



 Latest updates: <https://dl.acm.org/doi/10.1145/3765766.3765873>

POSTER

Can LLMs be Surprised? Evaluation and Analysis of Surprise Expression of LLMs



PDF Download
3765766.3765873.pdf
04 January 2026
Total Citations: 0
Total Downloads: 0

Published: 10 November 2025

[Citation in BibTeX format](#)

HAI '25: International Conference on
Human-Agent Interaction
November 10 - 13, 2025
Yokohama, Japan

Can LLMs be Surprised? Evaluation and Analysis of Surprise Expression of LLMs

Motoori Takeuchi

Kyoto University

Kyoto, Japan

takeuchi@sap.ist.i.kyoto-u.ac.jp

Keiko Ochi

Kyoto University

Kyoto, Japan

ochi.keiko.5f@kyoto-u.ac.jp

Koji Inoue

Kyoto University

Kyoto, Japan

inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara

Kyoto University

Kyoto, Japan

kawahara@i.kyoto-u.ac.jp

Abstract

We investigate methods for virtual agents or humanoid robots to express surprise in dialogue by using LLMs. We created a newly annotated dialogue dataset focusing on surprise. Our findings indicate that accurately expressing surprise in dialogue is a challenging task. They also suggest directions for improvement—such as more appropriate modeling of commonness—and identify the feature of surprise itself.

CCS Concepts

- Human-centered computing → Human computer interaction (HCI).

Keywords

Dialogue systems, LLM, Emotion, Surprise expression

ACM Reference Format:

Motoori Takeuchi, Koji Inoue, Keiko Ochi, and Tatsuya Kawahara. 2025. Can LLMs be Surprised? Evaluation and Analysis of Surprise Expression of LLMs. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3765766.3765873>

1 Introduction

In recent years, research related to emotions is active [1, 13] and various dialogue systems have been developed with the aim of expressing “empathy” towards users [2, 3, 6, 7, 9] and . However, many people would feel uncomfortable when inanimate robots expressed emotions, such as saying “I want to try that food too” or “I feel so sad I could cry,” in response to user utterances. Nonetheless, according to Descartes, one of the basic human emotions discussed in “The Passions of the Soul” is “surprise,” which is said to be triggered by “unexpected impression [14] ,” and it is expressed when encountering something that deviates from common sense. Therefore, it is expected that LLM-based dialogue systems capable of retaining vast amounts of data and knowledge are less likely

Table 1: Example of Japanese Empathetic Dialogues

Speaker	utterance
A	Listen, my son did it!
B	Did he do something wrong?
A	No, he got the top score in his grade on the exam.
B	Oh, your son really worked hard! (surprise)

to evoke discomfort when expressing such surprise compared to other kinds of emotions. We aim to develop a dialogue system that naturally expresses surprise, much like human beings. As the first step towards this goal, we conduct experiments on how LLMs express surprise in Japanese text dialogues and analyzes the features and tendencies of surprise responses by LLMs with three methods.

2 Dataset

We utilize a portion of the Japanese Empathetic Dialogue [10] dataset. The dialogue data is composed of two speakers, A and B, alternating speaking for two turns each, as shown in Table 1. Additionally, referring to several linguistic studies [4, 15] , we annotated the last turn of speaker B’s utterance with a “surprised” label. Using this method, we prepared 50 instances of positive labels (“surprised”) and 500 instances of negative labels (“not surprised”).

3 Methods for Surprise Expression

• Direct prediction using LLM (LLM-Direct)

First, we provide an LLM with a dialogue context up to A’s utterance in the second turn, and then LLM directly judges whether B would express a surprised reaction. The prompt included examples for few-shot learning as well as instructions for the Chain of Thought (CoT) [12]. In CoT, before outputting the judgment of the surprised reaction, the LLM is also instructed to output the reason for the judgment.

• Dialogue prediction by LLM (LLM-Predict)

The second method aims to replicate the surprise reaction that occurs when the content of the interlocutor’s utterance differs from what was predicted by the listener. Specifically, the LLM generates multiple utterance predictions corresponding to the interlocutor’s final input utterance. Then, the actual interlocutor’s utterance is compared with the predicted utterances, and if none of them are



This work is licensed under a Creative Commons Attribution 4.0 International License.
HAI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2178-6/25/11
<https://doi.org/10.1145/3765766.3765873>

Table 2: Experimental results of surprise expression models

Method	Recall	Precision	F1 score	False surprise
LLM-Direct	72.0	26.3	38.6	20.1
LLM-Direct (CoT)	66.7	20.6	31.4	25.9
LLM-Direct (Few-shot)	71.3	32.6*	44.7*	14.8*
LLM-Direct (CoT + Few-shot)	69.3	28.7	40.5	17.3
LLM-Predict	12.0	5.0	7.1	22.6
BERT-Finetuning	50.0	31.3	38.5	11.0

similar, it is judged as “surprise.” In this study, Comparison involves entailment recognition [16] and cosine similarity [11], considering the unpredictability of utterances [8]. Prior researches—such as Zenkyo et al. [5, 17]—have explored mechanisms that generate surprise by comparing predictions with actual results, while LLMs are not employed.

- **BERT finetuning**

The approach of fine-tuning a small-scale model was also examined, instead of using zero-shot or few-shot learning with LLMs. Here, we use a BERT model to judge whether or not to express a surprise reaction to a given utterance. For the base BERT model, we use “bert-base-japanese-whole-word-masking”¹.

4 Evaluations (Setup & Results)

The methods compared are as follows. The evaluation metrics were precision, recall, and F1-score of surprise response, as well as the rate at which surprises were incorrectly detected (hereafter referred to as the false surprise rate). We used GPT-4o as the LLM.

- **LLM-Direct:** Zero-shot GPT-4o
- **LLM-Direct (Few-shot):** 30 Few-shot GPT-4o
- **LLM-Direct (CoT):** Chain of Thought and Zero-shot GPT-4o
- **LLM-Direct (CoT + Few-shot):** Chain of Thought and 30 Few-shot GPT-4o
- **LLM-Predict:** Predict the next utterance of a dialogue using GPT-4o and make judgments based on the differences from the actual utterance.
- **BERT-Finetuning:** Fine-tuning BERT with training data

The results are shown in Table 2. The values for LLM-Direct, LLM-Direct (Few-shot), LLM-Direct (CoT), and LLM-Direct (CoT+Few-shot) represent the average results obtained from three experimental trials. Additionally, a *t*-test was conducted on three LLM-Directs (from LLM-Direct (CoT) up to LLM-Direct (CoT + Few-shot)) with the baseline LLM-Direct to test whether there is a significant difference in the metrics.

5 Analysis of Reasons of Surprise

By classifying the sources of surprise, the following main factors were identified: difficulty level, fear, sadness, and others. This confirms that LLMs are capable of inferring various causes of surprise from conversational context. Additionally, we analyze instances as shown in Table 3, where the method using CoT made incorrect output. From Table 3, it is speculated that incorrect judgments occur in utterances that do not fit the general reasons for surprise and are deemed “common” by the LLM, which has learned from vast

¹<https://github.com/cl-tohoku/bert-japanese>

Table 3: LLM-Direct(CoT + Few-shot) incorrect result

Speaker	utterance
A	There’s a troublemaker kid in the neighborhood, and I just can’t handle it anymore.
B	What? What kind of mischief is he up to?
A	He plays ball on the street all the time, and to top it off, he doesn’t even apologize when he hits someone’s door with the ball.
LLM output	
The presence of such mischievous kids in the neighborhood is indeed a troublesome situation for the local community. However, it is not uncommon for children to occasionally break the rules. While it may be a significant source of stress for A, it is not surprising.	

amounts of data. Some instances demonstrate that even with the same LLM, there can be differing interpretations of whether something is considered “surprising” or not, even for identical content. Furthermore, some explanations for judgments included statements like “there are partial elements of surprise, but overall it is not surprising.” It is still challenging for LLMs to make binary decisions on whether to express surprise or not for many cases. The use of a few-shot example can affect LLM judgments for such cases.

These experimental methods also have several limitations. First, surprise often depends on background knowledge and is expressed multimodally, so predicting it from textual context has inherent limitations. Second, since the proposed methods achieved relatively low scores and were tested under limited conditions, it remains uncertain whether they would perform similarly under different or more practical settings. Finally, human-agent interactions and evaluations may provide valuable directions for improvement, helping to deepen our understanding of surprise.

6 Conclusion

It was confirmed that LLMs can express surprise reactions, including reasons, in many situations. We will use these evaluation results to help LLMs acquire common sense closer to that of humans, leading to the expression of natural surprise reactions. Analysis of surprise expressions also leads to clarifying the characteristics of empathetic surprise and its relationship with human memory.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011.

References

[1] Md Ehtesham-Ul-Haque, Jacob D’Rozario, Rudaiba Adnin, Farhan Tanvir Utshaw, Fabiha Tasneem, Israt Jahan Shefa, and ABM Alim Al Islam. 2024. EmoBot: Artificial emotion generation through an emotional chatbot during general-purpose conversations. *Cognitive Systems Research* 83, 101168.

[2] Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI* 9 (2022), 933261.

[3] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 118–127.

[4] Eri Kato. 2020. On Interjection ‘uwa’ in Daily Conversation. In *Proceedings of Language Resources Workshop*, Vol. 5. 283–292 (in Japanese).

[5] Jun Kogami, Yutaka Miyaji, and Ken Tomiyama. 2010. Construction and Evaluation of a Virtual KANEI System for Robots. *Transactions of Japan Society of Kansei Engineering* 9, 4 (2010), 601–609.

[6] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *International Conference on Computational Linguistics (COLING)*. 4454–4466.

[7] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 8968–8979.

[8] Yuya Nakano, seiya Kawano, koichiro Yoshino, and Satoshi Nakamura. 2020. taiwa ni yotte aimaisei kaishou wo okonau sitsumon outou (Question Answering with Dialogue-Based Disambiguation). *The Special Interest Group Technical Reports NL of IPSJ* 2020, 3 (2020), 1–5 (in Japanese).

[9] Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the Power of Large Language Models for Empathetic Response Generation: Empirical Investigations and Improvements. In *Findings of Empirical Methods in Natural Language Processing (EMNLP-Findings)*. 6516–6528.

[10] Hiroaki Sugiyama, Hiromi Narimatsu, Masahiro Mizukami, Tsunehiro Arimoto, Yuya Chiba, Toyomi Meguro, and Hideharu Nakajima. 2021. Analysis of Subjective Evaluation for Fine-tuning Methods of Transformer encoder-decoder based Conversational Systems. In *Proceedings of The 35th annual conference of the Japanese society for artificial intelligence*. (in Japanese).

[11] Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information* 11, 9 (2020), 421.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS) 35* (2022), 24824–24837.

[13] Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint*. arXiv:2505.20511.

[14] Ichiro Yamane. 2005. Phenomenology of Surprise. *Journal of Sugiyama Jogakuen University. Humanities* 36 (2005), 13–28 (in Japanese).

[15] Nao Yoshida, Katsuya Takanashi, and Yasuharu Den. 2009. taiwa ni okeru aiduchi hyougen no nintei to sono mondaiten ni tsuite (Identification of Backchannel Expressions in Dialogue and Its Issues). *Proceedings of the Fifteenth Annual Meeting of the Association for Natural Language Processing* (2009), 430–433 (in Japanese).

[16] Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a Natural Language Inference Dataset Using Machine Translation. *The Special Interest Group Technical Reports NL of IPSJ 2020-NL-244*, 6 (2020), 1–8 (in Japanese).

[17] Masaki Zenkyoh and Ken Tomiyama. 2011. Surprise generator for virtual KANSEI based on human surprise characteristics. In *Human Interface and the Management of Information. Interacting with Information: Symposium on Human Interface, Held as Part of HCI International*. 190–198.