

A. Alibekov^{1,2} , G. Askar^{1,2} , Z. Makhataeva^{1,2*} ,
G. Nauryzbaikyzy^{1,3} , K. Sadirova^{1,3} , H. Varol^{1,2} 

¹ Private Institution “Institute of Smart Systems and Artificial Intelligence” (ISSAI),
Nazarbayev University, Astana, Kazakhstan

² Al-Farabi Kazakh National University, Almaty, Kazakhstan

³ K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan

*e-mail: zhanat.makhataeva@nu.edu.kz

MULTILINGUAL GENERATIVE AI AVATAR: EDGE AND CLOUD-BASED REAL-TIME DEPLOYMENTS

Abstract. This paper introduces a multilingual avatar system designed for real-time human-AI interaction in Kazakh, Russian, and English. AI-enhanced avatars present novel ways of communication and interaction between users and technology. This work follows current trends in Human-Computer Interaction (HCI), bringing innovations towards text-, voice-, and vision-supported embodied AI agents. We present and evaluate two practical configurations for the proposed multilingual avatar system (i.e., practical cloud-based avatar configuration and practical edge-deployed avatar configuration, both working in real-time). Both methods utilize models developed and fine-tuned at the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University, based in Astana, Kazakhstan. The cloud-based configuration uses ISSAI’s Large Language Model (LLM) “Oylan” and the Text-to-Speech/Speech-to-Text (TTS/STT) services of the ISSAI’s “MangiSoz” platform. On the other hand, the edge-deployed avatar configuration uses the ISSAI’s small vision-language model “Qolda” based on the InternVL3.5 architecture with a Qwen3-4B language backbone. For speech processing, edge-deployed configuration uses on-device speech processing tools (i.e., Whisper Turbo for Automatic Speech Recognition (ASR) and Matcha-TTS for speech synthesis). The avatar interface is built using the popular 3D-stylized avatar creation platform “Ready Player Me”. During the evaluation of 96 queries across three languages (specifically, 32 Kazakh-language queries, along with 32 Russian and 32 English translations), the edge-deployed avatar configuration showed a substantial performance advantage, reducing end-to-end response latency by 45.5% (2.28 s vs. 4.18 s) compared to the cloud-based avatar configuration. In a user study, seven participants tested the edge-deployed avatar configuration, and one participant tested the cloud-based configuration, which has higher response latency affecting its usability. The study results highlight the potential of a small language model (SLM) integrated with speech detection/generation models on the edge to create real-time, responsive, offline communication with avatar-enhanced AI agents within the trilingual environment of Kazakhstan.

Keywords: Human-AI Interaction, avatars, large language model (LLM), text-to-speech (TTS), speech-to-text (STT).

1. Introduction

Human-AI Interaction is considered as a new era in Human-Computer Interaction (HCI) [1]. AI avatars present a novel approach to enhancing voice- and text-based interaction with AI systems. These avatars incorporate visual and auditory modalities to facilitate natural, human-like communication, supported by features such as appearance, voice, and behavior. Ongoing technological advancements enable the development of increasingly realistic interactions through varied tones, inflections, facial expressions, gestures, and body language [2]. Over the years, numerous studies have explored how humans

intuitively interpret nonverbal messages conveyed through gestures, establishing communication channels in virtual social and collaborative environments [3]. Researchers envision that in the AI era, traditional human-to-human collaboration will eventually transform into human-AI interaction [4]. Since humans rely on spatial audio-visual cues, such as speech and facial expressions, for effective communication, efficient human-AI interaction requires information delivery modalities that extend beyond textual forms [5].

The evolution of AI-based avatars traces back to the creation of early chatbots [6], followed by the development of systems based on neural networks

and language models [7]. In 2023, the release of GPT-4, a Large Language Model (LLM) integrated with a chat-based interface, significantly accelerated research in this domain [8]. Concurrently, work in the field of HCI explored the concept of digital human-like agents, laying the foundation for modern avatar research [9]. Subsequent advancements in Natural Language Processing (NLP), Text-to-Speech (TTS), and Speech-to-Text (STT) technologies empowered digital assistants to recognize, transcribe, and generate human-like speech [10]. The convergence of these technological breakthroughs led to the emergence of AI avatars driven by LLMs and multimodal generative AI models [11]. These avatars are capable of operating in a human-like manner within shared virtual spaces, responding dynamically to descriptive prompts and contexts [12]. Although the foundational technologies existed previously, recent optimizations have yielded faster, more computationally efficient systems that are easier to integrate into diverse applications.

In general, cloud-based deployment is a common architectural approach for avatar systems, where avatar rendering, animation, and speech processing are implemented on remote cloud servers [13]. According to [14], paid Application Programming Interfaces (APIs) are supported by cloud services and capable of offering higher-quality outputs. However, they are prone to increased latency, which affects the system’s temporal responsiveness and reliability. In addition, paid API services raise privacy and data security issues related to user safety. For example, stored user voice recordings could be used to generate deepfakes, or data leakage could expose sensitive user information from conversations to third parties [15]. As a solution, edge deployment of avatar systems has been proposed as an alternative to cloud-based deployment. Specifically, edge-deployed avatar systems are capable of running all necessary components on the edge devices, such as user laptops or other mobile devices. They should have certain hardware capabilities, such as a GPU with sufficient computational power and memory. For example, in [16], the authors present an avatar system operating in Virtual Reality (VR), integrated with an edge-deployed Mistral 7B LLM, Whisper ASR, and Edge-TTS technologies via local API calls. For the User Interface (UI), the authors utilized the stylized avatar framework and lip synchronization technology to animate the avatar’s speech. The system features a control button to activate voice input, which stops automatically when the user finishes speaking. According to the authors,

the average latency, measured from the moment the user stops speaking to when the avatar begins its response, is 3.2 seconds.

In this work, we created a multilingual avatar system that leverages AI models and speech processing tools across cloud-based and edge-deployed configurations, both developed at the Institute of Smart Systems and Artificial Intelligence (ISSAI), based in Astana, Kazakhstan. In the Kazakhstani context, existing studies on avatars and AI have not yet demonstrated cloud-based and edge-deployed system configurations capable of real-time interaction in the Kazakh, Russian, and English languages. For the practical edge-deployed avatar configuration, we used ISSAI’s “Qolda”, a small vision-language model based on the InternVL3.5 architecture with a Qwen3-4B language backbone [17]. Qolda has 4.3 billion parameters. For edge speech processing, we used fine-tuned Whisper Turbo ASR and Matcha-TTS. For the practical cloud-based avatar configuration, we used ISSAI’s LLM “Oylan” API (<https://issai.kz/documentation/oylanApi>) and the TTS/STT services of ISSAI’s “MangiSoz” translation and speech processing platform running on the ISSAI server (<https://mangisoz.nu.edu.kz/backend/api-docs/en>). For the evaluation of the proposed avatar system configurations, we used 96 queries, with 32 queries in each language (i.e., Kazakh, Russian, and English). We designed a user study for the subjective evaluation of usability, cognitive load, and performance involving the edge-deployed avatar configuration. We analyzed overall response latency and computational resource utilization for both cloud-based and edge-deployed avatar configurations. In a user study, seven participants tested the edge-deployed avatar configuration, and one tested the cloud-based avatar configuration. Additionally, we investigated whether the benefits of reduced latency and offline accessibility justify the hardware requirements for edge-deployed avatar configuration.

The main contributions of this paper are the following. To the best of our knowledge, this is the first avatar system with two practical configurations (i.e., cloud-based and edge-deployed) and capable of real-time trilingual interaction in Kazakh, Russian, and English using locally developed AI models. This addresses a critical gap in the literature on low-resource Turkic languages, where no comparable system has previously been demonstrated. We evaluate the performance of cloud-based and edge-deployed avatar configurations that use different models, runtimes, and server conditions, yet both use a browser-

based 3D avatar rendering platform. Edge-deployed ISSAI avatar configuration shows 45.5% end-to-end latency reduction compared to the cloud-based ISSAI avatar configuration. In a user study involving seven participants, the edge-deployed avatar configuration was tested to evaluate its usability and effect on cognitive load. Meanwhile, one participant was offered the cloud-based avatar configuration due to its higher response latencies. Also, we performed a detailed component-level analysis of GPU resource allocation within edge-deployed ISSAI avatar configuration for concurrent multi-model deployment scenarios. The results demonstrate that 3D-stylized avatar rendering and AI inference can coexist on a single consumer-grade laptop GPU, which further can increase the accessibility of the avatar system.

The remainder of this paper is organized as follows: Section 2 provides an overview of existing research on avatars, covering their applications, advantages, and limitations. Section 3 details the proposed architecture and the methodology used for performance testing. Section 4 presents system evaluation and user study results, including latency analysis, resource usage, and performance comparisons across different languages and deployment architectures. Section 5 analyzes the trade-offs between cloud-based and edge-deployed avatar configurations, existing limitations, and offers recommendations for future research. Finally, Section 6 concludes the paper.

2. Literature review

Avatars are increasingly deployed across diverse domains, including education, healthcare, media, customer service, and gaming [18], [19]. In [20], researchers have developed avatars to interact with patients and assess their cognitive states. In [21], the authors present LLM-enhanced avatars designed for gaming environments to improve players' social communication skills through dialogue. The authors highlight the ability of avatars to provide detailed, real-time feedback that facilitates the improvement of learning skills. Many studies in the literature also explore the efficacy of avatar-based pedagogical agents. The authors in [22] outline a theoretical framework for the design of AI-enhanced educational avatars, emphasizing their capacity to support personalized and context-aware interactions through individualized learning, contextualized instructions, enhanced immersion, and the scaffolding of self-regulation skills. According to the authors, human learning strategies are effectively stimulat-

ed by interaction with AI-based avatars. However, valid concerns remain. For example, incorrect information generated by an LLM could be perceived as genuine and convincing when delivered by an AI-embodied avatar with a unique personality. Furthermore, when technologies such as LLMs, ASR, and cognitive services (including face recognition, emotion recognition, and user context evaluation) operate in the cloud, critical issues may arise, particularly regarding data security and user privacy.

In [23], the authors explore how an avatar's interactional cues and visual realism influence the overall perception of the system. The research indicates that visualizing the avatar's engagement through visual signals enhances the user's positive perception of the system. Regarding visual realism, studies suggest that photo-realistic avatars are not always perceived as the most convincing agents. For example, avatars that appear human but lack corresponding behavioral characteristics can fall into the "uncanny valley" thereby diminishing user trust [23]. Recent research has introduced sophisticated frameworks for generating highly realistic avatars in terms of appearance and behavior (e.g., OmniAvatar and WanS2V). In [24], the authors present a "Live Avatar" capable of speaking, reacting, and perceiving real-time context over extended durations. This approach employs an audio-driven avatar generation process based on a 14-billion-parameter diffusion model and a distributed inference paradigm that performs denoising computation across multiple GPUs. The system demonstrated end-to-end generation of 20 frames per second (FPS) using the computing power of five H800 GPUs. Although this avatar generation system runs rapidly, the latency for generating the first frame remains significant and requires further reduction.

In another work [25], the authors present a system that generates animations for photo-realistic 4D talking avatars. This framework transforms LLM-generated text input into corresponding audio-visual output assigned to the avatar. The proposed approach uses interconnected diffusion transformers to generate speech synchronized with lip movements, as well as facial and head gestures. The system presents "always-on" avatars that listen and react to the user's audio-visual input. In addition to active listening behaviors (e.g., back-channeling via nodding, smiling, eye contact, and mirroring expressions), the generated avatars deliver expressive reactions based on the transcribed text, introducing characteristics of empathetic dyadic interactions. The system is identity-specific, enabling the mod-

el to generate audio-visual animations for specific identities included in the training. For system execution, the authors utilized an A100 GPU, requiring approximately 200 ms to synthesize a 20-second audio-visual sequence with the interactive avatar.

3. Materials and methods

3.1. System Description

The ISSAI avatar is designed as a browser-based application integrating language models, speech processing technologies, lip sync, and 3D-stylized avatar. Two practical system configurations are developed that both use ISSAI's

AI models: 1) practical cloud-based ISSAI avatar configuration and 2) practical edge-deployed ISSAI avatar configuration (see Fig. 1). The cloud-based configuration uses ISSAI's LLM "Oylan" API (<https://issai.kz/documentation/oylanApi>) and TTS/STT services of ISSAI's "MangiSoz" platform designed for translation and speech processing applications (<https://mangisoz.nu.edu.kz/backend/api-docs/en>). The edge-deployed configuration uses ISSAI's "Qolda", a small vision-language model with 4.3 billion parameters and based on the InternVL3.5 architecture with a Qwen3-4B language backbone [17], fine-tuned Whisper Turbo ASR, and Matcha-TTS frameworks.

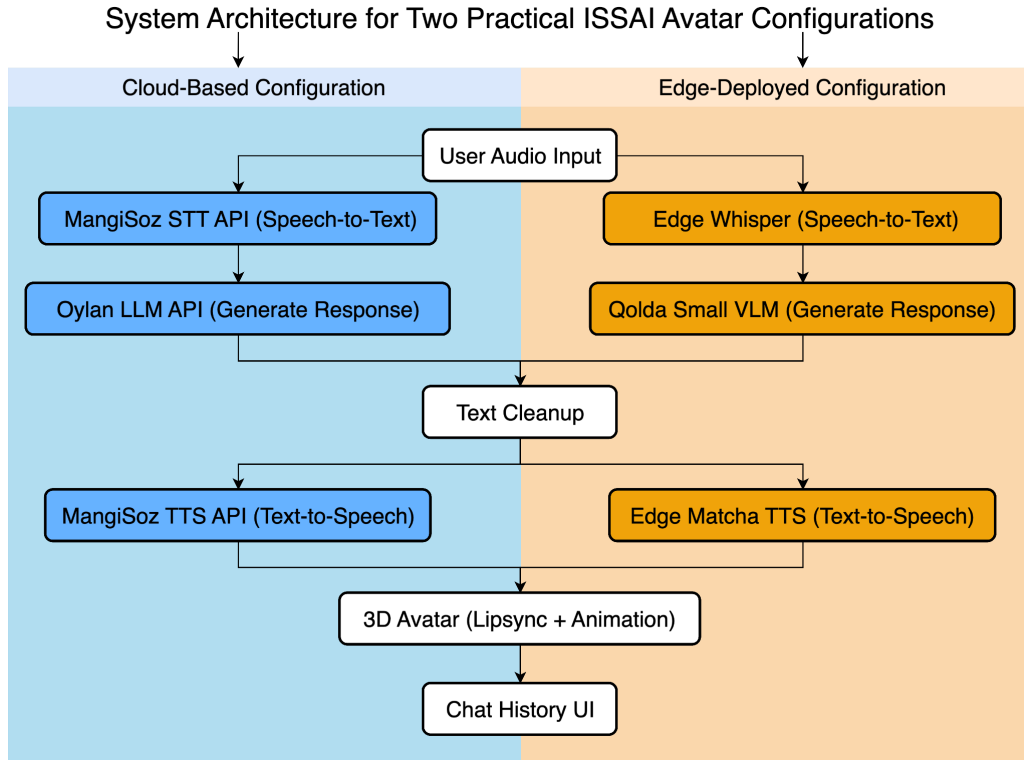


Figure 1. System architecture for cloud-based and edge-deployed ISSAI avatar configurations

The models were deployed as backend services using FastAPI and containerized with Docker. On the frontend, the system uses React 18, Vite, and Tailwind CSS. Users interact with the system via voice input. The avatar's synthesized speech is synchronized using lip-sync technology and delivered in the selected language. The system also displays the conversation history and the response times of the system components. The same hardware was

used for the frontend part of both cloud-based and edge-deployed ISSAI avatar configurations (i.e., ASUS ROG Zephyrus G16 laptop equipped with an NVIDIA RTX 4090 Mobile GPU, 16GB VRAM).

In the system's initial configuration screen (see Fig. 2), users can select the preferred language (Kazakh, Russian, or English), the avatar character, and the background environment (classroom, office, or outdoor scene). The main avatar interface displays

a 3D-stylized avatar within the selected office environment (see Fig. 3). A microphone button is located next to the text input field. When pressed, voice input is activated, allowing users to pose their queries to the avatar. Users can also type their queries directly into the text input field. The interface displays the conversation history between the user and the avatar, along with system response times. A performance metrics panel in the upper-right corner shows real-time latency measurements for system components.

3D-Stylized Avatar. For the avatar component, we utilized the Ready Player Me platform, which

supports web-based implementation. The avatars have a model size of 2.8 MB and load in less than 1 second. During testing on the ASUS laptop, rendering maintained a stable 60 FPS, with Central Processing Unit (CPU) usage ranging from $7 \pm 2.63\%$ (idle) to $12 \pm 2.63\%$ (active), while GPU usage for rendering remained $17 \pm 2.50\%$. Stylized 3D avatar rendering conserves GPU resources for AI inference, especially in an edge-deployed configuration. Facial animation frames are synchronized with audio playback via Web Audio API timestamps, with a latency of 50-100 ms between audio and visual output.

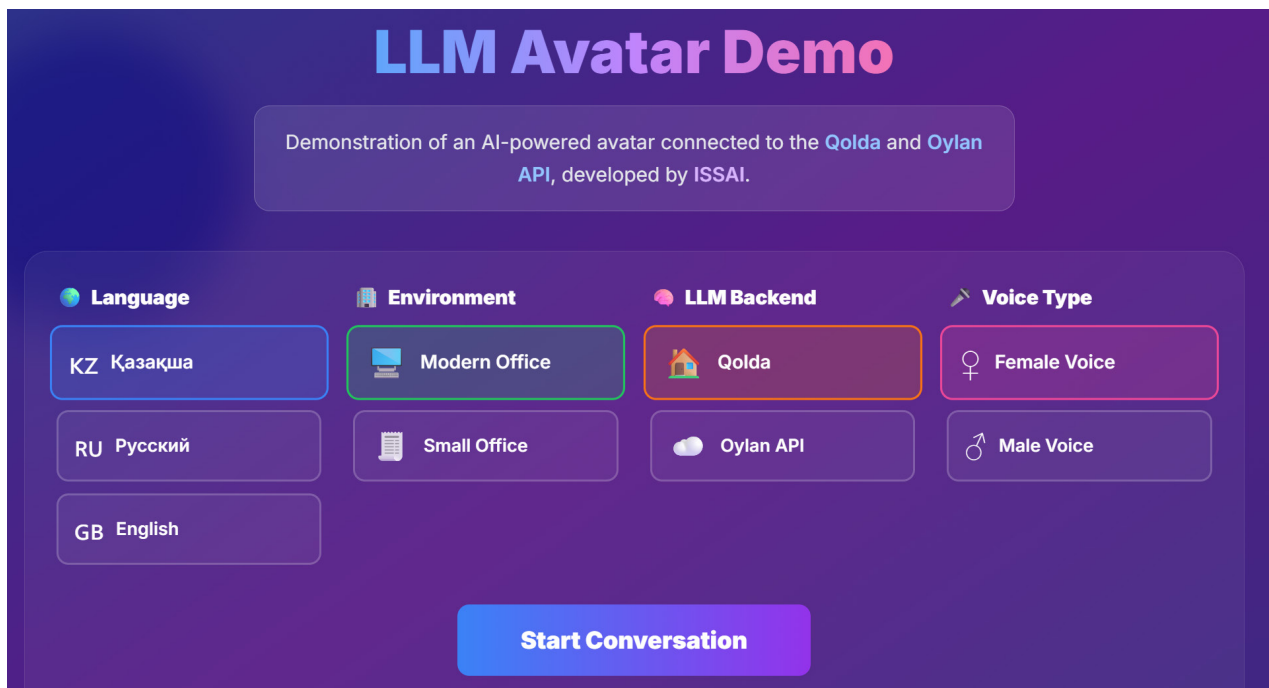


Figure 2. Initial configuration UI for selecting language, avatar, and environment

Speech Processing Pipeline. For speech processing, the system’s MediaRecorder captures audio at 48 kHz and converts the WebM input to WAV format (16 kHz mono PCM) with -3 dB normalization. In the cloud-based avatar configuration, speech processing is handled by MangiSoz API STT and TTS services available at <https://mangisoz.nu.edu.kz/backend/api-docs/en>. In the edge-deployed configuration, speech processes are handled by Whisper Turbo ASR and Matcha-TTS that operate directly on the laptop GPU alongside ISSAI’s Qolda. The

audio preprocessing steps on the edge-deployed avatar configuration include format conversion, 16 kHz resampling, normalization and transcription by Whisper Turbo ASR (note that Whisper Turbo requires 1.22 GB of VRAM). Matcha-TTS synthesizes audio on the edge and generates the phoneme timing required for the avatar’s lip synchronization. The Matcha-TTS model occupies 1.83 GB on VRAM. All used models have been fine-tuned by ISSAI to support the Kazakh, Russian, and English languages.



Figure 3. 3D-stylized avatar UI in the office environment communicating in Kazakh

Language Model. The cloud-based avatar configuration utilizes ISSAI’s Oylan API available at <https://issai.kz/documentation/oylanApi>. The edge-deployed avatar configuration employs ISSAI’s Qolda, a vision-language model based on the InternVL3.5 architecture with a Qwen3-4B language backbone [17]. Qolda comprises 4.3 billion parameters [17]. The model occupies 8.6 GB in FP16 GGUF format or 4.58 GB in INT8 GGUF format. In terms of performance, Qolda achieves an average accuracy of 71.64% on Kazakh text tasks across popular benchmarks, including MMLU, Winogrande, HellaSwag, ARC, GSM8K, and DROP [17]. Furthermore, it attains an accuracy of 60.44% on vision tasks such as AI2D, MMStar, RealWorldQA, and KazakhOCR (in “Think” mode) [17]. According to the authors, ISSAI’s Qolda shows a 19.64-point improvement over the base Qwen3-4B model on Kazakh benchmarks [17]. To address the resource constraints for the edge-deployed configuration, we investigated model quantization as a compression technique. During the design of the edge-deployed ISSAI avatar configuration, we explored two variants: Qolda-Full, operating at full FP16 precision (8.6 GB memory footprint), and Qolda-Quantized, utilizing INT8 quantization in GGUF format (4.58 GB memory footprint).

Inference in our avatar system is executed using Llama.cpp, which exposes an OpenAI-compatible

localhost endpoint. This architecture allows us to use identical client-side code for both Qolda and Oylan models, requiring only a change in the service URL. Llama.cpp manages model loading, KV-cache optimization, request queuing, and optional token streaming. Llama.cpp deployment with the GGUF format Qolda model was selected as the deployment method, rather than a safetensors format Qolda model deployed with LMDeploy, as reported in the original Qolda paper of ISSAI [17]. In our case, using LMDeploy caused the language model to consume more GPU VRAM. Memory-related limitations prevented us from running all AI components on a single consumer laptop GPU with 16 GB of VRAM when using LMDeploy. By using the Qolda model in FP16 GGUF format and the Llama.cpp deployment method, we were able to run all AI models on a consumer laptop with a single GPU and 16 GB of VRAM. Qolda’s `enable_thinking` parameter can activate internal chain-of-thought reasoning to enhance response quality while also increasing inference latency.

3.2. System Evaluation

To evaluate the performance of ISSAI avatar edge-deployed and cloud-based configurations, we designed a user study and conducted latency analysis comparing two practical configurations. The evaluation focused on three pipeline compo-

nents: STT, Language Model Inference, and TTS. We measured response times across 96 queries (i.e., 32 questions in Kazakh, 32 Russian, and English translations). Questions were selected to cover conversational, factual, reasoning, and practical tasks.

Example questions in three languages are shown in Table 1. The full set of questions in three languages used for the user study is available at the project’s GitHub repository https://github.com/IS2AI/multilingual_generative_AI_avatar.

Table 1. Example queries used for system evaluation in three languages

#	Kazakh	Russian	English
1	Сәлем! Қалың қалай?	Привет! Как дела?	Hello! How are you?
2	Егер менде 5 алма болса, екеуін берсем, нешеуі қалады?	Если у меня 5 яблок и я отдам 2, сколько останется?	If I have 5 apples and give away 2, how many are left?
3	Күн жүйесінде неше планета бар?	Сколько планет в Солнечной системе?	How many planets are in the solar system?
4	«Мектеп» сөзіне сөйлем құрастыршы.	Составь предложение со словом «школа».	Make a sentence with the word «school».

For the user study, we invited 7 participants (4 male and 3 female) from the community of students and researchers at Nazarbayev University (NU). The average age of the participants was 25.29 ± 3.95 years. Prior to the user studies, ethics approval was obtained from NU Institutional Research Ethics Committee (IREC). All participants were informed and provided consent. During the user study, participants were asked to read 32 queries to the avatar, composed of 8 questions in the conversational domain, 8 questions assessing factual knowledge, 8 questions with reasoning tasks, and 8 questions on general tasks. When the avatar returned an answer, participants rated the response on a scale from 1 to 5. A rating of 0 was assigned if the system failed to recognize the query, failed to respond, or hallucinated, as well as when a participant wished to repeat the query to receive a better response. Let us call this procedure an experimental task. The experimental task was performed in three languages. All 32 queries used in the experimental task were generated in Kazakh, then translated into Russian and English. Participants were divided into three groups, with each group completing the experimental task in all three languages but in different orders. For example, Group A participants completed the experimental task in the Kazakh language, then in Russian and English. Group B participants followed the order of Russian-English-Kazakh, while Group C participants completed the experimental task in the order of English-Kazakh-Russian.

After participants completed an experimental task in the first testing language, they were asked to

rate their cognitive state and evaluate the usability of the system. They then moved to the second language, followed by the same evaluation, and finally tested the system a third time in the third language. For the evaluation of cognitive load, we used the Unweighted/Raw NASA-Task Load Index (NASA-TLX), also referred to as RTLX [26], and for the usability assessment, we used the System Usability Scale (SUS) Questionnaire [27]. After completing the experimental task in each language, participants filled out both questionnaires before proceeding to the next language’s experimental task. The Raw NASA-TLX rating subjectively measures user-experienced workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration, with each scored from 0 to 100. SUS score evaluates the usability of the tested system in the range from 0 to 100.

The total end-to-end latency is calculated as a sum of the processing time for each component of the pipeline. For the edge-deployed avatar configuration, where all computations are deployed on the edge device, the latency (L_{edge}) is defined as:

$$L_{edge} = T_{STT} + T_{LLM} + T_{TTS} \quad (1)$$

On the other hand, the cloud-based avatar configuration latency (L_{cloud}) has additional network transmission delays (δ), which occur during the data transfer between MangiSoz and Oylan API, so the formula is defined as:

$$L_{edge} = L_{STT} + L_{LLM} + L_{TTS} \quad (1)$$

where

$$\begin{aligned} L_{STT} &= T_{STT} + \delta_1 \\ L_{LLM} &= T_{LLM} + \delta_2 \\ L_{TTS} &= T_{TTS} + \delta_3 \end{aligned}$$

In these formulas, T_{STT} , T_{LLM} , and T_{TTS} represent the raw model inference times, while δ_n denotes the network latency for the respective remote API calls. The values presented in Table 4 were calculated using the arithmetic mean (μ) and the sample standard deviation (σ):

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}} \quad (4)$$

The comparative efficiency of the edge-deployed ISSAI avatar configuration over the cloud-based ISSAI avatar configuration can be calculated using the Latency Reduction Efficiency (E) metric:

$$E = (1 - \frac{L_{edge}}{L_{cloud}}) * 100\% \quad (5)$$

4. Results

Table 2 shows the user ratings for the avatar's responses to the queries asked in the edge-deployed configuration tested during the user study. User ratings were averaged per language and per query category, with the specific query category and all categories together for each language; the means and standard deviations are reported in Table 2. The average rating for Kazakh queries was 4.18 ± 0.93 , slightly lower than those for Russian and English. The user ratings for the Russian and English queries were 4.38 ± 0.41 and 4.37 ± 0.88 , respectively. Users could also assign a value of '0' to ask the query again and get a better response. The average number of 0 ratings assigned by the 7 participants was $12.00 \pm$

3.15 for Kazakh, 6.29 ± 1.69 for Russian, and 7.86 ± 2.34 for English. Users 4 and 5 decided to ask some queries again, contributing more than 20 zero ratings in Kazakh. Analysis of daily language usage patterns revealed that both users use Kazakh 10% of the time, compared to Russian and English. The mean number of '0' ratings across the 32 queries per language is 37.5% for Kazakh, 19.7% for Russian, and 24.6% for English. The high retry rate of 37.5% for the Kazakh language could be attributed to limitations in both the speech recognition system and the language models generating responses in Kazakh.

Table 3 shows the cognitive load evaluation results. The mental demand rating for the Kazakh language testing of the system was 30.71 ± 25.13 , which is higher than the mental demand ratings for Russian and English, with corresponding values of 16.43 ± 11.25 and 15.71 ± 7.76 . Similarly, the physical demand rating for Kazakh is higher than for Russian and English, even though all reported values are in the range related to low physical demand. The rating for effort is higher for the Kazakh language than for Russian and English. The ratings for temporal demand and frustration are similar across all languages and have values corresponding to low levels of cognitive demand. The performance ratings for the Kazakh and English languages are in the same range, and higher than the performance for Russian. The corresponding visual representation of the cognitive load ratings is shown in Figure 4. Each participant in our user study completed the task in all three languages; the NASA-TLX ratings were analyzed using the non-parametric Friedman test [28]. The test revealed no statistically significant differences between languages for NASA-TLX cognitive load ratings (all $p > 0.05$). Specifically Friedman test applied on NASA-TLX ratings indicated the following results: mental demand [$\chi^2(2) = 2.55$, $p = 0.28$], physical demand [$\chi^2(2) = 0$, $p = 1$], temporal demand [$\chi^2(2) = 0.93$, $p = 0.627$], performance [$\chi^2(2) = 2.15$, $p = 0.341$], effort [$\chi^2(2) = 2.33$, $p = 0.311$], and frustration [$\chi^2(2) = 3.55$, $p = 0.17$].

Table 2. User ratings for each category queries (scale 1-5)

Category	Kazakh	Russian	English
Conversational	4.30±0.94	4.83±0.35	4.57±0.69
Logic & Mathematics	3.87±1.0	4.03±0.50	4.16±1.07
Factual Knowledge	4.41±0.84	4.07±0.63	4.35±0.84
Practical Tasks	4.13±0.93	4.60±0.16	4.39±0.91
Total	4.18±0.93	4.38±0.41	4.37±0.88

In addition to cognitive load feedback, 7 participants who tested the edge-deployed avatar configuration were also asked to rate the usability of the system using the SUS questionnaire. The results show that the highest SUS score, as rated by participants, is for English (90.71 ± 9.76). The SUS scores for Russian (88.93 ± 8.40) and Kazakh (86.79 ± 7.46)

are slightly lower than for English. Differences in SUS scores for three languages are not statistically significant (Friedman test showed $\chi^2(2) = 0.38$, $p = 0.827$). A SUS score of 80 or more is considered as high usability. As discussed by Sauro et al. [29], a SUS score above 80.3 suggests that the tested system will be recommended to friends.

Table 3. Raw NASA Task Load Index (RTLX) ratings (Mean $\pm$ STD)

Category	Kazakh	Russian	English
Mental Demand	30.71 $\pm$ 25.13	16.43 $\pm$ 11.25	15.71 $\pm$ 7.76
Physical Demand	20.0 $\pm$ 23.30	11.43 $\pm$ 7.89	11.43 $\pm$ 7.89
Temporal Demand	20.0 $\pm$ 23.30	15.0 $\pm$ 11.65	7.14 $\pm$ 15.08
Performance	33.43 $\pm$ 24.60	24.33 $\pm$ 25.27	34.31 $\pm$ 30.87
Effort	35.0 $\pm$ 20.53	21.43 $\pm$ 15.29	22.86 $\pm$ 15.32
Frustration	21.43 $\pm$ 24.01	17.86 $\pm$ 17.50	19.29 $\pm$ 16.13

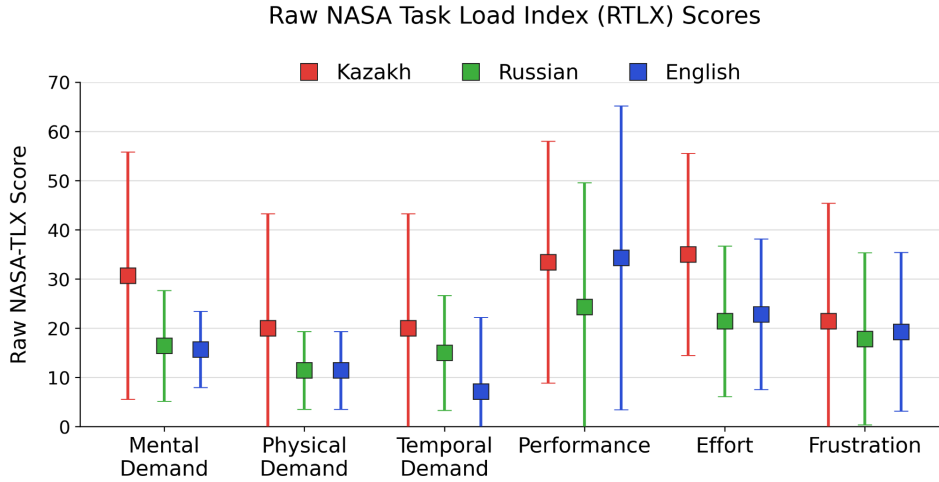


Figure 4. Raw NASA-TLX ratings for edge-deployed avatar configuration in Kazakh, Russian, and English languages

Table 4 summarizes the complete end-to-end pipeline latency analysis after the edge-deployed ISSAI avatar configuration was tested in a user study across three languages (32 queries for each language) by 7 participants and the cloud-based ISSAI avatar configuration was tested by one user. The cloud latencies were measured by sending the same set of queries from the client laptop to the hosted ISSAI services over a standard university network connection. The study did not systematically control or log network conditions, warm-up, server load, or

API-side caching during these measurements. Because cloud performance can vary with these factors, the cloud latency should be considered as an approximate baseline rather than a tightly controlled benchmark. The edge-deployed avatar configuration shows an average total response time of 2.28 s, while the cloud-based avatar configuration shows an average total response time of 4.18 s. The total response time of the edge-deployed configuration is 45.5% lower than that of the cloud-based configuration.

Speech-to-text Performance. The cloud-based STT deployment demonstrated better performance, reducing latency by 39.7% with an average of 0.38 s compared to the 0.63 s value of the edge-deployed avatar configuration. Such a difference could be caused by the computational advantage of the cloud servers for speech recognition (i.e., NVIDIA A100). Edge-deployed ISSAI avatar configuration might have computational constraints due to concurrent management of the STT, LLM, and TTS pipelines. However, the dedicated server-side resources used in the cloud-based ISSAI avatar configuration maintain a stable processing window of 0.33-0.44 s across all query categories. Ultimately, while the cloud-based avatar configuration performs better in specialized transcription speed, the edge-deployed avatar configuration achieves faster total system latency by optimizing for end-to-end response generation through its downstream LLM and TTS tasks. **Language Model Inference Performance.** Language model inference represents the most significant performance differentiator, with the edge-deployed avatar configuration achieving 60.6% faster inference (1.08 s average) than the cloud-based avatar configuration (2.74 s average). Performance gains are consistent across languages, with edge-deployed avatar configuration showing a 54.3% speedup in Kazakh (1.38 s vs 3.02 s), a 59.4% speedup in Russian (1.08 s vs 2.66 s), and a 69.4% speedup in English (0.78 s vs 2.55 s). Cloud-based avatar configuration inference latency varies significantly across query categories (ranging from 2.08 s to 3.48 s). On the edge-deployed avatar configuration, simple factual queries are processed in 0.34-1.01 s, while reasoning-intensive tasks take up to 2.37 s.

Text-to-Speech Performance. While the edge-deployed ISSAI avatar configuration achieves 43.4% lower average TTS latency (0.60 s edge-deployed vs. 1.06 s cloud-based configuration), their consistency across workloads differs significantly. Such performance remains stable across languages. Edge-deployed avatar configuration maintains faster synthesis for Kazakh (0.61 s vs 1.11 s), Russian (0.59 s vs 1.02 s), and English (0.61 s vs 1.05 s). The cloud-based avatar configuration TTS exhibits significant variability: standard responses range from 0.97 s to 1.31 s, due to the cumulative overhead of network transmission and downloading audio pay-

loads from remote servers. Edge-deployed ISSAI avatar configuration does not have these delays, maintaining stable synthesis times of 0.57-0.65 s regardless of query category. For complex or longer outputs, this results in nearly 2 times faster delivery of response on the edge-deployed ISSAI avatar configuration.

Each component value in Table 4 is a rounded mean, while each reported «Total» is computed from the raw per-query sums. Small differences of at most 0.11 s between a row’s listed components and the value reported in “Total” might come from the applied rounding. The “Total” values reported in Table 4 are measured directly with an end-to-end timer (input to final output), not by adding the STT, LLM, and TTS means. Because the end-to-end timer and the per-stage timers measure latency in slightly different ways, the reported “Total” can differ slightly from the sum of the three stages. On the edge, this is mainly because model-loading overhead is included in each stage’s timer but incurred only once across the full pipeline; on the cloud, where services stay warm, the difference is negligible. Table A in the supplementary material shows the average component and query category latency comparison, that is an extended version of Table 4.

GPU and VRAM Utilization. Performance evaluation on a consumer laptop with GPU (ASUS ROG Zephyrus G16) reveals that AI inference is the primary driver of resource consumption, while 3D avatar rendering incurs minimal overhead. Among the AI components, the LLM is the largest contributor to VRAM footprint at 9.07 GB during FP16 inference, with the KV cache for conversation history accounting for an additional 1.09 GB. VRAM values denote peak usage during FP16 inference. The 9.07 GB for language model inference comprises the FP16 model weights (about 8.6 GB) plus activation buffers; the KV cache and runtime for conversation history is listed separately (1.09 GB). As shown in Table 5, the Ready Player Me avatar maintains a stable 60 FPS while consuming only 17.0% GPU and <0.10 GB VRAM, leaving the vast majority of compute and memory resources available for the AI models. TTS engine records the highest peak GPU load at 79.4%, this represents a short-term processing spike during synthesis rather than sustained utilization.

Table 4. Average component latency comparison across 32 queries for 3 languages (Mean $\pm$ STD)

Language	STT		LLM		TTS		Total	
	Edge	Cloud	Edge	Cloud	Edge	Cloud	Edge	Cloud
KZ	0.67 $\pm$ 0.09	0.39 $\pm$ 0.04	1.38 $\pm$ 1.35	3.02 $\pm$ 1.98	0.61 $\pm$ 0.10	1.11 $\pm$ 0.29	2.63 $\pm$ 1.48	4.51 $\pm$ 2.31
RU	0.60 $\pm$ 0.07	0.35 $\pm$ 0.03	1.08 $\pm$ 0.93	2.66 $\pm$ 1.00	0.59 $\pm$ 0.08	1.02 $\pm$ 0.10	2.23 $\pm$ 1.04	4.03 $\pm$ 1.12
EN	0.61 $\pm$ 0.09	0.41 $\pm$ 0.09	0.78 $\pm$ 0.76	2.55 $\pm$ 1.38	0.61 $\pm$ 0.10	1.05 $\pm$ 0.18	2.00 $\pm$ 0.93	4.00 $\pm$ 1.65
All	0.63$\pm$0.09	0.38$\pm$0.05	1.08$\pm$1.01	2.74$\pm$1.46	0.60$\pm$0.09	1.06$\pm$0.19	2.28$\pm$1.15	4.18$\pm$1.70

Table 5. GPU and VRAM utilization

Component	GPU Utilization (%)	VRAM Usage (GB)
Avatar Rendering	17.0 $\pm$ 2.50	<0.10
ASR (Speech-to-Text)	46.1 $\pm$ 35.1	1.22
TTS (Text-to-Speech)	79.4 $\pm$ 5.23	1.83
Language Model Inference	74.3 $\pm$ 24.7	9.07
KV Cache & Runtime	-	1.09

5. Discussion

Modern avatar systems use different deployment configurations that could further affect their accessibility, performance, and usability across various applications. In this work, we developed and evaluated two practical ISSAI avatar approaches: cloud-based avatar configuration and edge-deployed avatar configuration. The cloud-based avatar configuration offers several distinct advantages. It imposes minimal computational requirements on client devices, enables access to significantly larger language models than those feasible for the edge-deployed configurations, and simplifies system maintenance. However, cloud-based ISSAI avatar configuration has inherent network latency in the interaction pipeline. Data transmission processes increase response latency of the cloud-based avatar configuration. Furthermore, cloud-based configuration requires a stable internet connection, recurring API costs that scale with the usage, and raises concerns regarding user data privacy. On the other hand, the edge-deployed ISSAI avatar configuration runs all computational components directly on the user device, resulting in reduced response latency. Furthermore, edge-deployed ISSAI avatar configuration supports offline interaction with the avatar. All user interactions, including speech audio and conversation history, remain entirely on-device, addressing data privacy

concerns. From an economic perspective, the edge-deployed ISSAI avatar configuration requires one-time hardware investment, reducing recurring API expenses.

Proposed edge-deployed ISSAI avatar configuration has reduced end-to-end response time (2.28 s) to 45.5% in comparison to cloud-based ISSAI avatar configuration (4.18 s). The 60.6% reduction in LLM inference latency (1.08 s vs. 2.74 s) represents the most substantial performance gain. The total response time of 2.28s of edge-deployed ISSAI avatar configuration shows that using a 3D rendering platform for 3D-stylized avatar (Ready Player Me) with optimized SLMs (Qolda 4.3 billion parameters) can offer an optimal balance between visual fidelity and conversational responsiveness.

By compressing Qolda’s memory footprint from 8.6 GB (FP16 GGUF) to 4.58 GB (INT8 GGUF), the edge-deployed configuration of ISSAI avatar can run on a consumer laptop with 8 GB GPU (e.g., NVIDIA RTX 4060 or 3060). In this study we used 16 GB GPU (i.e NVIDIA RTX 4090) and original Qolda in FP16 GGUF format. Our analysis shows that 3D-stylized avatar rendering consumes around $17.0 \pm 2.5\%$ of GPU resources, leaving $83.0 \pm 2.5\%$ dedicated to concurrent AI model inference. For standard text-based interactions, we deployed Qolda with Llama.cpp. This allowed all pipeline components (Whisper ASR, Qolda LLM, and Matcha

TTS) to run on the edge device reducing memory footprint and improving real-time performance of edge-deployed configuration of ISSAI avatar.

5.1. Limitations

Our study explored two practical configurations for the ISSAI avatar (i.e., cloud-based configuration and edge-deployed configuration). It has several limitations:

- **Hardware requirements.** The edge-deployed ISSAI avatar configuration requires dedicated hardware to support the operation of the edge-deployed language model, speech processing tools, and the visual part of the avatar system. Specifically, a GPU equivalent to the NVIDIA RTX 4090 Mobile or, at a minimum, a card with 16 GB VRAM is needed to run the original Qolda in FP16 GGUF format, satisfying the full FP16 requirement. For running the system with quantized Qolda satisfying the quantized INT8 requirement, a GPU equivalent to NVIDIA RTX 4060 or 3060, or at a minimum, a card with 8 GB VRAM, is needed.

- **Model size trade-offs.** While ISSAI's Qolda with 4.3 billion parameters achieves 71.64% accuracy on Kazakh text benchmarks, it has constraints on supporting broader knowledge and has limited reasoning capabilities compared to cloud-hosted models.

- **Visual realism.** The current system employs a stylized 3D avatar system rather than a more photo-realistic representation. The effects of avatar visual representation on user experience should be further explored through user studies.

- **Query dataset scope.** Our evaluation involved 96 queries in three languages (Kazakh, Russian, and English), covering topics that addressed conversational, factual, and reasoning tasks for the avatar. The query dataset should be expanded for more generalizable results.

- **User study design limitations.** The performed user study involved 7 participants (i.e., 7 participants tested the edge-deployed ISSAI avatar configuration and 1 participant tested the cloud-based ISSAI avatar configuration under the same conditions). Because the cloud-based ISSAI avatar configuration response time is almost twice that of the edge-deployed configuration, the cloud system was not offered to all 7 participants and was tested by only 1 participant. System response times and latencies significantly affect the perception of such conversational systems and can impact human cognition. Therefore, the decision was made to test the cloud-based avatar configuration only

once. On the other hand, when the edge-deployed avatar configuration was evaluated by 7 participants and the cloud-based avatar configuration by only 1 participant, the comparison could be underpowered and potentially confounded. Ideally, the evaluation procedure should have been repeated under paired conditions, including the same queries, the same speaker set, repeated runs, the same prompt settings, the same output-length constraints, and have paired statistical comparisons.

- **Limited participants pool.** In the performed user study, the pool of participants was limited to 7 participants, which creates a small sample size for the evaluation results. Participants are all residents of NU (i.e., researchers and students). This means a small set of people evaluated the avatar system, and, for more generalizable outcomes, the user study should involve people from different backgrounds and occupations. Language-proficiency variation should also be considered during the evaluation of the conversational systems involving speech processing technologies.

- **Other limitations.** The study has cloud benchmark limitations, as well as language model mismatch between cloud-based and edge-deployed ISSAI avatar configurations. The study involved a controlled noise environment in which system evaluations were conducted under laboratory conditions with limited background noise.

5.2. Directions for Future Research

There are several directions for future research of AI-enhanced avatar systems:

- **Functional specialization.** Future research can adapt AI avatars for specific roles (e.g., technical support or language learning) by fine-tuning SLMs on domain-specific corpora. User studies involving human participants could be designed to measure usability and performance of such systems in real-world scenarios.

- **Multimodality and language support.** Integrating computer vision tools would enable avatars to interpret gestures and facial expressions for richer interactions. Language support can be extended to other languages to further promote multilingual interaction with AI-enhanced avatar systems.

- **Language mode optimization.** Future work can also explore model compression techniques beyond INT8 quantization, such as pruning, knowledge distillation, and 4-bit quantization. Such optimization could lower hardware barriers enough for mobile device deployment of AI-enhanced interactive systems.

- Latency reduction. Future research works can also investigate the trade-off between response quality and the latency introduced by chain-of-thought reasoning when the thinking parameter is enabled in language models. Such investigations could drive the development of dynamic systems capable of adjusting their inference strategies based on query complexity.

6. Conclusions

In this paper, we presented and evaluated two ISSAI avatar system configurations using different models, runtimes, and server conditions: 1) cloud-based avatar configuration and 2) edge-deployed avatar configuration. Our evaluation involved an experimental task in which the participants in the user study asked 96 queries to the system in three languages (32 queries for each language). The results show that the end-to-end response latency (2.28 s) of the edge-deployed avatar configuration is 45.5% lower than that of the cloud-based avatar configuration (4.18 s). This improvement was consistent across all tested languages, with total latency reductions of 41.7% in Kazakh (2.63 s vs. 4.51 s), 44.7% in Russian (2.23 s vs. 4.03 s), and 50.0% in English (2.00 s vs. 4.00 s). These gains were driven by 60.6% faster language model inference and 43.4% faster TTS, even though cloud-based configuration's automatic speech recognition performance was 39.7% faster than the edge-deployed configuration's speech-to-text performance.

For the edge-deployed avatar configuration, we used ISSAI's Qolda in FP16 GGUF format taking 8.6 GB of VRAM (16 GB) and deployed it in Llama.cpp. This enabled the simultaneous execution of three AI models, including Qolda, Whisper Turbo, and Matcha-TTS, alongside 3D stylized avatar rendering on the consumer laptop with a single GPU. Because the avatar rendering consumed $17 \pm 2.5\%$

of GPU resources at a stable 60 FPS, $83 \pm 2.5\%$ of computational capacity remained fully available for AI inference. In our work, we showed that edge-deployed configurations of multilingual avatar systems are both technically feasible and have practical advantages. The combination of small language models, quantization techniques, and efficient 3D rendering creates an opportunity to build avatar systems that support real-time offline interactions. The implementation procedure, set of questions used in user study, and a demo video with a human asking queries from an edge-deployed ISSAI avatar are publicly available at the project's GitHub repository https://github.com/IS2AI/multilingual_generative_AI_avatar.

Acknowledgments

We extend our gratitude to the supervisor of the research grant, Professor Madina Mansurova, for her coordination and support.

Funding

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, A.A., G.A., Z.M. and H.A.V.; *Methodology*, A.A., G.A. and Z.M.; *Software*, G.A. and A.A.; *Validation*, G.A., A.A. and Z.M.; *Formal Analysis*, Z.M. and H.A.V.; *Investigation*, G.A. and A.A.; *Resources*, G.A. and A.A.; *Data Curation*, G.N. and K.S.; *Writing – Original Draft Preparation*, A.A., G.A., G.N. and Z.M.; *Writing – Review & Editing*, Z.M. and H.A.V.; *Visualization*, G.A. and A.A.; *Supervision*, H.A.V.; *Project Administration*, H.A.V.; *Funding Acquisition*, H.A.V.

References

1. S. Amershi et al., "Guidelines for human-AI interaction," in *Proc. 2019 CHI Conf. Human Factors Comput. Syst. (CHI)*, Glasgow, U.K., 2019, pp. 1–13, doi: 10.1145/3290605.3300233.
2. Y. Liu and K. L. Siau, "Human-AI interaction and AI avatars," in *HCI International 2023 — Late Breaking Papers* (Lecture Notes in Computer Science, vol. 14059), H. Degen, S. Ntoa, and A. Moallem, Eds. Cham, Switzerland: Springer, 2023, pp. 120–130, doi: 10.1007/978-3-031-48057-7_8.
3. R. K. Ghamandi et al., «Unlocking understanding: An investigation of multimodal communication in virtual reality collaboration,» in *Proc. 2024 CHI Conf. Human Factors Comput. Syst. (CHI)*, Honolulu, HI, USA, 2024, pp. 1–16, doi: 10.1145/3613904.3642491.

4. W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, «Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI,» *Int. J. Hum.-Comput. Interact.*, vol. 39, no. 3, pp. 494–518, 2023, doi: 10.1080/10447318.2022.2041900.
5. E. Luger and A. Sellen, «‘Like having a really bad PA’: The gulf between user expectation and experience of conversational agents,» in *Proc. 2016 CHI Conf. Human Factors Comput. Syst. (CHI)*, San Jose, CA, USA, 2016, pp. 5286–5297, doi: 10.1145/2858036.2858288.
6. R. S. Wallace, «The anatomy of A.L.I.C.E.,» in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber, Eds. Dordrecht, The Netherlands: Springer, 2009, pp. 181–210, doi: 10.1007/978-1-4020-6710-5_13.
7. A. Vaswani et al., «Attention is all you need,» in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.
8. J. Achiam et al., «GPT-4 technical report,» 2023, arXiv:2303.08774, doi: 10.48550/arXiv.2303.08774.
9. J. Cassell, «Embodied conversational agents: Representation and intelligence in user interfaces,» *AI Mag.*, vol. 22, no. 4, 2001, Art. no. 67, doi: 10.1609/aimag.v22i4.1593.
10. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, «wav2vec 2.0: A framework for self-supervised learning of speech representations,» in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12449–12460.
11. G. Molinari and F. Ciravegna, «Towards pervasive distributed agentic generative AI — A state of the art,» *ACM Comput. Surv.*, 2025, doi: 10.1145/3821566.
12. H. Wan et al., «Building LLM-based AI agents in social virtual reality,» in *Ext. Abstracts 2024 CHI Conf. Human Factors Comput. Syst. (CHI EA)*, Honolulu, HI, USA, 2024, pp. 1–7, doi: 10.1145/3613905.3651026.
13. Y. Song and W. Xiong, «Large language model-driven 3D hyper-realistic interactive intelligent digital human system,» *Sensors*, vol. 25, no. 6, 2025, Art. no. 1855, doi: 10.3390/s25061855.
14. E. Altarawneh, M. Jenkin, and I. S. MacKenzie, «An extensible cloud based avatar: Implementation and evaluation,» in *Recent Advances in Technologies for Inclusive Well-Being: Virtual Patients, Gamification and Simulation* (Intelligent Systems Reference Library, vol. 196). Cham, Switzerland: Springer, 2021, pp. 503–522, doi: 10.1007/978-3-030-59608-8_27.
15. H. Kim et al., «ProsocialDialog: A prosocial backbone for conversational agents,» in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Abu Dhabi, UAE, 2022, pp. 4005–4029, doi: 10.18653/v1/2022.emnlp-main.267.
16. M. Maslych, C. Pumarada, A. Ghasemaghahi, and J. J. LaViola Jr., «Takeaways from applying LLM capabilities to multiple conversational avatars in a VR pilot study,» 2025, arXiv:2501.00168, doi: 10.48550/arXiv.2501.00168.
17. B. Arystanbekov, A. Nurimanov, A. Maxutov, V. Albrekht, A. Kuzdeuov, and H. A. Varol, «Qolda: A Small Vision–Language Model for the Kazakh Language,» *IEEE Access*, vol. 14, pp. 46392–46414, 2026, doi: 10.1109/ACCESS.2026.3676919.
18. O. Allal-Chérif, R. Puertas, and P. Carracedo, «Intelligent influencer marketing: How AI-powered virtual influencers outperform human influencers,» *Technol. Forecast. Soc. Change*, vol. 200, 2024, Art. no. 123113, doi: 10.1016/j.techfore.2023.123113.
19. Y. Wang, L. Wang, and K. L. Siau, «Human-centered interaction in virtual worlds: A new era of generative artificial intelligence and metaverse,» *Int. J. Hum.-Comput. Interact.*, vol. 41, no. 2, pp. 1459–1501, 2025, doi: 10.1080/10447318.2024.2316376.
20. G. Stefanek and J. DeMuth, «An avatar-based framework using large language models for assessing patient cognitive state,» *Issues Inf. Syst.*, vol. 25, no. 2, pp. 190–202, 2024, doi: 10.48009/2_iis_2024_115.
21. F. Salvetti, B. Bertagni, I. Contardo, R. Gardner, J. Rudolph, and R. Minehart, «Feedback reimaged: Generative AI and conversational avatars as your new training partners,» in *Proc. Learning Ideas Conf. (Lecture Notes in Networks and Systems)*. Cham, Switzerland: Springer, 2024, pp. 133–144, doi: 10.1007/978-3-031-73427-4_13.
22. M. C. Fink, S. A. Robinson, and B. Ertl, «AI-based avatars are changing the way we learn and teach: Benefits and challenges,» *Front. Educ.*, vol. 9, 2024, Art. no. 1416307, doi: 10.3389/educ.2024.1416307.
23. M. Seymour, K. Riemer, and J. Kay, «Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence,» *J. Assoc. Inf. Syst.*, vol. 19, no. 10, 2018, Art. no. 4. Accessed: Sep. 19, 2026. [Online]. Available: <https://aisel.aisnet.org/jais/vol19/iss10/4>
24. Y. Huang et al., «Live Avatar: Streaming real-time audio-driven avatar generation with infinite length,» in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2026. [Preprint]. Available: arXiv:2512.04677, doi: 10.48550/arXiv.2512.04677.
25. A. Chatziagapi, L.-P. Morency, H. Gong, M. Zollhöfer, D. Samaras, and A. Richard, «AV-Flow: Transforming text to audio-visual human-like interactions,» 2025, arXiv:2502.13133, doi: 10.48550/arXiv.2502.13133.
26. S. G. Hart and L. E. Staveland, «Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,» in *Human Mental Workload* (Advances in Psychology, vol. 52), P. A. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: North-Holland, 1988, pp. 139–183, doi: 10.1016/S0166-4115(08)62386-9.
27. J. Brooke, «SUS: A ‘quick and dirty’ usability scale,» in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.
28. T. J. Cleophas and A. H. Zwinderman, «Non-parametric tests for three or more samples (Friedman and Kruskal-Wallis),» in *Clinical Data Analysis on a Pocket Calculator*, 2nd ed. Cham, Switzerland: Springer, 2016, pp. 193–197, doi: 10.1007/978-3-319-27104-0_34.
29. J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

Information about the authors:

Auyeskhan Alibekov is a Junior Data Scientist at the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University (NU) (Astana, Kazakhstan, auyeskhan.alibekov@nu.edu.kz). Auyeskhan Alibekov holds a B.S. degree in Computer Science from Astana IT University, Astana, Kazakhstan. ORCID ID: 0009-0005-7551-668X.

Galammadin Askar is a Master's student at Nazarbayev University (NU) and a Research Assistant at the Institute of Smart Systems and Artificial Intelligence (ISSAI) at NU (Astana, Kazakhstan, galammadin.askar@nu.edu.kz). Galammadin Askar holds a B.S. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. ORCID ID: 0009-0001-6304-3003.

Zhanat Makhataeva, PhD, is a Senior Data Scientist at the Institute of Smart Systems and Artificial Intelligence (ISSAI) of Nazarbayev University (NU) (Astana, Kazakhstan, zhanat.makhataeva@nu.edu.kz). Dr. Zhanat Makhataeva received a B.S. degree in Robotics and Mechatronics from Nazarbayev University (NU), Astana, Kazakhstan, in 2017, and the M.S. and Ph.D. degrees in Robotics Engineering from the School of Engineering and Digital Sciences at NU, in 2019 and 2023, respectively. During her doctoral studies, Zhanat became a Microsoft Research PhD Fellow 2021-2022 and had a 12-week Internship at Microsoft Research Cambridge, UK Laboratory. She was hosted by the Future of Work Research Group at Microsoft Research and presented their work at the CHI 2023 Conference in Hamburg, Germany. Her research interests include augmented reality, cognitive assistive systems, artificial intelligence, human memory augmentation, cognitive ergonomics, system testing and evaluation, and human information processing. ORCID ID: 0000-0001-9366-7047.

Guldana Nauryzbaikyzy, PhD, is a Senior Lecturer at the Faculty of Foreign Languages of K. Zhubanov Aktobe Regional University (Aktobe, Kazakhstan, gnauryzbaikyzy@zhubanov.edu.kz). Dr. Guldana Nauryzbaikyzy received the B.S. and M.S. degrees in Arts with a specialization in Foreign Philology from L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, in 2017 and 2020. In 2026, Dr. Guldana Nauryzbaikyzy received her PhD in Philology from K. Zhubanov Aktobe Regional University. Her research interests include linguacultural studies, cognitive linguistics, and discourse studies. ORCID ID: 0000-0001-9272-8952.

Kulzat Sadirova, Doctor of Philological Sciences, is a professor at the Department of Theoretical and Applied Linguistics, Faculty of Philology, at K. Zhubanov Aktobe Regional University (Aktobe, Kazakhstan, ksadirova@zhubanov.edu.kz). She received her degree in Kazakh Language and Literature from K. Zhubanov Aktobe Pedagogical Institute in 1993. She obtained her Candidate of Sciences degree in Philology in 1997 and her Doctor of Sciences degree in Philology in 2009 from the A. Baitursynuly Institute of Linguistics. She currently serves as a professor and Chair of the Dissertation Council for the PhD program "8D02304 – Philology" at K. Zhubanov Aktobe Regional University. Her research interests include linguistics, discourse studies, syntax, and turkology. She has published nearly 200 scientific and methodological works, including articles indexed in Scopus and Web of Science, as well as monographs and textbooks. She has supervised 5 PhD graduates. She was the principal investigator of the grant-funded research project "Onomastic Discourse: Mythological Aspect" (2021–2023), supported by the Science Committee of the Ministry of Education and Science of Kazakhstan, and contributed as a linguist and editor to the National Corpus of the Kazakh Language at the Til-Kazyna National Scientific and Practical Center. She has received several awards, including the State Scientific Scholarship for Leading Scientists (2024), "Best Researcher – 2024," and the title "Best University Teacher of the Republic of Kazakhstan – 2012". ORCID ID: 0000-0001-6092-8191.

Huseyin Atakan Varol, PhD, is a full professor of robotics at Nazarbayev University's School of Engineering and Digital Sciences (Astana, Kazakhstan, ahvarol@nu.edu.kz). He received his BS degree in mechatronics engineering from Sabanci University, Istanbul, Turkey, in 2005. He received the MS and PhD degrees both in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2007 and 2009, respectively. From August 2009 to August 2011, he was first a postdoctoral research associate and then a research assistant professor with the Center for Intelligent Mechatronics, Department of Mechanical Engineering, Vanderbilt University, Nashville, TN, USA. In 2011, he joined the faculty of Nazarbayev University, Astana, Kazakhstan, where he is a full professor of robotics and directs the Institute of Smart Systems and Artificial Intelligence (ISSAI). His research interests include artificial intelligence, soft robotics, machine learning, sensor fusion and tensegrity. He has published more than 160 technical papers on these topics in reputable international journals and conferences. He holds multiple international patents in the area of bionics. He has attracted multiple competitively-funded research grants from the Ministry of Education and Science of Kazakhstan, World Bank and Nazarbayev University Research Fund. ORCID ID: 0000-0002-4042-425X.

Submission received: 23 April, 2026.

Revised: 31 July, 2026.

Accepted: 31 July, 2026.

Supplementary materials

Table A. Average component and query category latency comparison (Mean $\pm$ STD).

Category & Language		STT		LLM		TTS		Total	
		Edge	Cloud	Edge	Cloud	Edge	Cloud	Edge	Cloud
Conversational	KZ	0.67 $\pm$ 0.09	0.37 $\pm$ 0.05	2.18 $\pm$ 1.87	3.48 $\pm$ 2.64	0.62 $\pm$ 0.12	1.06 $\pm$ 0.20	3.49 $\pm$ 2.07	4.92 $\pm$ 2.88
	RU	0.58 $\pm$ 0.06	0.33 $\pm$ 0.03	2.37 $\pm$ 1.84	2.10 $\pm$ 0.61	0.62 $\pm$ 0.14	1.00 $\pm$ 0.05	3.46 $\pm$ 1.93	3.42 $\pm$ 0.68
	EN	0.60 $\pm$ 0.10	0.37 $\pm$ 0.09	1.70 $\pm$ 1.85	2.72 $\pm$ 1.51	0.65 $\pm$ 0.15	1.02 $\pm$ 0.10	2.94 $\pm$ 2.09	4.10 $\pm$ 1.69
Logic & Mathematics	KZ	0.68 $\pm$ 0.08	0.42 $\pm$ 0.06	0.43 $\pm$ 0.32	3.15 $\pm$ 2.59	0.57 $\pm$ 0.06	1.31 $\pm$ 0.59	1.65 $\pm$ 0.40	4.88 $\pm$ 3.23
	RU	0.66 $\pm$ 0.10	0.36 $\pm$ 0.04	0.47 $\pm$ 0.62	2.15 $\pm$ 0.74	0.57 $\pm$ 0.06	1.01 $\pm$ 0.04	1.64 $\pm$ 0.72	3.52 $\pm$ 0.80
	EN	0.64 $\pm$ 0.10	0.43 $\pm$ 0.09	0.45 $\pm$ 0.61	2.16 $\pm$ 1.43	0.57 $\pm$ 0.07	1.00 $\pm$ 0.09	1.65 $\pm$ 0.76	3.58 $\pm$ 1.61
Factual Knowledge	KZ	0.66 $\pm$ 0.10	0.40 $\pm$ 0.04	1.01 $\pm$ 0.98	2.08 $\pm$ 0.76	0.59 $\pm$ 0.06	1.05 $\pm$ 0.26	2.22 $\pm$ 1.11	3.53 $\pm$ 1.06
	RU	0.60 $\pm$ 0.08	0.34 $\pm$ 0.01	0.57 $\pm$ 0.43	3.37 $\pm$ 1.48	0.57 $\pm$ 0.06	1.09 $\pm$ 0.25	1.70 $\pm$ 0.51	4.80 $\pm$ 1.72
	EN	0.63 $\pm$ 0.08	0.44 $\pm$ 0.08	0.34 $\pm$ 0.09	2.83 $\pm$ 1.42	0.59 $\pm$ 0.05	1.20 $\pm$ 0.47	1.56 $\pm$ 0.17	4.48 $\pm$ 1.97
Practical Tasks	KZ	0.66 $\pm$ 0.10	0.36 $\pm$ 0.03	1.89 $\pm$ 2.22	3.35 $\pm$ 1.94	0.65 $\pm$ 0.16	1.01 $\pm$ 0.12	3.14 $\pm$ 2.34	4.72 $\pm$ 2.09
	RU	0.58 $\pm$ 0.05	0.36 $\pm$ 0.04	0.92 $\pm$ 0.84	3.02 $\pm$ 1.18	0.60 $\pm$ 0.07	0.98 $\pm$ 0.05	2.11 $\pm$ 0.98	4.37 $\pm$ 1.27
	EN	0.59 $\pm$ 0.09	0.42 $\pm$ 0.09	0.62 $\pm$ 0.51	2.47 $\pm$ 1.17	0.64 $\pm$ 0.12	0.97 $\pm$ 0.08	1.86 $\pm$ 0.70	3.86 $\pm$ 1.35
Overall	KZ	0.67 $\pm$ 0.09	0.39 $\pm$ 0.04	1.38 $\pm$ 1.35	3.02 $\pm$ 1.98	0.61 $\pm$ 0.10	1.11 $\pm$ 0.29	2.63 $\pm$ 1.48	4.51 $\pm$ 2.31
	RU	0.60 $\pm$ 0.07	0.35 $\pm$ 0.03	1.08 $\pm$ 0.93	2.66 $\pm$ 1.00	0.59 $\pm$ 0.08	1.02 $\pm$ 0.10	2.23 $\pm$ 1.04	4.03 $\pm$ 1.12
	EN	0.61 $\pm$ 0.09	0.41 $\pm$ 0.09	0.78 $\pm$ 0.76	2.55 $\pm$ 1.38	0.61 $\pm$ 0.10	1.05 $\pm$ 0.18	2.00 $\pm$ 0.93	4.00 $\pm$ 1.65
	-	0.63$\pm$0.09	0.38$\pm$0.05	1.08$\pm$1.01	2.74$\pm$1.46	0.60$\pm$0.09	1.06$\pm$0.19	2.28$\pm$1.15	4.18$\pm$1.70