# Unified Spoken Language Understanding from Heterogeneous Data: Spoken NER and SA as a Case Study

Anonymous ACL submission

## Abstract

Spoken Language Understanding (SLU) plays a fundamental role in enabling intelligent systems to comprehend human speech. It involves multiple tasks, such as Automatic Speech Recognition (ASR), spoken Named Entity Recognition (NER), and Sentiment Analysis (SA). Existing approaches typically address these tasks in isolation using task-specific models. Such a paradigm limits the effective utilization of heterogeneous datasets across tasks, restricts cross-task interaction, and increases overall system complexity. In this work, we propose a unified framework that jointly models different SLU tasks within a single architecture, taking spoken NER and spoken SA as a case study. Specifically, we design a unified, task-agnostic representation that facilitates the effective use of heterogeneous datasets across multiple tasks. Building upon this representation, we further propose a unified generative approach that jointly models ASR, spoken NER, and SA. Extensive experiments on public SLU datasets demonstrate that our method achieves superior SLUE scores compared to prior methods. Notably, compared with several popular LLM-based methods, our method improves the SLUE score by more than 2 points while achieving up to a 5× improvement in efficiency.

## 1 Introduction

Spoken Language Understanding (SLU) (Lakhotia et al., 2021) aims to enable machines to comprehend human speech by converting audio signals into interpretable information (Wang et al., 2024). As a core component of intelligent systems, SLU bridges the gap between natural speech and machine understanding, supporting a wide range of real-world applications (Yuan et al., 2024) in scenarios such as lectures, meetings (Du et al., 2022), and daily chats (Chiba and Higashinaka, 2025), as shown in the left and right part of Figure 1.

SLU typically involves multiple tasks. In this work, we take spoken named entity recognition (NER) (Zhou et al., 2024) and spoken sentiment analysis (SA) (Mitsui et al., 2024) as a case study, following Shon et al. (2022). The primary step of SLU tasks is Automatic Speech Recognition (ASR). As shown in Figure 1, ASR transcribes speech into text. Spoken NER identifies named entities, such as persons, organizations, and locations, while spoken SA analyzes the sentiment expressed in the speech.

Existing SLU research typically handles different SLU tasks with separate models, designing task-specific modules for each task (Xu et al., 2021; Zou et al., 2021; Dai et al., 2023). For example, in spoken NER and SA tasks, SLUE (Shon et al., 2022) embeds NER into ASR through special tagging with CTC and uses a classification module for SA. While this approach offers a solution for handling SLU tasks, it still faces several challenges. First, training separate models for different SLU tasks overlooks potential interactions across tasks that could improve overall SLU performance. Second, models are typically trained on task-specific datasets and do not fully exploit heterogeneous data from multiple sources. Finally, using distinct architectures for different SLU tasks increases the complexity of the overall model design.

To enhance interactions between multiple SLU tasks and address them in a unified manner, Kumar et al. (2024) propose a unified model capable of jointly modeling several SLU tasks, including ASR, speaker change (Cheng et al., 2023), endpointing, and NER. This model facilitates interactions between tasks, enabling them to share and enrich semantic understanding. However, their approach requires aligned datasets with annotations for all these tasks simultaneously, which are extremely rare and costly to obtain, particularly for the specific tasks we focus on, i.e., spoken NER, and SA.

To address the above limitations, we propose a unified generative framework that jointly performs diverse SLU tasks within a single model architecture, promoting cross-task interactions and enables

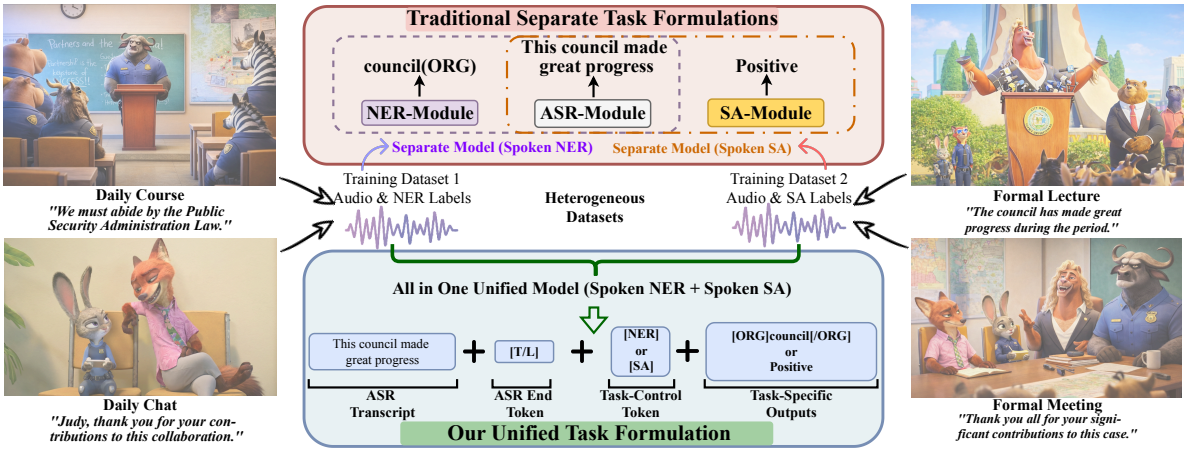


Figure 1: Comparison between traditional separate task formulations for different SLU tasks and our unified SLU formulation, using the speech input whose transcription is "This council made great progress" as an example.

effective use of heterogeneous datasets across tasks. Specifically, we first formulate different tasks into a unified formulation, enabling full utilization of heterogeneous datasets across tasks. As shown in the lower middle part of Figure 1, the unified formulation consists of four components: "ASR Transcript" which represents the transcribed speech, "ASR End Token" (i.e., "[T/L]"), which marks the end of the transcript, "Task-Control Token", which specifies the downstream understanding task (i.e., "[NER]" or "[SA]"), and "Task-Specific Outputs", which denote the corresponding task predictions. Based on this formulation, we propose a unified generative framework to jointly perform spoken NER and SA tasks, which facilitates the effective use of heterogeneous data. Extensive experiments on heterogeneous datasets demonstrate the effectiveness of our approach, resulting in overall improvements in SLU performance.

The contributions are summarized as follows:

- We propose a unified task formulation for spoken NER and SA. This facilitates the leverage of heterogeneous data across different tasks.
- We propose a generative framework that jointly models spoken NER and SA within a single architecture, enhancing multitask interaction through shared representations.
- We conduct extensive experiments on public SLU datasets and with popular LLM-based methods. Results on public datasets demonstrate superior performance and strong generalization, while comparisons with large LLM-based approaches show up to 5× higher efficiency, highlighting the practicality of our method for real-world SLU applications.

We have also provided an anonymous repository at https://anonymous.4open.science/r/ARR-January-Anonymous-Repo-2302 to further support the reproduction of this work.

## 2 Related Work

### 2.1 SLU Joint Modeling

SLU involves processing and comprehending human speech, encompassing core tasks such as ASR, NER, and SA. In this work, we target these three specific tasks. This section briefly reviews prior work on the joint modeling of ASR, NER, and SA, focusing on their integration strategies.

Joint ASR-NER modeling remains a well-studied topic in SLU. Prior research (Yadav et al., 2020; Chen et al., 2022) commonly inserts predefined tags (Ghannay et al., 2018; Yadav et al., 2020) into ASR transcripts to mark entity boundaries and types. This approach enables NER without extra modules and simplifies system design, but it can introduce ambiguity into ASR and lacks flexibility, which limits generalization to a unified framework.

Similarly, researchers have explored joint ASR-SA modeling. Earlier approaches treat ASR and SA as separate models, using ASR output as the input for SA (Bertero et al., 2016). More recent work shares the encoder between ASR and SA but employs separate decoders (Lu et al., 2020). While this allows for shared audio representations, it continues to rely on task-specific modules.

To address these, Kumar et al. (2024) proposed a unified model that supports the joint modeling of

multiple tasks. However, it suffers from the scarcity of datasets where all tasks are aligned.

Despite these, joint models still face limitations. Task interaction remains limited despite shared encoders, training is difficult due to different learning dynamics, and existing methods rely heavily on aligned data. To address these issues, we propose a unified modeling approach with a unified data formulation that better exploits heterogeneous datasets and encourages shared representations.

### 2.2 Generative Speech Language Models

Generative speech language models such as T5 (Raffel et al., 2020) and SpeechT5 (Ao et al., 2022) extend the "text-to-text" framework to speech, unifying tasks like recognition and translation. However, they struggle with SLU tasks due to the fundamental differences between speech and text.

Whisper (Radford et al., 2023), a state-of-the-art speech model, differs from T5. It is designed for transcription and translation, which limits its adaptability to SLU applications. Recent advancements like FLAN-T5 (Chung et al., 2024) aim to improve multitasking by aligning task descriptions with outputs, but these models focus primarily on text and require further adaptation for speech tasks.

In our approach, we integrate ASR, NER, and SA within a unified framework by leveraging an audio encoder for speech understanding and a decoder to generate outputs for all three tasks jointly. This unified design reduces reliance on costly aligned datasets, efficiently utilizes vast heterogeneous datasets, facilitates cross-task interaction by sharing common representations, and mitigates task imbalance to achieve superior overall performance.

## 3 Our Approach

The core concept of this work is to unify spoken NER and SA within a single generative framework. This approach aims to decrease reliance on rare and expensive multi-task aligned datasets while maximizing the utilization of extensive and accessible heterogeneous datasets. Furthermore, the framework enables models to achieve higher overall performance and improve individual task results by learning shared representations. Consequently, the method proves effective for complex and diverse real-world scenarios. In this section, we present a unified data formulation for tasks from heterogeneous data and introduce our proposed generative framework for spoken NER and SA.

### 3.1 Unified Data Formulation

Previous studies typically address spoken NER and SA separately (Shon et al., 2022) by training independent models or designing distinct modules for each task. This approach complicates model architecture and restricts the exploitation of shared characteristics across tasks. While Kumar et al. (2024) tackle multiple tasks simultaneously using an aligned dataset where each sample contains comprehensive annotations, this dataset remains private and expensive. Furthermore, it lacks open-source alternatives. Consequently, non-aligned heterogeneous datasets remain underutilized despite being more common and readily available.

To leverage these heterogeneous datasets effectively, we propose a unified data formulation following the template: "[ASR Transcript][ASR End Token][Task-Control Token][Task-Specific Outputs]". Here, "[ASR Transcript]" denotes the recognized text. "[ASR End Token]" (represented as [T/L]") serves as a delimiter separating the transcript from subsequent task results. "[Task-Control Token]" indicates the task name (specifically "[NER]" or "[SA]"), and "[Task-Specific Outputs]" signifies the corresponding task result.

The "Task Formulation" part of Figure 1 illustrates examples in both the traditional separate format and our proposed unified format. The separate format clearly generates three independent sequences, whereas the unified format executes multiple tasks directly within a single sequence. For instance, the system represents the sentence for the NER task as "This council made great progress[T/L][NER][ORG]council[/ORG]", and for the SA task as "This council made great progress[T/L][SA]Positive".

### 3.2 Unified Generative Framework

We propose a unified generative framework to process ASR, NER, and SA jointly. Figure 1 details the complete architecture and the inference process. Specifically, the audio feature extractor processes raw audio and passes it to the audio encoder to yield deep phonetic representations. If the subsequent decoder integrates natively with the encoder, it receives these representations directly. Otherwise, an adapter transforms the hidden dimensions to ensure compatibility.

During inference, the model first generates the speech transcript. The appearance of the ASR end token "[T/L]" signals the beginning of the

understanding tasks. At this stage, we insert a task-control token to specify the subsequent task. Inserting "[NER]" guides the model to identify entities and their types, such as persons, locations, or organizations. Inserting "[SA]" directs the model to output the sentiment of the speech. These control tokens lead the decoding process and ensure the generation of coherent, context-aware outputs based on the preceding transcriptions.

This unified framework enables the joint training of spoken NER and SA using widespread heterogeneous datasets. It encourages tasks to share common representations and maximizes the utility of causal reasoning within autoregressive generation processes. The following sections detail the architecture and training strategy.

**Model Architecture.** The model accepts a raw speech waveform $\mathbf{X}$ as input. This input is first converted into a log-mel spectrogram and then passed to the encoder-decoder for final autoregressive generation. The entire process proceeds as follows:

$$\mathbf{H} = \text{Encoder}_{Audio}(\text{Mel}(\mathbf{X})),\ \mathbf{H} \in \mathbf{R}^{l \times d} \quad (1)$$

$$\mathbf{Y} = \text{Decoder}_{Language}(\mathbf{H}, \mathbf{Y}_{<t}) \quad (2)$$

where $l$ denotes the downsampled sequence length, $d$ represents the encoder hidden dimension, and $\mathbf{Y}$ signifies the output token sequence, conditioned on the previously generated tokens $\mathbf{Y}_{<t}$.

**Dynamic Weighted Training Loss.** In multitask fine-tuning, balancing weights between different tasks proves crucial. Since ASR transcripts typically exceed the length of NER and SA outputs, the model tends to overemphasize ASR and neglect the understanding tasks. To address this issue, we design a dynamic loss function based on the ratio of each task's sequence length to the total sequence. Specifically, longer ASR outputs receive lower weights as they already benefit from substantial optimization during pre-training. Conversely, shorter NER and SA outputs receive correspondingly higher weights. This strategy ensures task balance during training. We define the total loss function as follows:

$$\mathcal{L}_{total} = W_{\text{task}}\mathcal{L}_{\text{task}} + W_{\text{ASR}}\mathcal{L}_{\text{ASR}} \quad (3)$$

where $\mathcal{L}_{\text{ASR}}$ denotes the ASR loss function, and $\mathcal{L}_{\text{task}}$ represents the loss for NER or SA. We employ cross-entropy as the base loss function and compute the corresponding weights $W_{\text{task}}$ and $W_{\text{ASR}}$ as follows:

$$W_{\text{task}} = \frac{\text{Len}_{\text{ASR}}}{\text{Len}_{\text{total}}} \quad W_{\text{ASR}} = \frac{\text{Len}_{\text{task}}}{\text{Len}_{\text{total}}} \quad (4)$$

where $\text{Len}_{\text{ASR}}$ represents the length of the transcription, $\text{Len}_{\text{task}}$ denotes the length of the task output, and $\text{Len}_{\text{total}}$ indicates the total length. All length measurements occur at the token level.

**Training Process.** During training, we acknowledge that the audio in our heterogeneous datasets originates from diverse sources, resulting in significant variations in audio features. To accommodate these differences, we fully fine-tune both the encoder and decoder. This approach enables the model to adapt effectively to diverse audio characteristics, thereby enhancing overall performance.

## 4 Experiments

**Datasets.** In line with recent studies (Denisov and Vu, 2023; Ayache et al., 2024), we base our primary experiments on SLUE-VoxPopuli (VP) and SLUE-VoxCeleb (VC) (Shon et al., 2022). These heterogeneous, open-source English datasets [1] facilitate spoken NER and SA tasks, originating from European Parliament recordings and YouTube conversations respectively. SLUE-VoxPopuli covers entity categories such as PLACE, QUANT, ORG, WHEN, NORP, PERSON, and LAW, while SLUE-VoxCeleb features positive, negative, and neutral sentiment labels. Table 1 presents all the statistics.

| Datasets | | Train | Dev | Test | Tasks |
|---|---|---|---|---|---|
| EN-SLUE VoxPopuli | #Sent | 5,000 | 1,753 | 1,842 | ASR NER |
| | #NE | 5,730 | 1,819 | 1,961 | |
| | Dur | 14.5h | 5.0h | 4.9h | |
| EN-SLUE VoxCeleb | #Sent | 5,777 | 1,454 | 3,553 | ASR SA |
| | Dur | 12.8h | 3.2h | 7.8h | |

Table 1: The dataset statistics include sentence counts (#Sent), entity counts (#NE), and audio duration (Dur).

**Implementation Details.** Our main experiments utilize whisper-medium [2] as the base model, which comprises a 24-layer Transformer block for both the encoder and decoder. We fine-tune the model for 100 epochs with a learning rate of 1e-5 and a batch size of 32. We repeat each experiment three times using different random seeds and report

[1] https://huggingface.co/datasets/asapp/slue
[2] https://huggingface.co/openai/whisper-medium

| Model | ASR-WER VP | ASR-WER VC | NER-F1 VP | SA-F1 VC | SLUE SCORE |
|---|---|---|---|---|---|
| Speech Foundation Models + Lightweight Prediction Head (Separate Models) | | | | | |
| SLUE (W2V2-B-LS960)† | 18.40 | 20.90 | 49.60 | 48.60 | 59.40 |
| SLUE (W2V2-L-LL60K)† | 12.10 | 13.80 | 50.50 | 50.10 | 62.55 |
| SLUR-PERB (HuBERT-large)* | 14.20 | 19.00 | 59.30 | 41.00 | 61.23 |
| SLUR-PERB (Wav2Vec2-large)* | 16.00 | 21.70 | 57.50 | 40.60 | 59.75 |
| SLUR-PERB (Whisper-medium)* | 12.50 | 15.00 | 63.10 | 49.60 | **66.31** |
| Speech Foundation Models + Complex Prediction Head (Separate Models) | | | | | |
| SLUR-PERB (HuBERT-large)* | 13.00 | 15.50 | 63.10 | 52.20 | 67.01 |
| SLUR-PERB (Wav2Vec2-large)* | 14.00 | 17.20 | 63.70 | 53.30 | 67.13 |
| SLUR-PERB (Whisper-medium)* | 13.20 | 14.90 | 64.10 | 51.00 | 67.01 |
| Ours (Whisper-medium) | 8.63 | 10.92 | 63.83 | 60.92 | **71.65** |
| Speech Foundation Models + Complex Prediction Head (Unified Models) | | | | | |
| Ours (Whisper-medium) | 8.77 | 11.47 | 69.95 | 61.60 | **73.81** |
| w/o Dynamic Loss | 8.97 | 10.93 | 68.69 | 61.74 | 73.49 |
| w/o Fine-tuning Encoder & Dynamic Loss | 8.08 | 11.05 | 65.18 | 61.22 | 72.27 |

Table 2: Performance of our proposed methods and other compared models across ASR, NER, and SA tasks, along with the overall SLUE SCORE on the test set. VP and VC denote the SLUE-VoxPopuli and SLUE-VoxCeleb datasets, respectively. Models marked with † and ∗ indicate the results in Shon et al. (2022) and Arora et al. (2024).

the averaged results. All of the models employ the Whisper tokenizer, and the complete parameter settings are provided in Appendix A.3 Table 9.

**Evaluation Metrics.** For ASR, we employ Word Error Rate (WER) for English data. For NER, consistent with prior studies (Ghannay et al., 2018; Yadav et al., 2020), we use Micro-F1 to measure performance. In the case of SA, we follow Bertero et al. (2016); Lu et al. (2020); Liu et al. (2022) and utilize Macro-F1. Additionally, we apply the SLUE SCORE metric (Shon et al., 2022) to evaluate overall performance on the two English datasets, defined as follows: $SLUE_{score} = \frac{1}{3}((100 - \frac{WER_{VP}+WER_{VC}}{2}) + F1_{VP} + F1_{VC})$ where $\mathrm{WER_{VP}}$ and $\mathrm{WER_{VC}}$ represent the WER score of the VP and VC datasets, $\mathrm{F1_{VP}}$ denotes the Micro-F1 score of the NER performance on VP, and $\mathrm{F1_{VC}}$ is the Macro-F1 score of SA on VC.

### 4.1 Main Results

Table 2 presents the main results on the SLUE-VoxPopuli and SLUE-VoxCeleb test sets.

**Results of Separate Models.** To explore the influence of different speech foundation models, task heads, and formulations, we analyze the first two major rows of separate models. The first is "Speech Foundation Models + Lightweight Prediction Head". It covers architectures that pair a speech foundation model with a lightweight prediction head. Examples include CTC for ASR, special tagging embedded in ASR for NER, and a linear classifier for SA. The second is "Speech Foundation Models + Complex Prediction Head". It distinguishes itself by utilizing a complex Transformer encoder-decoder architecture to execute tasks via generative methods. For the three "SLUR-PERB (*)" models, NER follows the formulation used in the lightweight prediction head, while SA generates the sentiment label at the beginning of the ASR transcript. For our baseline separate model, it first generates the ASR transcript and then performs NER or SA after the transcript.

First, models utilizing a complex prediction head consistently outperform those with a lightweight prediction head. This suggests that tasks require deep semantic understanding through autoregressive generation rather than simple label mapping. Furthermore, regarding speech foundation models, whisper-medium yields the best performance regardless of the prediction head type. Consequently, we select whisper-medium as the base for our unified model experiments. Second, we examine task formulations in "Speech Foundation Models + Complex Prediction Head (Separate Models)". The "SLUR-PERB (*)" models employ special tagging

| Model | ASR-WER VP | ASR-WER VC | NER-F1 VP | SA-F1 VC | SLUE SCORE |
|---|---|---|---|---|---|
| Speech Foundation Models + Large Language Models (Pipeline Models) | | | | | |
| Whisper-medium + Qwen2.5-3B (Zero-Shot) | 7.70 | 13.33 | 6.35 | 58.03 | 51.28 |
| Whisper-medium + Qwen2.5-7B (Zero-Shot) | 7.70 | 13.33 | 16.05 | 59.57 | 55.03 |
| Whisper-medium + Qwen2.5-3B | 7.70 | 13.33 | 53.71 | 66.26 | 69.81 |
| Whisper-medium + Qwen2.5-7B | 7.70 | 13.33 | 55.16 | 68.65 | **71.09** |
| Speech Foundation Models + Large Language Models (E2E Models) | | | | | |
| W2V-BERT-2.0 + Qwen2.5-3B | 16.14 | 23.24 | 60.24 | 60.73 | 67.09 |
| W2V-BERT-2.0 + Qwen2.5-7B | 16.40 | 22.94 | 61.02 | 62.80 | 68.05 |
| Whisper-medium + Qwen2.5-3B | 12.75 | 17.44 | 62.61 | 63.45 | 70.32 |
| Whisper-medium + Qwen2.5-7B | 11.63 | 14.16 | 64.05 | 63.27 | **71.47** |
| Speech Foundation Models + Complex Prediction Head (Unified Models) | | | | | |
| Ours (Whisper-medium) | 8.77 | 11.47 | 69.95 | 61.60 | **73.81** |

Table 3: Comparison of performance metrics for LLMs across ASR, NER , and SA on the test set.

for NER and prefix the SA label to the transcript. Conversely, our baseline avoids special tagging and appends task labels after the transcript. Our approach outperforms these baselines and validates our formulation. While NER performance remains similar, the special-tagging method degrades ASR performance by increasing WER. Regarding SA, positioning the label after the transcript enables the model to leverage the full context for better predictions.

Overall, our proposed formulations combining with generative method achieve great improvements. Thus, our unified experiments are on whisper-medium combined with our proposed methods in the following paragraph.

**Results of Unified Models.** Following the analysis above, we examine our proposed unified generative methods and compare them against separate models to determine if unifying spoken NER and SA yields mutual benefits.

Our unified models consistently outperform the separate models, confirming the effectiveness of combining tasks with heterogeneous datasets. These tasks mutually reinforce each other, leading to superior overall performance and improved individual results for both NER and SA. We attribute this success to two primary factors. First, the combination of datasets provides a larger training corpus with a broader range of entities and more complex sentiment variations than treating datasets independently. This also increases the complexity and diversity of the training data, introduces noise to some extent, reduces overfitting on a single task, and improves generalization. Second, All tasks share the parameters of the entire model, which strengthens the common representations between tasks and greatly enhances the interactivity.

Overall, these findings demonstrate that unifying tasks from heterogeneous datasets via our proposed formulation yields benefits. Furthermore, the results on Chinese datasets maintain the same trend, further confirming this conclusion. Detailed data can be found in Appendix A.1.

### 4.2 Comparison with LLM-based Methods

Here, we compare the models against LLM-based methods. For pipelines, an ASR model first transcribes speech into text, and the LLM processes this output for NER and SA tasks. The E2E models, which comprises an audio encoder, an adapter, and an LLM backbone, execute spoken NER and SA directly. Specifically, the pipeline configurations utilize whisper-medium paired with either Qwen2.5-3B-Instruct[3] (Yang et al., 2024) or Qwen2.5-7B-Instruct[4], covering both zero-shot and fine-tuned versions of the LLM. For the E2E models, we integrate W2V-BERT-2.0[5] and whisper-medium with Qwen2.5-3b-Instruct and Qwen2.5-7B-Instruct using an adapter. We first pre-train them for ASR on LibriSpeech-960 (LS960)[6] and then fine-tune them for spoken NER and SA. Table 3 presents the

[3]https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[5]https://huggingface.co/facebook/w2v-bert-2.0
[6]https://www.openslr.org/12/

| Model | Throughput ↑ | RTF ↓ | Latency ↓ | TFLOPS ↓ | SLUE SCORE ↑ |
|---|---|---|---|---|---|
| Speech Foundation Models + Large Language Models (Pipeline Models) | | | | | |
| Whisper-medium + Qwen2.5-3B | 2.29 | 0.0515 | 437.43 | 6.17 | 69.81 |
| Whisper-medium + Qwen2.5-7B | 2.32 | 0.0506 | 430.31 | 11.35 | 71.09 |
| Speech Foundation Models + Large Language Models (E2E Models) | | | | | |
| W2V-BERT-2.0 + Qwen2.5-3B | 1.13 | 0.1039 | 883.01 | 4.03 | 67.09 |
| W2V-BERT-2.0 + Qwen2.5-7B | 1.38 | 0.0854 | 726.23 | 7.49 | 68.05 |
| Whisper-medium + Qwen2.5-3B | 3.77 | 0.0312 | 265.03 | 4.26 | 70.32 |
| Whisper-medium + Qwen2.5-7B | 3.89 | 0.0302 | 256.93 | 8.01 | 71.47 |
| Speech Foundation Models + Complex Prediction Head (Unified Models) | | | | | |
| Ours (Whisper-medium) | **19.35** | **0.0052** | **51.59** | **3.55** | **73.81** |

Table 4: Efficiency performance of our proposed model and the LLM-based models.

results, and our detailed analysis follows.

First, regarding the "Speech Foundation Models + Large Language Models (Pipeline / E2E)" row, pipeline models achieve superior ASR performance because the speech foundation model focuses solely on transcription without interference from simultaneous understanding tasks. However, pipeline models excel at SA, while E2E models perform better on NER. We attribute this difference to the combined effects of task difficulty and the autoregressive generation mechanism. The pipeline approach generates two independent sequences, whereas E2E models output a single unified sequence. NER requires complex structured outputs with numerous label types, thus benefiting from the deep semantic understanding facilitated by the causal reasoning inherent in the autoregressive process during the generation of ASR. In contrast, SA is simpler and for samples where sentiment is evident, directly predicting proves more effective.

Second, our method outperforms all LLM-based models in overall performance despite having significantly fewer parameters. Specifically, compared to pipeline LLM-based models, our unified model achieves more balanced ASR performance and exhibits NER and SA trends similar to those discussed in the first point. Compared to E2E LLM-based models, our method achieves significantly better ASR and NER performance. We attribute this to the difficulty LLM backbones face in maintaining basic ASR performance. This deficiency further impacts NER performance. In contrast, although our model's fewer parameters lead to weaker SA performance, our proposed method requires no additional ASR pre-training and maximizes the utility of heterogeneous datasets to achieve the best overall results.

Overall, the superior performance further demonstrate the effectiveness of our proposed methods.

### 4.3 Efficiency

In real-world systems, efficiency is a key concern. Thus, we evaluate the efficiency of our methods and LLM-based methods on four metrics: Throughput, Real-Time Factor (RTF), Latency, and TFLOPS. All inferences use the same settings and are conducted on a single A100-40G GPU with a batch size of 32. Detailed results are presented in Table 4.

Our model demonstrates significantly higher efficiency compared to LLM-based methods. Specifically, we compare our approach with the end-to-end model "whisper-medium + Qwen2.5-7B," which achieves the best performance among LLM-based approaches. Our method delivers an overall performance improvement of approximately 2 points and runs about 5× faster, while consuming only half of the computational resources.

Overall, our method achieves a better performance with great efficiency which further demonstrates the feasibility in real-world scenarios.

### 4.4 Analysis Regarding Different Lengths

Here, we further analyze the relationship between performance and output length. Figure 2 evaluates the NER-F1, SA-F1, and the overall SLUE-SCORE corresponding to plots (a), (b), and (c) respectively, on different lengths measured by words.

First, comparing models with and without the two training strategies, the former achieves significantly ($p < 0.0001$) better NER performance. We attribute this to two factors. On one hand, the dynamic loss balances the optimization gap caused by

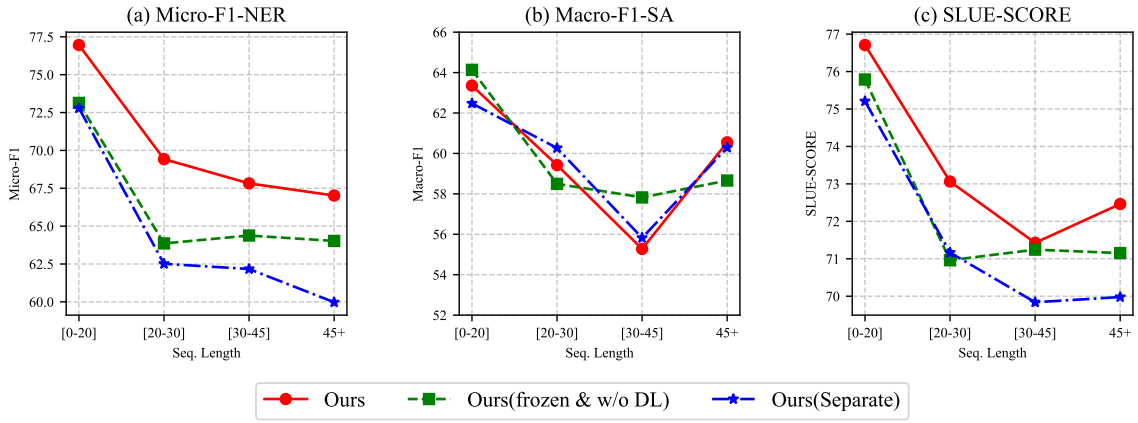


Figure 2: Comparison of Micro-F1, Macro-F1 and SLUE SCORE metrics across different sequence length categories on test set. The sequence lengths on the horizontal axis are computed on the ASR transcripts and measured in words.

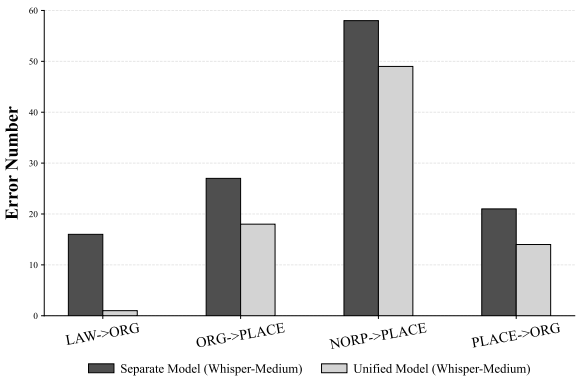


Figure 3: Different NER error patterns.

different output lengths in multi-task training. On another hand, activating the encoder captures common characteristics across heterogeneous datasets with distinct domains and audio features.

Second, compared to the baseline separate model, our unified models perform better on longer sequences, even without the strategies. We attribute this to the complementary data length distributions of the NER and SA datasets (averaging 24 and 27 words respectively) and help alleviate data sparsity.

Overall, these indicate that our unified methods with the two training strategies work effectively.

### 4.5 Analysis on NER Error Patterns

To analyze improvements, we compare NER error patterns between the unified model and the separate baseline. The notation "X→Y" denotes an entity labeled X misclassified as Y. Figure 3 shows the four error categories with the largest changes.

We observe that most errors arise from misclassifications as ORG or PLACE. First, incorrect ORG predictions mainly stem from nested entity structures. For example, LAW entities often combine an ORG with a legal article, and some ORG entities include PLACE names. Boundary detection errors therefore lead the model to predict only sub-entities or to hallucinate complete entities. Second, incorrect PLACE predictions are largely due to semantic overlap across entity types caused by limited use of contextual information. Without sufficient context, a country name may be interpreted as PLACE, NORP, or ORG. Our method leverages heterogeneous data to cover a broader range of such cases, and the unified task formulation enables downstream understanding based on the complete ASR transcript. This improves the model's ability to distinguish entities that exhibit semantic overlap when contextual information is lacking and leads to clear gains on these ambiguous error types, further demonstrating the effectiveness of our approach.

## 5 Conclusion

In this work, we propose a novel unified data formulation and generative framework that integrates spoken NER and SA within a single model architecture. This approach increases the utilization of heterogeneous datasets and facilitates rich, effective cross-task interactions through shared representations. We conduct extensive experiments on multilingual public datasets, where our methods consistently outperform strong baselines, demonstrating the effectiveness and robustness of our approach. Furthermore, we compare our method with LLM-based methods; our approach achieves not only higher performance but also greater efficiency, proving its viability for real-world scenarios.

## Limitations

Our current work faces a primary limitation. Specifically, while we consider spoken NER and SA to be the most critical SLU tasks and use them as case studies in this work, many other SLU tasks still need improvement. Consequently, in future work, we will explore more other SLU tasks and further extend our method.

## Ethics Statement

This work adheres to established ethical standards. All datasets are obtained and used in compliance with relevant usage guidelines, with no violations of privacy. We make consistent efforts to avoid bias or discriminatory outcomes throughout the research process. No personally identifiable information is involved. Elements in Figure 1 drawn from selected film or television works are used solely for academic presentation in the paper, not for any commercial or non-academic purposes, and strictly comply with fair use provisions under copyright law. All large language models are used only for text or image refinement and do not participate in the methods, code, experiments, or conclusions. No experiments raise privacy or security concerns. We remain committed to transparency and integrity throughout the research process.

## References


Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5723–5738. Association for Computational Linguistics.

Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan S. Sharma, Jee-weon Jung, Hira Dhamyal, William Chen, Suwon Shon, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2024. On the evaluation of speech foundation models for spoken language understanding. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11923–11938. Association for Computational Linguistics.

Gil Ayache, Menachem Pirchi, Aviv Navon, Aviv Shamsian, Gill Hetz, and Joseph Keshet. 2024. Whisperner: Unified open named entity and speech recognition. *Arxiv*, abs/2409.08107.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1042–1047. The Association for Computational Linguistics.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. AISHELL-NER: named entity recognition from chinese speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8352–8356. IEEE.

Luyao Cheng, Siqi Zheng, Qinglin Zhang, Hui Wang, Yafeng Chen, and Qian Chen. 2023. Exploring speaker-related information in spoken language understanding for better speaker diarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14068–14077. Association for Computational Linguistics.

Yuya Chiba and Ryuichiro Higashinaka. 2025. Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3687–3696, Abu Dhabi, UAE. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Chenquan Dai, Xiaobin Zhuang, and Jiaxin Cai. 2023. A survey on deep learning for chinese medical named entity recognition. In *Proceedings of the 9th International Conference on Computing and Artificial Intelligence, ICCAI 2023, Tianjin, China, March 17-20, 2023*, pages 472–476. ACM.

Pavel Denisov and Ngoc Thang Vu. 2023. Leveraging multilingual self-supervised pretrained models for sequence-to-sequence end-to-end spoken language understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Zhihao Du, Shiliang Zhang, Siqi Zheng, and Zhi-Jie Yan. 2022. Speaker overlap-aware neural diarization

for multi-party meeting analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7458–7469. Association for Computational Linguistics.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, pages 692–699. IEEE.

Shashi Kumar, Srikanth R. Madikeri, Juan Pablo Zuluaga-Gomez, Iuliia Thorbecke, Esaú Villatoro-Tello, Sergio Burdisso, Petr Motlícek, Karthik S, and Aravind Ganapathiraju. 2024. Tokenverse: Towards unifying speech and NLP tasks via transducer-based ASR. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 20988–20995. Association for Computational Linguistics.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and av-mixup consistent module. In *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*, pages 247–258. ACM.

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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7149–7153. IEEE.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. 2024. PSLM: parallel generation of text and speech with llms for low-latency spoken dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2692–2700. Association for Computational Linguistics.

Jinzhong Ning, Yuanyuan Sun, Bo Xu, Zhihao Yang, Ling Luo, and Hongfei Lin. 2024. Breaking the boundaries: A unified framework for chinese named entity recognition across text and speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1250–1260. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu Jeong Han. 2022. SLUE: new benchmark tasks for spoken language understanding evaluation on natural speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7927–7931. IEEE.

Bin Wang, Meishan Zhang, Hao Fei, Yu Zhao, Bobo Li, Shengqiong Wu, Wei Ji, and Min Zhang. 2024. Speechee: A novel benchmark for speech event extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 10449–10458. ACM.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14176–14184. AAAI Press.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 4268–4272. ISCA.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ziqi Yuan, Baozheng Zhang, Hua Xu, Zhiyun Liang, and Kai Gao. 2024. OpenVNA: A framework for analyzing the behavior of multimodal language understanding system under noisy scenarios. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 9–18, Bangkok, Thailand. Association for Computational Linguistics.

Shilin Zhou, Zhenghua Li, Chen Gong, Lei Zhang, Yu Hong, and Min Zhang. 2024. Chinese spoken named entity recognition in real-world scenarios:

Dataset and approaches. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1872–1884. Association for Computational Linguistics.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14665–14673. AAAI Press.

| **Datasets** | | **Train** | **Dev** | **Test** | **Tasks** |
|---|---|---|---|---|---|
| CH-CNERTA | #Sent | 34,102 | 4,440 | 4,445 | ASR NER |
| | #NE | 23,805 | 5,889 | 7,263 | |
| | Dur | 56.7h | 7.5h | 7.6h | |
| CH-SIMS v2.0 | #Sent | 2,722 | 647 | 1,034 | ASR SA |
| | Dur | 2.4h | 0.4h | 1.2h | |

Table 5: The dataset statistics include sentence counts (#Sent), the number of entities (#NE), and audio duration (Dur).

| **Model** | **NER F1** | **SA ACC** |
|---|---|---|
| Conformer† | 60.36 | - |
| Wav2Vec 2.0* | - | 65.38 |
| Whisper-medium | 59.63 | 69.05 |
| Ours (Whisper-medium) | **60.62** | **70.01** |

Table 6: Performance across NER and SA On CNERTA and CH-SIMS v2.0. Models marked with † and $*$ are from Ning et al. (2024) and Liu et al. (2022).

# A Appendix

## A.1 Results on Chinese Datasets

To further explore the generalization capabilities of our proposed method, we evaluate it on two Chinese datasets CNERTA and CH-SIMS v2.0 for spoken NER and SA. CNERTA v2.0, annotated based on Aishell-1, features Mandarin pronunciation across different regional accents and spans five distinct domains. CH-SIMS, derived from in-the-wild videos on platforms such as YouTube and Bilibili, presents more chaotic and diverse data covering eleven different scenarios. Both datasets offer challenging conditions suitable for rigorous evaluation. Table 5 details the dataset statistics, and Table 6 presents the evaluation results.

Comparisons with other models reveal that our method achieves improvements similar to those observed in the English datasets. Furthermore, the comparison between our separate and unified models shows stable gains across both tasks without imbalance. This result persists even though the Chinese datasets present greater difficulty and differ significantly in data volume. These findings collectively demonstrate the generalization capabilities of our proposed method.

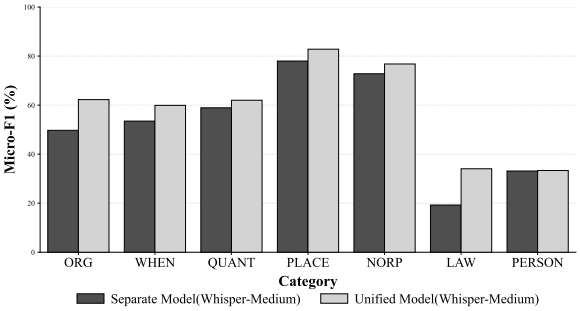


Figure 4: NER-Micro-F1 per entity type on test set.

| PLA | QUANT | ORG | WHEN | NORP | PER | LAW |
|---|---|---|---|---|---|---|
| 82.81 | 61.99 | 62.24 | 59.90 | 76.78 | 33.33 | 34.04 |

Table 7: NER-Micro-F1 per entity type on test set.

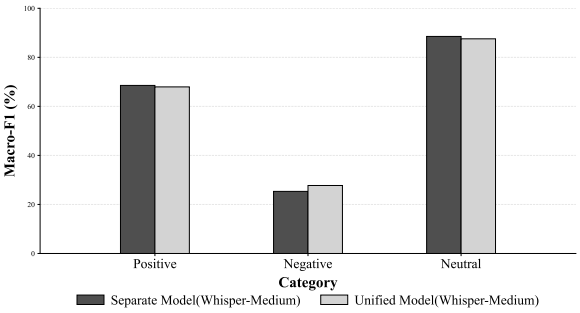


Figure 5: SA-Macro-F1 per sentiment class on test set.

| Positive | Negative | Neutral |
|---|---|---|
| 67.94 | 27.73 | 87.55 |

Table 8: SA-Macro-F1 per sentiment class on test set.

### A.2 Analysis on Each Label Type

We analyze the performance of each entity and sentiment type on our unified model. And metrics still follow the main experiment, for NER with Micro-F1 and SA Macro-F1. Table 7, Table 8 provide the results and Figure 4, Figure 5 detail the comparison between unified and separate methods. As shown in the two tables, for NER, the LAW and PERSON categories perform worse than others. For SA, the Negative sentiment performs the worst among the three sentiment classes. We attribute this primarily to label imbalance in the original training data, where LAW, PERSON, and Negative make up only 4.5%, 4.9%, and 2.1% of the datasets. Moreover, the two heterogeneous datasets originate from formal and daily scenarios. Consequently, LAW and PER entities are less common in the daily scenario dataset, while the Neutral sentiment is common in the formal scenario dataset. This makes it difficult to supplement the scarce examples.

### A.3 Detailed Parameter Settings

| **Whisper-medium** | |
|---|---|
| Encoder Blocks | 24 |
| Decoder Blocks | 24 |
| Hidden Size | 1024 |
| Fine-tune Epoch | 100 |
| Learning Rate | 2e-5 |
| Warm Ratio | 0.01 |
| Log Mel | 80 |
| Window Size | 25ms |
| Window Shift | 10ms |
| Num-beams | 1 |
| **Qwen2.5-3B / 7B-Instruct** | |
| Pretrain Epoch | 2 |
| Fine-tune Epoch | 20 |
| Learning Rate | 2e-4 |
| Lora-r | 32 |
| Lora-alpha | 32 |
| Lora-dropout | 0.05 |
| Temperature | 0.7 |
| Top-p | 0.9 |
| Repetition-penalty | 1.2 |
| Num-beams | 1 |

Table 9: Hyper-parameters and model details.