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

★ A realistic test set for SLU evaluation in a drink ordering scenario

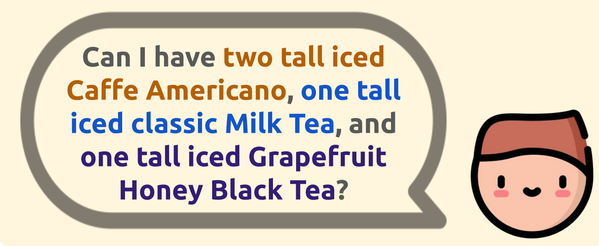
Motivation: Current LLM and speech assistant evaluations rely on controlled setups that do not reflect the variability of real user requests.

Real-world task focus: Drink ordering with diverse named entities, drink types, sizes, customizations, and brand-specific terminology.

Our dataset: English-Korean test set with speech recordings, transcriptions, and annotated NLU slots.

Supports ASR (speech, transcription),
SLU (speech, slots), and NLU (text, slots) evaluation.

Speech Features and Transcription



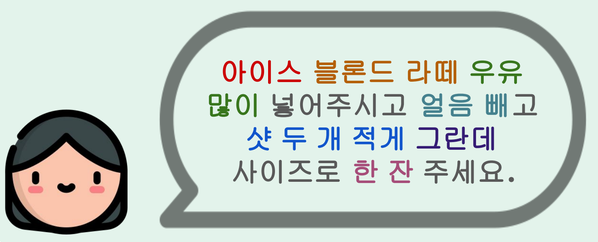
Structured Data

{"DRINK_TYPE": "cafe_americano", "NUMBER": "2", "SIZE": "tall", "TEMPERATURE": "ice"}

{"DRINK_TYPE": "classic_milk_tea", "NUMBER": "1", "SIZE": "tall", "TEMPERATURE": "ice"}

{"DRINK_TYPE": "grapefruit_honey_black_tea", "NUMBER": "1", "SIZE": "tall", "TEMPERATURE": "ice"}

Speech Features and Transcription

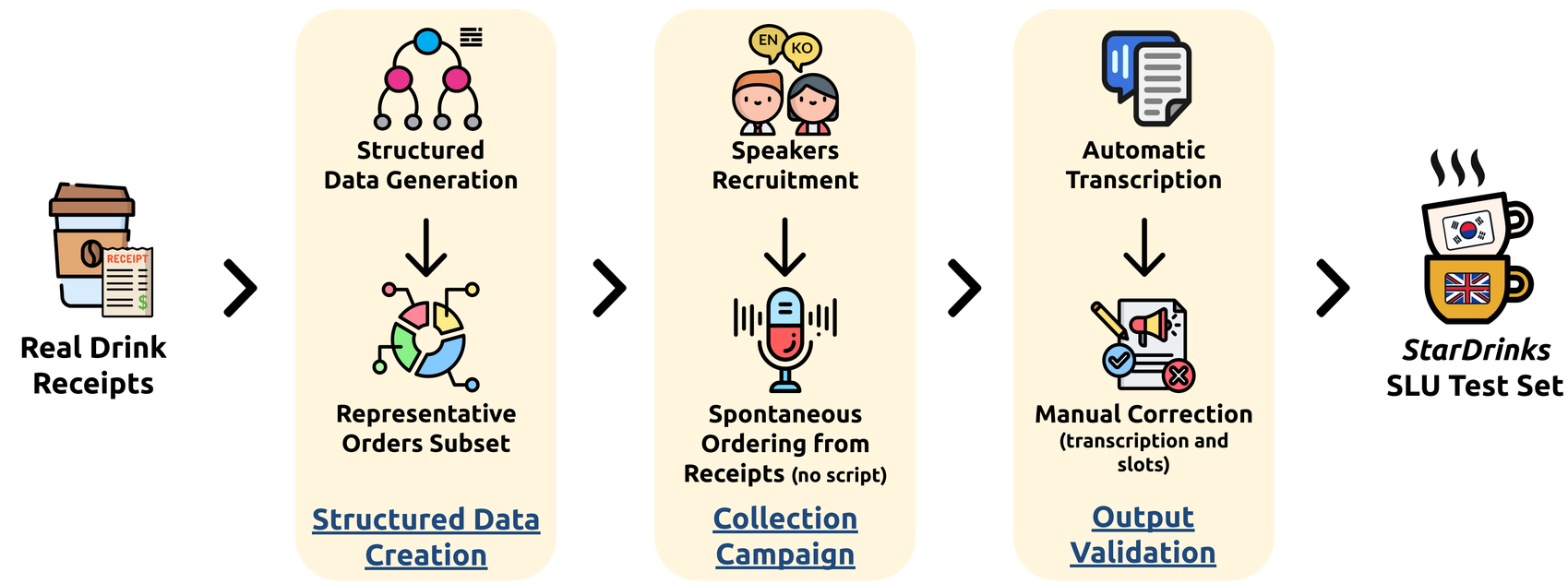


Structured Data

{ "TEMPERATURE": "ice",
"DRINK_TYPE": "blonde_cafe_latte",
"MILK_AMOUNT": "more_milk",
"WATER_ICE_AMOUNT": "no_ice",
"ADDING_SHOT": "2 less_shot",
"SIZE": "grande",
"NUMBER": "1"}

★ Dataset creation:

From real drink receipts (structured information) from a popular coffee chain, we build a representative collection of orders in natural language (speech and text).



★ Use case: coffee ordering agent

We showcase our test set in a coffee ordering agent made of an ASR component (whisper-large-v3) and an NLU component (ChatGPT-4o, in few shot setting).

- **ASR**: The ASR module struggles to adapt to unknown named entities, **highlighting the necessity of research on test-time adaptation approaches.**

	WER	CER
English	9.2	3.6
Korean	22.9	7.3

Table 1: ASR results

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?

Table 2: Some critical ASR mistakes from whisper-large-v3 on StarDrinks.

- **NLU/SLU**: While the reported Unordered Exact Match (UEM) accuracy can get as high as 87.06% for English and 89.83% for Korean, **this performance is still short of the near-perfect requirements for a deployed system.**

Configuration	ASR Model	EN (UEM %)	EN (Slot F1 %)	KO (UEM %)	KO (Slot F1 %)
Gold Trans. + 3-shots (NLU)	None	**87.0**	**98.0**	**89.8**	**98.7**
Gold Trans. + 0-shot (NLU)	None	71.7	94.5	85.7	97.7
ASR + 3-shots (SLU)	Whisper	84.3	97.3	84.7	97.4
ASR + 0-shots (SLU)	Whisper	60.0	89.96	67.8	93.7

Table 3: NLU/SLU results on the StarDrinks English and Korean test sets.

Paper:

Dataset:

Contact: dl_stardrinks@naverlabs.com