

Beyond the Known: Investigating LLMs Performance on Out-of-Domain Intent Detection

Pei Wang¹, Keqing He², Yejie Wang¹, Xiaoshuai Song¹, Yutao Mou¹,
Jingang Wang², Yunsen Xian², Xunliang Cai², Weiran Xu¹

¹Beijing University of Posts and Telecommunications

²Meituan Group

wangpei@bupt.edu.cn





Content

- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 



Content



- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 

Motivations

OOD Intent Detection: Identify when a user's query falls outside the predefined intents recognized by a system

LLM: The emergence of LLM injects new vitality into NLP tasks. Their superior zero-shot learning capability enables a new paradigm of NLP research and applications by prompting LLMs without finetuning.

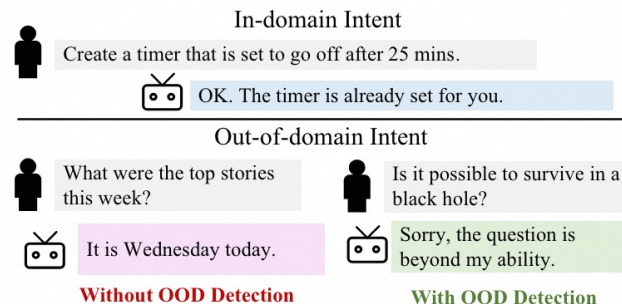


Figure 1: Explanation of the role of OOD intent detection in the TOD system. When the system encounters an intent that is beyond its supported intents, it can detect and friendly prompt the user.

What are the potential benefits and challenges that LLMs might meet in open-scenario intent detection?

Our research goal:

1. What are the potential positive and negative effects of large language models on the Out-ofDomain (OOD) detection task?
2. What are the strengths and weaknesses of large language models, compared with traditional finetuned models?
3. Why do large language models exhibit certain strengths and weaknesses?
4. How can we potentially address and improve these weaknesses?



Content



- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 

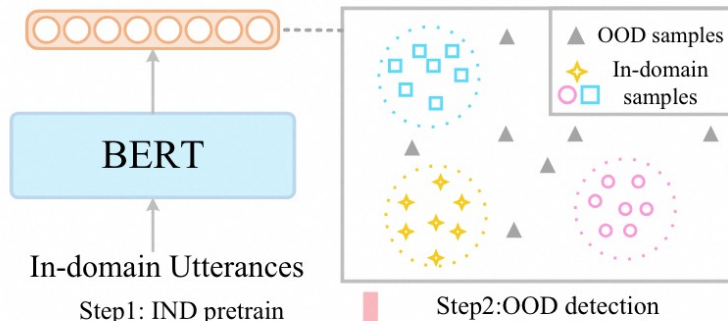
- **Innovative Frameworks:** Introduction of two LLM-based frameworks, ZSD-LLM and FSD-LLM, for improved OOD intent detection.
- **Comparative Performance Analysis:** A thorough comparison between ChatGPT and traditional discriminative models in OOD detection, highlighting the effectiveness of LLMs.
- **Extensive Analytical Experiments:** Conduct a variety of experiments to explore factors affecting OOD detection in LLMs, such as the number of IND intents and the impact of prompts.
- **Identification of Strengths and Weaknesses:** Summarize the strengths and weaknesses of ChatGPT in OOD detection tasks and future improvement directions



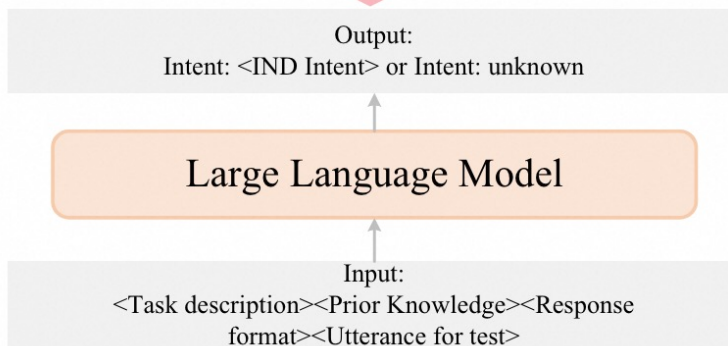
Content

- 1 Motivations
 - 2 Contributions
 - 3 Methodology**
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 

Traditional
discriminative
model-based
methods



The end-to-
end method
based on large
models



<Task description>

You are an out-of-domain intent detector, and your task is to detect whether the intents of users' queries belong to the intents supported by the system. If they do, return the corresponding intent label, otherwise return unknown. The supported intents include: [Intent 1], [Intent 2] ... [Intent N]

<Response format>

Please respond to me with the format of "Intent: XX" or "Intent: unknown"

<Utterance for test>

Please tell me the intent of this text: [Here is the utterance for text.]

ZSD-LLM

<Task description>

You are an out-of-domain intent detector, and your task is to detect whether the intents of users' queries belong to the intents supported by the system. If they do, return the corresponding intent label, otherwise return unknown. The supported intents include: [Intent 1] ([Example 1] [Example 2]...), [Intent 2] ([Example 1] [Example 2]...), ...The text in parentheses is the example of the corresponding intent.

<Response format>

Please respond to me with the format of "Intent: XX" or "Intent: unknown"

<Utterance for test>

Please tell me the intent of this text: [Here is the utterance for text.]

FSD-LLM

Figure 3: The demonstration of the two prompts we use to assist ChatGPT in performing OOD intent detection. FSD-OOD incorporates examples of intentions in the prompt as prior knowledge.



Content



- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup**
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 

- Datasets

- CLICK^[1]
- Banking^[2]

- Split: 25%, 50%, 75%

- Evaluation Metrics

- IND- Accuracy
- IND- F1
- OOD- Recall
- OOD- F1

[1] Stefan Larson, Anish Mahendran, Joseph Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In EMNLP/IJCNLP

[2] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. arXiv preprint arXiv:2003.04807.

- Main Results ➤ Results for ZSD

Model	Split = 25%						Split = 50%						Split = 75%					
	ALL		IND		OOD		ALL		IND		OOD		ALL		IND		OOD	
	ACC	F1	ACC	F1	Recall	F1	ACC	F1	ACC	F1	Recall	F1	ACC	F1	ACC	F1	Recall	F1
SCL	74.36	61.06	71.57	60.18	76.46	77.76	75.45	69.94	80.43	69.98	67.44	68.25	79.43	84.54	82.35	84.86	71.06	66.34
KNN-CL	88.21	77.65	78.16	77.94	91.77	92.36	81.98	83.67	85.13	83.72	85.25	81.96	81.69	70.21	86.30	86.03	85.13	71.88
UniNL	89.41	80.04	78.59	79.36	92.96	93.02	81.42	82.66	84.68	82.70	78.24	81.18	82.78	86.36	81.78	86.59	85.66	73.34
ChatGPT	47.5	42.68	73.16	42.02	39.09	55.17	46.46	54.91	71.32	55.47	22.24	33.77	50.58	56.97	62.85	57.58	15.62	22.03

Table 1: The performance comparison between ChatGPT and baselines of Banking. We select 25%, 50%, and 75% of all intents as IND intents. Three average values are taken for each experiment.

Model	Split = 25%						Split = 50%						Split = 75%					
	ALL		IND		OOD		ALL		IND		OOD		ALL		IND		OOD	
	ACC	F1	ACC	F1	Recall	F1	ACC	F1	ACC	F1	Recall	F1	ACC	F1	ACC	F1	Recall	F1
SCL	87.64	88.32	91.44	89.08	74.11	77.76	85.82	83.64	84.89	83.58	86.46	87.94	89.18	90.55	88.75	90.58	89.86	87.13
KNN-CL	92.04	83.31	84.86	82.99	93.85	94.97	90.33	88.52	88.53	88.47	91.57	91.95	89.18	92.03	88.49	92.10	92.30	77.66
UniNL	87.8	89.79	97.27	89.89	77.3	86.14	90.95	93.15	95.79	93.25	80.03	85.75	91.77	94.01	93.84	94.09	83.95	82.74
ChatGPT	63.86	58.86	81.26	58.51	58.15	71.82	59.84	69.9	82.4	70.14	37.29	51.43	64.24	70.18	74.79	70.44	33.16	41.71

Table 2: The performance comparison between ChatGPT and baselines of CLINC. We select 25%, 50%, and 75% of all intents as IND intents. Three average values are taken for each experiment.

- **ChatGPT Trails in IND Accuracy**
- **OOD Recall Significantly Lower**
- **Better Results on CLINC than Banking**

- Main Results
 - Results for FSD

IND num	Few-shot	ALL		IND		OOD	
		ACC	F1	ACC	F1	ACC	F1
5	0	77.19	69.06	96.36	66.09	72.61	83.92
	1	88.80	80.11	90.00	77.60	88.50	92.67
	3	85.89	77.02	96.02	74.34	83.33	90.41
	5	89.11	81.26	90.01	78.93	88.89	92.88
	UniNL	88.35	78.1	78.35	75.21	90.85	92.58
10	0	63.44	64.85	86.87	64.64	51.71	66.94
	1	79.19	74.75	84.00	73.85	76.77	83.75
	3	78.81	76.67	91.00	76.10	72.77	82.35
	5	82.89	88.17	89.80	79.57	79.50	86.18
	UniNL	84.17	78.11	74.55	77.1	89.05	88.25
20	0	61.76	63.01	74.78	63.11	48.54	60.96
	1	70.89	72.29	79.29	72.36	62.44	70.89
	3	75.95	76.30	80.81	76.24	71.07	77.35
	5	77.84	79.57	88.27	79.74	65.82	76.33
	UniNL	80.51	78.87	73.21	78.71	87.89	81.92
30	0	56.91	58.13	74.31	58.64	30.56	42.72
	1	69.06	72.00	78.64	72.28	54.40	63.64
	3	72.18	72.65	84.31	73.00	52.63	62.11
	5	74.74	81.38	91.95	82.03	47.62	61.86
	UniNL	80.76	83.07	82.53	83.29	78.09	76.61
40	0	55.80	60.26	67.25	60.75	31.86	40.66
	1	66.32	70.66	76.53	71.08	45.60	53.99
	3	69.08	75.32	83.12	75.87	40.91	53.11
	5	69.35	75.84	86.84	76.55	32.80	47.15
	UniNL	83.69	87.77	86.86	88.05	77.32	76.53

Table 4: Performance of ChatGPT under different few-shot settings with varying five sets of IND numbers.

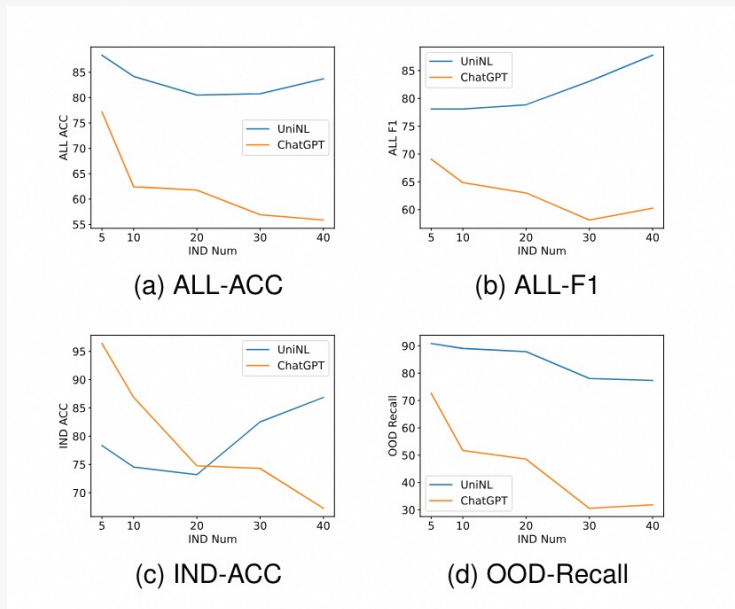
- (1) FSD-LLM demonstrates strong competitiveness compared to the baseline in situations with a limited number of INDs
- (2) The more the number of intents, the more demonstrations are needed for IND intent recognition.
- (3) Too many demonstrations may introduce noise into OOD detection.



Content

- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis**
 - 6 Challenges
 - 7 Conclusion
- 

- The effect of IND Intent number



- ChatGPT is sensitive to the number of intents compared with UniNL.
- The increase in intents leads to more severe confusion between labels.
- The increase in intents causes a sharp drop in the OOD-recall rate.

- Other LLM

Model	ALL		IND		OOD	
	ACC	F1	ACC	F1	Recall	F1
text-davinci-002	54.24	60.72	73.6	60.85	34.38	48.53
text-davinci-003	55.87	64.14	79.03	65.15	30.81	43.98
Claude	56.58	52.76	70.72	52.74	43.59	53.12
Llama2-70b-Chat	55.5	57.8	67.0	57.84	44.0	56.96
ChatGPT	61.76	63.01	74.78	63.11	48.54	60.96
GPT4	68.5	73.55	87.5	74.07	49.5	63.26

Table 5: OOD detection performance of six different LLMs.

- Other prompt

Model	ALL		IND		OOD	
	ACC	F1	ACC	F1	Recall	F1
prompt.original	53.10	51.32	56.82	51.18	49.47	56.62
prompt.detector	51.07	53.20	61.61	53.25	40.78	51.11
prompt.discovery	45.63	51.05	63.29	51.32	28.39	40.66
prompt.order	42.47	48.57	51.75	48.74	33.40	42.09
prompt.reason	48.67	47.65	55.47	47.56	42.03	50.92

Table 6: The performance of ChatGPT on various prompts.



Content



- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges**
 - 7 Conclusion
- 

- Conflict between Domain-Specific Knowledge and General Knowledge
false association, focus deviation, lack of domain knowledge.

(a) General knowledge vs domain-specific knowledge

False Association

Query: Can you tell me if my top-up has been cancelled?

Ture: OOD

Predict:

top_up_by_bank_transfer_charge.

Error:

the cancellation if accidental, which may be related to the charge.

Focus Deviation

Query: How do I transfer funds from my American Express into my account?

Ture:

supported_cards_and_currencies

Predict: OOD

Error:

No emphasis on which account is referred to as 'my account'.

Lack of Domain Knowledge

Query: What happened to where my top-up was canceled?

Ture: OOD

Predict: declined_transfer

Error:

charging and transfer are equivalent.

(b) Ambiguous label meaning

Query: I don't understand how to top up my account

Ture: OOD

Predict: top_up_by_bank_transfer_charge

Error:

Asking for The method of toping up is understood, but the label was misclassified.

(c) Task failed

Query: how many prime numbers are there between 0 and 100 ?

Ture: OOD

Predict: math

Error:

'Math' is not in our supported intents.

Challenge & Further Discussion

- Difficulty of Knowledge Transfer from IND to OOD
- Sensitivity to input length

- **Future Insights**

- ✓ injecting domain knowledge
- ✓ strengthening knowledge transfer from IND to OOD
- ✓ understanding long instructions.

Content

- 1 Motivations
- 2 Contributions
- 3 Methodology
- 4 Experimental Setup
- 5 Qualitative Analysis
- 6 Challenges
- 6 Conclusion**

In this paper, we conduct a comprehensive evaluation of ChatGPT for OOD intent detection. We first compare the performance of ChatGPT with traditional discriminative models and identify a significant performance gap. Additionally, we observe that ChatGPT excels in handling tasks with a small number of intents but struggles with tasks involving a large number of intents. While incorporating demonstration examples shows some improvements, there is still considerable room for enhancement. We recommend future research to focus on improving large-scale models for OOD tasks by incorporating domain-specific knowledge into the models and how to learn transfer relationship from OOD detection.

Thank You

THANKS!