GigaPevt: Multimodal Medical Assistant

Pavel Blinov[1] , **Konstantin Egorov**[1] , **Ivan Sviridov**[1] , **Nikolay Ivanov**[1] , **Stepan Botman**[1] , **Evgeniy Tagin**[1] , **Stepan Kudin**[1] , **Galina Zubkova**[1] and **Andrey V. Savchenko**[1,2]
[1]Sber AI Lab, Moscow, Russia
[2]HSE University, Nizhny Novgorod, Russia
{Blinov.P.D, Egorov.K.Ser, IAnatoSviridov, NikIvanov, SABotman, EyTagin, SSKudin, GVZubkova, AVladSavchenko}@sber.ru,

Abstract

Building an intelligent and efficient medical assistant is still a challenging AI problem. The major limitation comes from the data modality scarceness, which reduces comprehensive patient perception. This demo paper presents GigaPevt, the first multimodal medical assistant that combines the dialog capabilities of large language models with specialized medical models. Such an approach shows immediate advantages in dialog quality and metric performance, with a 1.18% accuracy improvement in the question-answering task.

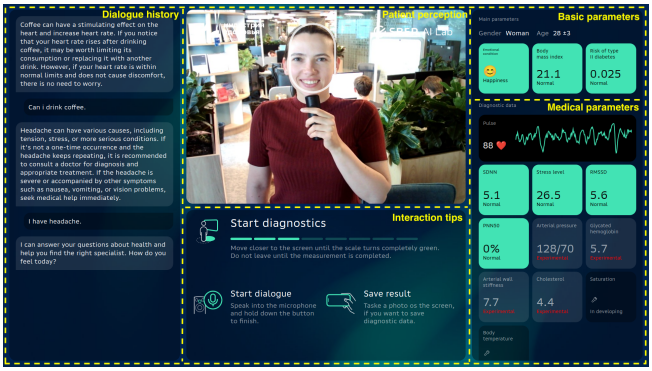

Figure 1: GigaPevt UI (English version for the demonstration)

1 Introduction

The availability of high-quality healthcare remains a significant social problem due to high-cost issues, shortage of medical staff, or inaccessibility in far-end regions. From a technological point of view, one possible solution lies in developing intelligent medical assistants. There are many attempts to build such an assistant in the form of a symptom checker application (Ada Health, K Health, Symptomate, Babylon Health) [Wallace *et al.*, 2022]. The major limitation of such systems is using the only text modality, where patients describe the symptoms and get answers in the chat. The system misses a significant part of the non-verbal doctor-patient communication context in that case. Another flaw of such AI assistants is the design in a survey systems paradigm with predefined questions that miss the human touch in the communication process.

However, the rapid development of large language models (LLMs) opens new ways to create next-generation assistants in healthcare. In this paper, we introduce and demonstrate[1] an innovative multimodal medical assistant called GigaPevt (Figure 1).[2] GigaPevt combines specialized medical models with the rich dialog capabilities of LLM, which account for broad context in visual, audio, and text modalities. That allows us to mitigate the problems mentioned above, significantly improves patient interaction experience, and boosts the quality of dialog.

2 GigaPevt Architecture

Given the assortment of potential deployment scenarios, we implemented GigaPevt as a client-server application, as shown in Figure 2. We implement a rich client written in Python to run lightweight models that require low latency for a comfortable user experience. Also, we use API services for face detection, Automatic Speech Recognition (ASR) for communication and Text-to-Speech (TTS) for providing voiceover for text responses[3]. Client logic handles data flow, client state management, and network interactions with the rest of the system.

We use Flask as a back-end framework that handles the server side. The server has a model manager, specialized models, and the GP Dialog Logic component responsible for the patient-assistant interaction.

3 Specialized Models

3.1 Video-Based Facial Analytics

We analyze facial videos in a frame-wise way to identify a user and predict the attributes list. The facial region is detected using the MediaPipe library [Lugaresi *et al.*, 2019],

[1]https://youtu.be/73fkCQeEqaQ

[2]The *Giga*Pevt name is a wordplay in many facets. First, the *GP* stands for General Practitioner. Second, in Russian, the word consonant to a therapist with the smaller metric prefix *Giga-* instead of *Tera-*. Third, it references the core Russian LLM *Giga*Chat.

[3]https://developers.sber.ru/docs/ru/salutespeech/category-overview

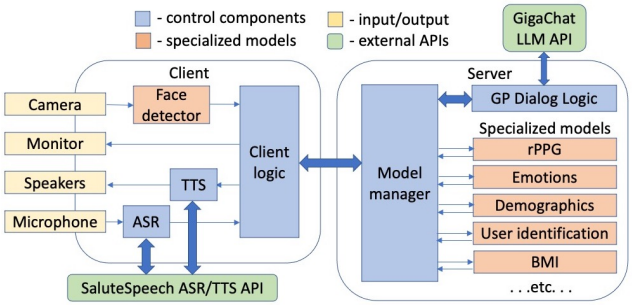

Figure 2: GigaPevt architecture

and the obtained facial image is fed into the corresponding neural network.

User Identification

We apply open-source face recognition model to identify, authorize, and manage the user session. The facial image is resized to 224x224 and fed into the fast OFAMobileNetV3 model [Makarov *et al.*, 2023] to extract facial descriptors. The latter model was pre-trained on the part of the VGGFace2 dataset [Cao *et al.*, 2018]. The training dataset contains feature vectors of known subjects obtained by the same procedure given their facial photos. Extracted facial embeddings are normalized using L_2 norm and matched with a training dataset by the k-NN with the Euclidean distance. In the experimental study, we gathered a testing dataset with 426 photos of 104 persons. As a result, the recognition accuracy equals 98.19%.

Socio-Demographic Model

We used a high-performance MobileNet-V1 model to predict age, gender, and ethnicity [Savchenko, 2021]. It was pre-trained to recognize faces from the VGGFace2 dataset [Cao *et al.*, 2018]. Unfortunately, the age groups in this dataset are very imbalanced, so the trained models work incorrectly for the faces of very young or old people. Hence, we decided to add all (15K) images from the Adience dataset [Eidinger *et al.*, 2014]. As the latter contains only age intervals, e.g., (60-100), we put all images from this interval to the average age, i.e., 80. The ethnicity classifier was trained on the subset of the UTKFace dataset [Zhang *et al.*, 2017] with different class weights to achieve better performance for imbalanced classes. As a result, we obtained gender classification accuracy (93.79%), which is 2.7% greater than the best-known accuracy of a much larger VGG-16 neural network in the DEX method [Rothe *et al.*, 2015]. The age prediction MAE (Mean Absolute Error) is equal to 5.74 years, which is 0.7 years better when compared to the DEX. Finally, the accuracy of ethnicity prediction is equal to 87.6%, which is on par with the best-known models [Savchenko, 2021].

Facial Expression Recognition

The EfficientNet-B0 model was used to predict facial emotions [Savchenko, 2022]. In particular, we pre-trained it to identify faces on the VGGFace2 dataset as in the previous subsection. Next, we add three heads for predicting eight basic expressions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise), valence, and arousal using the training part of the AffectNet dataset [Mollahosseini *et al.*, 2017]. In contrast to typical face recognition training, we use highly cropped faces in both training stages to remove hair, background, etc., and concentrate on facial expressions only. Thus, we obtained competitive results on the validation part of the AffectNet dataset. Our accuracy for eight classes equals 61.93%, 0.5-2% greater than that of models of similar size and efficiency [Pourmirzaei *et al.*, 2021; Savchenko, 2021].

Body Mass Index Model

The Body Mass Index (BMI) estimation is addressed as an image regression problem. The dataset for this task includes samples (face-BMI value) from Reddit,[4] a public social web platform enriched with the FIW-BMI dataset [Jiang *et al.*, 2019]. The detected faces are resized to 256×256 and fed into the ResNet34 model [He *et al.*, 2016] as a backbone with MSE as a metric and loss function. On the test set, we get the MSE equal to 4 units.

rPPG Model

The remote photoplethysmography (rPPG) model allows extracting a photoplethysmography signal using a video of a user's face. Our system used the Plane Orthogonal-to-Skin (POS) algorithm described in [Wang *et al.*, 2017]. This algorithm is the strong baseline in the field; it adopts signal processing techniques on subtle changes of the RGB values in the time domain. The algorithm runs based on a face landmark detection system, as it relies on averaging pixel values change over face regions of interest (cheeks, nose, and forehead), which are the most valuable for pulse wave detection [Wang *et al.*, 2018]. According to [Liu *et al.*, 2024], this method has high accuracy, robustness, and speed. Calculating rPPG also allows us to infer several derived parameters of the state of the cardiovascular system, such as heart rate, RR intervals (intervals between consecutive heartbeats), stress index, PNN50, and many others. All these health indicators allow GigaPevt to determine the user's condition more accurately and comprehensively, improving the patient interaction experience.

3.2 GP Dialog Logic

To manage and control the dialog, we used GigaChat Pro model[5]. As the assistant continuously receives a stream of additional parameters from sensors, we can substantially enrich the prompts. Even the simple inclusion of such augmented parameters in a text prompt is helpful. But our ultimate goal is to obtain controllable, high-quality dialog focused on the anamnesis collection. To achieve this goal, we resorted to the Retrieval-Augmented Generation (RAG) technique [Lewis *et al.*, 2020] as it can help supplement the user's request with the found relevant information and the Chain of Thoughts (CoT) [Wei *et al.*, 2022] as it was successfully used in challenging NLP scenarios [Li *et al.*, 2023]. We started to assemble a database of medical dialogues and questions. These improvements substantially boost the quality of answers and performance in downstream health-related tasks.

[4]https://www.reddit.com/r/progresspics

[5]https://developers.sber.ru/portal/products/gigachat

Question	GigaChat answer	LLaMA answer	GigaPevt answer
Is my heart rate comfortable for exercise?	To determine a comfortable heart rate when exercising, you need to consider your age and level of physical fitness. Generally, for people aged 20 to 59 years, exercising at a heart rate between 50% and 85% of your maximum heart rate is recommended.	The normal heart rate for an adult is between 60 and 100 beats per minute. If your heart rate is within this range, has a regular rhythm, and you do not feel discomfort during exercise, it may be comfortable for you. However, consult your doctor for testing and advice if you have any doubts.	No, your heart rate is currently not comfortable for exercise. Your heart rate is 115 beats per minute, which is higher than normal. For comfortable physical exercise, it is recommended to have a heart rate ranging from 60 to 100 beats per minute. I suggest you consult a doctor to determine the optimal level of physical activity for you.
What do you think? Am I obese?	To determine if you are obese, you need to know your weight and height. Obesity is defined as excess accumulation of fat, which can lead to various diseases and poor quality of life.	As an artificial intelligence model, I cannot determine your weight or estimate your body. However, if you suspect you are obese, it is best to see your doctor to determine your body mass index (BMI) and get advice. The most important thing is not to put off visiting your doctor if you have concerns about your health.	Based on your body mass index, which is within the normal range, you are not obese. However, if you have other signs or symptoms, such as fainting, dizziness, or shortness of breath, it is recommended that you consult your doctor.

Table 1: Examples of answers to health-related questions (translated from Russian)

Model	RuMedDaNet	RuMedNLI
GigaPevt	**93.36**	71.59
Human	92.97	**85.67**
GigaChat	92.18	71.21
ChatGPT-3.5	89.26	61.33

Table 2: LLMs performance metrics (Accuracy, %). The best results are shown in bold.

4 Experiments

For a diverse assessment of the GigaPevt assistant, we evaluate it from quantitative and qualitative sides.

4.1 Performance Evaluation

We evaluate the proposed assistant on the two most relevant downstream tasks from the RuMedBench [Blinov *et al.*, 2022]. The RuMedDaNet is a binary question-answering task with 512 test medical-related questions in Russian. As shown in Table 2, the GigaPevt model with the RAG technique substantially outperforms GigaChat and Human assessment by 1.18% and 0.39%, respectively. It is the most indicative task for assessing the increase in the quality of dialogue because it is close to our assistant's target scenario.

The RuMedNLI is a task of natural language inference with 1536 Russian text pairs to decide whether the second text in a pair is entailment, contradiction, or neutral to the first. For this task, the performance gain is positive but relatively moderate (0.38%).

4.2 Dialog Showcases

Table 1 lists several examples to assess the qualitative gap in the answers between the GigaChat model, open LLaMA [Touvron *et al.*, 2023] and GigaPevt. GigaChat and LLaMA answers are the base model answers in a zero-shot mode without additional information. GigaPevt answers is the model with visual input, RAG, and CoT for enriching our prompts. Tuned prompts have the following form: constitution (description of how the model should behave and its role) + visual information about a person (pulse, body mass index, etc.) + postfix (recommendations for model).

We can see that the model with prompt tuning, RAG, visual information from the camera, etc., allows us to answer questions that require a visual inspection immediately. Due to generally more awareness about the patient, our assistant gives more specific and concrete answers.

5 Conclusion and Future Work

This demo aims to inspire multimodality research in the medical domain. We demonstrate the prototype of medical assistant technology, which processes a rich multimodal context. Several experiments confirm assistant effectiveness.

We plan to use more advanced knowledge management techniques for LLMs, such as Advanced and Modular RAG [Gao *et al.*, 2023], to make our model more complete and focused on the medical domain. In addition to working with prompt enrichment, our strategy includes incorporating direct engagement with visual modality, such as utilizing images to summarize patient complaints that may be challenging to discern solely through textual descriptions. Also, our plan includes closer integration with a patient electronic health record for better context awareness and the ability to target more specific questions. Finally, it is crucial to perform voice analysis to recognize emotions, sensitively understand a patient, promptly manage the dialog, etc.

References

[Blinov *et al.*, 2022] Pavel Blinov, Arina Reshetnikova, Aleksandr Nesterov, Galina Zubkova, and Vladimir Kokh. Rumedbench: a Russian medical language understanding benchmark. In *International Conference on Artificial Intelligence in Medicine*, pages 383–392. Springer, 2022.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[Eidinger *et al.*, 2014] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[Jiang *et al.*, 2019] Min Jiang, Yuanyuan Shang, and Guodong Guo. On visual BMI analysis from facial images. *Image and Vision Computing*, 89:183–196, 2019.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[Li *et al.*, 2023] Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator. *arXiv preprint arXiv:2312.04474*, 2023.

[Liu *et al.*, 2024] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rPPG-Toolbox: Deep remote PPG toolbox. *Advances in Neural Information Processing Systems*, 36, 2024.

[Lugaresi *et al.*, 2019] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Proceedings of the Third workshop on computer vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[Makarov *et al.*, 2023] Ilya Makarov, Andrey V. Savchenko, and Lyudmila V. Savchenko. Fast search of face recognition model for a mobile device based on neural architecture comparator. *IEEE Access*, 11:65977–65990, 2023.

[Mollahosseini *et al.*, 2017] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[Pourmirzaei *et al.*, 2021] Mahdi Pourmirzaei, Gholam Ali Montazer, and Farzaneh Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv preprint arXiv:2105.06421*, 2021.

[Rothe *et al.*, 2015] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a single image. In *Proceedings of the International Conference on Computer Vision (CVPR) Workshops*, pages 10–15. IEEE, 2015.

[Savchenko, 2021] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *Proceedings of the 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021.

[Savchenko, 2022] Andrey V Savchenko. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of European Conference on Computer Vision (ECCV) Workshops*, pages 45–59. Springer, 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv.2302.13971*, 2023.

[Wallace *et al.*, 2022] William Wallace, Calvin Chan, Swathikan Chidambaram, Lydia Hanna, Fahad Mujtaba Iqbal, Amish Acharya, Pasha Normahani, Hutan Ashrafian, Sheraz R Markar, Viknesh Sounderajah, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ Digital Medicine*, 5(1):118, 2022.

[Wang *et al.*, 2017] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

[Wang *et al.*, 2018] Chen Wang, Thierry Pun, and Guillaume Chanel. A comparative survey of methods for remote heart rate detection from frontal face videos. *Frontiers in Bioengineering and Biotechnology*, 6:33, 05 2018.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5818. IEEE, 2017.