

Supplementary: Integrating Stickers into Multimodal Dialogue Summarization: A Novel Dataset and Approach for Enhancing Social Media Interaction

Yuanchen Shi
Soochow University
School of Computer Science and Technology
Suzhou, China
20227927002@stu.suda.edu.cn

Fang Kong*
Soochow University
School of Computer Science and Technology
Suzhou, China
kongfang@suda.edu.cn

ACM Reference Format:

Yuanchen Shi and Fang Kong*. 2024. Supplementary: Integrating Stickers into Multimodal Dialogue Summarization: A Novel Dataset and Approach for Enhancing Social Media Interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3664647.3680978>

1 Dialogue Length Statistics

Table 1 summarizes the syntactic tree depth of sentences used to generate dialogues in the MCDSCS dataset, along with the corresponding dialogue lengths.

Table 1: The number and percentage of syntactic dependency tree depth and dialogue length statistics.

Tree Depth/Dialogue Length	Quantity	Percent
2 / Short	922	16.68%
3 / Short	1805	32.66%
4 / Medium	1618	29.27%
5 / Medium	748	13.53%
6+ / Long	434	7.85%

2 Dialogue Generation COT Example

Figure 1 illustrates an example of generating a social media chat dialogue that adhere to the constraints of dialogue length, scenario, and characteristics using an initial sentence and GPT4 through COT. It also includes automatic logical error checking.

* Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
<https://doi.org/10.1145/3664647.3680978>

Initial sentence: 医生说很常见，在青少年中尤其如此。(The doctor said it's quite common, especially among teenagers.)	
Syntactic tree depth: 4	Dialogue length: Medium
Characteristics: Colloquial, Natural	Scenario: Social Media
Chain-Of-Thought: 生成一段长度中等的社交媒体上的闲聊对话，包含句子：“医生说很常见，在青少年中尤其如此。”要求对话自然且口语化，并且检查对话有没有逻辑错误。(Generate a chat dialogue of medium length on social media including the sentence, "The doctor said it's quite common, especially among teenagers." The dialogue needs to be natural and colloquial. Then check if there's any logical error in the dialogue.)	
Dialogue A: 我听说安迪最近总是感到疼痛。 (A: I heard that Andy has been experiencing pain recently.) B: 是的，他去看了医生。 (B: Yes, he went to see a doctor.) A: 医生说了什么? (A: What did the doctor say?) B: 他说安迪是发育期的疼痛。 (B: He said that Andy is experiencing developmental pain.) A: 这种情况常见吗? (A: Is this situation common?) B: 医生说很常见，在青少年中尤其如此。 (B: The doctor said it's quite common, especially among teenagers.) A: 那么，有什么特别的处理方法吗? (A: So, are there any specific treatment options?) B: 医生建议进行适当的运动和保持良好的饮食。 (B: The doctor recommended engaging in appropriate exercise and maintaining a healthy diet.)	
There's no logical error in this dialogue.	

Figure 1: An COT example for textual dialogue generation using GPT4, along with logical check.

3 Summary Generation COT Example

Figure 2 illustrates an example of extracting summaries from multimodal dialogues containing stickers using COT and the GPT4-Vision model. Compared to simple summarization prompts that consider both text and images, our COT produces more natural and conversational summaries, seamlessly integrating sticker content directly into the textual summary rather than rigidly mentioning specific stickers.



Figure 2: An COT example for getting summaries of multimodal dialogues containing stickers.

4 Experimental Results of VLMs

We randomly select 300 pieces of data from the MCDSCS dataset three times and conduct experiments using LLaVA-V1.6-7b, Yi-6b, and Qwen-VL, respectively, for each selection. As shown in Table 2, the results of the experiments are averaged over the three trials. Since these VLMs emphasize images, the generated summaries are significantly influenced by sticker images, resulting in more abstract summaries that have difficulty overlapping with the standard answers in terms of n-grams. Consequently, the BLEU and Meteor scores are very low, almost zero. Therefore, we only report the ROUGE scores here.

Table 2: Results of VLM experiments.

Model	R1	R2	RL
LLaVA-v1.6-7b	35.43	12.42	26.99
Yi-6b	10.05	2.76	7.54
Qwen-VL	16.96	4.74	13.61

5 Implementation Details

Each of the fine-tuned models is trained for 100 epochs on two NVIDIA-RTX3090 GPUs with the same parameters: training batch size of 16; validation and test batch size of 4; learning rate of 5×10^{-5} .