SpokenWOZ: A Large-Scale Speech-Text Dataset for Spoken Task-Oriented Dialogue Agents

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang and Yongbin Li

Task-oriented Dialogue Datasets

達摩院
ALIBABA DAMO ACADEMY

🔴 Human to Human

🔵 Human to Machine

Towards Spoken Dialogue

2014
DSTC2
citation: 524
SIGDIAL2014
amazon

2017
KVRET
citation: 302
SIGDIAL2017

2018
M2M
citation :141
NAACL2018
Google AI

2018
MultiWOZ1.0
citation :111
ACL2018
UNIVERSITY OF CAMBRIDGE

MultiWOZ 2.0
citation :665
EMNLP2018 best paper
UNIVERSITY OF CAMBRIDGE

MultiWOZ 2.1
Citation:123
LREC 2020
amazon

MultiWOZ 2.2
Citation:92
ACL2020 workshop
Google AI

MultiWOZ

2019
SGD
citation: 236
AAAI2020
Google AI

2021
ABCD
citation: 13
NAACL2021
COLUMBIA UNIVERSITY

2022
EVI
citation: 0
NAACL2022
PolyAI

2022
DSTC11
citation: 0
Proposal
Google AI

2022
SSTOD
citation: 0
ACL2022
達摩院
ALIBABA DAMO ACADEMY

time

ii). The trend of TOD datasets is towards to spoken setting

i). MultiWOZ datasets leading the development of task-oriented over past four years

Limitation of MultiWOZ

Text Modality Conversations are collected through text conversations in web chat rooms. The entire conversation is turn-by-turn.

Incomplete modeling ASR and TTS parts are not modeled. There is a lack of information after ASR, and errors will accumulate in subsequent modules.

Models trained through text dialogue have a large gap with spoken dialogue scenarios, which will increase the difficulty of application.

SpokenWOZ Dialogue Corpus

We introduce a large-scale speech-text TOD dataset named **SpokenWOZ**, which contains more than **203K** annotated utterances, **5,700** dialogues, and the associated **249 hours** of audios.

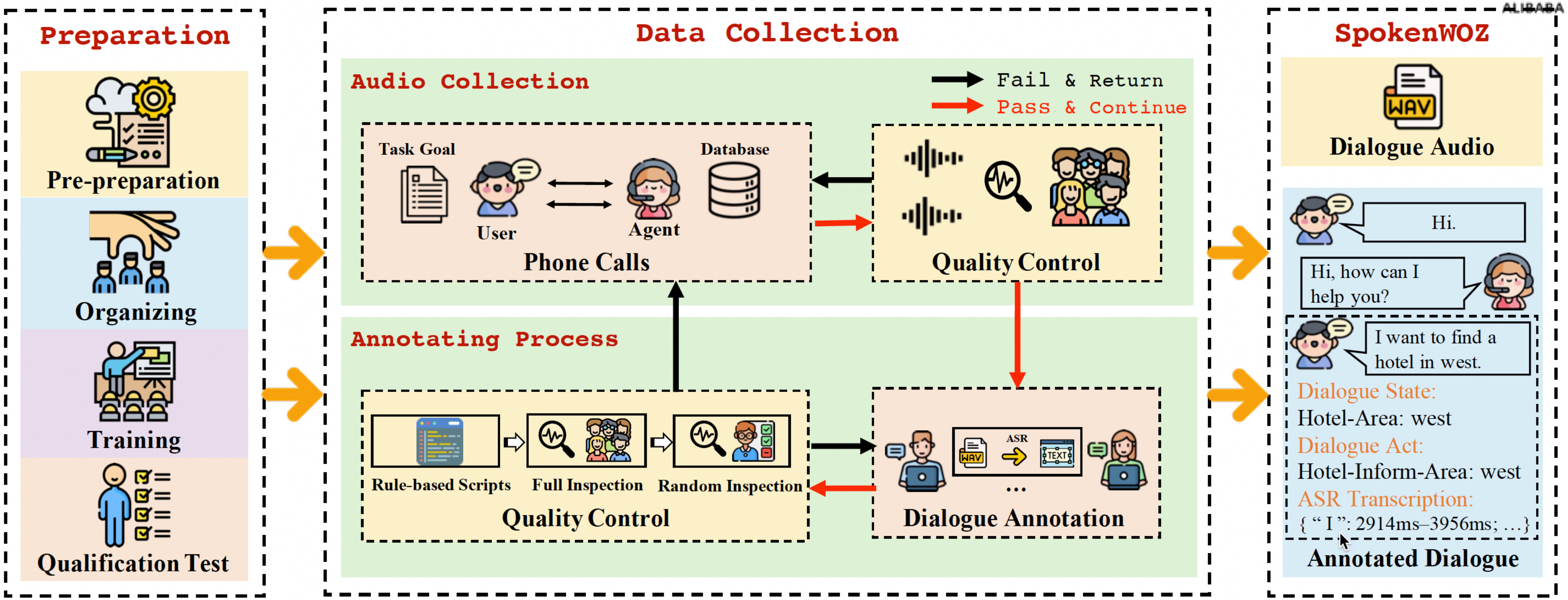
Dataset	Train	Dev	Test
Audio Hours	183	22	44
Dialogues	4,200	500	1000
Turns	149,126	18,384	35,564
Tokens	1,672,984	204,644	396,933
Avg. Turns	35.50	36.77	35.56
Avg. Tokens	11.21	11.13	11.16

Statistics of SpokenWOZ

Metric	DSTC2	KVRET	M2M	MultiWOZ	ABCD	DSTC10	**SpokenWOZ***
Dialogues	1,612	2,425	1,500	**8,438**	8,034	107	5,700
Turns	23,354	12,732	14,796	115,424	177,407	2,292	**203,074**
Domains	Single	**Multi**	**Multi**	**Multi**	**Multi**	**Multi**	**Multi**
Collection	H2M	**H2H**	M2M	**H2H**	**H2H**	**H2H**	**H2H**
Type	**Spoken**	Written	Written	Written	Written	**Spoken**	**Spoken**
Audio	✓	✗	✗	✗	✗	✗	✓
Cross-turn Slot	✗	✗	✗	✗	✗	✗	✓
Reasoning Slot	✗	✗	✗	✗	✗	✗	✓

Dataset statistics of SpokenWOZ and existing TOD datasets *: SpokenWOZ contains 4200 dialogues in the training set.

SpokenWOZ-Construction



Spoken dialogue collection

Organized **250 participants** to generate **5,700 conversations** by making phone calls

One participant assumes the role of the user and asks questions, Another participant plays an Agent to complete the user's needs and tasks and answer the user's questions.

Data annotation

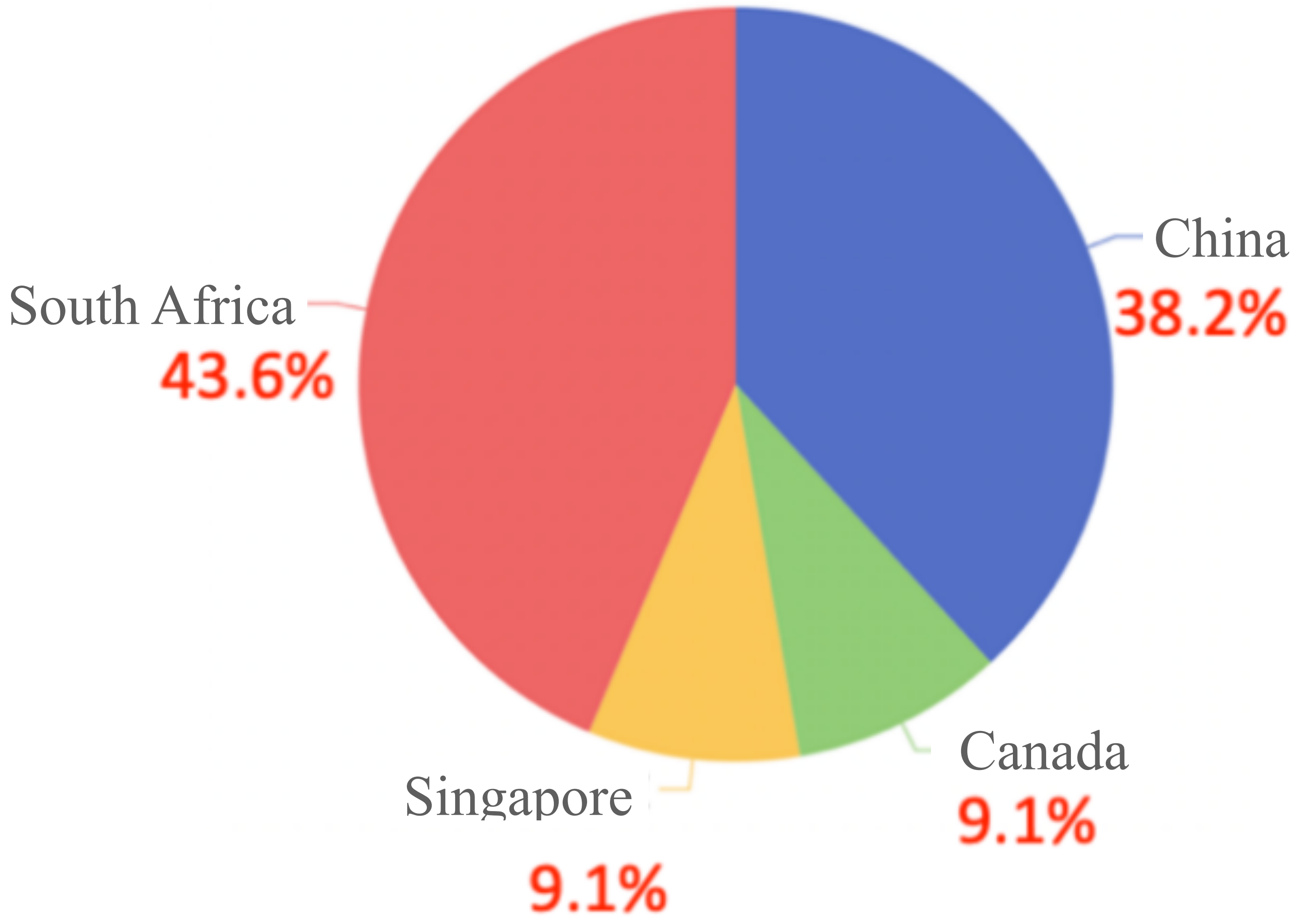
Trained **15 professional annotators** and use **multiple checks** to ensure annotation quality

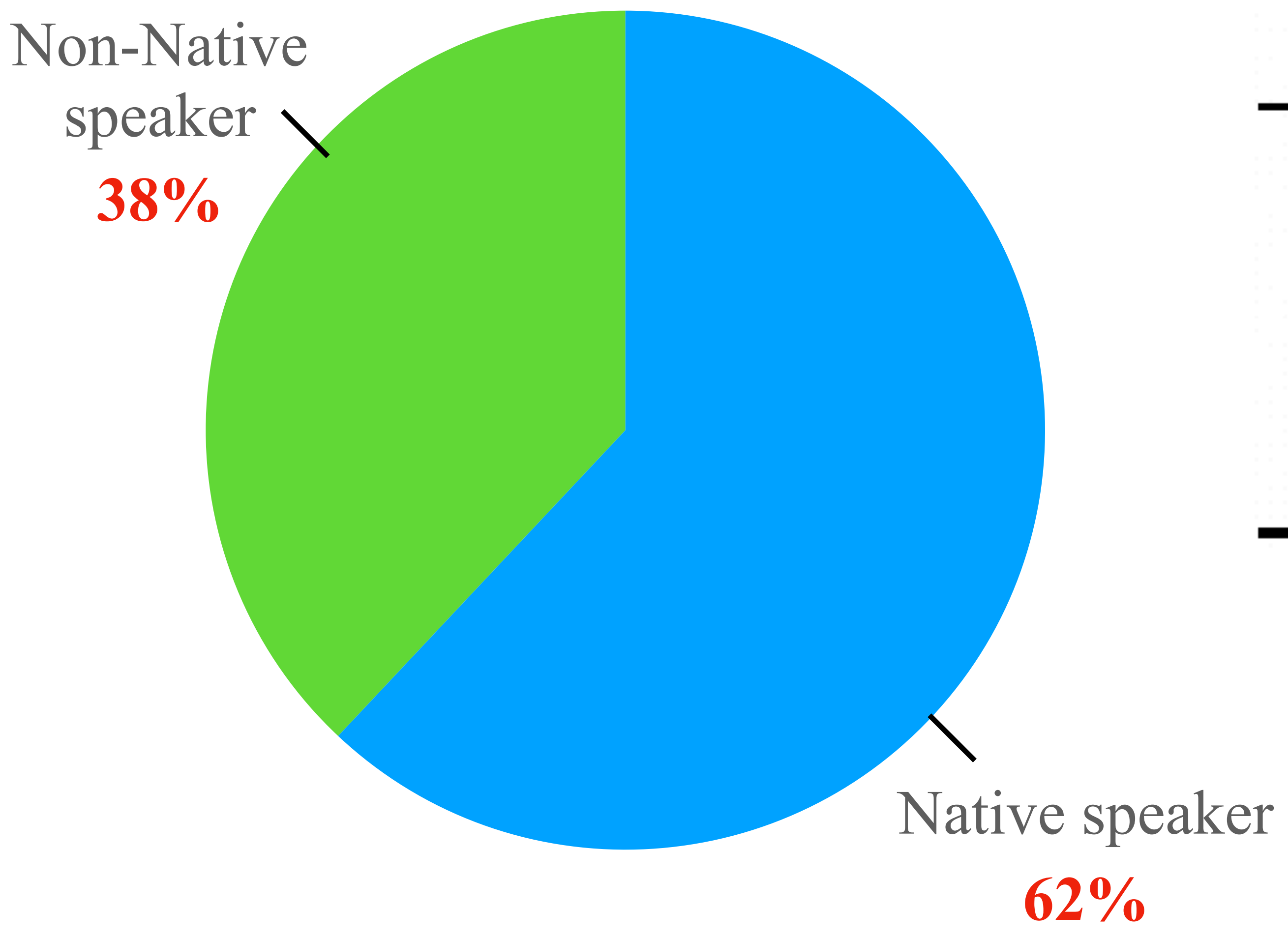
Inherited and expanded the annotation specifications of MultiWOZ and added dialogue action annotations such as Backchannel

SpokenWOZ-Challenges

Diversified Speech

We construct dialogue content using speakers in four countries and regions: mainland China, South Africa, Singapore, and Canada. The proportion is as follows:





Country	Dialogues	Percentage	People	Percentage
Canada	500	8.77%	60	24%
Singapore	500	8.77%	40	16%
China	2100	36.84%	30	12%
South Africa	2600	45.61%	120	48%

The origins diversity of SpokenWOZ. Participants come from four different countries to improve the diversity of spoken conversations

Completely Colloquial

The conversation is generated through real-time voice calls, which is obviously different from the characteristics of textual dialogue. For example, when making a reservation at a restaurant:

Textual expression in MultiWOZ	Spoken expression in SpokenWOZ
I Would like a cheap restaurant in the north area.	I would like a restaurant, hmmm, cheap one please, meanwhile in the south area, sorry, in the north.

SpokenWOZ-Challenges

Cross-Turn and Reasoning slots

Spoken dialogue also brings two new types of slot : multi-turn slots (cross-turn slots) and reasoning that require fine-grained memory slots (reasoning slot). These two slots are due to new challenges in memory and reasoning caused by incomplete spoken language and indirect expression.

Cross-Turn Slot

👤: Oh, my id number is **5 2 5** ~~8~~ (*“8” is missed by the ASR tool, but appears in the audio*).
(Dialog State: id_number = 5258)
🤖: So it’s 5 2 5 8.
👤: Yes. and then **5 7 6 3**.
(Dialog State: id_number = 52585763)
🤖: 5 7 6 3.
👤: And then **7 5 2 5**.
(Dialog State: id_number = 525857637525)
🤖: 7 5 2 5.
👤: I’m sorry, **7 5 to 4**.
(Dialog State: id_number = 525857637524)
🤖: Yes. okay, so it’s 7524.
👤: And then **double 9 0 3**.
(Dialog State: id_number = 5258576375249903)

Reasoning Slot

🤖: Uh, yes, and on which day please?
👤: Oh. yeah. And I think **today is Friday**, right. Uh we will be there **tomorrow**.
(Dialog State: Bookday = Saturday)

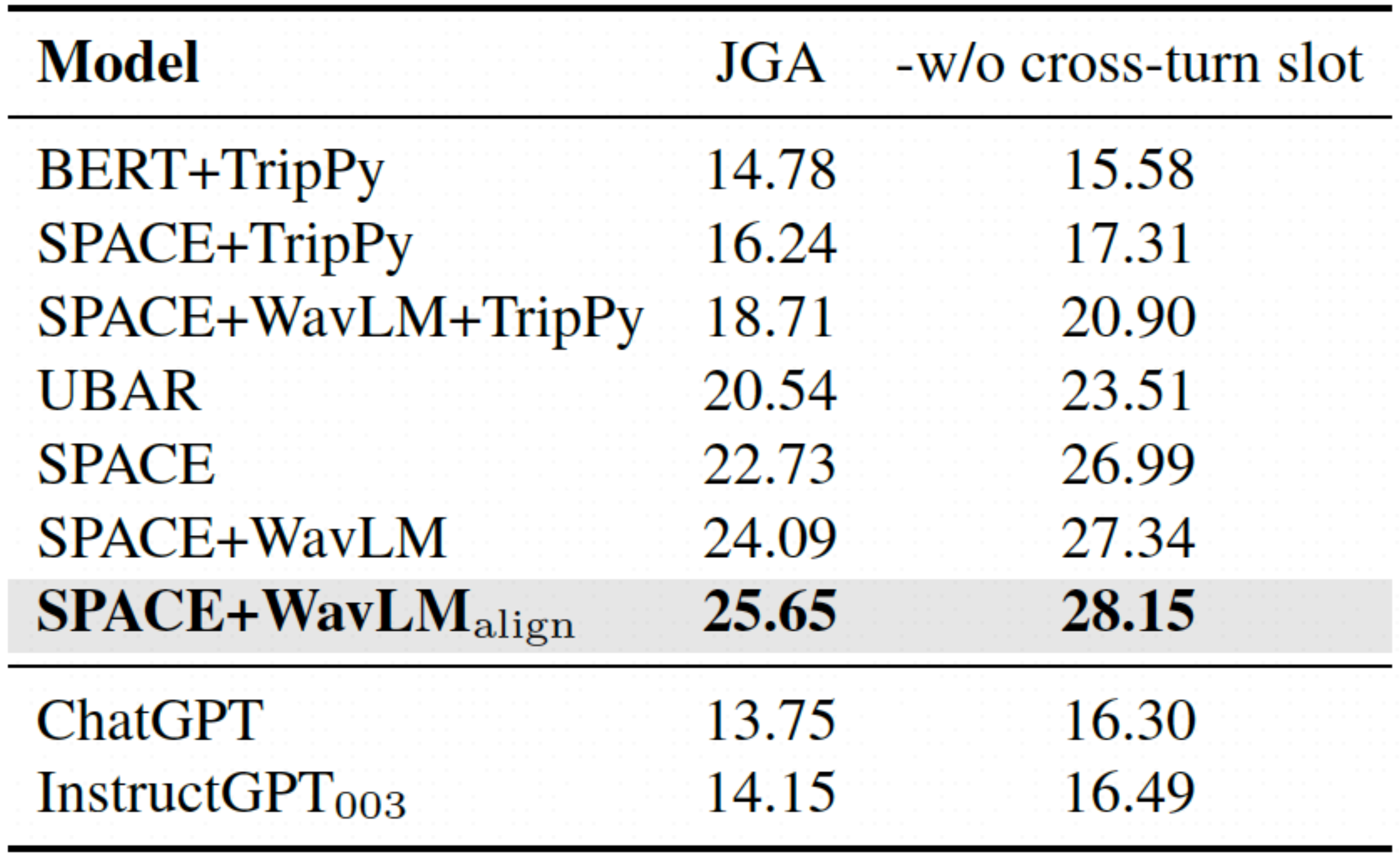
👤: Yeah. Yeah. Uh, can you book it for **me, my parents and my grandparents?**
🤖: Okay, so it’s **five** people in total.

👤: Yes, ma’am, the restaurant should serve **sushi** and should be in the center.
(Dialog State: Restaurant-Type = Japanese; Restaurant-Area = centre)
🤖: Just to confirm that the restaurants are serving Japanese food in the centre.
👤: That’s correct, man.

SpokenWOZ-Experiments

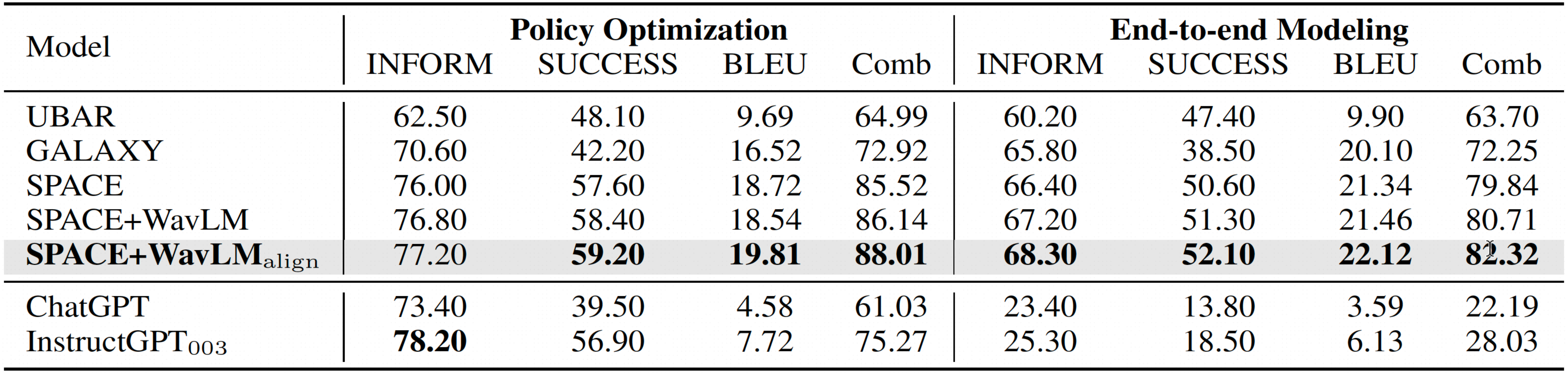
We considered various types of Dialogue Agents of different sizes, including single-modal models below 1B (UBAR, SPACE, etc.) and dual-modal models of sound and meaning (SPACE+WavLM), ChatGPT (gpt-3.5-turbo), and 175B InstructGPT$_{003}$(text-davinci-003)

In terms of evaluation indicators, the joint accuracy **JGA** (the proportion of all slots in the current round that are correct) is used for the DST task, and **Inform**, **Success** (completion rate) and **BLEU** are used for the reply generation task.



Model	JGA	-w/o cross-turn slot
BERT+TripPy	14.78	15.58
SPACE+TripPy	16.24	17.31
SPACE+WavLM+TripPy	18.71	20.90
UBAR	20.54	23.51
SPACE	22.73	26.99
SPACE+WavLM	24.09	27.34
SPACE+WavLM$_{align}$	**25.65**	**28.15**
ChatGPT	13.75	16.30
InstructGPT$_{003}$	14.15	16.49

DST experimental results with different methods



Model	Policy Optimization				End-to-end Modeling			
	INFORM	SUCCESS	BLEU	Comb	INFORM	SUCCESS	BLEU	Comb
UBAR	62.50	48.10	9.69	64.99	60.20	47.40	9.90	63.70
GALAXY	70.60	42.20	16.52	72.92	65.80	38.50	20.10	72.25
SPACE	76.00	57.60	18.72	85.52	66.40	50.60	21.34	79.84
SPACE+WavLM	76.80	58.40	18.54	86.14	67.20	51.30	21.46	80.71
SPACE+WavLM$_{align}$	77.20	**59.20**	**19.81**	**88.01**	**68.30**	**52.10**	**22.12**	**82.32**
ChatGPT	73.40	39.50	4.58	61.03	23.40	13.80	3.59	22.19
InstructGPT$_{003}$	**78.20**	56.90	7.72	75.27	25.30	18.50	6.13	28.03

Policy Optimization and End-to-end Modeling experimental results

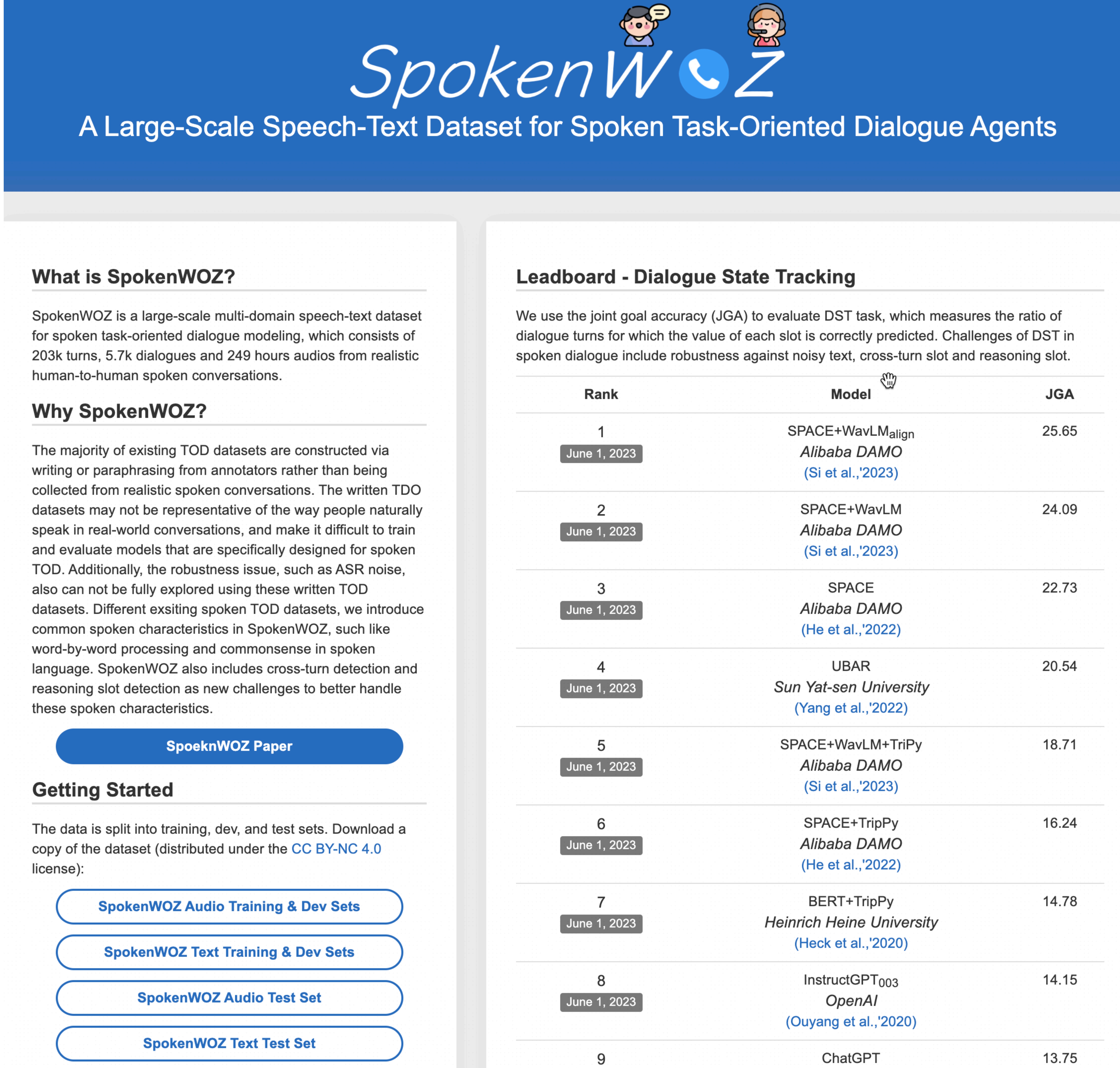
SpokenWOZ-Resource

Code

https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/spokenwoz

Leadboard

https://spokenwoz.github.io/SpokenWOZ-github.io/



A Large-Scale Speech-Text Dataset for Spoken Task-Oriented Dialogue Agents

What is SpokenWOZ?

SpokenWOZ is a large-scale multi-domain speech-text dataset for spoken task-oriented dialogue modeling, which consists of 203k turns, 5.7k dialogues and 249 hours audios from realistic human-to-human spoken conversations.

Why SpokenWOZ?

The majority of existing TOD datasets are constructed via writing or paraphrasing from annotators rather than being collected from realistic spoken conversations. The written TDO datasets may not be representative of the way people naturally speak in real-world conversations, and make it difficult to train and evaluate models that are specifically designed for spoken TOD. Additionally, the robustness issue, such as ASR noise, also can not be fully explored using these written TOD datasets. Different exsiting spoken TOD datasets, we introduce common spoken characteristics in SpokenWOZ, such like word-by-word processing and commonsense in spoken language. SpokenWOZ also includes cross-turn detection and reasoning slot detection as new challenges to better handle these spoken characteristics.

Getting Started

The data is split into training, dev, and test sets. Download a copy of the dataset (distributed under the CC BY-NC 4.0 license):

SpokenWOZ Audio Training & Dev Sets
SpokenWOZ Text Training & Dev Sets
SpokenWOZ Audio Test Set
SpokenWOZ Text Test Set

Leadboard - Dialogue State Tracking

We use the joint goal accuracy (JGA) to evaluate DST task, which measures the ratio of dialogue turns for which the value of each slot is correctly predicted. Challenges of DST in spoken dialogue include robustness against noisy text, cross-turn slot and reasoning slot.

Rank	Model	JGA
1 June 1, 2023	SPACE+WavLM$_{align}$ *Alibaba DAMO* (Si et al.,'2023)	25.65
2 June 1, 2023	SPACE+WavLM *Alibaba DAMO* (Si et al.,'2023)	24.09
3 June 1, 2023	SPACE *Alibaba DAMO* (He et al.,'2022)	22.73
4 June 1, 2023	UBAR *Sun Yat-sen University* (Yang et al.,'2022)	20.54
5 June 1, 2023	SPACE+WavLM+TriPy *Alibaba DAMO* (Si et al.,'2023)	18.71
6 June 1, 2023	SPACE+TripPy *Alibaba DAMO* (He et al.,'2022)	16.24
7 June 1, 2023	BERT+TripPy *Heinrich Heine University* (Heck et al.,'2020)	14.78
8 June 1, 2023	InstructGPT$_{003}$ *OpenAI* (Ouyang et al.,'2020)	14.15
9	ChatGPT	13.75

Thanks