LLMs Enable Context-Aware Augmented Reality in Surgical Navigation

Hamraz Javaheri
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
hamraz.javaheri@dfki.de

Omid Ghamarnejad
Allgemein-, Viszeral und Thoraxchirurgie, Chirurgische Onkologie, Klinikum Saarbrücken
Saarbrücken, Germany
omid.ghd@gmail.com

Paul Lukowicz
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany
paul.lukowicz@dfki.de

Gregor Alexander Stavrou
FEBS (HPB/SURGONC)
Klinikum Saarbrücken gGmbH
Allgemein-, Viszeral und Thoraxchirurgie, Chirurgische Onkologie
Saarbrücken, Germany
gstavrou@klinikum-saarbruecken.de

Jakob Karolus
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany
jakob.karolus@dfki.de

Abstract

Wearable Augmented Reality (AR) technologies are gaining recognition for their potential to transform surgical navigation systems. As these technologies evolve, selecting the right interaction method to control the system becomes crucial. Our work introduces a voice user interface (VUI) for surgical AR assistance systems (ARAS), designed for pancreatic surgery, that integrates Large Language Models (LLMs). Employing a mixed-method research approach, we assessed the usability of our LLM-based design in both simulated surgical tasks and during pancreatic surgeries, comparing its performance against conventional VUI for surgical ARAS using speech commands. Our findings demonstrated the usability of our proposed LLM-based VUI, yielding a significantly lower task completion time and cognitive workload compared to speech commands. Additionally, qualitative insights from interviews with surgeons aligned with the quantitative data, revealing a strong preference for the LLM-based VUI. Surgeons emphasized its intuitiveness and highlighted the potential of LLM-based VUI in expediting decision-making in surgical environments.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; **Empirical studies in HCI**.

Keywords

Large Language Models, Voice Control, Surgery, Augmented Reality

DIS '25, Funchal, Portugal

ACM ISBN 979-8-4007-1485-6/25/07
https://doi.org/10.1145/3715336.3735708

ACM Reference Format:
Hamraz Javaheri, Omid Ghamarnejad, Paul Lukowicz, Gregor Alexander Stavrou FEBS (HPB/SURGONC), and Jakob Karolus. 2025. LLMs Enable Context-Aware Augmented Reality in Surgical Navigation. In *Designing Interactive Systems Conference (DIS '25), July 05–09, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715336.3735708

1 Introduction

Wearable augmented reality (AR) has become a technology with vast potential for surgical navigation systems, promising enhanced precision and real-time guidance for medical professionals. Despite its promising capabilities, the effective integration of AR in critical domains, such as open surgery, has faced delays and notable challenges compared to other professional fields [16, 17]. One of the challenges associated with this integration delay is the current limitations in usable interaction methods to control the system [11, 20, 46] that could be easily used and adapted to critical domains [12]. Traditional input mechanisms, such as hand gestures or using combinations of eye gaze and virtual menus, prove impractical in areas where manual control is restricted and the AR view should not be occluded with too many virtual objects, hindering the full adaptation and utilization of wearable AR technology in these crucial fields [40, 54]. Among alternative interaction methods, voice-controlled assistants using speech commands stand out as a viable option, offering a hands-free and potentially intuitive means of interacting with AR systems during surgical procedures [8, 23, 24, 40]. While impressive in their ability to respond to voice commands, often lack contextual understanding and adaptability [21]. Furthermore, Relying solely on speech commands presents its own set of challenges, including the need to implement distinct keywords for each custom functionality, consequently increasing the complexity of the application for users [40]. In domains where simplicity, efficiency, and ease of use are mandatory, such as in critical surgical settings, introducing additional workloads or time-consuming processes may compromise the technology's adoption and effectiveness.

Despite recent improvements in interaction methods for AR applications [43] and in voice-controlled assistant systems[15], a gap still exists in the feasibility of using these techniques in critical domains, such as surgical navigation systems [12, 46]. The unique challenges presented in such vital domains have not yet received sufficient attention.

In this paper we presented a voice user interface (VUI) for a surgical AR-based assistance systems (ARAS), utilizing large language models (LLM), aiming to address the aforementioned limitations and enhance user experience in these critical settings. By harnessing the capabilities of speech recognition and LLMs, we introduce an intuitive method to control the system during surgical procedures helping to reduce the cogntitive load caused by interaction with the system.

We use LLM to perform function calls and control the system, enabling users to execute parallel functionalities based on the context of their request and the context presented in the application. We integrated our LLM-based VUI into the previously developed and clinically evaluated system, ARAS, designed to navigate pancreatic surgeries by enabling in-situ visualization of patient-specific 3D models of vessels and tumors, as well as access to supportive data.

We evaluated the feasibility of our LLM-based VUI for intraoperative interaction with ARAS throughout two studies, comparing it against previously tested VUIs for surgical AR applications using speech commands [13, 47]. Firstly, we conducted a user study with expert surgeons (N=9) involving surgical tasks to simulate the environment where ARAS would be normally used. We measured system usability, task execution time (TCT), cognitive load (NASA-RTLX), and conducted interviews to evaluate our method through quantitative and qualitative measures. Secondly, we employed both VUIs independently in two pancreatic surgeries to further evaluate these approaches in the users' end setup. Our findings highlight that the integration of LLM-based VUI led to faster task executions and reduced cognitive workload by generating more context-aware system outputs compared to VUI relied on speech commands.

Our work contributes to the design and development of VUIs specifically for time-critical, highly stressful, and demanding domains, such as surgery. These domains often do not fit general solutions but require systems that are carefully designed, implemented, and include user involvement throughout the process, as any unpredicted errors during system implementation could have significant consequences. To the best of our knowledge, our work is among the first studies to explore the potential of controlled usage of LLMs targeted and tested in a surgical environment beyond their commonly explored role as training and teaching tools [42, 50]. Our study showcases the feasibility of this approach in such a critical field by demonstrating not only the advantages of LLMs for users in terms of time and cognitive load reduction but also how LLMs can simplify the development process by combining and calling simple system functionalities to achieve more complex and context-aware system behavior.

2 Related Work

This related work section will first provide a comprehensive overview of AR applications in the surgical domain, followed by an exploration of interaction techniques within the medical field, and conclude with a dedicated chapter on natural communication and speech-Based interaction.

2.1 Augmented Reality in Surgery

Surgical assistance and simulation systems have undergone significant advancements with the integration of AR technologies [29, 33, 38, 52, 56]. While many approaches were implemented for display-based or handheld AR navigation systems for surgical procedures for example in laparoscopic surgery [45], the advancement of having wearable AR devices opens the door for further integration of technology in more challenging fields such as open surgery [17, 44, 48]. While the interaction principles to execute certain tasks with displays and touch-sensitive surfaces such as tablets have become traditional and well-accepted over the years, the control techniques for wearable AR devices have yet not been perfected. The restrictions available in critical fields such as open surgery in terms of available interaction and control methods, make the integration of this technology in a routine practice more difficult as the risks are very high and any distractions or errors caused by technological devices are not welcomed. In a study done by Saito et al. [46] it was demonstrated that using 3D holograms and hand manipulation caused higher cognitive workload compared to usual 2D supports as it required higher physical demand and effort.

2.2 Interaction Techniques in Medical Domain

The interaction techniques used in wearable AR technologies might vary depending on the task and used modalities [24]. While many interaction modalities could be used for a range of tasks in non-critical medical domains, such as training and simulations[5, 34]; the options for hygienic and sterilized surgical environments remain very restricted. Among different interaction and control modalities hand, voice, and foot input became possible options for interactions in the surgical environment as they do not require any direct contact with foreign objects [23]. Despite, the usability of these interaction methods, the environmental limitations during the surgical theater might not always allow practical usage of all these input modalities [40, 54]. Moreover, the limited number of different gestures that could be performed using only hand or foot inputs might restrict the system features [24] when compared to voice.

2.3 Natural Communication and Speech Based Interaction

The employment of voice input as a natural interaction, control, and communication method gathered extensive attention in the research. While the majority of applications focused on using voice in combination with other input modalities [27, 35, 43], the recent speech recognition algorithms and natural language processing provided mediums to use voice-based interaction as the sole input modality [49]. With the latest development of smart assistant systems and natural communication schemes through speech, voice input became popular especially where other input modalities could not be used [25]. With the outbreak of LLM, new possibilities have emerged to use natural communication schemes for interaction with assistant systems. Mahmood et al. [39] presented a LLM-powered conversational voice assistant that could be used in different areas. The combination of speech input and LLMs could be also used to achieve smarter assistant systems to control the system and perform certain system functionalities, as it was demonstrated by Dong et al. [14].

Despite the extensive research in the mentioned research fields, there has been a notable gap in exploring the various VC methods within AR environments. This gap in the literature is particularly noteworthy in critical domains such as open surgery, where practicality and safety become essential, and alternative input modalities may not be feasible.

3 Methodology

In this work we focused on finding an optimal interaction method for ARAS specifically designed for open pancreatic surgery, addressing the challenges associated with the impracticality of common input modalities in confined surgical spaces. This led us to explore and evaluate voice control (VC) methods for a more practical and efficient user experience in this surgical context. To guide our research, we addressed the following research questions over five consecutive phases as depicted in Figure 1.

Research Questions (RQs):

(1) What are the user and field-specific requirements in terms of interaction methods for an AR-based surgical navigation system?

Objective: To gather insights from surgeons to inform the design of the user interface and interaction with the AR system, ensuring it meets their practical requirements and enhances their ability to perform the surgery.
Method: Observations, interviews with surgeons, along with demographics of participating surgeons.

(2) How feasible is an LLM-based VUI for a surgical AR system, and how does it impact the user's cognitive load? How does this approach compare to previously tested VC methods, such as speech commands?

Objective: To gather insights from surgeons about the usability, cognitive workload, and their assessment of the LLM-based VUI method and compare it to the previously employed approaches.
Method: User study with surgeons in a simulated scenario involving surgically relevant system interaction tasks. Data collection using NASA_RTLX [22], and system usability scale (SUS) [9], and post-study interviews with surgeons, along with demographics of participating surgeons.

(3) How feasible is the LLM-based VUI in the users' end setup during surgery compared to speech commands? What are the users' (surgeons') reflections?

Objective: To gather insights from field surgeons about the usability of each VUI in an ecologically valid setup that involves highly stressful situations.
Method: Case study involving the employment of both VUIs, each in a pancreatic surgery session, and conducting postoperative interviews with surgeons about the interaction method used to control the surgical AR system.

We began our investigation by exploring the user and domain-specific requirements for a VUI for ARAS by interviewing experts from the field. Consequently, we developed two non-conversational VUI. Following the previous works on clinically tested voice interaction methods, our first VUI utilizes speech recognition and speech commands [13, 47] and serves as our baseline. The second VUI incorporates speech recognition, an LLM, and natural communication schemes to control the system. Internally, both VUIs have access to the same set of system functions of ARAS. In our first user study, we explored the usability of each VUI in a simulated environment and compared their performance in terms of added workload on surgeons while performing surgically relevant tasks, their usability, and conducted semi-structured interviews with participating surgeons. Finally, after proving the usability of both methods, we tested both VUIs in an ecologically valid environment during a clinical trial and conducted interviews with surgeons.

In this work, we solely focused on evaluating our proposed LLM-based VUI for ARAS in a time-critical domain, specifically in pancreatic surgery, and compared our approach to a conventional VUI using speech commands. We note that the evaluation of ARAS itself falls outside the scope of this paper.

4 System Design and Development

We used the Unity 3D game engine [4] for software design and development and used Microsoft HoloLens 2 device [3] as a wearable AR device. The selection of the HoloLens 2 for this application was mainly motivated by ethical, and safety considerations as it is CE-certified and was shown to have been successfully used in the medical domain before [28, 37, 46].

4.1 AR Assistance System

The ARAS software was designed and clinically evaluated as a supportive and navigation tool for open pancreatic surgery [30–32]. It was designed to contain two distinct feature sets to support surgeons throughout the surgery session: in-situ visualization and supportive data visualization (Figure 2). The system features segment-based visualization of a patient-specific 3D model, the ability to enable or disable modes between marker-based tracking for in-situ visualization, and maintaining the model's position and orientation. It also allowed for the visualization of supportive data, such as CT images and patient diagnoses. The designed user interface was aimed to provide means for surgeons to precisely execute features without interrupting the surgical flow.

4.2 Voice User Interfaces

During the pre-design stage, we interviewed four surgeons from the Hepatopancreatobiliary [1] and Visceral[2] fields (Table 1) to understand the user and domain-specific requirements in terms of interaction with ARAS. Furthermore, to enhance our understanding of the domain-specific constraints and environmental challenges we participated in 2 pancreatic tumor resection procedures as observers. The insights from interviews and observations made during the operation highlighted certain limitations regarding the feasible interaction method to control ARAS. With surgeons requiring both hands to perform intricate operations, incorporating hand gestures

[1]Hepatopancreatobiliary surgery consists of the general surgical treatment for benign and malignant diseases of the liver, pancreas, gallbladder, and bile ducts.
[2]Visceral surgery, also known as abdominal surgery, refers to surgery of the abdominal cavity and abdominal wall, endocrine glands, and soft tissue, including transplantation.

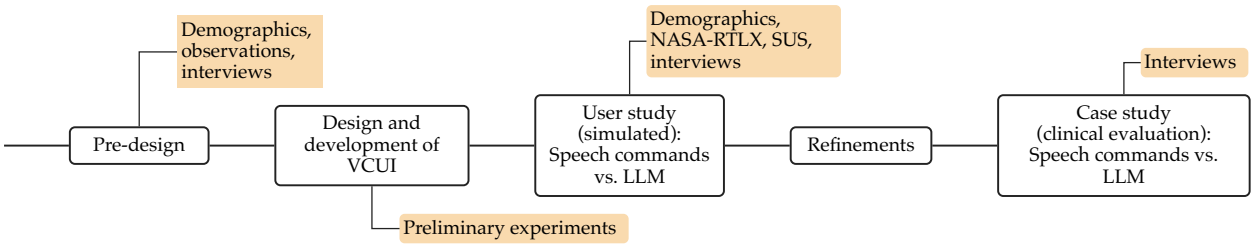
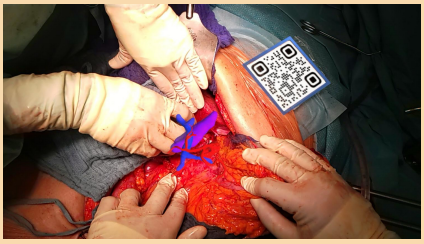
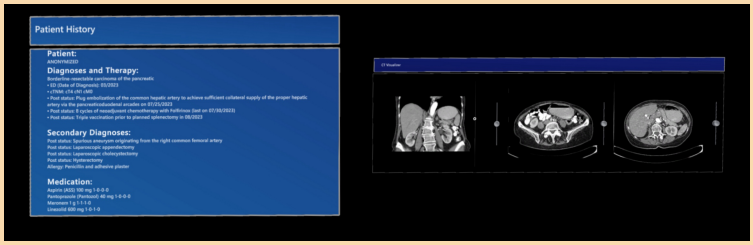


Figure 1: Chart showing different phases of our work from pre-design till the case study involving a clinical trial and associated data collection for each phase.



In-situ visualization

Features	Segment-based visuzalization
	Enable/disable marker tracking
	Freeze 3D model position and orientation



Supportive data visualization

Features	Manual repositioning
	Scrollable 3x DICOM visualizer
	Scrollable patient history panel

Figure 2: The characteristics of in-situ visualization and supportive data visualization feature set provided by the ARAS software.

is severely constrained. Furthermore, the physical setup around the operation field of the surgery table, where four surgeons and a nurse typically stand (as illustrated in Figure 3), results in a limited field of view for the AR application. Consequently, placing any virtual objects between surgeons and the operation area can obstruct the surgical view, introducing undesirable visual occlusion. Objects positioned behind other medical professionals are not only obscured but also challenging to interact with, given the physical constraints of extending hands and arms in a confined space. These insights underscore the necessity of an alternative user interface, leading us to explore and evaluate fully VUI for a more practical and efficient user experience in the context of surgery. Consequently, we developed two VUIs to interact with ARAS. While the first VUI follows the previously tested VC methods for the surgical domain using speech commands, our second VUI uses LLM and a natural communication scheme.

4.2.1 VUI with Speech Commands. In implementing the speech commands, we utilized the Windows speech recognition service (using IMixedRealityDictationSystem service) [1] and Microsoft Mixed Reality Toolkit, MRTK2 [2], to define specific keywords for triggering desired functionalities. These keywords were carefully selected in collaboration with surgeons from the field, ensuring relevance and ease of recall. A total of 34 unique keywords were assigned to trigger the system functionalities (Table 2). Designed as medical terms commonly used in daily practice or intuitive words, each keyword corresponded either to the names of anatomical structures available in the 3D model (Figure 4) or as intuitive words

Table 1: Profiles of interviewed surgeons before design and development process.

Age	Gender	Field	Surgical Experience (year)
51	Male	Hepatopancreatobiliary	23
53	Male	Hepatopancreatobiliary	25
40	Female	Hepatopancreatobiliary	13
33	Male	Visceral	4

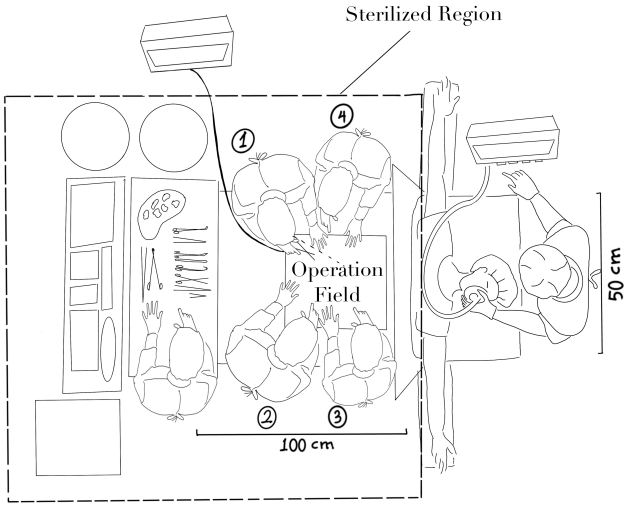


Figure 3: A top view of the surgical room setup. The placement of the medical staff around the table. The limited area around the operation field, and sterilization rules restrict the interaction with AR surgical assistance system. The surgeons are numbered based on their role in the surgery with Number 1 being the lead surgeon.

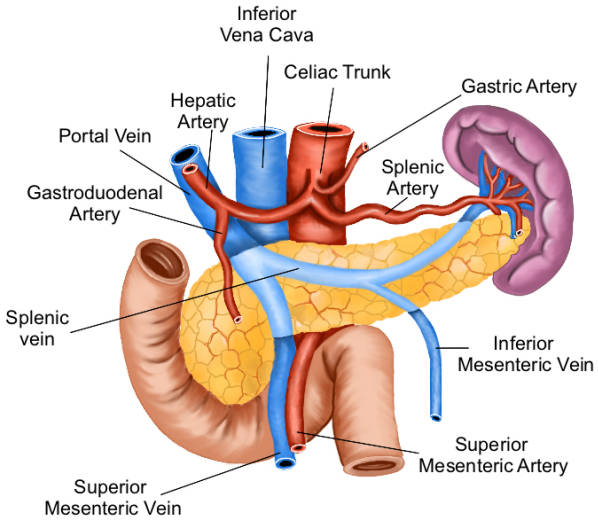


Figure 4: Anatomical drawing of 3D reconstructed segments in ARAS and their positions around the pancreas. The names of these segments were used as keywords for VUI using speech commands.

routinely used in a clinical environment or daily life. Recognizing the challenges of keyword recognition and aiming to enhance flexibility, some functions could be executed using multiple keywords or synonyms. For instance, the keyword "cancer" could be used interchangeably with the keyword "tumor" to enable or disable tumor visualization (Table 2). Additionally, all the keywords used for 3D visualization could be used in a combination of ON/OFF to either enable or disable some segments visualization or without these extensions as toggle behavior between on and off mode. To provide visual feedback to the user confirming the successful recognition of the spoken keyword, we used the MRTK tooltip component to be shown upon the detection of a speech command.

4.2.2 LLM Based VUI. We designed and implemented a system framework to automatically execute system functions upon user query stated via natural speech and speech recognition using the same Windows speech recognition service as the speech commands method [2]. We implemented a GPT communicator layer for Unity in C# to facilitate external API calls to chat-GPT from Unity and used GPT-3.5-turbo model to process the user request and return the system functions. Our LLM-based VUI framework consisted of four main components, a dynamic initial prompt generator, a dictation service, a response handler, and a GPT 3.5 model (Figure 5).

We developed an adaptive prompt generator to dynamically create initial prompts specific to each patient, aiming to provide patient-specific contextual information while defining the task for the LLM.

We used the patient's specific 3D model meshes to calculate the proximity and relational distance of each 3D object in the model, such as vessels and tumors, to other structures. We further provided the system with the patient diagnosis, and surgical resection guidelines, along with a list of system functions, and heuristic examples (Figure 5).

To mitigate hallucinations by LLMs, as demonstrated in other related studies [19, 36, 55], we implemented an auto-repeat function that resends the initial prompt to the chat. Our preliminary experiments revealed that the accuracy of LLM outputs decreases as more user inputs are added, causing the model to lose the context of the initial prompt—which contains critical patient-specific information and the main task. This leads to increased hallucination in the generated responses.

To counter this effect, we developed a mechanism that periodically reminds the LLM of the task and relevant information by resending the initial prompt after each user input. Once the user's request is processed, this reminder prompt is automatically sent to the chat, ensuring that the model retains the original context without the user noticing.

As a safeguard for occurred hallucinations which resulted in non-accurate response from LLM, we have implemented a reset function that would be used to correct the initial prompt and reset the chat. Like other system functionalities, the LLM could call upon this function based on the context of the user query. Users were informed about this functionality and instructed that if the system executed an incorrect or unexpected action, they could notify the LLM and specify the correct response for that situation. Apart from this reset function, no extra function or direct annotation to the study tasks (C5.1) was included in the initial prompt to avoid potential performance bias for the sake of the study.

As the GPT model only receives data in text format, we generated a JSON file containing all this information along with the requested task to return the appropriate system functions and variables based

Table 2: Implemented voice keywords and their functionalities

Type	Voice Keywords	Functionality
Patient information	Patient history / diagnosis / medication	Activates/Deactivates patient history panel
	CT / CT Image / Tomography / Computed Tomography / CT scans	Activates/Deactivates CT Visualizer
3D visualization	arteries/veins	Activates/Deactivates rendering of all arteries/veins
	Sternum	Activates/Deactivates rendering of the associated structure
	Celiac trunk	
	GDA	
	Mesenteric artery	
	Splenic artery	
	Gastric artery	
	Hepatic artery / liver artery	
	Portal vein	
	Vena cava	
	Splenic vein	
	Mesenteric vein	
	Tumor / lesion / cancer	
	Variation	Activates/Deactivates rendering of unusual anatomical structures
Control commands	Go up/down	Activates the automatic scrolling up/down of the CT slices
	Stop	Stops the automatic scrolling of CT slices
	Capture photo / hologram	captures a photo using HoloLens mounted camera with / without holograms
	Freeze	Freezes the 3D model position and orientation and disabling the marker tracking
	Marker tracking	Enables marker tracking
	Reset	Resets the app including the 3D model position and orientation relative to sternum marker

on the given sentence. The JSON format used for the initial prompt is given in Appendix A.

After successful initialization, the user could send a request to the application through voice query using a natural communication scheme. The dictation is activated using a single speech command called "Assistant" to avoid false queries. After activation of the dictation system, the system starts listening to the user query. During preliminary testing, we observed that most of the requests from users to the LLM last around ten seconds. Therefore, we set the default listening time to ten seconds. However, if the ten seconds is exceeded and the user is still speaking the system would wait for two seconds of silence before sending the query to the LLM. This way we made sure that the system would at least listen to the user query for ten seconds while also allowing for longer queries. The transcribed user query would then be formatted to JSON and sent to the GPT model via the GPT communicator. Upon receiving the response from GPT, the response is first validated, and if there is no error in the received format or request to reset the chat by LLM due to the user correction, then associated system functions are executed and presented to the user.

If there is a call to reset among the received functions from LLM, then the system stores the user interaction example along with the correction that the user provided to LLM. The active chat would be terminated another initial prompt would be generated using the recently added user example and a new chat would be initiated.

In addition, similar to tooltips used in the speech commands approach, we designed a virtual panel (Figure 6) to provide a real-time transcription of the user's voice input, to provide visual feedback about the recognition of the user's query. This decision was made to show the user about the recognized request and give the user a chance to correct it if it was detected wrong.

5 User Study 1: Evaluation During Simulated Surgical Scenarios

We employed a mixed-method evaluation approach, integrating both qualitative and quantitative data analyses, to comprehensively assess the feasibility of our proposed LLM-based VUI and draw comparisons with the VUI method utilizing speech commands. In this study, we compared and evaluated both methods in a simulated lab environment focusing on our two research questions (RQ2). We conducted a within-subject study, where all participants experienced both VUI (LLM and speech commands) for two different patient cases. To avoid potential bias the order of the VUI and the patient case were counterbalanced.

5.1 Study Design

Aligned with ARAS's primary function, the study tasks focused on adjusting the visualization of 3D model segmentations to guide various phases of surgery. This approach aimed to test the VUIs within their relevant context in a controlled, simulated environment.

Even though the tested system's functionality was limited, this task design aimed to focus more on interaction with the system as a result of cognitively challenging tasks which would be the normal case during the pancreatic surgery. The decision to trigger different system functions to visualize various combinations of structures depended on several factors, such as the relationship between structures and the tumor, patient history, surgical guidelines, and the current stage of the surgery. Any unnecessary virtual objects or structures visible in the AR view could confuse the surgeon or unnecessarily occlude the view. It is worth noting that while it is possible to implement separate functions for certain predefined guidelines, often the decision of which structure combination to

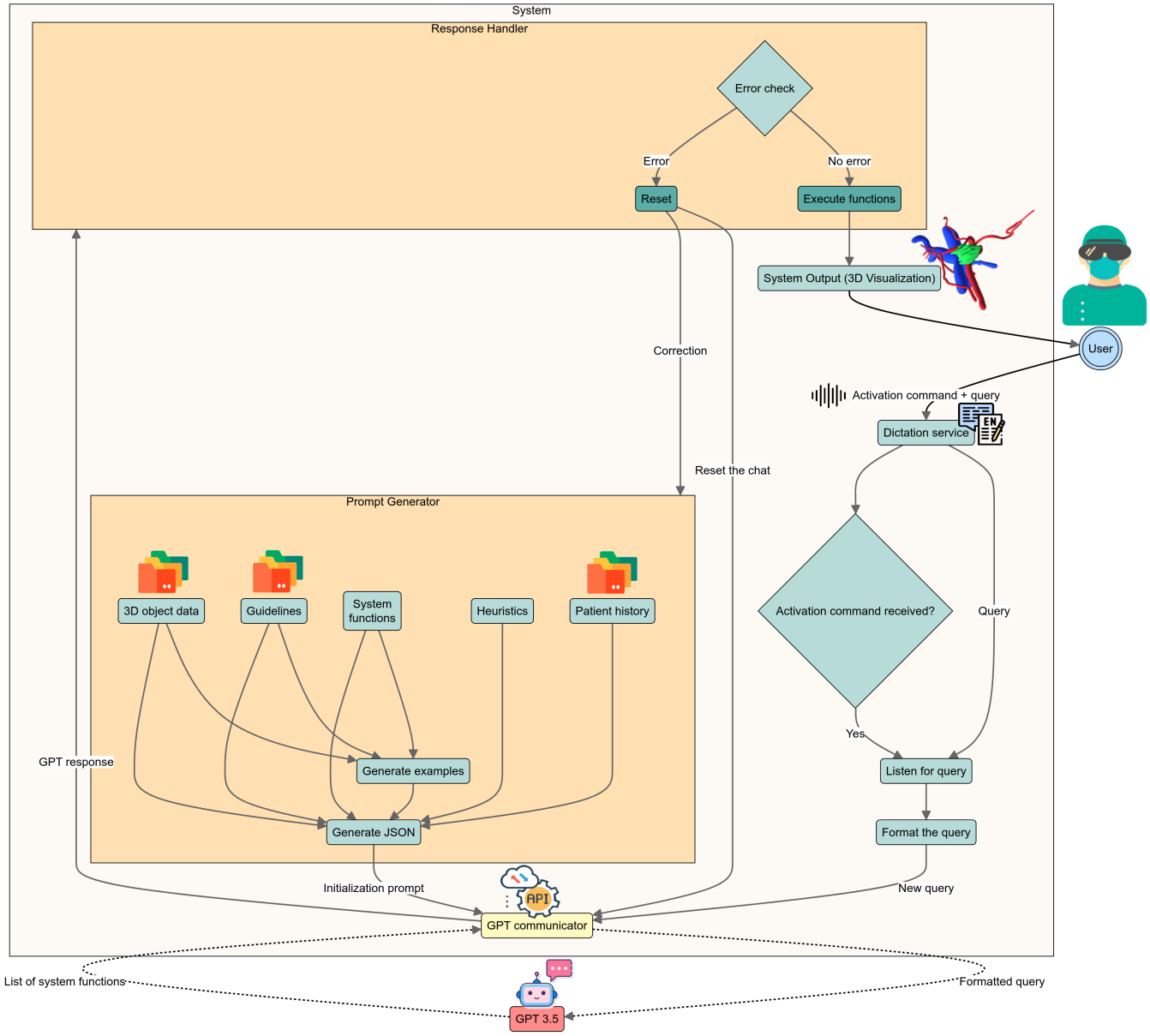


Figure 5: Overview of LLM-based VC framework. The system begins with loading patient files and function descriptions to generate an initial prompt. The system functions then is called upon the user's query via speech.

visualize is not definite, hard to implement, requires processing power on the operating device, and highly varies depending on each patient case and the stage of the operation.

Given the multifaceted nature of pancreatic surgery, which is divided into consecutive sessions for the preparation and resection of the vascular system and pancreas organ infiltrated with the tumor, we tailored the tasks to align with the interaction with the system during these distinct stages of each intraoperative session.

The task designs and their execution order were advised by experienced surgeons to simulate the progression of pancreatic surgery and the associated workload on surgeons. This approach aimed to replicate the decision-making process, considering the varying cognitive difficulty levels involved in identifying vital structures at different stages of the surgery.

The first two tasks targeted the commonly used first surgical approaches therefore the names of the structures to be visualized were given in the task description. Tasks 3 and 4 are performed in occasional situations based on the progress of the operation, however, due to the fix procedure performed in these tasks the names of the essential structures to be visualized were also given

in the task description. Tasks 5 and 6 were designed to address later stages of the operation where the surgeon is required to make complex decisions on which structures are needed to be observed to guide a critical phase of the operation such as the tumor resection phase. Therefore, in these last two tasks, the names of the structures to be visualized were neither given in the task description nor were annotated in the system as the decision to which structure to be visualized is subjective to the surgeon and might vary. While task 5 focuses on visualization of the structures that are affected by the tumor, task 6 focuses on visualization of the structures that need to be removed along with the tumor which is not always limited to those structures affected by the tumor. The task descriptions were as follows:

Tasks:

(1) **Kocher maneuver:** During this task participants were asked to only enable the visualization for the following structures: the tumor, inferior vena cava, and portal vein [53].
(2) **Preparation of the hepatoduodenal ligament:** During this task participants were asked to enable visualization for the following structures: portal vein, hepatic artery, and gastroduodenal artery [53].
(3) **Uncinate-first approach:** During this task participants were asked to enable visualization for the following structures: tumor, portal vein, and superior mesenteric artery [53].
(4) **Artery-first approach:** During this task participants were asked to enable visualization for the following structures: hepatic artery, gastroduodenal artery, celiac trunk, and superior mesenteric artery [53].
(5) **Tumor infiltration:** Participants were asked to only enable those structures that are infiltrated by the tumor. They were free to look at the CT images to decide or observe the 3D model as it would be the case for the real surgery.
(6) **Complex surgical decision:** During this task the participants were asked to evaluate and make decisions about which structures should be resected (surgically cut or removed along with the tumor to secure patient safety) during the pancreatic resection and only enable the visualization of those structures.

During this study, we used a medical manikin to simulate the surgical scenario where the AR assistance system (Figure 6). We used the reconstructed 3D models of two real patients with complex pancreas tumor localization with vascular involvement to efficiently address all the above-mentioned tasks. Both study groups included both patient cases. The participants were asked to stand around the table where the manikin was placed and position themselves as lead surgeon position (Figure 3, surgeon number 1) as the main decisions during the surgery and the above-mentioned tasks are usually performed and are decided by the lead surgeon.

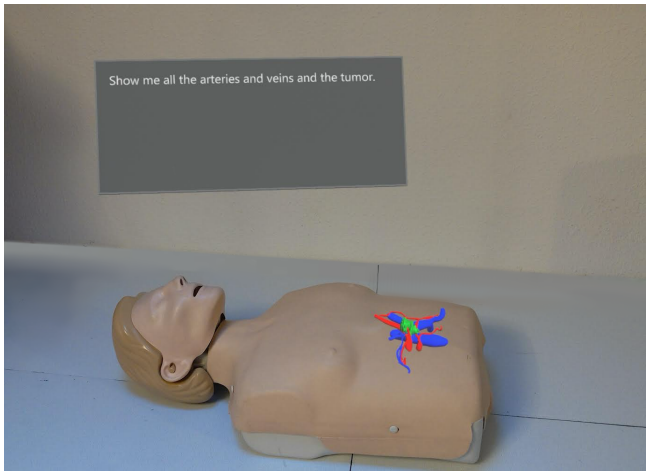


Figure 6: A captured image from the AR surgical assistance system using LLM-based VUI. The manikin and transcription panel were used in the first study to simulate the visualization of the overlayed patient-specific 3D model during the surgical session.

5.2 Measures and Data Analysis

We used quantitative and qualitative measures to evaluate different VUIs. TCT (measured in seconds) was recorded for each performed task. We also recorded the attempt count for the successful completion of each task.

We used the SUS [9] and NASA-RTLX [22] questionnaires to evaluate system usability and perceived cognitive workload after the completion of all tasks using each VC method. We concluded with a semi-structured interview with each participant to gather qualitative insights about their experiences with each method.

To analyze the quantitative data, we formulate the following hypotheses:

Hypotheses (Hs):

(1) The LLM-based VC leads to lower task completion times.
(2) The LLM-based VC leads to lower cognitive load as measured by NASA-RTLX.
(3) The LLM-based VC has better usability as measured by SUS.

After confirming the normality of the data, we thus conducted one-tailed paired t-test to confirm or reject our hypotheses. For H3, we conducted a one-tailed Wilcoxon signed rank test instead, since data normality was violated.

All interviews were transcribed verbatim. Given the volume of the data, we followed the pragmatic approach to qualitative analysis as recommended by Blandford *et al.* [7]. Initially, two researchers analyzed 25% portion of the data. Following this, we created a preliminary coding framework through iterative discussions. The remainder of the interview data was then divided equally among the two researchers for coding. In a concluding discussion, we refined the coding framework further and identified overarching themes.

5.3 Procedure

The study procedure started with participants completing demographic questionnaires to provide essential background information. To mitigate order bias and potential learning effects, all participants engaged with both studied VUIs in a counterbalanced order with a different case for each VUI. Prior to task execution, participants were familiarized with the AR system and its functionalities using each VUI. Data recording was initiated after participants confirmed their ability to successfully interact with the system. Each participant systematically performed all experiment tasks for two sessions, each session using a different VUI and different patient case. Task progression was subject to verbal confirmation of the participant regarding the accurate visualization of structures. Upon completion of the six tasks in each session, participants were prompted to fill out paper-printed questionnaires asking them to specifically answer the questions considering the experienced VUI but not the patient case. After completing both sessions, a brief semi-structured interview was conducted to gather insights about participants' opinions regarding each VUI. The whole study took approximately 40 minutes per participant. The study received the approval of the Medical Association of Saarland ethical committee board, ensuring that all aspects of the research adhered to established ethical guidelines.

5.4 Participants

Our study included nine volunteered experienced surgeons with a mean age of 42.44 (SD = 7.49) and 14.33 (SD = 7.35) years of average surgical experience. We compared the participant number with the participant number required for a usability evaluation. The number of participants in this study falls above the acceptable range of 4 ± 1 [10, 26], considering their high expertise in the field. The participants reported an average of 11 to 50 times experience with AR technology and 2 to 10 times experience with LLM or generative AI systems including chatbots and conversational assistance systems. The detailed characteristics of the participants are given in Table 3.

Table 3: Participant Characteristics (N=9). Likert scale values range from 1 to 5, with 1 being the least and 5 being the most frequent or proficient. SD = Standard deviation

Characteristic	Mean	SD
Age (years)	42.44	7.49
Surgical Experience (years)	14.33	7.35
How many times used AR (1-5 Likert Scale)	3.00	1.50
How many times used LLM (1-5 Likert Scale)	2.78	1.20
English Proficiency (1-5 Likert Scale)	3.89	0.60

5.5 Results

5.5.1 Quantitative Measures: TCT, NASA-RTLX, SUS. The participants completed all six tasks using both methods: LLM, with a mean attempt count of 1.074 (SD = 0.328), and speech commands, with a mean attempt count of 1.370 (SD = 0.784) for successful completion. While the average error rate for speech commands due to misrecognition or failure to recognize keywords was calculated at 27%, the average error rate in executing system functionalities when using the same speech recognition service with LLMs was 6%. As shown in Figure 7, individual one-tailed paired t-tests revealed that the LLM-based VC method yielded a significantly lower TCT for all tasks: Task 1 ($t(8) = -3.04$, $p < .01$), Task 2 ($t(8) = -2.33$, $p < .05$), Task 3 ($t(8) = -2.60$, $p < .05$), Task 4 ($t(8) = -2.56$, $p < .05$), Task 5 ($t(8) = -4.34$, $p < .01$), and Task 6 ($t(8) = -4.16$, $p < .01$). This result confirms H1.

The overall NASA-RTLX score was significantly lower for the LLM-based approach ($t(8) = -2.24$, $p < .05$), confirming H2. Specifically, it scored significantly lower for the subscales mental demand ($t(8) = -2.71$, $p < .05$), physical demand ($t(8) = -2.35$, $p < .05$), and effort ($t(8) = -2.36$, $p < .05$). We found no significantly lower score for LLM-based VUI for the other subscales, temporal demand, performance, and frustation. A visualization of this result can be found in Figure 8.

Using Bangor et al. [6] rating scale, the LLM-based VUI was classified as "excellent" with a SUS score of 87.78, and VUI using speech command was classified as "good" with and 79.17 SUS score. Despite the better performance of LLM, it did not yield a statistically significant higher SUS score. Thus, H3 could not be confirmed.

5.5.2 Interview Findings. Our analysis identified three main themes: User preferences and experience, context of use, and limitations and future improvements

User Preferences and Experience: Participants generally expressed a strong preference for the LLM VUI over traditional speech commands. They appreciated the flexibility and intuitiveness of the LLM, which allowed for more natural communication and seemed to reduce stress by accommodating various phrasings and intents without requiring specific keywords.

One surgeon reflected on the ability to articulate complex queries and receive accurate, contextually relevant information was particularly valued in high-stakes environments like operating rooms, saying:

> *I personally think it's much less stressful to have such language support because I can say whatever I want, how I ever want to phrase it and the system realizes what I want. (P1)*

Furthermore, they emphasized that not only the LLM-based VUI provided a more natural way of communicating with the system but also helped them to reduce the burden of thinking and making decisions about the requirements of the task by performing context-aware function calls. P2 pointed out this aspect stating:

> *For example, for tumor infiltrations, you have to first look at the tumor and which vascular system is infiltrated, then tell each time what to open or turn on [Speech Commands]. That's why I think it was much better with LLM, with large language model, because I simply asked and that showed. (P2)*

Context of Use: The interviews revealed that, while both VUIs are usable, the specific context and situations in which the system is employed may highlight the unique potential of each method. While the LLM-based method could be highly beneficial in stressful and time-critical conditions, the voice command might be a better option for surgeries that do not have time criticality. P8 and P9 highlighted this by following statements:

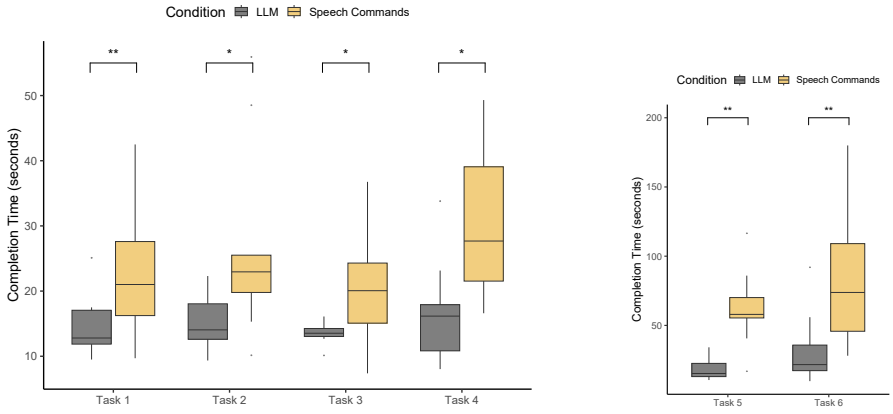


Figure 7: Task completion times for each task given the different VC methods (LLM, Speech commands). LLM yielded significantly lower completion times for all tasks (marked with * for $p < .05$ and with ** for $p < .01$).

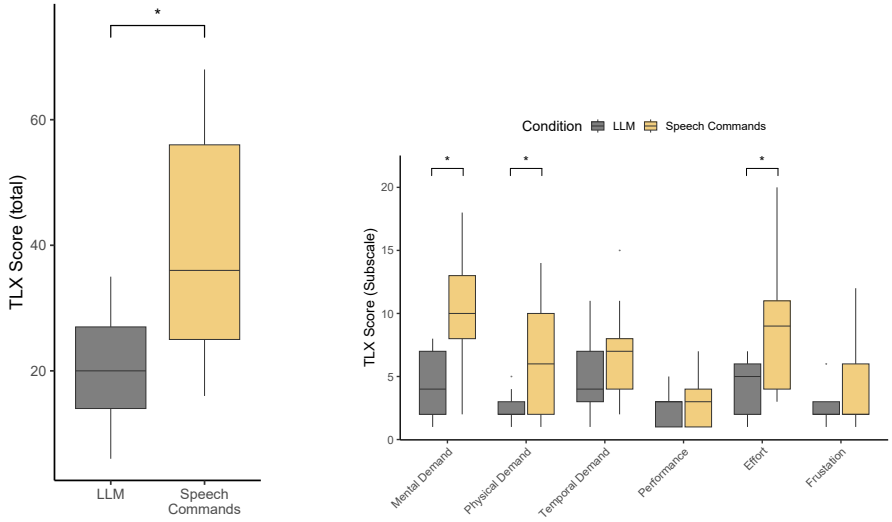


Figure 8: Total NASA-RTLX score (left) and individual subscale scores (right) given both VC methods (LLM, Speech commands). LLM yielded significantly lower scores (marked with *).

It depends on which time I have to use. So for example, if you say there's an emergency, see, then I would prefer the AI [LLM]. In obesity surgery I have time. I have no emergency. So I don't have tumor and I have more time. In the process of the operation you can say remove this, remove that. So step by step [Speech Comands]. (P9)

If the patient bleeds, even a few seconds wait is already annoying (...) nevertheless, even with bleeding I find the second one better [LLM], because you can say directly what you want to see without thinking. (P8)

Limitations and Future Improvements: The interviews revealed the importance of accurate dictation and robust speech recognition service for both VUIs. The visual feedback on the real-time transcription of the user query in the LLM-based method caused extra confusion, showing the potential misjudgment about the system's capability by the user. As users attempted to correct misinterpreted words upon observing incorrect transcriptions, the clarity of their requests diminished, leading to decreased performance in triggering the relevant functions by LLM. Conversely, the LLM would typically mitigate such errors by interpreting them as typographical mistakes, thereby maintaining higher accuracy in understanding and responding to user commands. P4 pointed out this matter by raising attention to the system's capability being affected by misinterpreted words due to the different pronunciations saying:

The less I say, it's supposed to be easier for the system to understand me, right? I don't know. I was thinking when I say too much then the system doesn't understand me because I said too much. With shorter words

you minimize the misunderstanding when someone pronounces it differently. (P4)

The surgeons also raised concerns about the default listening time that was set in the LLM-based approach. P3 suggested adopting an approach where the initialization and ending of the listening time could be activated by some keywords to refrain from waiting if the query time is less or longer than the default listening time, saying:

For example, you say assistant or something to begin. But can I also say end of sentence or so that I don't have to wait those couple of seconds. (P3)

Additionally, participants mentioned that the LLM-based VUI required a clear statement of the request. Despite the ease of use, the interaction would be even easier over time as one would learn how to clearly phrase their request. P3 reflected on this saying:

When I tried the second one [LLM], with the AI, I think it's even easier to use if you've done it before. Then you have the routine of what you have to say so that the device understands what I want, and then it's easier and quicker. I mean how I should formulate my question so that the system understands me and shows the result that I want. (P3)

5.6 Implications

Our study comparing LLM-based VUI to a VUI using speech commands has revealed implications for the integration of such technologies into surgical settings. These implications highlight the potential benefits and necessary refinements for practical application:

Potential superiority of LLM-based approach in critical surgical moments: The LLM-based approach demonstrated advantages, particularly when a decision-making situation was involved. It exhibited significantly reduced execution times across various tasks. As tasks increased in cognitive demand, particularly in Task 5 and Task 6, the disparity in execution times became more pronounced, reflecting the challenge of mental workload and decision-making when using speech commands. Moreover, assessments of cognitive load indicated a lower mental demand with the LLM-based approach. This convergence suggests that LLM-based VUI could offer a superior option in real surgical environments, where timely decisions are required during constrained time frames.

Need for system refinements prior to real-surgery evaluation: However, our findings also illuminate areas necessitating refinement before practical deployment in surgical settings. The real-time transcription panel introduced confusion as users attempted to rectify misinterpreted words, compromising the clarity of sentence context. Additionally, while LLM-based VUI facilitated quicker task completion, further enhancements in dictation service are essential to mitigate any remaining delays and optimize task execution times.

6 Case Study: Evaluation During Actual Surgery

Following the proven usability of both VUI systems in our initial user study (section 5), this study aimed to further evaluate these VUIs in a real surgical setup, addressing our third research question (RQ3). This phase sought to confirm our findings under actual surgical conditions, which can differ significantly from laboratory environments due to factors such as higher stress levels and time constraints. Building upon our findings from our first user study (Section 5.5), we have first performed refinements (Figure 1, Refinement) to our approach and later evaluated each VUI during a pancreatic tumor removal surgery (Figure 9).

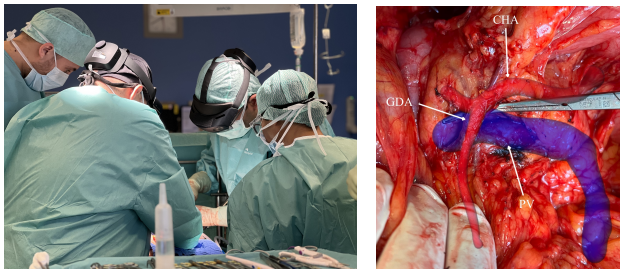


Figure 9: Actual pancreatic surgery. The right picture shows a snippet from the application view captured from a surgeon's device. GDA: Gastroduodenal Artery, PV: Portal Vein, CHA: Common Hepatic Artery

We performed the following refinements to our LLM-based VUI: The transcription panel providing real-time feedback to visualize the transcription of the voice query panel was removed, as it was observed to cause more confusion. Users attempting to correct what they perceived as misinterpreted words during their query can diminish the efficiency of the LLM method, as the context of the sentence may become unclear. Instead, we used conversational audio feedback similar to those commercially available conversational assistants such as Siri [3]. We used sound saying "OK" to indicate receiving the user query and "Please state your request differently" when the LLM response did not yield any of the defined system functions. Furthermore, we adapted the listening time after activation of the dictation service. The user request would automatically send to the LLM model after one and half seconds of silence without waiting for any further default time.

After performing refinements, we deployed our previously developed AR Assistance system designed to visualize 3D model of the patient during the surgery with the capability of both VUIs.

6.1 Study Design

To evaluate our LLM-based VUI and compare its outcomes, such as TCT and cognitive load, with the speech command during actual surgery, we exclusively employed qualitative measures and conducted post-operation interviews. This decision was driven by the inherent variability in each patient case, which might inevitably affect cognitive load measurements due to the unique nature of each surgical procedure. Similarly, task execution time would be influenced by the specifics of the surgery being performed. Consequently, a direct comparison between the two methods across different surgeries would not yield meaningful results. Thus, we chose to gather insights through interviews and observations, with

[3]Siri is Apple's voice-activated virtual assistant, available on iOS devices such as iPhones and iPads.

an observer researcher participating in the surgery sessions, making notes and observations about user interaction with the system. This approach allowed us to assess the impacts of each VUI, enabling us to collect qualitative data that could inform future improvements and implementations.

6.2 Procedure

To validate our approach in an ecologically valid environment, we conducted clinical trials involving the intraoperative evaluation in patients with underlying (borderline) resectable pancreatic tumors who required various types of pancreatic resection. The trials took place at Saarbrücken Klinikum hospital, certified to perform pancreatic tumor resection surgeries.

The study protocol received approval from the Ethics Committee of the Medical Association of Saarland under registration number: (registration number: 159/23). The protocol of our study was also registered at ClinicalTrials.gov under the registration number: NCT06208579.

Patients provided informed consent prior to surgery. All participating surgeons were fully briefed on the experimental nature of the method and the device used. They voluntarily agreed to use the system during the surgeries, assuming full responsibility for its operation and the outcomes. The surgeons were also informed that they could discontinue the use of the system at any point if necessary, without obligation.

To avoid first-time use bias two of the surgeons (P1, P2) who participated in study 1 participated in this study. Each surgery began with two surgeons equipped with our designed wearable assistance system with both VC method capabilities. In each session, surgeons were asked to use only one of the VC methods. However, they always had the choice to use the other method if it was essential for the course of the surgery. We conducted interviews with surgeons after each surgery session about their experience with each VUI.

6.3 Results

No technical difficulties regarding the VUIs were observed during both sessions and both surgeons used the system with the assigned VUI throughout the surgery.

Interviews with two surgeons who experimented with both VUI across two surgical procedures proved the feasibility of our LLM-based VUI in a real surgical environment in line with findings from our first study.

The LLM's capability to discern user context and analyze patient data facilitated the surgeons in adjusting the visualization of patient 3D models according to the tumor's proximity more efficiently specifically in the initial preparation phase of the operation. This benefit was encapsulated by a participating surgeon, who remarked:

> *Today's patient had an anatomical anomaly so we had to be more careful identifying the vessels during preparation so we don't damage them because they were so close to the tumor. So we had to change the visualization a lot. At that moment actually the LLM was a big help actually because it saved us a lot of times. (S2)*

S1 also reflected on this topic saying:

> *I think the biggest difference between two [VUIs] was during the initial phase of the operation where we usually use the system more to identify the vessels. But when the vessels are identified and already visible we don't interact with the system much. (S1)*

Unlike the speech command method which requires precise pronunciation of predefined keywords, the LLM system maintained a more natural communication flow. This aspect was profoundly appreciated, as S2 shared:

> *Voice commands also worked fine but the issue with the voice commands is sometimes when people are talking around the table you have to say a word 100 times until the system detects it. LLM is more forgiving if that's a correct word to use. (S2)*

Additionally, S1 reported on the benefits of using natural communication to control the system and also sharing information with other staff around the table. S1 reported:

> *When I say a single word usually other staff surgeons don't know what I am doing because they don't see what I see in the device. But when I talk to the system the way how I talk normally, then they know, ok, now I am trying to see where some vessels are when I say, for example, show me the vessels near the tumor or like I want to see mesenteric artery. (S1)*

S2 also highlighted the benefits of the performed refinements, noting the reduced confusion from removing the transcription panel and improved system response times:

> *This time with LLM system we didn't have to wait much for the system to react so it was way better and less annoying. Also, I think removing the panel was a good decision as I didn't see what the system understands so I didn't worry much about correcting my request and the system worked even better. (S2)*

7 Discussion

The introduction of AR-based surgical assistance systems has significantly transformed surgical practices, offering an enhanced level of precision and support. As these technologies evolve, the choice of interaction method to control the system becomes a pivotal consideration. Our study introduces a novel VUI for surgical ARAS using speech recognition and LLM and conducts a comparative analysis with the conventional VUI using speech recognition and speech commands, focusing on enhancing operational efficiency and user experience in the critical context of surgery. Importantly, we tested both methods in controlled laboratory settings and real surgical environments, offering a robust evaluation of their practical application and performance. This dual-context approach allowed us to gather comprehensive insights into the efficiency, cognitive load, user preferences, limitations, and situational applicability of each VC method.

7.1 Efficiency and Cognitive Load: LLMs versus Speech Commands

The SUS scores, along with the successful implementation of both VC methods in simulated and real surgical environments, demonstrate their usability and confirm their applicability in critical medical settings. However, the distinction in performance, especially in time-sensitive scenarios like the initial phases of surgical intervention, which is the most mentally demanding phase, underscores the criticality of choosing the right control and interaction method.

The use of LLMs significantly outperformed traditional speech commands in TCT. This efficiency is attributable to the LLMs' ability to generate context-aware outputs and execute multiple functions simultaneously to achieve a certain undefined functionality, a quality unattainable with keyword-specific speech commands without implementing further keywords to perform this task. As functionalities expand, the speech command method suffers from scalability issues, requiring an ever-increasing list of keywords. Conversely, LLMs streamline this process, enabling parallel function execution based on user requests without necessitating an extensive set of unique commands.

A more intriguing aspect of LLMs lies not just in determining which function to call based on user requests, but also in acting as an intelligent assistant and generating outputs which normally requires a complex decision-making process. This was particularly evident when LLM successfully generated the correct output to call certain system functionalities even when the function name or its specific purpose was not directly mentioned in the user's query. The capability of LLMs to generate context-aware outputs using all available information represents a significant advancement towards truly intelligent user interfaces and assistance systems.

A detailed analysis of user interaction logs with the LLM revealed its ability to successfully make decisions in numerous instances where users would otherwise have had to decide themselves. This difference in performance compared to speech commands was particularly evident. Despite the system functionalities being simple and identical in both cases, the reasons for employing these functionalities were often complex. With speech commands, the user (a surgeon) needed to decide and then instruct the system to make specific changes in visualizations using keywords. In contrast, the LLM-based system handled the decision-making process.

For example, in task 6, P2 asked the LLM, "Can you show me what should be resected?" Despite no specific information or annotation regarding the task description or the structures to be enabled in this context, the LLM correctly decided to display the tumor and the infiltrated veins and arteries within the resection margins by invoking multiple system functions simultaneously. This decision-making process is highly complex, relying on factors such as patient history, tumor position, and surgical guidelines regarding resection margins.

This feature demonstrated by the LLM not only reduced task completion times but also significantly diminished cognitive load. This was also evidenced by the lower scores in the NASA-RTLX, indicating a more intuitive and less burdensome interaction for the user—something unachievable with speech command methods without pre-implementing more complex functions into the system.

In applications such as surgical navigation systems, where system interaction is part of a decision-making process and this process depend heavily on the specific context of each patient case and scenario, pre-implementing an all-encompassing solution is very challenging. In such settings, predefining an comprehensive rule-based or intent-driven system is highly challenging due to the variability and nuance involved. Specialized voice assistance methods, such as those using machine learning [21], require training the system with specific user terminology for different scenarios. Additionally, these approaches come with significant processing costs that might affect the performance of wearable devices. Here, LLMs can provide significant benefits by offering a more generalizable, adaptive, context-aware, and efficient approach to managing complex tasks, even when the underlying system remains relatively simple in terms of functionality. Their flexibility to interpret diverse inputs, adapt to new scenarios, and support open-ended reasoning makes them well-suited for assisting in dynamic decision-making processes.

By analyzing patient-specific data and examples given in the dynamically generated initial prompt, LLMs can offer custom recommendations, enhancing the support system's utility in high-pressure situations. This capability to interpret context and call relevant function combinations offers surgeons a richer, more contextual understanding of the patient's data, including visualization of details in the 3D models, an attribution that with conventional voice assistant systems using speech commands cannot be achieved [21].

Despite the apparent advantages of the LLM method in facilitating quicker, multi-functional requests, our study also highlighted a perception mismatch among some participants. They perceived speech command execution as a faster method for task completion, despite objective evidence showing the LLM method reduced TCTs. This discrepancy may be linked to the system's default listening time (minimum ten seconds or wait for 2 seconds of silence if longer than ten seconds) following the user query. It suggests the necessity for an adaptive approach in managing the activation and deactivation of the system's listening duration for the LLM-based method to ensure the receipt of full user queries without a long wait. As a shorter listening period could prematurely send incomplete queries to the LLM, while a longer period might unnecessarily delay the system's response.

7.2 Pros and Cons: Balancing Control and Transparency

In our study we found out that, speech commands, with their direct and deterministic nature, afford users a clear understanding of cause and effect. This transparency in interaction fosters a sense of reliability and control, essential in high-stakes environments like surgery. However, this method's scalability and flexibility are constrained by the need to predefine every command, which can limit the system's responsiveness to complex or unforeseen requests.

On the other hand, LLMs represent a paradigm shift towards more fluid, conversational interactions. By understanding and processing natural language, these models offer a dynamic and flexible interface that can interpret a broad spectrum of user requests. However, this sophistication comes with a degree of opacity. The "black

box" nature of LLMs can obscure the pathway from request to action, potentially undermining user confidence if the system's reasoning and decision-making processes are not sufficiently transparent.

This lack of transparency can occasionally lead to doubts about the system's decisions, especially given the potential for hallucinations in LLMs. To mitigate these concerns, future system designs could adopt a reasoning-based approach. In scenarios involving decision-making or uncertainty, the system could present its reasoning to the user, allowing them to review and ultimately serve as the final judge.

Moreover, when LLMs are applied in more critical contexts, moving beyond basic roles such as VUIs for controlling 3D visualizations and into more complex functions as intelligent assistants, the potential impact of system errors might become greater. In these high-stakes applications, mistakes resulting from model hallucinations could lead to more serious consequences, such as incorrect analysis, flawed decision-making, or unintended system behavior. To address these risks, incorporating a human-in-the-loop approach becomes essential. This means that the system should not carry out any requested actions without first presenting the proposed response or decision to the user for review and explicit confirmation. By involving the user as the final authority in the decision-making process, the likelihood of critical failures can be significantly reduced, ensuring both greater accountability and system reliability.

Our study revealed that the LLM system could effectively compensate for errors in the dictation and speech recognition system. Unlike speech commands, which necessitate precise pronunciation, the LLM system can infer the user's intent by analyzing the context of the query rather than focusing on individual words. This capability significantly enhances the system's flexibility and user-friendliness.

However, it became evident that providing users with real-time visual feedback of transcription could inadvertently lead to misjudgments about the system's capabilities. Users attempting to correct what they perceive as misinterpreted words during their query can diminish the LLM method's efficiency, as the context of the sentence may become obscured. This observation underscores the critical need for designing user feedback mechanisms that do not compromise the clarity of communication or the efficiency of the system control method.

Furthermore, mitigating these challenges necessitates clear communication about the system's operational boundaries and capabilities. Users need to understand not just how to interact with the system, but also the underlying principles guiding its responses. This understanding is crucial for formulating effective requests, especially with LLMs, where the context and specificity of language can dramatically influence outcomes. Training and educational programs play a pivotal role in this regard, equipping users with the knowledge to navigate the system's complexities and leverage its full potential.

We believe that a hybrid approach, integrating both speech commands and LLM capabilities, emerges as a promising solution to balance control with transparency. By allowing users to switch between modes based on the task's complexity or urgency, such a system combines the directness of speech commands with the adaptability of LLMs. For routine tasks or when precision is paramount, predefined speech commands could offer the most efficient pathway. Conversely, for tasks that require time consuming decision-making process or when additional context is required, the LLM recommendations could provide a faster solution.

Implementing a hybrid model also entails designing interfaces that intuitively signal which mode is in operation, thereby maintaining user awareness and trust. Visual or auditory cues could indicate the system's current state, whether executing a direct command or processing a more complex LLM-based request. Moreover, offering users the ability to override or specify the control mode empowers them to use the system's capabilities depending to their immediate needs and preferences. By carefully navigating the trade-offs between control and transparency, and by fostering an environment of continuous learning and adaptation, we can develop systems that not only enhance surgical outcomes but also align with the users' operational and cognitive needs.

7.3 Ethical Considerations

Maintaining ethical standards is crucial for preserving trust in medical research and innovation. By adhering to ethical guidelines, researchers and practitioners demonstrate their commitment to prioritizing patient safety and well-being over technological advancements.

In this study, we emphasized ethical adherence throughout all stages. We ensured that the introduction of the AR system did not compromise the safety of patients or surgeons, nor did it undermine the integrity of the surgical process.

To achieve this, we initiated the study only after obtaining full approval from the relevant medical ethics review board. All participants, including surgeons and patients, were thoroughly informed about the study, with their participation contingent upon a clear explanation and the collection of informed consent. Patients were made aware of the potential risks, benefits, and alternatives to ensure their participation was both voluntary and fully informed.

We also ensured that neither the AR system nor the VC method used did replace the surgeon's judgment, maintaining human decision-making in surgical procedures, and surgeons retained complete control over the system, consistent with the Fundamental Principles of Ethics [51].

Our key takeouts from this study regarding ethical considerations for future studies are as follows: The system should function solely as a supplementary tool to assist the surgeon without replacing the surgeon's expertise. The surgeon must retain ultimate decision-making authority, ensuring patient safety and adapting to the unique aspects of each case. Human intuition and experience should remain the final safeguard in surgical procedures.

7.4 Limitations and Future Work

Even though the results of this study are promising steps towards using LLMs not only as a VUI but also as intelligent assistants in the medical domain, our findings are limited to the specific functionalities of ARAS we used in this study. While these functionalities are integral to most of surgical navigational applications, a broader understanding of LLM capabilities requires further research. This should involve more complex system functionalities and tasks to fully explore and validate the potential of LLMs in diverse and demanding scenarios. Furthermore, due to the different criticality

level and domain-specific requirements involved in each different type of surgical procedure, further domain-specific research is required to assess the generalizability of our findings and suitability of this method for other surgical domains that benefit from AR surgical navigation tools such as neurosurgery and orthopedics [18, 41]. In our future work, we intend to broaden the scope of our investigation into the capabilities and opportunities presented by LLMs in surgical assistance systems beyond our current application as a function caller and VC method. By leveraging the advanced natural language understanding and processing capabilities of LLMs, we hope to uncover new ways in which these models can contribute to the enhancement of surgical outcomes, efficiency, and safety.

8 Conclusion

Our comparative study of two VUIs within an AR-based surgical assistance system highlights the distinct advantages and considerations associated with speech commands and LLM. We found that the LLM-based VUI offered significant improvements in operational efficiency and reduced the cognitive load of users by allowing for natural, conversational interactions and the ability to generate context-aware system behavior by executing multiple functions concurrently. However, the choice between LLMs and speech commands is not clear-cut, despite higher preference towards LLM user preferences may vary based on perceived control, transparency, and the context in which the system is employed. While speech commands provide a sense of direct control and transparency, LLMs require clear instructions to function optimally, which can sometimes challenge users. The idea of a hybrid model emerges as a promising solution, aiming to combine the strengths of both approaches to cater to a broader range of needs and situations in surgical settings. Looking ahead, we plan to expand our exploration into the potential of LLMs as conversational assistants that not only could control the system but could participate more in the decision-making process, further enhancing the capabilities of surgical assistance systems. This study lays the groundwork for future advancements in surgical technology, emphasizing the importance of the involvement of end-users during design and evaluation and the need for systems that balance efficiency, cognitive ease, and adaptability to the fast-paced, complex nature of surgical environments.

Acknowledgments

We extend our sincere gratitude to all of the surgeons who collaborated on this project. Their valuable time, expertise, and contributions were instrumental in making this study possible. Lastly, we acknowledge the financial support provided by CrossAct Project 01IW25001, which made this work feasible.

References

[1] 2020. WindowsSpeechInputProvider Class - Mixed Reality Toolkit. https://learn.microsoft.com/en-us/dotnet/api/. Accessed: April 2024.
[2] 2023. Introduction to Mixed Reality Toolkit (MRTK) for Unity - MRTK2. https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/?view=mrtkunity-2022-05. Accessed: April 2024.
[3] 2023. Microsoft HoloLens 2. https://www.microsoft.com/en-us/hololens. Accessed on September 14, 2023.
[4] 2023. Unity 3D Engine. https://unity.com/. Accessed on September 13, 2023.
[5] Mareen Allgaier, Vuthea Chheang, Patrick Saalfeld, Vikram Apilla, Tobias Huber, Florentine Huettl, Belal Neyazi, I Erol Sandalcioglu, Christian Hansen, Bernhard Preim, et al. 2022. A comparison of input devices for precise interaction tasks in VR-based surgical planning and training. *Computers in Biology and Medicine* 145 (2022), 105429.
[6] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.
[7] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.
[8] Jakub Blokša. 2017. Design guidelines for user interface for augmented reality. *Masaryk University* (2017).
[9] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[10] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
[11] Enylton Machado Coelho, Blair MacIntyre, and Simon J Julier. 2005. Supporting interaction in augmented reality in the presence of uncertain spatial knowledge. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*. 111–114.
[12] Fabrizio Cutolo, Benish Fida, Nadia Cattari, and Vincenzo Ferrari. 2019. Software framework for customized augmented reality headsets in medicine. *IEEE Access* 8 (2019), 706–720.
[13] A Dias, D Männle, T Balkenhol, Jürgen Hesser, N Rotter, L Huber, O Hoffmann, A Schell, B Kramer, A Lammert, et al. 2021. Augmented Reality during Parotid Surgery: Real-Life Evaluation of Voice Control and User-Experience. *Laryngo-Rhino-Otologie* 100, S 02 (2021).
[14] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5792–5793.
[15] Faruk Lawal Ibrahim Dutsinma, Debajyoti Pal, Suree Funilkul, and Jonathan H Chan. 2022. A systematic review of voice assistant usability: An ISO 9241–11 approach. *SN computer science* 3, 4 (2022), 267.
[16] PJ " Eddie" Edwards, Manish Chand, Manuel Birlo, and Danail Stoyanov. 2021. The challenge of augmented reality in surgery. *Digital Surgery* (2021), 121–135.
[17] Benish Fida, Fabrizio Cutolo, Gregorio di Franco, Mauro Ferrari, and Vincenzo Ferrari. 2018. Augmented reality in open surgery. *Updates in surgery* 70, 3 (2018), 389–400.
[18] Andrew A Furman and Wellington K Hsu. 2021. Augmented reality (AR) in orthopedics: current applications and future directions. *Current reviews in musculoskeletal medicine* (2021), 1–9.
[19] Boris A Galitsky. 2023. Truth-o-meter: Collaborating with llm in fighting its hallucinations. (2023).
[20] Yahya Ghazwani and Shamus Smith. 2020. Interaction in augmented reality: Challenges to enhance user experience. In *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations*. 39–44.
[21] J Gowthamy, A Senthilselvi, Aniket Kumar, S Aakash, and Gandikota Sreedhar. 2023. Enhanced AI Voice Assistance using Machine Learning and NLP. In *2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, Vol. 1. IEEE, 1–5.
[22] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[23] Benjamin Hatscher and Christian Hansen. 2018. Hand, foot or voice: Alternative input modalities for touchless interaction in the medical domain. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 145–153.
[24] Julia Hertel, Sukran Karaosmanoglu, Susanne Schmidt, Julia Bräker, Martin Semmann, and Frank Steinicke. 2021. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. In *2021 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 431–440.
[25] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.
[26] Wonil Hwang and Gavriel Salvendy. 2010. Number of people required for usability evaluation: the 10±2 rule. *Commun. ACM* 53, 5 (2010), 130–133.
[27] Ajune Wanis Ismail and Mohd Shahrizal Sunar. 2015. Multimodal fusion: gesture and speech input in augmented reality environment. In *Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Information Systems (INNS-CIIS 2014)*. Springer, 245–254.
[28] Vladimir M Ivanov, Anton M Krivtsov, Sergey V Strelkov, Anton Yu Smirnov, Roman Yu Shipov, Vladimir G Grebenkov, Valery N Rumyantsev, Igor S Gheleznyak, Dmitry A Surov, Michail S Korzhuk, et al. 2022. Practical application of augmented/Mixed reality technologies in surgery of abdominal cancer patients. *Journal of Imaging* 8, 7 (2022), 183.
[29] Hamraz Javaheri, Omid Ghamarnejad, Ragnar Bade, Paul Lukowicz, Jakob Karolus, and Gregor Alexander Stavrou. 2024. Beyond the visible: preliminary evaluation of the first wearable augmented reality assistance system for pancreatic surgery. *International Journal of Computer Assisted Radiology and Surgery* (2024),

1–13.
[30] Hamraz Javaheri, Omid Ghamarnejad, Paul Lukowicz, Gregor Alexander Stavrou, and Jakob Karolus. 2024. ARAS: LLM-Supported Augmented Reality Assistance System for Pancreatic Surgery. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 176–180.
[31] Hamraz Javaheri, Omid Ghamarnejad, Paul Lukowicz, Gregor Alexander Stavrou, and Jakob Karolus. 2024. Design and Clinical Evaluation of ARAS: An Augmented Reality Assistance System for Open Pancreatic Surgery. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 376–385.
[32] Hamraz Javaheri, Omid Ghamarnejad, Paul Lukowicz, Gregor Alexander Stavrou, and Jakob Karolus. 2025. From Concept to Clinic: Multidisciplinary Design, Development, and Clinical Validation of Augmented Reality-Assisted Open Pancreatic Surgery. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, Yokohama, Japan. https://doi.org/10.1145/3706598.3713458
[33] Hamraz Javaheri, Omid Ghamarnejad, Rizky Widyaningsih, Ragnar Bade, Paul Lukowicz, Jakob Karolus, and Gregor Alexander Stavrou. 2024. Enhancing Perioperative Outcomes of Pancreatic Surgery with Wearable Augmented Reality Assistance System: A Matched-Pair Analysis. *Annals of Surgery Open* 5, 4 (2024), e516.
[34] Chaowanan Khundam, Varunyu Vorachart, Patibut Preeyawongsakul, Witthaya Hosap, and Frédéric Noël. 2021. A comparative study of interaction time and usability of using controllers and hand tracking in virtual reality training. In *Informatics*, Vol. 8. MDPI, 60.
[35] Minkyung Lee, Mark Billinghurst, Woonhyuk Baek, Richard Green, and Woontack Woo. 2013. A usability study of multimodal input in an augmented reality environment. *Virtual Reality* 17 (2013), 293–305.
[36] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.
[37] Kai Liu, Yuan Gao, Ahmed Abdelrehem, Lei Zhang, Xi Chen, Le Xie, and Xudong Wang. 2021. Augmented reality navigation method for recontouring surgery of craniofacial fibrous dysplasia. *Scientific Reports* 11, 1 (2021), 10043.
[38] Abel J Lungu, Wout Swinkels, Luc Claesen, Puxun Tu, Jan Egger, and Xiaojun Chen. 2021. A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery. *Expert review of medical devices* 18, 1 (2021), 47–62.
[39] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).
[40] Helena M Mentis, Kenton O'Hara, Gerardo Gonzalez, Abigail Sellen, Robert Corish, Antonio Criminisi, Rikin Trivedi, and Pierre Theodore. 2015. Voice or gesture in the operating room. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 773–780.
[41] Antonio Meola, Fabrizio Cutolo, Marina Carbone, Federico Cagnazzo, Mauro Ferrari, and Vincenzo Ferrari. 2017. Augmented reality in neurosurgery: a systematic review. *Neurosurgical review* 40 (2017), 537–548.
[42] Devi Prasad Mohapatra, Friji Meethale Thiruvoth, Satyaswarup Tripathy, Sheeja Rajan, Madhubari Vathulya, Palukuri Lakshmi, Veena K Singh, and Ansar Ul Haq. 2023. Leveraging Large Language Models (LLM) for the plastic surgery resident training: do they have a role? *Indian Journal of Plastic Surgery* 56, 05 (2023), 413–420.
[43] SS Muhammad Nizam, Rimaniza Zainal Abidin, Nurhazarifah Che Hashim, Meng Chun Lam, Haslina Arshad, and NAA Majid. 2018. A review of multimodal interaction technique in augmented reality environment. *Int. J. Adv. Sci. Eng. Inf. Technol* 8, 4-2 (2018), 1460.
[44] Philip Pratt, Matthew Ives, Graham Lawton, Jonathan Simmons, Nasko Radev, Liana Spyropoulou, and Dimitri Amiras. 2018. Through the HoloLens™ looking glass: augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels. *European radiology experimental* 2 (2018), 1–7.
[45] Long Qian, Jie Ying Wu, Simon P DiMaio, Nassir Navab, and Peter Kazanzides. 2019. A review of augmented reality in robotic-assisted surgery. *IEEE Transactions on Medical Robotics and Bionics* 2, 1 (2019), 1–16.
[46] Yu Saito, Maki Sugimoto, Satoru Imura, Yuji Morine, Tetsuya Ikemoto, Shuichi Iwahashi, Shinichiro Yamada, and Mitsuo Shimada. 2020. Intraoperative 3D hologram support with mixed reality techniques in liver surgery. *Annals of surgery* 271, 1 (2020), e4–e7.
[47] Claudia Scherl, David Männle, Nicole Rotter, Jürgen Hesser, Jan Stallkamp, Tobias Balkenhol, Lena Huber, Benedikt Kramer, Anne Lammert, and Annette Affolter. 2023. Augmented reality during parotid surgery: real-life evaluation of voice control of a head mounted display. *European Archives of Oto-Rhino-Laryngology* 280, 4 (2023), 2043–2049.
[48] Claudia Scherl, Johanna Stratemeier, Nicole Rotter, Jürgen Hesser, Stefan O Schönberg, Jérôme J Servais, David Männle, and Anne Lammert. 2021. Augmented reality with HoloLens® in parotid tumor surgery: a prospective feasibility study. *ORL* 83, 6 (2021), 439–448.
[49] George Terzopoulos and Maya Satratzemi. 2020. Voice assistants and smart speakers in everyday life and in education. *Informatics in Education* 19, 3 (2020), 473–490.
[50] Julian Varas, Brandon Valencia Coronel, IGNACIO VILLAGRáN, Gabriel Escalona, Rocio Hernandez, Gregory Schuit, VALENTINA DURáN, Antonia Lagos-Villaseca, Cristian Jarry, Andres Neyem, et al. 2023. Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM). *Revista do Colégio Brasileiro de Cirurgiões* 50 (2023), e20233605.
[51] Basil Varkey. 2021. Principles of clinical ethics and their application to practice. *Medical Principles and Practice* 30, 1 (2021), 17–28.
[52] Petr Vávra, Jan Roman, Pavel Zonča, Peter Ihnát, Martin Němec, Jayant Kumar, Nagy Habib, Ahmed El-Gendi, et al. 2017. Recent development of augmented reality in surgery: a review. *Journal of healthcare engineering* 2017 (2017).
[53] Donald Venes. 2017. *Taber's cyclopedic medical dictionary*. FA Davis.
[54] Juan Pablo Wachs, Mathias Kölsch, Helman Stern, and Yael Edan. 2011. Vision-based hand-gesture applications. *Commun. ACM* 54, 2 (2011), 60–71.
[55] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469* (2023).
[56] Zhen-yu Zhang, Wen-chao Duan, Ruo-kun Chen, Feng-jiang Zhang, Bin Yu, Yun-bo Zhan, Ke Li, Hai-biao Zhao, Tao Sun, Yu-chen Ji, et al. 2019. Preliminary application of mxed reality in neurosurgery: Development and evaluation of a new intraoperative procedure. *Journal of Clinical Neuroscience* 67 (2019), 234–238.

A Initial Promt Json Format

```
Initial_prompt = {
  "description": "Depending on given sentences, Return
  only appropriate method or methods from the executable
  methods list without explanation.",
  "executableMethods": [
    "function_A(variables)", ..., "function_X(variables)"],
  "organTypes": [
    "Organ_A", ..., "Organ_X"],
  "OrganCategories": [
    "Category_A", ..., "Category_X"],
  "distanceData": [
    " Organ_A": { " Organ_A": xx, ..., " Organ_x": xx},
    ...,
    " Organ_x": {" Organ_A": xx, ..., " Organ_x": xx}],
  "guidlines": [
    " rule_A":" description of rule_A,
    ...,
    " rule_x":" description of rule_x ],
 "sentencesAndResultsExamples": [
    { "sentence": "Show me all of the arteries",
      "result": "function_{xx} (variable_xx)"
    },
    ...,
    { "sentence": "Show me the infiltrated vessels",
      "result": {"function_{yy} (variable_y1, variable_y2)",
      ..., "function_{yy} (variable_y1, variable_y2)" }}]
}
```