MAC-SLU: MULTI-INTENT AUTOMOTIVE CABIN SPOKEN LANGUAGE UNDERSTANDING BENCHMARK

Yuezhang Peng[1], *Chonghao Cai*[1], *Ziang Liu*[2], *Shuai Fan*[3], *Sheng Jiang*[3], *Hua Xu*[3], *Yuxin Liu*[1], *Qiguang Chen*[2], *Kele Xu*[4], *Yao Li*[5], *Sheng Wang*[1,5], *Libo Qin*[2,*], *Xie Chen*[1,*]

[1] School of Computer Science, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2]School of Computer Science and Engineering, Central South University
[3]AISpeech Co., Ltd. [4] National University of Defense Technology, [5]Shanghai Aviation Electric Co., Ltd

ABSTRACT

Spoken Language Understanding (SLU), which aims to extract user semantics to execute downstream tasks, is a crucial component of task-oriented dialog systems. Existing SLU datasets generally lack sufficient diversity and complexity, and there is an absence of a unified benchmark for the latest Large Language Models (LLMs) and Large Audio Language Models (LALMs). This work introduces MAC-SLU, a novel Multi-Intent Automotive Cabin Spoken Language Understanding Dataset, which increases the difficulty of the SLU task by incorporating authentic and complex multi-intent data. Based on MAC-SLU, we conducted a comprehensive benchmark of leading open-source LLMs and LALMs, covering methods like in-context learning, supervised fine-tuning (SFT), and end-to-end (E2E) and pipeline paradigms. Our experiments show that while LLMs and LALMs have the potential to complete SLU tasks through in-context learning, their performance still lags significantly behind SFT. Meanwhile, E2E LALMs demonstrate performance comparable to pipeline approaches and effectively avoid error propagation from speech recognition. Code[1] and datasets[2] are released publicly.

Index Terms— Spoken Language Understanding, Large Language Models, Large Audio Language Models

1. INTRODUCTION

Spoken Language Understanding (SLU) is a conventional paradigm for task-oriented spoken semantics extraction, which has been widely applied in scenarios such as smart homes and automobiles to extract spoken commands from users for executing downstream tasks [1]. Traditional SLU employs a pipeline approach, first transcribing the user's spoken query into text via Automatic Speech Recognition (ASR), and then extracting semantic information through Natural Language Understanding (NLU), which typically includes Intent Classification (IC) and Slot Filling (SF) [2]. Subsequently, E2E SLU systems emerged to address the issue of error propagation from the ASR transcription process [3] and potentially incorporate pre-trained language models (e.g., RoBERTa [4]) to enhance performance [5]. Presently, the advancement of LLMs and LALMs offers the potential for a more flexible and precise SLU.

Nevertheless, research on LLM-based SLU still confronts the following two challenges: (1) Existing SLU datasets lack sufficient diversity and complexity. The widely used ATIS [6] and SNIPS [7] datasets contain only 16 and 7 intents, respectively, which limits the task's difficulty, resulting in existing models already achieving over 95% accuracy in both IC and SF [8]. SLURP dataset [9] increases the number of intent and slot categories but remains confined to single-intent SLU tasks. (2) There is a lack of a unified benchmark for state-of-the-art (SOTA) open-source LLMs and LALMs. Although some studies have preliminarily explored the performance of LLMs like ChatGPT and Llama on certain SLU tasks [10, 11], they have adopted different task formats (i.e., varying data formats, prompts, or training and alignment methods), leading to evaluation results that cannot be fairly compared. Furthermore, existing research has been limited to pipeline-based methods and LLMs, without exploring E2E approaches and the most advanced LALMs.

This work addresses these challenges with two steps. First, we introduce the **Multi-Intent Automotive Cabin Spoken Language Understanding** (**MAC-SLU**) dataset, a novel Chinese SLU corpus to overcome the complexity limitations of existing data. Derived from real-world automotive text commands with TTS-synthesized speech, it spans 8 domains, 81 intents, 192 slots, and includes multi-intent queries with up to 5 intents, creating a more rigorous testbed. Second, we establish a unified benchmark on MAC-SLU for SOTA open-source LLMs and LALMs. Standardizing formats, tasks, and evaluation methods enables fair comparisons and provides a dependable reference for the community. Our contributions include:

- We introduce MAC-SLU, a novel Chinese multi-intent SLU dataset including complex, multi-intent queries from a real-world automotive cabin domain. MAC-SLU enables a more chanllenging evaluation of the latest LLMs and LALMs.
- We provide a comprehensive benchmark for SOTA open-source and closed-source LLMs and LALMs, encompassing methods based on direct inference, in-context learning, and SFT, as well as both pipeline and E2E SLU task paradigms.
- Our experiments demonstrate that (1) existing LLMs and LALMs can complete parts of the IC or SF tasks through in-context learning, yet there remains a significant performance gap compared to in-domain SFT; (2) benefiting from the avoidance of error propagation, current LALMs can already achieve performance comparable to pipeline methods.

2. RELATED WORK

2.1. SLU Datasets

Existing SLU datasets primarily consist of single-intent datasets such as ATIS, SNIPS, FSC, and SLURP [6, 7, 3, 9], as well as

* Corresponding authors.

[1] `https://github.com/Gatsby-web/MAC_SLU`

[2] `huggingface.co/datasets/Gatsby1984/MAC_SLU`

Domain	**Intent**	**Slot**
Car Control	Car System Control, Car Body Control	Object, Location, Function, Application
Map	Navigation, Provide Address, Check Traffic	Destination Name, Destination Type, Waypoint Name
Music	Play Music, Query Music Information	Music Name, Music Type, Artist Name

Table 1. Examples of domains, intents, and slots in the MAC-SLU dataset. All data shown are English translations of the original Chinese.

Dataset	**Domain**	**Intent**	**Slot**	**Multi-intent**
ATIS [6]	1	16	41	✗
SNIPS [7]	2	7	4	✗
FSC [3]	2	6	2	✗
SLURP [9]	**18**	46	56	✗
MixATIS [12]	1	16	41	✓
MixSNIPS [12]	2	7	4	✓
MAC-SLU	8	**81**	**192**	✓

Table 2. Comparison of SLU datasets. MAC-SLU supports the largest intent/slot categories and complex multi-intent data.

multi-intent datasets like MixATIS and MixSNIPS [12]. The ATIS dataset, released in 1990, mainly contains voice queries for flight information and includes only 16 intent categories. More recent datasets like SNIPS and FSC are designed for smart home scenarios but also have a limited number of intent classes. Another commonly used large-scale SLU dataset is SLURP, which expands the number of intent categories to 46, significantly increasing the task's complexity. Nevertheless, SLURP does not contain multi-intent data, which limits its diversity. The MixATIS and MixSNIPS multi-intent datasets were constructed from the ATIS and SNIPS datasets, respectively, by connecting sentences with different intents using conjunctions (e.g.,"and"). However, constrained by the relative simplicity of the original ATIS and SNIPS datasets, the intent diversity in MixATIS and MixSNIPS is also limited. Furthermore, E2E Chinese SLU datasets are scarce, and the closest alternatives, CAIS [13] and ECDT-NLU[3], are text-based datasets for NLU tasks.

2.2. LLMs for SLU

Several studies have attempted to apply LLMs to SLU tasks. He [10] explored using in-context learning to enable ChatGPT to perform IC and SF, demonstrating that LLMs can achieve commendable performance in IC but still face challenges with the more complexly defined SF task. Yin [14] utilized an LLM fine-tuned with LoRA [15] for SLU tasks and achieved superior performance compared to methods trained from scratch or fine-tuned on masked language models. WHISMA [11], which employs a Whisper encoder [16] and a Llama-3 decoder [17] and is trained with modal alignment, is capable of performing zero-shot SLU tasks and has surpassed pipeline-based methods. Although these approaches all leverage LLMs for SLU tasks, they use different models, training methods, and data, making it impossible to compare them within a unified benchmark.

3. MAC-SLU DATASET

In this section, we provide details on text data collection and speech data synthesis, and a specific dataset analysis for MAC-SLU.

Intents	**Train**	**Dev**	**Test**	**Overall**	**Ratio (%)**
0	5305	419	26	5750	28.00
1	10018	768	826	11612	56.54
2	2183	166	246	2595	12.63
3	379	26	50	455	2.22
≥4	112	12	3	127	0.62
Total	17997	1391	1151	20539	100.00

Table 3. Distribution of samples by the number of intents in the train, dev, and test sets of MAC-SLU dataset, with totals and ratios.

3.1. Text Data Collection

The text data for the MAC-SLU dataset originates from real-world automotive cabin scenarios. We collected over 20,000 transcribed texts of Chinese spoken commands, along with their corresponding SLU parsing results. The training and validation sets consist of 17,997 and 1,391 samples, respectively, which were randomly selected and directly utilized the existing SLU parsing results as weakly labeled data. For the test set, we randomly partitioned 1,800 samples from the dataset. These samples were then manually reviewed and curated by three data annotators, who removed or corrected a small number of samples with parsing errors or blank intents. This process resulted in a final clean test set of 1,152 samples.

3.2. Speech Data Generation

We synthesized the corresponding Mandarin Chinese speech data for the transcribed texts using CosyVoice-2 [18]. The speaker embedding templates for the TTS process were derived from AIShell-1 [19], a widely used Chinese ASR dataset. Specifically, for each speaker in the train, dev, and test sets of AIShell-1, we randomly selected audio clips to create three distinct sets of speaker templates. When synthesizing each audio sample for MAC-SLU, a template was randomly chosen from the corresponding set to ensure speaker diversity and maintain isolation across different data splits.

3.3. Dataset Analysis

We compared the semantic diversity of MAC-SLU with commonly used datasets, including ATIS, SNIPS, FSC, and SLURP in Table 2. MAC-SLU is surpassed by SLURP only in the richness of its domains. This is primarily because MAC-SLU focuses on domains within automotive cabins, where the range of tools a user can invoke is relatively limited. In contrast, the smart-home scenario targeted by SLURP often involves a greater number of callable tools across more domains. However, in terms of the variety of intents and slots, MAC-SLU exceeds all existing single and multi-intent datasets, supporting more fine-grained intent and entity classification.

We present examples of domains, intents, and slots from the MAC-SLU dataset in Table 1 to facilitate an understanding of the

[3] conference.cipsc.org.cn/smp2019/evaluation.html

Category	Model	0-shot			5-shot			10-shot		
		Acc	F1	OA	Acc	F1	OA	Acc	F1	OA
NLU	Qwen3-1.7B	25.10	13.97	0.17	47.61	30.85	3.13	58.64	34.57	1.91
	Qwen3-4B	30.06	0.20	0.87	63.16	43.32	6.08	60.64	45.44	7.65
	Qwen3-8B	38.75	1.74	1.22	59.86	46.82	9.38	65.42	49.50	10.69
	Qwen3-32B	67.07	0.70	0.52	70.11	51.76	14.16	**70.37**	**55.09**	**14.42**
Pipeline (ASR + NLU)	Whisper + Qwen3-1.7B	24.07	3.24	0.09	46.83	27.22	2.09	56.30	31.00	1.65
	Whisper + Qwen3-4B	26.24	0.13	0.87	58.47	36.86	5.56	55.26	39.17	6.86
	Whisper + Qwen3-8B	35.88	1.47	1.48	54.91	39.84	8.43	60.73	42.72	9.30
	Whisper + Qwen3-32B	61.69	0.70	0.47	65.51	43.62	9.73	**66.12**	**47.38**	**10.86**
E2E SLU	Qwen2-Audio-Instruct	-	-	-	-	-	-	-	-	-
	Qwen2.5-Omni-3B	27.28	2.73	2.09	40.31	24.48	4.08	34.58	27.84	4.52
	Qwen2.5-Omni-7B	30.67	6.20	1.39	58.73	39.63	7.73	**62.47**	43.79	8.95
	Phi-4-Multimodal-Instruct	16.5	3.44	0.70	17.98	16.15	2.95	22.76	21.40	2.87
	GPT-4o-Audio	50.12	1.14	2.31	52.53	44.32	11.34	55.92	**46.45**	**12.21**
	Gemini-2.5-Flash	48.91	18.20	2.52	42.31	36.36	10.43	45.61	38.34	10.86

Table 4. In-context learning results for LLMs and LALMs on MAC-SLU dataset. Acc, F1, and OA represent Intent Classification Accuracy, Slot Filling F1-Score, and Overall Accuracy, respectively. The best results in each category are shown in **bold**. Qwen2-Audio-Instruct exhibited weak instruction-following abilities and failed to produce outputs in our required format.

specific data categorization. Due to space constraints, only a selection of examples is provided. The remaining domains include phone call, radio, weather, movies, and playback control.

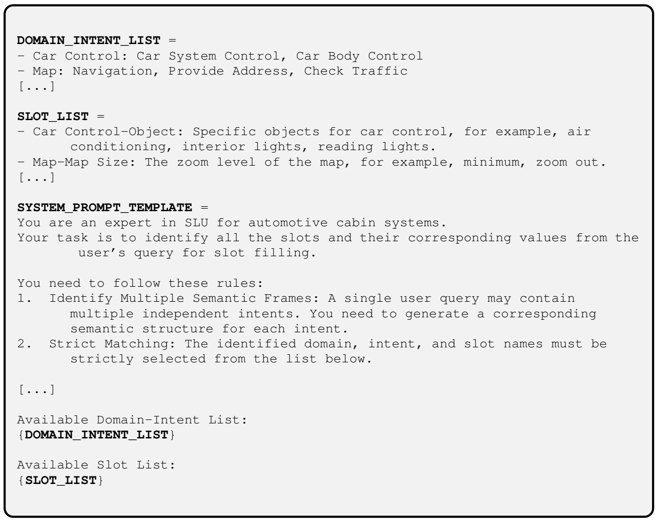
4. EXPERIMENTS

In this section, we evaluate the performance of existing LLMs and LALMs on the MAC-SLU dataset with direct inference, in-context learning, and SFT methods.

Models. Our experiments primarily target open-source LLMs and LALMs. For the LLMs experiments, we used various sizes of the Qwen3 [20]. For the multi-modal experiments, we investigated Qwen2-Audio-Instruct [21], Qwen2.5-Omni [22], Phi-4-multimodal-instruct [23], and MiniCPM-o-2_6 [24]. For the ASR model, we employed Whisper-Large-V3-Turbo [16] or Paraformer [25], which achieves a Character Error Rate (CER) of 10.40% or 3.64% on the MAC-SLU test set, respectively. As a supplement, we also tested the more powerful closed-source model, GPT4o-Audio-Preview-2024-12-17 and Gemini-2.5-Flash.

Implementation Details. Our experiments were conducted in two parts. For in-context learning methods, all experiments were run on Nvidia H20 GPUs, with inference accelerated using vLLM [26] deployment. For SFT methods, all experiments were conducted on Nvidia 3090 GPUs. We uniformly adopted the Llama-Factory [27] training framework and applied the LoRA [15] method to implement parameter-efficient fine-tuning. The LoRA rank and alpha were set to 16 and 32, respectively, following common default settings.

Evaluation Metrics. We followed the standard metrics for SLU tasks [28]. For the IC task, we calculated accuracy, and for the SF task, we calculated the F1 score. The Overall accuracy is calculated as the probability that IC and SF tasks are simultaneously correct.



```
DOMAIN_INTENT_LIST =
- Car Control: Car System Control, Car Body Control
- Map: Navigation, Provide Address, Check Traffic
[...]

SLOT_LIST =
- Car Control-Object: Specific objects for car control, for example, air
      conditioning, interior lights, reading lights.
- Map-Map Size: The zoom level of the map, for example, minimum, zoom out.
[...]

SYSTEM_PROMPT_TEMPLATE =
You are an expert in SLU for automotive cabin systems.
Your task is to identify all the slots and their corresponding values from the
       user's query for slot filling.

You need to follow these rules:
1.  Identify Multiple Semantic Frames: A single user query may contain
      multiple independent intents. You need to generate a corresponding
      semantic structure for each intent.
2.  Strict Matching: The identified domain, intent, and slot names must be
      strictly selected from the list below.

[...]

Available Domain-Intent List:
{DOMAIN_INTENT_LIST}

Available Slot List:
{SLOT_LIST}
```

Fig. 1. In-context learning prompt template for jointly SLU task. The intent lists, slot lists, and format are partially omitted for brevity.

4.1. In-Context Learning Results

LLMs and LALMs are capable of performing SLU tasks through in-context learning. Table 4 shows the performance of LLMs and LALMs under direct inference or in-context learning. For the IC results, our findings are similar to [10], namely that LLMs have the potential to complete the task using the in-context learning paradigm, and we extend this conclusion to the latest LALMs. For the SF results, we were surprised to find that with the carefully designed prompt shown in Figure 1, LLMs and LALMs were also able to provide correct answers. For text and speech inputs, our

Speech Query	Model Prediction	Ground Truth
Close the car window	{Intent1: Car Control: [{value: Body Control, name: intent}, {name: object, value: window}, {name: action, value: close}]}	{Intent1: Car Control: [{value: Body Control, name: intent}, {name: object, value: car window}, {name: action, value: close}]}
Turn on the rear windshield defroster	{Intent1: Car Control: [{value: Body Control, name: intent}, {name: action, value: turn on}, {name: position, value: rear}, {name: function, value: defrost}]}	{Intent1: Car Control: [{value: Body Control, name: intent}, {name: action, value: turn on}, {name: position, value: rear}, {name: function, value: windshield defrost}]}
I want to listen to Jay Chou's representative works	{Intent1: Music: [{name: action, value: I want to listen}, {name: artist, value: Jay Chou}, {name: album, value: representative works}, {value: Play Music, name: intent}]}	{Intent1: Music: [{name: action, value: listen}, {name: artist, value: Jay Chou}, {name: object, value: representative works}, {value: Play Music, name: intent}]}

Table 5. Comparison between fine-tuned Qwen2.5-Omni-7B predictions and ground truth. All data shown are English translations of the original Chinese. The differences are shown in blue.

system achieved F1 scores of up to 55.09% and 47.38%, respectively, which is significantly higher than 13.35% reported in [10]. However, even with the best-performing models like Qwen3-32B or GPT-4o-Audio, their Overall Accuracy did not exceed 15%, which reflects the challenges of the MAC-SLU dataset.

E2E LALMs can achieve performance comparable to that of pipeline systems. For speech input, Qwen2.5-Omni-7B slightly surpasses the pipeline system of a similar size (Whisper + Qwen3-8B), with advantages of 2% in the IC task and 1% in the SF task. This is attributed to the avoidance of ASR transcription error propagation in E2E SLU. However, when compared to the larger Qwen3-32B, the performance of Qwen2.5-Omni-7B still lags, indicating substantial potential for further performance improvements by scaling LALMs to larger sizes in the future. Furthermore, the closed-source GPT4o-Audio and Gemini-2.5-Flash demonstrate superior performance to open-source LALMs in the overall accuracy.

4.2. SFT Results

LLMs and LALMs that undergo in-domain SFT exhibit performance superior to methods based on in-context learning. Table 6 presents the performance of LLMs and LALMs after being fine-tuned on the training set, where all models show significant improvements. SFT boosted Qwen2.5-Omni-7B's performance over in-context learning, with increases of 29% in IC accuracy, 39% in SF F1 score, and 47% in overall accuracy. Currently, the most effective method for achieving optimal performance on SLU tasks with LLMs and LALMs remains fine-tuning on the training set. Enabling LLMs and LALMs agents to autonomously perform SLU tasks through in-context learning continues to pose a considerable challenge.

Error propagation in pipeline systems significantly degrades model performance. When processing text queries (CER=0%), the Qwen3-8B model achieves an overall accuracy of 60.73%, the highest among all SFT models. However, when integrated into a pipeline system, the performance of Qwen3-8B deteriorates by over 13% with transcripts from Paraformer (CER=3.64%) and by more than 25% with transcripts from Whisper (CER=10.40%). Inaccurate transcriptions from ASR substantially reduce the efficacy of LLMs on SLU tasks, resulting in final performance that is markedly inferior to that of LALMs.

Model	Acc	F1	Overall Acc
Qwen3-8B	90.91	84.69	60.73
Paraformer + Qwen3-8B	88.92	79.09	47.18
Whisper + Qwen3-8B	82.42	70.58	35.45
Qwen2-Audio-Instruct	86.46	79.46	49.87
Qwen2.5-Omni-3B	89.98	82.86	**56.56**
Qwen2.5-Omni-7B	**91.24**	**83.02**	55.60
MiniCPM-o-2_6	88.98	81.26	51.87
Phi-4-Multimodal-Instruct	81.69	74.12	37.97

Table 6. SFT results for LLMs and LALMs on MAC-SLU dataset. The ASR model was not fine-tuned, while the Qwen3-8B was fine-tuned on the NLU task. The best results are shown in **bold**.

4.3. Qualitative Analysis

A significant portion of observed errors arises because model outputs, while semantically correct, are phrased differently from the label. Our case study in Table 5 shows that in most instances, the models' responses convey the intended meaning but are penalized for lexical variation. For the SLU task, which aims to extract user semantics for downstream applications, these outputs are functionally correct and would be accepted by human evaluators. Therefore, standard SLU metrics relying on exact string matching likely underestimate the true intent comprehension capabilities of LALMs.

5. CONCLUSION

This paper introduced MAC-SLU, a novel Chinese multi-intent SLU dataset for automotive cabin scenarios, addressing the lack of diversity and complexity in existing datasets. Building upon this dataset, we established the unified benchmark for open-source LLMs and LALMs on SLU tasks. Our experiments demonstrate that while in-context learning shows potential, in-domain SFT remains crucial for optimal performance. Furthermore, E2E LALMs achieve performance comparable to traditional pipeline methods with LLMs by effectively mitigating ASR error propagation. Future work should focus on enhancing the in-context learning capabilities of models and exploring more semantically-aligned evaluation metrics.

6. REFERENCES

[1] Gokhan Tur and Renato De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.

[2] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al., "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM TASLP*, vol. 23, no. 3, pp. 530–539, 2014.

[3] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

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

[5] Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou, "ML-LMCL: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," *arXiv preprint arXiv:2311.11375*, 2023.

[6] Charles T Hemphill, John J Godfrey, and George R Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings Workshop*, 1990.

[7] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al., "SNIPS Voice Platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[8] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu, "A co-interactive transformer for joint slot filling and intent detection," in *Proc. ICASSP*. IEEE, 2021, pp. 8193–8197.

[9] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser, "SLURP: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[10] Mutian He and Philip N Garner, "Can ChatGPT detect intent? evaluating large language models for spoken language understanding," *arXiv preprint arXiv:2305.13512*, 2023.

[11] Mohan Li, Cong-Thanh Do, Simon Keizer, Youmna Farag, Svetlana Stoyanchev, and Rama Doddipatla, "WHISMA: A Speech-LLM to perform zero-shot spoken language understanding," in *Proc. SLT*. IEEE, 2024, pp. 1115–1122.

[12] Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," *arXiv preprint arXiv:2004.10087*, 2020.

[13] Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu, "CM-NET: A novel collaborative memory network for spoken language understanding," *arXiv preprint arXiv:1909.06937*, 2019.

[14] Shangjian Yin, Peijie Huang, Yuhong Xu, Haojing Huang, and Jiatian Chen, "Do large language model understand multi-intent spoken language?," *arXiv preprint arXiv:2403.04481*, 2024.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models.," *in Proc. ICLR*, vol. 1, no. 2, pp. 3, 2022.

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*. PMLR, 2023, pp. 28492–28518.

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al., "The Llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[18] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al., "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[19] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*. IEEE, 2017, pp. 1–5.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[21] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al., "Qwen2-Audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[22] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al., "Qwen2.5-Omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[23] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al., "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[24] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al., "MiniCPM-V: A GPT-4V Level MLLM on Your Phone," *arXiv preprint arXiv:2408.01800*, 2024.

[25] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proc. SOSP*, 2023, pp. 611–626.

[27] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," *arXiv preprint arXiv:2403.13372*, 2024.

[28] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu, "A survey on spoken language understanding: Recent advances and new frontiers," *arXiv preprint arXiv:2103.03095*, 2021.