A Methodology for Identifying Evaluation Items for Practical Dialogue Systems Based on Business-Dialogue System Alignment Models

Mikio Nakano[1,3], Hironori Takeuchi[2], Kazunori Komatani[3]

[1]C4A Research Institute, Inc., Setagaya, Tokyo, Japan
[2]Musashi University, Nerima, Tokyo, Japan
[3]SANKEN, Osaka University, Ibaraki, Osaka, Japan
mikio.nakano@c4a.jp, h.takeuchi@cc.musashi.ac.jp
komatani@sanken.osaka-u.ac.jp

Abstract

This paper proposes a methodology for identifying evaluation items for practical dialogue systems. Traditionally, user satisfaction and user experiences have been the primary metrics for evaluating dialogue systems. However, there are various other evaluation items to consider when developing and operating practical dialogue systems, and such evaluation items are expected to lead to new research topics. So far, there has been no methodology for identifying these evaluation items. We propose identifying evaluation items based on business-dialogue system alignment models, which are applications of business-IT alignment models used in the development and operation of practical IT systems. We also present a generic model that facilitates the construction of a business-dialogue system alignment model for each dialogue system.

1 Introduction

Traditionally, in the dialogue systems research community, user satisfaction (Walker et al., 1997; Ultes and Maier, 2021; Pan et al., 2022) and user experience (Clark et al., 2019; Følstad and Taylor, 2021; Johnston et al., 2023; Minato et al., 2023) have been widely used as metrics for evaluating dialogue systems. With recent advancements in dialogue system technology, particularly the development of large language models (LLMs), it has become possible to develop dialogue systems with high scores in these metrics (Hudeček and Dusek, 2023; Iizuka et al., 2023).

However, in developing and operating practical systems, it is necessary to consider various factors other than the aforementioned metrics. For instance, a chatbot using Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) can generate natural responses based on the contents of a database, but there is still a possibility of generating responses that are inconsistent with the database contents. Therefore, there are risks associated with using such a system for customer service. Additionally, when using an LLM on one's own hardware, substantial hardware resources are required, resulting in high running costs. Consequently, if the anticipated benefits do not exceed these costs, it is difficult to continue operating the system.

In addition to LLMs, various new technologies have been proposed for dialogue systems, but not all are used in practical systems. We suspect that one reason for this is the difference between the evaluation metrics used in the research community and those used to evaluate practical systems. So it is crucial to identify evaluation items for building and operating practical systems.

Dybkjær and Bernsen (2002) and McTear (2004) mention requirements for dialogue systems in explaining dialogue systems development life cycles. McTear (2004) discusses the need for considering requirements from not only users but also operators, but how to list all the requirements is not discussed. Nakano et al. (2024) categorize evaluation items for dialogue systems from the system owner's perspective into benefits, costs, and risks, and they include items that do not have a positive correlation with user satisfaction or user experience. However, the methodology for identifying all evaluation items for individual dialogue systems has not been presented.

In this paper, we apply business-IT alignment models (Hinkelmann et al., 2016) to dialogue systems. Business-IT alignment models are widely used to link business goals, business processes, and applications to facilitate the examination and evaluation of business systems by various stakeholders. We call the results of the application of business-IT alignment models to dialogue systems **Business-Dialogue System Alignment Models** (hereafter Business-DS Alignment Models). By applying these models to individual dialogue systems to create a business-DS alignment, it becomes possible

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 225–237
May 27–30, 2025. ©2025 Association for Computational Linguistics

to list evaluation items specific to each dialogue system.

Furthermore, to facilitate the creation of the business-DS alignment model for an individual dialogue system, this paper proposes a *generic model for business-DS alignment*. By applying this generic model to individual dialogue systems, it is possible to create an alignment model tailored to each system, which can then be used to identify the corresponding evaluation items.

It should be noted that, while this paper uses the term *business*, it is not limited to the narrow sense of business. Instead, it encompasses all practical dialogue system development and operation. For example, the same analytical approach can be applied to systems developed and operated by non-profit organizations or local governments.

2 Previous Work

2.1 Evaluating Dialogue Systems

As previously mentioned, user satisfaction (Walker et al., 1997; Pan et al., 2022; Ultes and Maier, 2021) and user experience (Clark et al., 2019; Følstad and Taylor, 2021; Johnston et al., 2023; Minato et al., 2023) are commonly used metrics for evaluating dialogue systems. User satisfaction is measured by integrating factors such as the degree of task completion and the cost incurred by the user to achieve the task (Walker et al., 1997). User experience is generally measured through subjective evaluations. Post-interaction surveys are often used to ask questions such as whether the interaction with the system was enjoyable or if the user would like to converse with the system again.

However, there are also studies addressing important factors that cannot be measured by these metrics alone. One such factor is development cost. Recent dialogue system technologies often utilize models trained with annotated data. Using active learning to achieve higher accuracy with a smaller amount of annotations is proposed (Asghar et al., 2017; Hiraoka et al., 2017; Tur et al., 2005). Additionally, end-to-end learning for building dialogue systems (Lowe et al., 2017; Wen et al., 2017) can reduce development costs by eliminating the need for annotations. Furthermore, research is also being conducted to reduce hardware costs during operation (Pandelea et al., 2022).

In addition, recent neural dialogue generation and dialogue systems using large language models may include offensive or discriminatory language in their utterances. Methods for avoiding such utterances are also proposed (Xu et al., 2021; Sun et al., 2022; Ziems et al., 2022; Henderson et al., 2018).

However, no methodology has been proposed to identify all the items to evaluate when developing and operating practical dialogue systems.

2.2 Business-IT Alignment Model

To identify all the evaluation items, it is necessary for various stakeholders involved in the development and operation to overview and evaluate the project from their respective perspectives. This requires a comprehensive view of the entire project.

In the context of IT systems in general, not limited to dialogue systems, discussing systems from both managerial and developmental viewpoints is referred to as *business-IT alignment*. To achieve this, the relationships between business goals, business processes, and applications are represented in what is called a *business-IT alignment model*.

In a business-IT alignment model, it is possible to represent not only the IT system itself but also its development and operation. Vicente et al. (2013) created models for operation and Mayer et al. (2019) created models for risks.

There is also research on modeling business-IT alignment for AI service systems that use machine learning (Takeuchi and Yamamoto, 2019). Additionally, meta-models that integrate multiple models related to AI service systems have been proposed (Husen et al., 2024; Takeuchi et al., 2024).

However, dialogue systems are different from typical AI service systems in that they intensively interact with humans. Therefore, the aforementioned models cannot be directly applied to dialogue systems.

3 Proposed Methodology

3.1 Overview

We propose a methodology in which various stakeholders involved in a dialogue system development and operation project can overview and evaluate the project from their respective perspectives by constructing a business-DS alignment model. Based on this model, we identify comprehensively the evaluation items.

A business-dialogue system alignment model consists of the services provided by the dialogue system, values, risks, and costs. Each of these components is broken down into finer elements

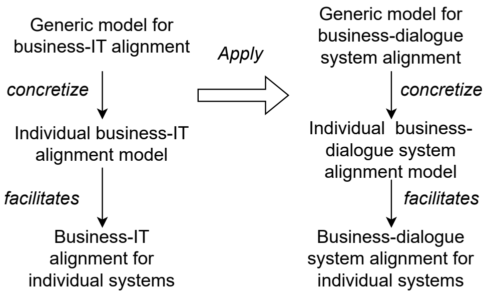
Generic model for business-IT alignment
concretize
Individual business-IT alignment model
facilitates
Business-IT alignment for individual systems
Apply
Generic model for business-dialogue system alignment
concretize
Individual business-dialogue system alignment model
facilitates
Business-dialogue system alignment for individual systems

Figure 1: Relationships among the business-IT and business-DS alignment models.

Notation	Element	
☺	Business actor	A business entity that is capable of performing behavior.
⬭	Business service	An explicitly defined exposed business behavior.
⇨	Business process	A sequence of business behaviors that achieves a specific outcome such as a defined set of products or business services.
▭	Business object	A concept used within a particular business domain.
⬭	Application service	An explicitly defined exposed application behavior.
▭	Application component	An encapsulation of application functionality aligned to implementation structure, which is modular and replaceable.
⌂	Application function	Automated behavior that can be performed by an application component.
▭	Data object	Data structured for automated processing.
⎈	Driver	An external or internal condition that motivates an organization to define its goals.
◎	Goal	A high-level statement of intent, direction, or desired end state for an organization and its stakeholders.
🔍	Assessment	The result of an analysis of the state of affairs of the enterprise with respect to some driver.

Table 1: ArchiMate elements.

and represented using a modeling language called ArchiMate (The Open Group, 2019). By further integrating these elements and expressing the relationships between them, the overall model can be represented. This allows for the enumeration of the values, risks, and costs associated with the target dialogue system.

However, constructing a business-DS alignment model from scratch is difficult for researchers in the dialogue system community. Therefore, we propose a generic model for business-DS alignment. Applying this generic model to individual dialogue systems makes it easy to create an alignment model tailored to each system, which can then be used to list evaluation items. Figure 1 illustrates the relationship among business-IT alignment models and business-DS alignment models.

Notation	Relation	
——→	Triggering	A temporal or causal relation.
- - - - ▷	Realization	A left-hand-side element plays a role in the creation or achievement.
◆——	Composition	A left-hand-side element comprise one or more other elements.
- - - - ->	Influence	A left-hand-side element affects the implementation of motivation elements.
·········>	Access	A left-hand-side element observes or acts upon right-hand- side elements.
●——→	Assignment	A left-hand-side element allocates responsibility or execution.

Table 2: ArchiMate relations.

3.2 Generic Model for Business-Dialogue System Alignment

The generic model for business-DS alignment consists of the generic model of values, the generic model of risks, the generic model of costs, and the generic model of the services provided by dialogue systems (hereafter, we simply call this *the generic model of dialogue systems*). We illustrate these using ArchiMate. The explanations of the ArchiMate elements and relationships are shown in Tables 1 and 2, respectively.

3.2.1 Generic Model of Values

The values of dialogue systems are defined from various perspectives. We consider that it consists of *user value*, *quality value*, and *business value*, and further enumerate their sub-elements. Figure 2 is the ArchiMate illustration for these.

User value User value refers to the benefits that users obtain. Following Aaker (2014), we consider the following four elements as components of the value model:

- **Functional value**: The utility obtained from the functions of the service: e.g., achieving tasks or effectively practicing dialogue.
- **Emotional value**: The special emotions brought about by the process and experience of using the service: e.g., enjoying the conversation.
- **Self-expressive value**: The state where users can express their ideal selves through the use of the service: e.g., feeling satisfied with one's ability to effectively use the dialogue system.
- **Social value**: The identity or sense of belonging gained from using the service: e.g., feeling satisfied being part of a group that uses the same dialogue system.

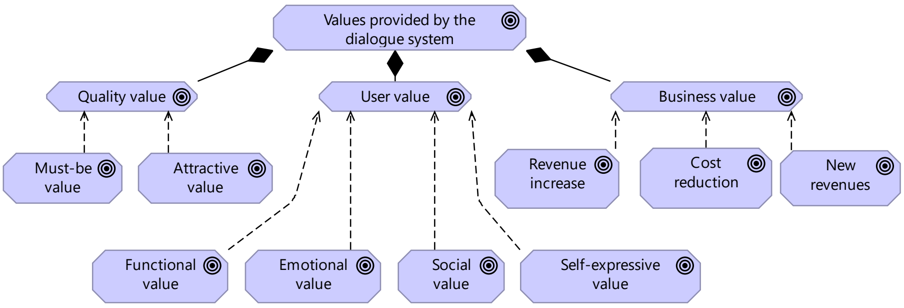

Figure 2: Generic model of values.

Quality value Quality value refers to the value that users obtain from the high quality of the service. Based on the quality model called the Kano Model (Kano et al., 1984; Mikulić and Prebežac, 2011), we decompose quality value into the following elements.

- **Must-be value**: This value leads to dissatisfaction if not fulfilled but does not significantly increase satisfaction when fulfilled. In the context of dialogue systems, this includes the ability to complete tasks reliably and the system not crashing.

- **Attractive value**: This value does not cause dissatisfaction if not fulfilled, but significantly increases satisfaction when it is. For dialogue systems, this includes the ability to engage in natural, human-like conversation, such as fluency and appropriate timing and prosody.

Business value Business value refers to the value obtained by the operators or owners of the dialogue system. The following three elements are considered sub-components:

- **Revenue increase**: This includes the increase in sales of products incorporating the dialogue system and the increase in sales of products recommended by the dialogue system.

- **Cost reduction**: This refers to the reduction in labor costs achieved by replacing tasks previously performed by humans with the dialogue system.

- **New revenue**: This includes revenue from service fees for using the dialogue system, income from displaying advertisements to dialogue systems users, and revenue from selling collected dialogue data.

Here, quality value demonstrates attributes such as "whether not providing it poses a risk" or "whether providing it leads to opportunities." On the other hand, business value can be seen as what the provider gains in exchange for delivering user value (Perri, 2018).

Here, we have listed quality value, user value, and business value in parallel. However, enhancing quality value and user value can lead to an increase in the number of users and usage frequency, which in turn may lead to revenue increase, cost reduction, and new revenues. These relationships vary depending on the individual system.

Note that we do not limit the dialogue systems targeted in this study to task-oriented dialogue systems. Non-task-oriented dialogue systems can also have various values. For example, in the case of a system that allows users to chat with a well-known character (Akama et al., 2017; Han et al., 2022), users can gain emotional value by enjoying casual conversations. Additionally, since the system can promote the character, the system owner can achieve a revenue increase.

3.2.2 Generic Model of Risks

In recent years, there have been many concerns about the risks associated with AI, including generative AI. In this context, principles for the societal implementation of AI are being considered not only by academic organizations but also by national and international institutions. This study views the failure to adhere to these principles as a risk.

Many principles have been established as guidelines, but they vary in granularity and comprehensiveness, and comparisons are being made (Jobin et al., 2019). In our study, the principles mentioned in more than one-third of the 84 guidelines investigated by Jobin et al. (2019) are considered components of risk, and we apply these principles to dialogue systems.

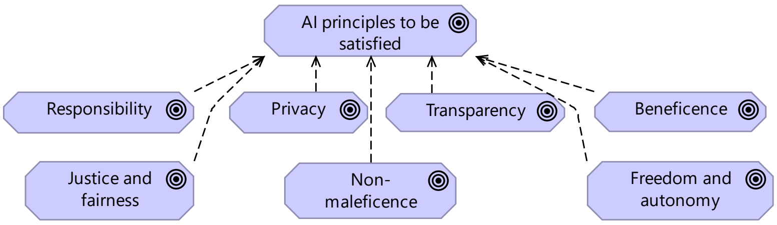

Figure 3: Generic model of risks. Not satisfying the AI principles causes risks.

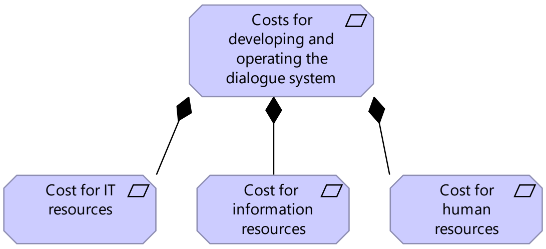
Costs for developing and operating the dialogue system
Cost for IT resources
Cost for information resources
Cost for human resources

Figure 4: Generic model of costs.

Transparency: The dialogue system can explain why it behaved in a certain way.

Justice and fairness: It does not make utterances based on biased thinking.

Non-maleficence: There is no risk of generating defamatory utterances, producing incorrect utterances, or copyright violation.

Responsibility: Responsibility is clearly assigned when problems arise.

Privacy: There is no risk of leakage of personal information, speech, or facial images contained in the dialogue content.

Beneficence: The dialogue system has a positive impact on users and society.

Freedom and autonomy: There is no risk of being used for criminal purposes.

When developing or operating dialogue systems, if there is a possibility that these principles could be compromised, it is considered to be a risk.

Figure 3 illustrates this generic model of risks.

3.2.3 Generic Model of Costs

In the practical implementation of any system, not limited to dialogue systems, development and operational costs are required. These costs can be broken down as follows:

Cost for human resources: This includes human resources for initial system development, system testing, system modifications after the start of operation, and human resources for handling issues and troubleshooting.

Cost for information resources: This involves the creation of annotated data for model building, and the creation of data used as references for writing rules.

Cost for IT resources: This includes computing resources needed for initial system development, server usage fees, external API service usage fees, and application registration fees.

Figure 4 illustrates this generic model of costs.

3.2.4 Generic Model of Dialogue Systems

Below we enumerate the elements related to a dialogue system. This is based on the AI service system description by Takeuchi et al. (2024).

User: The user of the dialogue system.

Operator: The person or entity operating or owning the dialogue system.

User activities using the dialogue system: Activities performed by the user using the dialogue system, such as performing tasks, practicing having a conversation, and enjoying a conversation.

Operator activities using the dialogue system: Activities performed by the operator using the dialogue system, such as providing information and obtaining information from users.

Dialogue services: Services provided by the dialogue system, such as providing information at any time and providing the joy of conversation.

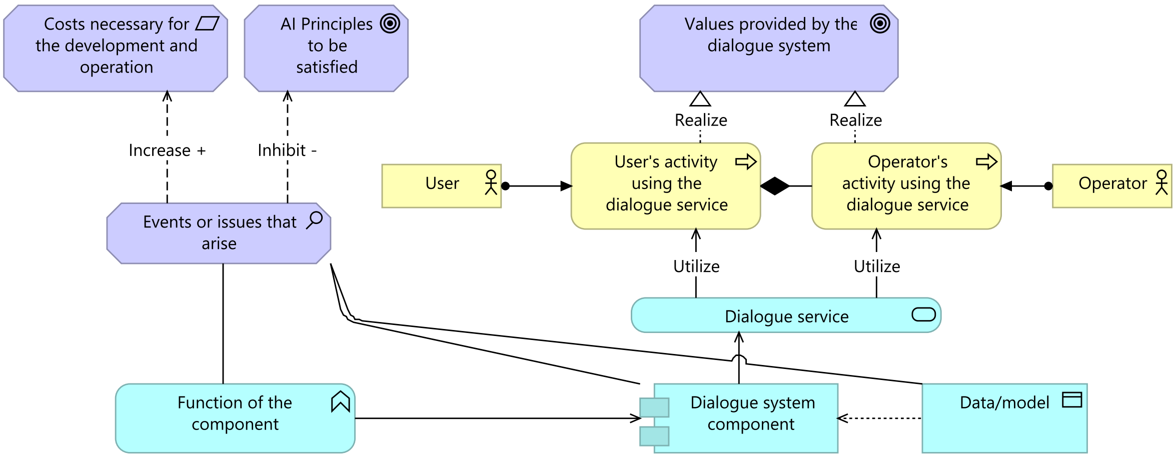
Costs necessary for the development and operation
AI Principles to be satisfied
Values provided by the dialogue system
Increase +
Inhibit -
Realize
Realize
User
User's activity using the dialogue service
Operator's activity using the dialogue service
Operator
Events or issues that arise
Utilize
Utilize
Dialogue service
Function of the component
Dialogue system component
Data/model

Figure 5: Generic model of dialogue systems. A solid line without direction denotes a general relationship.

Dialogue system components: Components within the dialogue system, such as language understanding component, dialogue management component, and information search component.

Component functions: Functions of the dialogue system components, such as language understanding, dialogue management, and information search.

Data/models: Models used by dialogue system components and the data to train these models, such as language understanding model and training data for it.

Observed events/issues: Possible events or issues regarding data/models, application components, or functions, such that annotated data for language model training is necessary and that the language generation component might generate incorrect statements.

Figure 5 illustrates this generic model of dialogue systems.

3.3 Creating a Business-Dialogue System Alignment Model and Identifying Evaluation Items

To create a business-DS alignment model, we will apply the general model described in Section 3.2 to the target dialogue system. In practice, each dialogue system will be represented using ArchiMate, illustrating its relationships with value, cost, and risk elements. Elements not related to these will be excluded.

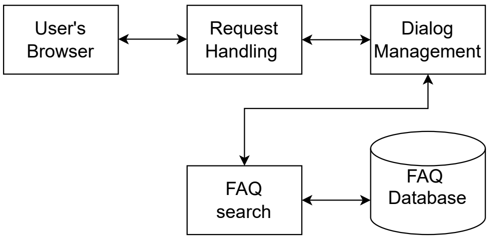

Figure 6: Architecture of the FAQ chatbot as a case study.

In the explanation below, we use a simple FAQ (Frequently-Asked Questions) chatbot as a case study. This chatbot uses an FAQ database containing question-and-answer pairs to respond to user queries via text input and output. It performs example-based question answering (Banchs and Li, 2012; Inaba and Takahashi, 2016). The system operates on a server, and users access it through a browser without entering a user ID. The chatbot comprises a web server for handling requests, a simple dialogue management module based on a state transition model, and an FAQ search module, as shown in Figure 6. The dialogue management module generates initial responses and handles situations where no FAQ match is found. The FAQ search module uses Sentence-BERT (Reimers and Gurevych, 2019) to match the input sentence with example questions, extracts the relevant FAQ, and returns it to the dialogue management module.

We first tailor the generic model of dialogue systems to the target system (Figure 7). In the case of the FAQ chatbot, it becomes as follows:

- *User* is the user of the dialogue system to seek information.

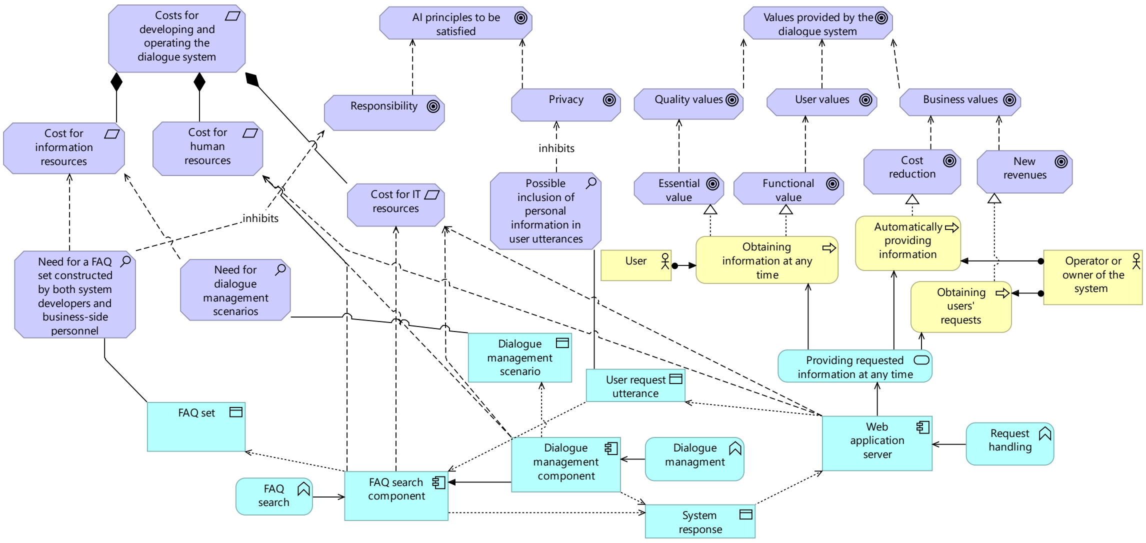
Costs for developing and operating the dialogue system
AI principles to be satisfied
Values provided by the dialogue system
Responsibility
Privacy
Quality values
User values
Business values
Cost for information resources
Cost for human resources
inhibits
Cost for IT resources
Possible inclusion of personal information in user utterances
Essential value
Functional value
Cost reduction
New revenues
inhibits
User
Obtaining information at any time
Automatically providing information
Operator or owner of the system
Need for a FAQ set constructed by both system developers and business-side personnel
Need for dialogue management scenarios
Obtaining users' requests
Dialogue management scenario
User request utterance
Providing requested information at any time
FAQ set
Web application server
Request handling
Dialogue management component
Dialogue managment
FAQ search
FAQ search component
System response

Figure 7: Business-dialogue system alignment model for FAQ Chatbot.

- *Operator* is the operator or the owner of the system who provides information.
- *User activity using the dialogue system* is obtaining information at any time.
- *Operator activities using the dialogue system* are automatically providing information and obtaining users' requests.
- *Dialogue service* is a service that provides information at any time.
- *Dialogue system components* are web application server, dialogue management component, and FAQ search component.
- *Component functions* are request handling, dialogue management, and FAQ search.
- *Data/models* are user request utterances, system responses, the dialogue management scenario, and the FAQ set.
- *Observed events/issues* are the need for a FAQ set, the need for dialogue management scenarios, and the possible inclusion of personal information in user utterances.

Then these are linked to the elements of values, risks, and costs by the following steps.

(1) Derive costs from observed events/issues in the development, operation, and usage of each component.

In the case of the FAQ chatbot, costs are required for developing and operating each component. Additionally, the need for a FAQ set and dialogue management scenarios incurs development and maintenance costs.

(2) Identify principles that are hindered by observed events in the development, operation, and usage of each component as risks.

In the case of the FAQ chatbot, the inclusion of personal information in user utterances poses a risk to privacy protection. On the contrary, since responses are pre-written in the FAQ database, the risk of incorrect answers, biased responses, or responses containing slander is low. Also, since the creation of the FAQ set involves cooperation between dialogue system developers/operators and business-side personnel, there is a risk of unclear responsibility for the content.

(3) Identify business value from activities associated with the dialogue system development operators.

In the case of the FAQ chatbot, automating information providing reduces labor costs. Additionally, analyzing user requests can reveal user needs, leading to new revenue opportunities.

(4) Identify user value from user activities using the dialogue system and the business value influenced by that user value.

In the case of the FAQ chatbot, the ability to obtain information provides functional value to the user.

(5) Identify quality value from user activities using the dialogue system and the business value influenced by that user value.

In the case of the FAQ chatbot, the ability to obtain information at any time without service interruption provides essential value to the user.

In this way, the values, risks, and costs of individual dialogue systems are enumerated and identified as evaluation items. The resulting business-DS alignment model for the FAQ chatbot written in ArchiMate is shown in Figure 7.

Additional case studies can be found in Appendix A.

4 Limitations and Discussion

Although the case studies suggested that our approach is promising, there may be values, risks, and costs that have not been considered, necessitating continuous review. Particularly with advancements in technology like LLMs, which enable more natural conversations, new risks that were previously unconsidered may arise.

As stated earlier, academic research has often used user satisfaction and user experience as evaluation metrics. Roughly speaking, user satisfaction relates to functional value, self-expressive value, and social value. User experience relates to emotional value, must-be value, attractive value, nonmaleficence, justice and fairness, and transparency. Our analysis identified evaluation items beyond these, so it became possible to consider user satisfaction, user experience, and other evaluation items all at once. We hope this leads to new research themes.

In planning the actual system development, it is necessary to balance values, risks, and costs. For example, while showing many advertisements might increase business value, it could decrease emotional value and pose risks to hinder nonmaleficence. Similarly, using a low-performance model to reduce costs can decrease must-be value. A balanced system design considering all evaluation items is necessary, and our approach enables such a balanced design by identifying evaluation items from various perspectives.

In some cases, it is desirable to integrate these evaluation items into a single-dimensional evaluation scale. However, the prioritization of these items must be determined by the consensus of various stakeholders, including the system owner. We hope business-DS alignment models help the facilitation among the stakeholders.

The evaluation items obtained using the methodology proposed in this paper do not necessarily allow for a quantitative assessment of dialogue systems. However, in many cases, various IT-related technologies are proposed and utilized without quantitative evaluation. In addition, focusing only on quantifiable evaluation items and ignoring other items have the risk of falling into the well-known *McNamara fallacy* (also known as the *quantitative fallacy*). We believe that instead of focusing solely on fields where quantitative evaluation is feasible through small-scale experiments, dialogue system researchers should also consider evaluation items that are difficult to quantify. This approach may lead to the development of more practical technologies.

While it is practically impossible to quantitatively demonstrate the superiority of our methodology, we aim to showcase its effectiveness by applying it to the development of a variety of practical dialogue systems and evaluating it from multiple perspectives.

Business-IT alignment models on which our methodology is based may not be familiar to dialogue system engineers, making it potentially challenging to construct a business-DS alignment model. Therefore, we believe it is effective to present a simpler model. As an alternative approach, it is also possible to consider developing human resources who can construct business-DS alignment models while communicating with various stakeholders.

5 Concluding Remarks

This paper proposed a methodology to identify evaluation items for dialogue systems based on business-DS alignment models. Although the methodology presented in this paper needs improvement through more case studies. Nevertheless, we believe that it serves as a useful first step.

Besides the future work already mentioned, We plan to analyze the issues that prevent commercializing systems in the research stage.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number JP22H00536.

References

David Aaker. 2014. *Aaker on branding: 20 principles that drive success*. Morgan James Publishing.

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. Generating stylistically consistent dialog responses with transfer learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–412, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yoshihiko Asao, Julien Kloetzer, Junta Mizuno, Dai Saiki, Kazuma Kadowaki, and Kentaro Torisawa. 2020. Understanding user utterances in a dialog system for caregiving. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 653–661, Marseille, France. European Language Resources Association.

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83, Vancouver, Canada. Association for Computational Linguistics.

Rafael E. Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Jerome R Bellegarda. 2013. Spoken language understanding for natural interaction: The siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, pages 3–14. Springer.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In *Proceedings of AAMAS '14*, page 1061–1068, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Laila Dybkjær and Niels Ole Bernsen. 2002. The dialogue engineering life-cycle. In *A Festschrift for Professor Haldur Õim*, pages 103–125. the Department of General Linguistics 3, University of Tartu.

Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. *Quality and User Experience*, 6(1).

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 123–129, New York, NY, USA. Association for Computing Machinery.

Knut Hinkelmann, Aurona Gerber, Dimitris Karagiannis, Barbara Thoenssen, Alta Van der Merwe, and Robert Woitsch. 2016. A new paradigm for the continuous alignment of business and IT: Combining enterprise architecture modelling and enterprise ontology. *Computers in Industry*, 79:77–86.

Takuya Hiraoka, Graham Neubig, Koichiro Yoshino, Tomoki Toda, and Satoshi Nakamura. 2017. Active learning for example-based dialog systems. In *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, pages 67–78. Springer.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Jati H. Husen, Hironori Washizaki, Jomphon Runpakprakun, Nobukazu Yoshioka, Hnin Thandar Tun, Yoshiaki Fukazawa, and Hironori Takeuchi. 2024. Integrated multi-view modeling for reliable machine learning-intensive software engineering. *Software Quality Journal*, 32:1239 – 1285.

Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. 2023. Clarifying the dialogue-level performance of GPT-3.5 and GPT-4 in task-oriented and non-task-oriented dialogue systems. In *Proceedings of the AAAI Fall Symposia*, volume 2, pages 182–186.

Michimasa Inaba and Kenichi Takahashi. 2016. Neural utterance ranking model for conversational dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 393–403, Los Angeles. Association for Computational Linguistics.

Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2021. *A Job Interview Dialogue System with Autonomous Android ERICA*, pages 291–297. Springer Singapore, Singapore.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5).

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, Di Jin, Patrick Lange, Shaohua Liu, Sijia Liu, Daniel Pressel, Hangjie Shi, Zhejia Yang, Chao Zhang, Desheng Zhang, Leslie Ball, Kate Bland, Shui Hu, Osman Ipek, James Jeun, Heather Rocker, Lavina Vaz, Akshaya Iyengar, Yang Liu, Arindam Mandal, Dilek Hakkani-Tür, and Reza Ghanadan. 2023. Advancing open domain dialog: The fifth Alexa Prize socialbot grand challenge. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Noriaki Kano, Nobuhiku Seraku, Fumio Takahashi, and Shinichi Tsuji. 1984. Attractive quality and must-be quality. *Journal of the Japanese Society for Quality Control*, 14(2):147–156. (in Japanese).

Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 2016. Small talk improves user impressions of interview dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 370–380, Los Angeles. Association for Computational Linguistics.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS 2020*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Nicolas Mayer, Jocelyn Aubert, Eric Grandry, Christophe Feltus, Elio Goettelmann, and Roel J. Wieringa. 2019. An integrated conceptual model for information system security risk management supported by enterprise architecture management. *Softw. Syst. Model.*, 18(3):2285–2312.

Michael F. McTear. 2004. *Dialogue Engineering: The Dialogue Systems Development Lifecycle*, pages 129–161. Springer London, London.

Josip Mikulić and Darko Prebežac. 2011. A critical review of techniques for classifying quality attributes in the Kano model. *Managing Service Quality: An International Journal*, 21(1):46–66.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2023. Design of a competition specifically for spoken dialogue with a humanoid robot. *Advanced Robotics*, 37(21):1349–1363.

Mikio Nakano, Hisahiro Mukai, Yoichi Matsuyama, and Kazunori Komatani. 2024. Evaluating dialogue systems from the system owners’ perspectives. In *In Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*.

Yan Pan, Mingyang Ma, Bernhard Pflugfelder, and Georg Groh. 2022. User satisfaction modeling with domain adaptation in task-oriented dialogue systems. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 630–636, Edinburgh, UK. Association for Computational Linguistics.

Vlad Pandelea, Edoardo Ragusa, Tom Young, Paolo Gastaldo, and Erik Cambria. 2022. Toward hardware-aware deep-learning-based dialogue systems. *Neural Comput. Appl.*, 34(13):10397–10408.

Melissa Perri. 2018. *Escaping the Build Trap: How Effective Product Management Creates Real Value*. O’Reilley Media.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Huai-Hung Huang. 2018. Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching. In *Proc. Interspeech 2018*, pages 1006–1010.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Hironori Takeuchi, Jati H. Husen, Hnin Thandar Tun, Hironori Washizaki, and Nobukazu Yoshioka. 2024. Enterprise architecture-based metamodel for machine learning projects and its management. *Future Generation Computer Systems*, 161:135–145.

Hironori Takeuchi and Shuichiro Yamamoto. 2019. Business AI alignment modeling based on enterprise architecture. In *Proceedings of the 11th KES International Conference of Intelligent Decision Technologies*, pages 155–165. Springer.

The Open Group. 2019. *ArchiMate® 3.1-A pocket guide*. Van Haren.

Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186.

Stefan Ultes and Wolfgang Maier. 2021. User satisfaction reward estimation across domains: Domain-independent dialogue policy learning. *Dialogue and Discourse*, 12(2):81–114.

Marco Vicente, Nelson Gama, and Miguel Mira da Silva. 2013. Modeling ITIL business motivation model in ArchiMate. In *Proceedings of IESS 2013*, volume 143, pages 86–99. Springer.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Recipes for safety in open-domain chatbots. *Preprint*, arXiv:2010.07079.

Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2019. *An Open-Source Dialog System with Real-Time Engagement Tracking for Job Interview Training Applications*, pages 199–207. Springer International Publishing, Cham.

Jie Zeng, Yukiko Nakano, and Tatsuya Sakato. 2023. Question generation to elicit users' food preferences by considering the semantic content. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 190–196, Prague, Czechia. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A Additional Case Studies

A.1 Dialogue Systems Analyzed

In addition to the FAQ chatbot that was analyzed in Section 3.3, we analyzed the system listed below. We selected these systems because they are already in commercial service or close to practical use. Note that we do not assume the same settings as the systems referenced in the literature.

Speech-based assistant on smartphones This works as an embedded application of smartphones and performs question answering, controlling applications, and other tasks like Apple's iPhone Siri (Bellegarda, 2013). The input modality is speech and the output modalities are speech, displaying on the smartphone, and application control. It uses proprietary speech recognition. Wake words are recognized on the device and other user utterances are recognized on the server. It also uses proprietary server-based language understanding using BERT or others. Dialogue management and response generation are rule-based and run on the server. Speech synthesis is device-embedded.

Job interview practice system This system is designed for practicing job interviews (Inoue et al., 2021; Yu et al., 2019; Su et al., 2018) by interacting with a virtual agent. The system operates on a server and is accessed via a browser. The input modalities are speech and facial images, and the output modalities are speech and virtual agents. It uses commercial server-based speech recognition and speech synthesis. Language understanding, dialogue management, and language generation use an API-based commercial LLM service (such as OpenAI's ChatGPT[1]). The virtual agent runs on the browser.

Interview dialogue system for understanding user status This is a virtual agent dialogue system designed to engage with users, asking about their lifestyle and health status while conversing with them (DeVault et al., 2014; Asao et al., 2020). To ensure continuous use, the system aims to make the dialogues enjoyable for the users (Kobori et al., 2016). The system operates on a server and is accessed through a browser. Input modalities are speech and facial images and the output modalities are speech and virtual agents. It uses server-based commercial speech recognition and language understanding, and device-embedded speech synthesis. It also uses scenario-based dialogue management running on the server. The virtual agent runs on a browser.

Conversational recommender system This system engages in dialogue to elicit user preferences and experience (Zeng et al., 2023), and based on this information, recommends products (Jannach et al., 2021; Gao et al., 2021). It operates on a server. The input and output modality is text. It uses a crowd service for language understanding and state transition model-based dialogue management (e.g., Google Dialogflow[2]).

A.2 Evaluation Items for Example Dialogue Systems

Table 3 shows the elements of the generic model for business-DS alignment and their relation to each example system. The factors listed under "common to all system" are those shared by all systems. We show this table instead of the comprehensive ArchiMate representations for simplicity.

Relatively minor risks have been omitted. For instance, even if rule-based utterance generation is used, there is a possibility that the person writing the rules might create biased or offensive utterance templates. However, this risk is generally low because checks are usually conducted before the system is deployed.

In contrast, response generation using LLMs carries a higher risk because it cannot be pre-checked. However, compared to other applications, job interview practice systems have relatively low actual harm even if the LLM generates inappropriate utterances. Considering the development cost, using an LLM is reasonable.

These case studies have suggested that, based on the business-DS alignment models, it is possible to identify the costs, risks, and values of individual dialogue systems. They also allow for highlighting potential issues and comparing systems from various perspectives.

[1] https://openai.com/index/chatgpt/

[2] https://cloud.google.com/dialogflow

Elements in the generic model			Common to all systems	Example dialogue system			
				Speech-based assistant on smartphones	Job interview pratice system	Interview dialogue system for understanding user status	Conversational recommender system
Values provided by the dialogue system	Business Value	Revenue increase		Increase in the sales of a product integrated with the system			Increase in the sales of recommended products
		Cost reduction			Reduction in labor costs	Reduction in labor costs	Reduction in labor costs
		New revenues	Reuse of collected dialogue data		Dialogue system usage fee		
	User value	Functional value		Can obtain desired information	Can effectively practice dialogues		Can receive product recommendations tailored to the user's preferences
		Emotional value			Not embarrassing because no one else can hear	Can enjoy conversation	
		Self-expressive value					
		Social					
	Quality Value	Must-be value	System does not stop	Can accomplish task with high probability	Can accomplish task with high probability		Can accomplish task with high probability
		Attractive value	Can engage in natural, human-like conversations				
AI principles to be satisfied	Transparency				Risk that the behaviors of the LLM cannot be explained		
	Justice and fairness				Risk of LLM making utterances based on biased thinking		
	Non-maleficence				Risk of LLM making defamatory or incorrect utterances		
	Responsibility				Risk that responsibility sharing between external services and the system is not clear	Risk that responsibility sharing between external services and the system is not clear	Risk that responsibility sharing between external services and the system is not clear
	Privacy			Risk of the leakage of personal information contained in the user's speech or utterance content	Risk of the leakage of personal information contained in the user's speech, facial images, and utterance content	Risk of the leakage of personal information contained in the user's speech, facial images, and utterance content	Risk of the leakage of personal information contained in the user's utterance content
	Beneficence						
	Freedom and autonomy				The risk of not being able to control the content generated by an LLM		
Costs for developing and operating the dialogue system	Cost for human resources		- Initial system development and system testing - System modifications after the operation starts - Troubleshooting and issue resolution				
	Cost for information resources			Annotated data for model construction and response generation rules		Annotated data for model construction and response generation rules	Annotated data for model construction
	Cost for IT resources		- Computational resources required for initial system development - Server cost		SaaS usage fees (speech recognition, speech synthesis, and LLM)	SaaS usage fees (speech recognition and language understanding)	SaaS usage fees (language understanding)

Table 3: Evaluation items for example dialogue systems.