Sentiment Reasoning for Healthcare

Khai-Nguyen Nguyen[*1], **Khai Le-Duc**[*2,3],
Bach Phan Tat[4], **Duy Le**[5], **Long Vo-Dang**[6], **Truong-Son Hy**[7]
[1]College of William and Mary, USA [2]University of Toronto, Canada
[3]University Health Network, Canada [4]KU Leuven, Belgium
[5]Bucknell University, USA [6]University of Cincinnati, USA
[7]University of Alabama at Birmingham, USA
knguyen07@wm.edu duckhai.le@mail.utoronto.ca
https://github.com/leduckhai/Sentiment-Reasoning

Abstract

Transparency in AI healthcare decision-making is crucial. By incorporating rationales to explain reason for each predicted label, users could understand Large Language Models (LLMs)'s reasoning to make better decision. In this work, we introduce a new task - Sentiment Reasoning - for both speech and text modalities, and our proposed multimodal multitask framework and **the world's largest multimodal sentiment analysis dataset**. Sentiment Reasoning is an auxiliary task in sentiment analysis where the model predicts both the sentiment label and generates the rationale behind it based on the input transcript. Our study conducted on both human transcripts and Automatic Speech Recognition (ASR) transcripts shows that Sentiment Reasoning helps improve model transparency by providing rationale for model prediction with quality semantically comparable to humans while also improving model's classification performance (**+2% increase in both accuracy and macro-F1**) via rationale-augmented fine-tuning. Also, no significant difference in the semantic quality of generated rationales between human and ASR transcripts. All code, data (five languages - Vietnamese, English, Chinese, German, and French) and models are published online.

1 Introduction

Sentiment analysis plays a pivotal role within the healthcare domain. In healthcare customer service, it facilitates real-time evaluation of customer satisfaction, enhancing empathetic and responsive interactions (Xia et al., 2009; Na et al., 2012). Moreover, sentiment analysis aids in monitoring the emotional well-being of patients (Cambria et al., 2012a), including those with mental health issues such as suicide (Pestian et al., 2012). However, these studies only work on text-only sentiment analysis instead of speech-based sentiment analysis.

Despite its potential, speech sentiment analysis presents several technical challenges. First, emotions conveyed through speech are subjective (Wearne et al., 2019), complex (Golan et al., 2006), and dependent on speaking styles (Shafran and Rose, 2003), making accurate sentiment classification difficult even for humans (Kuusikko et al., 2009), thereby necessitating the role of explainable artificial intelligence (AI). Second, given the critical nature of healthcare decisions, where errors can have severe consequences, transparency in AI decision-making is essential to build trust among machines, healthcare professionals, and patients (Antoniadi et al., 2021).

To tackle challenges above, reasoning in AI is crucial for sentiment analysis because it enables deeper understanding beyond surface-level sentiment polarity via the textual explanations. Recent works on Chain-of-Thought (CoT) distillation (Wadhwa et al., 2024; Chen et al., 2024; Hsieh et al., 2023; Ho et al., 2022) have revealed that training generative small language models (SLMs) on rationale-augmented targets (the CoT from larger models is provided along side with the target label) can help the SLM (1) perform better and (2) acquire the ability to generate rationale. Our work leverage these findings and prepare a set of human-labeled rationale to train our sentiment analysis models to do Rationale Generation and enhance their performance (Section 4.4 and 4.5). By incorporating rationales to explain reason for each predicted sentiment label, users could understand the model's reasoning, facilitating better decision-making based on the classification results. Therefore, we introduce a novel multimodal framework for a novel task: Sentiment Reasoning, which comprises of two tasks: (i) Sentiment Classification, in which the model learns to output the **sentiment label** (POSITIVE, NEUTRAL,

[*]Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1175–1202
July 28-30, 2025

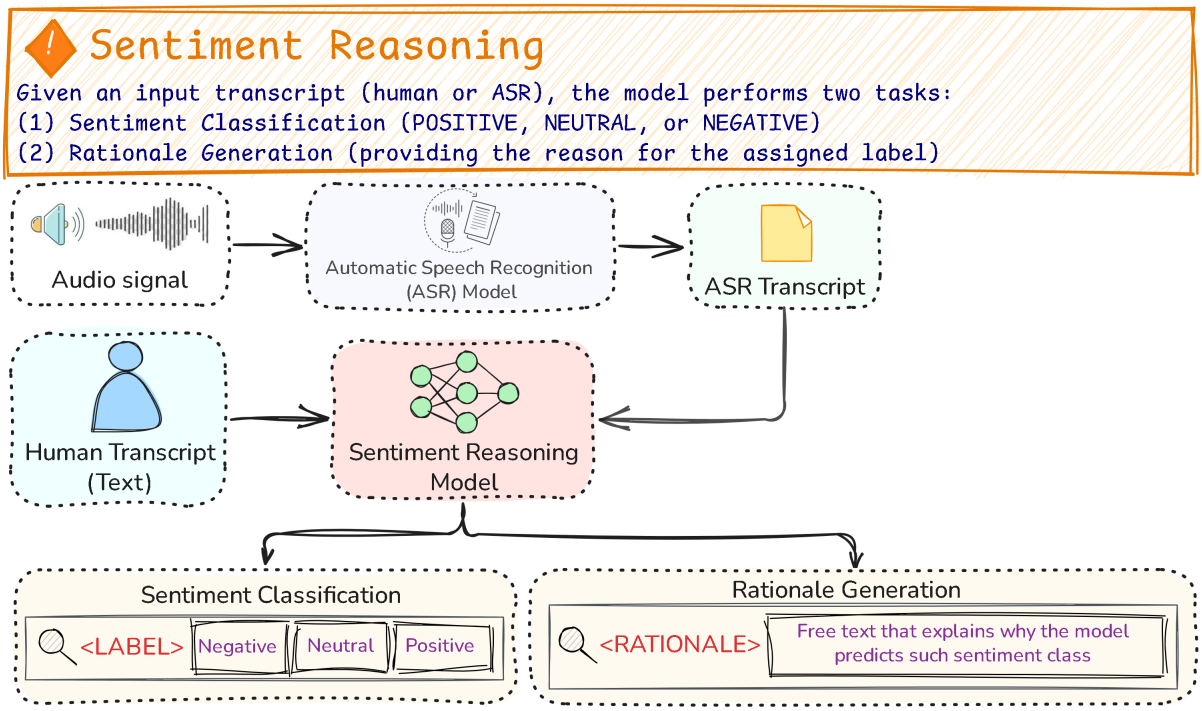

Figure 1: Visualized pipeline for Sentiment Reasoning . Given an input transcript (either human transcript or ASR transcript), the model learns to output the **sentiment label** (POSITIVE, NEUTRAL, or NEGATIVE) and its **rationale** (the reason for this label). It comprises of two tasks: (1) Sentiment Classification and (2) Rationale Generation . Traditional sentiment analysis only includes Sentiment Classification task, while our framework generates corresponding rationale to explain the reason behind each predicted sentiment label. 9 examples with sentiment labels and their corresponding rationales in our dataset are shown in Table 7 in the Appendix.

or NEGATIVE), and (ii) Rationale Generation, in which the model generates rationale (the free-form text that explains reason for this label). Our contributions are as follows:

1. We introduce a new task: Sentiment Reasoning for both speech and text modalities, along with the world's largest multimodal sentiment analysis dataset, supporting five languages (Vietnamese, English, Chinese, German, and French)
2. We propose our novel multimodal speech-text Sentiment Reasoning framework
3. We empirically evaluate the baselines on our dataset using state-of-the-art backbone models
4. We provide in-depth analysis of rationale / Chain-of-Thought (CoT)-augmented training

All code, data and models are published online.

2 Data

2.1 Data Collection

The dataset employed for constructing the Sentiment Reasoning dataset was *VietMed* (Le-Duc, 2024), a large and publicly accessible medical ASR dataset. The dataset comprises real-world doctor-patient conversations. We then annotated *sentiment labels* (POSITIVE, NEUTRAL, or NEGATIVE) and their corresponding *rationales* (the reason for this label). We then manually translate the transcripts from Vietnamese into other four languages: English, Chinese (Simplified and Traditional), German, and French, making the dataset six times larger. The full dataset (with 5 languages) includes 30000 samples, making it **the largest multimodal sentiment analysis dataset**, to the best of our knowledge (see Table 2). Our paper focuses mainly on the **Vietnamese subset** (Section 5) and the **English subset** (Appendix D).

2.2 Data Annotation

The annotation task consists of two primary steps. First, annotators are required to perform Sentiment Classification. Second, annotators are instructed to provide a rationale behind each class (Rationale Generation). To ensure consistency, our TESOL-certificated professional linguist has developed an initial guideline inspired by (Chen et al., 2020), which was also adopted by various well-known works (Shon et al., 2022, 2023), and revised it fre-

quently if necessary. Details of data annotation pipeline, annotation guidelines, data imbalance, translation annotation, and translaton quality control are shown in Appendix Section B.

2.3 Data Quality Control

During the independent annotation process conducted by three annotators, we observed a low inter-annotator agreement (Cohen's kappa coefficient below 0.5 for the inter-annotator agreement between the two annotators), a common occurrence in real-world datasets as noted by Chen et al. (2020). To address this issue, we implemented an alternative label merging approach. We convened a discussion meeting involving the three annotators and two reviewers (one professional linguist and one with a biomedical background). Each annotator was required to justify their chosen sentiment label and its corresponding rationale. A label and its rationale were selected based on the consensus of all three annotators and two reviewers, rather than a majority vote, as employed in other studies (Aziz and Dimililer, 2020; Saleena et al., 2018).

2.4 Data Statistics

Split	Label	Count	Percentage
Train	Neutral	2844	49.94%
	Negative	1694	29.74%
	Positive	1157	20.32%
Test	Neutral	958	43.88%
	Negative	701	32.11%
	Positive	524	20.01%

Table 1: Distribution of sentiment labels in the dataset for a single language. The real size of the dataset is 6 times larger when accounting all 5 languages - English, Chinese (Simplified and Traditional), German, and French.

Table 1 shows the distribution of sentiment labels in the dataset. This reflects the dataset's slight emphasis on neutral content, typical in medical conversations involving explanations and advice.

It should be noted that the statistics are reported **for a single language**, meaning that the real size of the dataset is 6 times larger when accounting all 5 languages.

3 Sentiment Reasoning Framework

3.1 Informal Definition

As shown in Figure 1, in Sentiment Reasoning , given an input transcript (either human transcript or ASR transcript), the model learns to output the **sentiment label** (POSITIVE, NEUTRAL, or NEGATIVE) and its **rationale** (the reason for this label). It comprises of two tasks: Sentiment Classification and Rationale Generation .

3.2 Formal Definition

Let $x_1^T := x_1, x_2, ..., x_T$ be an audio signal of length T. Let C be the set of all possible sentiment classes, we should build a speech-based Sentiment Reasoning model f that both estimates the probability $p(c|x_1^T)$ for each $c \in C$ and generates its rationale sequence r_1^M of M length.

The decision rule to predict a sentiment class is:

$$x_1^T \rightarrow \hat{c} = \arg\max_{c \in C} f(c|x_1^T) \quad (1)$$

The decision rule to generates the corresponding rationale sequence is:

$$x_1^T \rightarrow r_1^M = \arg\max_{r^*} h(r^*|x_1^T) \quad (2)$$

For text-based Sentiment Reasoning, the input audio signal x_1^T could be replaced with a word sequence (human transcript) w_1^N of length N, thus ASR model is not needed.

3.3 ASR Model

An ASR model aims to convert audio signal into text by mapping an audio signal x_1^T to the most likely word sequence w_1^N. The relation w^* between the acoustic and word sequence is:

$$w^* = \arg\max_{w_1^N} p(w_1^N|x_1^T) \quad (3)$$

3.4 Language Model for Sentiment Reasoning

3.4.1 Sentiment Classification

Let the transcribed audio signal (ASR transcript) w_1^N serve as the input for the Sentiment Classification model g, which maps w_1^N to a class label $\hat{c}$:

$$w_1^N \rightarrow \hat{c} = \arg\max_{c \in C} g(c|w_1^N) \quad (4)$$

g is trained to minimize a loss function $\mathscr{L}(g(w_1^N), \hat{c})$. The optimal parameters θ of the model are found by solving the optimization problem $\min_\theta \mathscr{L}(g(w_1^N; \theta), \hat{c})$. Once trained, the model can predict the class of the transcribed audio signal by evaluating $\hat{c} = g(w_1^N)$.

Dataset	Venue	#Samp.	#Lang.	Domain
Mosi (Zadeh et al., 2016)	IEEE	3k	1	Vlog
CMU-MOSEI (Bagher Zadeh et al., 2018)	ACL	23k	1	Various
MELD (Poria et al., 2019)	ACL	13k	1	TV Series
IEMOCAP (Busso et al., 2008)	Springer	12k	1	General
SEMAINE (McKeown et al., 2012)	IEEE	1k	1	Simulation
Sentiment Reasoning (ours)	-	**30k**	**5**	Medical

Table 2: Data statistics comparison based on the number of samples and languages. Our dataset with 5 languages (Vietnamese, English, Chinese, German and French) includes 30000 samples, making it **the largest multimodal sentiment analysis dataset**.

3.4.2 Rationale Generation

Let the transcribed audio signal (ASR transcript) w_1^N serve as the input for the Rationale Generation model h, which maps w_1^N to a rationale sequence r_1^M of M length:

$$w_1^N \rightarrow r_1^M = \arg\max_{r^*} h(r^*|w_1^N) \quad (5)$$

h is trained to minimize a loss function $\mathscr{L}(h(w_1^N), r_1^M)$. The optimal parameters θ of the model are found by solving the optimization problem $\min_\theta \mathscr{L}(g(w_1^N;\theta), r_1^M)$. Once trained, the model can generate rationale of the transcribed audio signal by evaluating $r_1^M = h(w_1^N)$.

4 Experimental Setups

4.1 ASR Model

We employed hybrid ASR setup using wav2vec 2.0 encoder (Le-Duc, 2024) to transcribe speech to text. The final ASR model has 118M trainable parameters and Word-Error-Rate (WER) of 29.6% on the test set. Details of ASR experiments are shown in Appendix C.1.

4.2 End-to-end Sentiment Classification

We fine-tuned two well-known models, PhoWhisper (Le et al., 2024) and Qwen2-Audio (Chu et al., 2024), for the end-to-end spoken sentiment analysis task. PhoWhisper is trained large-scale ASR training set consisting of 844 hours of Vietnamese audio, while Qwen2-Audio is trained on more than 500 hours of audio. We use the base version of PhoWhisper with 74M parameters, while Qwen2-Audio has 8.2B parameters.

4.3 Language Model for Sentiment Reasoning

4.3.1 Encoder

The encoder architecture is naturally well-suited for Sentiment Classification, which can be reformulated into the classical classification task. To this end, we directly apply a linear classifier to the output of the encoders. However, encoders can not generate rationales. As such, **they serve as baselines in our experiments**.

We use **phoBERT** (110M params) (Nguyen and Nguyen, 2020), RoBERTa (Liu et al., 2019) pre-trained on 20GB Vietnamese text, and **ViHealthBERT** (110M params) (Minh et al., 2022), phoBERT trained on 32GB of Vietnamese text in the healthcare domain. For ViHealthBERT, we report the syllable version which achieved better performance than the word version.

4.3.2 Generative Models

We reformulated Sentiment Classification into a text-to-text problem, where given the input transcript w_1^N, the generative model g and the predicted sentiment class c, we have $g(w_1^N) = c$ with $c \in C = \{"0", "1", "2"\}$ where "0","1","2" corresponds to the labels *NEGATIVE*, *NEUTRAL* and *POSITIVE*.

Encoder-Decoder: **BARTpho** (139M params) (Tran et al., 2022a) is the Vietnamese variant of BART (Lewis et al., 2019) trained on 20GB of Vietnamese text from Wikipedia and news corpus. **ViT5** (223M params) (Phan et al., 2022) is the Vietnamese version of T5 (Raffel et al., 2020) trained on 71GB of Vietnamese text from CC100 (Conneau et al., 2019).

Decoder: We use **Vistral-7B-Chat** (Nguyen et al., 2023) and **vmlu-llm**[1]. Both models have Mistral-7B(Jiang et al., 2023a) as their backbone. These models were chosen based on their performance on the **vmlu benchmark** (**V**ietnamese **M**ultitask **L**anguage **U**nderstanding)[2].

4.4 Training with Rationale

Previous works (Wadhwa et al., 2024; Chen et al., 2024; Hsieh et al., 2023; Ho et al., 2022) have shown that rationale-augmented targets consistently improve the performance of generative lan-

[1]https://huggingface.co/vtrungnhan9/vmlu-llm
[2]https://vmlu.ai/leaderboard

guage models. Our rationale-augmented training methods are based on, to our knowledge, the current state-of-the-art CoT-distillation approaches for each architecture.

(i) **Multitask Training** (Hsieh et al., 2023): We train our encoder-decoders using distilling step-by-step. Distilling step-by-step is a multitask training approach that prepends particular prefixes to the input, guiding the model to output either the answer or generate a rationale. Hsieh et al. found that it consistently improves encoder-decoders performance compared with single-task training which treats rationale and label predictions as a single task.

(ii) **Post-thinking** (Chen et al., 2024): For decoder-based models, we augment the training targets by append the human rationale to the label (<LABEL> <RATIONALE>) in a single prompt. Previous works have shown that post-thinking achieved impressive performance (Chen et al., 2024; Wadhwa et al., 2024) and compared to pre-thinking where the model first generates its CoT then provide the label (<RATIONALE> <LABEL>), post-thinking is more stable and token-efficient (Chen et al., 2024; Wadhwa et al., 2024) as the model suffers less from hallucination, consistently yields better performance and is more resource efficient as users can already retrieve the target label from the first generated token.

4.5 Rationale Format

While the rationale in our dataset were re-labeled by humans, we are also interested in **whether a different and more detailed rationale format would help the models learn better**. To this end, we further study the effects of the format of the rationale on the performance of the generative models. In particular, given the human rationale and human label, we further prompt GPT-3.5-turbo to enhance the rationale into two different format:

Elaborated rationale: An elaborated version of the human rationale that is 1-2 sentence(s) long, grounded on the provided human rationale and the sentiment label.

CoT rationale: A step-by-step, elaborated version of the human rationale, which includes the following steps: (1) identifies the medical entity, (2) extracts the progress of the corresponding medical entity in the transcript, and (3) provides the elaborated rationale on the sentiment grounded on the provided human rationale, the sentiment label, and information from steps (1) and (2). This approach is inspired by aspect-based sentiment instruction-tuning approaches (Varia et al., 2022).

4.6 Evaluation Metrics

For Sentiment Classification task, we employ accuracy and class-wise F1 score. For Rationale Generation , we employ ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score (Lin, 2004). Also, we employ BERTScore (Zhang et al.) which captures the contextual and semantic nuances. BERTscore has shown to correlate well with human judgment.

5 Results and Analysis

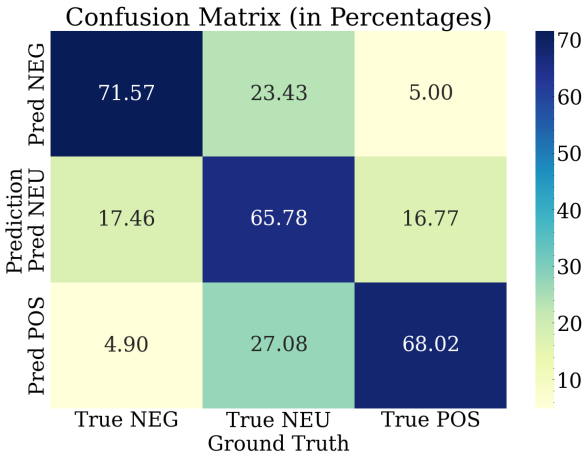

Figure 2: Confusion matrix of the predicted classes versus the actual labels on human transcript, obtained from Vistral7B trained with human rationale

We evaluate and analyze our models performance on Table 3. Based on the obtained results, we make the following observations:

1. Encoders are efficient yet effective Sentiment Classification **baselines**: Encoder models yields the best performance compared to their encoder-decoder and decoder counterparts, with high accuracy scores (> 0.665) and stable F1 scores (macro F1 of both models > 0.665). We further observe that **domain-specific encoders yield notably better performance**, with ViHealthBERT outperforming phoBERT in accuracy (+0.8%) and macro F1 (+0.9%).

2. ASR errors have a marginally negative impact on Sentiment Classification **performance**: For a fair comparison in real-world environments, WERs for human annotators on a standard conversational spontaneous English ASR dataset range from 5% to 15% (Stolcke and Droppo, 2017) while

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-Lsum	BERTscore
Encoder (Label Only)										
PhoBERT	0.6674	0.6969	0.6607	0.6377	0.6651					
ViHealthBERT	0.6752	0.6970	0.6718	0.6535	0.6741					
Encoder-Decoder (Label Only)										
ViT5	0.6628	0.6922	0.6687	0.6007	0.6545					
BARTpho	0.6523	0.6870	0.6571	0.5841	0.6427					
Decoder (Label Only)										
vmlu-llm	0.6592	0.6768	0.6769	0.5911	0.6483					
Vistral7B	0.6716	0.6858	0.6771	0.6398	0.6676					
Encoder-Decoder (Label + Rationale)										
ViT5	0.6633	0.6936	0.6572	0.6335	0.6615	0.3910	0.2668	0.3653	0.3660	0.8093
BARTpho	0.6619	0.7029	0.6460	0.6265	0.6585	0.3871	0.2613	0.3658	0.3683	0.8077
Decoder (Label + Rationale)										
vmlu-llm	0.6729	0.7039	0.6714	0.6307	0.6687	0.3947	0.2467	0.3789	0.3796	0.8086
Vistral7B	0.6812	0.7152	0.6765	0.6425	0.6781	0.4155	0.2788	0.3880	0.3900	0.8101

Table 3: Baseline performance of encoders, encoder-decoders, and decoders on the Vietnamese human transcript. From left to right is: Accuracy, F1-{negative, neutral, positive, macro}, ROUGE-{1, 2, L, Lsum}, BERTscore. The **Label Only** models are models trained only with the label, serving as the baseline, while **Label + Rationale** indicates models trained with rationale. As the **Label Only** models are not trained to generate rationale, we do not evaluate them on ROUGE and BERTscore.

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-LSum	BERTscore
Encoder (Label Only)										
PhoBERT	0.6166	0.6418	0.6231	0.5658	0.6102					
ViHealthBERT	0.6198	0.6307	0.6261	0.5934	0.6167					
Encoder-Decoder (Label Only)										
ViT5	0.6157	0.6412	0.6258	0.5523	0.6064					
BARTpho	0.6056	0.6364	0.6156	0.5311	0.5944					
Decoder (Label Only)										
vmlu-llm	0.6216	0.6296	0.6551	0.5186	0.6011					
Vistral7B	0.6299	0.6377	0.6537	0.5609	0.6174					
Encoder-Decoder (Label + Rationale)										
ViT5	0.6189	0.6305	0.6286	0.5837	0.6143	0.3571	0.2202	0.3350	0.3366	0.8044
BARTpho	0.6129	0.6523	0.6028	0.5665	0.6072	0.3956	0.2652	0.3728	0.3774	0.8106
Decoder (Label + Rationale)										
vmlu-llm	0.6395	0.6585	0.6557	0.5723	0.6289	0.3853	0.2386	0.3663	0.3671	0.8092
Vistral7B	0.6354	0.6485	0.6479	0.5892	0.6285	0.3558	0.2237	0.3343	0.3394	0.7994

Table 4: Baseline performance of encoders, encoder-decoders, and decoders on the Vietnamese ASR transcript. Further information about our metrics can be found in Table 3.

more challenging real-world ASR datasets are between 17% and 31% (Mulholland et al., 2016). Given the complexity of real-world medical conversations, WER of 29.6% by our ASR model is within an acceptable range. Despite the WER of 29.6%, the performance drop in macro F1 scores is small (absolute value of only about 5%).

3. Rationale-augmented training improve model performance: Consistent with previous findings, performing CoT-augmented training on both encoder-decoders and decoders improve our models performance compared to the baseline. We further conducted a Student's t-test (Student, 1908) and found that the gains are statistically significant for $\alpha = 0.1$. This pattern holds for the results in Table 5. We observe a decline in all of our models performance on ASR data which is anticipated due to its WER of 29.6 %. Nonetheless, the models trained with rationale perform noticeably better than models without, with an average absolute accuracy gain of +0.85%, absolute macro F1 gain of +1.4%, and relative macro F1 gain of +2.5%.

4. The format of post-thinking rationale doesn't affect the generative models performance: We study the effects of the format of post-thinking rationale on the performance of generative models on Table 5 and observe that it is unclear whether there is a performance gain from more elaborated rationales. This result agrees with previous findings (Wadhwa et al., 2024).

5. Models are likely to misclassify *POSITIVE* and *NEGATIVE* transcripts as *NEUTRAL*: We

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1
Encoder-Decoder (Label + Rationale)					
ViT5_human	0.6633	0.6936	0.6572	0.6335	0.6615
ViT5_elaborate	0.6661	0.6903	0.6799	0.5985	0.6562
ViT5_cot	0.6619	0.6968	0.6552	0.6237	0.6586
BARTpho_human	0.6619	0.7029	0.6460	0.6265	0.6585
BARTpho_elaborate	0.6564	0.7031	0.6528	0.5870	0.6476
BARTpho_cot	0.6464	0.6922	0.6611	0.5287	0.6273
Decoder (Label + Rationale)					
Vistral7B_human	0.6812	0.7152	0.6765	0.6425	0.6781
Vistral7B_elaborate	0.6688	0.6846	0.6647	0.6564	0.6685
Vistral7B_cot	0.6706	0.6725	0.6807	0.6477	0.6670
vmlu-llm_human	0.6729	0.7039	0.6714	0.6307	0.6687
vmlu-llm_elaborate	0.6867	0.7203	0.6868	0.6353	0.6808
vmlu-llm_cot	0.6821	0.6966	0.6779	0.6711	0.6819

Table 5: Performance of generative models on the different rationale formats on our test set. Human/elaborate/CoT specifies the format of rationale the model was trained on. Details in Section 4.5

study the confusion matrix of our best model on human transcript, Vistral7B finetuned with human rationale, on Figure 2. We observe a notable misclassification tendency between *NEUTRAL* and the other two classes (23.43% and 27.08% with *NEGATIVE* and *POSITIVE* respectively). On the other hand, we found that models can easily distinguish *NEGATIVE* transcripts from *POSITIVE* ones. This reflects the ambiguity of sentiment analysis data. Furthermore, given the slightly imbalanced nature of our dataset with fewer *POSITIVE* examples, its average F1 score is the lowest among the three labels across all models.

6. Analysis of Generated Rationale: Compared to human rationale, we observe from Table 3 and Table 4 that the models trained with rationale have high BERTscore (around 0.8) with low ROUGE score, indicating that while the vocabulary used in the rationale is different, the overall semantic of the generated rationale remains similar to that of humans. Also, no noticeable changes in the semantic quality of rationale between human transcripts and ASR transcripts because BERTScore is still about 0.8 on both settings.

7. Results on end-to-end audio language models We report the results for end-to-end spoken sentiment analysis on PhoWhisper (Le et al., 2024) and Qwen2-Audio (Chu et al., 2024). Based on the results in Table 6, we make two observations: First, the performance of PhoWhisper is sub-optimal which we attribute to the fact that it was pre-trained for ASR-based tasks. Second, we found that **rationale-augmented training can also increase the** Sentiment Classification **performance** for audio language models.

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1
PhoWhisper	0.4651	0.4393	0.5277	0.3328	0.4333
Decoder (Label only)					
Qwen2-Audio	0.5815	0.5707	0.6150	0.5208	0.5688
Decoder (Label + Rationale)					
Qwen2-Audio	0.5884	0.5875	0.6131	0.5337	0.5781

Table 6: Performance of audio language models

6 Conclusion

In this work, we introduce a new task - Sentiment Reasoning - for both speech and text modalities, along with the framework and **the world's largest multimodal sentiment analysis dataset**. In Sentiment Reasoning, given an input transcript (human transcript or ASR transcript), the model learns to output the sentiment label (POSITIVE, NEUTRAL, or NEGATIVE) and its rationale (the reason for this label). It comprises of two tasks: Sentiment Classification and Rationale Generation.

We menticulously evaluate the use of rationale during training to improve our models' interpretability and performance. We found that rationale-augmented training improves model performance in Sentiment Classification in both human and ASR transcripts (**+2% increase in both accuracy and macro-F1**). We found that the generated rationales have different vocabulary to human rationale but with similar semantics. Finally, we found no major difference in the semantic quality of generated rationales between human and ASR transcripts.

7 Acknowledgement

We thank Anh Totti Nguyen at Auburn University and Jerry Ngo at MIT for insightful feedback.

8 Limitations

Hybrid ASR: This study utilized the hybrid ASR system, which is generally recognized as superior in performance compared to the attention-based encoder-decoder or end-to-end ASR systems (Lüscher et al., 2019; Prabhavalkar et al., 2023; Raissi et al., 2023). However, the hybrid ASR requires multiple steps, beginning with acoustic

feature extraction and progressing through GMM-HMM modeling before transitioning to DNN-HMM modeling, which complicates reproducibility for non-experts.

Cascaded speech sentiment analysis approach: While we do report the results for end-to-end systems, our main focus in this paper is on cascaded speech sentiment analysis for Sentiment Reasoning . This approach uses a previously trained ASR model to generate ASR transcripts that are subsequently input into a language model (LM) for downstream Sentiment Classification and Rationale Generation tasks. Consequently, the weights in the ASR model remain unchanged while the LM weights are updated. In this setting, only semantic features from speech are utilized, omitting other trainable acoustic features, like prosody, tones, etc. In spoken language processing, where semantic features play a more important role than other acoustic features, cascaded approach is prefered due to its straightforwardness, simplicity and superior accuracy (Lu, 2023; Bentivogli et al., 2021; Tran et al., 2022b; Tseng et al., 2023). Future work should consider the end-to-end sentiment analysis task, where weights in both the ASR model and LM are updated simultaneously, as it might hold promise for improved performance.

References

T A Al-Qablan, M H Mohd Noor, M A Al-Betar, and A T Khader. 2023. A survey on sentiment analysis and its applications. *Neural Computing and Applications*, 35(29):21567–21601.

Shivaji Alaparthi and Manit Mishra. 2020. Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey.

Shivaji Alaparthi and Manit Mishra. 2021. BERT: a sentiment analysis odyssey. *J. Mark. Anal.*, 9(2):118–126.

Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen. 2013. Can I hear you? sentiment analysis on medical forums. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 667–673, Nagoya, Japan. Asian Federation of Natural Language Processing.

Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. 2021. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088.

Oscar Araque, Ignacio Corcuera-Platas, J Fernando Sánchez-Rada, and Carlos A Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.*, 77:236–246.

Roza H Hama Aziz and Nazife Dimililer. 2020. Twitter sentiment analysis using an ensemble weighted majority vote classifier. In *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 103–109. IEEE.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? 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 2873–2887, Online. Association for Computational Linguistics.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, Chong Zhou, John Yen, Greta E Greer, and Kenneth Portier. 2013. Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 413–417.

Leo Breiman. 2017. *Classification and regression trees*. Routledge.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Erik Cambria, Tim Benson, Chris Eckl, and Amir Hussain. 2012a. Sentic proms: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, 39(12):10533–10543.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012b. The hourglass of emotions. In *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, February 21-26, 2011, revised selected papers*, pages 144–157. Springer.

Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. 2020. A large scale speech sentiment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6549–6555.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. 2022. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Xiao Chen, Sihang Zhou, Ke Liang, and Xinwang Liu. 2024. Post-semantic-thinking: A robust strategy to distill reasoning capacity from large language models. *arXiv preprint arXiv:2404.09170*.

Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2018. Deep neural networks for emotion recognition combining audio and transcripts. In *Interspeech*, pages 247–251.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

M D Deepa. 2021. Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(7):1708–1721.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

K Devipriya, D Prabha, V Pirya, and S Sudhakar. 2020. Deep learning sentiment analysis for recommendations in social applications. *Int J Sci Technol Res*, 9(1):3812–3815.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sefik Emre Eskimez, Zhiyao Duan, and Wendi Heinzelman. 2018. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103. IEEE.

G.D. Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Usha Devi Gandhi, Priyan Malarvizhi Kumar, Gokulnath Chandra Babu, and Gayathri Karthick. 2021. Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM). *Wirel. Pers. Commun.*

Ofer Golan, Simon Baron-Cohen, Jacqueline J Hill, and Yael Golan. 2006. The "reading the mind in films" task: complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience,*, 1(2):111–123.

Irving John Good. 1952. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2225. NIH Public Access.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

M Hoang, O A Bihorac, and J Rouces. 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023b. Mistral 7b. *Preprint*, arXiv:2310.06825.

Lakshmish Kaushik, Abhijeet Sangwan, and John H L Hansen. 2017. Automatic sentiment detection in naturalistic audio. *IEEE ACM Trans. Audio Speech Lang. Process.*, 25(8):1668–1679.

J D M W C Kenton and L K Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Eesung Kim and Jong Won Shin. 2019. Dnn-based emotion recognition based on bottleneck acoustic features and lexical features. In *ICASSP 2019-2019 IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6720–6724. IEEE.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Senthil Kumar and B Malarvizhi. 2020. Bi-directional LSTM-CNN combined method for sentiment analysis in part of speech tagging (PoS). *International Journal of Speech Technology*, 23:373–380.

Sanna Kuusikko, Helena Haapsamo, Eira Jansson-Verkasalo, Tuula Hurtig, Marja-Leena Mattila, Hanna Ebeling, Katja Jussila, Sven Bölte, and Irma Moilanen. 2009. Emotion recognition in children and adolescents with autism spectrum disorders. *Journal of autism and developmental disorders*, 39:938–945.

Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. 2019. Incorporating end-to-end speech recognition models for sentiment analysis. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7976–7982. IEEE.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. PhoWhisper: Automatic Speech Recognition for Vietnamese. In *Proceedings of the ICLR 2024 Tiny Papers track*.

Khai Le-Duc. 2024. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai. 2018. An attention pooling based representation learning method for speech emotion recognition.

Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng. 2019. Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6675–6679. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiting Lu. 2023. *Improving cascaded systems in spoken language processing*. Ph.D. thesis.

Zhiyun Lu, Liangliang Cao, Yu Zhang, Chung-Cheng Chiu, and James Fan. 2020. Speech sentiment analysis via pre-trained features from end-to-end ASR models. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention. In *Proc. Interspeech 2019*, pages 231–235.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet, and Florence Galtier. 2014. Patient's rationale: Patient knowledge retrieval from health forums.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. ViHealthBERT: Pre-trained language models for Vietnamese in health text mining. In *Proceedings*

of the Thirteenth Language Resources and Evaluation Conference, pages 328–337, Marseille, France. European Language Resources Association.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Matthew Mulholland, Melissa Lopez, Keelan Evanini, Anastassia Loukina, and Yao Qian. 2016. A comparison of asr and human errors for transcription of non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5855–5859. IEEE.

Jin-Cheon Na, Wai Yan Min Kyaing, Christopher SG Khoo, Schubert Foo, Yun-Ke Chang, and Yin-Leng Theng. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. In *The Outreach of Digital Libraries: A Globalized Resource Network: 14th International Conference on Asia-Pacific Digital Libraries, ICADL 2012, Taipei, Taiwan, November 12-15, 2012, Proceedings 14*, pages 189–198. Springer.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Chien Van Nguyen, Thuat Nguyen, Quan Nguyen, Huy Nguyen, Björn Plüster, Nam Pham, Huu Nguyen, Patrick Schramowski, and Thien Nguyen. 2023. Vistral-7b-chat - towards a state-of-the-art large language model for vietnamese.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *AMIA annual symposium proceedings*, volume 2005, page 570. American Medical Informatics Association.

Nir Ofek, Cornelia Caragea, Lior Rokach, Prakhar Biyani, Prasenjit Mitra, John Yen, Kenneth Portier, and Greta Greer. 2013. Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In *2013 international conference on social intelligence and technology*, pages 109–113. IEEE.

Stefan Ortmanns, Hermann Ney, and Xavier Aubert. 1997. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language*, 11(1):43–72.

Subarno Pal, Soumadip Ghosh, and Amitava Nag. 2018. Sentiment analysis in the light of LSTM recurrent neural networks. *Int. J. Synth. Emot.*, 9(1):33–39.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. *Preprint*, arXiv:2205.06457.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tina Raissi, Christoph Lüscher, Moritz Gunz, Ralf Schlüter, and Hermann Ney. 2023. Competitive and resource efficient factored hybrid hmm systems are simpler than you think. In *Interspeech*, Dublin, Ireland.

T-YLPG Ross and GKHP Dollár. 2017. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988.

Nabizath Saleena et al. 2018. An ensemble classification system for twitter sentiment analysis. *Procedia computer science*, 132:937–946.

Abeed Sarker, Diego Mollá-Aliod, and Cécile Paris. 2011. Outcome polarity identification of medical papers. In *Proceedings of the Australasian language technology association workshop 2011*, pages 105–114. Australian Language Technology Association.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Izhak Shafran and Richard Rose. 2003. Robust speech detection and segmentation for real-time asr applications. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Hashim Sharif, Fareed Zaffar, Ahmed Abbasi, and David Zimbra. 2014. Detecting adverse drug reactions using a sentiment classification framework.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2023. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937.

Suwon Shon, Pablo Brusco, Jing Pan, Kyu J Han, and Shinji Watanabe. 2021a. Leveraging pre-trained language model for speech sentiment analysis. In *Interspeech 2021*, ISCA. ISCA.

Suwon Shon, Pablo Brusco, Jing Pan, Kyu J Han, and Shinji Watanabe. 2021b. Leveraging pre-trained language model for speech sentiment analysis. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 566–570. International Speech Communication Association.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Phillip Smith and Mark Lee. 2012. Cross-discourse development of supervised sentiment analysis in the clinical domain. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 79–83, Jeju, Korea. Association for Computational Linguistics.

Marina Sokolova, Stan Matwin, Yasser Jafer, and David Schramm. 2013. How joe and jane tweet about their health: mining for personal health information on twitter. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 626–632.

Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. BERT for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Akana Chandra Mouli Venkata Srinivas, Ch Satyanarayana, Ch Divakar, and Katikireddy Phani Sirisha. 2021. Sentiment analysis using neural network and LSTM. *IOP Conf. Ser. Mater. Sci. Eng.*, 1074(1):012007.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. *Interspeech*.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Ivan J Tashev and Dimitra Emmanouilidou. 2019. Sentiment detection from ASR output. In *2019 International Conference on Information Technologies (InfoTech)*. IEEE.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022a. Bartpho: Pre-trained sequence-to-sequence models for vietnamese. *Preprint*, arXiv:2109.09701.

Viet Anh Khoa Tran, David Thulke, Yingbo Gao, Christian Herold, and Hermann Ney. 2022b. Does joint training really help cascaded speech translation? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4480–4487.

Yuan Tseng, Cheng-I Jeff Lai, and Hung-yi Lee. 2023. Cascading and direct approaches to unsupervised constituency parsing on spoken sentences. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. 2018. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2022. Instruction tuning for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2210.06629*.

Esaú Villatoro-Tello, S Pavankumar Dubagunta, Julian Fritsch, Gabriela Ramírez-de-la Rosa, Petr Motlicek, and Mathew Magimai-Doss. 2021. Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. In *Interspeech*, pages 1927–1931.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2024. Investigating mysteries of cot-augmented distillation. *arXiv preprint arXiv:2406.14511*.

Chengyi Wang, Yu Wu, Shujie Liu, Jinyu Li, Yao Qian, Kenichi Kumatani, and Furu Wei. 2021a. UniSpeech at scale: An empirical study of pre-training method on large-scale speech recognition dataset.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021b. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings*

of Machine Learning Research, pages 10937–10947. PMLR.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:581–591.

Travis Wearne, Katherine Osborne-Crowley, Hannah Rosenberg, Marie Dethier, and Skye McDonald. 2019. Emotion recognition depends on subjective emotional experience and not on facial expressivity: evidence from traumatic brain injury. *Brain injury*, 33(1):12–22.

Xixin Wu, Songxiang Liu, Yuewen Cao, Xu Li, Jianwei Yu, Dongyang Dai, Xi Ma, Shoukang Hu, Zhiyong Wu, Xunying Liu, et al. 2019. Speech emotion recognition using capsule networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6695–6699. IEEE.

Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors.

Lei Xia, Anna Lisa Gentile, James Munro, and José Iria. 2009. Improving patient opinion mining through multi-step classification. In *Text, Speech and Dialogue: 12th International Conference, TSD 2009, Pilsen, Czech Republic, September 13-17, 2009. Proceedings 12*, pages 70–76. Springer.

Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. 2019. Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimed. Syst.*, 26(4):431–451.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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

Zixing Zhang, Bingwen Wu, and Björn Schuller. 2019. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709. IEEE.

Chunjun Zheng, Chunli Wang, and Ning Jia. 2022. A two-channel speech emotion recognition model based on raw stacked waveform. *Multimedia Tools and Applications*, 81(8):11537–11562.

Contents

1 Introduction **1**

2 Data **2**
- 2.1 Data Collection . . . 2
- 2.2 Data Annotation . . . 2
- 2.3 Data Quality Control . . . 3
- 2.4 Data Statistics . . . 3

3 Sentiment Reasoning Framework **3**
- 3.1 Informal Definition . . . 3
- 3.2 Formal Definition . . . 3
- 3.3 ASR Model . . . 3
- 3.4 Language Model for Sentiment Reasoning . . . 3
  - 3.4.1 Sentiment Classification . . . 3
  - 3.4.2 Rationale Generation . . . 4

4 Experimental Setups **4**
- 4.1 ASR Model . . . 4
- 4.2 End-to-end Sentiment Classification . . . 4
- 4.3 Language Model for Sentiment Reasoning . . . 4
  - 4.3.1 Encoder . . . 4
  - 4.3.2 Generative Models . . . 4
- 4.4 Training with Rationale . . . 4
- 4.5 Rationale Format . . . 5
- 4.6 Evaluation Metrics . . . 5

5 Results and Analysis **5**

6 Conclusion **7**

7 Acknowledgement **7**

8 Limitations **7**

A Related Works **16**
- A.1 Multimodal Speech Sentiment Analysis . . . 16
- A.2 ASR-based Speech Sentiment Analysis . . . 16
- A.3 Speech Sentiment Analysis in Healthcare . . . 17

B Details about Data **18**
- B.1 Data Annotation Pipeline . . . 18
- B.2 LLM Prompt for Pre-labeling . . . 18
- B.3 Annotation Guidelines . . . 18
  - B.3.1 Output Annotation . . . 18
- B.4 Annotation Flowchart . . . 19
- B.5 Data Imbalance Discussion . . . 19
- B.6 Translation Annotation Process and Translation Quality Control . . . 19
- B.7 Data Samples . . . 20

C Details about Experimental Setups **22**
C.1 Details of ASR Experiments 22
C.2 Training Setup 22
C.3 Student's T-Test 22

D Results on English subset **24**

E Results on end-to-end audio language models **26**
E.1 Encoder-Based 26
E.2 Audio LLMs 26

F Error Analysis **27**

A Related Works

A.1 Multimodal Speech Sentiment Analysis

It is widely known that there have been two research directions in the field of speech sentiment analysis, as also confirmed by Chen et al. (2020).

- **Single modality model (unimodal)**: In speech sentiment analysis, single modality models focus on utilizing a single type of data to predict sentiment. These models may rely exclusively on acoustic features, such as pitch, tone, and rhythm, to infer emotional states from spoken language (Li et al., 2019, 2018; Wu et al., 2019; Xie et al., 2019). Alternatively, they might use raw waveforms (Tzirakis et al., 2018; Zheng et al., 2022; Villatoro-Tello et al., 2021) or the textual content of transcripts to predict sentiment (Lakomkin et al., 2019). The strength of single modality models lies in their simplicity and specialization, allowing them to hone in on specific attributes of the data source they are designed for. However, this specialization can also be a limitation, as these models might miss out on the richer, more nuanced information that can be gleaned from combining multiple data types. Despite this, single modality models remain a fundamental approach in the field, providing valuable insights and serving as a benchmark for more complex multimodal systems.

- **Multimodality models**: In speech sentiment analysis, multimodality models leverage the combined strengths of both acoustic and textual data to provide more accurate and nuanced sentiment predictions. While traditional models might rely solely on either the acoustic features—such as tone, pitch, and rhythm—or the text derived from speech transcripts, multimodal models integrate these two data streams. This integration allows for a more holistic understanding of sentiment, as it captures the emotional cues present in the speaker's voice along with the contextual and semantic content of the spoken words. By maximizing the mutual information between these modalities, multimodal models can better discern subtleties in speech that single modality models might miss, leading to accuracy improvements (Kim and Shin, 2019; Cho et al., 2018; Gu et al., 2018; Eskimez et al., 2018; Zhang et al., 2019).

Our dataset is ideal for both single modal and multimodal research, as it includes both acoustic and text features.

A.2 ASR-based Speech Sentiment Analysis

Speech sentiment analysis on ASR transcripts is a field that aims to interpret and classify sentiments conveyed in spoken language. As technology advances, ASR systems have become increasingly proficient at transcribing spoken words into text with high accuracy (Schneider et al., 2019; Baevski et al., 2020, 2019; Wang et al., 2021b; Chen et al., 2022; Wang et al., 2021a), providing a rich source of data for sentiment analysis. Sentiment analysis algorithms then analyze the transcribed text from speech signal, utilizing language models as decoders to detect positive, negative, or neutral sentiments (Lu et al., 2020; Shon et al., 2021a; Wu et al., 2022; Tashev and Emmanouilidou, 2019; Kaushik et al., 2017).

In the era of deep learning, as surveyed by Al-Qablan et al. (2023), many researchers have been applying deep learning methods to the sentiment analysis process on transcript, leading to the development of various models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BLSTM) (Araque et al., 2017; Devipriya et al., 2020; Yadav and Vishwakarma, 2020). CNNs, primarily used for image processing, have been adapted for text by treating sentences as sequences of words and applying convolutional filters to capture local features. This approach helps in identifying crucial patterns within the text that are indicative of sentiment (Kumar and Malarvizhi, 2020; Wang et al., 2020). On the other hand, RNNs are designed to handle sequential data by maintaining a hidden state that captures the history of previous inputs, making them suitable for understanding the context and temporal dependencies in sentences. However, traditional RNNs face challenges with long-term dependencies due to issues like vanishing gradients, which is where LSTMs come in. LSTMs, an advanced form of RNNs, address these issues by incorporating gates that regulate the flow of information, allowing them to maintain and update long-term dependencies effectively. Furthermore, BLSTMs enhance this by processing the input sequence in both forward and backward directions, thus capturing dependencies from both past and future contexts simultaneously. This bidirectional approach is especially useful for

sentiment analysis, where the interpretation of a word can depend heavily on both preceding and succeeding words. Together, these architectures provide powerful tools for sentiment analysis, each contributing unique strengths that can be leveraged depending on the specific requirements and characteristics of the data at hand (Gandhi et al., 2021; Pal et al., 2018; Srinivas et al., 2021).

Developed by Google, BERT (Bidirectional Encoder Representations from Transformers) (Kenton and Toutanova, 2019) revolutionized NLP tasks by enabling models to understand the context of words in a sentence more effectively through its bidirectional training approach. Unlike previous models that read text input sequentially, BERT reads the entire sequence of words at once, capturing the full context and nuances of language. This capability allows BERT to excel in sentiment analysis, where understanding the subtleties of human emotion and opinion is paramount (Alaparthi and Mishra, 2020; Deepa, 2021). BERT's pre-training on vast amounts of text data, followed by fine-tuning on specific sentiment analysis tasks, further enhances its performance. By leveraging its powerful language representations, BERT can handle the complexities of sentiment analysis, such as sarcasm, idiomatic expressions, and context-dependent sentiment shifts, making it a preferred choice for applications ranging from social media monitoring to customer feedback analysis. The model's ability to generalize across various domains and languages also contributes to its widespread adoption, offering robust and scalable solutions for sentiment analysis in diverse settings (Hoang et al., 2019; Xu et al., 2019; Sousa et al., 2019; Alaparthi and Mishra, 2021).

A.3 Speech Sentiment Analysis in Healthcare

Sentiment analysis in healthcare is an emerging field that leverages NLP and machine learning techniques to analyze and interpret the emotional tone conveyed in biomedical textual data. This technology is particularly useful for understanding patient feedback, monitoring public health trends, and improving patient-provider communication. By analyzing large volumes of data from sources such as social media, online reviews, electronic health records (EHRs), and patient surveys, sentiment analysis can provide valuable insights into patient experiences, satisfaction levels, and overall public sentiment towards healthcare services and policies. For instance, analyzing patient reviews on healthcare platforms can help identify common concerns and areas needing improvement, allowing healthcare providers to address issues proactively and enhance the quality of care. Additionally, sentiment analysis can play a critical role in mental health monitoring by detecting signs of distress or dissatisfaction in patient communications, enabling timely intervention and support. As this technology continues to evolve, it holds the promise of transforming healthcare by fostering a more patient-centric approach, enhancing service delivery, and ultimately improving patient outcomes (Denecke and Deng, 2015). However, the sentiments expressed in clinical narratives have not been extensively analyzed or exploited, based on the total number of previous works we have identified to the best of our knowledge:

- Sentiment analysis from the medical web: Most sentiment analysis research in the medical domain focuses on web data, such as medical blogs and forums, to mine patient opinions or assess quality (Ali et al., 2013; Xia et al., 2009; Na et al., 2012; Sokolova et al., 2013; Biyani et al., 2013; Ofek et al., 2013; Smith and Lee, 2012; Sharif et al., 2014; Melzi et al., 2014).

- Sentiment analysis from biomedical literature: In addition to the analysis of medical social media data, biomedical literature has been examined concerning the outcomes of medical treatments. Within this framework, sentiment denotes the results or efficacy of a treatment or intervention (Niu et al., 2005; Sarker et al., 2011).

- Sentiment analysis from medical text (except biomedical literature): Several researchers have focused on leveraging supplementary sources of medical texts to implement sentiment analysis and emotion detection methodologies, suicide notes or patient questionnaire for example (Pestian et al., 2012; Cambria et al., 2012a; Liu and Singh, 2004; Cambria et al., 2012b).

To the best of our knowledge, no literature among those cited has addressed speech sentiment analysis specifically within the domain of healthcare.

B Details about Data

B.1 Data Annotation Pipeline

We use LLM pre-labeling as it helps speed up the labeling process through providing the annotators with the initial sentiment labels and the corresponding rationales. In the relabeling process, annotators go through each sample and inspect it manually. If the annotators deem the label and the rationale is appropriate, they can quickly move to the next sample. If not, the annotators can update the label and rationale to be more appropriate.

The data annotation process is as followed. First, all the subtitles are separated into different chunks. These segments are subsequently input into gpt-3.5-turbo, which conducts a weakly supervised 3-label classification task to categorize each segment as *NEGATIVE, NEUTRAL, or POSITIVE*. In addition to the sentiment label, gpt-3.5-turbo also provides a brief synthetic rationales for the classification, such as 'Negative medical condition' or 'Objective description'. The labels and rationales generated by gpt-3.5-turbo are subsequently reviewed and independently corrected by a team of 3 developers.

B.2 LLM Prompt for Pre-labeling

gpt-3.5-turbo

Annotate the sentiment (neutral, positive or negative) of the following sentence and provide a very short justification. The procedure is as follows:
1. If the segment shows clear emotional signs, annotate based on these signs.
2. If no emotional markings are present, determine if the segment is an objective description. Positive for beneficial facts/features, negative for detrimental facts/features, and neutral otherwise.
3. If not objective, check if there's a preference expression. Positive for likes or positive views, negative for dislikes or negative views, and neutral if no preference is expressed.
4. If too short to determine sentiment, label as neutral.
{3 in-context learning examples}

B.3 Annotation Guidelines

The definition of "sentiment" encompasses both "emotions" and "facts" in our work. Existing works (Chen et al., 2020; Mohammad, 2016; Shon et al., 2021b, 2022, 2023) use both emotions and facts for sentiment labeling.

- Emotion: Existing literature includes "emotion" as part of "sentiment" (Chen et al., 2020; Shon et al., 2021b; Mohammad, 2016) and sentiment analysis can be considered a more abstract level of emotion recognition, e.g. polarity of emotions (Mohammad, 2016).
- Facts: Many sentiment analysis systems require statements that describe events/situations to be given a sentiment label (Chen et al., 2020; Mohammad, 2016).

The annotation task consists of two primary steps. First, annotators are required to perform Sentiment Classification. Second, annotators are instructed to provide a rationale behind each class (Rationale Generation).

To ensure consistency, our TESOL-certificated professional linguist has developed an initial guideline inspired by (Chen et al., 2020), which was also adopted by various well-known works (Shon et al., 2022, 2023), and revised it frequently if necessary as followed:

B.3.1 Output Annotation

The *NEGATIVE* label is for chunks that discuss negative, serious diseases, disorders, symptoms, risks, negative emotions, or counter-positive statements (e.g. "This would NOT bring a good outcome"). It also applies to incomplete chunks where the amount of negativity is greater than the amount of positivity.

The *NEUTRAL* label is for incomplete chunks where the ratio of negativity is equal to the ratio of positivity, as well as chunks that describe processes, ask questions, provide advice, or are too short.

The *POSITIVE* label is for chunks that discuss positive outcomes, recovery processes, positive emotions, or counter-negative statements (e.g. "This will *reduce* discrimination"). It also applies to incomplete chunks where the ratio of positivity is greater than the ratio of negativity.

It is important to note that all chunks are considered independent, even though they may be incomplete and related to preceding or following chunks. Given that this data is derived from spoken language, the chunks contain a significant amount of filler words, which are disregarded in the labeling

process. The majority of the *NEUTRAL* labels are attributed to chunks that involve sharing advice or descriptions. Additionally, the presence of modal verbs (e.g., should, would, need) often indicates advice sharing, thereby classifying the chunk as *NEUTRAL* regardless of its content.

B.4 Annotation Flowchart

Inspired by the well-known annotation flowchart provided by Chen et al. (2020), we asked annotators to adopt the annotation flowchart and we ,if necessary, revised as follows:

1. Does the segment exhibit distinct emotional cues indicative of sentiment, such as laughter for positive affect or yelling for negative affect?
    - **Yes** – Annotate the corresponding class and also note that:
        - (a) In some instances, individuals may laugh to mitigate the discomfort associated with delivering negative statements. In such cases, it should be classified as neutral.
        - (b) If individuals exhibit a sneer (a smile or laughter with a mocking tone), the corresponding sentiment should be classified as negative in such instances.
    - **No** - Jump into Step 2
2. Does the segment provide an objective account of the facts?
    - **Yes** - If the segment lists several positive attributes (e.g., good progress in medical treatment, good signs of health improvement), it is classified as positive. Conversely, if it lists several negative attributes, it is classified as negative. In the absence of a clear preponderance of either, the segment is considered neutral.
    - **No** - Jump into Step 3
3. Does the segment exhibit a preference?
    - **Yes** - If the subjective opinion or preference conveys a like or dislike, or expresses a positive (e.g., "it is beneficial that...") or negative sentiment, it should be annotated accordingly.
    - **No** - It's neutral
4. If the utterance is insufficient in length to accurately assess sentiment, it should be classified as neutral.

B.5 Data Imbalance Discussion

As shown in Table 1, *NEUTRAL* category is the most predominant, accounting for a significant portion of the dataset. With 3802 instances for both train and test set, *NEUTRAL* sentiments make up approximately half of the dataset. This prevalence of *NEUTRAL* sentiment is expected, as also seen by a real-world conversational dataset (Chen et al., 2020), given the nature of medical consultations, which often involve objective descriptions, explanations, and advice. The *NEGATIVE* category is the second most common, with around 2395 instances. *NEGATIVE* sentiments include discussions about serious diseases, negative emotions, and adverse medical outcomes. The substantial presence of negative sentiments reflects the medical context, where discussions about illnesses and symptoms are common. The *POSITIVE* category, while the least common, still represents a significant portion of the dataset with 1681 instances. *POSITIVE* sentiments typically involve discussions about recovery processes, positive outcomes, and favorable emotions.

A slight bias in the distribution of the labels towards NEUTRAL in our dataset (49.94% in the train set, 43.88% in the test set) reflects the nature of real-world medical conversations, rather than a weakness of our work. For context, in comparable real-world sentiment analysis datasets such as Switchboard-Sentiment (Chen et al., 2020), the distribution is as follows: 30.4% of the speech segments are labelled as POSITIVE, 17% of the segments are labelled as NEGATIVE, and 52.6% of the segments are labelled as NEUTRAL.

To address this labeling bias issue, future works can leverage techniques for fine-tuning models in data imbalance regimes, such as focal loss (Ross and Dollár, 2017), class weighting (King and Zeng, 2001).

B.6 Translation Annotation Process and Translation Quality Control

The data were initially translated from the source language into target languages (many-to-many) using the Gemini Large Language Model (LLM). Following the annotation process by **?**, the LLM-generated translated transcripts were treated as outputs from a *real* human annotator. In the data qual-

ity process, five human annotators manually corrected and then cross-verified *all* these translations based on the context of the whole conversation. Only transcripts that received consensus approval from multiple annotators were retained, resulting in an inter-annotator agreement of 100%.

All human annotators possessed a professional language proficiency of C1 or higher (or HSK5 for Chinese) in their respective working languages. Additionally, each annotator had completed basic medical training and demonstrated substantial knowledge of medical terminology in their selected language. Furthermore, they were either currently pursuing or had completed undergraduate or graduate studies in countries where their chosen language is predominantly spoken.

B.7 Data Samples

Table 7 shows 9 examples with 3 samples per sentiment label in our dataset. As the Vietnamese transcripts are obtained from short-formed audio, the transcripts contain characteristics of spoken language which serve as noises to the model (e.g. stuttering, hesitation, etc). **In our English translation, we aim to retain these properties, leading to unnatural, incomplete sentence with broken wording**.

Figure 3 shows 3 examples per sentiment label for all languages: Vietnamese, English, Chinese (Simplified and Traditional), German and French.

Transcript	ENG Translation	Label	Rationale
bệnh nhân sẽ có những cái rối loạn về mặt cảm xúc đôi khi có những bệnh nhân đã rơi vào trạng thái trầm cảm và đôi khi	The patient will suffer from emotional disorder and sometimes depression	NEG.	Emotional disorder
não đột quỵ đó thì nó liên quan đến việc hình thành các cục máu đông và việc cục máu đông đã nó trôi ra là đi	Stroke is related to the formation of blood clots and the fact that these blood clots travel	NEG.	Negative medical condition
nhầm lẫn với một cái nhóm thuốc khác đó là nhóm thuốc gọi là thuốc chống tiểu cầu tiểu cầu mà cụ	It's often confused with antiplatelet drugs	NEG.	Confusion
điểm cần thiết phải lưu tâm rõ ràng là cái người là bị béo phì đó	A crucial point is that the overweight patient	NEU.	Sharing advice
ra đó là cái hormone cortisol trong máu cũng như là hormone về catecholamine nó	The cortisol hormone in blood as well as catecholamine	NEU.	Objective description of hormones
có thể gọi đây là thuốc lẫn máu hay là một số cái tên khác mà thì nó có thể	You could call these blood-thinning drugs or other names, and it can	NEU.	Objective description
của nó không có cao nhưng mà rất là hình thức thì rất là may mắn là những năm gần đây thì mình có một cái nhóm thuốc khác	It is not expensive, luckily, in recent years there are another group of medicine	POS.	Expressing luck
để mà giảm xóa bỏ cái chuyện hình thành cái cục máu đông đó hiện ta sẽ dùng một số biện pháp trong đó thì chủ	To reduce and eliminate the formation of these blood clots, we use several measures, one of which is	POS.	Avoid forming blood clots
nhóm thuốc này á thì nó là rất là lâu đời và nó không có mất tiền rất là rẻ là	This group of drugs has been around for a very long time and is very cheap, with no cost	POS.	Long-standing and inexpensive medication

Table 7: 9 examples with 3 samples per sentiment label and its corresponding rationale

text	label	rationale	rationale_english	English	Chinese	raditional_chines	French	German
gì có phải là do cái cơn khó thở hay là còn có chuyện gì khác đôi khi nó có thể là hai ba nguyên nhân cùng một lúc nó	negative	lo lắng và không chắc chắn	worry and uncertainty	Is it due to shortness of breath, or is there something else going on? Sometimes it can be two or three causes at the same time.	是呼吸困难造成的，还是其他原因？有时可能是两个或三个原因同时发生。	是因呼吸困難所致，抑或其他原因？有時可能是兩個或三個原因同時發生。	Est-ce dû à un essoufflement, ou y a-t-il autre chose qui se passe ? Parfois, il peut y avoir deux ou trois causes en même temps.	Liegt es an der Atemnot, oder gibt es noch etwas anderes? Manchmal können es auch zwei oder drei Ursachen gleichzeitig sein.
chưa mạch máu của chúng ta vấn đề gì chưa tim của chúng ta có vấn đề chưa hoặc là chúng ta có cần có những cái	neutral	mô tả khách quan	objective description	Is there anything wrong with our blood vessels? Is there anything wrong with our heart? Or do we need anything?	我们的血管没有问题吗？我们的心脏没有问题吗？或者我们是否需要一些东西？	我們的血管沒有問題嗎？我們的心臟沒有問題嗎？或者我們是否需要一些額外的檢查或治療？	Y a-t-il un problème avec nos vaisseaux sanguins ? Y a-t-il un problème avec notre cœur ? Ou avons-nous besoin de quelque chose ?	Stimmt etwas mit unseren Blutgefäßen nicht? Stimmt etwas mit unserem Herzen nicht? Oder brauchen wir etwas?
một số các cái giải pháp điều trị nó vừa tin cậy với mình vừa an toàn với mình mà nó có thể đồng hành với mình trong cái	positive	sự tự tin, an toàn	confidence, safety	Some treatment solutions are both reliable and safe for me, and they can accompany me in the	一些治疗方案对我来说既可靠又安全，而且可以陪伴我一起	一些治療方案對我來說既可靠又安全，而且可以伴隨我一起	Certain solutions de traitement sont à la fois fiables et sûres pour moi, et elles peuvent m'accompagner dans le	Einige Behandlungslösungen sind sowohl zuverlässig als auch sicher für mich, und sie können mich auf dem Weg begleiten

Figure 3: Some samples from our dataset with versions all available languages.

C Details about Experimental Setups

C.1 Details of ASR Experiments

We employed hybrid ASR setup using wav2vec 2.0 encoder (Le-Duc, 2024) to transcribe speech to text. First, we generated alignments obtained by using Gaussian-Mixture-Model/Hidden-Markov-Model (GMM/HMM) as labels for wav2vec 2.0 (Baevski et al., 2020) neural network training. The labels used in the acoustic modeling are context-dependent phonemes, triphones in this case. In GMM/HMM process, we used a CART (Classification And Regression Tree) (Breiman, 2017) to tie the states s, resulting 4501 CART labels:

$$p(x_1^T|w_1^N) = \sum_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, w_1^N)$$
$$= \sum_{[s_1^T]} \prod_{t=1}^{T} \underbrace{p(s_t|s_{t-1}, w_1^N)}_{\text{transition prob.}} \cdot \underbrace{p(x_t|s_t, s_{t-1}, w_1^N)}_{\text{emission prob.}} \tag{6}$$

After inputting CART labels for hybrid wav2vec 2.0 training, we employed frame-wise cross-entropy (fCE) loss (Good, 1952) to train the acoustic model.

To transcribe speech given the acoustic observations, the acoustic model and n-gram language model (Ney et al., 1994) should be combined based on the Bayes decision rule using Viterbi algorithm (Forney, 1973) which recursively computes the maximum path to a find best-path in the alignment graph of all possible predicted words to the acoustic observations:

$$w_1^N = \arg\max_{N, w_1^N} p\Big(\prod_{n=1}^{N} p(w_n|w_{n-m}^{n-1}) \cdot \max_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, w_1^N)\Big) \tag{7}$$

Finally, acoustic model and n-gram language model pruning (beam search) is used to only focus on the most promising predicted words at each time step t (Ortmanns et al., 1997).

The final ASR model has 118M trainable parameters and Word-Error-Rate (WER) of 29.6% on *VietMed* test set.

C.2 Training Setup

Our encoders and encoder-decoders were trained on a cluster of 2 NVIDIA A40s with 46 GBs of memory. All models were trained on 30 epochs with with a learning rate of $2e$-5 and batch size of 64. We evaluated every epoch with early stopping with patience = 3.

For the decoder-based LLMs, due to their massive number of parameters, we use LoRA (Hu et al., 2021) for fine-tuning with hyperparameters: the rank of the update matrices $r = 8$, and the LoRA scaling factor $\alpha = 3$. We train our LLMs for 5 epochs with learning rate 2e-4.

We use the best model checkpoints for evaluation. Note that we do not perform hyperparameter tuning as we only aim to provide the initial benchmark results as well as studying the effects of CoT-augmented finetuning.

C.3 Student's T-Test

A Student's t-test, is a statistical method used to compare the means of one or two populations through hypothesis testing. It can assess whether a single group mean differs from a known value (one-sample t-test), compare the means of two independent groups (independent two-sample t-test), or determine if there is a significant difference between paired measurements (paired or dependent samples t-test). Figure 4 below is the code for reproducing Student's t-test experiments.

```python
import scipy.stats as stats

alpha = 0.05
label = [0.6628, 0.6523, 0.6592, 0.6716]
rationale = [0.6427, 0.6441, 0.6812, 0.6729]

def analyze_t_test(t_statistic, p_value, alpha):
  print(f"T-statistic: {t_statistic}")
  print(f"P-value: {p_value}")
  if p_value < alpha:
    print("The two populations are significantly different.")
  else:
    print("The two populations are not significantly different.")

t_statistic, p_value = stats.ttest_rel(label, rationale)
analyze_t_test(t_statistic, p_value, alpha)
```

Figure 4: Python code for reproducing Student's t-test experiments

D Results on English subset

We randomly sampled 50 transcripts and check their quality. We further train English models on this English subset of our dataset to ensure full usability.

The result of our experiments is in Table 8. More information on the models used can be found in the same table. Overall, we found that rationale-augmented training also help boost the model's performance. This finding is consistent with what when observed in our experiments in Section 5.

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1
Encoder (Label Only)					
mBERT (Devlin et al., 2018)	0.6001	0.5972	0.6320	0.5408	0.5900
BERT (Devlin et al., 2018)	0.6143	0.6338	0.6245	0.5653	0.6079
Encoder-Decoder (Label Only)					
mT0 (Muennighoff et al., 2022)	0.6216	0.6303	0.6418	0.5670	0.6130
Flan-T5 (Chung et al., 2022)	0.6157	0.6295	0.6385	0.5462	0.6048
Encoder-Decoder (Label + Rationale)					
mT0 (Muennighoff et al., 2022)	0.6175	0.6495	0.6253	0.5535	0.6094
Flan-T5 (Chung et al., 2022)	0.6326	0.6487	0.6390	0.5978	0.6285
Decoder (Label only)					
Mistral7B (Jiang et al., 2023b)	0.6290	0.6536	0.6322	0.5850	0.6236
Decoder (Label + Rationale)					
Mistral7B (Jiang et al., 2023b)	0.6454	0.6768	0.6364	0.6176	0.6436

Table 8: Baseline performance of encoders, encoder-decoders, LLMs on the English human transcript. Further information about our metrics can be found in Table 2.

E Results on end-to-end audio language models

E.1 Encoder-Based

PhoWhisper (Le et al., 2024): Whisper (Radford et al., 2022) trained on an 844-hour Vietnamese dataset. It achieves the state-of-the-art performance on various Vietnamese ASR datasets. To perform Sentiment Classification, we attach a classification head to the encoder layer of the model. We use the PhoWhisper base version in our experiments.

E.2 Audio LLMs

Qwen2-Audio (Chu et al., 2024): the state-of-the-art Qwen large audio-language model with 7B parameters, pretrained on diverse audio datasets. We use the Instruct version of this model. We finetune Qwen2-Audio on both *Label Only* and *Label + Rationale* settings in our experiments. It should be note that the rationales are text-only.

F Error Analysis

We report our best model's misclassified transcripts with the highest label confidence (defined as the softmax of the logits of the model prediction) in Table 9. By analyzing at the model's rationale, we hypothesize that the model is confounded by the appearance of certain keywords that elicit either extremely positive (hữu ích (helpful)) or negative, disease-related words and sentiment which pushes the model away from the *NEUTRAL* label.

Transcript	Label	Pred.	Human Rationale	Model Rationale
VI: trả lại cho họ chất lượng cuộc sống bình thường như bao người khác là được nghe được nói thế nhưng điều kỳ diệu đã **ENG:** give them back a normal quality of life like everyone else, but a miracle has happened	NEU.	POS.	Mô tả khách quan (Objective description)	chất lượng cuộc sống bình thường (normal quality of life)
VI: những chia sẻ vô cùng hữu ích và thiết thực vừa rồi ạ có thể thấy là hầu hết người bệnh nằm điều trị trong **ENG:** with the extremely useful and practical shares shared just now, it can be seen that most of the patients are in hospital for treatment)	NEU.	POS.	Mô tả khách quan (Objective description	thông tin hữu ích và thiết thực (useful and practical information)
VI: khám suốt tiểu đường nó vẫn mệt mỏi vô khám tai biến bộ não vô khám nhưng mà xương thì nó loãng xương rất là nhiều **ENG:** even after being examined for diabetes, she still feels tired, has had a stroke, and has not been examined for stroke, but her bones have a lot of osteoporosis	NEU.	NEG.	Mối quan tâm và vấn đề sức khỏe (Health concerns and problems)	triệu chứng tiêu cực của bệnh tiểu đường và loãng xương (negative symptoms of diabetes and osteoporosis)

Table 9: Some misclassified transcripts from our best model with high confidence (>0.99). VI means the Vietnamese transcript, EN means the transcript translated to English