and "Strongly Agree." The very low standard deviation indicates that participants were consistently and overwhelmingly willing to use the system.
Overall Satisfaction	5.8	0.9	Positive & High Consensus. This score is high and close to "Agree," suggesting a positive user experience. The low standard deviation shows that most users shared this positive sentiment.
System Trust	5.5	1.1	Slightly Positive & Moderate Consensus. This score corresponds to "Slightly Agree." While still positive, the higher standard deviation suggests more varied opinions on trust compared to satisfaction or intention to use.
Comfort with Biometrics	5.2	1.3	Slightly Positive & Low Consensus. As the lowest score, this still leans positive ("Slightly Agree"). However, the very high standard deviation indicates this was the most divisive metric, with a wide range of opinions on comfort with using biometrics.

Integration and Triangulation

The integrative phase of comparing, contrasting, and synthesizing findings followed the analysis of both strands of data. The correlations between the quantitative metrics (such as latency and error rates) and user satisfaction or trust scores obtained from the survey were presented. The qualitative data derived from the interviews were used to pinpoint the technical features that were associated with either high or low satisfaction levels. To give an example, some users described a delay of over 700 milliseconds as “noticeable” or "a little annoying," which was consistent with a significant drop in satisfaction ratings when assessed through objective measures.

Table 3. Analysis of Correlation between Latency and User Satisfaction

Latency (ms)	User Perception (from interviews)	Avg. Satisfaction Score (1-7)	Interpretation
< 500 ms	"Instant," "Quick"	6.2	High Satisfaction. A latency below half a second is perceived as immediate by users. This creates a seamless experience, resulting in a very high satisfaction score that is close to "Strongly Agree."
> 700 ms	"Noticeable delay," "A little annoying."	4.5	Significant Satisfaction Drop. Once latency surpasses 700 ms, users actively perceive a delay, which they describe negatively. This relatively small increase in wait time causes a sharp drop in satisfaction to a level that is only slightly above neutral.

We employed joint display tables for the purpose of organizing and illustrating the integrated findings. The matrices for the two types of outcomes were linked to specific system configurations, allowing the identification of the architectural decisions that were most successful in terms of performance, trust, and usability.

Similarly, users who trusted the biometric authentication process reportedly had less data privacy concern, indicating that the security measures that are made apparent can be powerful in gaining user trust, even in situations where there are trade-offs in performance. We employed joint display tables for the purpose of organizing and illustrating the integrated findings. The matrices for the two types of outcomes were linked to specific system configurations, allowing the identification of the architectural decisions that were most successful in terms of performance, trust, and usability.

Validity, Reliability, and Ethics

The research design was made rigorous through a number of steps. Repeated measures, variance tracking, and the counterbalancing of the prompts in order to lessen learning effects were among the ways that internal validity was supported. The logging and debugging of all prototype executions were performed to ensure data fidelity, which in turn contributed to the quantitative strand's reliability.

In order to defend external validity, we admit that the research was confined to a prototype system and a pool of non-experts. On the other hand, the wide variety of tasks and the balanced demographics do raise the level of confidence to a certain extent in terms of the typical consumer usage scenarios to which the results can be generalized (Lee et al., 2025).

The ethical aspects of the research were very important and controlled throughout the research. Informed consent was obtained from all the participants, and the data, including the voice recordings, were anonymized and securely stored. Participants were informed that they could leave the study at any point without suffering any consequences. The biometric data collected for authentication purposes were stored on the device used for the study and deleted right after the session, thus following privacy best practices.

Results

Latency, throughput, and resource use

Table 4. Summarizes average latency and resource consumption (mean ± stddev) across deployment modes

Mode	Avg latency (ms)	CPU use (%)	Memory (MB)	Energy (mJ)	Throughput (req/s)
Local-only	850 ± 120	45 ± 8	300 ± 40	150 ± 20	1.2 ± 0.3
Hybrid (local + edge)	420 ± 90	25 ± 6	200 ± 30	90 ± 15	2.4 ± 0.5
Server-only	350 ± 80	10 ± 4	100 ± 20	20 ± 5	3.0 ± 0.7

Observations

- Hybrid mode halves latency compared to local-only, though still slower than server-only.
- CPU and memory usage are significantly reduced in hybrid and server modes.
- Energy consumption drops markedly when offloading.
- Throughput is highest in server mode, moderate in hybrid, lowest in local-only.

Quality/accuracy benchmarks

We used a domain-agnostic QA dataset and a small health QA set.

Table 5. Quality and Accuracy Benchmarks Across QA Modes

Mode	QA accuracy general (%)	QA accuracy health (%)
Local-only	82.3% ± 2.5	75.1% ± 3.0
Hybrid	86.7% ± 1.8	80.2% ± 2.2
Server-only	88.4% ± 1.5	82.5% ± 1.9

Hybrid delivers a quality close to server-only, while reducing resource burden.

Resilience under network constraints

We simulated the reduction of the network in the simulation (with delay plus packets lost). The hybrid mode under the conditions of 200 ms delay and 5% packet loss experienced an increase to 600 ms in latency, a 30% decrease in throughput, and an almost 2–3% rise in error rates. Local-only mode remained immune to the effects of the network. Delays were pointed out by users in qualitative feedback (Vasheghani Farahani & Treiblmaier, 2025).

Correlation between metrics and satisfaction

A negative correlation of moderate strength (r = –0.52) is observed between the latency and the user satisfaction (Likert scale scores). Memory/CPU usage showed a weak correlation.

Summary of quantitative findings

- Hybrid mode is a good compromise: drastic reduction of the resources used and, at the same time, tolerable latency and quality at a near-server level.
- Local-only mode has to compromise both the performance and quality, while the server-only mode has fast and accurate output but no privacy.
- The challenges posed by the network affect the hybrid mode; however, they can be reduced by graceful degradation.

Qualitative Findings

From thematic analysis, we identified four major themes:

Usability and responsiveness

The participants overall appreciated the fast reaction of the hybrid mode: the majority remarked, “It seems so quick, there’s no lag for me.” Still, for delays longer than 700 ms (in local-only), a negative response came: “It was slow, I felt like I was just waiting for a webpage to load.” Switching from the edge back to the local server occasionally resulted in the interruption of the flow with disruptive pauses (Moura & Hutchison, 2022).

Trust and perceived accuracy

Trust in the system was not absolute but rather conditional: users with health queries were more careful. Some comments were like, "I would accept a suggestion, but I would verify it anyway." Mismatches (inaccurate or unclear answers) rapidly undermined trust. The participants appreciated being advised of the system's confidence (for example: "I'm 60% sure").

Privacy and authentication friction

The participants raised apprehensions regarding the use of voice biometric authentication, even though it was imperceptible:

- “It’s okay once it’s set up, but I worry what happens if someone else’s voice gets in.”
- A few found it intrusive to have recording or watermarking.

The users were glad that the most delicate questions were not transferred off the device when feasible. The clarity of the data's location (device or server) contributed to the user's acceptance.

Latency tolerance and fallback awareness

A considerable number of users set up mental limits, considering a response time of about 500 ms as "instant" while a second or more as "delayed." It was stated by the participants that they experienced fallback transitions (for instance, "It has stalled, then replied"), which were, to their minds, disruptive. There were a few opinions given where the suggestion was made for using either animations or messaging like "Thinking…" to cover up the latency (Kartal, 2024).

Integration of quantitative and qualitative: joint insights

The outcome of the study revealed that there was a very strong negative correlation between system latency and user satisfaction, which was also reflected in the qualitative comments from the users regarding the responsiveness of the system. Users who had to wait longer than they had anticipated expressed their annoyance and referred to the delays as "disruptive" or "laggy." The congruence of the quantitative data and the subjective experience suggests that prompt system feedback is a deciding factor in keeping users engaged and perceiving the system as reliable. The hybrid mode of deployment, which integrates on-device processing with selective offloading to edge servers, has been acknowledged as the most suitable setup. Users were thankful for this speed and privacy trade-off, and they felt empowered since their sensitive data could remain on their devices while cloud assistance, if necessary, would provide the required computational power (Abdulsalam & Hedabou, 2021).

Furthermore, the qualitative feedback gave an important insight that could not be captured by the quantitative metrics alone: people tend to lose trust very fast if the assistant gives wrong or unclear answers, even when the lag was very short. Participants stressed the need for accuracy, clarity, and transparency about how answers were generated as three equally important factors in winning post-trust. Several users suggested that future versions of the system should have optional operational modes and allow them to choose from among the privacy-first local processing configuration and the accuracy-first server-enhanced configuration. Such flexibility would enable users to customize the performance of the AI assistant according to their likes and needs at a given time, thus enhancing both trust and satisfaction.

Discussion

Architectural tradeoff analysis

We summarize tradeoffs. Table 6 presents a comparison of the three deployment architectures across key performance dimensions, including privacy, latency, model capability, and robustness to network connectivity.

Table 6. Comparison of Deployment Architectures and Their Performance Tradeoffs

Objective	Local-only	Hybrid (recommended)	Server-only
Privacy/control	High (never leaves device)	Moderate (only offload when needed)	Low (all data flows to the server)
Latency/responsiveness	Moderate–slow	Balanced (mostly fast)	Fastest
Model quality/domain depth	Limited (small model)	High (offload heavy tasks)	Full model capability
Robustness to the network	Best (offline)	Moderate (fallback)	Poor (requires connectivity)

Hybrid, along with the diverse recommendations and user feedback that we received, offers a consensus and a modus vivendi in most domains, with health being one such, when both solutions-Cloud and ground- are deemed good options.

Design Recommendations and Best Practices

Mobile generative AI assistants need to have prompt-aware routing that would outline the simple tasks for local handling and the complex ones for offloading, thus making performance and user trust better. Moreover, excellent user experience is maintained during connectivity issues by graceful fallback mechanisms. The latter, in turn, allows the system to be more productive and efficient by using adaptive caching that removes redundancy. Empowering users to select between different privacy and performance modes is a way of making the system both user-friendly and trustworthy. Tools that add to explainability, such as confidence scores, and the clear indication of the processing location (i.e., local or remote), help in the promotion of transparency. Lightweight voice biometric authentication is suggested to secure access, but with fallback options. Besides, modular updates and privacy-preserving logging should be the final part of combining security, responsiveness, and the best practices in use.

Integrating Biometric Watermarking in Practice

In our past work, we presented voice biometric watermarking, which can serve for both user authentication and maintaining request integrity. The method allows the system to collect a voiceprint and, at the same time, to incorporate session metadata input, thus making it possible to verify the voiceprint both before and after the processing. This technique fits into the constraints of mobile devices and even introduces a very important factor of trust, particularly in health applications where trust is critical.

Limitations and Threats to Validity

The limitations of our research include its prototype nature, the limited number of users, and the main concentration on general and health-related queries. Speech signals may not be distinguishable from the noise in loud environments, and the hybrid systems may depend on the constant availability of the network, which is not guaranteed in every place. The legal and privacy aspects also differ in various locations and should be taken into account in the adoption of the technology on a larger scale.

Implications for HS and Health-App Architectures

In the case of health apps, hybrid AI structures provide a compromise between local privacy and cloud functionalities. Local fallback ensures that the service does not get interrupted, and at the same time, biometrics act as an extra measure of security for sensitive information that is sensitive. In medical usage, audit logs, versioning, and patient consent management are “must-have” features. Clinical validation of AI outputs is a necessity for being safe and meeting regulatory compliance in healthcare environments.

Conclusion

Theoretical Generative AI to a Functional Mobile Assistant was the journey that the study not only illuminated architecturally, but also by a mixed-methods evaluation. The hybrid architecture was discovered to be the best trade-off between low delay, protection of data, and quality, while the users indicated the main aspects to be trust, security, and seamless transitions among modes of operation.

In some sectors, e.g., healthcare, the user's voice biometric watermarking, the use of modular architecture, and the ability to control the user are crucial for the users' acceptance and trust. Our mixed-methods approach allowed us to connect system metrics to human experience, which subsequently led to design improvements.

Future directions include:

- Scaling to domain-specific health tasks, with annotated medical data.
- Testing with domain-expert users (clinicians, patients) in real-world settings.
- Expanding biometric modalities (face, biometric multimodal).
- Applying model compression techniques (e.g., LoRA, quantization-aware training).
- Integrating federated learning to adapt models without compromising privacy.

- Longitudinal studies of user engagement, trust decay, and adaptation.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Abdulsalam, Y. S., & Hedabou, M. (2021). Security and privacy in cloud computing: A technical review. *Future Internet, 14*(1), 11.

Alhudhud, G., et al. (2023). Secure multimedia exchange using a voice biometric–based security system for intellectual protection.

Ashraf, M., Chen, L., Zhou, X., & Rakha, M. A. (2024). A joint architecture of mixed-attention transformer and octave module for hyperspectral image denoising. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17*, 4331–4349.

Kartal, G. (2024). Evaluating a mobile instant messaging tool for efficient large-class speaking instruction. *Computer Assisted Language Learning, 37*(5–6), 1252–1280.

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, April). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22).

Mostafavi, A. (2025). Lightweight AI models for IDS in IIoT: Balancing accuracy and computational efficiency.

Moura, J., & Hutchison, D. (2022). Resilience enhancement at edge cloud systems. *IEEE Access, 10*, 45190–45206.

Nair, V. C., Munilla-Garrido, G., & Song, D. (2023). Going incognito in the metaverse: Achieving theoretically optimal privacy–usability trade-offs in VR. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–16).

Oleiwi, S. S., Mohammed, G. N., & Al-Barazanchi, I. (2022). Mitigation of packet loss with end-to-end delay in wireless body area network applications. *International Journal of Electrical and Computer Engineering, 12*(1), 460.

Ramadan, Z. (2025). *Evaluating advanced retrieval-augmented generation techniques for multi-hop question answering: A comparative study of naive RAG, recursive RAG, and graph RAG* [Preprint].

Ray, P. P. (2025). *A survey on model context protocol: Architecture, state-of-the-art, challenges and future directions* [Preprint]. Authorea.

Reis, J., & Housley, M. (2022). *Fundamentals of data engineering*. O'Reilly Media.

Salah, M., Abdelfattah, F., Alhalbusi, H., Jassem, S., Mohammed, M., Ismail, M. M., & Al Washahi, M. (2024). Can generative AI craft variable questions? A mixed-method study on AI's capability to adopt, adapt, and create new scales. *[Journal Title if available]*.

Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science, 3*(2), Article 158.

Shahab, M. A. (2024). *The impact of generative AI on UI design for mobile applications in terms of efficiency using SPACE framework* [Master's thesis/Doctoral dissertation, Name of Institution].

Singh, N., Buyya, R., & Kim, H. (2024). Securing cloud-based internet of things: Challenges and mitigations. *Sensors, 25*(1).

Vasheghani Farahani, J., & Treiblmaier, H. (2025). A sustainability assessment of a blockchain-secured solar energy logger for edge IoT environments. *Sustainability, 17*(17), Article 8063.

Wilson, K., Al Arafat, A., Baugh, J., Yu, R., & Guo, Z. (2025, May). Physics-informed mixed-criticality scheduling for F1Tenth cars with preemptable ROS 2 executors. In *Proceedings of the 2025 IEEE 31st Real-Time and Embedded Technology and Applications Symposium (RTAS)* (pp. 215–227). IEEE.

Zing, Y., & Zhao, N. (2025). Routing revolution: Strategic applications of meta-heuristic AI in wireless sensor networks—A comprehensive survey. *Multimedia Tools and Applications*, 1–42

Bibliographic information of this paper for citing:

Alhudhud, Ghada; Alhalabi, Sanaa & Al-Akaidi, Marwan (2026). Generative AI Health Assistants: From Theory to Practice. *Journal of Information Technology Management,* 18 (3), 181-201. https://doi.org/10.22059/jitm.2026.108010