

CMPA-SLU: A Benchmark for Chinese Mobile Personal Assistant Spoken Language Understanding

Di Wu^{1*}, Zhongqiu Li^{1*}, Zetao Zhu^{1,2,3*}, Sishi Xiong¹, Ruiyu Fang¹, Shuangyong Song^{1†}

¹Xingchen AGI Lab, China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd

²Beijing University of Posts and Telecommunications

³School of Artificial Intelligence

Abstract

China has over 1.116 billion mobile Internet users, whose demands continue to evolve rapidly. However, existing Chinese spoken language understanding (SLU) benchmarks are outdated and cover limited domains, failing to reflect real-world mobile assistant scenarios. To this end, we introduce CMPA-SLU¹, a Chinese SLU dataset for mobile personal assistants collected from real-world applications and annotated with rigorous quality control. CMPA-SLU contains 50 domains, 247 intents, and 74 slot types, expanding the diversity of existing Chinese SLU benchmarks. To better simulate practical deployment scenarios, we further introduce a Partial Category setting, where candidate labels are dynamically filtered using a Bidirectional Guided Embedding (BGE)-based retrieval strategy. We further adopt a label mapping strategy for joint intent and slot prediction within a single generation process. We evaluate multiple large language models (LLMs) under both Full Category and Partial Category settings. Experimental results show that the Partial Category setting improves performance by reducing the candidate label space.

1 Introduction

China has over 1.116 billion mobile Internet users (China Internet Network Information Center, 2025), whose demands continue to evolve rapidly. With the increasing capabilities of smartphones, spoken language understanding (SLU), including intent detection and slot filling (Qin et al., 2021b; Weld et al., 2022), has become a fundamental component of mobile personal assistants.

However, existing Chinese SLU datasets are outdated and suffer from limited domain coverage, as shown in Table 1. Widely used benchmarks, including CAIS (Liu et al., 2019), SMP2019, and

Dataset	Domain	Intent	Slot
CAIS	1	14	24
SMP2019	29	23	59
SMP2020	45	59	80
MASSIVE-ZH	18	60	55
CMPA-SLU (Ours)	50	247	74

Table 1: Statistics of domains, intents, and slot types in Chinese SLU datasets.

SMP2020 (Xie et al., 2023), were released over six years ago. Although MASSIVE-ZH (FitzGerald et al., 2023) was released in 2023, it is translated from SLURP (Bastianelli et al., 2020) (2020), and thus does not reflect native Chinese user behavior. Consequently, existing datasets exhibit limited diversity in domains, intents, and slots, and fail to capture evolving real-world user needs. Moreover, their label schemas are insufficient for emerging applications, such as “Turn Off Video AI Subtitles”.

To this end, we construct CMPA-SLU, a Chinese SLU dataset for mobile personal assistants. All utterances in the dataset were recently collected from real-world applications, followed by anonymization, deduplication, and manual annotation to ensure data quality. It contains 50 domains, 247 intents, and 74 slot types, surpassing existing Chinese SLU datasets in scale.

To better simulate practical deployment scenarios, we further introduce a Partial Category setting, where candidate intent and slot labels are dynamically filtered using a Bidirectional Guided Embedding (BGE)-based retrieval strategy (Xiao et al., 2024). In addition, we adopt an intent-slot schema mapping (IS-Map) strategy to support joint intent and slot prediction within a unified generation process. Finally, we evaluate multiple large language models (LLMs) (He et al., 2024; Liu et al., 2025; Wang et al., 2025) on CMPA-SLU under both Full Category and Partial Category settings.

*Equal contribution.

† Corresponding author.

¹<https://github.com/XingChen-AGI/CMPA-SLU>

2 Dataset Construction Process

2.1 Data Preprocessing

After data collection, regular-expression-based anonymization is applied, followed by deduplication to remove exact duplicate utterances, thereby reducing the risk of data leakage during dataset splitting. During manual annotation of intents and slots, annotators review each utterance individually and perform additional anonymization when necessary. Utterances containing sensitive information are discarded to ensure user privacy.

2.2 Category Predefinition

We define an initial set of intent and slot labels based on real-world mobile personal assistant applications. During annotation, annotators and data reviewers may encounter previously unseen or undefined labels. Such utterances are temporarily set aside and later revisited for final label confirmation, with the label schema updated when necessary.

2.3 Annotation Procedure

We employ eight native annotators with prior experience in SLU data annotation to annotate and review the dataset. The annotation process consists of three stages:

Stage 1: First-round annotation. The dataset is divided into six subsets, each assigned to one annotator. Annotators are instructed to flag utterances whose intents or slots fall outside the candidate label set and forward them to the inspectors (Stage 3).

Stage 2: Second-round verification. Two senior reviewers independently verify the first-round annotations, each responsible for three subsets. Identified errors are corrected at this stage.

Stage 3: Third-round sampling inspection. A 10% random sample is inspected by two PhD students specializing in SLU. In addition to reviewing sampled data, they address flagged utterances by refining annotation guidelines or updating intent and slot categories. Subsets that do not meet quality standards are returned to Stage 1 for re-annotation until requirements are satisfied.

After multiple iterations, the final numbers of domains, intents, and slot types are summarized in Table 1, and the frequency distributions across intent and slot categories are illustrated in Figure 1.

As shown in Table 1 and Figure 1, the CMPA-SLU dataset exhibits substantial class diversity and a pronounced long-tail distribution. Specifically, 40

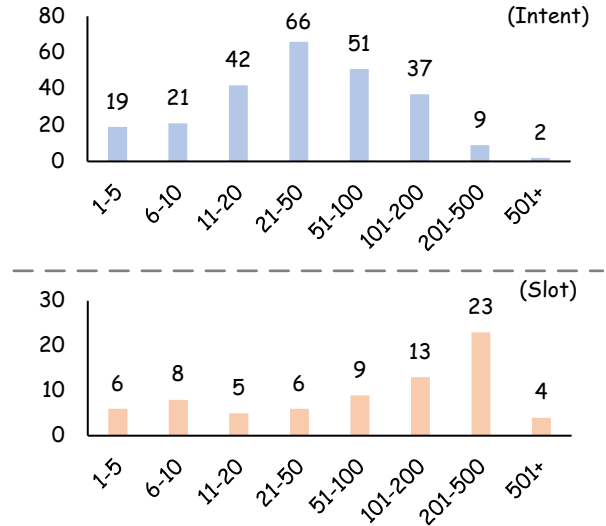


Figure 1: Frequency distributions of sample counts across different categories.

intent classes contain only 1–10 instances, accounting for 16.19% of intent classes but only 1.49% of total samples. Similarly, 14 slot types (18.92%) account for only 0.05% of all samples. Additional statistics are provided in Appendix A.

In real-world scenarios, some user utterances do not express a clear intent or fully specify user needs, such as “Never mind” and “Why is it like this?”. We retain these cases and annotate them as *Other intent*. This design better reflects the complexity of real-world SLU scenarios and distinguishes our dataset from existing SLU benchmarks in Chinese and other languages, which typically do not explicitly consider such cases (Van Der Goot et al., 2021; Aimaiti et al., 2024; Wu et al., 2026). Furthermore, since the data is collected from recent real-world mobile applications, the labels are more aligned with the usage habits of Chinese mobile users compared to existing SLU datasets. Examples are provided in Appendix A.

2.4 Candidate Category Pre-Extraction

As shown in Table 1 and Figure 1, CMPA-SLU contains a large and diverse set of intent and slot categories. In practical applications, pre-filtering candidate intents and slots is commonly adopted to reduce computational overhead, especially for large-scale LLMs. Therefore, we construct a candidate label set for each sample using a three-stage pre-filtering strategy.

Stage 1: Label Index Construction. We randomly sample 2,000 utterances from CMPA-SLU to train a BGE-based embedding model that maps

utterances, intents, and slot labels into a shared representation space. Based on the learned embeddings, we construct a label index for efficient candidate retrieval.

Stage 2: Candidate Retrieval. Based on the constructed index, we retrieve the top-20 intent candidates with the highest similarity to each utterance. The top-20 candidate set achieves 100% recall, indicating that the retrieval module successfully includes the ground-truth intent for every utterance. The retrieved intent labels, together with their associated slot types, constitute the initial candidate set. In real-world scenarios, however, the number of candidate categories can vary across utterances, while models trained under a fixed-category assumption may suffer from performance degradation when the candidate set size changes. To address this issue, we further introduce a candidate refinement stage.

Stage 3: Candidate Confirmation. To simulate variations in candidate set sizes in practical scenarios, we randomly sample X intent categories from the retrieved candidates to construct the final candidate set. To ensure completeness, the ground-truth intent and its associated slot labels are added when absent. In addition, we include the “*Other Intent*” label to support unseen or out-of-scope intents in open-set SLU scenarios.

Under the Partial Category setting, the average numbers of candidate intent categories and slot types per sample are 11.51 and 6.57, respectively.

2.5 Dataset Splitting

We consider two evaluation settings for CMPA-SLU: Full Category and Partial Category. In the Full Category setting, models have access to the complete label space, whereas in the Partial Category setting, only a subset of candidate labels is provided. All other experimental settings remain identical. The dataset is split into training, validation, and test sets with a ratio of 6:2:2, resulting in 4,685 training samples, 1,550 validation samples, and 1,585 test samples.

3 Method

3.1 Task Formulation

We formulate the SLU task as a question answering (QA) problem (Wu et al., 2022; Song et al., 2024; Xiong et al., 2024; Wu et al., 2025a; Jiang et al., 2026). Given an utterance U and a task prompt P , the model predicts the corresponding intent and

slot-level semantic structure:

$$O_I, O_S = M(P, U \mid \mathcal{L}), \quad (1)$$

where $\mathcal{L}$ denotes the candidate label set, including both intent labels and slot types. O_I represents the predicted intent label, and O_S denotes the predicted slot annotations with their corresponding entities. The candidate label set $\mathcal{L}$ can vary across different settings, and is either fully available in the Full Category setting or dynamically constructed in the Partial Category setting.

3.2 Prompt Template

We design a structured prompt template consisting of five components: role definition, task description, candidate label set, output format specification, and utterance, as illustrated in Figure 2. The prompt structure remains identical across settings, with the difference being the candidate label set.

```

#### Role Definition:
You are an assistant for spoken language understanding, equipped with intent detection and slot filling capabilities.

#### Task Description:
Given an input utterance, you are required to identify the underlying intent and extract the corresponding slot-value pairs.

#### Candidate:
The candidate intent and slot categories are provided as follows: {...}.

#### Format:
Intent: [Slot1: Value1, Slot2: Value2, ...]

#### Utterance:
[...]
```

Figure 2: Prompt Template. Detailed candidate lists and utterances are omitted for brevity.

To support joint intent and slot prediction, we incorporate the predefined mappings between intents and their associated slot types into the prompt. Specifically, for each candidate intent, its corresponding slot types are provided as a paired set of candidate labels, explicitly conveying the intent-slot correspondence to the model. Unlike existing methods that treat intent and slot labels independently (Wu et al., 2022; Xing et al., 2024; Qin et al., 2025), our approach explicitly exploits their correspondence during candidate construction and generation. In addition, the candidate labels are consistently aligned with the predefined output format to facilitate structured and reliable generation.

Model	Full Category					Partial Category				
	I(Acc)	I(F1)	S(F1)	O(Acc)	PFL0Ps	I(Acc)	I(F1)	S(F1)	O(Acc)	PFL0Ps
GPT-4o	84.16	64.98	71.43	73.75	-	87.07	72.78	72.01	77.10	-
GPT-5	87.44	72.83	73.26	77.29	-	89.97	72.99	74.59	78.80	-
Qwen2.5-7B	95.96	92.14	86.61	89.40	1143.33	96.03	91.77	86.48	89.31	125.64
LLaMA3.1-8B	95.39	91.20	85.89	89.02	1380.56	96.40	91.70	85.33	89.09	141.61
Deepseek-r1-8B	94.95	90.83	85.45	88.14	1378.95	95.21	88.16	84.91	87.32	139.83
Mistral-7B	95.71	89.37	86.55	90.28	1645.16	97.03	93.68	89.90	91.17	163.75

Table 2: Results of different LLMs on CMPA-SLU under both Full Category and Partial Category settings.

4 Experiment

4.1 Experimental Setup

We evaluate GPT-4o (Hurst et al., 2024), GPT-5 (Leon, 2025; Bubeck et al., 2025), Qwen2.5-7B (Hui et al., 2024), LLaMA3.1-8B (Grattafiori et al., 2024), DeepSeek-R1-8B (Guo et al., 2025), and Mistral-7B (Jiang et al., 2023a) on CMPA-SLU. GPT-4o and GPT-5 are evaluated using prompt-based inference. Each open-source model is fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Detailed experimental settings are provided in Appendix B.

4.2 Results and Analysis of Different Settings

Table 2 reports the performance of multiple LLMs on CMPA-SLU under the Full Category and Partial Category settings.

1. GPT-4o and GPT-5 perform worse than several LoRA-tuned open-source LLMs under both settings, indicating that CMPA-SLU remains challenging even for strong models.

2. Compared with the Full Category setting, the Partial Category setting substantially reduces computational overhead as the candidate intent and slot spaces are relatively constrained, lowering the number of categories that the model needs to distinguish. This reduction in the prediction space not only improves efficiency but also enhances accuracy, as it reduces prediction uncertainty and allows the model to focus more on semantic matching within the candidate set. Most LLMs achieve better performance across all four metrics under the Partial Category setting, suggesting that a smaller candidate label space facilitates more precise intent and slot prediction. Qualitative examples are shown in Figure 3, where Mistral-7B produces incorrect predictions under the Full Category setting but correctly identifies both intent and slots under the Partial Category setting.

(a)	Intent: Stop Playback
	Slot: Stop Playback
Utterance	I don't feel like listening to music today.
	Let's turn off NetEase Cloud Music.
(b)	Intent: Close App
	Slot: App: NetEase Cloud Music

Figure 3: Comparison of Mistral-7B predictions under the Full Category (a) and Partial Category (b) settings. Red indicates incorrect predictions, while green indicates correct predictions.

3. Figure 4 compares the F1 scores of Mistral-7B across different intent and slot categories under the two settings. The Partial Category setting achieves more balanced performance and outperforms the Full Category setting on most categories, for challenging and low-frequency labels. Detailed experimental results are provided in Appendix C.

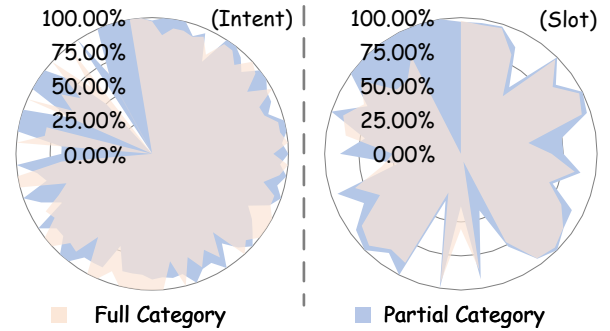


Figure 4: Comparison of F1 scores for intent and slot categories under Full and Partial settings of Mistral-7B.

4.3 Results and Analysis of Different Prompts

Table 3 compares our prompt design with previous prompting methods, including UGEN (Wu et al., 2022) and DC-Instruct (Xing et al., 2024), on Mistral-7B under the Partial Category setting.

UGEN and DC-Instruct model the interaction between intent detection and slot filling through constructing multiple questions, which leads to a

Method	I(Acc)	I(F1)	S(F1)	O(Acc)	PFLOPs
UGEN	96.53	92.37	87.68	89.59	235.81
DC-Instruct	96.78	92.84	87.84	90.09	304.23
IS-Map(Ours)	97.03	93.68	89.90	91.17	163.75

Table 3: Results of different prompts under Partial Category setting (Mistral-7B).

substantial increase in training FLOPs. In contrast, we present all candidate intents and their corresponding slot candidates to the model in a combined format. This enables the model to explicitly perceive the direct correlation between the two, eliminating the need for multi-turn dialogue to establish the association between intent and slot recognition. Thanks to this streamlined design, the proposed IS-Map achieves lower computational overhead while delivering superior performance.

5 Conclusion

We introduce CMPA-SLU, a Chinese SLU dataset collected from real-world applications with rigorous quality control, substantially expanding existing Chinese SLU benchmarks in domains, intents, and slot types. To better reflect practical deployment scenarios, we further introduce a Partial Category setting based on candidate label pre-selection. Experimental results on multiple LLMs show that CMPA-SLU remains challenging, while reducing the candidate label space improves intent and slot prediction performance. In addition, the proposed IS-Map prompt supports efficient joint prediction within a unified generation process.

Limitations

Although CMPA-SLU covers a wide range of domains and intent categories, it is currently limited to Chinese mobile assistant scenarios. As user behavior and language usage may vary across different languages and cultural contexts, the generalizability of CMPA-SLU to multilingual settings remains unexplored. Future work may extend the dataset to additional languages and regions to support broader SLU research.

Acknowledgements

This work was supported by the Innovation Studio for Beijing Outstanding Young Engineers.

Ethical Considerations

The dataset constructed in the manuscript has undergone rigorous deidentification and does not raise any ethical concerns.

References

- Ainikaerjiang Aimaiti, Di Wu, Liting Jiang, Gulin-geer Abudouwaili, Hao Huang, and Wushour Silamu. 2024. An uyghur extension to the massive multilingual spoken language understanding corpus with comprehensive evaluations. In *Proc. Interspeech 2024*, pages 3525–3529.
- Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 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.
- S’ebastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupasca, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K Spears, Derya Unutmaz, Kevin Weil, Steven Yin, and Nikita Zhivotovskiy. 2025. Early science acceleration experiments with gpt-5. *arXiv preprint arXiv:2511.16072*.
- China Internet Network Information Center. 2025. [The 56th statistical report on internet development in china \[in chinese\]](#). Technical report, China Internet Network Information Center (CNNIC).
- 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 Natara-jan. 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 539 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 180 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

- E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5467–5471.
- Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huinan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, and 17 others. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 398 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023a. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Lighting Jiang, Di Wu, Zhe Li, Yanbing Li, and Hao Huang. 2026. 2i-instruct: Generative joint empathy detection and empathy intent classification via inter-task and inter-instance interactions. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18097–18101. IEEE.
- Lighting Jiang, Di Wu, Bohui Mao, Yanbing Li, and Wushour Slamu. 2023b. Empathy intent drives empathy detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290.
- Lighting Jiang, Di Wu, Haoxiang Su, Xiaoyong Guo, Shuangyong Song, and Yanbing Li. 2025. A label co-occurrence transformation network for joint empathy detection and empathy intent classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Maikel Leon. 2025. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620.
- Xinzhang Liu, Chao Wang, Zhihao Yang, Zhuo Jiang, Xuncheng Zhao, Haoran Wang, Lei Li, Dongdong He, Luobin Liu, Kaizhe Yuan, Han Gao, Zihan Wang, Yitong Yao, Sishi Xiong, Wenmin Deng, Haowei He, Kaidong Yu, Yu Zhao, Ruiyu Fang, and 35 others. 2025. Training report of telechat3-moe. *arXiv preprint arXiv:2512.24157*.
- Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu. 2019. Cm-net: A novel collaborative memory network for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1051–1060.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Jinhui Pang, Huinan Xu, Shuangyong Song, Bo Zou, and Xiaodong He. 2022. Mfdg: A multi-factor dialogue graph model for dialogue intent classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 691–706. Springer.
- Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021a. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197. IEEE.
- Libo Qin, Fuxuan Wei, Qiguang Chen, Jingxuan Zhou, Shijue Huang, Jiasheng Si, Wenpeng Lu, and Wanxiang Che. 2025. Croprompt: Cross-task interactive prompting for zero-shot spoken language understanding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021b. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv preprint arXiv:2103.03095*.
- Feifan Song, Lianzhe Huang, and Houfeng Wang. 2024. A unified framework for multi-intent spoken language understanding with prompting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9966–9970. IEEE.
- Rob Van Der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet   st  n, Marija Stepanovi  , Alan Ramponi,

- Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.
- Zihan Wang, Xinzhang Liu, Yitong Yao, Chao Wang, Yu Zhao, Zhihao Yang, Wenmin Deng, Kaipeng Jia, Jiabin Peng, Yuyao Huang, Sishi Xiong, Zhuo Jiang, Kaidong Yu, Xiaohui Hu, Fubei Yao, Ruiyu Fang, Zhuoru Jiang, Ruiting Song, Qiyi Xie, and 19 others. 2025. Technical report of telechat2, telechat2. 5 and t1. *arXiv preprint arXiv:2507.18013*.
- Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.
- Di Wu, Ruiyu Fang, Liting Jiang, Shuangyong Song, Xiaomeng Huang, Shiquan Wang, Zhongqiu Li, Lingling Shi, Mengjiao Bao, Yongxiang Li, and Huang Hao. 2025a. Multi-intent spoken language understanding: a survey of methods, trends, and challenges. *Vicinatearth*, 2(1):20.
- Di Wu, Liting Jiang, Ruiyu Fang, Jing Bian, Hongyan Xie, Haoxiang Su, Hao Huang, Zhongjiang He, Shuangyong Song, and Xuelong Li. 2026. Introducing visual scenes and reasoning: A more realistic benchmark for spoken language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33899–33907.
- Di Wu, Liting Jiang, Bohui Mao, Hongyan Xie, Haoxiang Su, Zhongjiang He, Ruiyu Fang, Shuangyong Song, Hao Huang, and Xuelong Li. 2025b. Int: Establishing information transfer for multilingual intent detection and slot filling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15120–15142.
- Di Wu, Liting Jiang, Lili Yin, Zhe Li, and Hao Huang. 2024a. Cea-net: a co-interactive external attention network for joint intent detection and slot filling. *Neural Computing and Applications*, 36(22):13513–13525.
- Di Wu, Liting Jiang, Lili Yin, Kai Wang, Haoxiang Su, Zhe Li, and Hao Huang. 2024b. Dual level intent-slot interaction for improved multi-intent spoken language understanding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12301–12305. IEEE.
- Yangjun Wu, Han Wang, Dongxiang Zhang, Gang Chen, and Hao Zhang. 2022. Incorporating instructional prompts into a unified generative framework for joint multiple intent detection and slot filling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7203–7208.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- Bo Xie, Xiaohui Jia, Xiawen Song, Hua Zhang, Bi Chen, Bo Jiang, Ye Wang, and Yun Pan. 2023. Recomif: Reading comprehension based multi-source information fusion network for chinese spoken language understanding. *Information Fusion*, 96:192–201.
- Bowen Xing, Lizi Liao, Minlie Huang, and Ivor Tsang. 2024. Dc-instruct: An effective framework for generative multi-intent spoken language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14534.
- Sishi Xiong, Yu Zhao, Jie Zhang, Li Mengxiang, Zhongjiang He, Xuelong Li, and Shuangyong Song. 2024. Dual prompt tuning based contrastive learning for hierarchical text classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12146–12158.

A Category Distribution and Data Examples

The slot frequency distribution of the CMPA-SLU is presented in Table 4, while the intent frequency distributions are reported in Tables 5, 6, and 7.

As shown in these tables, both intent and slot distributions are highly imbalanced, which aligns with the nature of user requests in real-world applications. Notably, the CMPA-SLU dataset introduces an *Other Intent* category, which covers user utterances involving casual conversation or non-task-oriented expressions. As shown in Figure 5(a), the user utterance “We’re off tomorrow.” does not convey any actual task need. Similar to existing SLU datasets, CMPA-SLU also includes conventional intent requests, such as *Set Alarm* shown in Figure 5(b). Moreover, our dataset is more novel, featuring functionalities that have only recently been introduced in mobile devices, such as “Turn off AI subtitles” (Figure 5(c)). Meanwhile, it is also more aligned with the usage habits of Chinese mobile users, containing operation requests for popular applications such as “WeChat” (Figure 5(d)).

B Experimental Setup Details

Following Haihong et al. (2019); Qin et al. (2021a), we use accuracy (Acc) to evaluate intent detection performance, the F1 score to measure the quality of slot filling, and sentence-level Acc to assess overall

Slot	Count	Slot	Count	Slot	Count
Time	1741	City	1233	Numerical Value	1214
Location	536	Event	470	App Name	436
Destination Station	435	Check-in Date	419	WeChat Contact	418
Departure Time	413	Departure Station	391	Requirements	378
Target Area	348	Source Area	348	Amount	322
Target Length	312	Source Length	312	Maximum Price	312
Query Item	286	Target Weight	248	Source Weight	248
Song Name	235	Type	228	Singer	223
Star Rating	216	Check-out Date	213	Number of Rooms	202
Contact Person	200	Destination	192	Competition Name	164
Video Name	162	Match Time	158	Number of Stay Days	150
Matter	138	Passenger (Flight)	116	Target Currency	110
Facility Type	108	Source Currency	108	Passenger (Ride)	104
Target	103	Scene	94	Content	91
Minimum Price	90	Departure Location	68	Travel Mode	68
Text Content	64	Hotel Name	61	Guest Name	58
Book Title	54	SMS Content	50	Target Time	46
Source Time	46	Competing Party	40	Playback Tool	30
Stock Name	28	File Name	18	Date	14
Card Sender	12	Card Recipient	12	Product Name	12
File Type	10	Stock Price Threshold	10	Change Direction	10
Holiday	8	Refresh Rate Value	6	Store Name	6
Quantity	6	Program Name	6	Rider	4
Gregorian Year	4	Gregorian Date	4	Shopping Platform	2
Year	2	Start Time	1		

Table 4: Frequency counts of slot types in CMPA-SLU.

(a)	We're off tomorrow. Intent: Other Intent Slot: []
(b)	Set an alarm for 6:30 tomorrow morning. Intent: Set Alarm Slot: [Time: 6:30 tomorrow morning]
(c)	I don't want the AI subtitles anymore, turn them off. Intent: Turn Off Video AI Subtitles Slot: []
(d)	Send Alice 300.75 yuan on WeChat. Intent: WeChat Transfer Slot: [WeChat Contact: Alice, Value:300.75]

Figure 5: CMPA-SLU dataset examples.

performance. In addition, as shown in Figure 1 and Appendix A, the constructed CMPA-SLU dataset exhibits class imbalance for intent labels. Therefore, we also report the intent F1 score.

Traditional small models (Pang et al., 2022; Wu et al., 2024b; Jiang et al., 2025; Wu et al., 2024a) or pretrained methods (Jiang et al., 2023b; Wu et al., 2025b) typically formulate intent detection as clas-

sification over a fixed label space and slot filling as sequence labeling. However, CMPA-SLU’s Partial Category setting introduces dynamic candidate sets, where the number of candidates varies across samples. Therefore, models need to perform semantic matching within a given candidate space rather than classification over a fixed label space, requiring a different modeling framework. Our experiments mainly focus on evaluating LLM-based SLU under dynamic candidate spaces, and thus we select LLMs with contextual understanding and instruction-following capabilities for experiments.

Closed-source models: GPT-4o (version: 2023-05-15, temperature: 0.7), and GPT-5 (version: 2024-12-01-preview, temperature: 0.7).

Open-source models: Qwen2.5-7B, Mistral-7B, LLaMA3.1-8B, and Deepseek-r1-8B. For training, we use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5e-5, training each model for 3 epochs with a batch size of 1. The LoRA rank is set to 8, and the dropout rate is 0.1. To ensure reproducibility, we use a fixed

random seed of 42 across all models. All experiments are conducted on a single NVIDIA A100 GPU (40GB).

C Details of Category F1 Scores

Detailed experimental results corresponding to Figure 4 are presented in Figures 6 and 7. Figures 6 reports the F1 scores of individual intent categories under the Full Category and Partial Category settings, respectively, while Figure 7 presents the F1 scores of slot categories under both settings.

Combined with the frequency statistics in Tables 4, 5, and 6, these results indicate that the improvements achieved by Mistral-7B under the Partial Category setting are broadly distributed across categories rather than concentrated in only a few cases. The performance gains are notable for low-frequency categories.

Intent	Count	Intent	Count
Other Intent	1133	Book Hotels	569
Convert Area Units	348	Convert Length Units	312
Check PM Index	304	Convert Weight Units	248
Play Music	236	Book Train Tickets	232
Check Weather	232	Book Flight Tickets	230
Recharge phone with amount	214	Close App	200
Watch Sports Events	186	Play Video	162
WeChat Transfer	162	Set Alarm	156
Turn Off Alarm	147	Disable Driving Mode	144
Disable Simple Mode	140	Recharge phone	140
Search the Internet	138	Stop Playback	136
Turn Off Wi-Fi	134	Turn Off Airplane Mode	134
Make Payment	130	Play News	130
Exit Smart Voice	126	Check Alarm	125
Check UV Index	124	Alipay Payment	124
Check Data Usage	124	Check Flight Itinerary	124
Turn Off Flashlight	122	Turn Off Bluetooth	120
Check Schedule	120	Resume Playback	116
Turn Off Vibration	114	Uninstall App	112
Set Phone Ringtone	112	Image Search	111
Exchange Rate Inquiry	110	Check Nearby Places	108
Set Scheduled Power On/Off	108	WeChat Payment	108
Turn Off Mobile Data	104	Delete Schedule	104
WeChat Sports	104	Check Train Itinerary	104
WeChat Receive Payment	100	Exit Navigation	100
Turn Off Do Not Disturb Mode	98	Enable SIM 2 Data	98
Check Battery	98	Enable Silent Mode	96
Turn On Wi-Fi	94	Turn Off GPS	94
Lower Phone Volume	94	Enable Red Packet Grabber	94
Turn Off Screen	90	Open Phone Volume Settings Panel	90
Turn Off Hotspot	84	Enable Mobile Data	84
Enable Voice-Controlled Camera	84	Turn Off Phone	84
Check Current Location	82	Alipay Features	82
Mini Programs/Official Accounts	80	Enable Hotspot	80
Set Schedule	78	Check Route to a Place	78
Open App	76	WeChat Send Red Packet	74
Set Phone Font	72	Open Screen Brightness Settings Panel	70
Disable Voice-Controlled Camera	70	Modify Schedule	70
WeChat Voice Call	70	Open Bluetooth Settings Panel	68
Set Display Settings	68	Play Next Track	68
Set App Notifications	66	View Moments	66
Open Do Not Disturb Mode Settings Panel	64	Enable Simple Mode	64

Table 5: Frequency counts of intents in CMPA-SLU (Part I).

Intent	Count	Intent	Count
Send Message via WeChat	64	Make a Call	63
Play Previous Track	62	Top Up Data Usage	60
Delete Alarm	58	Check Hotel Itinerary	58
Open Health Code	56	Reduce Phone Font Size	54
Open Airplane Mode Settings Panel	54	Disable SIM 1 Data	54
Search for Novels	54	Open Settings	52
Set Theme	52	Check Phone Model	52
Enable Driving Mode	52	Send SMS	50
Enable Vibration	50	Tell a Joke	50
Turn On Bluetooth	48	Set One-Handed Mode	48
WeChat Video Call	48	Tell a Story	48
Set Standby Wallpaper	48	Open Wi-Fi Settings Panel	46
Enable Airplane Mode	46	Turn Off Power Saving Mode	46
Check Phone Software Version	46	Convert Time Units	46
WeChat Scan	46	Voice Wake-Up	46
Open Hotspot Settings Panel	44	Enable Do Not Disturb Mode	44
Enable Power Saving Mode	44	Increase Phone Font Size	42
Enable SIM 1 Data	42	Set Phone Data Plan	42
WeChat Health Code	42	Alipay Scan	42
Enable Wireless Charging	42	Alipay Receive Payment	40
Voice Settings	40	Turn On Flashlight	38
Turn Off Silent Mode	38	Set Default Apps	38
Enable Smart Scenarios	38	Return to Home Screen	38
Disable SIM 2 Data	36	Set Navigation Keys	36
Check Phone Number	35	Delete SMS	34
Increase Phone Volume	34	Set Lock Screen Wallpaper	34
Check Phone IMEI	34	Record Audio	34
Check Exercise Index	34	Set Alarm	32
Set System Update	32	Check Phone Processor	32
Enable Infrared Function	32	Install App	31
Open GPS Settings Panel	28	Brighten Screen	28
Set Face Recognition	28	Search for Files	28
Display My WeChat QR Code	28	Read Screen Aloud	28
Set Desktop Wallpaper	26	Set Mobile Network	26
Enable Developer Mode	26	Check Dressing Index	26
Check Fishing Index	26	Hail a Ride	26
Delete Phone Number	24	Start Data Backup	24
Countdown	24	Check Phone Memory	22
Check Phone Storage	22	Take Night Photos	22
Report a Bug	22	Submit Suggestions/Requests	22
Set Screen Brightness to Minimum	22	Dim Screen	20
Set Connection to PC	20	Activate Voice Assistant	20

Table 6: Frequency counts of intents in CMPA-SLU (Part II).

Intent	Count	Intent	Count
Check Phone Balance	20	Check Phone Screen Resolution	20
Take Photos	20	Set Phone Volume to Minimum	20
Turn Off Ultra Power Saving Mode	20	Open Negative One Screen	18
Factory Reset	18	Find My Phone	18
Start Split Screen	18	Start Backup Restoration	18
Set Red Packet Grabber	18	Set Date and Time	18
Start Screen Recording	18	Check Phone Feature Help Manual	18
Set Phone Volume to Maximum	18	Take Selfie	17
Set Lock Screen Unlock Method	16	Enable Multitasking	16
Enable HD Calling	16	Turn Off NFC	16
View Monitored Stocks	16	Search by Keyword in Specified App	16
Set Phone Fingerprint	14	Check Phone Name	14
Enable Screen Casting	14	Restart Phone	14
Take Screenshot	14	Check Car Wash Index	14
Stop Monitoring Stock	14	Mobile Game Recommendations	14
Set App Data Management	12	Set Screen Assistant	12
Lock Desktop Layout	12	WeChat Send Contact Card	12
Generate Images from Text	12	Alipay Ant Forest Collect Energy	12
Use NFC Wallet Feature	12	Check Current Time	12
Arithmetic Operations	11	Turn On GPS	10
App Cloning	10	Phone Function Self-Check	10
Set Input Method	10	Check Traffic Violations	10
Set Screen Brightness to Maximum	10	Monitor Stock Prices	10
Turn Off Video AI Subtitles	10	Turn Off Screen Translation	10
Record Video	8	Switch Phone Language	8
Use ShareMe to Share Files	8	Adjust Phone Refresh Rate	8
Convert Relative Date to Specific Date	8	Navigate	6
Record Access Card	6	Check Address	6
Organize Phone Desktop	6	Open Sidebar	6
Order Food Delivery	6	Check TV Programs	6
Set Desktop	4	Start Long Screenshot	4
Turn Off Future Mode	4	Check Stock	4
Turn On NFC	4	Enable Ultra Power Saving Mode	4
Enable Video AI Subtitles	4	Search Historical Prices and Lowest Price	4
Convert Solar Calendar to Lunar Calendar	4	Check Which Day a Holiday Falls On	4
Check Holidays on a Specific Date	4	Check Days Until a Specific Holiday	4
Check Lunar Calendar Zodiac and Animal Sign	4	Open NFC Settings Panel	2
Search for Images in Apps	2	Buy Things	2
Enable Voice Conversation Translation	2	Convert Date to Day of the Week	2
Translate Content on Screen	2		

Table 7: Frequency counts of intents in CMPA-SLU (Part III).

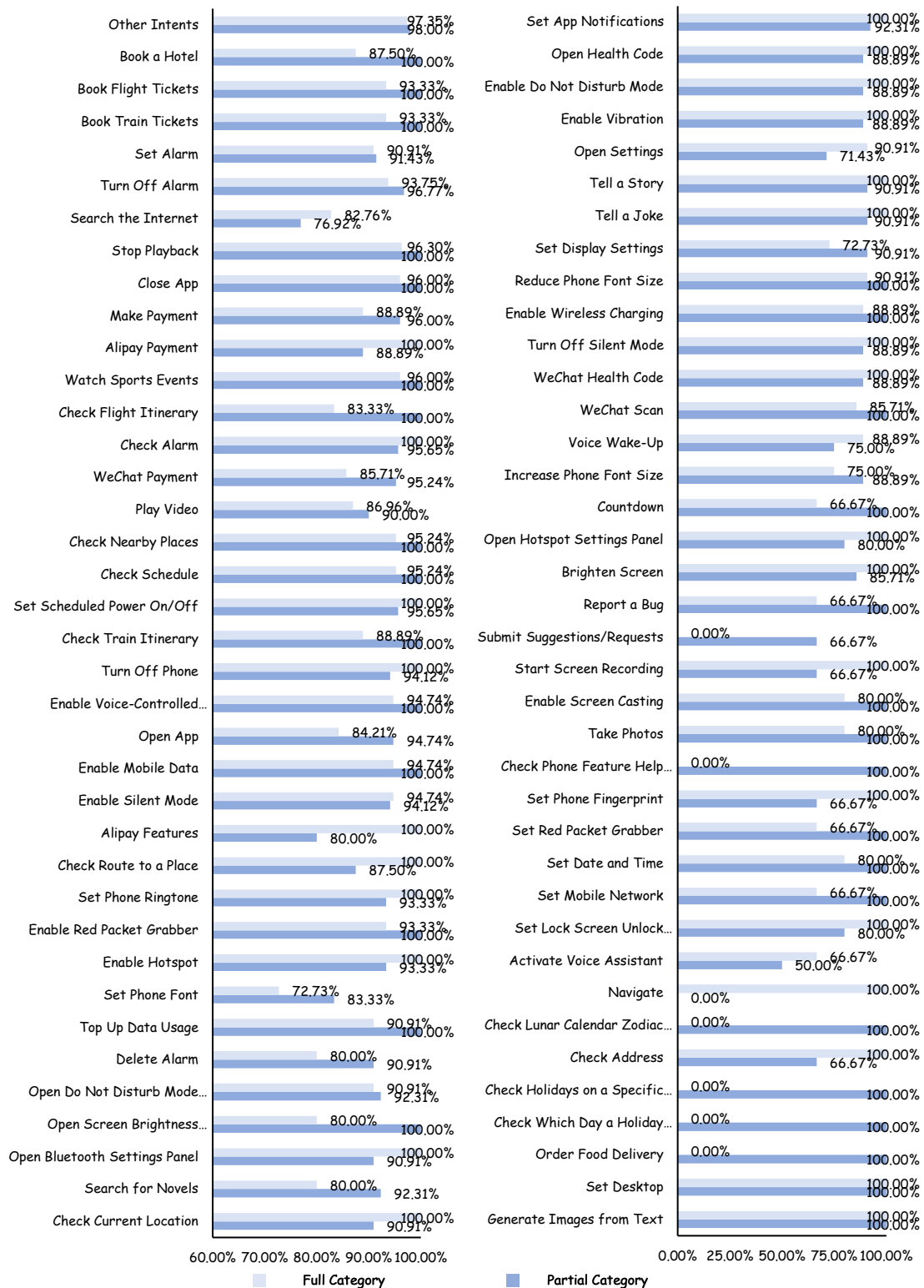


Figure 6: Comparison of F1 scores for intent under Full and Partial settings of Mistral-7B. Results that are identical have been omitted.

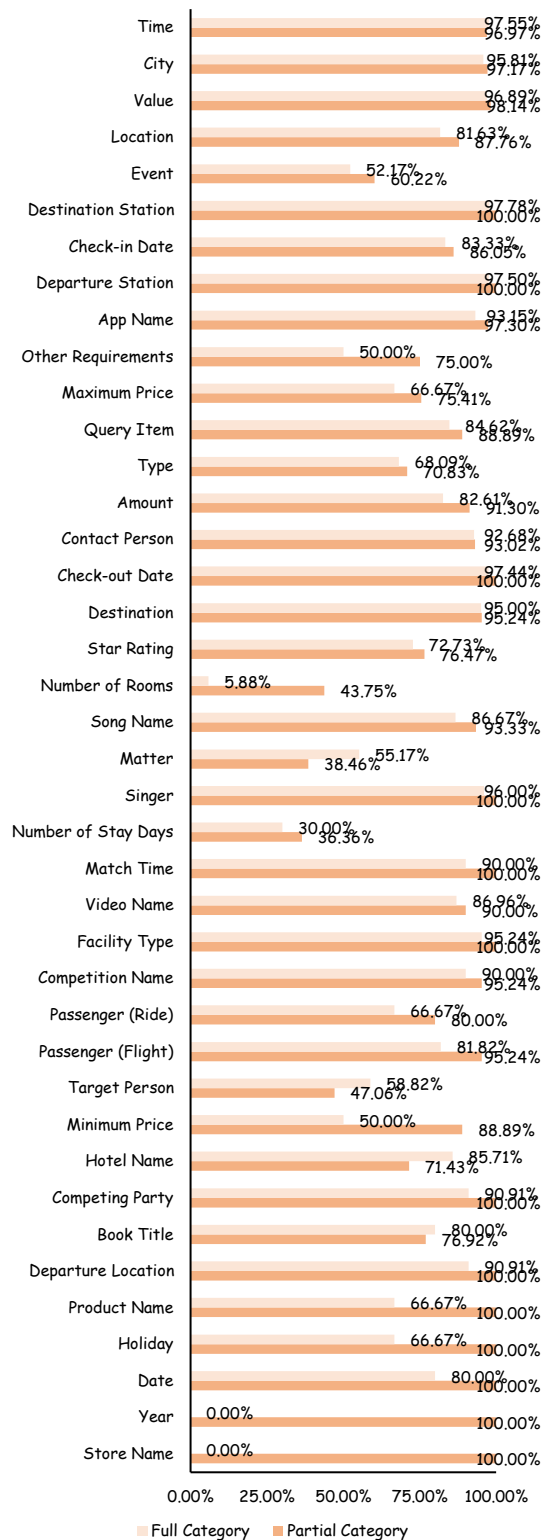


Figure 7: Comparison of F1 scores for slot under Full and Partial settings of Mistral-7B. Results that are identical have been omitted.