Towards Zero-Shot SLU: An Empirical Study of Competing Architectural Paradigms

Beomseok Lee[1,2,3], Marco Gaido[2], Ioan Calapodescu[3], Laurent Besacier[3], Matteo Negri[2]

[1]University of Trento, [2]Fondazione Bruno Kessler, [3]NAVER LABS Europe

Abstract

Spoken Language Understanding (SLU) is crucial for enabling natural voice interactions with modern devices. However, traditional supervised models fail to generalize to new domains due to two key challenges: the prohibitive cost of data annotation and the inherent difficulty of transferring domain-specific intents. While the rise of Large Language Models (LLMs) offers a promising solution through zero-shot inference, the zero-shot SLU capabilities of emerging speech-enabled LLMs have remained largely unexplored. To address this gap, this paper provides the first comprehensive assessment, focusing on intent classification (IC), the first key sub-task of SLU, across 13 languages. We systematically evaluate a range of architectures, including cascaded, end-to-end, and hybrid systems for zero-shot SLU. Our analysis identifies the hybrid approach as the most effective architectural design for end-to-end SLU, and assesses multilingual transfer capabilities. The findings offer a detailed map of the challenges and opportunities, highlighting which models and settings are most promising for zero-shot SLU.

1 Introduction

Spoken Language Understanding (SLU) is the task of extracting semantic meaning and intent from spoken language (Tur and De Mori, 2011). While automatic speech recognition (ASR) and speech translation (ST) respectively focus on generating textual transcriptions and translations, SLU maps the input utterance to a structured, machine-readable representation highlighting intents (e.g., the overall goal of a request, like `book_flight`) and slots (the critical details to fulfill the request, like `origin_city: Boston` and `destination_city: San Francisco`). SLU is a key element of human-computer interaction, powering voice interactions on everyday devices like smartphones and digital assistants (Bellegarda, 2014; Marge et al., 2022).

Despite its widespread adoption, building robust SLU systems is challenging due to the high cost of collecting and labeling the data required for training dedicated models for specific domain and language settings (Lee et al., 2025). The challenge is exacerbated by the nature of existing datasets, which are often narrowly focused on specific intents and slots, preventing models from generalizing effectively to new tasks (Peng et al., 2025; Qin et al., 2021). By eliminating the need for domain/language-specific training data, zero-shot learning provides an alternative approach to address these data limitations.

The recent success of Large Language Models (LLMs), with their remarkable zero-shot capabilities thanks to their broad pretraining on diverse data, has opened a new frontier for this problem. Building on this, prior works have established a strong precedent by demonstrating impressive zero-shot NLU performance (Qin et al., 2025; Mirza et al., 2024; He and Garner, 2023). For instance, the results exhibited by Mirza et al. (2024) on the textual components of SLU benchmarks (FitzGerald et al., 2023; Coucke et al., 2018; Budzianowski et al., 2018) are particularly encouraging.

More recently, while some studies by Cho et al. (2024); Li et al. (2024) claim zero-shot SLU capabilities, their approaches fall short of a true zero-shot definition. This is because their models were pre-trained on either out-of-domain SLU datasets or on the transcripts of in-domain SLU datasets. Moreover, the zero-shot evaluation was conducted on relatively limited sets of target intents. It also remains an open question how effectively these zero-shot capabilities would transfer to the more challenging end-to-end (E2E) SLU task, which must handle the complexities of raw audio signals directly. The E2E approach can directly leverage non-textual acoustic cues like prosody, which may be relevant for intent classification but absent from a text transcript (Lugosch et al., 2019). For an am-

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 1–7

biguous utterance like '*Turn it up*', the textual information alone is insufficient. However, the presence of background audio provides the crucial context to correctly classify the intent as `audio_volume_up`. This motivates our focus on speech-enabled Large Language Models (SpeechLLMs), which integrate from powerful speech models and LLMs and show promising performance in various speech tasks (Arora et al., 2025).

Since Mirza et al. (2024) have already been highlighted that LLMs encounter significant difficulties in addressing the complex slot filling (SF) task in zero-shot settings, as we also confirmed in preliminary investigations, we focus our investigation into zero-shot SLU on intent classification (IC), where zero-shot LLM performance is considerably better. Specifically, by comparing 6 different systems belonging to 3 architectural designs in 13 languages, our study demonstrates that:

- Hybrid systems, which feed both automatic transcripts and continuous representations by the speech encoder into the LLM, are best suited for zero-shot SLU, effectively bridging the modality gap where end-to-end systems falter.
- Multilingual capabilities, assessed through performance of different language families, are strongly dependent on both the LLM's innate multilingual proficiencies and the speech encoder's per-language accuracy, as well as the language families covered by the data used to train the speech encoder-LLM interface in E2E models.

2 Zero-shot SLU

We evaluate a diverse set of architectures for zero-shot SLU, including cascaded, end-to-end, hybrid, and commercial systems. Unless specified otherwise, Llama-3.1-8B-Instruct[1] serves as the default LLM for these architectures.

Cascaded systems are evaluated in two configurations, each consisting of an off-the-shelf Speech Foundation Model (SFM) followed by the LLM. The two SFMs used for transcription are Whisper-large-v3[2] (Radford et al., 2023) and Seamless-m4t-v2-large[3] (Barrault et al., 2023). These models were chosen to represent distinct architectural paradigms, as Whisper employs a Transformer encoder (Vaswani et al., 2017) while Seamless-m4t utilizes a Conformer encoder (Gulati et al., 2020). In this pipeline, the input speech is first transcribed by one of the SFMs, and the resulting text is subsequently processed by the LLM to perform NLU.

End-to-end (E2E) systems are represented by two instances. A custom-built SpeechLLM utilizing a Seamless-m4t encoder and a pre-existing SpeechLLM employing a Whisper encoder. Both systems utilize continuous speech representations, while we do not explore discrete token-based systems due to their known inferior performance in SLU tasks (Wang et al., 2025). We specifically developed the first system to maintain full control over the training data and thus ensure a strict zero-shot scenario for the SLU task. The system utilizes Seamless-m4t as the speech encoder and Llama LLM. To bridge these heterogeneous models, we introduce a Transformer-encoder interface that projects speech features into the LLM's embedding space similar to Verdini et al. (2025). Constrained by practical considerations of limited computing resources, we freeze the parameters of both the speech encoder and the LLM, training only the randomly initialized interface via instruction tuning (Zhang et al., 2025). Crucially, the interface is trained solely using ASR data and objectives, without incorporating any SLU-related data or supervision.

As a third-party E2E SpeechLLM for our comparative evaluation, we included Qwen2-Audio-7B-Instruct model.[4] This model has a different architecture, composed of a Whisper encoder and a Qwen2 LLM (Chu et al., 2024). Unlike our minimally-trained system, it has been extensively fine-tuned on a large-scale, 510k-hour supervised speech-text dataset, representing a common paradigm for developing powerful, publicly available models. While not directly comparable to our setup, its performance provides a valuable benchmark for understanding how heavily fine-tuned systems operate in a zero-shot SLU scenario.

Hybrid systems condition their output on both the raw speech and its transcription, in contrast to E2E systems that process audio directly. For this category, we employ the Desta2.5-Audio-Llama model,[5] which integrates the full Whisper-large-v3 and LLaMA-3.1-8B-Instruct models (Lu et al.,

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2] https://huggingface.co/openai/whisper-large-v3

[3] https://huggingface.co/facebook/seamless-m4t-v2-large

[4] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[5] https://huggingface.co/DeSTA-ntu/DeSTA2.5-Audio-Llama-3.1-8B

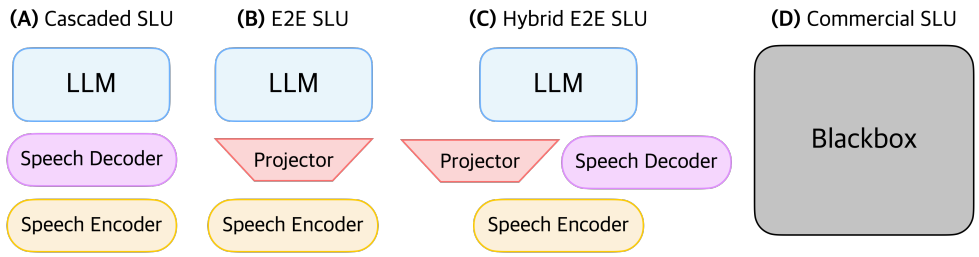

Figure 1: Comparison of different architectures

2025) instruction tuned with 7K hours of speech-text data. This architecture leverages the explicit text transcription from the Whisper component alongside its acoustic representations, providing a meaningful comparison point that occupies a middle ground between purely cascaded and E2E systems.

Commercial system Gemini-2.5-flash-lite[6] (Comanici et al., 2025) serves as a state-of-the-art benchmark in our study, allowing us to contextualize our results. Although its architecture and training data are opaque, this black-box comparison is valuable for understanding the performance gap between proprietary and open-source models. Since the model is multimodal, we evaluate it on both NLU (text input) and SLU (speech input). This also allows us to measure the modality gap within a single commercial system and compare these results directly against our open-source configurations.

3 Prompting strategy

Recognizing the critical role of prompting in the zero-shot performance of LLMs (Reynolds and McDonell, 2021), we conducted a preliminary study to establish an optimal prompt design. This investigation was performed on a representative subset of the SLURP development dataset (Bastianelli et al., 2020) to isolate the effects of different prompting strategies.

3.1 Plain prompts and constrained decoding

Our initial approach utilized a plain prompt, detailed in Table 1, which instructs the model to select an intent from a given list. While this method achieved 57.23% IC accuracy, a qualitative analysis of the outputs revealed a tendency for the model to hallucinate intents not present in the candidate list. To mitigate this, we enforced constrained decoding following Willard and Louf (2023) to limit the model's output to only valid intents. Unexpectedly, this intervention proved detrimental, reducing accuracy to 52.4%. We hypothesize that mechanically restricting the output space is an insufficient remedy. It cannot compensate for a lack of contextual understanding in the prompt, particularly when faced with fine-grained, semantically similar labels (e.g., `alarm_set` vs. `alarm_query`, `play_music` vs. `music_query`). This outcome highlights the need for a more contextually rich prompting strategy.

3.2 Chain-of-Thought inference

To elicit robust zero-shot intent classification performance, we design a multi-step prompting strategy inspired by the Chain-of-Thought (CoT) method (Wei et al., 2022). As shown in Table 1, instead of directly asking for an intent classification, our prompt instructs the model to reason about the speaker's underlying intention first. This initial step compels the model to generate its own understanding of the context. This self-augmented reasoning then serves as a richer foundation for the final classification step. This method gave 65.06% accuracy, boosting the performance of the plain prompt by 8%. Additionally, for SLU, we enriched the prompt by asking for transcription of the input utterance, in order to enable better error analysis by helping to disentangle failures caused by incorrect speech transcription from those caused by flawed semantic reasoning.

4 Experimental settings

4.1 Datasets and metrics

To evaluate zero-shot SLU performance under challenging conditions, we utilize the test splits of three datasets: SLURP (Bastianelli et al., 2020), MASSIVE (FitzGerald et al., 2023), and Speech-MASSIVE (Lee et al., 2024). SLURP is an English dataset centered on an in-home personal robot assistant, encompassing 18 domains and 60 intents

[6] https://deepmind.google/models/gemini/flash-lite/

Turn	Role	Plain
1	System	You are a helpful assistant. You have to choose one option from the list. Answer only with the given option. Given the following sentence, choose a single intent of the sentence from the following intent list. [datetime_query], ... , [lists_query]
	User	Sentence: [gold_transcript] Intent:
Chain-of-Thought NLU Inference		
1	System	You are a helpful assistant.
	User	Explain speaker's intention of saying [gold transcript]
2	System	You are a helpful assistant. Choose only one of the following intent categories: [datetime_query], ... , [lists_query]
	User	Based on the speaker's utterance and your understanding, choose the single action you should take for the speaker. You must answer from given intent categories only. Answer with the intent only.
Interface Training		
	User	Transcribe speech into text. <begin_of_speech>[speech_features]<end_of_speech>
Chain-of-Thought SLU Inference		
1	System	You are a helpful assistant. The utterance is provided within the tags <begin_of_speech>and <end_of_speech>
	User	1. Turn the speaker's utterance into text. 2. Explain speaker's intention.<begin_of_speech>[speech_features]<end_of_speech>
2	Both	Same prompt as NLU inference step #2.

Table 1: Prompt strategy for the inference and the training

with recorded speech, transcript and semantic annotations. MASSIVE is a multilingual text dataset that expands SLURP's scenarios into 52 languages, while Speech-MASSIVE provides corresponding speech data for 12 of those languages. Across all evaluations in this work, the SLURP test split is used for English NLU and SLU evaluation. MASSIVE is used for non-English NLU evaluation and Speech-MASSIVE is used for non-English SLU evaluation.

For training the interface layer of our custom SpeechLLM, we used the English portion of Common Voice 17.0 developed by Ardila et al. (2020) (2.6K hours) exclusively for an ASR task. We report intent accuracy in all the different model settings and languages.

4.2 E2E system Traning

To train the custom-built E2E system of Section 2, we configure the interface architecture to have an input layer, four Transformer Encoder layers, and an output layer. The encoder layers have a hidden size of 768 and an intermediate dimension of 3072. The output layer projects the resulting features to a dimension of 4096, matching the embedding space of the Llama-3.1-8B-Instruct, totaling 33M trainable parameters. Throughout all training experiments for this component, we used a constant learning rate of 1e-5.

5 Results

We present the full results of our comparative study in Table 2. The NLU-only model (A), serving as our topline, demonstrates the capability of an off-the-shelf LLM on this task, achieving a 52.8% average intent accuracy across 13 languages, with a performance of 61.03% in English. Performance varies significantly by language; English achieves the highest accuracy, whereas languages such as Turkish, Hungarian, Vietnamese, and Arabic are among the lowest-performing. These results establish a crucial baseline, indicating that zero-shot intent classification is challening even without the audio component, particularly in a multilingual context.

When moving to the cascaded SLU approach, models (B) and (C) respectively exhibit performance drops of 2% and 4% compared to the NLU topline, a degradation that we attribute to the Word Error Rate (WER) of the upstream ASR model. While Whisper exhibits a better average WER overall ((I) vs. (J)), for languages where Seamless-m4t has a lower WER (e.g., Arabic, Hungarian, Vietnamese), this ASR advantage directly translates into superior SLU performance. Consequently, for these languages, the Seamless-m4t-based cascaded system (C) outperforms the Whisper-based one (B). This proves that, unsurprisingly, a critical dependency for cascaded approaches exists where ASR performance limits the ceiling for final accuracy while the LLM cannot recover transcription errors.

The transition to an E2E approach reveals a more substantial performance decline. Both E2E models, (D) and (E), show significant drops from the cascaded systems, highlighting the inherent difficulty of aligning raw speech representations with LLM embeddings. Of the two, model (E) shows the weakest performance overall. Ultimately, these

	System	Model	en*	ar	de	es	fr	hu	ko	nl	pl	pt	ru	tr	vi	avg all	avg w/o en
(A)	NLU	Llama-3.1-8B	61.03	45.16	53.33	53.90	54.27	49.66	54.84	54.00	53.09	52.59	55.38	51.34	47.81	52.80	52.11
(B)	Cascaded	Whisper + Llama-3.1-8B	54.70	38.47	52.56	53.33	52.86	46.91	52.19	52.86	52.05	51.58	54.67	48.89	44.28	50.41	50.05
(C)		Seamless + Llama-3.1-8B	51.04	40.99	51.75	51.82	50.34	47.48	45.56	52.22	50.27	45.16	51.61	47.51	44.75	48.50	48.29
(D)	E2E	Custom SpeechLLM	47.73	16.85	47.57	33.42	42.37	29.02	16.44	47.34	33.29	28.61	32.31	20.95	22.60	32.19	30.90
(E)		Qwen2-Audio-7B	32.70	9.52	33.99	33.66	33.52	4.81	23.94	22.60	10.96	26.63	29.25	7.60	3.67	20.99	20.01
(F)	Hybrid	Desta 2.5-Audio	52.97	38.26	54.30	53.26	52.25	46.87	53.33	54.07	53.30	52.56	55.25	49.60	44.01	50.77	50.59
(G)	Commercial	Gemini-2.5 flash-lite NLU	75.18	64.86	72.63	70.51	71.62	70.41	73.67	71.72	72.19	71.08	72.46	71.92	70.85	71.47	71.16
(H)		Gemini-2.5 flash-lite SLU	56.76	56.86	70.01	69.54	67.05	62.14	61.13	67.92	66.58	69.13	71.05	67.92	61.74	65.22	66.42
(I)	ASR	Whisper	17.91	34.19	11.84	8.95	11.09	20.98	26.42	10.52	12.58	12.11	8.99	18.06	14.94	16.04	15.89
(J)		Seamless	24.66	32.78	13.01	10.32	13.73	18.18	42.39	11.43	14.89	24.63	11.71	19.27	11.75	19.13	18.67

Table 2: Zero-shot Intent Classification (IC) Accuracy (%) and ASR Word Error Rate (WER, %) across all models. Rows (A)-(H) report IC accuracy (↑), while rows (I)-(J) report ASR WER (↓). The English (en) test set from SLURP contains 13,078 samples with multiple recordings per transcript; all other languages have 2,974 samples with a single recording per transcript.

findings demonstrate that the E2E paradigm, despite its theoretical advantages, is not yet as robust as cascaded systems in a purely zero-shot context.

Among the open-source models, the hybrid E2E SLU model (F) shows highly promising results (50.77 in average), generally matching the performance of the cascaded systems (50.41 in average). For some languages (German, Dutch, Polish), it even surpasses the cascaded accuracy by more than 1%, occasionally matching the NLU topline performance. While these gains are modest, the hybrid approach's true strength appears to have its robustness to ASR errors. This suggests that **the hybrid architecture can partially mitigate ASR error propagation. By leveraging both acoustic features and the transcribed text, it can provide a richer, more resilient context to the LLM than text alone.**

For the commercial model, we observe the same trend: the SLU task is more challenging than the NLU task ((G) vs. (H)). Although a significant performance gap exists between the commercial and open-source models, a direct comparison between the models remains difficult due to the proprietary nature of the commercial model. However, this analysis highlights both the current performance gap and promising architectural directions for the open-source community, especially considering the likely disparities in training data scale.

To better understand the factors influencing multilingual performance, we performed a correlation analysis between SLU system's accuracy and two key variables: the LLM's NLU performance across languages and the speech encoder's ASR performance. The Whisper-based systems ((B) and (F)) exhibit remarkably strong correlations. The cascaded system (B) shows correlations of 0.90 with the LLM and 0.71 with ASR, while the hybrid system (F) has similarly high correlations of 0.83 (LLM) and 0.71 (ASR). In comparison, the Seamless-m4t-based models show slightly weaker, yet still significant, correlations. The cascaded model (C) has correlations of 0.72 (LLM) and 0.65 (ASR), and the E2E model (D) has a correlation of 0.63 with the LLM's performance. These findings quantitatively demonstrate that **robust multilingual zero-shot SLU is heavily dependent on two distinct factors: the innate multilingual capabilities of the backbone LLM and the per-language quality of the speech encoder**.

Additionally, our English-only training setup for the E2E interface (D) allows us to probe the limits of its cross-lingual generalization. The results reveal a clear linguistic dependency: the performance drop for closely related languages like German and Dutch is minimal (4.1%), whereas all other languages exhibit a much larger degradation (20%). However, the accuracy for all other languages is close to that of SFMs (Whisper and Seamless-m4t), demonstrating that the model partially retains the ability to process languages unrelated to the one used to train the interface. While this highlights the insufficiency of single-language training for true multilingual performance, it also positively indicates that targeted interface training based on the language families of interest could be a key strategy for improving these E2E models. Consequently, a promising direction for future work is to strategically develop language-family-specific interfaces to improve performance. Furthermore, building a single, robust multilingual interface may benefit from techniques that isolate the different language families, such as building mixture-of-experts interfaces where each expert processes a specific language family. We leave the investigation of this idea and effective multilingual interface training to future work.

In summary, our findings align with and extend

previous work showing that SLU is more difficult than NLU (Huang et al., 2023), demonstrating that this challenge is significantly amplified in a zero-shot context. We demonstrate that hybrid systems are the most promising architecture for addressing the task, closely followed by cascaded pipelines. In fact, the hybrid architecture consistently emerged as the most robust approach, demonstrating higher resilience to ASR errors than cascaded approaches. On the contrary, current E2E models are lagging behind by a significant margin. We can conclude that much research is still needed for them to close this gap, e.g. by means of more advanced interface designs or training strategies. Furthermore, achieving true multilingual performance remains a significant challenge. The results are heavily dictated by language-specific ASR quality for cascaded systems and linguistic proximity to the training data of the interface between the SFM encoder and the LLM for E2E models.

6 Conclusion

In this paper, we conducted a systematic evaluation of zero-shot Spoken Language Understanding (SLU) for intent classification (IC) using speech-enabled Large Language Models (SpeechLLMs). Our analysis covers diverse model architectures across 13 languages. By comparing text-only NLU toplines against cascaded, end-to-end (E2E), and hybrid systems, we aimed to map the current landscape, identify the most effective architectural designs, and understand the challenges of multilingual transfer.

Our findings reveal several key insights. First, we quantitatively demonstrate that hybrid architectures currently offer the most robust and promising solution for zero-shot SLU. They effectively match the performance of strong cascaded systems while showing greater resilience to upstream ASR errors. In contrast, current end-to-end models lag significantly behind, highlighting the substantial challenge of aligning speech and text representations. Second, our multilingual analysis highlights that performance is far from uniform. It is heavily influenced by two distinct factors: the innate multilingual capabilities of the LLM and the speech encoder's ASR capabilities for each language. We also show that zero-shot cross-lingual transfer from an English-only trained model is limited, primarily benefiting only linguistically similar languages.

For future works, as our work focused exclusively on intent classification, extending this analysis to the more complex task of slot filling is a critical next step. Furthermore, to unlock the potential of end-to-end SLU systems in multilingual settings, dedicated strategies account for language similarities should be investigated for training SpeechLLM interfaces.

Acknowledgements

This paper has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People).

References

Rosana Ardila and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Siddhant Arora and 1 others. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv:2504.08528*.

Loïc Barrault and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv:2312.05187*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on EMNLP*, pages 7252–7262.

Jerome R. Bellegarda. 2014. Spoken language understanding for natural interaction: The siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14, New York, NY. Springer New York.

Paweł Budzianowski and 1 others. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on EMNLP*, pages 5016–5026, Brussels, Belgium.

Jaejin Cho and 1 others. 2024. Zero-shot intent classification using a semantic similarity aware contrastive loss and large language model. In *2024 IEEE ICASSP*.

Yunfei Chu and 1 others. 2024. Qwen2-audio technical report. *arXiv:2407.10759*.

Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

Alice Coucke and 1 others. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv:1805.10190*.

Jack FitzGerald and 1 others. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proc. of the 61st ACL*, pages 4277–4302.

Anmol Gulati and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Mutian He and Philip N. Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. In *Interspeech 2023*, pages 1109–1113.

He Huang, Jagadeesh Balam, and Boris Ginsburg. 2023. Leveraging pretrained asr encoders for effective and efficient end-to-end speech intent classification and slot filling. In *Interspeech 2023*, pages 2933–2937.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. 2024. Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond. In *Proc. Interspeech 2024*, pages 817–821.

Beomseok Lee, Marco Gaido, Ioan Calapodescu, Laurent Besacier, and Matteo Negri. 2025. Speech foundation models and crowdsourcing for efficient, high-quality data collection. In *Proceedings of the 31st COLING*, pages 6816–6826.

Mohan Li, Cong-Thanh Do, Simon Keizer, Youmna Farag, Svetlana Stoyanchev, and Rama Doddipatla. 2024. Whisma: A speech-llm to perform zero-shot spoken language understanding. In *2024 IEEE SLT*.

Ke-Han Lu and 1 others. 2025. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv:2507.02768*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech 2019*, pages 814–818.

Matthew Marge and 1 others. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech Language*, 71:101255.

Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. 2024. ILLUMINER: Instruction-tuned large language models as few-shot intent classifier and slot filler. In *Proc. of the 2024 LREC-COLING*, pages 8639–8651.

Jing Peng and 1 others. 2025. A survey on speech large language models for understanding. *Authorea Preprints*.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth IJCAI-21*, pages 4577–4584.

Libo Qin and 1 others. 2025. Croprompt: Cross-task interactive prompting for zero-shot spoken language understanding. In *2025 IEEE ICASSP*, pages 1–5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI*.

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

Ashish Vaswani and 1 others. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Francesco Verdini and 1 others. 2025. How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not. In *Proc. Interspeech 2025*.

Dingdong Wang, Junan Li, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen Meng. 2025. Speech discrete tokens or continuous features? a comparative analysis for spoken language understanding in speechllms. *arXiv:2508.17863*.

Jason Wei and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv:2307.09702*.

Shengyu Zhang and 1 others. 2025. Instruction tuning for large language models: A survey. *arXiv:2508.17863*.