ScreenGPT – Developing and Evaluating a Custom Large Language Model to Support Cervical Cancer Prevention

Ph.D. thesis

Angyal Viola

Health Sciences Division

Supervisor: Prof. Dr. Dinya Elek, PhD

Official reviewers: Dr. Tóth Zoltán, PhD
Dr. Molnár László, PhD

Head of the Complex Examination Committee:
Prof. Dr. Maurovich Horvát Pál, MD, PhD, DSc

Members of the Final Examination Committee:
Dr. Kaposi Ambrus, PhD
Dr. Girasek Edmond, PhD

Budapest, 2026

1. Introduction

In the field of artificial intelligence (AI), natural language processing (NLP) is a key subdomain that focuses on enabling computers to interact using human languages. Within NLP, generative artificial intelligence (GenAI) refers to models that can generate new and coherent language outputs. GenAI systems can be further categorized into small language models (SLMs) and large language models (LLMs). LLMs play an increasing role in patient-facing applications. They can power conversational agents and have the potential to increase accessibility to healthcare information for populations.

Cervical cancer screening has been widely recognized as one of the most effective and broadly implemented preventive health measures worldwide. However, many individuals lack adequate knowledge regarding the importance of regular preventive screening and lack of trust in healthcare systems. Additionally, cultural beliefs and stigma surrounding gynecological care further prevent them from timely screening, highlighting the urgent need for patient education and accessible, evidence-based information.

Addressing these gaps requires innovative, multi-dimensional strategies, including the integration of digital health technologies, such as a web-based AI conversational agent powered by an LLM. Given the widespread recommendation for cervical cancer screening beginning at age 30 (25 in Hungary) and the target demographic's digital literacy, this topic was selected.

2. Objectives

The primary objective of this study was to investigate whether an AI-based conversation agent system could improve access to relevant cervical cancer prevention information by providing personalized advice tailored to individual user needs.

3. Methods

3.1. Backend

As part of the implementation process, we registered for the Streamlit base package and utilized its framework to develop our web application. For conversational capabilities, we integrated the OpenAI API (GPT-4/4o), enabling natural language interaction through GPT-based models. Prompt engineering techniques were used to customize the model for cervical cancer screening related topics. The prompts were developed in accordance with international and national guidelines, including the recommendations of the American Cancer Society and the World Health Organization for cervical cancer screening and the management of precancerous lesions, as well as evidence drawn from relevant peer-reviewed scientific literature. In addition, two Hungarian guidelines were incorporated, namely the Professional Protocol of the Hungarian Ministry of Health and the recommendations of the Hungarian National Public Health and Pharmaceutical Center.

3.2. Frontend

Page-level configuration, including layout and display parameters, was accomplished using a function provided by the Streamlit framework. This allowed us to define the application's title, icon, layout orientation, and initial state. Furthermore, we implemented a flexible column-based layout with dynamically adjustable widths, enabling multiple interface components to be presented side by side. A landing page was designed to serve as the entry point of the application. The application was named ScreenGPT.

3.3. User experience questionnaire

A user experience questionnaire has been put in the menu options, and those who tried the ScreenGPT web application

were able to fill out the questionnaire after that. The questionnaire was specifically developed for this study. Responses were captured using a four-point scale offering clear, interpretable results.

3.4. Evaluation of the system's responses by two independent medical doctors

To comprehensively assess the performance of the ScreenGPT system in addressing inquiries related to cervical cancer, an independent evaluation was conducted by two medical doctors. A set of thirty frequently asked questions concerning cervical cancer screening was gathered from reputable sources, including the Hungarian National Public Health Center, PennMedicine, and the Mayo Clinic. The questions were answered by the ScreenGPT system, then the responses were evaluated by the medical doctors based on 3 key criteria which were Accuracy, Clarity, and Usefulness. Each response was rated on a 4-point Likert scale.

3.5. Statistical analysis

Descriptive statistics were used to summarize the tester demographics. To examine associations between age groups and user experience variables, the Chi-Square test was applied ($p < 0.05$). Gender-related differences in user experience were assessed using the Mann–Whitney U test. To analyze the evaluations provided by the two medical doctors, both descriptive statistics and the Wilcoxon Signed-Rank Test were employed. All the statistical analyses were performed via the SPSS 29 (IBM®) software package.

4. Results

4.1. ScreenGPT system architecture

The architecture of ScreenGPT was designed to integrate a user-friendly web interface with a flexible backend and evidence-

based knowledge sources, ensuring reliable and personalized conversational support for cervical cancer screening (Figure 1.).

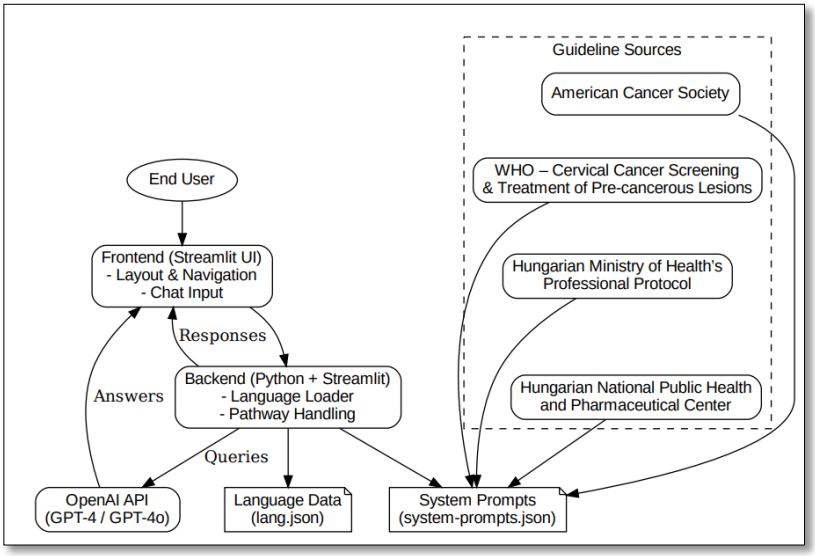

Figure 1. ScreenGPT system architecture

4.2. ScreenGPT user pathway

When entering the system, the user selects the preferred language (English or Hungarian) and then chooses a thematic topic (currently cervical cancer screening). Based on the user's profile (male, female under 25, or female 25 and older), the application directs the user to a personalized pathway defined in JSON files. The conversation progresses through guided options presented as buttons, followed by a free-text chat input.

4.3. System performance metrics

Response time was assessed using Python's datetime.now function across 11 test iterations. Each measurement captured the time interval between receiving a response from the OpenAI API and rendering it on the Streamlit user interface. Notably, a

gradual increase in average response time was observed across sequential interactions, which may be attributed to the system's need to reprocess the full conversational context at each step (Figure 2.).

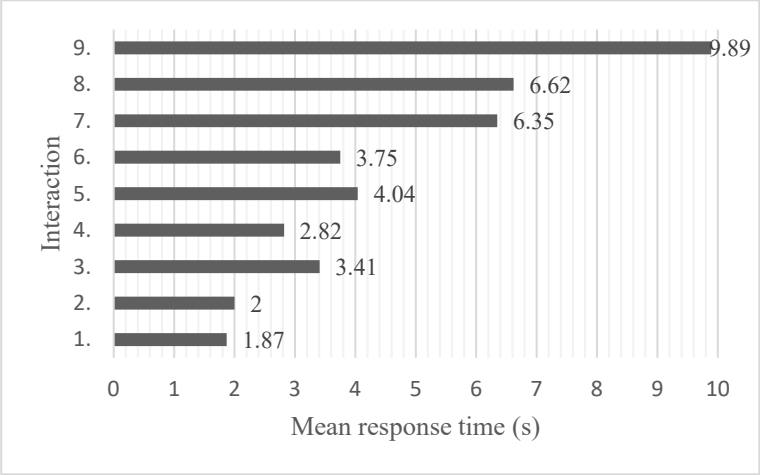

Figure 2. System response time per interaction

4.4. Results of the user experience questionnaire

Data were collected over a three-week period, during which a total of 115 individuals tested the application, comprising 87 females (76%) and 28 males (24%). The mean age of participants was 41.2 years (SD = 13.5), with ages ranging from 18 to 74 years.

Out of 115 individuals who participated in the testing of the web application, 112 completed the feedback form, offering essential data for assessing the application's performance and overall user experience.

Regarding the parameters assessed in the user experience questionnaire, we identified a statistically significant association with the Chi-Square Test ($p < 0.05$) between age groups and the

perceived personalization of the responses, with a p-value of 0.047 (Table 1.).

			Personalization of advice		
Age groups	**N**	**%**	**Mean**	**SD**	**P value**
18-24	22	19.64%	2.18	0.64	0.047
25-34	21	18.75%	2.25	1.00	
35-44	25	22.32%	2.20	0.83	
45-54	28	25.00%	2.17	0.83	
55+	16	14.29%	2.55	0.93	

Table 1. Personalization of advice with Chi-Square Test

We observed a significant association ($p = 0.037$) with Mann–Whitney U test between gender and the perceived customization of the system. Furthermore, an almost significant ($p = 0.069$) association was noted for the variable "intention for future reuse," though the p-value did not reach the conventional threshold for statistical significance (Table 2.).

	Gender	**N**	**Mean**	**SD**	**P value**
Customization of the system	Male	35	2.51	0.66	0.037
	Female	77	2.22	0.64	
	Total	112			
Future reuse	Male	35	2.86	0.73	0.069
	Female	77	2.57	0.83	
	Total	112			

Table 2. Customization of the system and Future reuse with Mann–Whitney U test

Analyzing further the data, we noticed that younger participants, mean age of under 40 years, generally reported a higher level of engagement with the interface, describing it as more interactive and user-friendly compared to testers over 40 years (Figure 3). This trend may be attributed to younger individuals' greater familiarity with digital platforms and their comfort in navigating AI-driven tools. In contrast, older users, while still able to use the application, often perceived it as less intuitive.

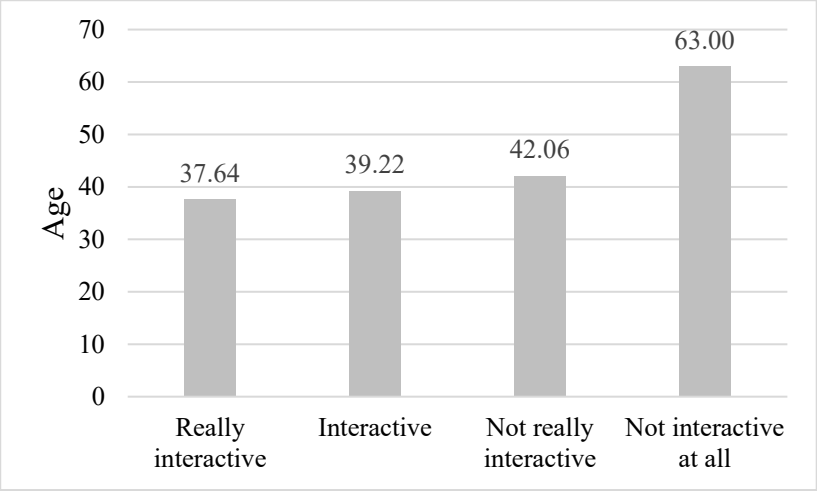

Figure 3. Web Application Interactivity Perceived by Different Age Groups

It is also notable that both male and female testers reported that the information provided by the system was highly relevant and informative (Figure 4.). Female participants emphasized the

importance of information related to screening intervals and early detection. Male participants highlighted the usefulness of content related to HPV vaccination and its role in cancer prevention.

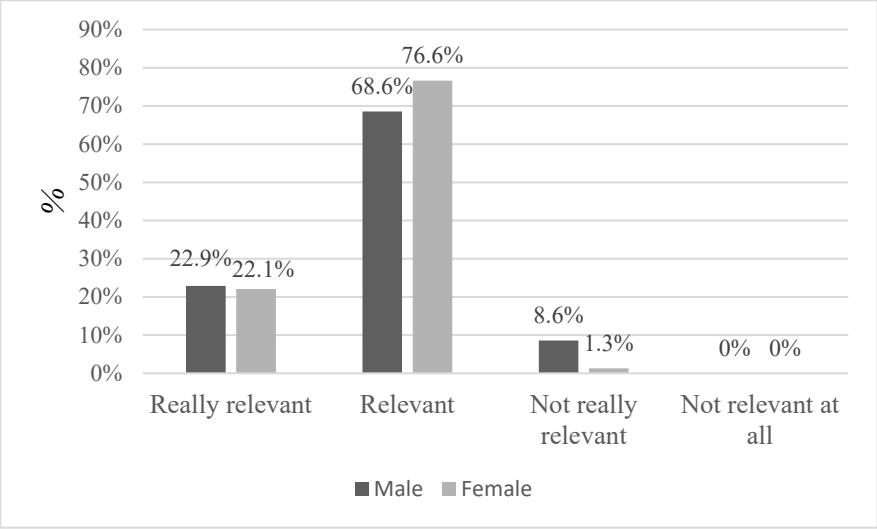

Figure 4. Gender-Based Perceptions of Information Relevance Provided by ScreenGPT

4.5. Results of the system's response evaluation by medical doctors

For each evaluation dimension (Accuracy, Clarity, and Usefulness), we computed the mean score and standard deviation across all questions. The mean scores were high for both MD1 and MD2 (all > 3.5), and the standard deviations were low. Due to the extremely low variability in the ratings, it was

not possible to compute a reliable correlation coefficient. Specifically, the Intraclass Correlation Coefficient (ICC) could not be calculated. Consequently, we employed the Wilcoxon Signed Ranks Test, which is designed to assess whether the median difference between two paired sets of observations is statistically indistinguishable from zero. The lack of statistically significant differences in Accuracy, Clarity, and Usefulness indicates that the responses evaluated by MD1 and MD2 are essentially equivalent. This indicates that the evaluations are consistent, as both sets of answers received very similar scores, demonstrating reliability and consistency in the system's responses (Table 3.).

	MD1		MD2		
	Mean	**SD**	**Mean**	**SD**	**p-values**
Accuracy	3.83	0.38	3.57	0.50	0.059
Clarity	3.80	0.55	4.00	0	0.063
Usefulness	3.97	0.18	4	0	0.317

Table 3. Independent Evaluations by Medical Doctors

5. Conclusions

The development of ScreenGPT, an AI-driven web application designed to support and strengthen public health efforts in cervical cancer prevention, represents an important step toward the integration of LLMs into digital health interventions. Building upon the capabilities of GPT-4 and GPT-4o, the system was deployed through Streamlit, ensuring accessibility and ease of use for both healthcare providers and the general public. Our evaluation encompassed technical validation, structured beta testing with potential users, and external review by two independent medical doctors. Collectively, these assessments highlight both the promise and the limitations of the current

prototype. A key contribution of this work lies in demonstrating the feasibility of adapting cutting-edge LLMs to a focused public health challenge. Unlike many generic AI health chatbots, ScreenGPT was purpose-built to promote cervical cancer awareness, risk assessment, and preventive behaviors.

From a technical standpoint, the system demonstrated stable functionality and consistent response generation across a range of test queries. The deployment on Streamlit proved to be a practical choice, offering rapid prototyping capabilities and straightforward accessibility without the need for complex installations. Importantly, the system's reliance on GPT-4 and GPT-4o models enabled ScreenGPT with contextual reasoning and natural language generation abilities that far exceed traditional rule-based or static FAQ-style systems.

This study demonstrated that an AI-based conversation agent, tailored specifically to cervical cancer screening, can enhance the accessibility and comprehensibility of clinical guidelines, thereby support patient education and encouraging user engagement with personalized, guideline-based health information. Findings from the user experience questionnaire provided empirical support for this hypothesis, with participants across genders perceiving the information as relevant and informative. Age and gender were significantly associated with perceptions of personalization and customization, while younger participants, though not at a statistically significant level, tended to report greater interactivity and accessibility. These insights highlight the importance of age-sensitive design considerations in the development of digital health interventions.

The study sample, predominantly aged 25–54, closely aligns with the WHO-recommended target group for cervical cancer screening, further justifying the choice to focus the system on this domain. By addressing a critical gap in patient awareness and health literacy, ScreenGPT distinguishes itself from prior

AI-based approaches, which have largely concentrated on image-based diagnostics, by instead offering a patient-centered educational tool.

The findings are consistent with the broader literature, which supports the effectiveness of conversational AI agents in simplifying complex medical content, enhancing health literacy, and promoting health behavior change. Nonetheless, the evidence base remains limited, with many studies, like the present one, constrained by sample size, methodological variation, and limited generalizability. These challenges underscore the need for rigorous future research, particularly randomized controlled trials, to establish the clinical utility of LLM-based applications with greater confidence.

ScreenGPT may serve as valuable entry points for enhancing awareness of cervical cancer screening and supporting women in preparing for informed decision-making. However, these tools should not be regarded as standalone or authoritative sources of information, as they may omit critical content or contextual nuances. Instead, ScreenGPT and similar applications are best positioned as adjuncts to, rather than replacements for evidence-based guidance and counseling provided by qualified health care professionals.

Future evaluations should expand to encompass cost-effectiveness, privacy, security, and ethical considerations, which will be critical to the responsible integration of LLM-driven tools into healthcare systems. As such, technologies become increasingly embedded in both clinical practice and public health, ensuring equitable, transparent, and safe implementation will be essential for maximizing their potential to improve patient outcomes and reduce barriers to life-saving preventive care.

6. Bibliography of the candidate's publications

Publications related to the thesis:

Angyal Viola, Bertalan Ádám, Domján Péter, Feith Helga Judit, Dinya Elek
Development of a Questionnaire for Assessing the Use of ChatGPT in Primary and Secondary
Disease Prevention
FRONTIERS IN PUBLIC HEALTH 13 Paper: 1709611, 10 p. (2026)
Közlemény: 36831463 | Szakcikk (Folyóiratcikk) | Tudományos
Scopus - Public Health, Environmental and Occupational Health SJR indikátor: Q1
IF: 3,2

Angyal Viola, Bertalan Ádám, Domján Péter, Dinya Elek
Exploring the possibilities and limitations of customized large language model to support and
improve cervical cancer screening
BMC MEDICAL INFORMATICS AND DECISION MAKING 25: 1 Paper: 242, 16
p. (2025)
Közlemény: 36226867 | Szakcikk (Folyóiratcikk) | Tudományos
Scopus - Computer Science Applications SJR indikátor: Q1
Scopus - Health Policy SJR indikátor: Q1
Scopus - Health Informatics SJR indikátor: Q2
IF: 3,8

Bertalan Ádám, **Angyal Viola**, Domján Péter, Sædder Eva Aggerholm, Király Gyula, Erdélyi Lóránd, Gyimesi Nóra, Dinya Elek
Development of an Artificial Intelligence Powered Medication Risk Score Calculator
Application
BASIC & CLINICAL PHARMACOLOGY & TOXICOLOGY 137: 4 Paper: e70109, 15 p. (2025)
Közlemény: 36346440 | Szakcikk (Folyóiratcikk) | Tudományos
Scopus - Medicine (miscellaneous) SJR indikátor: Q2
Scopus - Pharmacology SJR indikátor: Q2
Scopus - Toxicology SJR indikátor: Q2
IF: 3,3

Angyal Viola, Bertalan Ádám, Dinya Elek
A ScreenGPT rendszer megbízhatóságának vizsgálata méhnyakrák prevencióval kapcsolatos információátadás támogatásában
In: V. Életmódorvosi és II. Mozgás Receptre Konferencia Magyar Életmód Orvostani Társaság (2025) 102 p. pp. 6-7.
Közlemény: 36381127 | Absztrakt | Kivonat (Egyéb konferenciaközlemény) | Tudományos

Kőrös-Angyal Viola, Bertalan Ádám, Domján Péter, Dinya Elek
CN10 Evaluating prompt engineering techniques for artificial intelligence-generated cervical cancer screening information: A comparative content analysis
ANNALS OF ONCOLOGY 36: Suppl. 2 pp. S1518-S1519. (2025)
Közlemény: 36704634 | Absztrakt | Kivonat (Folyóiratcikk) | Tudományos

Angyal Viola, Bertalan Ádám, Domján Péter, Dinya Elek
ScreenGPT – A mesterséges intelligencia alkalmazásának lehetőségei és korlátai a primer, szekunder és tercier prevencióban [ScreenGPT-The opportunities and limitations of artificial intelligence in primary, secondary and tertiary prevention] **ORVOSI HETILAP** 165: 16 pp. 629-635. (2024)
Közlemény: 34817394 | Szakcikk (Folyóiratcikk) | Tudományos
Scopus - Medicine (miscellaneous) SJR indikátor: Q4 IF: 0,9

Angyal Viola, Dinya Elek
Mesterséges intelligencia alapú GPT-modellek felhasználásának lehetőségei és korlátai egészségügyi alkalmazások fejlesztésében
IME 23: KSZ 1. pp. 23-29. (2024)
Közlemény: 34854971 | Szakcikk (Folyóiratcikk) | Tudományos

Publications not related to the thesis:

Domján Péter, **Angyal Viola**, Vingender István
Concentration and geospatial modelling of Health Development Offices' accessibility for the total and elderly populations in Hungary BMC PUBLIC HEALTH 25 Paper: 1466, 16 p. (2025)
Közlemény: 36115753 | Szakcikk (Folyóiratcikk) | Tudományos
Scopus - Public Health, Environmental and Occupational Health SJR indikátor: Q1
IF: 3,6

Domján P, **Angyal VZs**, Vingender I
The role and potential of Health Development Offices (HDOs) in maintaining the health of older people in Hungary
DEVELOPMENTS IN HEALTH SCIENCES 7: 2 pp. 38-45. (2024) Közlemény: 35623540 | Szakcikk (Folyóiratcikk) | Tudományos

Angyal Viola Zsuzsanna
Homoktövis (Hippophae rhamnoides L.) antioxidáns kapacitásának és antibakteriális hatásának vizsgálata
In: Áldorfai György; Tóth Tamás (szerk.) "SZIE kiváló tehetségei" konferencia előadásainak összefoglaló kiadványa
Gödöllő, Magyarország: Szent István Egyetemi Kiadó (2018) 157 p. pp. 9-9.
Közlemény: 3344984 | Absztrakt | Kivonat (Könyvrészlet) | Tudományos

Angyal Viola Zsuzsanna, Kovács Mónika, Stefanovitsné Bányai Éva
Homoktövis (Hippophae rhamnoides L.) antioxidáns kapacitásának és antibakteriális hatásának vizsgálata

In: Fodor Marietta; Bodor Péter (szerk.) ITT - Ifjú Tehetségek Találkozója Budapest, Magyarország: Szent István Egyetem, Budai Karok (volt KÉE) (2017) 162 p. pp. 72-75. Közlemény: 3303936 | Konferenciaközlemény (Könyvrészlet) | Tudományos

∑IF: 14,8