StarDrinks: An English and Korean Test Set for SLU Evaluation in a Drink Ordering Scenario

Marcely Zanon Boito, Caroline Brun, Inyoung Kim, Denys Proux, Salah Ait-Mokhtar, Nikolaos Lagos, Jean-Luc Meunier and Ioan Calapodescu
NAVER LABS Europe, France
contact: marcely.zanon-boito@naverlabs.com

Abstract

LLMs and speech assistants are increasingly used for task-oriented interactions, yet their evaluation often relies on controlled scenarios that fail to capture the variability and complexity of real user requests. Drink ordering, for example, involves diverse named entities, drink types, sizes, customizations, and brand-specific terminology, as well as spontaneous speech phenomena such as hesitations and self-corrections. To address this gap, we introduce *StarDrinks*, a test set in English and Korean containing speech utterances features, transcriptions, and annotated slots. Our dataset supports speech-to-slots SLU, transcription-to-slots NLU, and speech-to-transcription ASR evaluation, providing a realistic benchmark for model robustness and generalization in a linguistically rich, real-world task.

Keywords: natural language understanding, spoken language understanding, automatic speech recognition, language models

1. Introduction

LLMs and speech assistants have rapidly advanced in recent years, enabling increasingly natural and efficient interactions between humans and machines. These systems are now deployed in numerous everyday applications, including personal assistants, customer service bots, and automated ordering systems. Despite these impressive developments, the evaluation of such models is still largely conducted under controlled or simplified conditions that fail to reflect the diversity and unpredictability of real-world usage (Lunardi et al., 2025; Wei et al., 2024; Xiong et al., 2025). Standard benchmarks often rely on clean, well-structured inputs and limited vocabularies, which do not capture the full range of linguistic variability, noise, and contextual complexity encountered in spontaneous human speech.

One particular deployment setting where these limitations become particularly evident is drink ordering, which, while seemingly simple, involves complex linguistic phenomena and decision-making structures. Understanding and processing such requests requires a natural language understanding (NLU) model to interpret the semantic intent behind user utterances, and in the case of spoken language understanding (SLU), we add to this challenge the handling of speech input, including hesitations, self-corrections, disfluencies, and prosodic cues. Real drink orders often include multiple attributes such as drink type, size, temperature, milk preference, flavorings, and toppings, expressed in varying orders or using brand-specific terminology, further complicating semantic parsing.

Despite the importance of NLU in task-oriented dialogue systems, there is a notable shortage of datasets that reflect realistic spoken interactions, and, to the best of our knowledge, no publicly available dataset exists for spoken drink ordering. As a result, current models are often trained and evaluated on limited or artificial benchmarks, performing well in controlled settings but struggling to generalize to authentic, complex user interactions.

To address the gap in realistic evaluation resources, we release *StarDrinks*, a test set in English and Korean for the drink-ordering scenario. *StarDrinks* contains speech utterance features, corresponding transcriptions, and annotated slots, making it suitable for multiple evaluation settings. It can be used for speech-to-slots SLU evaluation of assistants, transcription-to-slots NLU evaluation, and speech-to-transcription ASR evaluation in a scenario that requires generalization to previously unseen named entities. The dataset is collected from authentic drink orders, capturing diverse named entities, complex combinations of order attributes, and natural speech phenomena such as hesitations, self-corrections, and disfluencies. By providing this rich and challenging test set, *StarDrinks* enables more realistic assessments of LLMs and spoken language assistants, supporting the development of more robust, context-aware systems capable of handling real-world interactions.

This paper is organized as follows. Section 2 reviews existing datasets for NLU and SLU. Section 3 describes the dataset creation process in detail. The resulting dataset is presented in Section 4, followed by a use case involving a drink-ordering agent in Section 5. Finally, Section 6 concludes the paper.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5749–5756
11–16 May 2026.

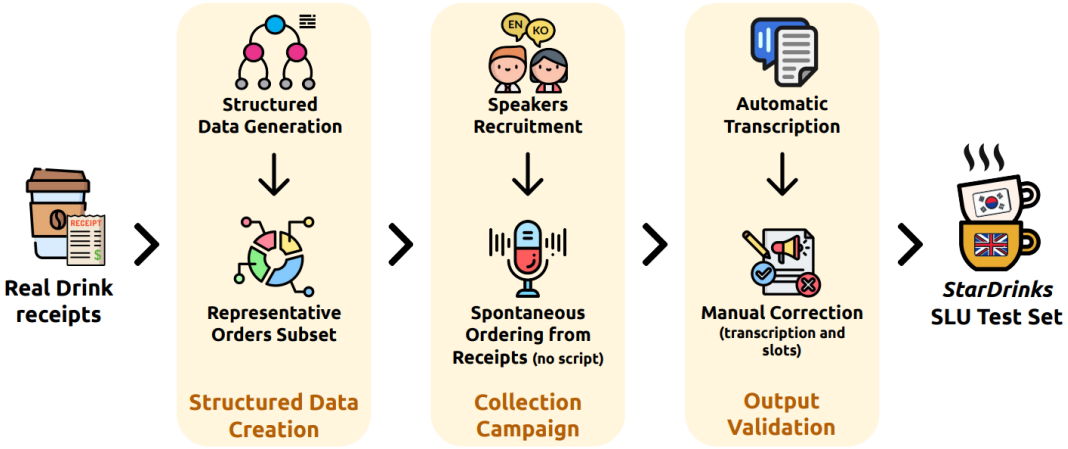

Figure 1: *StarDrinks* data generation pipeline overview.

2. Related Works

SLU and NLU datasets typically consist of speech and text data, respectively, accompanied by annotations for intent classification and/or slot filling tasks. Moreover, datasets can be either *single* or *multi-turn*, meaning that the task is accomplished in either a single or through multiple turns of interaction, respectively. In this paper, we focus on the single-turn setting, where a customer places their drink order in a single interaction. Our dataset also does not include intent classification labels, since the intent is always the same: to order drinks.

Although many datasets were released in recent years, NLU, and particularly SLU, remain scarcely covered, specially in multilingual settings. We now highlight relevant NLU/SLU resources we are aware of. *SNIPS* (Coucke et al., 2018) is a single-turn dataset covering 7 intents, such as booking a restaurant or rating a book. The *ATIS* dataset (Hemphill et al., 1990) is a popular NLU/SLU benchmark containing spoken queries related to air travel, annotated with intents (e.g., *flight*, *airfare*) and slots such as departure city and date. The *SLURP* corpus (Bastianelli et al., 2020) offers a large-scale, multi-domain resource with real spoken utterances for virtual assistant scenarios, annotated with 18 intent types and over 50 slot categories. *MASSIVE* (FitzGerald et al., 2023) and *Speech-MASSIVE* (Lee et al., 2024) extent a subset of SLURP to 51 text languages, and 12 speech languages, respectively.

The *Fluent Speech Commands (FSC)* dataset (Lugosch et al., 2019) consists of spoken smart home commands (e.g., "turn on the lights in the kitchen") annotated with triplets representing action, object, and location, effectively encoding intents and slots. The *SmartLights* dataset (Coucke et al., 2018; Gupta et al., 2020) provides spoken commands for smart home device control, annotated for intent and slot-based semantics, serving as a realistic testbed for voice assistant systems.

The *Spoken Task Oriented Parsing (STOP)* (Tomasello et al., 2023) dataset is a large semantically complex end-to-end spoken language dataset for end-to-end semantic parsing. It contains over 200,000 audio files from over 800 different speakers, recorded through Amazon's Mechanical Turk. The text utterances and semantic parses are taken from TOPv2 (Chen et al., 2020), covering 8 different domains: alarm, event, messaging, music, navigation, reminder, timer and weather.

Finally, The *FoodOrdering dataset* (A. Rubino et al., 2022) is an English NLU dataset semantically close to *StarDrinks*. It is a task-oriented parsing resource focused on the food-ordering domain, drawing utterances and annotations from menus of five representative different venues. Human-generated data was crowd-sourced via Mechanical Turk, where participants crafted natural language requests for orders serving one or multiple people based on the provided menus, with the resulting utterances then manually annotated into machine-executable formats.

We highlight that from these aforementioned works, only *Speech-MASSIVE* present speech for the Korean language, and none covers the spoken drink ordering scenario.

3. Dataset Creation

Figure 1 presents the general pipeline for creating *StarDrinks*. We now provide a brief overview, with further details being presented in the following sections. Starting from real drink order receipts collected from a popular coffee chain in South Korea, we extracted structured data and selected a representative subset of entries that both covered the

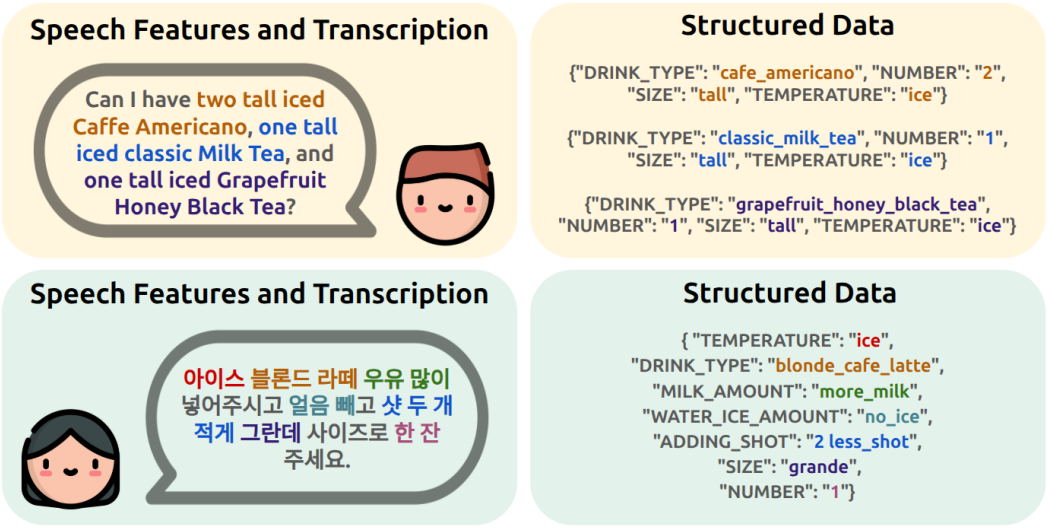

Figure 2: *StarDrinks* data examples from the English (top) and Korean (bottom) splits.

entire menu and reflected the overall distribution of orders (Section 3.1). We recruited native English and Korean speakers on the Prolific platform, assigning each participant a subset of receipts and asking them to record utterances ordering the corresponding items (Section 3.2). From the collected speech, we generated transcripts using a state-of-the-art ASR model, followed by manual correction of both transcripts and slot annotations to produce the final output (Section 3.3). The final *StarDrinks* SLU test set contains speech recordings, corresponding transcriptions, and slot annotations for each drink order.

3.1. Structured Data Generation

We started the data generation process with 2,500 samples corresponding to real drink order receipts, which we used in order to create a large set of synthetic variations of drink order structures. Using a semantic schema that represents possible values for drink attributes (see Table 5), we replaced elements such as drink types, sizes, temperatures, options (e.g., milk type, syrup amount), and quantities with compatible alternatives. Low-frequency structures were prioritized to enhance diversity, resulting in a final output of up to 83,974 structures. From these, we sampled a subset that includes at least one occurrence of all attribute-value pairs in the drink orders, yielding 326 structures that are used for the Prolific annotation process.

3.2. Data Collection

To gather speech recordings aligned with our collection of drink order receipts, we recruited participants via the Prolific platform.[1] Two separate data collection campaigns were conducted, one for English and one for Korean. The participants were compensated through the standard recommendation proposed by the platform, corresponding to £10.66 per hour.

Before starting a session, the participants were presented with some general guidelines. Participants were asked to make a drink order in a natural manner according to the information on the screen. We also mentioned that they could use common synonyms or abbreviations for drink and option names.

During a recording session, the recording screen presented a given receipt in English or Korean (see Figure 3). We instructed each participant to record one utterance per drink that included all ordering options—quantity, size, customizations, and temperature. After recording, participants verified the utterance by using the "Replay" button and were allowed to re-record if they were not satisfied with the result.

For the English dataset, we recruited 32 participants located in the United Kingdom, aged between 21 and 50 years. All participants were native English speakers. Recordings were required to be made on a desktop or laptop computer equipped with a microphone. Each participant viewed 10 receipts in a single session, and participation was restricted to one session per individual. The demographic composition of this panel was 56.3% male and 43.7% female. In terms of ethnicity, 68.8% identified as White, 15.6% as Asian, 12.5% as Black, and 3.1% as mixed ethnicity. On average, participants spent 66.4 seconds recording each receipt.

For the Korean dataset, fewer eligible participants were available in the platform. Consequently, some selection criteria were relaxed. Participants were recruited from both the United Kingdom and

[1] https://www.prolific.com/

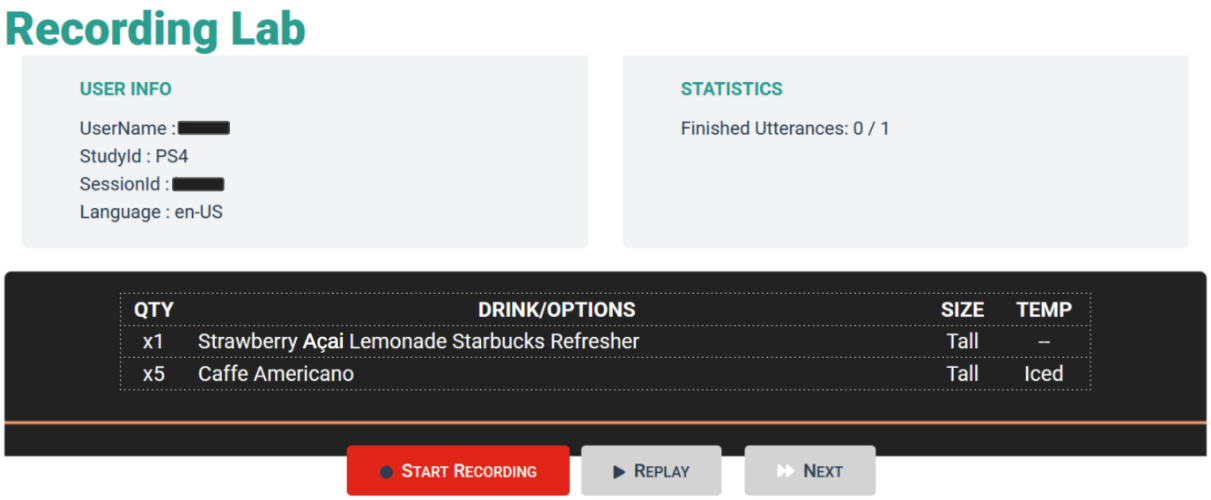

Figure 3: An example from our recording session for English.

the United States, provided they were native Korean speakers with English as a second language. Each participant could complete up to three sessions, while the age range criteria remained unchanged. This panel was composed of 29 participants with 31% male and 69% female. The average time spent recording each receipt was 51.7 seconds.

3.3. Output Validation

We collected 291 audio files for English, and 295 for Korean. We used whisper-large-v3 (Radford et al., 2023) to produce *approximated* speech transcripts, since this model is very competent in English and Korean ASR. We then manually annotated the whisper output in order to validate that i) the speech corresponds to the structured data slots; ii) the speech corresponds to the transcript. In both cases, we corrected the slots and the transcripts when discrepancies are found, and we removed the utterance from the test set if we judged that the participants did not perform the task correctly. This annotation was performed by the authors of this paper, including fully fluent English speakers and one native Korean speaker.

We highlight that while the design of our data collection campaign forces users to produce spontaneous speech (only a receipt is shown), we did not encourage users to produce disfluent speech, which we find to be a natural consequence of collecting speech in this setting. Moreover, we did not annotate disfluent speech, as whisper is trained to ignore and correct such disfluencies. However, qualitative inspection shows that most English outputs contain either hesitation markers or repetitions.

	En	Ko
# validated samples	255	295
# total slots	2,058	2,423
# speakers	32	29
Duration	53 min	45.7 min
Avg. utterance length	12.4 s	9.3 s

Table 1: Statistics over *StarDrinks*.

	WER	CER
English	9.2	3.6
Korean	22.9	7.3

Table 2: ASR results for whisper-large-v3.

4. The *StarDrinks* Test Set

The *StarDrinks* dataset was designed to evaluate speech assistants in realistic scenarios, focusing on drink ordering in English and Korean. Each order can include up to six drinks. As a SLU dataset, it consists of speech utterance features paired with gold-standard transcriptions and structured NLU outputs. While centered on a single intent, drink ordering, the dataset features 15 distinct slots, encompassing 45 unique drink types along with their various customization options. Figure 2 presents examples from the dataset. Table 5 presents all existing slot values. Statistics are reported in Table 1.

5. Use Case: Drink Ordering Agent

To demonstrate the utility of our test set, we evaluated a drink ordering agent built from state-of-the-art speech and text systems. We first present ASR results in Section 5.1, illustrating the challenge of recognizing unknown named entities after training.

Configuration	ASR Model	English		Korean	
		UEM (%)	Slot F1 (%)	UEM (%)	Slot F1 (%)
Gold Trans. + 3-shots (NLU)	None	**87.06**	**98.04**	**89.83**	**98.76**
Gold Trans. + 0-shot (NLU)	None	71.76	94.51	85.76	97.75
ASR + 3-shots	Whisper-large-v3	84.31	97.37	84.75	97.45
ASR + 0-shot	Whisper-large-v3	60.00	89.96	67.80	93.72

Table 3: NLU/SLU results on the *StarDrinks* English and Korean test sets with GPT-4o model.

Reference	**Whisper's Output**
Please can I have two cafe americano size tall and iced please?	Please can I have two caffi americanas size tall and iced please?
Hi, can i have one grande iced decaf americano one extra shot thank you?	Hi, can I have one grand eyes decaf americano one extra shot thank you?
Can I get a tall strawberry yogurt?	Can I get a tool to roll over your gut?
Can I have one youthberry tea size tall with ice and one yuja mint tea size tall with ice?	Can I have one yuleberry tea size tall with ice and one yucca mint tea size tall with ice?

Table 4: Some critical ASR mistakes from whisper-large-v3 on *StarDrinks*.

This is followed by slot filling results from both text and ASR output in Section 5.2. The agent is designed to generate NLU slots from speech input, using whisper-large-v3 for ASR and GPT-4o (Achiam et al., 2023) as the language model.

5.1. ASR results

We present the zero-shot whisper-large-v3 performance in Table 2. We compute WER and CER scores using the HuggingFace evaluate library[2] after normalizing the input using the MMS normalization scripts from Pratap et al. (2024).

We observe that the test splits are challenging for whisper to correctly transcribe: we reach a WER of 9.2% for English and 22.9% for Korean. Indeed, qualitatively, we observe that while whisper is very competent producing fluent output, since its language model did not train with this domain-specific vocabulary, it struggles to produce approximations to these new named entities. Some examples for English are given in Table 4.

Our results highlight the ongoing challenge of adapting ASR systems to previously unseen vocabulary. In this deployment setting, one could argue that the menu is known and could be part of the system's adaptation data for fine-tuning. However, we argue that we should always expect menu changes and the introduction of new items. Therefore, we believe that a more promising open-ended solution for this problem is the research focused on context-biasing and test-time ASR adaptation (Lin et al., 2024; Mittal et al., 2023; Yoon et al., 2024). By releasing our test set, we aim to encourage further research on these directions.

[2] https://github.com/huggingface/evaluate

5.2. Slot filling

We now report on our *drink ordering agent use case* NLU and SLU experiments conducted with GPT-4o as language model. For SLU, we use the ASR model from Section 5.1. We present results for the *StarDrinks* test set in both English and Korean.

In order to build our NLU component, GPT-4o was prompted to perform slot filling either on original transcriptions (NLU) or on automatic transcriptions generated by the ASR model (SLU). This prompt included the NLU schema with all slot types and possible slot values, and it was designed with either no examples (0-shot) or three examples (3-shot) of input and expected output (see Figure 4).[3]

We report results using two metrics: UEM and Slot F1. UEM (Unordered Exact Match) accuracy measures the percentage of utterances for which the entire set of predicted slot-value pairs exactly matches the reference annotation, disregarding the linear order of the pairs (Rubino et al., 2022). It provides a strict end-to-end measure of understanding correctness. Slot F1, instead, measures slot-level performance.

Results for English and Korean across all settings are presented in Table 3. As expected, the highest performance is achieved with gold transcriptions, as the ASR inevitably adds noise to the input of the NLU module. For 0-shot, replacing gold transcripts by whisper-large-v3 results in an UEM accuracy reduction of 11.76 points for English, and 17.96 points for Korean. Surprisingly, this gap is much smaller in the 3-shot setting, being of only 2.75 for English, and 5.08 for Korean. This could hint to the NLU module becoming more robust to noise in this

[3] Additional 6 and 10 shot setups were evaluated on English NLU, but they did not yield clear improvements over the 3-shot setting.

Slot Name	Explanation	Possible Values
ADDING_SHOT	Options for adjusting the number of espresso shots	1 to 6 extra_shot, 1 to 6 less_shot, no_shot
BEAN	Type of coffee bean used	blonde, decaf, half_decaf, regular
CHOCOLATE_AMOUNT	Intensity of chocolate flavor	light, rich
CUSTOMS	Drink specific customizations	extra_mango_juice, in_ice_cup, less_vanilla_cream_base, no_condensed_milk, no_tea
DRIZZLE_AMOUNT	Adjustments to drizzle toppings like caramel or chocolate	extra_chocolate_drizzle, less_caramel_drizzle, less_chocolate_drizzle, more_caramel_drizzle, more_chocolate_drizzle, regular_chocolate_drizzle
DRINK_TYPE	The specific type of beverage being ordered	milk_or_steam_milk, coffee_of_the_day, cafe_americano, blonde_cafe_americano, iced_coffee, mint_blend, youthberry, english_breakfast, chamomile_blend, hibiscus_blend, cafe_latte, jeju_organic_green_tea, decaffeinated_cafe_latte, cold_brew, blonde_cafe_latte, starbucks_double_shot, mango_passion_fruit_blended, espresso_frappuccino, cold_brew_latte, chai_tea_latte, grapefruit_honey_black_tea, signature_chocolate, vanilla_cream_cold_brew, starbucks_dolce_latte, decaffeinated_starbucks_dolce_latte, caramel_macchiato, white_chocolate_mocha, blonde_starbucks_dolce_latte, yuzu_mint_tea,blonde_vanilla_double_shot_macchiato, white_chocolate_mocha_frappuccino, latte_made_with_jeju_organic_matcha, classic_milk_tea,dolce_cold_brew, strawberry_acai_lemonade_starbucks_refresher, pink_drink_with_strawberry_acai_starbucks_refresher, java_chip_frappuccino, mango_dragonfruit_lemonade_starbucks_refresher, decaffeinated_caramel_macchiato, chocolate_cream_chip_frappuccino, cream_frappuccino_made_with_jeju_organic_matcha, caramel_frappuccino, strawberry_delight_yogurt_blended, earl_gray_vanilla_tea_latte, purple_drink_with_mango_dragonfruit_starbucks_refresher
MILK_AMOUNT	Adjustments to the amount of milk	less_milk, more_milk
MILK_TYPE	Type of milk used	low_fat, non_fat, oat, oat_milk, regular_milk, soy
NUMBER	The quantity of drinks to order	1 to 12
SIZE	Drink size options	grande, short, tall, venti
SYRUP_AMOUNT	Adjustments to syrup quantity	no_syrup, no_vanilla_syrup
SYRUP_TYPE	Syrup flavor	coffee, hazelnut, vanilla
TEMPERATURE	Serving temperature	hot, ice
WATER_ICE_AMOUNT	Adjustments to water or ice levels	extra_ice, less_ice, less_water, more_water, no_ice
WHIP_CREAM_AMOUNT	Adjustments to the whipped cream topping	less_espresso_whip, less_whip, more_espresso_whip, more_whip, no_whip, regular_espresso_whip

Table 5: *StarDrinks* slot types, their meaning, and possible values.

```
messages=[
{"role": "system", "content": "You are a natural language understanding expert."},
{"role": "user", "content": f'''Here is a list of JSON dictionaries that describe drink orders of various DRINK_TYPE:
{schema}.

Analyze the sentence: {nlu_input} to output a corresponding list of structured JSON orders that is compliant with
the schema. Make sure the JSON dictionaries of the list are correct JSON objects. Each attribute value in these
dictionaries is atomic, it is not a list. Output only the JSON order list, no input sentence, no explanations.

Here are some samples:

nlu_input: "I would like to order a hot tall Caffe Americano and a cold Venti Chamomile Blend Brewed Tea, please."
output: [{{"DRINK_TYPE": "cafe_americano", "NUMBER": 1, "TEMPERATURE": "HOT", "SIZE": "TALL"}},
        {{"DRINK_TYPE": "chamomile_blend", "NUMBER": 1, "TEMPERATURE": "ICE", "SIZE": "VENTI"}}]

nlu_input: "Could I please have one tall, hot Starbucks Dolce Latte?"
output: [{{"DRINK_TYPE": "starbucks_dolce_latte", "NUMBER": 1, "TEMPERATURE": "HOT", "SIZE": "TALL"}}]

nlu_input: "Could I place an order for one tall Vanilla Cream Cold Brew, two tall pink drink strawberry açai
refresher, and 5 hot tall Hibiscus Tea?"
output: [{{"DRINK_TYPE": "vanilla_cream_cold_brew", "NUMBER": 1, "SIZE": "TALL"}},
        {{"DRINK_TYPE": "pink_drink_with_strawberry_acai_starbucks_refresher", "NUMBER": 2, "SIZE": "TALL"}},
        {{"DRINK_TYPE": "hibiscus_blend", "NUMBER": 5, "TEMPERATURE": "HOT", "SIZE": "TALL"}}] '''
```

Figure 4: An example of prompt for NLU (3-shots, English).

setting, potentially *guessing* or correcting incorrect transcriptions given by the ASR module.

Regarding the 3-shot configuration, we find that it consistently outperformed the 0-shot setting: we observe an NLU UEM accuracy improvement of 15.3 points for English, and 4.07 for Korean. This confirms the benefit of few-shot prompting. Finally, we observe that the Korean results are generally of higher quality than the English ones (+2.77 UEM accuracy points). We believe this could be due to English being a language that allows for a more flexible expression of the items in an order, making this test set slightly more challenging.

Although the best UEM accuracy scores appear

reasonably high, they remain relatively low for practical user-facing applications, where near-perfect understanding is required. In those settings, Substantial amounts of adaptation data, whether collected in-domain or synthetically generated, would likely be required to achieve significant improvements in the agent's performance.

6. Conclusion

In this paper we presented *StarDrinks*, a test set consisting of spontaneously uttered drink orders, their transcripts and NLU slot values. It provides data for speech-to-slots SLU, text-to-slots NLU and speech-to-transcription ASR evaluation in both English and Korean.

To showcase its usefulness, we presented a coffee ordering agent baseline by plugging whisper-large-v3 to GPT-4o. We observe that the ASR module struggles to adapt to unknown named entities, highlighting the necessity of research on test-time adaptation approaches. Regarding NLU/SLU results, we observe that, while the reported UEM accuracy can get as high as 87.06% for English and 89.83% for Korean for a given setting, this performance is still short of the near-perfect requirements for a deployed system.

We hope our test set will provide a realistic and challenging adaptation setting for NLU and SLU models, supporting the development of more robust, context-aware systems capable of handling real-world interactions. The dataset is available for download at https://europe.naverlabs.com/stardrinks.

7. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guan-Ting Lin, Wei Ping Huang, and Hung-yi Lee. 2024. Continual test-time adaptation for end-to-end speech recognition on noisy speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20003–20015, Miami, Florida, USA. Association for Computational Linguistics.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. On robustness and reliability of benchmark-based evaluation of llms. *arXiv preprint arXiv:2509.04013*.

Ashish Mittal, Sunita Sarawagi, Preethi Jyothi, George Saon, and Gakuto Kurata. 2023. Speech-enriched memory for inference-time adaptation of ASR models to word dictionaries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14820–14835, Singapore. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

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*, pages 28492–28518. PMLR.

Melanie Rubino, Nicolas Guenon des Mesnards, Uday Shah, Nanjiang Jiang, Weiqi Sun, and Konstantine Arkoudas. 2022. Cross-TOP: Zero-Shot Cross-Schema Task-Oriented Parsing. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 48–60, Hybrid. Association for Computational Linguistics.

Fangyun Wei, Xi Chen, and Lin Luo. 2024. Rethinking generative large language model evaluation for semantic comprehension. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52525–52558. PMLR.

Lang Xiong, Nishant Bhargava, Wesley Chang, Jianhang Hong, Haihao Liu, and Kevin Zhu. 2025. Stealtheval: A probe-rewrite-evaluate workflow for reliable benchmarks. *arXiv preprint arXiv:2509.00591*.

Eunseop Yoon, Hee Suk Yoon, John Harvill, Mark Hasegawa-Johnson, and Chang D. Yoo. 2024. LI-TTA: Language Informed Test-Time Adaptation for Automatic Speech Recognition. In *Interspeech 2024*, pages 3490–3494.

8. Language Resource References

Melanie A. Rubino, Nicolas Guenon des mesnards, Uday Shah, Nanjiang Jiang, Weiqi Sun, and Konstantine Arkoudas. 2022. Cross-TOP: Zero-shot cross-schema task-oriented parsing. In *Proceedings of the Third Workshop on Deep Learning*

for Low-Resource Natural Language Processing, pages 48–60, Hybrid. Association for Computational Linguistics.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, Verena Rieser, and Oliver Lemon. 2020. Slurp: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262.

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1478–1482.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Praveen Gupta, Anurag Gupta, et al. 2020. Speech intent recognition for smart assistants: A review. In *Proceedings of the IEEE International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 489–495.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop*, pages 96–101. ACL.

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

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

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, Robin Algayres, Tu Ahn Nguyen, Emmanuel Dupoux, Luke Zettlemoyer, and Abdelrahman Mohamed. 2023. Stop: A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 991–998.