ASSESSMENT OF CONVERSATIONAL CHATBOTS FOR DRIVING SUPPORT APPLICATIONS

J. Novotný[*], *F. Kotas*[*]

Abstract: The study evaluated chatbots designed as conversational assistants for drivers with the aim of reducing driver fatigue through appropriate and undemanding conversation. The introductory part included a questionnaire survey focused on identifying preferred topics of conversation while driving, which provided a basis for selecting the content focus of the experiment. This was followed by a subjective evaluation of selected chatbots using user experience metrics, focusing on their comprehensibility, naturalness, and ability to maintain attention without increasing mental load. The final part used the QFD method to assess the extent to which individual chatbot procedures and features met the set criteria. The output was a comparison of chatbots and a determination of their suitability for further extensive experiments.

Key words: *artificial intelligence, chatbot, driver fatigue, QFD*

Received: January 16, 2025 **DOI:** 10.14311/NNW.2025.35.001
Revised and accepted: February 24, 2025

1. Introduction

Artificial intelligence, and specifically advanced chatbots, are a dynamically developing field. In the automotive industry, chatbots are beginning to be applied in a number of areas. They can also play an important role directly in the vehicle while driving in the form of conversations with the driver.

The use of chatbots could have an impact on driver performance and especially driver fatigue. That could have a significantly positive impact on the level of safety on the roads. It remains a fact that fatigue is responsible for around 20% of serious accidents worldwide [1]. Monotonous driving increases the risk, and the emergence of partially autonomous driving, meanwhile, increases the risk even more by placing lower demands on the driver.

Studies have shown the positive effect of driver conversation on the suppression of fatigue symptoms. Therefore, the application of advanced voice assistants (chatbots) is proposed as a preventive measure, especially for drivers travelling alone.

[*]Jan Novotný – Corresponding author; Filip Kotas; Czech technical University in Prague, Faculty of Transportation Sciences, Department of Vehicles, Horská 3, 128 08 Prague 02, Czech Republic, E-mail: jan.novotny@cvut.cz, kotasfil@cvut.cz

For this purpose, an experiment comparing the current widely used chatbots in terms of car-related conversational topics was designed, using a high-fidelity simulator to create a sense of realism. A survey about conversation topics while driving was done in order to determine relevant topics for experimental conversations.

In the experiment, participants tested selected chatbots and rated them based on several UX criteria. According to the results, the most suitable chatbot were selected for its application in a vehicle simulator for further studies and experiments in this research area.

2. Related Work

2.1 Effect of Communication on Driving Performance and Fatigue

Cognitive workload is the specific level of mental effort required to perform or think through given activities. If too much cognitive load is placed on the driver, whether due to secondary activities such as operating a mobile phone or infotainment system, or due to a crying child or an argument, there is a risk of driver overload, inattention, and increased risk of an accident.

On the other hand, if the cognitive load on the driver is minimal, driving is monotonous and not demanding, there is also a risk of inattention, as well as drowsiness, microsleeps and even falling asleep. Thus, again, there is an increased risk of an accident.

Therefore, when applying voice assistant (or any conversation) in current vehicles, it is necessary to balance on a fine line in terms of what is a beneficial distraction for the driver and an aid against fatigue, and what is already an excessive distraction that distracts from driving. Driver distraction is another concept that is very important in terms of safe driving.

The situation in which the voice assistant would communicate with the driver is also a factor. In more demanding driving situations with increased cognitive load, any voice system should create a pause in the conversation and let the driver concentrate on the task at hand [2]. In their study, Lindstrom et al. also address dialogue adaptation based on the current cognitive load. When observing driver-passenger dialogues, as cognitive load increases, the flow of the conversation systematically changes and disrupts, indicating the need to interrupt the dialogue due to the need to focus [3]. Similarly, Lundholm Fors et al. document prolonged pauses in dialogue during demanding driver tasks. Furthermore, when changing the topic, the co-driver looks at the situation and the driver's workload. This means that voice systems should be able to adapt the length of pauses as well as the appropriateness of the start of the conversation [2].

Another experiment showed, using the DALI and SASSI questionnaires, that the pro-active voice assistant is perceived similarly positively by participants as the passive voice assistant in terms of cognitive load and likability. This means that a more active system, developing the conversation on its own, is as well accepted by users as more conservative assistants [4].

2.2 Driver Fatigue

There are a number of systems in vehicles that attempt to detect driver fatigue, either passively (by analyzing changes in driving behavior) or actively (by using cameras to monitor the driver's facial features and eyes). From the perspective of driver condition monitoring research, heart electrical signal monitoring electrocardiogram (ECG) and brain activity monitoring electroencephalogram (EEG) may also be used. However, application in a real vehicle is problematic. EEG is accompanied by uncomfortable wearing, motion artifacts, and inaccuracies in data interpretation [5, 6]. Direct ECG monitoring is also unsuitable due to the need to apply sensors to the head, so wearable electronics may be an option.

The current state of camera monitoring is also unsatisfactory, as manufacturers are only just beginning to deploy it [7]. Higher-level autonomous driving will also not be deployable for the foreseeable future. This is also only possible in certain environments, states, and types of communication in terms of infrastructure [8]. On the contrary, only partial vehicle automation is even more dangerous in terms of the driver's attention and fatigue. These are all arguments as to why the driver's own perception of fatigue is still very important.

The perception of fatigue is very individual, but accident statistics indicate that drivers' judgment is not very reliable. Drivers defend themselves against fatigue in a variety of ways, which vary in their effectiveness. The most appropriate method is regular breaks in driving and sufficient hydration. However, car and truck drivers often combat fatigue with other methods.

According to a study by Vanlaar et al. involving 750 respondents from the province of Ontario, 34.2% of respondents chose talking to a passenger as a measure to combat fatigue. Only an open window or a fan running strongly achieved a higher percentage of responses (43.7%) [9].

According to a study by Jellentrup et al., mobile telephony is also effective in delaying fatigue during monotonous driving. According to the measurements, drivers were more alert and awake for up to 20 minutes longer than in the reference trips [10]. However, talking on the phone while driving has several negative effects, including increased distraction [11, 12]. A conversation with a voice assistant can have positive effects on driver fatigue and driving performance, especially when using partial automated driving [13, 14].

The studies mentioned above confirm the positive effect of driver talking as a measure against fatigue symptoms. Whether it is talking to a passenger, talking on the phone, or even talking to themselves, the effectiveness against fatigue is both subjectively and objectively high. Therefore, voice assistants have a high potential to have a positive effect in combating fatigue symptoms, especially for drivers travelling alone.

A driver without a passenger on a monotonous journey is more likely to succumb to fatigue and have an accident [15,16]. Therefore, interacting with a voice assistant can replace interaction with a passenger and thus keep the driver's cognitive load above a dangerously low level.

2.3 Survey about Conversation Topics while Driving

A questionnaire was created to determine the relevant topics of conversation when driving a car. What information would the driver like to be informed about while driving. A total of 50 respondents completed the questionnaire. The average age of the respondent was 41 years. The primary question was which of the following categories would they prefer as content for a spontaneous voice message from the navigation system?

The most preferred is the content around traffic and weather situations, as seen in Fig. 1. Furthermore, vehicle information is relevant. Another interesting finding was the preferred length of a voice message – the most preferred was 3 statements, with a length approximately 10 seconds. Obviously, it depends on the driving situation, but this is the average preference.

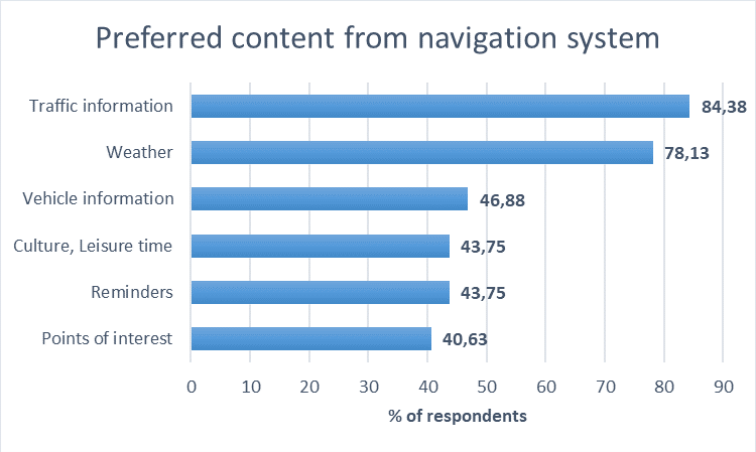

Fig. 1 *Preferred content from navigation system.*

Another issue was the preference for topics for open discussion during the drive. Two types of conversation behavior can be observed – giving commands and asking questions. There are either request for information about weather, navigation, requests to play a song, or setting up a climacontrol, etc. Or people want to use the assistant to get information, topics of their interest – for example, discussions about movies, food, sports, and hobbies, etc, as seen in Fig. 2. Currently, mainly "instruction" chatbots function in cars. In this research, we focus on the possible use of a full conversational AI chatbot to enable a full flow of conversation.

2.4 Advanced NLP Chatbots

A chatbot is a computer program or application designed to interact with people through text or voice communication. It is programmed to simulate a conversation and provide automated responses to the user's questions or requests.

Nowadays, we commonly use text-based or voice-based bots to retrieve information. The most common ones include Google Assistant, Apple's Siri, or the voice assistants provided by individual car brands. These include BMW Intelligent

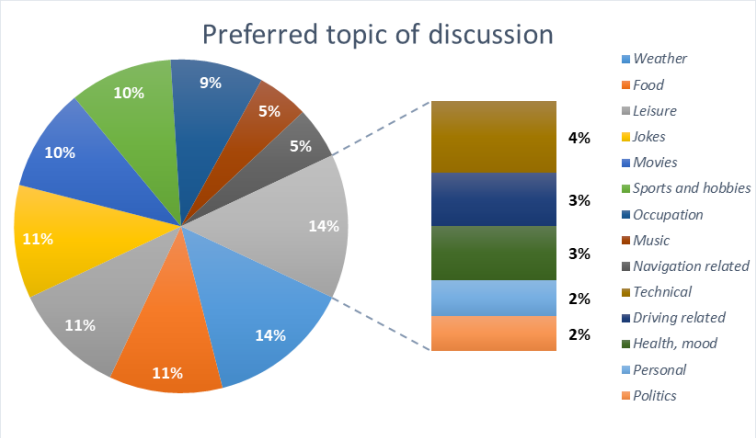

Fig. 2 *Preferred topic of discussion.*

Personal Assistant, Mercedes-Benz MBUX, and Skoda Laura. However, these assistants are generally seen as tools focused on performing short tasks or commands, rather than engaging in full conversations.

For deeper conversations, discussions, or more complex interactions, chatbots rely on artificial intelligence (AI) and machine learning, along with natural language processing (NLP) algorithms. These algorithms then allow computers to fully understand and process the written or spoken word [17].

The most widely used technology for such tasks today is the large language model (LLM), which is a powerful AI model trained on massive text datasets to understand and generate human-like language. This enormous amount of data was used to form a "deep learning neural network," which is a complex, hierarchical neural network of algorithms inspired by the human brain. Neural networks work by predicting the next part of the text. The text is converted, divided into numbers, called tokens. The subsequent transformation layers include "self-attention", which is a mathematical expression (Eq. 1) of which previous words are important for the subsequent context [18].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q are queries, K are keys, V are values, and d_k is the dimension (size) vector.

Following the attention mechanism, the signal passes through small, fully connected neural networks to help the model construct richer internal representations.

$$\text{FFN}(x) = W_2 f\left(W_1 x + b_1\right) + b_2, \tag{2}$$

where FFN is a feed-forward network, $W_1 x + b_1$ is the first linear layer, f is the activation function, $W_2 f(\cdot) + b_2$ is the second linear layer.

Thanks to the aforementioned training, the prediction and selection of the output token occur. It excels at many tasks, such as translation, summarization, and

conversation, without the need for task-specific training. These models can also handle various data types, including images and audio [19].

However, even these advanced chatbots have limitations. For example, insufficient NLP performance may make it difficult for a chatbot to carry on a complex, multi-topic conversation. Another issue can arise with voice bots, such as when the bot does not know how long to listen to the user, which may lead to interruptions or no response at all [20].

Despite these limitations, chatbots are growing in popularity and are used across various industries, including customer service, business, healthcare, and education. Some of the most well-known chatbots today include ChatGPT, DeepSeek, Claude, Google Gemini, Microsoft Copilot, and Pi.

2.5 High-Fidelity HMI Simulators

Advanced vehicle simulators are used for experiments where it is necessary to replicate a large number of tests or perform tests that would be too dangerous in a real environment. Simulators are particularly suitable for testing driver fatigue and its consequences.

Simulators are advanced systems relying on the complex interconnection of hardware, software, and their architecture. From the perspective of HMI testing, it is important to have the correct settings and to use appropriate metrics to measure the required parameters. In terms of fatigue measurement, for example, it is possible to monitor eye parameters, heart rate, or deviation from the center of the lane [21].

An advanced vehicle simulator containing the interior of a Škoda Superb III car, with three 75-inch LCD televisions for projecting driving scenes, can be used for experiments (Fig. 3). The simulator's interior features a replica of the seats, dash-

Fig. 3 *High-fidelity Simulator in CTU Prague.*

board, speedometer, and infotainment displays found in the actual vehicle. This makes the vehicle environment and simulated driving experience highly immersive. This simulator was also used in other various research, described in detail by El Hamdani et al. [22]

3. Experimental Evaluation

3.1 Chosen Chatbots

The choice of chatbots for this experiment is based on the previous research, general popularity, and their availability. In terms of focus, we targeted chatbots that have a general aptitude for providing quality conversation on general topics related to driving in a vehicle. After considering all the facts, four chatbots were finally selected for the experiment:

- **ChatGPT-5** – Currently the best known and most used AI-based chatbot from the company OpenAI. Like many others, ChatGPT uses LLM, which is called GPT-5. In addition to generating text, it has a voice interface and can also generate images and graphs.

- **DeepSeek** – A slightly younger chatbot with its DeepSeek V3 and R1 model surprised with its problem-solving capability compared to OpenAI's older GPT3 model. Originating in China, however, it is not entirely clear how data is processed through the service, and this may raise privacy concerns.

- **Google Gemini Flash 2.5** – Developed by Google, Gemini currently runs on the Gemini 2.5 model. It has a large memory for remembering conversations for deeper contextual understanding. Gemini also integrates seamlessly with other Google apps like Gmail, Google Maps, and more, enhancing its functionality and user experience.

- **Claude** – An AI chatbot developed by Anthropic that uses the LLM Claude 3. It focuses primarily on safe, ethical, and reliable communication, and is designed to be useful for tasks such as writing, text analysis, programming, or creative content creation.

- **Pi** – "Personal intelligent" is a conversational chatbot developed by Inflection AI. It uses an advanced LLM focused on empathy, support, and natural dialogue, making it different from tech-oriented chatbots. It focuses on conducting personal conversations, promoting mental well-being, and friendly, human conversation.

Other chatbots were also considered, such as Grok and Claude, but these have a strict limit on the number of queries. Perplexity was not considered suitable for simple conversation.

3.2 Chosen Driving Topics

Based on the questionnaire statements in Section 2.3, we selected the topics that participants in the experiment would be asked to briefly converse about with the chatbots:

- Weather in general,
- Traffic information – finding and discussing routes,
- Vehicle, driving information,
- Requests to find a hotel, petrol station, shop, fast food, etc. in a certain area,
- Points of interest along the way – conversation about a certain city, castle, etc.
- Leisure – free topics.

3.3 Chosen Measurements and Criteria

Although interest in chatbots and their evaluation criteria is growing, a comprehensive and widely accepted evaluation framework is still lacking [23]. As we review various research papers, we encounter a range of criteria and metrics used to assess chatbots from different perspectives. One of the most used evaluation methods, which appears in many studies, is user experience (UX).

UX refers to how individuals perceive and interact with a product, system, or service. It goes beyond basic functionality, focusing on how practical, user-friendly, and efficient interaction feels from the user's point of view [20]. It's important to note, however, that this is a subjective form of evaluation.

For this study, we focused on evaluating conversation quality, features, and capabilities. For completeness, other relevant criteria could include aspects such as quality of voice interaction, user interface, privacy, and security. In Tab. I, we list the selected metrics and criteria, with the evaluation scale ranging from 0 to 9.

3.4 Experiment Design

During the experiment, the participant is given the task to converse with 4 chatbots and evaluate them in depth. It all takes place while driving in a high-fidelity simulator with a real vehicle interior, with the use of a microphone, loudspeaker, and voice synthesis.

First, the participant evaluates the importance of the evaluation criteria (1–10). This will answer how important the chatbot characteristics are, and also, this will be further used for QFD analysis.

The participant converses with the chatbot for approximately 10 minutes on the selected topics listed in the previous chapter. At the end of each conversation, the participant is asked about the overall impression and rates the listed criteria (0–9).

Criteria	Description
Speed of response	Subjective evaluation of the proband
Naturalness of language	An evaluation of how much the chatbot sounds like a human
Contextual understanding	To what extent does the chatbot remember the previous part of the conversation
Depth of answers	Evaluation of whether the chatbot can have a deeper debate or only give superficial answers
Appropriateness of the answer	Evaluating the extent to which the chatbot responds accurately, clearly and relevantly to a given question
Quality of the supported language	Evaluation of grammar, sentence structure, etc.

Tab. I *Selected criteria with description.*

The order of chatbots alternates. Each conversation in the chatbot has an initial prompt:

- The user is in the process of driving, and your job is to respond briefly and clearly.
- Answers must be no longer than three sentences.
- Conversation is in Czech.

The prompt in the translated version:

> *"Pretend you are having a conversation in a car, respond as if you were talking to a friend. Do not respond in a machine-like manner. We are on the route between A and B, so do not send any links or images, and try to keep your answers short, to a maximum of two to three sentences."*

After the conversation with all the chatbots and their evaluation, there is a final chatbot preference question and a space for comments.

4. Results

A total of 20 probands participated in the experiment, including 12 men and 8 women with an average age of 32.3 years. Most of the probands already had some degree of experience with chatbots, with the majority mentioning the popular ChatGPT.

In Tab. II, we can see the average values for the importance of individual criteria and the average values for the evaluation of chatbots. Value 1 being the worst, value 10 being the best rating.

Criteria	Importance	Gemini	ChatGPT	DeepSeek	Pi
Speed of response	6.50	7.50	7.96	7.88	6.25
Naturalness of language	6.63	8.33	8.38	8.30	3.63
Contextual understanding	9.62	8.40	8.33	8.38	3.78
Depth of answers	8.88	7.25	8.13	6.75	4.00
Appropriateness of the answer	9.80	7.50	8.63	6.25	3.63
Quality of the supported language	6.25	8.88	8.72	8.75	4.12
Completely wrong answer	–	0.38	0.13	1.13	1.50

Tab. II *Criteria results.*

The last row in the table contains the average number of "completely incorrect answers" per proband. These "completely incorrect answers" are factual errors, such as completely incorrect opening hours for a particular establishment, recommendations for restaurants that do not exist, or completely inaccurate information about the current weather. Responses during conversations that had factual value were subject to fact-checking.

Values in Tab. II are visualized in the following graphs Fig. 4, 5, 6 including standard deviations, which indicate how much each parameter varied between the participants.

In terms of importance, the respondents considered the criteria "appropriateness of the answer" and "contextual understanding" to be the most important. On the other hand, they considered "quality of the supported language" and "naturalness of language" to be less important.

Looking at the results more widely, we can see that there are only slight differences between Gemini, ChatGPT, and DeepSeek, but Pi lags far behind, as it is not as well-known as the other chatbots, and its quality level is much lower according to

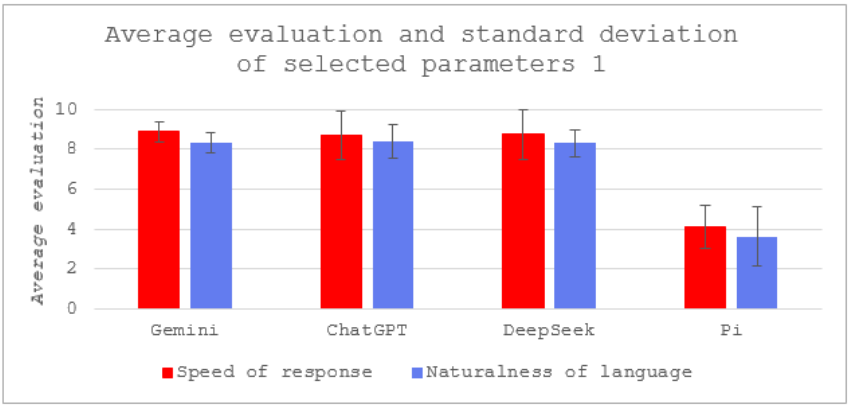

Fig. 4 *Average evaluation and standard deviation of selected parameters 1.*

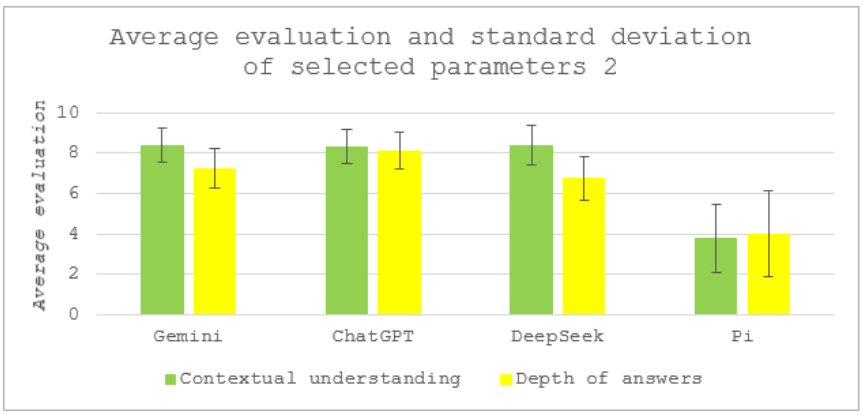

Fig. 5 *Average evaluation and standard deviation of selected parameters 2.*

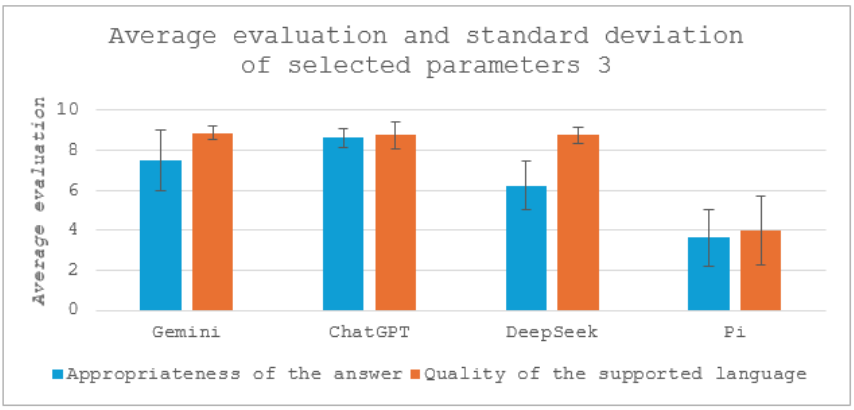

Fig. 6 *Average evaluation and standard deviation of selected parameters 3.*

the results. Pi lagged in all evaluated criteria, with the biggest problems mentioned being incomplete answers, declension, and incorrect information provided.

The Gemini performed most averagely, neither surprising nor disappointing in any way, compared to its two competitors. The only negative observation concerned the constant sending of links to places and restaurants, which was undesirable in the initial setting.

DeepSeek was the least researched chatbot of the trio. It surprised positively with its contextual understanding of text and handled slang language well, but on the other hand, it sometimes provided false information. DeepSeek sometimes takes prompts too literally and does not fully understand the spoken reference.

ChatGPT received the best average rating, excelling particularly in the criterion of "appropriateness of the answer," which was the most important criterion from the perspective of the test probands. It also had the lowest number of "completely wrong answers" and generally handled the naturalness and quality of language very well.

5. QFD Analysis

Quality function deployment is an analytical method that uses a structured process to translate customer requirements into technical solutions. It may also be used to evaluate how well a particular solution or product meets specific criteria (customer requirements). A structured matrix, known as the "house of quality", is used to evaluate the data [24].

The evaluation criteria C_i have a certain priority for the customer, scored on a scale of 1–10. Their corresponding correlation, R_{ij}, represents the strength of the relationship between customer requirements i and technical descriptor j. It is evaluated using a discrete scale of 0, 1, 3, and 9, based on expert assessment.

Typical R_{ij} values are

- strong = 9,
- medium = 3,
- weak = 1,
- none = 0.

The importance of each technical descriptor j is calculated as a weighted sum of evaluation criteria C_i and their corresponding correlations R_{ij}.

This relationship can be expressed as:

$$W_j = \sum_{i=1}^{m} C_i \, R_{ij}, \tag{3}$$

where W_j represents the absolute weight of the technical descriptor j.

To express the relative importance of individual technical descriptors, the absolute weights are normalized as:

$$P_j = \frac{W_j}{\sum C_i} \cdot 100, \tag{4}$$

where P_j denotes the relative weight expressed as a percentage, W_j is the importance of the selected technical feature, and the $\sum C_i$ is the maximum ideal customer priority.

In the case of chatbot evaluation, we can use the QFD method to assess how the tested chatbots overall fulfill the given criteria. The correlation values were divided according to the average rating given by participants. If the rating exceeded 8, a full correlation of 9 was assigned; if the rating exceeded at least 4, a correlation of 3 was assigned. Lower ratings received a correlation of 1.

The QFD matrix with the results is presented in Tab. III. The values C_i represent the priority of individual evaluation criteria, while R_{ij} expresses how each model j performs with respect to the criterion metric i. The total score W_j is calculated as a weighted sum of these values, and P_j represents the corresponding normalized percentage.

The QFD matrix illustrates the differences between individual chatbots more clearly than a table with averages. Using this, ChatGPT is clearly the most successful, meeting 92.29% of the maximum requirements.

Metric	Priority $\langle 1;10\rangle$	Ideal	Gemini	ChatGPT	DeepSeek	Pi
Speed of response	6.50	9	3	3	3	3
Naturalness of language	6.63	9	9	9	9	1
Contextual understanding	9.62	9	9	9	9	3
Depth of answers	8.88	9	3	9	3	3
Appropriateness of the answer	9.80	9	3	9	3	1
Quality of the supported language	6.25	9	9	9	9	3
Hallucinations/complete inaccuracy	8.50	9	9	9	3	1
TOTAL	$\langle 1;9\rangle$	505.62	354.54	466.62	303.54	118.68
Percentage		100	70.12	92.29	60.03	23.47

Tab. III *QFD matrix with analysis results.*

6. Discussion

At the beginning of the study, we created a questionnaire that focused on preferred topics of conversation. We also compiled an overview of chatbots, the topic of driver fatigue, the influence of communication, and designed an experiment.

The experiment confirmed our initial assumption that ChatGPT and Google Gemini chatbots would likely achieve the best results. UX metrics were used for evaluation, which is a subjective form of evaluation. Participants also rated (1–10) the evaluation criteria in terms of relevance. This makes it clear which requirements for chatbots in a car are considered most important – "appropriateness of the answer" and "contextual understanding". On the other hand, speed of answers, naturalness, and quality of supported language were surprisingly considered not particularly important.

The most popular chatbot, ChatGPT, achieved the highest average score across nearly all criteria. The greatest difference in both average score and standard deviation was observed in the "appropriateness of answer" criterion, where the chatbot stayed within the boundaries defined by the prompt; this difference can be seen in Fig. 6.

Google Gemini and DeepSeek chatbots achieved very similar results across all criteria. Both outperformed ChatGPT in the "quality of supported languages" and "contextual understanding" criteria. On the other hand, they performed significantly worse in the "depth of answer" category, where their responses were often very superficial and frequently came across more like those of a voice assistant. A significant range in standard deviation emerged in the "appropriateness of answer" criteria, where both chatbots failed to adhere to the boundaries set by the initial prompt. Despite the prompt, Google Gemini chatbot included links to Google Maps in its responses, displayed weather information, or shared links to restaurants. The DeepSeek chatbot, on the other hand, often took the prompt very literally, and upon verification, its responses were frequently not based on accurate

information. For example, it provided incorrect information about the opening hours of various establishments or venues.

The Pi chatbot lagged behind the others, falling short in all criteria and confirming that more popular and larger chatbots perform much better. In Fig. 4, 5, and 6, we can see that Pi consistently received the lowest average ratings and often had the greatest standard deviation.

The chatbots were tested in Czech, which could have had a significant impact on the criteria of "naturalness of language" and "quality of the supported language". This was most noticeable in the Pi chatbot, which supported Czech but clearly had problems with declension and grammar. However, the responses of chatbots in terms of the actual content of their answers and occasional hallucinations were not language-dependent. The experiment clearly showed the quality of the answers generated by chatbots.

Finally, we performed a QFD analysis to evaluate how well individual chatbots meet UX criteria. ChatGPT came out on top in the QFD, followed by Gemini, DeepSeek, and Pi.

7. Conclusion

Based on the study, it can be said that among the selected chatbots, ChatGPT achieved, on average, the best results, proving to be the most comprehensive, stable, and responsive to a wide range of stimuli. Google Gemini and DeepSeek achieved good results in "quality of supported languages" and "contextual understanding" criteria, but on the other hand, they were lagging in "appropriateness of answer" criteria. Based on this, we consider ChatGPT to be the better chatbot due to its generally more consistent results. Pi achieved the worst results, which may have been partly due to the chosen language. These results were also confirmed by QFD analysis, where ChatGPT achieved the highest rating of 92.29%, meeting the most criteria, followed by Google Gemini, DeepSeek, and finally Pi.

The results show that chatbots at their current level achieve sufficiently high-quality results to be used as conversational assistants for drivers as a method of reducing fatigue in a laboratory testing environment to verify their impact on drivers. For large-scale, serial deployment in vehicles in actual traffic, these tools are still inconsistent and will require further development and optimization.

The main objective of this research will be to test the applicability of the method while conducting fatigue-oriented experiments to verify whether such communication has a positive or negative effect in such a situation.

References

[1] GUANGNAN ZHANG, YAU K.K.W., XUN ZHANG, YANYAN LI. 2016. Traffic accidents involving fatigue driving and their extent of casualties. *Accident Analysis & Prevention*, 87, 2016, pp. 34–42, doi: 10.1016/j.aap.2015.10.033.

[2] FORS K.L., VILLING J. Reducing cognitive load in in-vehicle dialogue system interaction. In: *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, SemDial*, 2011, pp. 55–62.

[3] LINDSTRÖM A., VILLING J., LARSSON S., SEWARD A., ÅBERG N., HOLTELIUS C. The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In: *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[4] SCHMIDT M., MINKER W., WERNER S. User acceptance of proactive voice assistant behavior. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung,* 2020, pp. 18–25.

[5] BOUCHNER P., NOVOTNÝ S., JIŘINA M. Identification of driver's drowsiness using driving information and EEG, *Neural Network World*, 2010, 20(6), pp. 773–791.

[6] BOUCHNER P. A complex analysis of the driver behavior from simulated driving focused on fatigue detection classification; *WSEAS Transactions on Systems*, 2006, 5(1), pp. 84–91.

[7] SPURNÝ P., ANDRŠ J., BOUCHNER P., PUČELÍK J., ROKYTA R. Testing a system for predicting microsleep, *Lékař a technika*, 2016, 46(2), pp. 51–54.

[8] What is Level 4 autonomous driving [online]. Available at: https://www.macnica.co.jp/en/business/maas/columns/144666/

[9] VANLAAR W. Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors. *Journal of safety research*, 2008, 39.3, pp. 303–309.

[10] JELLENTRUP N., METZ B., ROTHE S. Can talking on the phone keep the driver awake?: results of a field-study using telephoning as a countermeasure against fatigue while driving. In: *2nd International conference on driver distraction and inattention.* 2011.

[11] SAXBY D.J., MATTHEWS G., NEUBAUER C. The relationship between cell phone use and management of driver fatigue: It's complicated. *Journal of Safety Research*, 61, 2017, pp. 129–140. doi: 10.1016/j.jsr.2017.02.016.

[12] MAHAJAN K., LARGE D.R., BURNETT G., VELAGA N.R. Exploring the effectiveness of a digital voice assistant to maintain driver alertness in partially automated vehicles. *Traffic Injury Prevention*, 2021, 22, 5 pp. 378–383, doi: 10.1080/15389588.2021.1904138.

[13] JIEUN LEE, TOSHIAKI HIRANO, TOMOYA HANO, MAKOTO ITOH. Conversation during Partially Automated Driving: How Attention Arousal is Effective on Maintaining Situation Awareness. In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, IEEE, Bari, Italy, pp. 3718–3723, doi: 10.1109/SMC.2019.8914632.

[14] LARGE D.R., BURNETT G., ANTROBUS V., SKRYPCHUK L. Driven to discussion: engaging drivers in conversation with a digital assistant as a countermeasure to passive task-related fatigue. *IET Intelligent Transport Systems*, 2018, 12, 6, pp. 420–426, doi: 10.1049/iet-its.2017.0201.

[15] LEE C., ABDEL-ATY M. Presence of passengers: does it increase or reduce driver's crash potential?, *Accident Analysis & Prevention*, 2008, 40.5, pp. 1703–1712.

[16] MAHACHANDRA M., PRASTAWA H., MUFID A.H. Effect of passenger presence towards driving performance level using kss and cnc indicators. In: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 2020. p. 012056.

[17] What is natural language processing? IBM; [online]. Available at: https://www.ibm.com/topics/natural-language-processing

[18] How does ChatGPT work?; [online]. Available at: https://zapier.com/blog/how-does-chatgpt-work/

[19] NAVEED H., KHAN A.U., QIU S., SAQIB M., ANWAR S., USMAN M., AKHTAR N., BARNES N., MIAN A. A comprehensive overview of large language models (arXiv:2307.06435v10), 2024, https://arxiv.org/abs/2307.06435.

[20] Bang J., Ahn S. UX Design and Evaluation on Conversational Bot Supporting Multi-Turn and Multi-Domain Dialogues, In: *International Conference on Platform Technology and Service (PlatCon)*, Jeju, Republic of Korea, 2022, pp. 92–97, doi: 10.1109/PlatCon55845.2022.9932091.

[21] BOUCHNER P., NOVOTNÝ S. System with Driving Simulation Device for HMI Measurements, In: *WSEAS Transactions on Systems*. Athens: WSEAS Press, 2005, pp. 287–293. ISBN 960-8457-29-7.

[22] EL HAMDANI S., BOUCHNER P., KUNCLOVÁ T., LEHET D. he Impact of Physical Motion Cues on Driver Braking Performance: A Clinical Study Using Driving Simulator and Eye Tracker, *Sensors*. 2023, 23(1), ISSN 1424-8220.

[23] PETOUSI D., KATIFORI V., ROUSSOU M., IOANNIDIS Y. The dialogue facilitator bot: Reflections on design and evaluation, International Conference on Interactive Media, In: *Smart Systems and Emerging Technologies (IMET)*, Limassol, Cyprus, 2022, pp. 1–8, doi: 10.1109/IMET54801.2022.9930025.

[24] DOMINGOS A.S., SILVA J.C.M. PEREIRA J.A. On the use of the quality function deployment matrix for flexible and quantitative prioritization, *Journal of Advanced Management Science*, 2017, 5, 5, pp. 401–408.