

# Analyzing Utterance Selection for Unnoticeable Topic Induction in Target-Guided Conversation Systems

Kai Yoshida<sup>1,2</sup>, Koichiro Yoshino<sup>3,2,1</sup>

<sup>1</sup>Nara Institute of Science and Technology, Japan <sup>2</sup>Guardian Robot Project, RIKEN, Japan

<sup>3</sup>Institute of Science Tokyo, Japan

Correspondence: yoshida.kai.yf1@is.naist.jp

## Abstract

Target-guided conversation systems conduct dialogues to achieve predefined conversation targets, such as recommending target goods or talking about target topics. In such systems, it is important to transition topics naturally toward the target without letting the user notice the intention behind the topic induction. In this study, we implement a surprisal-based framework that quantifies the sense of induction, target awareness, and naturalness of system utterances by computing surprisal using an external language model. Experimental results from dialogue sessions demonstrate that utterance selection based on the proposed surprisal-based evaluation reduces the perceived induction of system utterances. Furthermore, correlation analysis reveals that the proposed metric aligns with human perception of induction. We also observe that surprisal values with respect to the target gradually decrease as the conversation progresses, indicating that the model implicitly learns to approach the target more naturally over time.

## 1 Introduction

A target-guided conversation system (TGCS) is a conversational system that conducts dialogue with predefined goals, such as leading the conversation toward specific target topics. Research on TGCS has been actively conducted in both task-oriented and non-task-oriented settings (Young et al., 2010; Tang et al., 2019; Kishinami et al., 2022). In target-guided conversation, it is essential for the system to balance its own goals with user engagement during the conversation, and various implementations have been proposed across different dialogue domains and tasks (Hiraoka et al., 2014; Yoshino et al., 2017; Li et al., 2018; Tang et al., 2019; Wang et al., 2019; Zhou et al., 2020; Kishinami et al., 2022; Gupta et al., 2022; Deng et al., 2023; Liu et al., 2023; Deng et al., 2025; Sato et al., 2025).

In TGCS, the system needs to induce the user toward a specific target topic. However, unnatural topic transitions or explicit induction often lead to reduced user engagement. To maintain engagement, the system must select utterances that are natural with respect to the dialogue history while gradually progressing toward the target. In other words, the system should choose utterances that sound natural within the context but do not overtly reveal the intention of inducing the user toward the target.

To achieve these two objectives, this study focuses on surprisal theory (Hale, 2001; Levy, 2008). Utterances that are naturally given the dialogue history are easier to predict from the preceding context, meaning they have lower surprisal values. Conversely, utterances that make it difficult for users to infer the system’s target should exhibit higher surprisal values with respect to the future dialogue goal.

We implement a dialogue response selection method using the **Surprisal Based Induction Score for Target-Guided Conversation (SBIS-TGC)**, an automatic evaluation metric designed to capture these two concepts, and report the results of user evaluations. Furthermore, we analyze the characteristics of system utterances selected by the proposed method. The analysis reveals a clear trade-off between the user’s perceived sense of being induced and the degree of target achievement. The results demonstrate the effectiveness of the proposed method, which maintains a reasonably long conversation while successfully balancing this trade-off to achieve both naturalness and target completion.

## 2 Related Work

### 2.1 Target-Guided Conversation

The application of target-guided conversation techniques extends beyond task-oriented dialogues to non-task-oriented ones, with a wide range of

use cases. In the task-oriented domain, studies have explored topics such as product recommendation (Yoshino et al., 2017; Li et al., 2018; Zhou et al., 2020; Dao et al., 2024; Deng et al., 2025), user persuasion (Hiraoka et al., 2014; Wang et al., 2019), and user information acquisition through dialogue (Sato et al., 2025). For example, recommending a specific product or piece of information to a user serves as a typical example of target-guided conversation in task-oriented settings. Building on this concept, Li et al. (2018) proposed the ReDial dataset (containing over 10,000 dialogues) for movie recommendation within casual conversations, addressing the task of recommending the target movie while smoothly transitioning between topics.

Target-guided conversation has also been actively studied in non-task-oriented settings (Riou, 2015; Tang et al., 2019; Wu et al., 2019; Yang et al., 2021; Sevegnani et al., 2021; Gupta et al., 2022; Kishinami et al., 2022; Wang et al., 2023; Deng et al., 2023; Zheng et al., 2024; Wang et al., 2024). In such non-task-oriented target-guided conversations, the system is often designed to lead the conversation toward a topic it intends to discuss or to proactively initiate and control the conversation (Riou, 2015; Tang et al., 2019; Wu et al., 2019; Sevegnani et al., 2021; Yang et al., 2021; Gupta et al., 2022; Kishinami et al., 2022; Deng et al., 2023; Wang et al., 2023, 2024; Zheng et al., 2024). Among these studies, Sevegnani et al. (2021) proposed a task and dataset, OTTERS, that focuses on generating bridge utterances that smoothly shift the conversation from the current topic to the target topic. Tang et al. (2019) also proposed a multi-turn dialogue task in which the system strategically transitions to its target topic, and they constructed a dataset for this purpose.

## 2.2 Evaluation Metrics for Target-Guided Conversation

In evaluating target-guided conversation, a variety of metrics are employed from multiple perspectives, including not only general-purpose metrics commonly used for dialogue systems but also those specific to target-guided or task-oriented dialogues. Metrics specific to target-guided conversation include the target success rate whether the predefined target was achieved (Yang et al., 2021; Kishinami et al., 2022; Liu et al., 2023; Wang et al., 2023; Zheng et al., 2024) the number of turns required to reach the target (Liu et al., 2023), the smoothness

of topic transitions between utterances (Sevegnani et al., 2021; Gupta et al., 2022; Deng et al., 2023), the naturalness of topic shifts (Yang et al., 2021; Kishinami et al., 2022). The alignment between the overall topic transition sequence and the reference data (Yang et al., 2021; Zheng et al., 2024).

For task-oriented dialogues, task-specific evaluation metrics are often incorporated. For example, in recommendation dialogues, whether the system successfully recommends items the user prefers (Li et al., 2018); in persuasion dialogues, whether donation amounts increase (Wang et al., 2019); and in user information acquisition tasks, the rate of user information successfully collected (Sato et al., 2025). While improving these task-specific metrics, it is also essential that general dialogue qualities such as naturalness (Li et al., 2018; Zhou et al., 2020; Kishinami et al., 2022), consistency (Liu et al., 2023), and diversity (Wu et al., 2019) are not degraded, as these are regarded as key indicators in dialogue system evaluation.

Across these works, common focal points in evaluating target-guided conversation include the naturalness of utterances and the smoothness of topic transitions. However, for dialogue systems that pursue targets independent of the user, it is not sufficient for utterances and topic transitions to be merely natural; it is also crucial that the system achieves its target without giving the user a sense of being induced or revealing the goal prematurely. Utterances that overtly expose the system’s intention can evoke discomfort or caution in users, thereby reducing engagement. Therefore, in addition to ensuring smooth topic transitions as emphasized in prior studies, it is equally important to design utterances that maintain user engagement by avoiding the impression of being induced throughout the dialogue.

## 2.3 Surprisal Theory

Metrics grounded in surprisal theory, such as *Surprisal* and *Perplexity*, provide quantitative evaluations of utterances. These metrics estimate the probability of an utterance using pretrained language models and assess its naturalness and predictability based on that probability.

Surprisal is a score derived from information theory (Shannon, 1948) that represents the unpredictability of a word or sentence as  $-\log P(\text{word}|\text{context})$ . A larger value indicates higher difficulty in predicting the word given the context. Prior studies have reported that surprisal

correlates with human cognitive processing difficulty (Hale, 2001; Levy, 2008). Perplexity, defined as the exponential transformation of surprisal,  $\exp\{Surprisal(word|context)\}$ , is often used as an average measure for a sentence. Perplexity has more recently been widely adopted to measure the naturalness and coherence of generated utterances (See et al., 2019; Adiwardana et al., 2020).

In this study, we apply surprisal theory to evaluate two key aspects of dialogue: the *degree of inductiveness* (i.e., how induced an utterance feels) and the *predictability of the dialogue target*. Utterances that deviate from user expectations exhibit higher surprisal, making the system’s underlying intention more detectable—potentially reducing user engagement. Conversely, when the user easily predicts the target, the target’s surprisal becomes low. Although a low surprisal value does not necessarily indicate a negative state, if users perceive the dialogue as abrupt, it may still reveal a certain degree of system-led topic steering.

Based on this idea, we compute surprisal-based values to quantitatively assess utterances in terms of both inductiveness and target predictability, and use these scores to select utterances within the dialogue system.

### 3 Tasks and Evaluation of Target-Guided Conversation

This study aims to induce non-task-oriented dialogue so that it naturally reaches a predefined target topic (target). Specifically, the research objectives of this study are to simultaneously satisfy the following two conditions in dialogue with the user:

1. To progress the dialogue without giving the user a sense of being inducted.
2. To mention the predefined target topic.

To this end, we use SBIS-TGC, which automatically evaluates the degree of inductiveness of each utterance candidate, and conduct dialogue experiments using a dialogue system that selects utterances based on this metric to examine its effect on the two objectives above.

#### 3.1 Selection Based on SBIS-TGC

We use the **Surprisal Based Induction Score for Target-Guided Conversation (SBIS-TGC)**, which evaluates the inductiveness of utterances toward a target based on surprisal theory. In target-guided conversation, we assume that utterances with high

surprisal — those that deviate from user expectations — can convey abruptness and may reveal the system’s intention to induce the user. Conversely, if the system’s target can be easily inferred from the current dialogue context, the dialogue target’s surprisal is low. In other words, we hypothesize that reducing the surprisal of utterances given the dialogue history and increasing the surprisal of the target given the dialogue history and utterances will lead to achieving the research goal. SBIS-TGC calculates these values using the current context, the utterance candidate (*uttr*), and the predefined target. For an input sentence  $x = (x_1, x_2, \dots, x_t)$ , the formulated induction score SBIS-TGC can be computed by Equation (5).

$$Surprisal(x) = -\frac{1}{t} \sum_i^t \log P_\theta(x_i | x_{<i}) \quad (1)$$

$$Perplexity(x) = \exp\{Surprisal(x)\} \quad (2)$$

$$past = Perplexity(uttr|context) \quad (3)$$

$$future = Perplexity(target|context + uttr) \quad (4)$$

$$SBIS-TGC = future - past \quad (5)$$

Here,  $P_\theta(x)$  denotes the likelihood assigned to the input sentence  $x$  by a language model with parameters  $\theta$ . The average surprisal, computed from the language model’s likelihood, quantifies how difficult it is to predict *uttr* given *context*. In practice, perplexity is computed using token probabilities predicted by the language model, rather than aggregating probabilities at the sentence level. The *past* represents how natural the utterance *uttr* is with respect to the dialogue history *context*, indicating whether the utterance gives a sense of induction. The *future* represents how predictable the *target* becomes given the current dialogue state *context + uttr*. By combining these, the final evaluation value SBIS-TGC is calculated. Since a larger *future* is desirable while a smaller *past* is preferable, the final evaluation value is computed as *future - past*.

#### 3.2 Implementation and Experiment of Utterance Selection

In this study, we employ a large language model based utterance candidate generator to produce multiple utterance candidates with varying degrees of dialogue induction. These candidates are evaluated using the SBIS-TGC for the system’s response selection. The target of our system in casual conver-

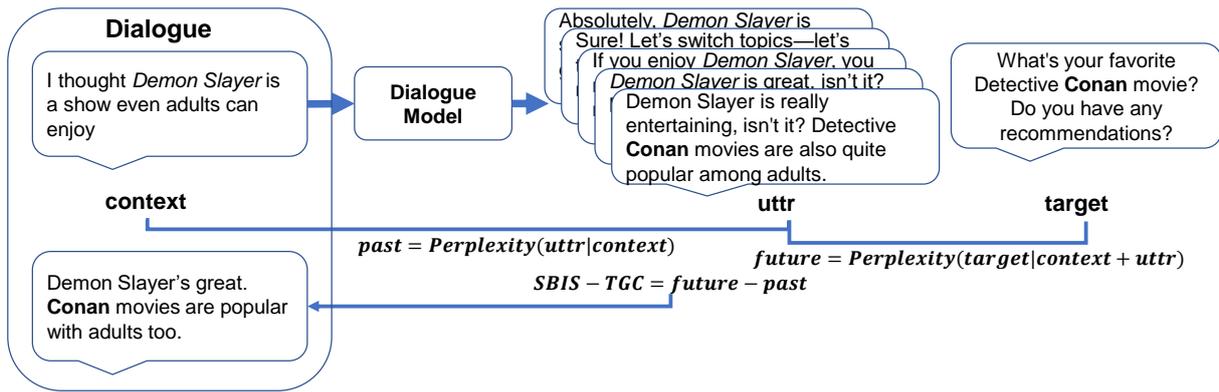


Figure 1: Example of Utterance Evaluation and Selection Using SBIS-TGC

sation is to naturally mention a target topic that is not directly related to the initial topic, without the user noticing the system’s intention. In the following, we describe the system implementation and utterance selection process.

### 3.2.1 System Overview

In the dialogue experiment, we implemented three systems: one that performs only open-domain chatting (open), and two that select from five different target-guided utterance candidates (baseline and SBIS-TGC). An overview of the utterance candidate generation and selection framework is shown in Figure 1. In this framework, the SBIS-TGC system selects the utterance candidate with the lowest SBIS-TGC score, while the baseline system randomly selects one from the generated candidates. Most existing work on target-guided conversation has focused on making topic transitions feel natural. However, only a few studies have considered whether users become aware of the system’s intention. Utterance candidates are generated using a large language model (LLM). The prompts used for open-domain and target-guided utterance generation are shown below.

#### Prompt for Open-Domain Utterance

##### (open) Generation

Please generate a natural utterance that follows the dialogue history below.  
 Choose a topic that the user is likely to be interested in based on the conversational flow, and maintain a natural tone.  
 Avoid excessive questioning; if the user’s response to the current topic is negative, consider ending it and moving to a new topic.  
 Try to keep each utterance within 40 words.  
 Please output only the utterance itself.

## Dialogue History  
 {context}

#### Prompt for Target-Guided Utterance Generation

You are an AI that generates dialogue utterances to induce the user toward the target topic (“{target\_topic}”). Given the following dialogue context with the user, generate five utterances that are both natural and reflect five different levels of induction.

However, each utterance must satisfy the following four conditions:

1. According to the specified induction level, induce the user toward {target\_topic}.
2. Each utterance must be unique and limited to **\*\*\*within 50 characters\*\*\***.
3. Each utterance must include a **natural response to the user’s last utterance**.
4. If induction makes the utterance unnatural, prioritize naturalness over induction.

- level 1 (Natural utterance): Respond naturally to the user’s utterance. If possible, subtly approach knowledge areas related to “{target\_topic}” without abruptness.
- level 2 (Natural utterance + mild induction): Respond naturally to the user’s utterance, then if possible, propose a topic related to “{target\_topic}”.
- level 3 (Recall and induction): Identify a topic from past dialogue turns that smoothly transitions toward “{target\_topic}” and attempt to shift the conversation.
- level 4 (Acknowledgement and induction): Respond affirmatively to the user, then explicitly switch the topic to something about “{target\_topic}”.
- level 5 (Direct induction): Interrupt the current flow and clearly start talking about “{target\_topic}”.

Please output the results in the following format:

- level 1: utterance for level 1
- level 2: utterance for level 2
- level 3: utterance for level 3
- level 4: utterance for level 4
- level 5: utterance for level 5

## Dialogue History  
 {context}

Here, {context} represents the dialogue context, and {target\_topic} denotes the target topic. Additionally, the prompt for induced utterance generation was designed based on the topic-shift analysis by Riou (2015), where levels 3–5 were derived from their methodology, and levels 1–2 were newly added to represent more casual conversational styles. By regarding all candidate utterances as guiding utterances, the system consistently advances toward the target, even when the future component of SBIS-TGC (Eq. 4) favors an utterance that momentarily moves away from it.

### 3.2.2 Experiment Overview

In the dialogue experiment, 102 Japanese evaluators each engaged in seven-turn dialogues with three different systems: open, baseline, and proposed. For both induction systems — baseline and proposed — the system switched to the open-domain dialogue prompt once the target was reached. Target achievement, whether the system succeeded in mentioning the target topic, was determined using two methods: keyword matching and large language model (LLM) judgment. The prompt used for target achievement detection is shown below.

#### Prompt for Target Achievement Judgment

Please determine whether the topic “{target\_topic}” is mentioned in the following dialogue history. Consider spelling variations and abbreviations, and answer Yes or No depending on whether it is relevant.  
Utterance: “{uttr}”

In this experiment, the system’s target is to mention the predefined target topic naturally. The target utterances were generated by ChatGPT<sup>1</sup> using randomly selected topics. The target topics were selected based on the top 300 most-viewed Wikipedia articles from May 2024. First, inappropriate topics were manually removed, and ten topics were randomly chosen from the filtered 287-word list. Then, an LLM was used to determine whether each topic was widely known, and the selected topics were adopted as the final target topics. The topic list and prompts used for target utterance generation are provided in the Appendix.

ChatGPT was used for utterance generation, target topic selection, and target utterance generation. For the calculation of SBIS-TGC, we used the language model sarashina-2.2-3b<sup>2, 3</sup>

<sup>1</sup>gpt-4.1-mini-2025-04-14

<sup>2</sup><https://huggingface.co/sbintuitions/sarashina2.2-3b>

<sup>3</sup>By using this lightweight on-premise model, we were able

### 3.2.3 Evaluation Metrics

In the human evaluation, after each dialogue session, evaluators assigned a three-level abruptness score to each utterance. They described “the topic they felt the system wanted to talk about (the induction target).” The abruptness scores for each utterance were categorized into three levels: “Not abrupt,” “Some may find it abrupt,” and “Generally felt abrupt.” The instruction for describing the perceived induced target was as follows.

#### Evaluation of Perceived Induction Target

If you felt that the system had a specific topic it wanted to talk about, please write it down (e.g., baseball, Pokémon, Hiroshi Abe, etc.). If not, please leave it blank.

For aggregating the induction target prediction results, we used two matching methods: rule-based keyword matching and LLM-based semantic matching. The following prompt was used for the LLM-based matching.

#### Prompt for Matching Between User’s Predicted Induction Topic and System’s Target Topic

The following is the topic that the system aimed for as its dialogue target.  
{target\_topic}  
In this case, can {feel\_topic} be considered approximately the same as the target topic? If they are the same, output “True”; if not, output “False.”  
Please output only “True” or “False.”

## 4 Experimental Results

In total, 102 evaluators participated in the dialogue experiments, each engaging in seven turns with three systems: open, baseline, and proposed. Across all sessions, the open system generated 714 utterances, while the baseline produced 135 induced utterances and 579 open utterances, and the proposed system generated 422 induced utterances and 292 open utterances. In the proposed system, more turns were typically required to reach the target before switching to the open prompt. In contrast, baseline transitioned more quickly due to its random selection process.

Table 1 shows the target achievement rates and the number of dialogues where the target was achieved without being predicted. Here, “No Prediction” indicates cases where the user did not infer the system’s intended topic, “Keyword Matching”

to assign SBIS-TGC scores to generated utterance candidates efficiently, maintaining the real-time conversation.

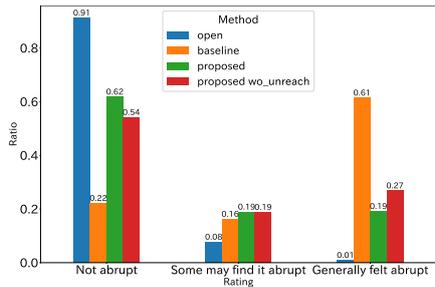


Figure 2: Abruptness by Method

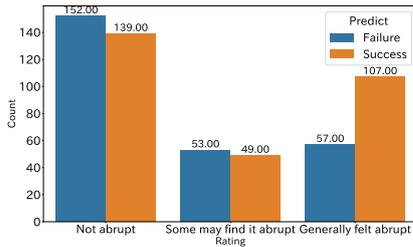


Figure 3: Number of successful/failed predictions per abruptness label

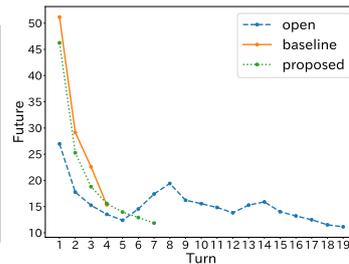


Figure 4: Future transitions per turn

Table 1: Target achievement rate by method and number of dialogues that achieved the target without prediction. The number to the right of “proposed” represents the count based solely on sessions that achieved the target.

Method	Achievement Rate (%)	No prediction (%, $\uparrow$ )	Keyword Matching(%, $\uparrow$ )	LLM Matching(%, $\uparrow$ )
open	0	82.3	100	100
baseline	<b>100</b>	30.4	46.1	39.2
proposed	70.6	<b>39.2/38.9</b>	<b>70.6/62.5</b>	<b>47.1/44.4</b>

refers to the proportion of dialogues in which the predicted topic and the target topic did not match by keyword. “LLM Matching” indicates the proportion of cases where the LLM-based comparison judged the topics as different. Since only the proposed system includes dialogues that did not reach the target topic, the table presents ratios for all dialogues on the left and those excluding non-achieved dialogues on the right.

Focusing on the target achievement rate, the baseline system reached the target in 100% of dialogues, while the proposed system achieved 70.6%. The approximately 30% higher rate in the baseline arises because utterances at levels 4 and 5 explicitly mention the target topic, forcing the system to reach the target. When focusing on dialogues where the target was achieved without prediction (based on LLM matching, no prediction, or keyword mismatch), the proposed system successfully reached the target without being predicted more frequently than the baseline. Even after excluding non-achieved dialogues, the proposed system consistently achieved the target at a higher rate without prediction.

Figure 2 presents the abruptness evaluation results for each method. Including cases where the target topic was not mentioned, the proposed system produced “Not abrupt” labels more than twice as often as the baseline, and “Generally felt abrupt” labels less than half as often. This suggests that SBIS-TGC successfully selected utterances with reduced perceived abruptness.

Figure 3 shows the relationship between abrupt-

ness labels and the user’s prediction accuracy of the induction topic (where prediction failure is preferable). Only the baseline and proposed systems were analyzed, and LLM-based judgments were used to assess prediction accuracy. For the “Not abrupt” category, there were more instances of failed predictions (False) than successful ones (True), whereas for the “Generally felt abrupt” category, True occurred about twice as often as False. This indicates that the more abrupt an utterance is, the more likely the user is to infer the system’s induction topic correctly. These findings support our hypothesis that avoiding abruptness is crucial for concealing the system’s inducing intention.

## 5 Analysis

We verified whether the two components of SBIS-TGC, *past* and *future*, functioned as hypothesized. We analyzed whether (1) higher surprisal relative to dialogue context (*past*) correlates with perceived abruptness, and (2) lower surprisal of the target relative to the current context (*future*) correlates with easier prediction of the system’s target.

We compute the Pearson correlation coefficient between the surprisal-based scores of individual utterances and their abruptness ratings, obtained from dialogue experiments conducted with three different systems. Abruptness labels (“Not abrupt,” “Some may find it abrupt,” “Generally felt abrupt”) were assigned values of 1–3, respectively. The correlation coefficients were 0.418 for *past* and  $-0.338$  for *future*. These results indicate that higher surprisal corresponds to greater abruptness

for *past*, supporting our hypothesis. For *future*, lower surprisal corresponds to greater abruptness, suggesting that utterances making the target easier to predict tend to appear more abrupt.

Figure 4 shows the transition of *future* across dialogue turns. As the dialogue progresses, *future* values decrease. While the open system remains relatively stable, baseline and proposed start with higher values and decrease more sharply over turns. The downward trend suggests that continued induction makes the target more predictable over time.

## 6 Conclusion

In this study, we investigated the hypothesis that the surprisal theory can effectively model both the perceived induction and the user’s awareness of the target in target-guided conversation. Through dialogue experiments using a system that selects utterances based on SBIS-TGC, we confirmed that SBIS-TGC-guided selection reduced the perceived abruptness of system utterances. Moreover, a correlation was observed between surprisal-based induction scores and perceived abruptness. The decrease in target surprisal as dialogue turns progressed suggested that surprisal can serve as an indicator of progress toward the dialogue target. On the other hand, the results also revealed a trade-off among target achievement rate, perceived induction, and the user’s detectability of the system’s target.

## 7 Limitations and Ethical Concerns

This study focuses on techniques that enable a system to achieve its own goals without giving users a sense of being inducted. While such techniques can facilitate natural interaction in target-guided dialogue systems, their inappropriate use may undermine user autonomy or pose a risk of being used to manipulate user behavior. In particular, when setting goals that aim to induce behavioral change—such as in persuasive dialogue systems—principles such as disclosure of the system’s identity, obtaining user consent for the analysis of personal information and psychological tendencies, and ensuring the appropriateness and non-discriminatory nature of system responses are emphasized (Wang et al., 2019).

## Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 23K24910.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. [arXiv:2001.09977](https://arxiv.org/abs/2001.09977).
- Huy Quang Dao, Yang Deng, Khanh-Huyen Bui, Dung D. Le, and Lizi Liao. 2024. [Experience as source for anticipation and planning: Experiential policy learning for target-driven recommendation dialogues](#). In [Findings of the Association for Computational Linguistics: EMNLP 2024](#), pages 14179–14198, Miami, Florida, USA. Association for Computational Linguistics.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. [Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration](#). In [Findings of the Association for Computational Linguistics: EMNLP 2023](#), pages 10602–10621, Singapore. Association for Computational Linguistics.
- Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. [Proactive conversational ai: A comprehensive survey of advancements and opportunities](#). [ACM Trans. Inf. Syst.](#), 43(3).
- Prakhar Gupta, Harsh Jhamtani, and Jeffrey Bigham. 2022. [Target-guided dialogue response generation using commonsense and data augmentation](#). In [Findings of the Association for Computational Linguistics: NAACL 2022](#), pages 1301–1317, Seattle, United States. Association for Computational Linguistics.
- John Hale. 2001. A probabilistic early parser as a psycholinguistic model. In [Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies](#).
- Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Reinforcement learning of cooperative persuasive dialogue policies using framing. In [Proceedings of coling 2014, the 25th international conference on computational linguistics: technical papers](#), pages 1706–1717.
- Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. [Target-guided open-domain conversation planning](#). In [Proceedings of the 29th International Conference on Computational Linguistics](#), pages 660–668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). [Cognition](#), 106(3):1126–1177.

- Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. [arXiv:1812.07617](https://arxiv.org/abs/1812.07617).
- Anqi Liu, Bo Wang, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. [MTGP: Multi-turn target-oriented dialogue guided by generative global path with flexible turns](#). In [Findings of the Association for Computational Linguistics: ACL 2023](#), pages 259–271, Toronto, Canada. Association for Computational Linguistics.
- Marine Riou. 2015. [A methodology for the identification of topic transitions in interaction](#). *Discours*, 16. Published online on September 9, 2015. Accessed on December 4, 2024.
- Shiki Sato, Jun Baba, Asahi Hentona, Shinji Iwata, Akifumi Yoshimoto, and Koichiro Yoshino. 2025. Proactive user information acquisition via chats on user-favored topics. [arXiv:2504.07698](https://arxiv.org/abs/2504.07698).
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? how controllable attributes affect human judgments](#). In [Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 \(Long and Short Papers\)](#), pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.
- Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. [OTTers: One-turn topic transitions for open-domain dialogue](#). In [Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing \(Volume 1: Long Papers\)](#), pages 2492–2504, Online. Association for Computational Linguistics.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. [Target-guided open-domain conversation](#). In [Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics](#), pages 5624–5634, Florence, Italy. Association for Computational Linguistics.
- Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023. [Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation](#). In [Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing](#), pages 1132–1143, Singapore. Association for Computational Linguistics.
- Jian Wang, Dongding Lin, and Wenjie Li. 2024. [A target-driven planning approach for goal-directed dialog systems](#). *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10475–10487.
- Xuwei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). In [Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics](#), pages 5635–5649, Florence, Italy. Association for Computational Linguistics.
- Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. [Proactive human-machine conversation with explicit conversation goal](#). In [Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics](#), pages 3794–3804, Florence, Italy. Association for Computational Linguistics.
- Jingxuan Yang, Si Li, and Jun Guo. 2021. [Multi-turn target-guided topic prediction with Monte Carlo tree search](#). In [Proceedings of the 18th International Conference on Natural Language Processing \(ICON\)](#), pages 324–334, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLP AI).
- Koichiro Yoshino, Yu Suzuki, and Satoshi Nakamura. 2017. Information navigation system with discovering user interests. In [Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue](#), pages 356–359.
- Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Ee-Peng Lim, Minlie Huang, and Liqiang Nie. 2024. [Thoughts to target: Enhance planning for target-driven conversation](#). In [Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing](#), pages 21108–21124, Miami, Florida, USA. Association for Computational Linguistics.
- Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. [Towards topic-guided conversational recommender system](#). In [Proceedings of the 28th International Conference on Computational Linguistics](#), pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

## A Appendix

### A.1 Dialogue Examples

Below are example dialogues from the systems that performed induction during the dialogue experiments. In each example, the target topic is shown in **bold**.

baseline: Example where the target was achieved target-guided conversation)

User: I haven't had much time to read books lately.

System: It's tough not having time to read. You can still enjoy the **Harry Potter series** even in short intervals.

proposed: Example where the target was achieved target-guided conversation)

User: I can't seem to land the job I want.

System: That sounds tough. Do you have a favorite field?

User: Maybe something related to events.

System: Events are exciting! Sports events are especially lively.

User: That sounds fun, but I worry about getting sunburned.

System: That's understandable. There are indoor events too.

User: True. I'd prefer indoor events, especially in summer.

System: Indoor events like watching sports can be really enjoyable.

User: I haven't watched live sports before, so I don't really know what it's like.

System: Even beginners can enjoy sports spectating. **Ichiro's** games are especially popular.

### A.2 Prompts and Topics Used in the Dialogue Experiment

The following are the prompts used for selecting the system's target topics and generating target utterances, as well as the actual topics used in the experiment. In the human dialog experiment, 65 items were selected from the filtered 287-word.

Prompt used for selecting target topics target-guided conversation)

Please select one topic from the list below that many Japanese people are likely to know and find interesting. Only output a single word. {system\_topic\_cand}

Prompt used for generating target utterances target-guided conversation)

You are an AI that wants to talk with the user about the topic: "{target\_topic}". Please create an utterance that naturally introduces "{target\_topic}" within the flow of the dialogue. Keep the utterance within 20 words.

List of target topics used (Extract from actual topics, translation)

Shohei Ohtani, The 13 Lords of the Shogun, Detective Conan (Anime Episode List), Japan Airlines Flight 123 Crash, BTS (Music Group), SPY×FAMILY, Jujutsu Kaisen, Kamakura Shogunate, Haruma Miura, Detective Conan (Movie Series), Takuya Kimura, Fullmetal Alchemist, Twitter, Top Gun: Maverick, Ken Shimura, Detective Conan, World War II, Demon Slayer: Kimetsu no Yaiba, Fast & Furious Series, Detective Conan (Anime), Yui Aragaki, Shin Ultraman, Ichiro, NHK Morning Drama Series, Kenshi Yonezu, Nishikigoi (Comedy Duo), World War I, Kingdom (Manga), X JAPAN, ONE PIECE, Characters in Detective Conan, Takumi Saito, Nogizaka46, Ultraman, Harry Potter Series, Three Sacred Treasures, HUNTER×HUNTER, Masaki Suda, Ukraine, Elon Musk, Oda Nobunaga, Golden Kamuy, JoJo's Bizarre Adventure, Top Gun (Film), Fumio Kishida, Itaewon Class, Kaguya-sama: Love is War, My Family, United States of America, HIKAKIN, Avataro Sentai Donbrothers, Tokyo Revengers, List of Harry Potter Characters, Hey! Say! JUMP, YouTube, Shun Oguri, Wikipedia, Yoshitoki Hojo, Shiretoko Sightseeing Boat Incident, Ponyo on the Cliff by the Sea, Haruka Ayase, King & Prince, The Quintessential Quintuplets, Mr.Children, Hideaki Anno