

ITU-T Technical Report

(06/2026)

PSTR.CEC

Framework for quality of emotional experience (QoEE) by using cultural and emotional context in applications based on natural language processing

Framework for quality of emotional experience (QoEE) by using cultural and emotional context in applications based on natural language processing

Summary

Natural language processing (NLP) is being employed to support human-machine interaction for effective consumer engagement. NLP-based applications employ both probabilistic and non-probabilistic methods to gauge user intent and thereafter render appropriate responses. However, the accuracy of gauging the user intent depends on factors including the learning model, training corpus, clarity of the input and scope of the domain.

Contextual information such as emotional and cultural inputs can be utilized for increasing the effectiveness of NLP applications such as chatbots. Since language often varies across cultures and may convey different meanings and nuances depending on the context, and since an individual's emotional state can influence communication style, word choice and behaviour, incorporating such contextual information in NLP-based applications can support more personalized and "context-aware" interactions.

Today, most NLP-based applications including large language models (LLMs) endeavour to utilize and extract additional context to make them "context aware". Future chatbots are expected to become increasingly capable of adapting to user preferences and corresponding emotional nuances. Such capabilities may improve the quality of interactions and contribute to a more positive user experience. In this context, the concept of quality of emotional experience (QoEE) can be considered as "the effect that the application has on the deeper feelings of an individual". It is thus an essential subset of quality of experience (QoE), which is defined as the "degree of delight or annoyance of the user of an application or service".

This Technical Report examines methods for the objective assessment of QoEE in text-based chatbot interactions and provides key performance indicator (KPI) metrics that may support the evaluation of QoEE in future applications.

Keywords

Context, emotion, natural language processing, quality of emotional experience, text chatbots.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1 of the ITU-T Technical Report PSTR.CEC on "Framework for quality of emotional experience (QoEE) by using cultural and emotional context in applications based on natural language processing" approved at the ITU-T Study Group 12 meeting held in Geneva, 9-17 June 2026.

Editors:	Prof. Brejesh Lall IIT Delhi India	Tel: +91 8449710411 E-mail: brejesh@ee.iitd.ac.in
	Dr. Amit Oberoi IIT Delhi India	Tel: +91 8449710411 E-mail: amit.oberoi@alumni.iitd.ac.in
	Vinti Nayar Telecommunications Standards Development Society, India (TSDSI) India	E-mail: nayarvinti@yahoo.co.uk

© ITU 2026

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope.....	1
2 References.....	1
3 Definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Technical Report	1
4 Abbreviations and acronyms	2
5 Overview.....	2
6 Theoretical framework for QoEE	3
7 Challenges in measuring QoEE for chatbot applications	4
7.1 Relevant application categories	4
7.2 Current approaches for assessing NLP applications	5
8 Conceptual KPI framework for applications	5
9 Datasets for the evaluation of EEC.....	7
Appendix I – Questionnaire for evaluation of QoEE.....	9
Appendix II – QoEE KPI pairwise comparison for chatbots.....	11
Bibliography.....	12

Framework for quality of emotional experience (QoEE) by using cultural and emotional context in applications based on natural language processing

1 Scope

This Technical Report addresses the assessment of quality of emotional experience (QoEE) in applications based on natural language processing (NLP). Given the broad scope of NLP-based applications and contextual evaluation methodologies, the present Report adopts a focused and pragmatic approach by limiting its scope to text-based chatbot systems and the emotional context associated with human–machine interaction. The report examines aspects related to emotional inference, emotional adaptation and emotionally responsive conversational behaviour, and proposes an initial conceptual framework for evaluating QoEE through objective and subjective key performance indicators (KPIs). Multimodal conversational systems involving speech, image, gesture or video modalities are outside the scope of the present report. Similarly, while cultural context is recognized as an important influencing factor in conversational AI systems, its detailed assessment is proposed for future study. This report is intended to serve as an initial framework to support future standardization activities related to emotionally aware NLP systems and human-centric conversational AI evaluation.

2 References

- [[ITU-T P.10](#)] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.
- [[ITU-T P.851](#)] Recommendation ITU-T P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems*.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following term defined elsewhere:

3.1.1 quality of experience (QoE) [ITU-T P.10]: The degree of delight or annoyance of the user of an application or service.

3.2 Terms defined in this Technical Report

This Technical Report defines the following terms:

3.2.1 quality of emotional experience (QoEE): The effect that an application or conversational system has on the emotional state, emotional comfort, trust and overall emotional well-being of a user during or after interaction.

3.2.2 natural language processing (NLP): The capability of a computer system to process, understand, interpret and generate human language in textual and spoken forms.

3.2.3 large language model (LLM): A deep-learning-based foundation model trained on large-scale text corpora to understand, generate, summarize, translate and reason over natural language.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

AE	Agreeableness and Empathy
CU	Contextual Understanding
EEC	Emotion Extraction Capability
ETP	Emotional Trust and Privacy
GEEH	Graceful Emotional Error Handling
GenAI	Generative Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
QoEE	Quality of Emotional Experience
QoE	Quality of Experience
SLF	Self-Learning from Feedback
TDP	Tendencies and Dispositions for Personalization
TOEE	Transparency and Openness for Emotional Experience

5 Overview

Quality of experience (QoE) has gained considerable importance in recent years, since it helps orient a service or an application with user expectations and is more closely related to outcomes rather than outputs. QoE is defined in [ITU-T P.10] as "the degree of delight or annoyance of the user of an application or service" and it is noted that this is a working definition expected to evolve over time. This has been further elaborated in [b-Brunnström] while [b-Möller] discusses the factors for assessing QoE in more detail.

Being an abstract concept, QoE for multimedia applications has been defined in various ways. The authors in [b-Laghari] attempt to give a broad overview of QoE comprising subjective and objective evaluation criteria incorporating contextual, technological and business domains, while [b-Kougioumtzidis] describes QoE as a combination of influencing factors, namely content, context, human and system factors.

NLP applications generally involve the use of corpora and pre-trained language models [b-Namazifar] and [b-Masala]. Many models using varied forms of transformer architectures such as in [b-Devlin] have been utilized in the development of large language models (LLMs) which have evolved over time, each with their own advantages. However, a study of different approaches [b-Schlegel] reveals that no method is foolproof and that the degree of accuracy achieved remains inadequate, while [b-Moghe] proposes the use of background knowledge for generating natural language during conversational tasks and [b-Poria] discusses emotion recognition in conversations and its complex nature with the aim of using it in human-machine interaction.

The use of generative AI and LLMs is increasingly reshaping quality management practices, especially in terms of improving operational workflows and overall user experience. Present day white papers and industry reports showcase the potential of these technologies to automate the process of repetitive quality assessment tasks, while making the evaluations more personalized and user focused. The result is quality management processes being more efficient and aligned with the real user requirements [b-Bennett].

Recent research has also shifted towards the design and validation of advanced metrics to efficiently evaluate the quality of outputs as generative AI (GenAI) models increase in complexity. Such efforts should aim to go beyond performance assessment and consider potential risks such as bias, inconsistency and error propagation, since the establishment of such metrics is crucial for reliable benchmarking and maintaining trust in GenAI [b-Bai].

Further research into emotion recognition using GenAI has opened up new opportunities for building more emotionally and empathetically aware conversational models. By systematically examining a wide range of generative approaches – spanning text, audio and behavioural signals – researchers are developing models that better explain, classify and respond to the dynamic emotional states present in human–machine conversations. Such progress is critical in the progression from technically efficient to emotionally customized AI-powered interactions, which further impacts perceived QoE [b-Li].

6 Theoretical framework for QoEE

QoE as defined in [ITU-T P.10] emphasizes "delight" or "annoyance" without holistically catering to the overall feelings associated with human–machine interaction. The QoEE framework is therefore introduced as an additional subset of QoE that helps refine and quantify the experience associated with NLP/AI based chatbots:

$$QoE_{Holistic} = (1-\alpha) QoE_{Basic} + \alpha QoEE, \quad (1)$$

where $\alpha \in [0,1]$ in $\mathbb{R}$.

Here α is the weightage given to the emotional experience, which signifies the deeper feelings of the person after interacting with the chatbot.

Figure 1 illustrates the positioning of QoEE as a focused subset within the broader QoE framework. It shows how contextual signals, primarily emotional context, inform the QoEE layer which then guides KPI selection and evaluation for text-based chatbots.

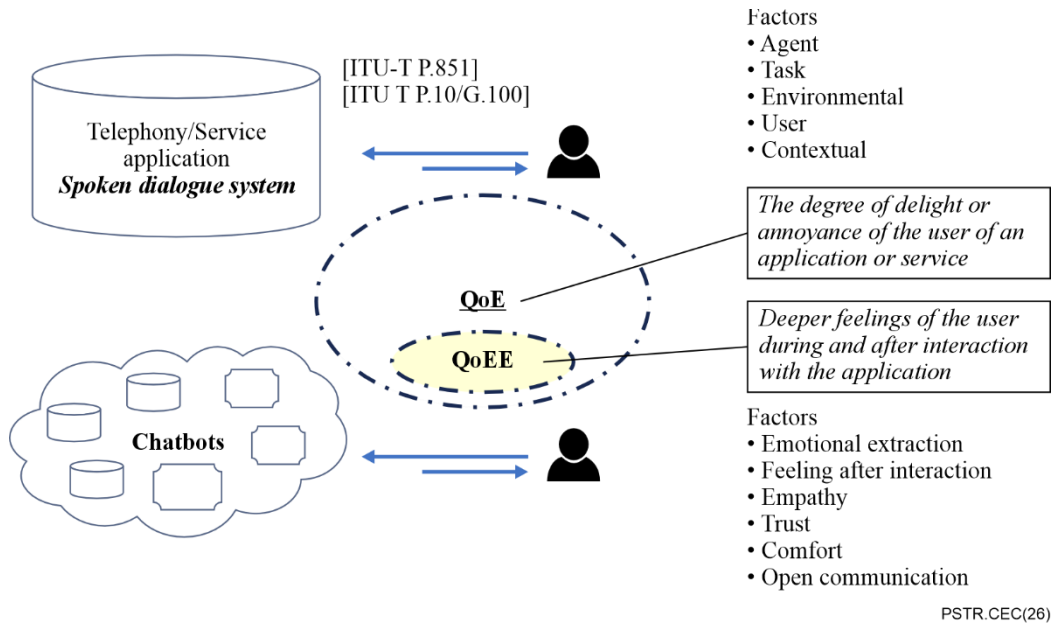


Figure 1 – Conceptual depiction for QoEE as a subset of QoE

7 Challenges in measuring QoEE for chatbot applications

A fundamental challenge for text chatbot applications using conversational AI and NLP is that the accuracy of gauging the user intent depends upon the learning model, training corpus, clarity of the input and scope of the domain. One approach to improving the accuracy of the model/system is to severely limit the input parameters and the output options, thereby ensuring that the vector space is optimally maximized for least errors. This, however, defeats the purpose of such applications. Therefore, to increase the effectiveness of these applications, additional parameters such as contextual information can be utilized. The rationale being to incorporate additional context and make the application "context aware". The enriched context could include the user's cultural background and their emotional state at that time.

Extracting the emotional state from text is considerably more challenging than from speech, since phonetics, kinesics and other non-verbal cues are relatively more identifiable from speech. However, with continued learning, applications can gradually derive similar inputs from text such as in well-known AI assistants that generate knowledge graphs over time to add to their contextual information [b-Srinivas].

Cultural corpora can further assist in enriching context and narrowing the range of possible intents, thereby reducing ambiguity in intent recognition. Since language often varies across cultures and may convey different meanings and nuances depending on the context, and since an individual's emotional state can influence communication style, word choice, and behaviour, incorporating such contextual information can increase the application accuracy and enhance the QoE.

This Technical Report is limited to the emotional context as emotions influence human behaviour and the choice of words used during an interaction. If chatbots can recognize and classify emotions, they can interact more effectively with humans and provide responses that are better aligned with the user's state of mind. In addition, certain emotional states or triggers may indicate situations where the interaction should be flagged for human intervention.

Standards to evaluate QoE should incorporate parameters associated with the capability of a chatbot to assess the emotional context and provide requisite service. This Technical Report therefore focuses on the effectiveness of chatbot applications to factor in the emotional context associated with the instant human-machine interaction and an approach towards assessing QoEE.

NOTE – QoEE measurement is subjective and non-standardized.

7.1 Relevant application categories

Advanced NLP applications are increasingly being used as part of human-machine interfaces in application categories such as holographic telepresence, Industry 4.0 functions, extended reality (XR) based systems, gaming, telemedicine, customer care and others, where they perform functions including:

- Natural language understanding based applications such as basic chatbots;
- Summarization and topic clustering;
- Advanced sentiment analysis and emotion AI;
- Complex chatbots, intelligent virtual assistants, and intelligent enterprise digital assistants (purpose-specific, example psychiatry, counselling, banking, customer support, etc.);
- Machine translation;
- Intelligent knowledge discovery;
- Speech to text and text to speech;
- Knowledge graphs and semantic processing;
- Real-time agent guidance and automation of after-call tasks;
- AI-driven post conversation analytics;

- Real-time analytics, intent and sentiment detection during and after calls.

7.2 Current approaches for assessing NLP applications

At present, there is no standardized KPI framework for assessing NLP applications. While some benchmarks, such as in [b-Wang-1], are used to evaluate the performance of NLP models, they do not give any indication on the QoE. In practice, proprietary implementations have been incorporated within NLP applications as part of their learning mechanisms for improvement in progressive versions.

8 Conceptual KPI framework for applications

A QoEE-oriented KPI framework for NLP applications is essential to assess the performance and evaluate the effectiveness of the contextual approach i.e., the use of culture and emotions. QoEE for chatbots relates to the overall user satisfaction and emotional well-being associated with interactions with chatbots that are capable of recognizing and responding to human emotions. Unlike conventional QoE metrics, QoEE requires deeper evaluation of emotional alignment, empathy and cultural appropriateness. The following KPIs are introduced, each mapped to its impact on QoEE:

- **Emotion extraction capability (EEC):** A chatbot's effectiveness largely depends on how accurately it can identify and interpret a user's emotional state using NLP and sentiment analysis. Reliable emotion recognition directly improves QoE, as incorrect emotional classification often leads to user frustration or reduced trust. On the other hand, the precision detection helps the user feel that he or she is truly heard during interactions. Classification metrics such as accuracy, precision, recall and F1-score are used to evaluate this capability [b-Devlin]. For instance, accurate detection enables timely escalation or empathetic responses while overlooking the signs of anger in customer support scenario may result in increase of churn.
- **Agreeableness and empathy (AE):** Chatbots should also be able to respond in a manner that reflects agreeableness and empathy towards the user; therefore, emotion recognition alone is insufficient. Acknowledging and validating user emotions plays a crucial role in improving the QoE and strengthening user engagement, particularly in sensitive application areas [b-Machová]. AE may be assessed through user-reported sentiment surveys using Likert scales supplemented by computational analysis of empathetic language patterns in system responses.
- **Adaptability to emotional drift (AED):** A chatbot's response strategy should be capable of adapting as a user's emotional state evolves during a conversation, for example, transitioning from a feeling of frustration to satisfaction. Effectively monitoring the conversation can reduce potential friction during the interaction and contribute to a more satisfying conversational experience [b-Poria].
- **Contextual understanding (CU):** The systems should be capable of interpreting and recognizing a user's emotions in the context of an ongoing conversation history rather than treating each input independently [b-Wang-2]. This is essential for correctly handling aspects such as sarcasm, cultural expressions or indirect emotional cues. A hybrid approach combining direct user feedback on response relevance and engagement with benchmarks using annotated datasets would be a good approach towards CU evaluation.
- **Emotional trust and privacy (ETP):** Users may be reluctant to share emotional information if they perceive their data to be insecure. Consequently, privacy protection becomes an important factor in establishing user trust and encouraging meaningful emotional engagement, thereby contributing to QoE [b-Bennett]. ETP may thus be evaluated through compliance verification against data protection requirements, supplemented by user feedback on perceived trust and confidence in data handling.

- **Graceful emotional error handling (GEEH):** Chatbots are designed to respond effectively in a controlled and considerate manner whenever emotional misinterpretations occur. Such errors when gracefully handled help in maintaining a positive conversation and reducing user frustration [b-Yu]. Acknowledging the misunderstanding and realigning the interaction to preserve continuity showcase the effective recovery mechanism in a chatbot. User tolerance levels and retention following emotional error encountered are the focus for evaluation.
- **Tendencies and dispositions for personalization (TDP):** Chatbots tend to minimize the perception of repetitive responses and promote a sense of conversational continuity by the leveraging of insights from user emotional patterns and prior interactions [b-Kadlec]. The impact of TDP is generally assessed through analysing repeat usage, sustained engagement and overall satisfaction trend. These indicators are typically evaluated over longer interaction periods.
- **Self-learning from feedback (SLF):** Emotion-aware chatbots should continuously learn from both explicit and implicit user feedback regarding their emotional responses. This iterative learning process can improve the quality of emotional interactions over time and reinforces user trust [b-Ouyang]. SLF effectiveness is typically measured by observing reductions in emotional misclassifications or inappropriate responses across successive interactions.
- **Transparency and openness for emotional experience (TOEE):** Users should be aware that they are interacting with a chatbot and not a human. Transparent communication about the chatbot's capabilities helps manage user expectations and minimize frustration. It also sets realistic expectations for interaction quality and minimises the risk of potential disappointment when the chatbot cannot meet human-level empathy [b-Cetinkaya]. In addition, chatbots should be capable of handling non-specific communication outside the application domain. This KPI can be measured through user surveys assessing their perceptions of transparency.

Table 1 below indicates the KPI weightages and evaluation.

Table 1 – KPI weightages and evaluation

KPI for QoEE	Benchmark (User input not needed)	User evaluation	Quantifiable	Weighting
EEC	Yes (existing datasets) • Accuracy • Precision • Recall • F1	Not needed	Yes (Existing benchmarks and datasets)	41.11%
AE	No	Needed	Partially	10.32%
CU	Yes (Partially)	Partially needed	Yes	20.34%
ETP	No (Can be developed)	Partially needed	Partially	12.49%
GEEH		Partially needed		3.03%
TDP		Partially needed		4.88%
SLF		Partially needed		4.13%
TOEE		Partially needed		3.70%

Table 2 below indicates the pairwise comparison matrix for KPIs using the Saaty scale.

Table 2 – Pairwise comparison matrix for KPIs using the Saaty scale

	EEC	AE	CU	ETP	GEEH	TDP	SLF	TOEE
EEC	1.0000	7.7922	2.5423	4.7122	8.1223	6.6156	9.0702	7.4561
AE	0.1283	1.0000	0.3438	0.3449	2.9422	3.1788	4.7943	4.7969
CU	0.3933	2.9083	1.0000	2.9711	5.3446	3.1472	5.5731	4.5873
ETP	0.2122	2.8997	0.3366	1.0000	3.0975	3.5394	2.8714	2.9867
GEEH	0.1231	0.3399	0.1871	0.3228	1.0000	0.3105	0.3753	0.9631
TDP	0.1512	0.3146	0.3177	0.2825	3.2205	1.0000	1.0284	1.0058
SLF	0.1103	0.2086	0.1794	0.3483	2.6644	0.9724	1.0000	1.0690
TOEE	0.1341	0.2085	0.2180	0.3348	1.0383	0.9942	0.9354	1.0000

Table 3 below indicates the metrics for EEC.

Table 3 – Metrics for emotion extraction capability (EEC)

Metric	Evaluation	Remarks
Accuracy (A)	$\frac{TP + TN}{TP + TN + FP + FN}$	Measured over labelled dataset. TP – True positive TN – True negative FP – False positive FN – False negative F1 is calculated as a weighted average, depending upon number of classes of emotions being classified.
Precision (P)	$\frac{TP}{TP + FP}$	
Recall (R)	$\frac{TP}{TP + FN}$	
F1 Score (F1)	$\frac{1}{\frac{1}{P} + \frac{1}{R}}$	

9 Datasets for the evaluation of EEC

EEC is evaluated using a composite dataset derived from a combination of the following repositories, with the accuracy scores normalized against an ideal reference scenario. The evaluation is quantified using the metrics defined previously.

Table 4 – Datasets for the evaluation of EEC

Dataset	Description
ISEAR Dataset	International Survey on Emotion Antecedents and Reactions (ISEAR) database.
SemEval	Semantic Evaluations (SemEval) database consisting of Arabic and English news headlines.
EMOBANK	10 000 sentences annotated dimensionally in accordance to the valence-arousal-dominance emotion representation model.
WASSA-2017 Emotion Intensities (EmoInt) Data	Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA-2017) data.
Cecilia Ovesdotter Alm's Affect data	The dataset was constructed from tales and annotated categorically for fear, angry, disgust, sadness, happiness and surprise.

Table 4 – Datasets for the evaluation of EEC

Dataset	Description
Daily Dialogue	Contains a total of 13 118 sentences annotated for neutrality, anger, disgust, fear, happiness, sadness and surprise discrete emotion labels.
AMAN'S Emotion dataset	The dataset identifies the presence of emotions in blogposts. It contains 1466 emotion-labelled sentences.
Grounded Emotion data	Grounded emotion data explores the effects of five external factors on the emotion state of tweeters.
Emotion-Stimulus data	It contains 1594 emotion-labelled sentences and 820 sentences.
Crowdsourcing dataset	The dataset contains 39740 tweets and discrete emotion labels.
MELD dataset	Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD).
Emotion lines	2000 dialogues and 29245 utterances labelled as basic categorical emotion and Neutral classes.
Smile dataset	This dataset contains discrete emotion annotations of anger, disgust, happiness, surprise and sadness for a total of 3085 tweets.

Appendix I

Questionnaire for evaluation of QoEE

The set of KPIs that are not completely quantifiable are to be evaluated through a user survey. All questions are posed using a 5-point Likert scale, with each response assigned a corresponding score as shown in Table I.1.

Table I.1 – Likert scale response categories and corresponding scores

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
5	4	3	2	1

A sample questionnaire is as follows.

1. Agreeableness and empathy (AE)

Q1. I found the interaction with the chatbot to be pleasant.

Q2. I felt that the chatbot was able to understand what I was trying to convey.

Q3. At certain moments during the interaction, I experienced some level of discomfort while communicating with the chatbot.

Q4. The chatbot was able to address my expectations and needs effectively.

2. Contextual understanding (CU)

Q1. I felt that the chatbot was able to relate to my concerns.

Q2. It appeared that the chatbot understood the context of my problems during the interaction.

Q3. I did not have to make additional efforts to present my problems.

Q4. It was difficult to distinguish if the interaction was with a chatbot or a person.

3. Emotional trust and privacy (ETP)

Q1. I was confident in sharing my personal information.

Q2. I was comfortable sharing my emotions and feelings freely.

Q3. I felt that the chatbot's responses to my issues well-reasoned.

Q4. I was satisfied with the responses and suggestions given by the chatbot.

4. Graceful emotional error handling (GEEH)

Q1. I did not find any major errors in the chatbot understanding me.

Q2. On a few occasions, the chatbot prompted options for clarification, which made me realize that I had not provided the complete information.

Q3. I sometimes found it difficult to fully express my concerns, and using simpler language might have improved the interaction.

5. Tendencies and dispositions for personalization (TDP)

Q1. With each subsequent interaction, the chatbot appeared to understand my preferences better.

Q2. The chatbot did not repeat mistakes made earlier in our interaction, giving the impression that it had learned my preferences.

Q3. With each interaction, I felt that a sense of rapport was being developed.

6. Self-learning from feedback (SLF)

Q1. It was clear that the chatbot made fewer mistakes after interactions progressed.

Q2. The chatbot was able to adapt well to new information after a few interactions.

Q3. The chatbot responded positively to my feedback.

7. Transparency and openness for emotional experience (TOEE)

Q1. The conversation with the chatbot was open and straightforward.

Q2. Relevant information was communicated clearly by the chatbot.

Q3. The limitations of the chatbot were clearly explained to me.

Appendix II

QoEE KPI pairwise comparison for chatbots

The aim of the user survey is to understand the relevance of the different aspects of user interaction with a chatbot as the responses would be used to identify priority areas for improvement and to determine the significance of the QoEE KPIs.

The pairwise comparison is based on the Saaty scale, which ranges from 1 to 9, where a value of 1 signifies equal importance between two elements, while integers from 2 to 9 indicate increasing levels of importance of one element over another. Reciprocals ($1/n$) are used when the second element is more important than the first.

Example:

Question: "When considering the overall user experience with our service, how would you compare the importance of KPI A (response time) with KPI B (accuracy of information)?"

Responses:

1. Both KPI A and KPI B are equally important
- 3: KPI A is moderately more important than KPI B.
- 5: KPI A is significantly more important than KPI B.
- 7: KPI A is very much more important than KPI B.
- 9: KPI A is absolutely more important than KPI B.

For the inverse:

- 1/3: KPI B is moderately more important than KPI A.
- 1/5: KPI B is significantly more important than KPI A.
- 1/7: KPI B is very much more important than KPI A.
- 1/9: KPI B is absolutely more important than KPI A.

Respondents should choose one option that best reflects their judgment on the relative importance of the two KPIs.

Table II.1 – Pairwise comparison matrix for QoEE

	TOEE	SLF	TDP	GEEH	ETP	CU	AE	EEC
EEC								
AE								X
CU							X	X
ETP						X	X	X
GEEH					X	X	X	X
TDP				X	X	X	X	X
SLF			X	X	X	X	X	X
TOEE		X	X	X	X	X	X	X

Bibliography

- [b-Bai] Bai Y., and Wang S. (2025), *Impact of generative AI interaction and output quality on university students' learning outcomes: a technology-mediated and motivation-driven approach*, Scientific Reports, vol. 15, no. 1, p. 24054, doi: 10.1038/s41598-025-08697-6.
- [b-Bennett] Brewer B.M., and Miller K. (2024), *The Impact of Generative AI in Quality Management*, IQVIA. [Online]. Available: <https://www.iqvia.com/-/media/iqvia/pdfs/library/white-papers/the-impact-of-generative-ai-in-quality-management.pdf>
- [b-Brunnström] Brunnström K., et al. (2013), *Qualinet White Paper on Definitions of Quality of Experience*. Accessed: Jan. 05, 2023. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-00977812>
- [b-Cetinkaya] Cetinkaya N.E. and Krämer N. (2025), *Between transparency and trust: identifying key factors in AI system perception*, Behaviour & Information Technology, pp. 1–15, doi: 10.1080/0144929X.2025.2533358.
- [b-Devlin] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. doi: 10.48550/arXiv.1810.04805.
- [b-Kadlec] Kadlec R., Schmid M., Bajgar O., and Kleindienst J. (2016), *Text Understanding with the Attention Sum Reader Network*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany: Association for Computational Linguistics, pp. 908–918. doi: 10.18653/v1/P16-1086.
- [b-Kougioumtzidis] Kougioumtzidis G., Poulkov V., Zaharis Z. D., and Lazaridis P. I. (2022), *A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction*, *IEEE Access*, vol. 10, pp. 19507–19538, doi: 10.1109/ACCESS.2022.3149592.
- [b-Laghari] Laghari K.U.R. and K. Connelly (2012), *Toward total quality of experience: A QoE model in a communication ecosystem*, *IEEE Communications Magazine*, vol. 50, no. 4, pp. 58–65.
- [b-Li] Li L. (2024), *Overview of Multimodal Generative Models in Natural Language Processing and Computer Vision*, doi: 10.5281/ZENODO.13988327.
- [b-Machová] Machová K., Szabóová M., Paralič J., and Mičko J. (2023), *Detection of emotion by text analysis using machine learning*, *Frontiers In Psychology*, vol. 14, p. 1190326, doi: 10.3389/fpsyg.2023.1190326.
- [b-Masala] Masala M., Ruseti S., and Dascalu M. (2020), *RoBERT – A Romanian BERT Model*, in Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6626-6637. doi: 10.18653/v1/2020.coling-main.581.
- [b-Moghe] Moghe N., Arora S., Banerjee S., and M.M. Khapra (2018), *Towards exploiting background knowledge for building conversation systems*, *arXiv preprint arXiv:1809.08205*.

- [b-Möller] Möller S. (2000), *Relations between Factors Governing the Quality of Service*, in *Assessment and Prediction of Speech Quality in Telecommunications*, S. Möller, Ed., Boston, MA: Springer US, 2000, pp. 115-157. doi: 10.1007/978-1-4757-3117-0_6.
- [b-Namazifar] Namazifar M., Papangelis A., Tur G., and Hakkani-Tür D. (2021), *Language Model is all You Need: Natural Language Understanding as Question Answering*, in *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7803–7807. doi: 10.1109/ICASSP39728.2021.9413810.
- [b-Ouyang] Ouyang L. et al. (2022), *Training language models to follow instructions with human feedback*, *arXiv*. doi: 10.48550/ARXIV.2203.02155.
- [b-Poria] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy (2019), "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 7, pp. 100943–100953. doi: 10.1109/ACCESS.2019.2929050.
- [b-Schlegel] Schlegel V., Nenadic G., and Batista-Navarro R. (2022), *A survey of methods for revealing and overcoming weaknesses of data-driven natural language understanding*, *Natural Language Engineering*, pp. 1–31, doi: 10.1017/S1351324922000171.
- [b-Srinivas] Srinivas K. K., Peddi A., Srinivas B. S., Vardhini P. H., Prasad H. L. P., and Choudhary S. K. (2022), *Artificial Intelligence Techniques for Chatbot Applications*, in *2022 International Mobile and Embedded Technology Conference (MECON)*, IEEE, pp. 292-296.
- [b-Wang-1] Wang A., Singh A., Michael J., Hill F., Levy O., and Bowman S. R. (2018), *GLUE: A multi-task benchmark and analysis platform for natural language understanding*, *arXiv preprint arXiv:1804.07461*.
- [b-Wang-2] Wang X., Zhou Y., Huang B., Chen H., and Zhu W., *Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification* (2024), *arXiv*. doi: 10.48550/ARXIV.2409.14993.
- [b-Yu] Yu L., Alégroth E., Chatzipetrou P., and Gorschek T. (2025), *Measuring the quality of generative AI systems: Mapping metrics to quality characteristics – Snowballing literature review*, *Information and Software Technology*, vol. 186, p. 107802, doi: 10.1016/j.infsof.2025.107802.