The Limits of Post-hoc Preference Adaptation: A Case Study on DSTC12 Clustering

Jihyun Lee[1], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jihyunlee, gblee}@postech.ac.kr

Abstract

Understanding user intent in dialogue is essential for controllable and coherent conversational AI. In this work, we present a case study on controllable theme induction in dialogue systems using the DSTC12 Track 2 dataset. Our pipeline integrates LLM-based summarization, utterance clustering, and synthetic preference modeling based on should-link and cannot-link predictions. While preference signals offer moderate improvements in cluster refinement, we observe that their effectiveness is significantly constrained by coarse initial clustering. Experiments on the Finance and Insurance domains show that even authentic human labeled preference struggle when initial clusters do not align with human intent. These findings highlight the need to incorporate preference supervision earlier in the pipeline to ensure semantically coherent clustering.

1 Introduction

Understanding user intent in open-domain or task-oriented conversations has traditionally relied on supervised intent classification(Hemphill et al., 1990; Eric et al., 2019). However, these approaches often assume a fixed set of discrete intent categories and lack flexibility when transferred to real-world customer dialogues, where user queries span a continuum of fine-grained themes. To address this, recent work has explored theme induction as a more flexible alternative, allowing systems to discover and assign user-centered thematic labels to dialogue segments without relying on predefined taxonomies (Gung et al., 2023).

Early approaches to intent understanding relied on supervised classification with annotated datasets, using techniques like attention-based models (Goo et al., 2018) or semantic lexicon-enhanced embeddings (Kim et al., 2016; Fan et al., 2020). However, collecting labeled data at scale is costly, making it difficult to apply such models to new domains. To address this, unsupervised intent induction methods have emerged, typically using clustering algorithms (Koh et al., 2023) or embedding refinement (Perkins and Yang, 2019) to group utterances without labels. While effective in narrow settings, these methods often struggle with domain transfer and fine-grained intent variation (Zhang et al., 2024; Koh et al., 2023). As a more flexible alternative, recent work has explored theme induction (Gung et al., 2023), enabling the discovery of latent topics without fixed taxonomies—an idea further developed in the DSTC12 Track 2 task (Organizers, 2025), which introduces user-defined pairwise preferences to guide theme clustering.

To address the DSTC12 Track 2 task, we adopt two-stage pipeline: first performing unsupervised clustering of utterances, then refining the clusters using post-hoc preference adapting. Our system comprises (1) summarization-based input compression, (2) initial utterance clustering, (3) pseudo labeling preference using a fine-tuned large language model (LLM) classifier, and (4) preference-guided post-processing. To train the preference model, we fine-tune the LLM on should-link and cannot-link examples generated from distance-based heuristics within the training domain. Once trained, the model is used to generate preference labels for a different domain in a zero-shot setting to guide its clustering process. These predicted preferences are used to adjust the clusters by reassigning individual utterances, aiming to better reflect human interpretations of thematic coherence.

However, despite its modular appeal, our experiments reveal that post-hoc preference processing fails to reliably improve clustering quality. As shown in our analysis, even accurate preference predictions cannot override structural errors from the initial clustering phase. In particular, when the initial clusters misrepresent the semantic granularity expected by users (e.g., grouping together utterances with subtly distinct intents), preference

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 36–43

signals are often ineffective or misapplied. These findings suggest that controlling thematic granularity in dialogue clustering cannot be deferred to post-processing alone, and underscore the importance of integrating user preferences more holistically into theme detection systems.

2 DSTC12 Task Track2

The DSTC12 Track 2 challenge focuses on *Controllable Conversational Theme Detection*. Given a set of dialogue utterances, the goal is twofold: (1) to cluster utterances into semantically coherent themes, and (2) to assign concise, natural language labels to each theme. A key aspect of this task is controllability: the desired granularity of clustering must be inferred from user-provided preferences indicating whether two utterances should or should not belong to the same theme.

2.1 Inputs

Participants are provided with the following resources for the training and development phases:

- A set of themed utterances, each with full dialogue context.
- Pairwise *user preference data* that indicates whether two utterances should be grouped together (should-link) or separately (cannot-link).
- Gold theme labels for evaluation on the dev set (hidden for test).
- A theme label writing guideline that outlines acceptable forms, including brevity, event-oriented verb phrases, and avoidance of context-sensitive terms.

2.2 Outputs

The expected system outputs are:

- A clustering of the themed utterances into distinct themes.
- A concise natural language label for each theme cluster, following the provided style guidelines.

2.3 Evaluation

Evaluation consists of two components:

- **Clustering quality**: measured by Normalized Mutual Information (NMI) and clustering accuracy (ACC) based on gold theme assignments.
- **Label quality**: measured by Cosine similarity (Sentence-BERT (Reimers and Gurevych, 2019)), ROUGE scores, and a private LLM-based metric that checks guideline adherence.

The challenge setting emphasizes generalization, as the test set comes from an unseen domain. Therefore, systems are expected to perform zero-shot transfer using only the train/dev domains for tuning and validation.

3 Approach

Our approach to the DSTC12 controllable conversational theme detection task consists of four main components: (1) input compression through summarization, (2) pseudo-labeling of should-link and cannot-link pairs, (3) post-clustering refinement, and (4) theme label generation via LLM prompting. For both summarization and label generation, we employ `mistralai/Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), an instruction-tuned language model. We illustrate the overall process in Figure 1.

3.1 Dialogue Summarization for Input Compression

To reduce noise and standardize input semantics, we first apply an LLM-based summarization step to each target utterance using the surrounding dialogue context. While the original DSTC12 setup uses only the single user utterance where the theme is annotated, we hypothesized that incorporating preceding dialogue context could provide valuable cues about user intent. Therefore, instead of clustering based solely on the raw user turn, we summarize the full context into a single sentence that captures the core intention.

This summarization step is designed to remove speaker-specific fillers, overly fine-grained details, and disfluencies, while preserving the semantic intent necessary for accurate theme clustering. We initially expected that this abstraction would help produce more coherent clusters by reducing irrelevant lexical variation.

We use the following prompt to generate concise summaries of the user's intent from the dialogue context:

Summarization Prompt

The following is a conversation between a user and a system. Based on the entire dialogue, summarize the user's intent in a single concise sentence.

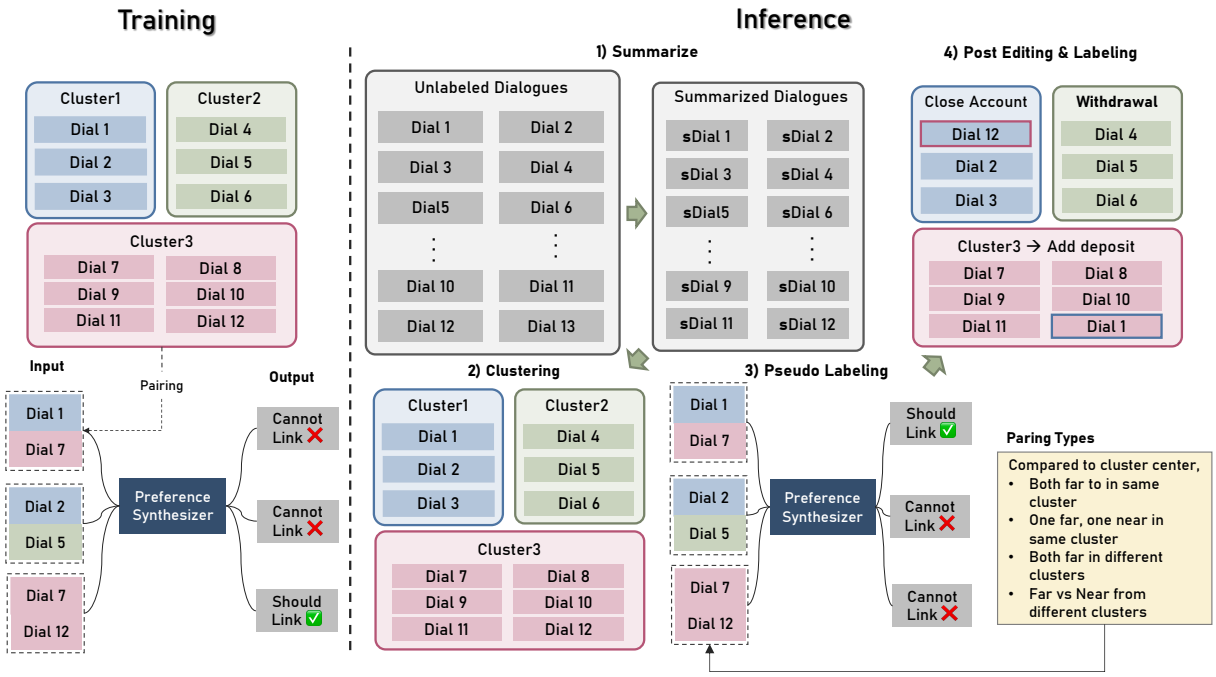

Figure 1: Overview of our pipeline. We train a preference model using cluster-based pairings, then apply it during inference to refine clustering results by predicting should-link/cannot-link pairs and adjusting utterance assignments accordingly.

> The summary must start with "User wants ..." or "User needs ...", and it should be concise and to the point. Output only a JSON object in the following format. Do not include any additional explanations or comments.
> **Format:**
> `{"summary": "<summary sentence>"}`
> **Dialogue:** `{dialogue}`

3.2 Pseudo-Labeling of Should-Link and Cannot-Link Pairs

To post-process the clustering results in alignment with human preferences, we train a pseudo-labeling model that classifies utterance pairs as either should-link or cannot-link, using supervision derived from human-annotated preferences. Specifically, we fine-tune a `LLaMA-3.1-8B-Instruct` (Grattafiori et al., 2024) model to determine whether two given dialogue contexts should belong to the same thematic cluster, generating structured outputs: `should-link` or `cannot-link`.

To construct the training dataset, we leverage the ground-truth cluster labels provided in the Banking domain. For each cluster, we compute embeddings for all utterances and calculate the cluster centroid by averaging the embeddings of utterances sharing the same label.

Within each cluster, utterances are categorized as either *near* or *far* based on their cosine distance to the centroid—specifically selecting the closest and farthest $k\%$, respectively. To ensure an informative and challenging training set, we sample a subset of contrastive utterance pairs likely to be difficult for the model, focusing on edge cases requiring fine-grained distinctions. We use the following types of pairs:

- **same_far_far**: Two *far* utterances from the same cluster (labeled `should-link`).
- **same_far_near**: One *far* and one *near* utterance from the same cluster (labeled `should-link`).
- **diff_far_far**: Two *far* utterances from different clusters (labeled `cannot-link`).
- **diff_far_near**: One *far* utterance from one cluster and one *near* utterance from a different cluster (labeled `cannot-link`).

During inference, we apply the same distance-based sampling strategy to identify utterance pairs that are likely to be misclustered. The trained model then predicts pairwise preferences, which are used to refine the clustering output. For each predicted `should-link` pair found in different clusters, we relocate one utterance to the cluster of its paired utterance to enforce co-membership. For

each `cannot-link` pair found in the same cluster, we move one utterance to the next most similar cluster based on centroid similarity, thus enforcing separation. This post-processing adjustment helps align the clustering structure more closely with human interpretations of thematic boundaries. We set k to 20% for both training and inference sampling.

3.3 Theme Label Generation

Lastly, after reassigin the cluster label with psuedo labels, we generate a short natural language label using an instruction-tuned LLM. Given a set of utterances within a cluster, we prompt the model to summarize the common customer intent using a constrained format. The prompt enforces the following requirements:

- The label must follow the structure: `verb + object` (e.g., *reset password*).
- All words must be in lowercase and free of punctuation.
- The label must contain a single verb and a 1–2 word noun phrase.
- The label should reflect the customer's intended action.

This approach aligns with the DSTC12 guideline for theme label writing and ensures consistency across generated labels.

4 Experiments

4.1 Experimental Setup

Dataset. We use the Banking (train) portion of the DSTC12 controllable conversational theme detection dataset, which consists of 2,504 themed utterances across 933 dialogues. Each utterance is annotated with a theme label and accompanied by surrounding dialogue context.

To train our preference synthesize model, we construct pairwise preference examples (`should-link` or `cannot-link`) from the training data. After removing duplicate prompts, we obtain 53,264 training instances, each consisting of a comparison between two utterances and a corresponding preference label. We evaluate our model on the Finance and Insurance splits of the DSTC12 dataset. Both domains are unseen during training. Note that we excluded the human-labeled preference datasets for the Finance and Insurance domains to evaluate performance in a truly unseen environment.

Clustering. We perform initial theme clustering over utterance embeddings using the KMeans algorithm. Each utterance is embedded using the `sentence-transformers/all-mpnet-base-v2` (Reimers and Gurevych, 2019) model, resulting in a fixed-dimensional vector representation. To determine the number of clusters k, we apply a silhouette-based selection method: for $k \in [15, 30]$, we compute the silhouette score for each candidate value and choose the k that yields the highest score. The selected number of clusters is then used to fit a KMeans model with `k-means++` initialization and a fixed random seed for reproducibility.

Training Details. We fine-tuned a `LLaMA-3.1-8B-Instruct` model using the HuggingFace Trainer[1] with LoRA (Hu et al., 2021) adaptation on a single A100-80GB GPU. Training was performed for one epoch with a learning rate of 1e-4 and batch size of 8 per device. LoRA was applied to the `q_proj` and `v_proj` modules with rank 8, $\alpha = 16$, and a dropout rate of 0.05.

Evaluation Metrics. To evaluate clustering and labeling performance, we report the following metrics:

- **Accuracy**: The proportion of utterances assigned to the correct cluster, assuming an optimal one-to-one mapping between predicted clusters and gold labels.
- **Normalized Mutual Information (NMI)**: Measures the mutual dependence between predicted and gold clusters. NMI is normalized between 0 (no mutual information) and 1 (perfect match), and is invariant to label permutations.
- **ROUGE-1 / ROUGE-2 / ROUGE-L**: These metrics assess lexical overlap between predicted theme labels and gold labels. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L captures the longest common subsequence.
- **Cosine Similarity**: The average cosine similarity between each utterance embedding and the centroid of its assigned cluster. This metric reflects intra-cluster semantic cohesion in the embedding space.

[1] https://huggingface.co/

Model Variant	Accuracy	NMI	ROUGE-1	ROUGE-2	ROUGE-L	Cosine Sim.	Clusters
Domain: Finance (Cluster num : 34)							
Baseline	41.74%	56.95	44.10%	23.43%	44.01%	51.70%	
+ Pseudo Preference	43.59%	57.74	41.35%	20.82%	41.10%	51.91%	25
+ Human Preference	**48.23%**	**61.97**	**49.35%**	**26.62%**	**48.47%**	**56.63%**	
+ Summarize	39.88%	41.87	32.03%	14.11%	31.69%	40.46%	
+ Pseudo Preference	37.80%	40.53	35.77%	18.48%	34.89%	42.94%	26
+ Human Preference	36.75%	40.48	30.82%	12.08%	29.53%	42.44%	
Domain: Insurance (Cluster num : 27)							
Baseline	36.16%	50.63	29.89%	9.97%	28.83%	44.21%	
+ Pseudo Preference	41.49%	50.76	**31.62%**	**11.77%**	**31.41%**	46.04%	26
+ Human Preference	**42.16%**	**52.14**	27.44%	9.33%	26.06%	**47.60%**	
+ Summarize	38.03%	39.48	24.07%	7.67%	24.07%	36.55%	
+ Pseudo Preference	35.71%	38.03	18.78%	6.85%	18.58%	34.20%	26
+ Human Preference	36.46%	40.66	22.44%	7.52%	22.16%	37.67%	

Table 1: Evaluation of different model variants across the **Finance** and **Insurance** domains in the DSTC12 theme detection task. Accuracy and NMI assess clustering quality, while ROUGE and cosine similarity evaluate the natural language quality of theme labels.

- **Clusters**: The number of clusters selected during inference, determined automatically via silhouette analysis.

Models. We experiment with combinations of the following components:

- **Summarization**: Each dialogue is abstracted using an LLM to a concise form starting with "User wants..." or "User needs...", preserving the core intent while removing surface-level noise (Section 3.1).
- **Human Preference**: Gold pairwise constraints derived from given dataset, which contains `should-link` and `cannot-link` pairs. The number of oracle pairs was 347 (Finance) and 282 (Insurance).
- **Pseudo Preference**: Automatically generated pairwise preferences using a preference synthesize model. These were used to guide post-clustering reassignment. We generated 1,836 pairs for Finance and 1,888 for Insurance.

4.2 Main Results

Table 1 presents the performance of different model variants across the Finance and Insurance domains. We initially hypothesized that incorporating LLM-based summarization and pseudo label preference refinement would improve clustering quality and label generation. However, the empirical results reveal several unexpected trends.

First, LLM-based summarization consistently degraded performance across both domains. While intended to reduce lexical variability, the summarization process often produced overly generic descriptions that failed to preserve the underlying intent of the original utterances. As a result, crucial semantic cues were lost, making it harder to distinguish thematically distinct examples during clustering (Section 5.1).

Second, pseudo labeled preference pairs offered limited improvements over the baseline. In some cases, it slightly boosted accuracy or label quality, but the gains were inconsistent and notably weaker than those achieved using gold (human) preference pairs. This gap highlights the challenge of training a generalizable preference predictor to unseen domain.

Finally, we observe that the predicted number of clusters tended to be underestimated, particularly in the Finance domain where the system often selected 25–26 clusters compared to the gold 34. This under-segmentation likely stemmed from the lack of user-preferred granularity being reflected during the clustering stage, leading to coarse groupings that failed to capture fine-grained thematic distinctions. These findings highlight the importance of incorporating preference signals earlier in the pipeline, a point we further explore in Section 5.

5 Analysis

In this section, we investigate the sources of failure observed in our main results by analyzing the effects of summarization, pseudo labeling, and clustering performance. We provide case studies and discuss potential directions for improvement.

5.1 Summarization

In Table 2, we illustrate how using full dialogue context for summarization—rather than focusing solely on the user turn where the theme label is

Examples of Summarization	
Original	*User: My email address is Hawthorne Thornton at ... dot com.* *System: I will get this right out to you. Also, you are trying for a ten thousand dollar loan currently.* *User: Yeah, how long's it gonna take for me to know if I get approved?* *System: You'll get a letter in the mail.* ***User: Hopefully it's ten thousand... In the meantime, what address do you have on file for me? I just wanna make sure it's the right one.***
Label	`get account info`
Summary	`User wants to apply for a corporate credit card.`
Original	*System: Sir I do think that Elgin is a wonderful town but I've lived here my whole life so I might be biased...* *User: I think that sounds like a very good idea. Very wise. Yes, a very wise idea. but I'm still not sure if this is a risk. I guess what I really need to is to talk to someone about the risks involved...* *System: Sure. It sounds to me you're asking if we have a risk specialist that you could speak with? Am I understanding that correctly?* *User:* ***Yes a rest risk specialist! That's exactly what I need!...***
Label	`request call transfer`
Summary	`User wants to assess the risks involved in opening a second store location in Elgin before deciding on a lease.`

Table 2: Examples where LLM-based summarization includes excessive contextual information, potentially reducing clustering accuracy.

assigned—can negatively impact clustering. While the initial motivation for incorporating previous dialogue was to better capture the user's intent, we observed that the resulting summaries often included excessive background rather than highlighting the intention expressed in the current turn.

For example, in the first case, the summary reflects the broader discussion about applying for a corporate credit card, rather than the user's immediate request to verify their mailing address. Similarly, in the second example, the summary emphasizes the user's interest in evaluating business risks in Elgin, but overlooks the specific request to speak with a risk specialist made in the labeled turn. These cases suggest that focusing too heavily on prior context can dilute the turn-level signal needed for accurate theme clustering. To address this issue, future summarization approaches should center the summary around the labeled turn, using surrounding context only to disambiguate or clarify intent—not to replace it.

5.2 Limitations of Pseudo-Labeled Preferences

To assess the accuracy of the pseudo labeling model, we compare its predictions against gold preference label in the Finance and Insurance domains (Table 3). The accuracy hovers around 50%, suggesting challenging in alignment with human preferences for unseen domains.

Domain	Finance	Insurance
Accuracy (%)	50.58	49.11

Table 3: Accuracy of pseudo labeling model on the unseen domains.

Examples of Synthesized Preference Prediction	
Label	`inquire about plans`
Utt 1	Could you tell me when my auto policy premium is due?
Utt 2	Well, I needed to cancel one of my insurance plans.
Prediction	`should-link` (correct)
Label	`update account information`
Utt 1	Hey I would like to my home address.
Utt 2	Can I update my billing frequency then?
Prediction	`should-link` (correct)
Label	`start/change/cancel plan`
Utt 1	Life insurance. Add a policy the cheapest one you have. Have young son who is an adult coming back home. Out of drug rehab again.
Utt 2	Hello, Sarah. I would like to cancel my auto insurance.
Prediction	`cannot-link` (incorrect)
Label	`get plan info`
Utt 1	Yes my name is Jack and I got a flyer for you guys saying that you offer homeowner's insurance in my area and I just wanted to see what you could offer me.
Utt 2	OK, and what would the annual rate be, if I decided to pay it all at once?
Prediction	`cannot-link` (incorrect)

Table 4: Examples of pseudo labeling model predictions. Top two rows show correctly predicted `should-link` cases, while bottom two rows show incorrect `cannot-link` predictions.

Table 4 analyzes common success and failure cases of the pseudo-labeled preference model. In particular, labels covering multiple intents (e.g., `start/change/cancel plan`) pose challenges, as the model tends to treat these actions as distinct. In contrast, it performs reliably on simpler intents such as `update account information`. These findings suggest that zero-shot generalization is challenging, as clustering standards assumed by users may vary across domains—highlighting that even minimal in-domain preference data can help the model better align with human judgments of appropriate clustering boundaries.

Label	Predictions
apply for loan	apply for loan, business loan inquiry, inquire about sba seven a loan
check credit card balance	check business silver card balance, check credit card balance
cancel/order check	order checks, cancel checks, update account
change account or card pin	change pin number
get debt income ratio	debt to income ratio
request call transfer	None
get net income	None

Table 5: Examples of label-to-prediction mappings in the finance domain.

Label	Predictions
change password	reset password
file life claim	get life insurance info, file life claim, enroll in life insurance
get pet quote	get pet insurance quote, inquire about pet insurance
create account	create account, set up account
pay bill	pay bill, understand cost
get homeowner quote	None
file poperty claim	None

Table 6: Examples of label-to-prediction mappings in the insurance domain.

5.3 Importance of Initial Clustering

Lastly, we apply human-annotated intent preferences to the clustering output to assess the importance of initial cluster quality. Specifically, Tables 5 and 6 present a comparison between gold labels and the predicted clusters after incorporating human-provided should-link and cannot-link constraints. Despite applying these authentic preferences during post-processing, we still observe substantial mismatches, indicating that preference-based refinement alone may not resolve structural issues in the initial clustering.

For example, in the Finance domain, utterances labeled as *apply for loan* are split into clusters like *business loan inquiry* and *SBA loan*, while *check credit card balance* appears as variants such as *check business silver card balance*. Some intents, like *request call transfer* and *get net income*, are missing altogether.

These results suggest that when the initial clustering does not align with the semantic scope assumed by the preference supervision, post-processing becomes ineffective. Even correct preference signals cannot recover from such misaligned segmentations. These findings highlight the need for future work to incorporate user preferences earlier in the pipeline—particularly during the embedding and clustering stages—to better estimate the number of clusters and achieve semantically aligned groupings.

6 Conclusion

Motivated by the need for controllable coherent theme induction in dialogue systems, we explore the use of pseudo-labeled preference post-processing to refine initial clustering outputs. Our findings reveal that while preference-based post-processing provides a structured way to improve cluster quality, its effectiveness is fundamentally constrained by the quality of the initial clustering. Through extensive experiments on the Finance and Insurance domains in the DSTC12 dataset, we observe that coarse-grained or misaligned clusters severely limit the corrective power of preference modeling. These results highlight the critical importance of aligning initial representations with user-intended semantics, suggesting that improvements to clustering quality may yield greater benefits than post-hoc refinement alone.

Limitations

While our analysis provides insights into the limitations of post-hoc preference modeling, our approach has several constraints. First, the pseudo preference labels are generated using in-domain data and a fine-tuned LLM, which not generalize well to other domains without additional supervision. Second, we employ a fixed clustering backbone and only apply preferences as a refinement step—more tightly coupled clustering and preference modeling might yield better results.

Acknowledgements

This work was supported by the following research programs: the Smart HealthCare Program funded by the Korean National Police Agency (KNPA) (No. RS-2022-PT000186, 47.5%), the ITRC (Information Technology Research Center) Program through the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (Ministry of Science and ICT) (No. IITP-2025-RS-2024-00437866, 47.5%), and the Artificial Intelligence Graduate School Program at POSTECH through

the IITP grant funded by the Korea government (MSIT) (No. RS-2019-II191906, 5%).

References

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Lu Fan, Guangfeng Yan, Qimai Li, Han Liu, Xiaotong Zhang, Albert YS Lam, and Xiao-Ming Wu. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1050–1060.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, et al. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of NAACL-HLT*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023. Natcs: eliciting natural customer support dialogues. *arXiv preprint arXiv:2305.03007*.

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 Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *2016 IEEE spoken language technology workshop (SLT)*, pages 414–419. IEEE.

Hyukhun Koh, Haesung Pyun, Nakyeong Yang, and Kyomin Jung. 2023. Multi-view zero-shot open intent induction from dialogues: Multi domain batch and proxy gradient transfer. *arXiv preprint arXiv:2303.13099*.

DSTC12 Organizers. 2025. Dstc12: Controllable conversational theme detection track. https://dstc12.dstc.community/.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. *arXiv preprint arXiv:1908.11487*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shun Zhang, Jian Yang, Jiaqi Bai, Chaoran Yan, Tongliang Li, Zhao Yan, and Zhoujun Li. 2024. New intent discovery with attracting and dispersing prototype. *arXiv preprint arXiv:2403.16913*.