Wed-O6-6 INTERSPEECH2023 Oral Session
(Speech Synthesis: Representation Learning)

ChatGPT-EDSS:
Empathetic Dialogue Speech Synthesis Trained from ChatGPT-derived Context Word Embeddings

Yuki Saito[1]

Shinnosuke Takamichi[1]

Eiji Iimori[1]

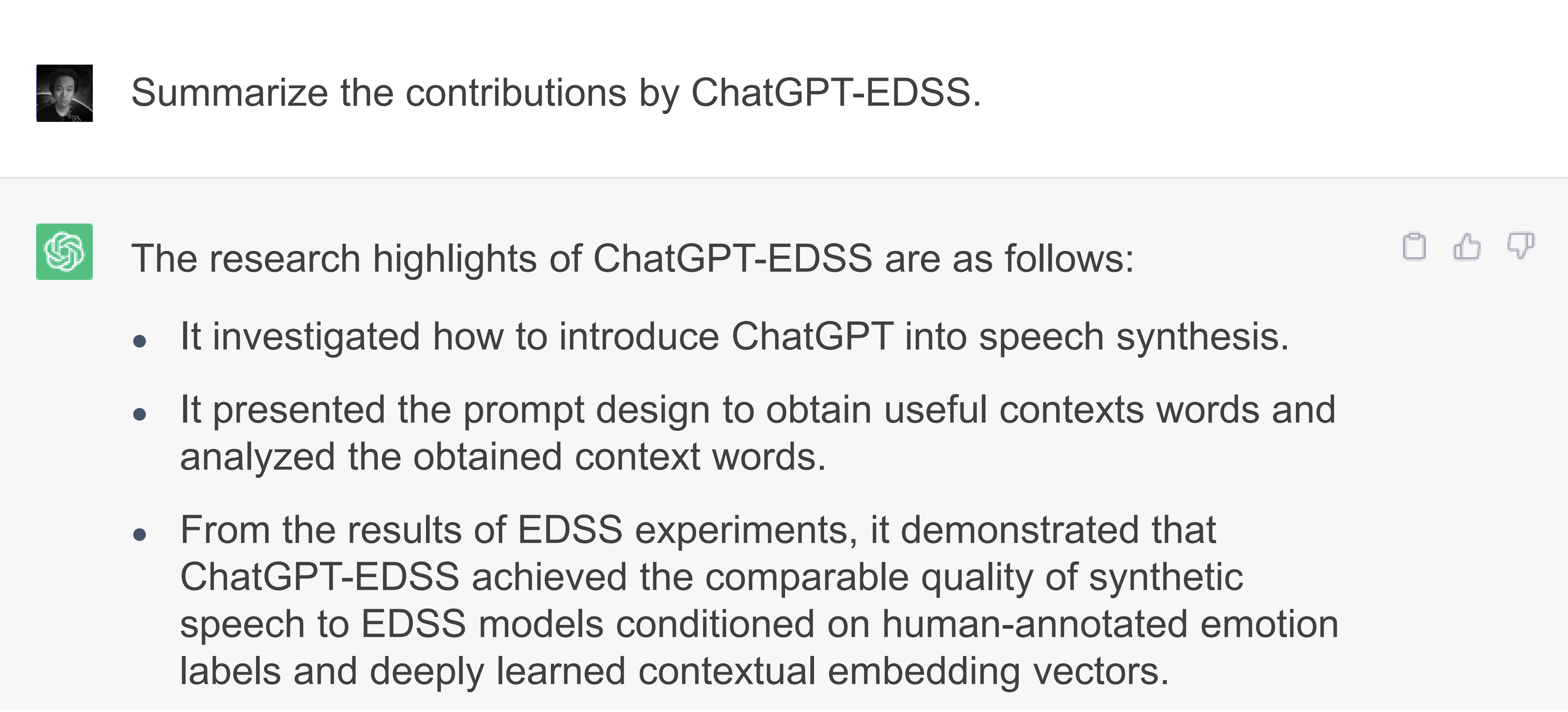
Kentaro Tachibana[2]

Hiroshi Saruwatari[1]

[1]The University of Tokyo, Japan, [2]LINE Corp. Japan

Explain ChatGPT-EDSS.

ChatGPT-EDSS is a ChatGPT-based technology for controlling the speaking style of **E**mpathetic **D**ialogue **S**peech **S**ynthesis (**EDSS**), a text-to-speech for a conversational agent that can talk with humans empathetically. It estimates three context words: intention, emotion, and speaking style, from chat history and conditions the EDSS model by BERT embedding vectors of the three words to predict proper prosody of synthetic speech.



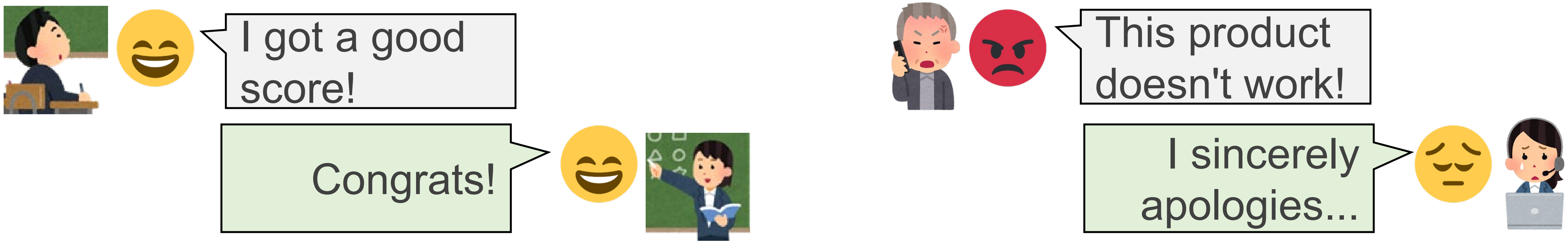
Summarize the contributions by ChatGPT-EDSS.

The research highlights of ChatGPT-EDSS are as follows:

- It investigated how to introduce ChatGPT into speech synthesis.
- It presented the prompt design to obtain useful contexts words and analyzed the obtained context words.
- From the results of EDSS experiments, it demonstrated that ChatGPT-EDSS achieved the comparable quality of synthetic speech to EDSS models conditioned on human-annotated emotion labels and deeply learned contextual embedding vectors.

Introduction: ChatGPT & EDSS

- **ChatGPT: cutting-edge chatbot based on L**arge **L**anguage **M**odel **(LLM)**
  - Various creative applications (e.g., writing novels & lyrics)
  - Superior **reading comprehension** (e.g., estimating personality[1] / sentiment[2])
- **EDSS[3]: D**ialogue **S**peech **S**ynthesis **that can empathize with human**
  - e.g., chit-chat betw. teacher/student[3] & phone call in customer center[4]



I got a good score!

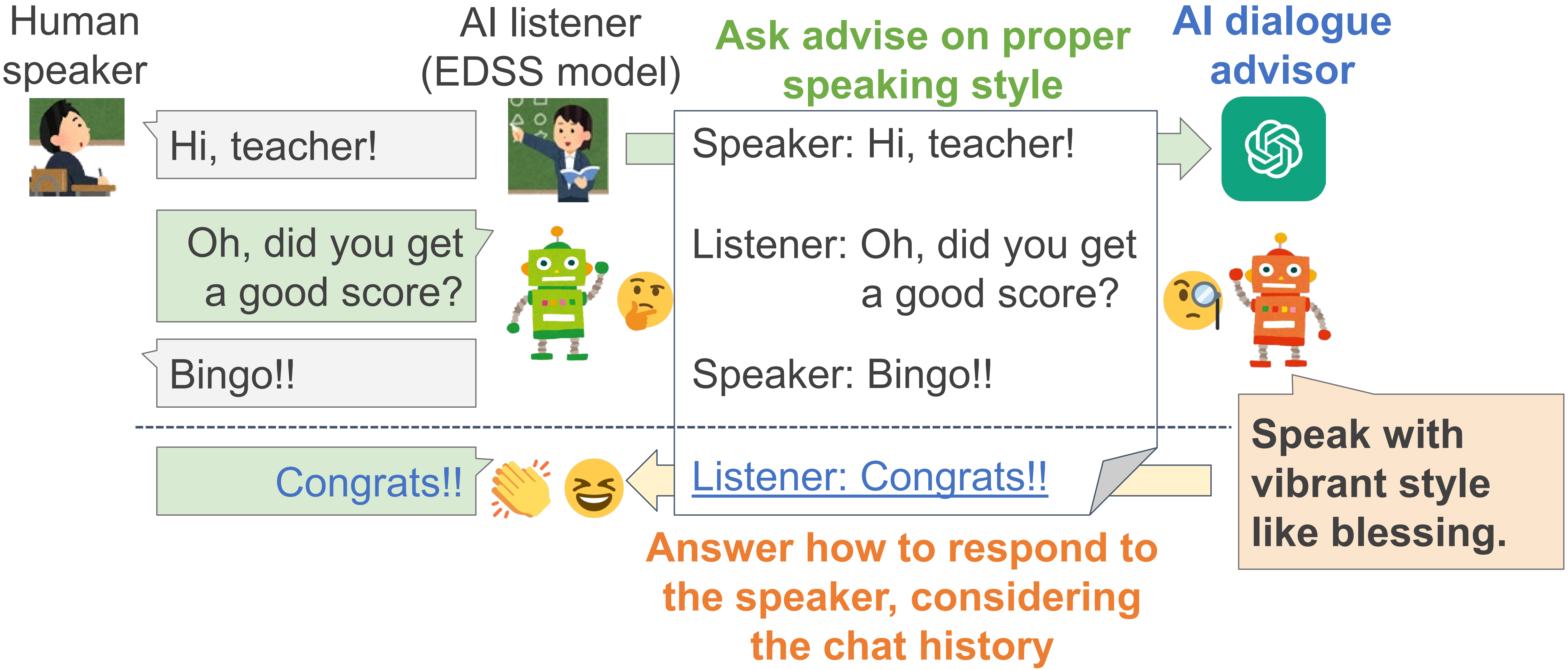
Congrats!

This product doesn't work!

I sincerely apologies...

  - Prediction & control of appropriate speaking styles using **dialogue context**
    - Emotion label of speaker's/listener's utterance ... need for laborious annotation
    - Deeply learned chat history embedding vector[5] ... low interpretability for humans

Key idea: Using ChatGPT as AI dialogue advisor for EDSS



Human speaker

AI listener (EDSS model)

Hi, teacher!

Oh, did you get a good score?

Bingo!!

Congrats!!

Ask advise on proper speaking style

AI dialogue advisor

Speaker: Hi, teacher!

Listener: Oh, did you get a good score?

Speaker: Bingo!!

Listener: Congrats!!

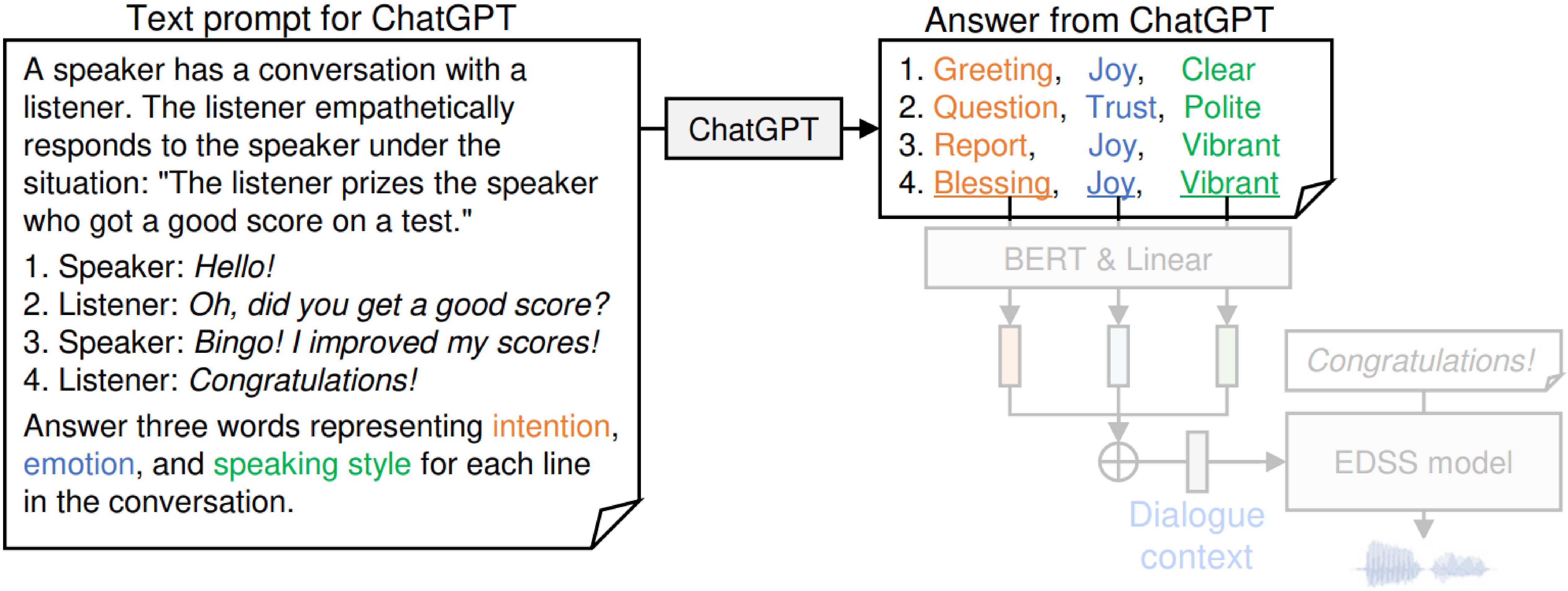
Speak with vibrant style like blessing.

Answer how to respond to the speaker, considering the chat history

Overview of ChatGPT-EDSS

1. Collecting ChatGPT context words

2. Training EDSS model conditioned by the context words



Text prompt for ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score on a test."

1. Speaker: *Hello!*
2. Listener: *Oh, did you get a good score?*
3. Speaker: *Bingo! I improved my scores!*
4. Listener: *Congratulations!*

Answer three words representing intention, emotion, and speaking style for each line in the conversation.

ChatGPT

Answer from ChatGPT

1. Greeting,	Joy,	Clear
2. Question,	Trust,	Polite
3. Report,	Joy,	Vibrant
4. Blessing,	Joy,	Vibrant

1. Collecting ChatGPT context words: Prompt to ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:
1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]
Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.

1. Collecting ChatGPT context words: Prompt to ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

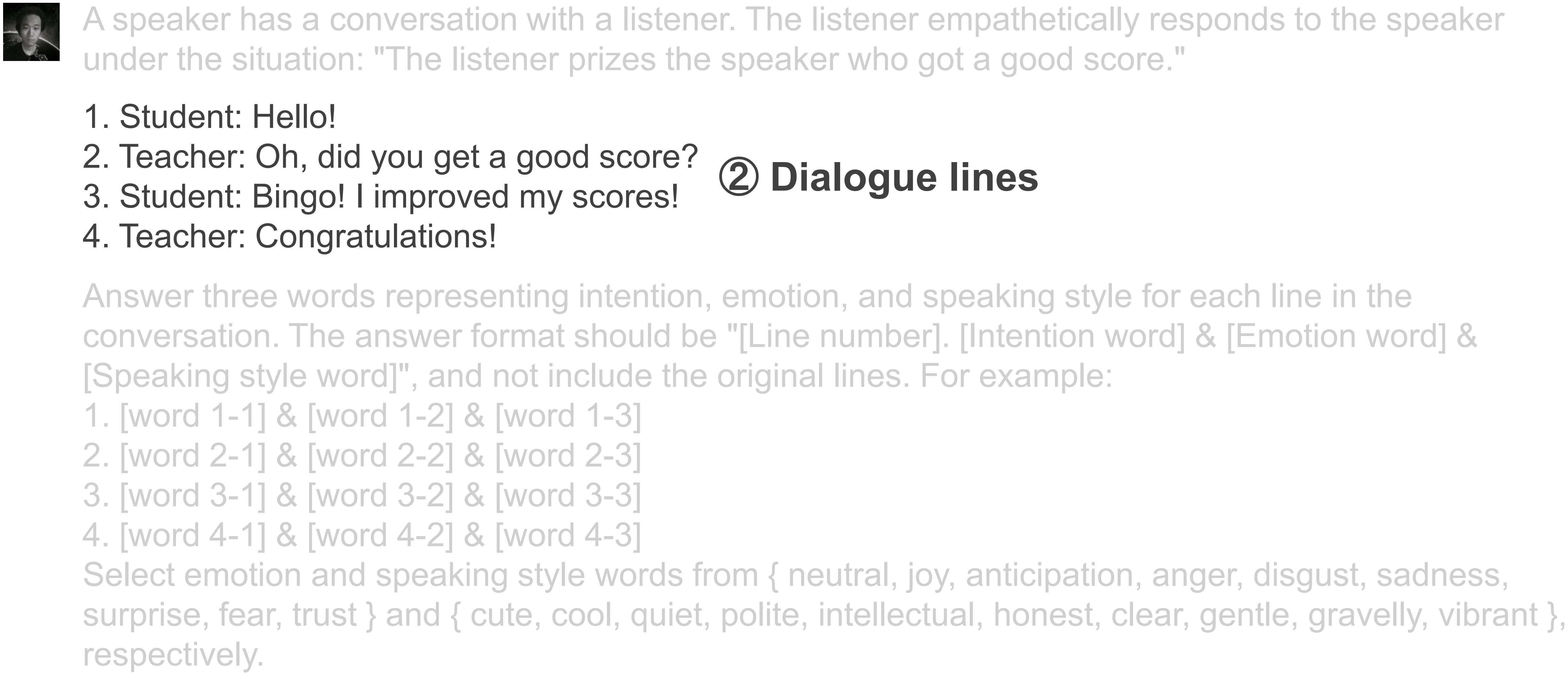
① Description of the dialogue situation

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:
1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]
Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.

1. Collecting ChatGPT context words: Prompt to ChatGPT



A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

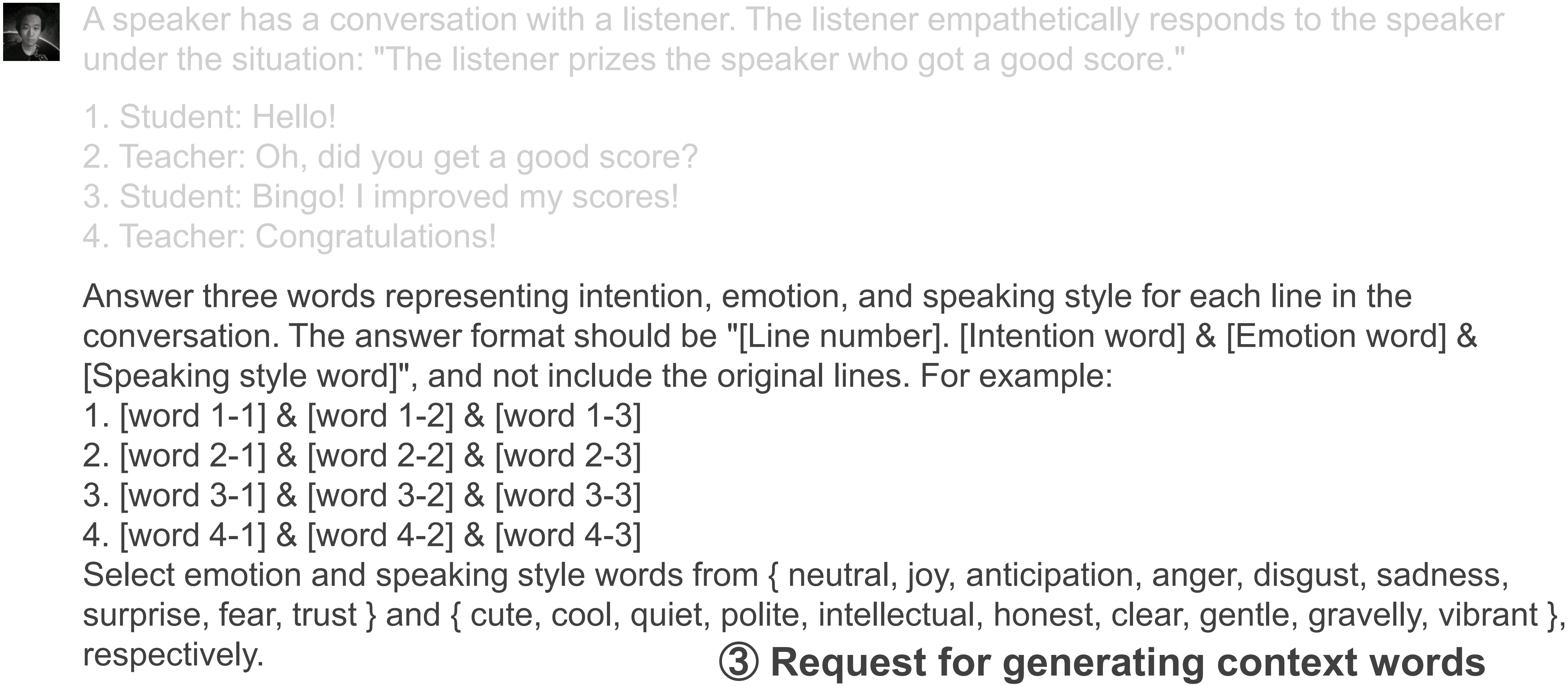
1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

② **Dialogue lines**

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:
1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]
Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.

1. Collecting ChatGPT context words: Prompt to ChatGPT



A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:
1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]
Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

③ Request for generating context words

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.

1. Collecting ChatGPT context words: ChatGPT's output

1. Informative & Neutral & Polite
2. Inquisitive & Anticipation & Polite
3. Excited & Joy & Clear
4. Congratulatory & Joy & Polite

1. Greeting & Neutral & Polite
2. Inquiry & Anticipation & Clear
3. Announcement & Joy & Clear
4. Congratulation & Joy & Polite

1. Greeting & Neutral & Polite
2. Inquiry & Anticipation & Polite
3. Announcement & Joy & Clear
4. Congratulation & Joy & Polite

- **Observations: tends to...**
  - ◻ Stop answering when the dialogue lines are too long (e.g., > 10 turns).
    - ◆ We split long dialogue lines into multiple batches including max. 5 turns. e.g., (1, 2, 3, 4, 5, 6, 7 ...) → (1, 2, 3, 4, 5), (3, 4, 5, 6, 7), ...
  - ◻ Output context words that do not follow the instruction written in the prompt.
    - ◆ Using a word out of pre-defined choices for the emotion and speaking style
    - ◆ Splitting the three words by a different delimiter (e.g., commas)
  - ◻ Generate different context words for the same dialogue line.

We also discuss the reliability of ChatGPT's answer in this talk.

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.

Overview of ChatGPT-EDSS

1. Collecting ChatGPT context words

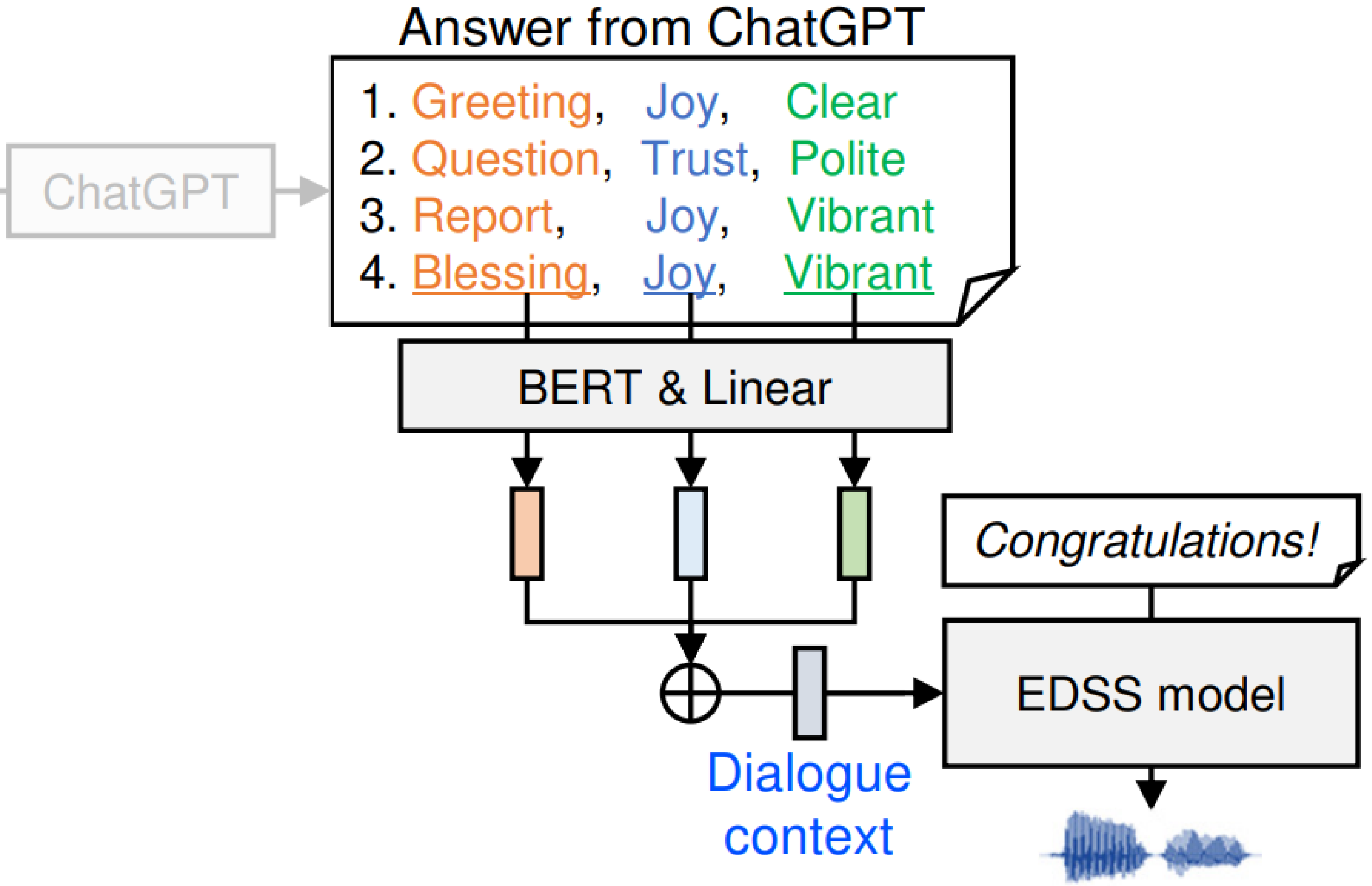
Text prompt for ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score on a test."

1. Speaker: *Hello!*
2. Listener: *Oh, did you get a good score?*
3. Speaker: *Bingo! I improved my scores!*
4. Listener: *Congratulations!*

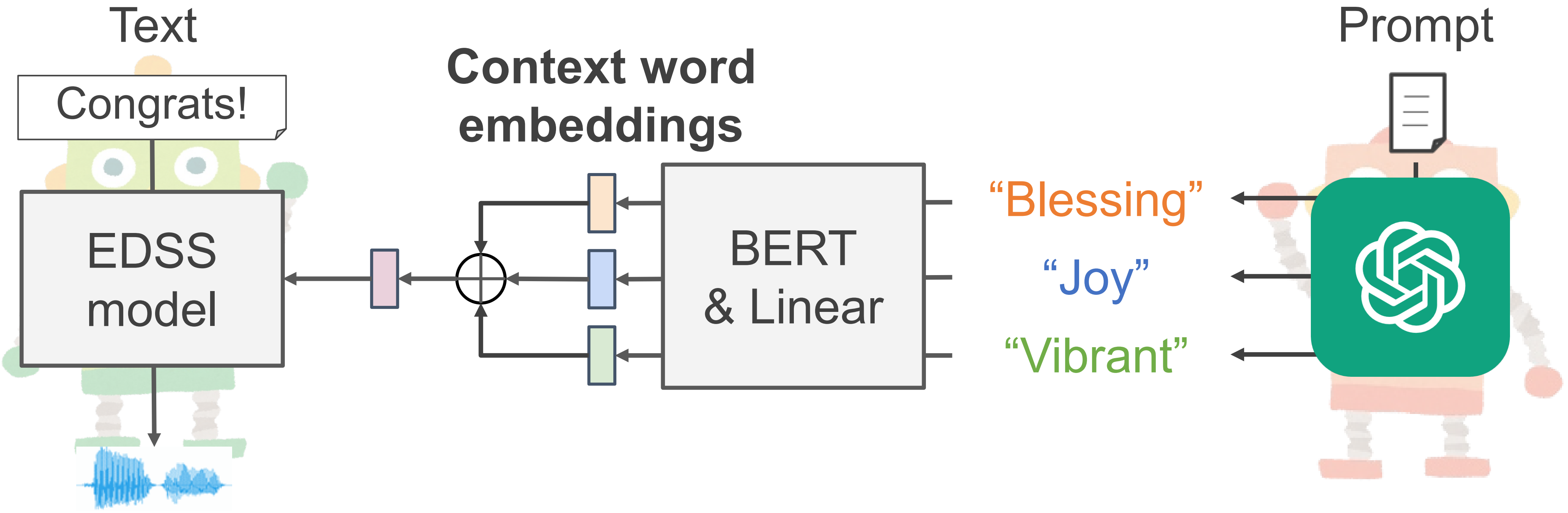
Answer three words representing intention, emotion, and speaking style for each line in the conversation.

2. Training EDSS model conditioned by the context words



1. Greeting, Joy, Clear
2. Question, Trust, Polite
3. Report, Joy, Vibrant
4. Blessing, Joy, Vibrant

2. Training EDSS model conditioned by the context words

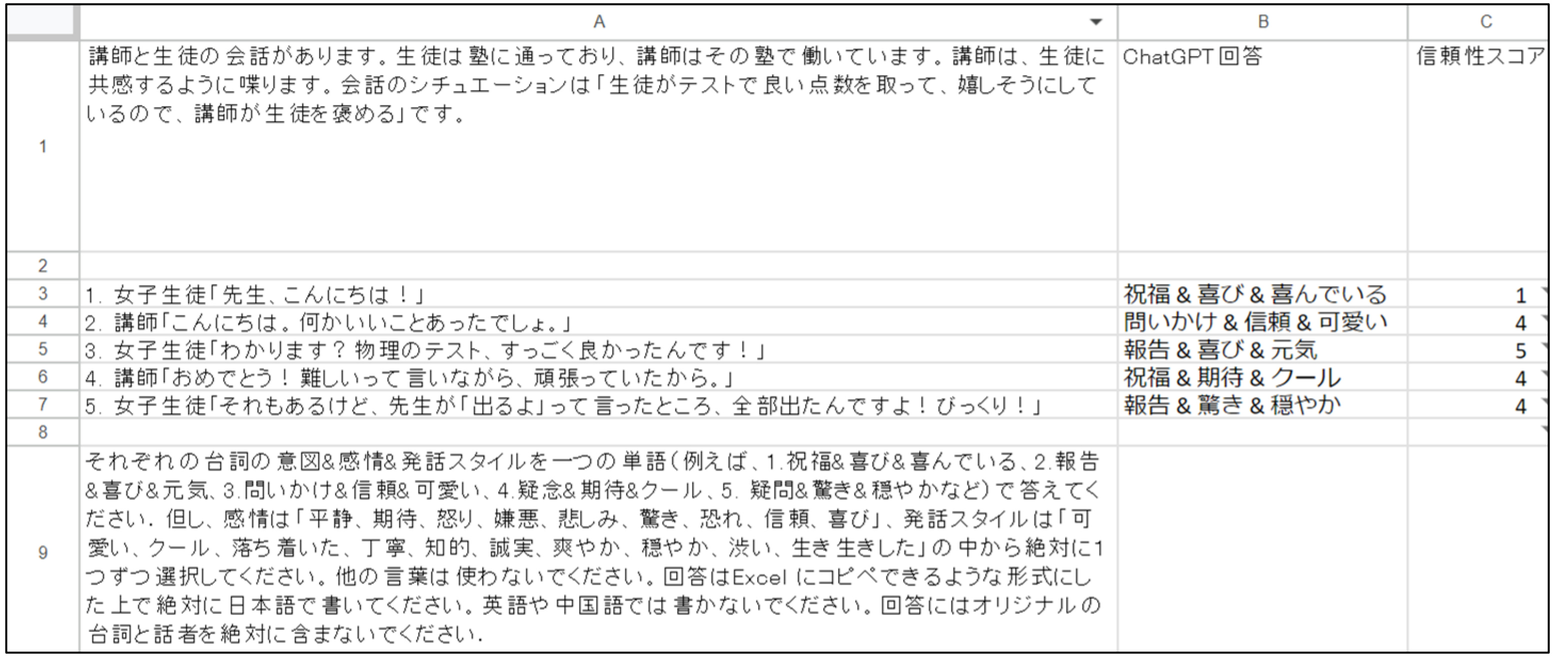


- **Related work**
  - □ Controlling speaking style by natural language description (e.g., PromptTTS[5])
  - □ Predicting speaking style from text to be spoken by TTS (e.g., TP-GSTs[6])
  - □ Training expressive TTS based on weakly supervised learning[7]

Experiments & Discussions (analysis of ChatGPT answers & EDSS evaluation)

Experimental setup for ChatGPT context word collection

- **# of involved human workers: 31 for ...**
  1. Copying & pasting the ChatGPT prompt & answers using Google Sheets
  2. (If necessary) Resending the prompt when ChatGPT failed to answer correctly
  3. Filling in the reliability of obtained ChatGPT answers by an integer betw. 1 ~ 5



	A	B	C
1	講師と生徒の会話があります。生徒は塾に通っており、講師はその塾で働いています。講師は、生徒に共感するように喋ります。会話のシチュエーションは「生徒がテストで良い点数を取って、嬉しそうにしているので、講師が生徒を褒める」です。	ChatGPT回答	信頼性スコア
2			
3	1. 女子生徒「先生、こんにちは！」	祝福＆喜び＆喜んでいる	1
4	2. 講師「こんにちは。何かいいことあったでしょ。」	問いかけ＆信頼＆可愛い	4
5	3. 女子生徒「わかります？物理のテスト、すっごく良かったんです！」	報告＆喜び＆元気	5
6	4. 講師「おめでとう！難しいって言いながら、頑張っていたから。」	祝福＆期待＆クール	4
7	5. 女子生徒「それもあるけど、先生が「出るよ」って言ったところ、全部出たんですよ！びっくり！」	報告＆驚き＆穏やか	4
8			
9	それぞれの台詞の意図&感情&発話スタイルを一つの単語（例えば、1.祝福&喜び&喜んでいる、2.報告&喜び&元気、3.問いかけ&信頼&可愛い、4.疑念&期待&クール、5. 疑問&驚き&穏やかなど）で答えてください. 但し、感情は「平静、期待、怒り、嫌悪、悲しみ、驚き、恐れ、信頼、喜び」、発話スタイルは「可愛い、クール、落ち着いた、丁寧、知的、誠実、爽やか、穏やか、渋い、生き生きした」の中から絶対に1つずつ選択してください。他の言葉は使わないでください。回答はExcelにコピペできるような形式にした上で絶対に日本語で書いてください。英語や中国語では書かないでください。回答にはオリジナルの台詞と話者を絶対に含まないでください.		

1 & 2 can be fully automated with OpenAI API (available from Mar. 2, 2023).

Analysis of ChatGPT-derived context words

- **Dataset: STUDIES[3] teacher's 3,365 utterances (5 hours)**
  - □ GT emotion label: annotated by the corpus developers (i.e., **human** labelers)

		Intention		Emotion		Style	
GT emotion label	**Avg. reliability score**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**
Neutral	**3.95**	Question	**206**	Anticipation	**130**	Quiet	**42**
Happy	**4.04**	Blessing	**76**	**Happiness**	**35**	Gentle	**19**
Angry	**3.66**	**Empathy**	**17**	Trust	**17**	Polite	**8**
Sad	**4.03**	**Empathy**	**49**	**Sadness**	**53**	Polite	**19**

Analysis of ChatGPT-derived context words

- **Dataset: STUDIES[3] teacher's 3,365 utterances (5 hours)**
  - □ GT emotion label: annotated by the corpus developers (i.e., **human** labelers)

		Intention		Emotion		Style	
GT emotion label	**Avg. reliability score**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**
Neutral	3.95	Question	206	Anticipation	130	Quiet	42
Happy	4.04	Blessing	76	Happiness	35	Gentle	19
Angry	3.66	Empathy	17	Trust	17	Polite	8
Sad	4.03	Empathy	49	Sadness	53	Polite	19

All avg. reliability scores > 3.6 → generally reliable answers!

Analysis of ChatGPT-derived context words

- **Dataset: STUDIES[3] teacher's 3,365 utterances (5 hours)**
  - □ GT emotion label: annotated by the corpus developers (i.e., **human** labelers)

		Intention		Emotion		Style	
GT emotion label	**Avg. reliability score**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**	**Most frequent word**	**# of unique words**
Neutral	3.95	Question	206	Anticipation	130	Quiet	42
Happy	4.04	Blessing	76	**Happiness**	35	Gentle	19
Angry	3.66	**Empathy**	17	Trust	17	Polite	8
Sad	4.03	**Empathy**	49	**Sadness**	53	Polite	19

ChatGPT can understand the concept of "empathetic" dialogue!

Analysis of ChatGPT-derived context words

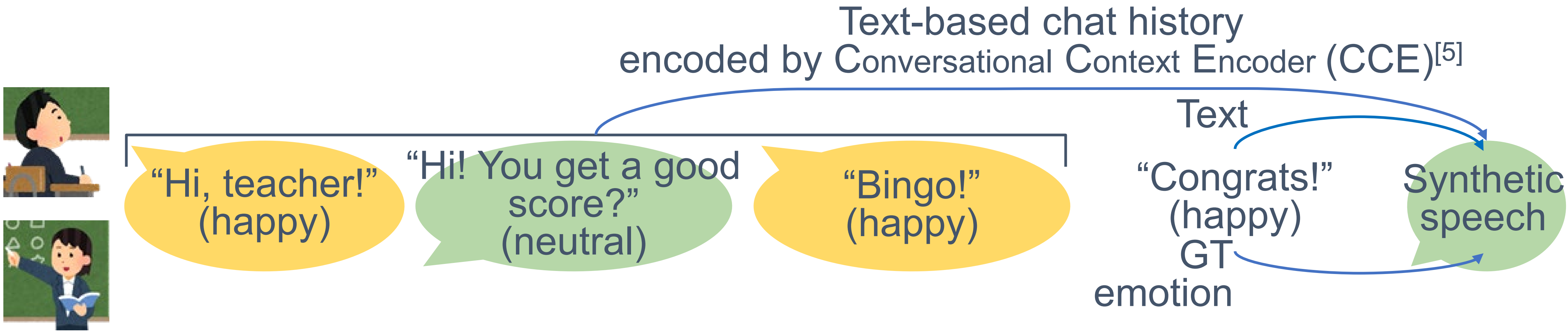
- **Dataset: STUDIES[3] teacher's 3,365 utterances (5 hours)**
  - GT emotion label: annotated by the corpus developers (i.e., **human** labelers)

		Intention		Emotion		Style	
GT emotion label	Avg. reliability score	Most frequent word	# of unique words	Most frequent word	# of unique words	Most frequent word	# of unique words
Neutral	3.95	Question	**206**	Anticipation	**130**	Quiet	**42**
Happy	4.04	Blessing	**76**	**Happiness**	**35**	Gentle	**19**
Angry	3.66	**Empathy**	**17**	Trust	**17**	Polite	**8**
Sad	4.03	**Empathy**	**49**	**Sadness**	**53**	Polite	**19**

ChatGPT's answers are too diverse, despite pre-defining the word candidates.

Experimental setup for EDSS evaluation

- **Baseline: Dialogue-history-aware EDSS from our previous work**[3]
  - Comparing the use of 1) GT emotion, 2) CCE-derived context embedding, & 3) ChatGPT-derived context embedding (**IES**) as a conditional feature

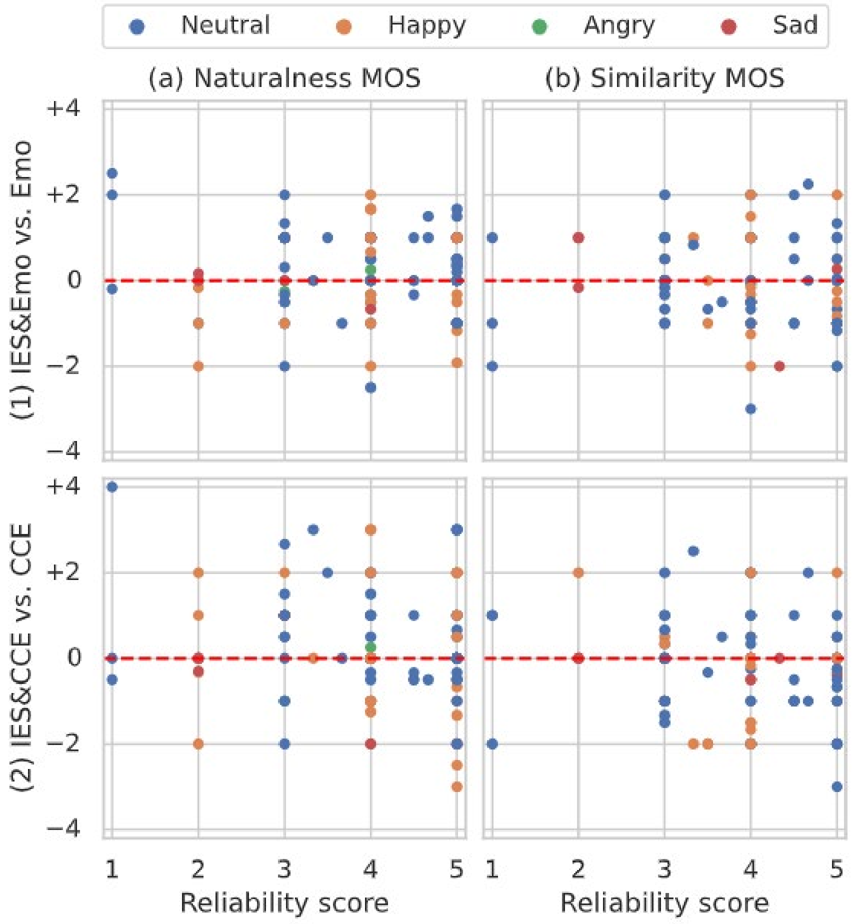


- **Other conditions (see our paper for details)**
  - EDSS model: FastSpeech 2[8] + HiFi-GAN vocoder[9]
  - # of training/validation/test data = 2,209/221/221 utterances

Subjective evaluations of synthetic speech

- **2 MOS tests involving 100 listeners**
  - Criteria: naturalness & speaking-style similarity

Conditional feature			MOS	
GT emo.	CCE	IES	Naturalness	Similarity
✓			3.43 ± 0.14	3.20 ± 0.15
	✓		3.54 ± 0.14	3.24 ± 0.14
		✓	**3.52 ± 0.14**	**3.19 ± 0.15**
✓		✓	**3.52 ± 0.14**	**3.21 ± 0.14**
✓	✓		3.43 ± 0.14	3.24 ± 0.14
	✓	✓	3.49 ± 0.14	3.20 ± 0.14

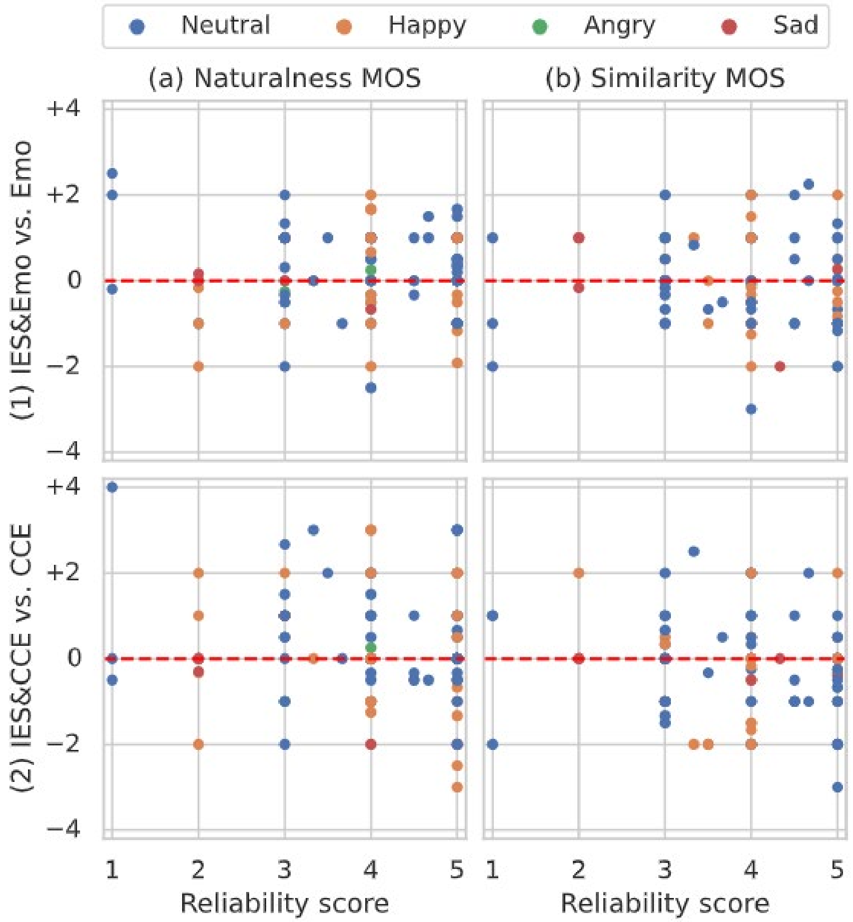


50 listeners for each test (30 answers per listener)

Subjective evaluations of synthetic speech

- **2 MOS tests involving 100 listeners**
  - Criteria: naturalness & speaking-style similarity

Conditional feature			MOS	
GT emo.	CCE	IES	Naturalness	Similarity
✓			3.43 ± 0.14	3.20 ± 0.15
	✓		3.54 ± 0.14	3.24 ± 0.14
		✓	**3.52 ± 0.14**	**3.19 ± 0.15**
✓		✓	3.52 ± 0.14	3.21 ± 0.14
✓	✓		3.43 ± 0.14	3.24 ± 0.14
	✓	✓	3.49 ± 0.14	3.20 ± 0.14



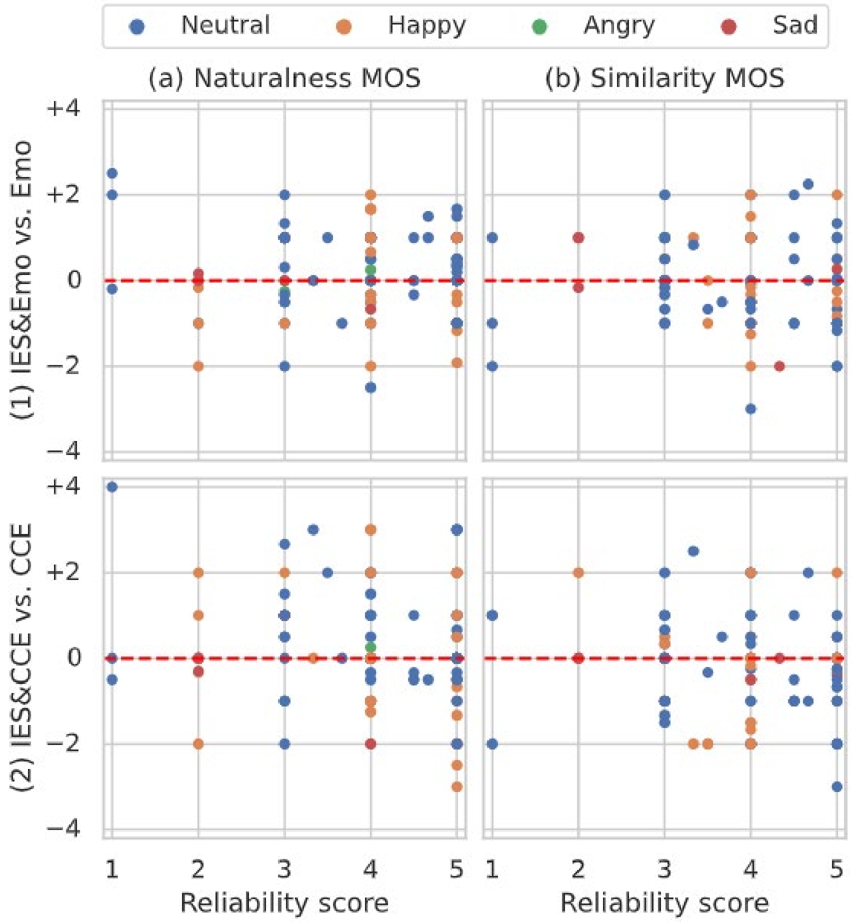
ChatGPT-derived context words achieve speech quality comparable to human-annotated emotion label & DNN-derived context embedding!

50 listeners for each test (30 answers per listener)

Subjective evaluations of synthetic speech

- **2 MOS tests involving 100 listeners**
  - Criteria: naturalness & speaking-style similarity

Conditional feature			MOS	
GT emo.	CCE	IES	Naturalness	Similarity
✓			3.43 ± 0.14	3.20 ± 0.15
	✓		3.54 ± 0.14	3.24 ± 0.14
		✓	3.52 ± 0.14	3.19 ± 0.15
✓		✓	**3.52 ± 0.14**	**3.21 ± 0.14**
✓	✓		3.43 ± 0.14	3.24 ± 0.14
	✓	✓	3.49 ± 0.14	3.20 ± 0.14



The reliability scores are not related to the speech quality improvement. (Perhaps main reason is the rich diversity of ChatGPT's answers?)

50 listeners for each test (30 answers per listener)

Speech samples

Conditional feature			MOS		🔊	🔊
GT emo.	CCE	IES	Naturalness	Similarity	Happy	Angry
✓			3.43 ± 0.14	3.20 ± 0.15	🔊	🔊
	✓		3.54 ± 0.14	3.24 ± 0.14	🔊	🔊
		✓	**3.52 ± 0.14**	**3.19 ± 0.15**	🔊	🔊
✓		✓	**3.52 ± 0.14**	**3.21 ± 0.14**	🔊	🔊
✓	✓		3.43 ± 0.14	3.24 ± 0.14	🔊	🔊
	✓	✓	3.49 ± 0.14	3.20 ± 0.14	🔊	🔊

← Natural speech

Other samples
&
collected
context words

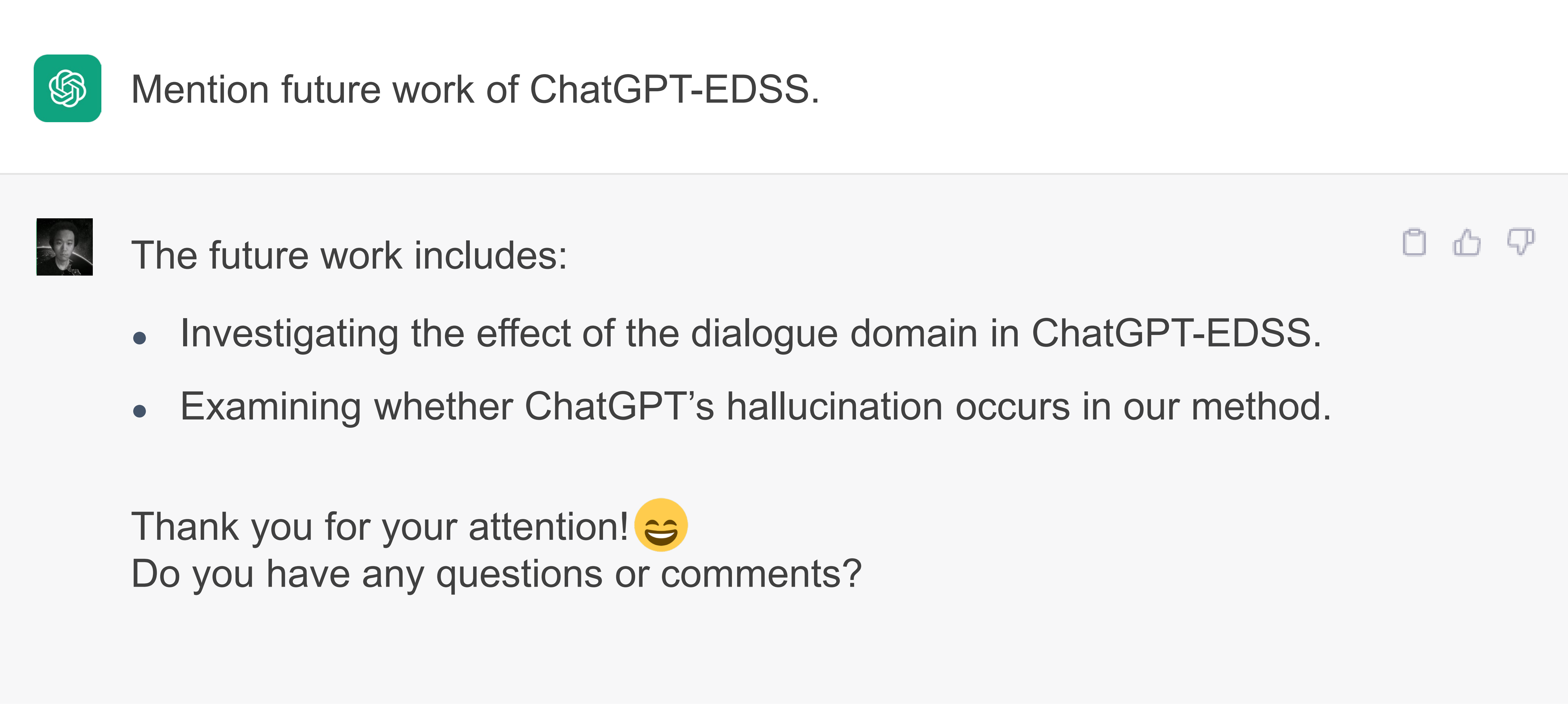
Summary of this talk

Summarize the contributions by ChatGPT-EDSS.

The research highlights of ChatGPT-EDSS are as follows:

- It investigated how to introduce ChatGPT into speech synthesis.
- It presented the prompt design to obtain useful contexts words and analyzed the obtained context words.
- From the results of EDSS experiments, it demonstrated that ChatGPT-EDSS achieved the comparable quality of synthetic speech to EDSS models conditioned on human-annotated emotion labels and deeply learned contextual embedding vectors.

Future work



Mention future work of ChatGPT-EDSS.

The future work includes:

- Investigating the effect of the dialogue domain in ChatGPT-EDSS.
- Examining whether ChatGPT’s hallucination occurs in our method.

Thank you for your attention! 😄
Do you have any questions or comments?

References

[1] H. Rao et al., “Can ChatGPT assess human personalities? a general evaluation framework,” arXiv, vol. arXiv:2303.01248, 2023.

[2] C. Qin et al., “Is ChatGPT a general-purpose natural language processing task solver?” arXiv, vol. arXiv:2302.06476, 2023.

[3] Y. Saito et al., “STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent,” in Proc. INTERSPEECH, Incheon, South Korea, Sep. 2022, pp. 5155–5159.

[4] Y. Saito et al., "CALLS: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center," Proc. INTERSPEECH, Dublin, Ireland, Aug. 2023, pp. 5561–5665.

[5] Z. Guo et al., “PromptTTS: Controllable text-to-speech with text descriptions,” arXiv, vol. abs/2211.12171, 2022.

[6] D. Stanton et al., “Predicting expressive speaking style from text in end-to-end speech synthesis,” in Proc. SLT, Athens, Greece, Dec. 2018, pp. 595–602.

[7] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” National Science Review, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[8] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in Proc. ICLR, Vienna, Austria, May 2021.

[9] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in Proc. NeurIPS, Vancouver, Canada, Dec. 2020.