Can LLMs Understand the Implication of Emphasized Sentences in Dialogue?

Guan-Ting Lin, Hung-yi Lee
Graduate Institute of Communication Engineering, National Taiwan University
Taiwan
{f10942104, hungyilee}@ntu.edu.tw

Abstract

Emphasis is a crucial component in human communication, which indicates the speaker's intention and implication beyond pure text in dialogue. While Large Language Models (LLMs) have revolutionized natural language processing, their ability to understand emphasis in dialogue remains unclear. This paper introduces **Emphasized-Talk**, a benchmark with emphasis-annotated dialogue samples capturing the implications of emphasis. We evaluate various LLMs, both open-source and commercial, to measure their performance in understanding emphasis. Additionally, we propose an automatic evaluation pipeline using GPT-4, which achieves a high correlation with human rating. Our findings reveal that although commercial LLMs generally perform better, there is still significant room for improvement in comprehending emphasized sentences[1].

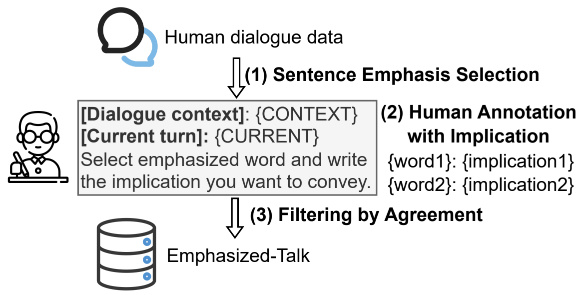

Figure 1: The illustration of Emphasized-Talk data collection pipeline.

1 Introduction

Emphasis plays a key role in communication by highlighting important parts of dialogue. Recognizing these emphatic cues is crucial for achieving natural interactions, as they help reinforce intentions, and express nuances essential for understanding conversations. For instance, altering the emphasis in the sentence *I never said he stole my bag* from *he* to *my* can drastically change its meaning. Emphasis can be represented in various ways, such as quotation marks, bold text, italics, capitalization, and underlining.

Large Language Models (LLMs) (Wei et al., 2022; Touvron et al., 2023a; OpenAI, 2023) have revolutionized natural language processing by leveraging vast amounts of unlabeled text data. They can generate and understand natural language accurately, capturing nuances and complex relationships. As a result, LLMs are widely used in applications such as dialogue systems and virtual assistants. Despite their capabilities, the ability of LLMs to grasp the *subtle nuances and complex meanings conveyed through* ***emphasis*** *in human dialogue remains unclear.*

To address this gap, we introduce a novel benchmark evaluation dataset, **Emphasized-Talk**, which features real dialogue samples with annotations capturing the implications of emphasis as interpreted by humans. Emphasized-Talk includes the same dialogue context and current sentence but with different words or phrases emphasized. This dataset is designed to test whether LLMs can accurately understand the meaning and intention behind emphasis in dialogue. Note that in this work, we use `quotation mark` ("") to highlight the emphasis in text sentences. Our study includes both open-sourced and commercial LLMs of varying sizes to provide a comprehensive analysis of their current capabilities and limitations.

Furthermore, we propose an automatic evaluation method that leverages the capabilities of GPT-4 (OpenAI, 2023) to assess the performance of these models. By comparing various automatic evaluation methods, we find that GPT-4's evaluations correlate well with human evaluation scores. Our findings indicate that open-source LLMs generally struggle to understand the implications of emphasized sentences. While commercial LLMs

[1] https://github.com/DanielLin94144/Emphasized-Talk

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13391–13401

perform better, there is still room for improvement in their comprehension of emphasized text.

Our contributions can be summarized as below:

1. We present a novel benchmark evaluation dataset, **Emphasized-Talk**, featuring real dialogues annotated with emphasis implications, providing a critical resource for testing LLMs' understanding of the emphasized text in conversations.

2. Our study includes a comparative analysis of various LLMs, assessing their ability to interpret and generate emphasis in dialogue. This highlights the performance differences across models of different scales.

3. We introduce an automated evaluation pipeline using GPT-4 with our benchmark, enabling efficient and consistent assessment of models, which significantly reduces the need for manual testing.

2 Related works

Human Communication Beyond Textual Information: In addition to text, the extra tags like emotion (Pohl et al., 2017; Xue et al., 2023), emoji (Demszky et al., 2020), sentiment (Lin et al., 2024b), prosody (Ward et al., 2012; Lin et al., 2024a), and non-lexical verbal sounds (Ward, 2006) are critical in understanding and generating human dialogue. These cues provide essential context and pragmatic use to the interaction, often conveying nuances that textual information alone cannot fully capture (Ward, 2004). For instance, emotions and sentiment can indicate the speaker's attitude and feelings, while speaking style and prosody can reveal additional layers of meaning, such as sarcasm, emphasis, or urgency. This work focuses on emphasis as the non-textual cue in human communication, aiming to understand the implication meaning of beyond text. The quotation mark ("") is used to simulate the expression of emphasis.

Emphasis in Dialogue: In human communication, whether spoken or written, emphasis is a crucial tool for enhancing meaning, indicating emotional states, or highlighting important dialogue segments. In dialogue systems, understanding the subtleties conveyed through emphatic cues is essential for achieving truly natural interactions (Pierrehumbert and Hirschberg, 1990; Wagner and Watson, 2010; Jackson, 2016). Emphasis aids in disambiguating syntactic structures, reinforcing intentions, and expressing nuances that are vital for comprehensive understanding in conversational contexts (Buchanan, 2013). Due to the importance of emphasis, studies on emphasis detection (Arons, 1994; Suni et al., 2017; Talman et al., 2019; Zhang et al., 2018; Vaidya et al., 2022; Morrison et al., 2024), emphatic Text-to-Speech synthesis (Fernandez and Ramabhadran, 2007; Seshadri et al., 2022; Stephenson et al., 2022), and Speech-to-Speech translation (Goldman et al., 2016; de Seyssel et al., 2023) are long-standing research topics to capture emphasis. This work focus on evaluating the LLMs' ability to understand the high-level meaning and intention of emphasis.

3 Dataset: Emphasized-Talk

The same context and current text with different emphasized words can lead to different implied meanings. Since no existing dataset contains varying emphasized words and meanings, we have created the first such dataset. Figure 1 illustrates the data collection. To build a real-world dialogue dataset with emphasized sentences, we initially used the DailyTalk (Lee et al., 2023) dataset, an open-domain multi-turn spoken dialogue dataset, as our content source. We then adopt the following strategies to create the Emphasized-Talk data:

3.1 Sentence Emphasis Selection

For each dialogue, we select the current sentence to be emphasized based on the following criteria: **(1) Sufficient dialogue context**: The current turn is selected only if there are more than two preceding dialogue turns, ensuring adequate contextual information. **(2) Availability of emphasized targets**: Since emphasizing function words like punctuation, articles, and proper nouns rarely affect sentence meaning, we select sentences containing more than four non-functional words and proper nouns for further consideration.

3.2 Human Annotation with Implication

Selecting *which words and phrases to emphasize* is non-trivial. In human communication, emphasis placement depends on the dialogue context and the information the speaker wishes to convey. In this work, ***human annotators choose where to place emphasis and document the implied meaning behind their choices***, making the data more pragmatic and reflective of real-world dialogue. On average, the emphasized fragments consist of

Model	MOS		BERT$_{f1}$		auto-gpt4-gt		auto-gpt4	
	score	rank	score	rank	score	rank	score	rank
ChatGPT	3.59	2	88.7	1	2.94	2	3.83	2
Claude 3 Sonnet	3.73	1	88.5	2	3.03	1	3.86	1
Llama 3-70B-chat	3.51	3	88.2	3	2.79	3	3.39	3
Llama 2-70B-chat	3.21	6	86.5	6	2.54	4	2.92	5
Llama 3-8B-instruct	3.41	4	87.7	4	2.36	6	3.06	4
Llama 2 7B-chat	2.61	7	86.3	7	1.71	7	1.85	7
Mistral 7B-instruct	3.29	5	87.4	5	2.47	5	2.82	6
Spearman's rank corr coef	-		0.964		0.857		0.964	
p-value	-		4.5×10^{-4}		1.4×10^{-2}		4.5×10^{-4}	

Table 1: The MOS, BERTscore, auto-gpt4-gt, and auto-gpt4 score of different LLMs.

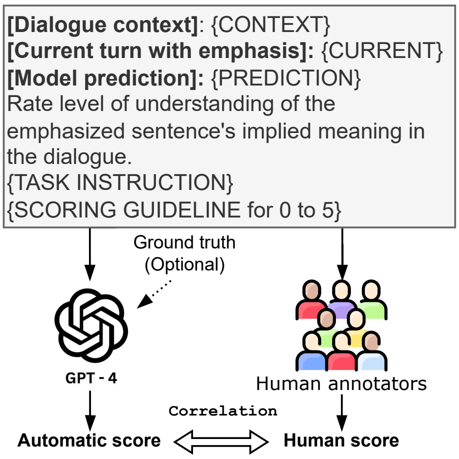

Figure 2: Illustration of automatic and human evaluation of the model's predicted implications.

1.15 words, indicating that annotators mostly emphasize single words, with occasional emphasis on phrases. Each annotator selects two different words or phrases to emphasize and note the implied meaning. The task template and instructions are shown in Appendix Figure 3.

3.3 Filtering by Agreement

When multiple annotators select the same word or phrase to emphasize, there are often several annotations of implied meanings. If the implied meanings are significantly different between annotators, indicating a lack of consensus or too many possible implied meanings based on annotators' backgrounds, we check the semantic similarity among the annotated meanings using GPT-4 (see Appendix B for details). Samples without agreement are filtered out. The percentage of samples that passed the filtering is 79.1%. Overall, we collect 984 dialogue sample pairs. We show some examples in Appendix Table 5.

4 Evaluation

4.1 Large Language Models to be Evaluated

In our evaluation framework, LLMs only interpret the emphasis as conveyed in the input, but do not detect the emphasis. We use both open-source and closed commercial LLMs for evaluation. Specifically, we evaluate different versions of Llama (Touvron et al., 2023a) (Llama 2 (Touvron et al., 2023b) and Llama 3 (AI@Meta, 2024)) and various model sizes (ranging from 7B to 70B parameters). In addition to Llama, we experiment with Mistral 7B (Jiang et al., 2023). All models used are the instruction-tuned chat versions, not the pre-trained models. For commercial LLMs, we evaluate Claude 3 Sonnet[2] and ChatGPT 3.5 (`gpt-3.5-turbo-0125`)[3]. For all the LLMs, we provide the following prompt to generate the implied meaning of the emphasized sentence, where `{CONTEXT}` and `{CURRENT}` are placeholders that vary for each sample.

```
[Dialogue context]: {CONTEXT}
[Current turn with emphasis]: {CURRENT}
The emphasized information is indicated by the quotation mark "". Implication meaning refers to the key intention the speaker wants to specifically highlight beyond the original text, which is not simply paraphrasing the original text. Use a simple and concise sentence to describe the specific highlighted information. Directly output the implication meaning of the current turn
```

[2]https://www.anthropic.com/news/claude-3-family
[3]https://openai.com/index/chatgpt/

Context	0: Umm excuse me, are there any more shopping carts? 1: Yes, you can find it at the entrance.
Current turn	0: but there isn't a single one right now.
Predicted implication	"single": The customer needs an available shopping cart for individual use. "right now": There are no available shopping carts at the moment.
Ground truth implication	"single": Not even one cart is available. "right now": The lack of carts is an immediate issue.

Table 2: Qualitative example with Emphasizing on different targets in the current turns. 0 and 1 denote the speaker's identity. For implication rows, the word or phrase within the quotation mark refers to the emphasized target, and the sentence after the colon is the implication. The **Predicted implication** here is from the Claude 3 Sonnet model.

`sentence.`

4.2 Evaluation Metrics

For human evaluation, we recruit human annotators to rate the scoring from 0 to 5, given the rating guideline (the detailed instruction is shown in Appendix Figure 4). Each sample is rated by three annotators, the average score is the Mean Opinion Score (MOS).

However, human evaluation incurs high costs and significant time requirements, requiring skilled evaluators who can understand and analyze the subtle nuances of language and context within the dialogues. Therefore, we propose using the state-of-the-art LLM, GPT-4 (OpenAI, 2023), for automatic evaluation. The automatic evaluation methods are illustrated in Figure 2. We adopt a similar LLM automatic evaluation method as in previous works (Chiang and Lee, 2023a,b), instructing the LLM to predict ratings based on guidelines. This is denoted as **auto-gpt4** score. Given potential uncertainties in the LLM's comprehension of emphasized sentences, we also use a similarity score compared to *ground truth* annotations, employing both the BERTScore (**BERT**$_{f1}$) and GPT-4 (**auto-gpt4-gt**). BERTScore (Zhang et al., 2020) is a metric for evaluating text generation quality by comparing generated texts to reference texts using BERT embeddings. Unlike traditional metrics like BLEU, which rely on exact matches, BERTScore uses contextual embeddings to capture semantic similarities. As for **auto-gpt4-gt**, Appendix D describes more details on the prompt template for GPT-4 evaluation.

5 Results

5.1 LLMs' Performance on Implication Modeling

Table reports show the scores of different LLMs. Commercial LLMs generally outperform open-source ones in MOS. The top model is Claude 3 Sonnet with a MOS of 3.73, followed by ChatGPT at 3.59. Among open-source LLMs, Llama 3-70B scores 3.51, close to ChatGPT. The Llama 3 series consistently outperforms the Llama 2 series of similar sizes. Mistral 7B ranks between Llama 2-7B and Llama 3-8B, even surpassing Llama 2-70B. None of the LLMs reach the optimal score of 5, with the best at 3.73, indicating room for improvement in understanding emphasized sentences. More error analyses are shown in Appendix A.

Auto eval	Pearson's	Kendall's
BERT$_{f1}$	0.313	0.203
auto-gpt4-gt	0.568	0.268
auto-gpt4	**0.643**	**0.327**

Table 3: The Pearson's and Kendall's correlation coefficient between human MOS and of three automatic evaluation methods.

5.2 Correlation Between Automatic Evaluation Score and Human Score

We follow Chiang and Lee (2023b) to calculate the dataset-level Pearson's correlation and document-level Kendall's correlation. Pearson's measures overall alignment between automatic scores and human ratings, while Kendall's assesses quality differentiation between models for the same input. Table 3 shows the correlations of various evaluation methods. Auto-gpt4 scores correlate well with human ratings, with a Pearson's coefficient of 0.643 and a Kendall's coefficient of 0.327. Both auto-gpt4-gt and BERT$_{f1}$ also show positive correlations, but their coefficients are much lower. This suggests that GPT-4's direct analysis is more effective than using ground truths as references, which can be influenced by human annotators' writing styles and biases. Therefore, the auto-gpt4 score is more reliable for a higher correlation with human ratings.

5.3 Relative Ranking for Automatic Evaluation Methods

The ranking order of LLMs by automatic evaluation scores aligns closely with human scores. Spearman's rank correlation between human MOS and three automatic methods shows high coefficients: 0.964 for auto-gpt4 and BERT_{f1}, and 0.857 for auto-gpt4-gt. Auto-gpt4 and BERT_{f1} had only two incorrect model rankings. Despite low Pearson's and Kendall's correlations, BERT_{f1}'s rankings are similar to human MOS. Overall, auto-gpt4 achieves both the highest score and rank correlation.

5.4 Qualitative Example

Table 2 shows examples of the same dialogue context with different emphasized words or phrases. For the emphasized word *single*, the Claude 3 Sonnet model incorrectly interprets it as referring to a cart for individual use, whereas it actually refers to the number of carts. For the emphasized phrase *right now*, the model correctly understands it implies immediate use. This demonstrates that while commercial LLMs can infer different meanings based on emphasis, their accuracy depends on the complexity and depth of the context.

6 Conclusion

This work explores the ability of LLMs to interpret emphasized text in dialogues. We introduced the novel Emphasized-Talk dataset with the same dialogue context and current turn sentence with different emphasized words or phrases. Furthermore, we study the automatic evaluation methods using GPT-4, showing a high correlation between human rating and auto-gpt4. Our analysis shows that while commercial LLMs outperform open-sourced models, there is still room for improvement. We encourage future LLM research to evaluate the model on the proposed benchmark and automatic evaluation method to enhance the LLM's dialogue ability.

Limitations

The primary limitations of this paper are as follows:

1. This study exclusively utilizes GPT-4 for automatic evaluation, relying on previous works (Chiang and Lee, 2023b; Liu et al., 2023) that have demonstrated its effectiveness for scoring. However, the performance of other LLMs for scoring is not explored in this paper.

2. While speech emphasis is a common feature of human communication, this work simplifies emphasis by using additional quotation marks in text. Future research should investigate the use of direct speech input for spoken dialogue to better capture natural communication.

3. The study employs quotation marks to indicate emphasis within text sentences and explicitly instructs that the emphasized information is enclosed by these marks. However, other methods of emphasis, such as bold text or capitalization, are not explored in this paper.

Acknowledgement

The authors want to thank Cheng-Han Chiang for his insightful discussion. Guan-Ting Lin is supported by the NTU GICE and the NTU Ph.D. scholarship.

References

AI@Meta. 2024. Llama 3 model card.

Barry Arons. 1994. Pitch-based emphasis detection for segmenting speech recordings. In *ICSLP*, pages 1931–1934. Citeseer.

Ray Buchanan. 2013. Conversational implicature, communicative intentions, and content. *Canadian Journal of Philosophy*, 43(5-6):720–740.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2023. Emphassess : a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *Preprint*, arXiv:2312.14069.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Raul Fernandez and Bhuvana Ramabhadran. 2007. Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In *6th ISCA Workshop on SpeecSpeh Synthesis*, volume 90.

Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, and Junichi Yamagishi. 2016. The SIWIS Database: A Multilingual Speech Database with Acted Emphasis. In *Proc. Interspeech 2016*, pages 1532–1535.

Rebecca Claire Jackson. 2016. *The pragmatics of repetition, emphasis and intensification*. University of Salford (United Kingdom).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. 2024a. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. *arXiv preprint arXiv:2402.12786*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-Yi Lee, and Ivan Bulyko. 2024b. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10316–10320.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Max Morrison, Pranav Pawar, Nathan Pruyne, Jennifer Cole, and Bryan Pardo. 2024. Crowdsourced and automatic speech prominence estimation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12281–12285. IEEE.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse.

Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond just text: semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):1–42.

Shreyas Seshadri, Tuomo Raitio, Dan Castellani, and Jiangchuan Li. 2022. Emphasis Control for Parallel Neural TTS. In *Proc. Interspeech 2022*, pages 3378–3382.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model. In *Proc. Interspeech 2022*, pages 3383–3387.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pre-trained contextualized word representations. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 281–290, Turku, Finland. Linköping University Electronic Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mithilesh Vaidya, Kamini Sabu, and Preeti Rao. 2022. Deep learning for prominence detection in children's read speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8157–8161. IEEE.

Michael Wagner and Duane G Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody 2004, International Conference*.

Nigel Ward. 2006. Non-lexical conversational sounds in american english. *Pragmatics & Cognition*, 14(1):129–182.

Nigel G Ward, Alejandro Vega, and Timo Baumann. 2012. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54(2):161–174.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Qian Chen, and Lei Xie. 2023. E-chat: Emotion-sensitive spoken dialogue system with large language models. *arXiv preprint arXiv:2401.00475*.

Long Zhang, Jia Jia, Fanbo Meng, Suping Zhou, Wei Chen, Cunjun Zhang, and Runnan Li. 2018. Emphasis detection for voice dialogue applications using multi-channel convolutional bidirectional long short-term memory network. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 210–214. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

A Analysis on Unsatisfied Model Prediction

We analyze the cases resulting in unsatisfactory ratings (average MOS score below 2). The examples are shown in Table 4. We observe that Claude 3 Sonnet overinterprets the emphasized information in an extremely contrasting way, but this is not relevant to the dialogue context. In contrast, the Llama 2-7B and Mistral-7B models fail to focus on the current turn and the emphasized word, merely summarizing the entire dialogue. These results indicate that small open-sourced LLMs struggle to understand emphasized text and its implications, and sometimes even fail to follow instructions. On the other hand, commercial LLMs can follow instructions but may overinterpret the emphasized text. The ideal interpretation should be reasonable and relevant to the dialogue context.

B Prompt for Checking Agreement

The prompt for checking agreement is: `For each sample, given a set of text sentences, check whether all the following sentences are semantically close to each other or not. You should also consider differences in subtle and nuanced meanings. You should provide an explanation, and then output yes if all sentences are semantically close to each other; otherwise, output no.`

C Inter-annotator Agreement for Human Evaluation Score

To measure the inter-annotator agreement, we use Krippendorff's alpha score[4]. From the rating of all the models, Krippendorff's alpha score is 0.255, which means moderate agreement among annotators. The result is reasonable since there is certain freedom for the implication meaning, depending on the background and personality of each annotator.

D Details of Automatic Evaluation

Following Chiang and Lee (2023b), we first ask the GPT-4 to analyze input samples and then predict scores. We request GPT-4 to output 3 outputs and average the score to reduce the score variation due to randomness. For the auto-gpt4 score, we use a similar prompt similar as the instruction in Figure 4. For the auto-gpt4-gt score, GPT-4 rates the similarity between ground truths and model predictions, using task instructions, grading definitions, dialogue context, and the emphasized current turn sentence. The prompt is as follows:

```
The task is modeling the implication meaning of the emphasized sentence. We are checking how semantically close and subtle meaning differences between the ground truth sentences and model prediction. For the subtle and nuanced meaning, focusing on the intention and highlighted information of the speaker. You must follow the following steps to provide the score: First, analyze and explain the sentences with the above definition. Second, output just the number of scores from the range of integers from 0 to 5:
0: No semantic similarity; the model prediction completely diverges from the ground truth in meaning and nuance.
```

[4]https://github.com/pln-fing-udelar/fast-krippendorff

Failure cases
0: can anyone tell me where I can buy the ingredients to make Chinese food? 1: There is a supermarket in Kensington High Street. I'll come and give you a hand. 0: That is "good" of you. [`current turn`] **Claude 3 Sonnet's Prediction:** The speaker is expressing sarcasm or lack of sincerity in response to the offer of help. **Llama 2-7B's Prediction:** The speaker is offering to help the listener find the supermarket and purchase the ingredients needed for Chinese cooking. **Mistral-7B's Prediction:** The speaker is offering to help the person find ingredients for Chinese food.

Table 4: Failure cases in which the model prediction receives a low rating from the human.

```
1: Very low semantic similarity; only a few elements match, with significant differences in meaning and nuance.
2: Low semantic similarity; some parts match, but there are notable differences in meaning and nuance.
3: Moderate semantic similarity; many parts match, but some differences in meaning and nuance are present.
4: High semantic similarity; most parts match, with minor differences in meaning and nuance.
5: Perfect or near-perfect semantic similarity; the model prediction closely mirrors the ground truth in both meaning and nuance.
The response must be in valid JSON format as below, which can be correctly parsed by json.loads() in python: "analysis": explanation, "score": number
```

For the BERT score, we average the scores if multiple ground truths are available[5].

E Dataset License

We release the Emphasized-Talk dataset under the MIT license.

F Details of Human Annotation Process

We assign three annotators for each assignment. All are based in the United States with HIT approval rates higher than 98%, given that the corpus is in American English. Each test contains 20 samples for evaluation. We pay the annotators 2.5 USD for each test. On average, based on the time for annotating and reading the content, it takes 10 minutes on one test, so the hourly wage is around 15 USD.

[5] https://huggingface.co/spaces/evaluate-metric/bertscore

Annotation for Implied Semantic Meaning of Emphasized Sentence in Dialogue

[Background]

Given the same sentence, the emphasis or stress placed on different words in a sentence can significantly alter its semantic meaning. In human dialogue, people can intentionally choose the words to emphasize for conveying meaning beyond pure text content.

[Instructions and Guidelines for the task]

In this task, there are two speakers (speaker 0 and 1). Given a dialogue context and the current sentence, please do the following actions:

1. Read the context and current sentence: Imagine you are the speaker of current sentence.

2. Choose “two different” word to place emphasis:
(a) Note that you should **"choose the words that emphasizes on it will cause obviously different meaning than pure text content".**
(b) Emphasizing on the part of speech like preposition, conjunction, and article can’t make too much difference compared to the original sentence meaning.
(c) **Directly copy and paste the word or phrase in the current sentence to align the format.**

3. Write down the corresponding implied meanings (that is, the reason you want to emphasize on that word) for each two emphasis:
(a) Write the implied meaning in text. Please use a **"ONE concise and grammatically correct sentence without clause" to describe the implied meaning.**
(b) The implied meaning **don't need to include all the information in original sentence, but the most important meaning** you want to convey.
(c) You can add adverb to describe more details.

Context:
0: I've got a complaint about the noise next door.
1: Yes, it's most irritating.
0: Some people aren't very considerate.
Current sentence:
1: We're going to do something about it.

Chosen emphasis word 1

Don't type! Directly copy and paste the selected emphasized word

Implied meaning 1

Write the implied meaning here...

Chosen emphasis word 2

Don't type! Directly copy and paste the selected emphasized word

Implied meaning 2

Write the implied meaning here...

Figure 3: The template for selecting emphasized words and documenting their implied meanings.

Rate level of understanding of the emphasized sentence's implied meaning in the dialogue

[Background]

Given the same sentence, the emphasis or stress placed on different words in a sentence can significantly alter its semantic meaning. In human dialogue, people can intentionally choose the words to emphasize for conveying meaning beyond pure text content. Emphasizing different words in a sentence changes its deeper meaning. This task is about understanding the deeper implied meaning.

[NOTE]: the ideal implication: Not just paraphrasing, focusing on intention and highlighted information, not too lengthy containing too much information.

[Instructions and Guidelines for the task]

(1) Read the dialogue context and the current turn. There are two speakers (0 and 1).

(2) The emphasized word is marked by the quotation mark "" in the current turn sentence.

(3) Imagine you are the current turn's speaker. You role is to rate how well the model prediction correctly capture the implication meaning of the emphasized sentence. **You must follow the below rating criterion to give the score.**

- 0: No Understanding - Implication is irrelevant or nonsensical.
- 1: Minimal Understanding - Barely relates to the emphasis, mostly misses the implied meaning.
- 2: Partial Understanding - Captures some implied meaning but is unclear or missing key points.
- 3: Moderate Understanding - Understands the main implied idea but misses some nuances or containing some irrelvant information.
- 4: Good Understanding - Mostly correct, reflects the implied meaning with minor irrelvant information.
- 5: Excellent Understanding - Correctly captures the implied meaning and intention behind the emphasis without irrelvant information.

Dialogue context:
0: what can i do for you, madam?
Current sentence with emphasized word in "":
1: i'd like to buy a toy car for my "son".

Model prediction of implied meaning:
The speaker is looking for a toy car for their child.

How well the model prediction correctly capture the implication meaning of the emphasized sentence.
( ) **0: No Understanding** - Implication is irrelevant or nonsensical.
( ) **1: Minimal Understanding** - Barely relates to the emphasis, mostly misses the implied meaning.
( ) **2: Partial Understanding** - Captures some implied meaning but is unclear or missing key points.
( ) **3: Moderate Understanding** - Understands the main implied idea but misses some nuances or containing some irrelvant information
( ) **4: Good Understanding** - Mostly correct, reflects the implied meaning with minor irrelvant information.
( ) **5: Excellent Understanding** - Correctly captures the implied meaning and intention behind the emphasis without irrelvant information.

Figure 4: The template for human evaluation, including instructions and grading policy.

Example 1	
Input	1. The speaker is expressing genuine happiness and appreciation for the compliment about their country. 2. I am genuinely happy to hear that you love japan. 3. I express my happiness upon hearing your admiration for japan.
Output	Output: Yes Explanation: All three sentences convey the speaker's genuine happiness in response to a compliment about Japan. They all express positive feelings related to the admiration of Japan.
Example 2	
Input	1. You should have expressed your deep feelings for her to me. 2. The relationship is serious and significant. 3. You should have informed me about your romantic feelings for her.
Output	Output: No Explanation: The second sentence discusses the seriousness of the relationship in general, which differs from the specific recommendation of expressing feelings in the first and third sentences.

Table 5: The examples of filtering by agreement process.

Example of auto-gpt4-gt score	
Input	**Dialogue context**: What sort of music do you like? **Current turn**: I love just about "anything" by prince. **Model prediction**: The speaker has a strong preference for and enjoyment of Prince's music. **Ground truth**: The speaker's affection for prince's music encompasses a wide range, indicating a broad and inclusive taste.
Output	**Analysis**: The ground truth emphasizes the breadth of the speaker's affection towards Prince's music, highlighting an inclusive and wide-ranging taste. The model prediction focuses more narrowly, noting the speaker's strong preference and enjoyment, but doesn't capture the aspect of 'broad and inclusive taste'. Therefore, while both sentences refer to the speaker's positive feelings towards Prince's music, the model prediction lacks the subtlety about the speaker's taste being wide and inclusive. **Score**: 3

Table 6: Example input and output for the `auto-gpt4-gt` score.