DSTC 2025

The Twelfth Dialog System Technology Challenge (DSTC12)

Proceedings of the Workshop

August 28, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-330-2

Introduction

We are excited to welcome you to DSTC-12, the Twelfth Dialog System Technology Challenge. This year the workshop is being held on August 28th, 2025 at SIGDial 2025.

The DSTC shared tasks have provided common testbeds for the dialog research community since 2013. This year, the program includes two invited talks from Verena Rieser and Milica Gašić, two track presentations, four paper presentations and a panel with Ryuichiro Higashinaka, Laurent Prévot, Tetsuro Takahashi, Milica Gašić.

We had two tracks this year: Track 1: Dialog System Evaluation: Dimensionality, Language, Culture and Safety, organized by John Mendonca, Lining Zhang, Rahul Mallidi, Alon Lavie, Isabel Trancoso, Luis Fernando D'Haro, João Sedoc Track 2: Controllable Conversational Theme Detection Track, organized by Igor Shalyminov, Hang Su, Jake Vincent, Siffi Singh, Jason Cai, James Gung, Raphael Shu, Saab Mansour

The track timelines are: Jan – April 2025 training phase April 2025 testing phase May 2025 results announcement

We are very proud of this years DSTC and would like to thank some key players. We would like to thank our session chairs (Emre Can, Alexandru Coca), our panelist moderator Luis Fernando D'Haro, all our track organizers and a big thank you to all the track participants for their fine work. Finally we would like to thank the wider DSTC community for helping making DSTC a success and helping it to grow for over 10 years.

Behnam Hedayatnia, General Chair

Michel Galley, Publicity Chair

Raghav Gupta, Publication Chair

Zhang Chen and Yun-Nung (Vivian) Chen, Workshop Co-Chairs

Organizing Committee

General Chair

Behnam Hedayatnia, Apple

Workshop Chairs

Vivian Chen, National Taiwan University
Zhang Chen, National Taiwan University, Taiwan

Publication Chair

Raghav Gupta, Google

Publicity Chair

Michel Galley, Microsoft

Program Committee

Program Chairs

Yun-Nung Chen

Michel Galley, Raghav Gupta

Behnam Hedayatnia

Chen Zhang

Reviewers

Namo Bang

Qian Chen, Anoop Cherian, Paul A. Crook

Suvodip Dey

Michel Galley, Kallirroi Georgila

Behnam Hedayatnia, Vojtech Hudecek, Vojtech Hudecek

Seongho Joo

Seokhwan Kim, Sarvesh Kirthivasan

Aditya Nair

Alexandros Papangelis, Baolin Peng

Saurav Sahay, Sashank Santhanam, Harsh Sharma, Prachee Sharma

David Thulke

Bin Wang, George Z Wei

Qi Zhu

Invited Talk

Intentional, Plural, Deep: The Foundations of Beneficial AI Alignment

Verena Rieser
Google Deepmind

August 28, 2025 – Time: **09:45 – 10:25** –

Abstract: We talk constantly about making AI aligned,but we rarely ask the most important questions: aligned to what, and to whom? This keynote deconstructs our current assumptions and proposes that truly beneficial AI requires moving through three distinct depths of alignment. The first is Intentionality: we must stop assuming alignment will emerge on its own and start making deliberate choices about the goals we set. The second is Plurality: we must abandon the search for a single gold standardof human values and engineer systems that thrive on diversity. The third and most profound is Depth: we must push beyond optimizing for clicks and convenience and instead align AI with the complex, often contradictory, nature of long-term human well-being. This is a framework for the next frontier of AI — one that builds technology to support our human nature, not exploit it.

Bio: Verena Rieser is a Senior Staff Research Scientist at Google DeepMind, where she founded and lead the VOICES (Voices-of-all in alignment) team. Their mission is to ensure that powerful AI models like Gemini are developed responsibly, making them safe and genuinely useful for diverse communities worldwide. Verena's career has been driven by a fascination with conversational AI. Verena has pioneered research in dialogue systems and natural language generation, always with a focus on applications that create societal benefit. Before joining Google, Verena was a full professor directing the NLP lab at Heriot-Watt University and held a Royal Society Leverhulme Senior Research Fellowship

Invited Talk

Dimensions of intelligence

Milica Gašić
Heinrich-Heine-University Düsseldorf

August 28, 2025 – Time: **15:00 – 15:40** –

Abstract: Large language models (LLMs) have transformed the area of artificial intelligence, achieving or surpassing human performance in a number of natural language processing tasks. Despite this tremendous success, they lack the ability to model multi-turn goal-directed conversation, they are largely uncalibrated and there is little insight into the way they operate. In this talk, I will present (1) multi-turn optimisation which integrates emotion, (2) confidence-based learning based on insights from human psychology and (3), in the direction of explainability, observations from topological data analysis applied to LLMs. To conclude, I will hypothesise how combining ideas from LLMs and task-oriented systems can lead to conversational agents encompassing a large spectrum of desired properties.

Bio: Milica Gašić is a Professor in Dialog Systems and Machine Learning at Heinrich Heine University. Her research focuses on fundamental questions of human-computer dialogue modelling and lie in the intersection of Natural Language Processing and Machine Learning. She is a recipient of the European Research Council Starting Grant and the Alexander von Humboldt Sofja Kovalevskaja Award. Prof. Gašić is a member of the International Scientific Advisory Board of DFKI, a member of ACL, a member of ELLIS and a senior member of IEEE. She served as the vice-president of SIGDIAL 2021-2025.

Table of Contents

Neural Models and Language Model Prompting for the Multidimensional Evaluation of Open-Ended Conversations

Michelle Elizabeth[*,1,2], **Alicja Kasicka**[*,1], **Natalia Krawczyk**[*,1],
Magalie Ochs[2], **Gwénolé Lecorvé**[1], **Justyna Gromada**[1], **Lina M. Rojas-Barahona**[1]
[1]Orange Research [2]Aix-Marseille University
michelle.elizabeth@orange.com, alicja.kasicka@orange.com, natalia1.krawczyk@orange.com,
magalie.ochs@lis-lab.fr, gwenole.lecorve@orange.com, justyna.gromada@orange.com, lina.rojas@orange.com

Abstract

The growing number of generative AI-based dialogue systems has made their evaluation a crucial challenge. This paper presents our contribution to this important problem through the Dialogue System Technology Challenge (DSTC-12, Track 1), where we developed models to predict dialogue-level, dimension-specific scores. Given the constraint of using relatively small models (i.e. fewer than 13 billion parameters) our work follows two main strategies: employing Language Models (LMs) as evaluators through prompting, and training encoder-based classification and regression models. Our results show that while LM prompting achieves only modest correlations with human judgments, it still ranks second on the test set, outperformed only by the baseline. The regression and classification models, with significantly fewer parameters, demonstrate high correlation for some dimensions on the validation set. Although their performance decreases on the test set, it is important to note that the test set contains annotations with significantly different score ranges for some of the dimensions with respect to the train and validation sets.

1 Introduction

Real-life dialogues are unpredictable and dynamic, making them difficult to reproduce in static corpora. Consequently, dialogue systems are typically evaluated with either simulated users or real users (Zhu et al., 2022). However, a significant gap exists between these approaches, leading to unrealistic simulations or subjective human evaluations (Cordier et al., 2023; Elizabeth et al., 2025). Despite its subjectivity, human evaluation is preferred. In the seminal framework PARADISE (Walker et al., 1997), subjective metrics, such as user satisfaction, were estimated based on objective metrics through linear regression. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ibarz et al., 2018) utilizes regression models as reward models to evaluate the output of Language Models (LMs) (Ouyang et al., 2022) for better alignment to human preferences. This may provide a rationale for high correlation between LM and human judgments (Kazi et al., 2024; Gunasekara et al., 2021). These results suggest that regression models can be a promising approach to conversation evaluation.

Track 1 of DSTC-12 "Dialog System Evaluation: Dimensionality, Language, Culture and Safety" (Mendonça et al., 2025) focuses on automatic evaluation of open-domain dialogues for ten dimensions, at the dialogue level. The challenge incorporates widely-used dimensions such as overall quality, *Relevance*, and *Proactivity*, alongside less conventional ones including *Empathy*, *Trust*, and *Skill*. This provides a valuable opportunity to assess the correlations between human judgments and automatic evaluation techniques across each dimension. For this purpose, we present four distinct approaches for dialogue-level evaluation that were submitted to the challenge by our team *ORALIS*.

This work covers three possible representations of the scores and one combination of these approaches:

i the most straightforward approach, treating scores as real numbers and predicting them through a *regression* task;

ii treating scores as classes, since they are integers that correspond to categories of evaluation (such as good, average, poor), which leads to training *classifiers*;

iii treating scores as tokens among others as handled in autoregressive LMs, and thus using *LM prompting* to generate scores.

iv a final strategy, referred to as the *hybrid* approach, consists of mixing predictions from diverse approaches for various dimensions.

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 1–16

According to the results, none of our systems outperforms the baseline (a prompted *Llama-3.1-8B-Instruct*[1] LM) in terms of average *absolute* correlation on the test set. However, our approaches outperform the baseline on most individual dimensions. The LM prompting system shows better generalization to the test set than other approaches, performing better on some dimensions than on the validation set. The regression and classification models demonstrate strong *positive* correlations with human scores on the validation set but achieve lower *absolute* correlation scores when applied to unseen examples, suggesting overfitting. The classification approach, while ranking lowest overall alongside the hybrid method, excels on six dimensions including *Empathy*, outperforming all other approaches in terms of number of winning dimensions. The hybrid method, which selects the best-performing approaches on the validation set (combining LM prompting and regression while excluding classification), does not generalize well to the test data. These results can also be explained by the fact that there are inconsistencies between the training-validation sets and the test set, especially regarding the score distribution and score ranges as depicted in Figure 2 and Figure 3.

The paper is organized as follows: Section 2 provides a literature review on dialogue evaluation; Section 3 introduces the datasets and dimensions used in our work, while Section 4 details the four implemented evaluators. Finally, Section 5 reports the results of the validation set (as used to develop the evaluators) as well as on the test set (as used to rank the submitted evaluators in the challenge).

2 Related Work

This section discusses evaluation paradigms, recent advances in automatic and LM-based metrics, current multi-dimensional frameworks and open challenges.

Open-ended conversational AI systems require multi-dimensional assessment due to the complex nature of dialogue, where multiple valid responses exist for any given context. Key dimensions include *coherence* (the contextual appropriateness and logical consistency of responses (Bao et al., 2021)), *engagement* (sustaining user interest (Venkatesh et al., 2018)), *informativeness* (providing relevant content (Bao et al., 2021)), *specificity* (context-tailored responses (Harrison et al., 2023)), *consistency* (avoiding contradictions (Bao et al., 2021)), and *factual correctness* (minimizing hallucinations and ensuring accurate information (Bao et al., 2021)). Additional dimensions include *fluency*, *personality*, and *context management* (maintaining memory across multiple turns (Wang et al., 2024)).

Traditional reference-based metrics, e.g. BLEU or ROUGE, show weak correlation with human judgments in open-domain dialogue (Liu et al., 2016; Saleh et al., 2020). While human evaluation remains most reliable (Li et al., 2019; Venkatesh et al., 2018), it is costly, time-consuming, and can suffer from inconsistency (Ji et al., 2022; Smith et al., 2022).

Researchers have developed various automatic metrics to overcome evaluation challenges. Embedding-based metrics capture semantic similarity beyond surface-level lexical overlap but struggle with catching conversational nuances. Regression models, inspired by PARADISE (Walker et al., 1997) and RLHF (Christiano et al., 2017), and classification models are constrained by availability and quality of training data. Reference-free metrics like FED (Mehri and Eskénazi, 2020) evaluate responses in context with better human alignment. LM-based evaluation uses LMs as judges, showing stronger correlation with human ratings (Lin and Chen, 2023; Yu et al., 2024), despite challenges including self-preference bias (Chen et al., 2025) and sensitivity to response characteristics and context complexity (Xu et al., 2025).

The adoption of LMs as judges (Gunasekara et al., 2021; Kazi et al., 2024) enables scalable evaluation, although concerns about annotation quality persist. Small LMs offer cost-effective alternatives and have recently shown strong potential as capable judges. Although they may seem less accurate due to their size, recent work (Deshpande et al., 2024) shows that well-aligned LMs with around 3B parameters can achieve the performance of much larger systems.

Evaluation campaigns like DSTC-9 (Gunasekara et al., 2021) and DSTC-11 (Soltau et al., 2023) have advanced the field through interactive and multilingual evaluation tracks. However, the most successful systems still rely heavily on fine-tuned generative models or LMs for data augmentation, and achieving high correlation with human judgments remains a challenge, especially for multi-dimensional conversation aspects.

Modern evaluation frameworks assess conversa-

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

tion quality across several interdependent dimensions like coherence, engagement, and context management (Bao et al., 2021; Harrison et al., 2023; Wang et al., 2024). Multi-dimensional LM-based approaches (e.g. LLM-Eval (Lin and Chen, 2023), MT-Bench (Bai et al., 2024), KIEval (Yu et al., 2024)) offer thorough assessments yet face challenges with bias, generalizability, and scalability.

Despite significant progress in automatic evaluation of conversational systems, current methods still face limitations in robustness, interpretability, and scalability, highlighting the need for improved multi-dimensional approaches that reliably reflect human perceptions of conversational quality across diverse scenarios.

3 Datasets and Dimensions

We use three datasets in this work: DSTC-12 (Mendonça et al., 2025), the official competition dataset; CONTURE (Gunasekara et al., 2021), which was used in the DSTC-9 evaluation campaign and FED (Mehri and Eskenazi, 2020), another dataset published for dialogue evaluation research.

As detailed later in this section, these datasets predominantly contain open-ended human-machine dialogues annotated by humans on dialogue-level for various evaluation metrics, although the FED dataset also includes human-human dialogues.

3.1 DSTC-12 Dataset and Metrics

DSTC-12 (Mendonça et al., 2025) is an official dataset released as part of the competition. It contains 185 open-domain human-machine dialogues in English. Each dialogue covers a wide variety of everyday topics such as personal stories, preferences and recommendations, and fact-based planning queries. Detailed dataset statistics are displayed in Table 1. What is worth noting is the significant difference between the length of utterances in DSTC-12, compared to other datasets. Still, in all datasets, the average number of words in machine turns is significantly greater than in human utterances.

The dialogues were evaluated by human annotators across ten dimensions. According to the organizers, human annotators were either MTurk workers or lab members, thus different dialogues might have been annotated by different annotators. This combination of research staff and online workers might raise concerns about potential inconsistencies in how dimensions were scored across conversations.

Unfortunately, the annotated data is highly imbalanced, as not all dialogues have scores assigned for each evaluation dimension. While each dialogue received scores for at least four dimensions, the coverage varies considerably. Only 54 out of 185 dialogues (29%) are annotated with all ten dimensions, while the majority include annotations for only four or five. At the dimension level, annotation counts are also unevenly distributed: most dimensions are well represented with over 120 annotations, whereas *Overall* appears in less than one-third of the dialogues. These patterns are illustrated in Figure 1. The test set consists of 120 dialogues, with varying coverage of annotations as in the training set.

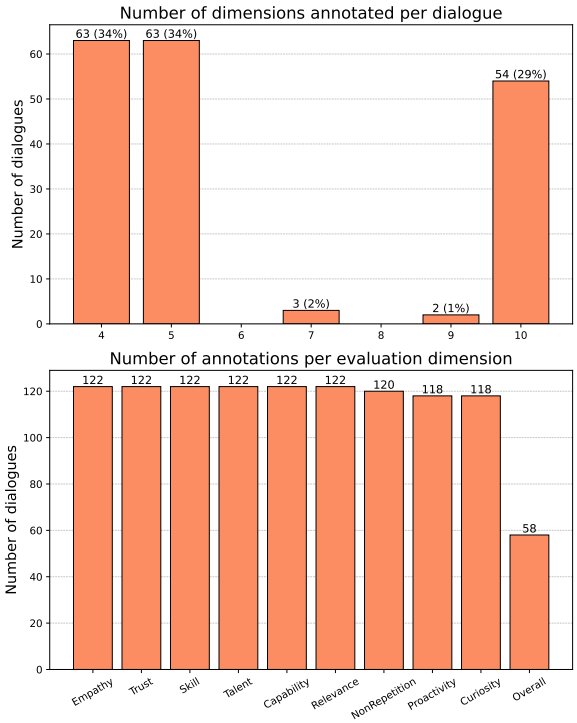


Figure 1: Distribution of the dialogues in the DSTC-12 dataset (train/validation) based on the scores (top), and the dimension (bottom).

In addition, dimensions have different score ranges. Their names, along with their ranges, are: *Empathy* (1-12), *Trust* (0-5), *Curiosity* (0-100), *Proactivity* (0-100), *NonRepetition* (0-100), *Relevance* (0-100), *Overall* (0-100), *Skill* (0-5), *Talent* (0-5), and *Capability* (0-5). For dimensions such as *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*, the majority of the human scores are between 6

	DSTC-12		FED	ConTurE
	train	test		
#Dialogues	185	120	125	119
#Ann. per Dialogue	1	1	5	3
Avg. #turns	15	21	6	9
Avg. #words per turn (H)	25	51	6	7
Avg. #words per turn (M)	130	193	12	19

Table 1: Statistics for all the datasets (Ann. stands for annotations, H stands for human and M stands for machine).

FED & ConTurE		DSTC-12
Inquisitive	—>	Curiosity
Avg(Informative, Coherence)	—>	Relevance
Topic depth	—>	Talent
Flexible	—>	Proactivity
Diverse	—>	Non-repetition
Likeable	—>	Empathy
Consistent	—>	Trust
Understanding	—>	Capability
Error recovery	—>	Skill

Table 2: FED & ConTurE to DSTC-12 mapping.

and 10, despite the score range being 0-100. The distributions of the scores are uneven, particularly for dimensions with the 0-100 range, see Appendix A, Figure 2. In the test set, the score ranges are between 1-5 for *Skill, Talent, Capability, Trust, and Overall* while the range is 1-10 for *Empathy, Relevance, NonRepetition, Proactivity and Curiosity*, which differs from the score ranges observed in the training set. See Appendix A, Figure 3.

Metrics: The challenge assesses the evaluators based on the mean *absolute* Spearman correlation with human judgments. In our experiments, we also consider the mean *positive* correlation with human judgments.

3.2 FED and ConTurE

The official dataset released for the challenge, i.e. DSTC-12, is rather small, containing only 185 dialogues. Each dialogue was annotated by only one evaluator. To increase data diversity and avoid overfitting when training our regression and classification models, we utilized two other open-domain human–machine dialogue datasets: ConTurE (Gunasekara et al., 2021), with 119 dialogues, which was proposed in DSTC-9 Track 3, and FED (Mehri and Eskenazi, 2020), with 125 dialogues, introduced at SigDial 2020. In all these three datasets, the dialogues predominantly involve interactions between a human and a machine, although the FED dataset also includes human-human interactions with one participant simulating a machine. The conversations cover a variety of everyday topics such as personal preferences, opinions, popular culture, and general knowledge, resembling natural and informal interactions.

In contrast to DSTC-12 dataset, the ConTurE dataset has each dialogue annotated by three different raters, while in FED dataset each dialogue was annotated by five different evaluators. To ensure that every dialogue contributes exactly one score per dimension to our models, just as in the DSTC-12 dataset, and to prevent over-representation of multi-rated dialogues, we first averaged the dimension scores in both ConTurE and FED. Then, we mapped the dimension names from the additional datasets to match those in the DSTC-12 dataset. This mapping was based on the heuristics shown in Table 2, where we paired each metric from ConTurE and FED with the DSTC-12 dimension that best reflected its core intent.

In both external datasets, the dimensions are annotated on different scales: for most dimensions, annotations are on a 3-point scale, except for Consistent, which is binary, and Overall quality, which is on a 5-point scale.

After averaging the raw scores per dialogue, we applied a linear rescaling function to fit each dimension into the corresponding dimension range in the DSTC-12 dataset:

$$q = (p - A)\,\frac{D - C}{B - A} + C \qquad (1)$$

where p is the original value in $[A, B]$ and q is the mapped value in $[C, D]$. Finally, we rounded q to the nearest integer to obtain the final score on the target dimension scale in the DSTC-12 dataset. We split the official DSTC-12 dataset equally into train and validation sets for each dimension. The training set was further enhanced by adding the dialogues from the two additional datasets.

4 Evaluators

In this section, we first describe the baseline system (provided by the challenge organizers) and then present the evaluation systems submitted to the challenge. Our first approach is based on the LM-as-a-judge paradigm, namely *LM Prompting*. The next two systems are classic neural models: *regression* and *classification*. Finally, the last system is a hybrid evaluator that combines the *regression* model with the *LM Prompting* system.

4.1 Baseline

This system was proposed by the organizers and it is a fine-tuned *Llama-3.1-8B-Instruct*[2] pretrained model for content safety classification, prompted for dialogue-level evaluation. For all of the dimensions the same prompt template was used, which included all evaluation dimensions in one prompt.

4.2 LM Prompting

We tested various prompting methods: (i) basic prompt with just the name of the dimension, see example in Appendix C.1; (ii) zero-shot learning with the definition of the dimension, see example in Appendix C.2; (iii) one-shot learning, see example in Appendix C.3; (iv) few-shot learning (with 3 samples), see example in Appendix C.4 and (v) self-consistency prompting, see example in Appendix C.5.

We assigned various roles to the LM (such as "crowd-worker", "expert", or "human evaluator") and provided task descriptions with score ranges at varying levels of detail.

For one-shot and few-shot learning methods, we provided examples with their assigned scores in two formats: either as conversation excerpts or as summarized dialogue. For the few-shot learning, we randomly sampled three dialogues: one with the lowest possible score, one from the median, and one with the highest possible score. For one-shot learning, we randomly sampled a dialogue with a score around the median. In self-consistency prompting, the LM was provided with a short description of the meaning of the scores within the score range, as well as was asked to check the validity of its response and fix it, if needed, before responding.

In every prompting method, the LM was asked to return the score along with a short explanation for the given score. Various versions of the prompts were evaluated on each dimension separately, and different prompts were selected for later study based on these preliminary results (listed in the Appendix B).

We explored several dialogue context strategies to feed the LM: the last 40% of the conversation, the first 40% of the conversation, the first 20% and last 20% of the conversation, a summarized version of the dialogue, or the full dialogue. The summarised version was obtained by summarising each utterance using *Llama 3.1 8B Instruct* model, as it performs well across variety of tasks (Grattafiori et al., 2024). We tested two different summarisation prompts for sentences whose length exceeded 200 words, in conversations with more than 3000 words in total. These prompts mainly differed by the maximum number of tokens in their summarised versions: either 50 (*summarisation 1*) or 150 (*summarisation 2*). The exemplary prompt can be found in the Appendix, section B.11.

We considered three state-of-the-art LMs: *Deepseek Llama 8B*[3], *Deepseek Qwen 7B*[4], and *Qwen 2.5 7B Instruct 1M*[5]. These models were selected based on their demonstrated effectiveness on multiple natural language processing tasks (Guo et al., 2025; Yang et al., 2025).

We tested various combinations of prompting on the validation set. We modified the prompt, the dialogue context, and utilized distinct LMs. Then, we selected the best performing combination for each dimension, i.e., achieving the highest *positive* correlation values with human annotations. The chosen configuration for each dimension is shown in Table 3.

Analysis of the standard deviation values for correlation results for different models (average std=0.09) and for different dialogue contexts (average std=0.12) implies that the latter, on average, impacts the final correlation result slightly more than the choice of the model.

Systems' performances on the validation set and test set are presented in Table 4 and Table 5, respectively.

4.3 Regression

We trained a regression model for each dimension. The model's architecture consists of a regression layer on top of a ModernBERT Large encoder (Warner et al., 2024). It is worth noting that ModernBERT has a context limit of 8K tokens allowing for encoding a larger dialogue context, in contrast to BERT-family models (Devlin et al., 2019), which are limited to only 512 tokens. ModernBERT is also notably smaller than LMs, with fewer than 1 billion parameters (395 million). We utilized the score ranges provided in the challenge

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[4] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M

Dimension	Prompting	Dialogue part	Language Model
Empathy	zero-shot	last 40%	Qwen 2.5 7B Instruct
Trust	zero-shot	full conversation	Qwen 2.5 7B Instruct
Skill	zero-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Talent	zero-shot	summarisation 1	Qwen 2.5 7B Instruct
Capability	zero-shot	summarisation 1	Qwen 2.5 7B Instruct
*Capability	zero-shot	summarisation 2	Deepseek Qwen 7B
Relevance	few-shot	summarisation 1	Deepseek Llama 8B
*Relevance	few-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
NonRepetition	few-shot	first 40%	Deepseek Qwen 7B
*NonRepetition	few-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Proactivity	zero-shot	first 40%	Deepseek Llama 8B
*Proactivity	zero-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Curiosity	zero-shot	summarisation 2	Deepseek Qwen 7B
*Curiosity	zero-shot	first 40%	Deepseek Llama 8B
Overall	zero-shot	full conversation	Qwen 2.5 7B Instruct

Table 3: LM prompting approach: Chosen methods and models for each evaluation dimension, * refers to the combination that obtained the highest absolute correlation on the validation dataset.

dataset and the mean-square error as the loss function. To generalize better and avoid overfitting, we utilized CONTURE and FED datasets in addition to the DSTC-12 dataset, using the mapping introduced in Section 3.2. Regression performance on the validation and test sets is presented in Table 4 and Table 5, respectively. A later experiment with varying values of weight decay for training, did not show any considerable improvement in the correlations on the test set.

4.4 Classification

Similar to the regression system, we trained individual classifiers for each dimension on our combined training set. All dialogues were encoded using *Sentence-BERT (SBERT)*[6]. Since we require discrete categories, each integer score was rescaled to an integer range using Equation 1 and rounded to the nearest integer.

Model development followed a two-stage grid search. In the first stage, we explored different class ranges and selected [0, 8] based on validation performance. In the second stage, we tuned *Multi-Layer Perceptron (MLP)* hyperparameters separately for each dimension to maximize *positive* validation Spearman correlation. To reduce overfitting, we specifically optimized regularization and training duration balancing convergence with generalization. The best hyperparameters for each dimension were used to train the final classifiers and predict dimension scores on the test set. Predicted scores were rescaled back into the original ranges, using the inverse of Equation 1. The validation and test performance of our classifiers trained with SBERT encodings are presented in Table 4 and Table 5, respectively.

To further address potential overfitting, we applied ModernBERT encodings as in the regression models, combined with label smoothing. However, these changes resulted in slightly lower test set correlations, suggesting that increased model capacity was not sufficient to improve performance.

4.5 Hybrid

The hybrid system combines methods that performed well on the validation set for each dimension, as shown in Table 4. When selecting these methods, we limited our choice to only regression and LM prompting approaches, excluding the classification method from consideration. This decision was based on the prioritisation of approaches with stronger contextual understanding and generalization capabilities. In the regression system we utilized ModernBERT that has a larger context window, in comparison to the SBERT model used in the classifier. This allows us to process more dialogue context. We strategically chose LM prompting for several dimensions due to its demon-

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Dimension	LM prompting	Regression	Classification	Hybrid
Empathy	0.3	0.23	**0.35**	0.3
Trust	**0.38**	-0.02	-0.07	0.38
Skill	**0.33**	-0.09	-0.06	0.33
Talent	0.26	**0.41**	0.15	**0.41**
Capability	0.17	**-0.21**	0.00	0.17
Relevance	0.19	**0.79**	0.71	0.79
NonRepetition	0.16	**0.75**	0.68	0.75
Proactivity	0.01	**0.79**	0.66	0.79
Curiosity	-0.02	**0.68**	0.65	0.68
Overall	0.4	0.27	**0.49**	0.4
Abs. Average	0.22	**0.42**	0.38	**0.5**

Table 4: Correlation between the gold labels and system's outputs on the validation set for each system. **Bold** values indicate the highest *absolute* correlation across all systems.

strated ability to generalize well to unseen data (Wang et al., 2023). Additionally, our classification approach returns discrete integer values, e.g., on the scale 0-8, requiring mapping to, e.g., the 0-100 scale, and potentially introducing approximation errors, while both regression and prompting methods produce continuous values within the desired range without the need for additional mapping. This combination of enhanced contextual processing and a potential for better generalization influenced the choice of methods for our hybrid system.

It is worth noting that all three approaches, i.e. LM prompting, regression, and classification, were submitted to the challenge separately.

The regression system was chosen for the following dimensions: *Talent*, *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*. For the remaining dimensions, i.e. *Empathy*, *Trust*, *Skill*, *Overall*, and *Capability*, the LM-prompting was chosen as it obtained the most promising results on the validation set, see Table 4. The results of our hybrid system on the test set are shown in Table 5. This system underperformed on this dataset in comparison to its scores on the validation set.

5 Results

The results of our systems on the test set are presented in Table 5, along with the baseline approach published by the DSTC-12 challenge organizers.

The baseline is based on the LM-as-a-judge approach, similar to one of our systems; however, it uses a different LM and different prompt.

The *absolute* average correlation on the test set for all systems is relatively low, between 0.14 and 0.15, while the baseline achieves 0.17. This represents a significant decrease from the validation set, where the regression and hybrid systems achieved values between 0.4 and 0.5 (see Table 4).

None of our systems achieved a higher average *absolute* score than the baseline; however, our approaches outperform the baseline on most of the individual dimensions. The baseline has higher scores only for the *NonRepetition* and *Overall* dimensions. Nevertheless, the difference on the *NonRepetition* dimension is significant enough to influence the absolute average score for the whole system.

Each of our approaches outperforms the baseline on multiple dimensions in terms of the *absolute* score. The classification approach performs best, in terms of number of winning dimensions, exceeding the baseline on six dimensions, while the LM prompting, regression, and hybrid approaches each outperform on five dimensions. All four of our systems outperform the baseline on *Empathy*, *Capability* and *Proactivity*, and three of them excel on *Talent* as well.

Performance patterns vary across dimensions. The classification approach maintains its strength for *Empathy* from validation to test set in terms of *absolute* correlation, though with reduced values. For *Talent* and *Capability*, the regression system outclasses other approaches across both sets. However, some dimensions show inconsistent results, for example, LM prompting excels on *Trust* on the validation set but its performance drops significantly on the test set. On the test set, the regression system shows the opposite trend for this dimension,

Dimension	LM prompting	Regression	Classification	Hybrid	**Baseline**
Empathy	-0.08	**0.17**	**-0.17**	-0.08	0.06
Trust	0.01	**0.2**	0.13	0.01	-0.11
Skill	**-0.22**	0.07	-0.02	**-0.22**	-0.1
Talent	0.05	**0.24**	0.22	**0.24**	0.1
Capability	0.13	**0.24**	0.12	0.13	0.07
Relevance	0.08	-0.1	**-0.28**	-0.1	0.23
NonRepetition	0.11	0.14	-0.0	0.14	**0.39**
Proactivity	**-0.15**	0.08	0.2	0.08	-0.02
Curiosity	**0.37**	0.09	0.08	0.09	0.23
Overall	**0.31**	0.13	-0.17	**0.31**	**0.38**
Abs. Average	**0.15**	**0.15**	0.14	0.14	**0.17**

Table 5: Correlation between the gold labels and systems' outputs on the test set. **Bold** values indicate the highest *absolute* correlation across all systems.

performing better than on the validation set.

We observe significant performance decrease between validation and test sets for several dimensions for regression and classification systems, suggesting potential overfitting. The regression system shows drastic decreases for *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*, despite achieving correlations of 0.68-0.79 on the validation set. The classification system demonstrates similar patterns on the same dimensions, with correlations of 0.65-0.71 on the validation set. Nevertheless, it maintains superior performance for *Relevance* on the test set.

Interestingly, LM prompting demonstrates the opposite pattern for some dimensions, performing better on the test set than on the validation set. It achieved the highest *absolute* correlations on test set for *Proactivity*, *Curiosity*, and the *Overall* dimension, despite weaker results on the validation set.

Inspecting both Table 4 and Table 5 raises concerns about why some dimensions show *negative* correlation values. One possible explanation lies in the conceptual mismatch between how LMs and humans interpret evaluation metrics. The inconsistent score ranges between the training and test set also leads us to question the quality of the annotations. Dimensions may have been understood differently by annotators and models, leading to inconsistent judgments that weakened or even inverted expected correlations. Evaluation systems often reflect individual user experiences shaped by emotion and subjectivity, making consistent human assessment especially difficult (Fan and Luo, 2020). Furthermore, scoring chatbot responses remains a fundamentally subjective and challenging task even for human evaluators, which increases the likelihood of annotation noise in human labels (Yuwono et al., 2019).

6 Conclusions and Future Work

In this paper, we present four distinct dialogue-level evaluators for different dimensions that were submitted to the DSTC-12 challenge. We explored distinct prompting strategies, including varying the dialogue context across different LMs. We also trained very small regression and classification models on the challenge dataset enriched with other evaluation datasets (CONTURE and FED). We also considered a hybrid system that combines the LM prompting and regression approaches. Furthermore, we analyzed the data and found that there are inconsistencies between the training-validation sets and the test set, in terms of the score distribution and score ranges. Although our systems did not outperform the baseline, classical approaches, such as regression and classification, show interesting results, competitive with larger models of 7 and 8 billion parameters used in LM prompting approach.

In terms of future work, we first suggest enhancing the quality of the dataset in a dedicated annotation campaign. Second, we would like to explore domain adaptation techniques for training models on similar but larger datasets from distinct sources (such as the *DSTC-11* dataset) to overcome data scarcity.

7 Limitations

The scaling laws have shown that the impressive capabilities of LLMs are highly influenced by three factors: the size of the model, the size of the dataset, and the amount of computing power used for training (Kaplan et al., 2020). All LMs used in our experiments have fewer than 13B parameters. The regression and classification models have fewer than 1B parameters.

We tuned our systems to maximize the *positive* correlation, however the systems were ranked based on the *absolute* correlation.

Moreover, the dataset provided in the challenge is quite small, making it difficult to use for training regression and classification models. The mapping we made between the annotation of the additional datasets and the DSTC-12 dataset is entirely subjective, which may require in depth investigation to study the impact of various mappings. It would have been beneficial to have the instructions provided to human annotators for a more accurate mapping as well as to define the dimensions more accurately for the LM prompting method.

Finally, there are concerns regarding the consistency of the annotations for certain dimensions, since their score ranges vary significantly between the train-validation and the test sets.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *Annual Meeting of the Association for Computational Linguistics*.

Siqi Bao, H. He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zhengyu Niu. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *AACL/IJCNLP*.

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. 2025. Do llm evaluators prefer themselves for a reason? *arXiv preprint arXiv:2504.03846*.

Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741.

Thibault Cordier, Tanguy Urvoy, Fabrice Lefèvre, and Lina M. Rojas Barahona. 2023. Few-shot structured policy learning for multi-domain and multi-task dialogues. In *EACL*.

Darshan Deshpande, Selvan Sunitha Ravi, Sky CH-Wang, Bartosz Mielczarek, Anand Kannappan, and Rebecca Qian. 2024. Glider: Grading llm interactions and decisions using explainable ranking. *arXiv preprint arXiv:2412.14140*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondrej Dusek, and Lina M. Rojas-Barahona. 2025. Exploring react prompting for task-oriented dialogue: Insights and shortcomings. In *IWSDS*.

Yifan Fan and Xudong Luo. 2020. A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1202–1209. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and 1 others. 2021. Overview of the ninth dialog system technology challenge: Dstc9. *Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vrindavan Harrison, Rishi Rajasekaran, and M. Walker. 2023. A transformer-based response evaluator for open-domain spoken conversation. *arXiv.org*.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *ArXiv*, abs/1811.06521.

Tianbo Ji, Yvette Graham, Gareth J. F. Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. *Annual Meeting of the Association for Computational Linguistics*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tür, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 913–920. IEEE.

Margaret Li, J. Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv.org*.

Yen-Ting Lin and Yun-Nung (Vivian) Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *NLP4CONVAI*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Conference on Empirical Methods in Natural Language Processing*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and M. Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *SIGDIAL Conferences*.

John Mendonça, Lining Zhang, Rahul Mallidi, Luis Fernando D'Haro, and João Sedoc. 2025. Overview of dialog system evaluation track: Dimensionality, language, culture and safety at dstc 12. In *DSTC12: The Twelfth Dialog System Technology Challenge*, 26th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Avignon, France.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart M. Shieber. 2020. Probing neural dialog models for conversational understanding. *NLP4CONVAI*.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and J. Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *NLP4CONVAI*.

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Wei Han, and Yuan Cao. 2023. DSTC-11: Speech aware task-oriented dialog modeling track. In *Proceedings of the 11th Dialog System Technology Challenge (DSTC-11)*, pages 226–234, Prague, Czech Republic. Association for Computational Linguistics.

Anu Venkatesh, Chandra Khatri, A. Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, R. Prasad, Ming Cheng, Behnam Hedayatnia, A. Metallinou, Rahul Goel, Shaohua Yang, and A. Raju. 2018. On evaluating and comparing conversational agents. *arXiv.org*.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: a framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, page 271–280, USA. Association for Computational Linguistics.

Jun Wang, Jiamu Zhou, Muning Wen, Xiaoyun Mo, Haoyu Zhang, Qiqiang Lin, Cheng Jin, Xihuai Wang, Weinan Zhang, and Qiuying Peng. 2024. Hammerbench: Fine-grained function-calling evaluation in real mobile device scenarios.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025. Does context matter? contextual-judgebench for evaluating llm-based judges in contextual settings. *arXiv preprint arXiv:2503.15620*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *Annual Meeting of the Association for Computational Linguistics*.

Steven Kester Yuwono, Biao Wu, and Luis Fernando D'Haro. 2019. Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer.

Qi Zhu, Christian Geishauser, Hsien chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format. *arXiv preprint arXiv:2211.17148*.

A Additional figures

We provide the distribution of human annotations by dimension for both datasets provided by the organizers: the training set in Figure 2, and test set in Figure 3.

B Selected Prompts

In this section we present the selected prompts.

B.1 Relevance

```
You are an expert evaluator tasked with assessing the relevance of chatbot's answers.
Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.
The final output should include the score (0-100) and your explanation for the given score.
Here are the examples of the excerpts of the conversations and the score these conversations received. Chatbot's and user's utterances are separated using "';"'.
Excerpt from the example conversation: "'{excerpt1}"'
Score for the example conversation: "'{score1}"'
(...)
The conversation for evaluation:
{conversation}
```

B.2 Proactivity

```
Act like a human evaluator tasked with assessing the proactivity of chatbot queries.
Proactivity refers to the system's ability to anticipate user's future problems, needs, and changes. A proactive chatbot often takes initiative and guides the conversation.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot is not proactive at all, and 100 suggests that the chatbot often takes initiative and anticipates the needs of the user.
The final output should include the score (0-100) and your explanation for the given score.
The conversation for evaluation:
{conversation}
```

B.3 NonRepetition

```
Act like a human evaluator tasked with assessing the chatbot's ability to avoid repeating responses within a conversation.
Non-repetition refers to the system's ability to avoid repeating information or questions the user has already provided. A chatbot with strong non-repetition capabilities ensures a smoother conversation by recognising and adapting to previously shared inputs.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot often repeats itself, and 100 suggests that the chatbot has strong non-repetition capabilities.
The final output should include the score (0-100) and your explanation for the given score.
Here are the examples of the summaries of the conversations (you will be evaluating a full conversation, not the summary) and the score these conversations received.
Summary of the example conversation: "'{summary1}"'
Score for the example conversation: "'{score1}"'
(...)
The conversation for evaluation:
{conversation}
```

B.4 Trust

```
You are an expert evaluator tasked with assessing how trustworthy the chatbot seems to the user.
Trustworthy chatbot is a chatbot that seems sincere, reliable, and honest, whose responses seem true and not harmful or intended to trick the user.
The final output should include the score (from the range 0-5) and your explanation for the given score.
The conversation for evaluation:
```

Figure 2: Distribution of human annotations by evaluation dimension in the training set.

{conversation}

B.5 Skill

You are a human evaluator tasked with assessing the *skill* of the chatbot in this dialogue.
Skill means how well the chatbot executes the task or responds to the user's input. Consider how accurate, clear, and appropriate the responses are.
Give a score between 0 and 5, and provide a short explanation for your score.
Dialogue:
{conversation}

Figure 3: Distribution of human annotations by evaluation dimension in the test set.

B.6 Capability

You are a human evaluator tasked with assessing the capability of responses.
Evaluate only capability (how effectively the chatbot fulfils user needs and achieves the purpose of the conversation). Do not assess any other dimension. Focus only on whether the chatbot meets or exceeds the user's expectations.
Give a score between 0-5 and a brief explanation for your score.
Dialogue to evaluate:
{conversation}

B.7 Empathy

You are an expert evaluator tasked with assessing the level of empathy of the chatbot in the conversation. Chatbot that displays high levels of empathy is the one that shows understanding, awareness, sensitivity to the feelings, thoughts, and experience of the user.
The final output should include the score (from the range 1-12) and your explanation for the given score.
The conversation for evaluation:
{conversation}

B.8 Curiosity

You are an expert evaluator tasked with assessing the curiosity of the chatbot in the conversation. Curiosity refers to how well the chatbot engages the user and shows interest in the responses by asking questions encouraging further interactions.
The final output should include the score (from the range 0-100) and your explanation for the given score.
The conversation for evaluation:
{conversation}

B.9 Talent

You are a crowdworker asked to rate the chatbot's *talent* in this conversation.
Talent means how naturally or intelligently the chatbot handles the conversation.
Was it thoughtful, clever, or showed any spark of conversational ability? Use your instinct- if it felt smart or interesting, that's talent.
Give a score from 0 to 5 and a short reason for your choice.
Dialogue:
{conversation}

B.10 Overall

Evaluate the following conversation between a user and a chatbot. The evaluation should be for the responses generated by the chatbot.
Give an integer score the scale of 0-100 to evaluate the overall impression, where 0 indicates the worst score possible and 100 indicates the best score possible.
The final answer must contain an integer in the range 0-100 and the reason for giving the score.
Here is the conversation to evaluate:
{conversation}

B.11 Summarisation prompt

Prompt:
You are an expert copywriter tasked with shortening a chatbot's utterances from a conversation between a chatbot and a user.
Objective:
Shorten the chatbot's response while preserving its original communication style and all relevant details necessary for later evaluation. Ensure that the short version remains faithful to the chatbot's intent, tone, and structure.
Guidelines:
- Retain all details that could be useful for evaluating the chatbot's performance.
- Encode proper names that are irrelevant to the evaluation (e.g., specific phone models) using placeholders like [model-name1].
- Return the shortened dialogue as a string.
- The summary must not exceed 50 words.
Chatbot's utterance to shorten:
{conversation}
Output: A concise yet comprehensive concise version of the chatbot's response (max 50 words).

C LM Prompts examples

In this section we present some outputs of the distinct prompt strategies.

C.1 Basic prompt example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers. Assess only the chatbot, not the user. The final output should include the score (from the range 0-100) and your explanation for the given score.
The conversation for evaluation:
{conversation}

C.2 0-shot learning example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers.
Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.
The final output should include the score (0-100) and your explanation for the given score.
The conversation for evaluation:
{conversation}

C.3 1-shot learning example

You are an expert evaluator tasked with assessing the relevance of chatbot's answers.
Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.
The final output should include the score (0-100) and your explanation for the given score.
Here is an example excerpt of the conversation and the score this conversation received. Chatbot's and user's utterances are separated using "';"'.
Excerpt from the example conversation: "'{excerpt}"'
Score for the example conversation: "'{score}"'
The conversation for evaluation:
{conversation}

C.4 Few-shots learning example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers.
Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.
Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.
The final output should include the score (0-100) and your explanation for the given score.
Here are the examples of the excerpts of the conversations and the score these conversations received. Chatbot's and user's utterances are separated using "';"'.
Excerpt from the example conversation: "'{excerpt1}"'
Score for the example conversation: "'{score1}"'
Excerpt from the second example conversation: "'{excerpt2}"'
Score for the second example conversation: "'{score2}"'
Excerpt from the third example conversation: "'{excerpt3}"'
Score for the third example conversation: "'{score3}"'
The conversation for evaluation:
{conversation}

C.5 Self-consistency prompting example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers. Assess only the chatbot, not the user.
Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.
Rate the chatbot's relevance on a scale from 0 to 100, where:
- 0-20: Very low relevance - The chatbot's responses are mostly irrelevant or off-topic. Users may find the answers confusing or unhelpful.
- 21-40: Low relevance - The chatbot provides some relevant information, but many responses are not aligned with the user's queries. Users may struggle to find useful insights.
- 41-60: Moderate relevance - The chatbot's answers are somewhat relevant, with a mix of useful and irrelevant information. Users may find some value but will likely encounter inconsistencies.
- 61-80: High relevance - The chatbot generally provides relevant and useful answers. Most responses align well with user queries, though occasional irrelevant information may still appear.
- 81-100: Very high relevance - The chatbot consistently delivers highly relevant and useful responses. Users can rely on the answers to be directly related to their queries, enhancing their experience significantly.
Return the score (0-100) along with a concise explanation of why the chatbot received that score.

```
Think like a domain expert and check the validity
of your score. Fix the score if needed.
Dialogue for Evaluation:
{conversation}
```

CATCH: A Controllable Theme Detection Framework with Contextualized Clustering and Hierarchical Generation

Rui Ke[1], Jiahui Xu[1], Kuang Wang[1], Shenghao Yang[1], Feng Jiang[2]*, Haizhou Li[1,3]
[1]SRIBD, School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong
[2]Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology
[3]Department of ECE, National University of Singapore
jiangfeng@suat-sz.edu.cn

Abstract

Theme detection is a fundamental task in user-centric dialogue systems, aiming to identify the latent topic of each utterance without relying on predefined schemas. Unlike intent induction, which operates within fixed label spaces, theme detection requires cross-dialogue consistency and alignment with personalized user preferences, posing significant challenges. Existing methods often struggle with sparse, short utterances and fail to capture user-level thematic preferences across dialogues. To address these challenges, we propose CATCH (Controllable Theme Detection with Contextualized Clustering and Hierarchical Generation), a unified framework that integrates three core components: (1) context-aware topic representation, which enriches utterance-level semantics using surrounding topic segments; (2) preference-guided topic clustering, which jointly models semantic proximity and personalized feedback to align themes across conversations; and (3) a hierarchical theme generation mechanism designed to suppress noise and produce robust, coherent topic labels. Experiments on a multi-domain customer dialogue benchmark demonstrate that CATCH achieves state-of-the-art performance in both theme classification and topic distribution quality. Notably, it ranked second in the official blind evaluation of the DSTC-12 Controllable Theme Detection Track, showcasing its effectiveness and generalizability in real-world dialogue systems.

1 Introduction

In real-world customer service scenarios such as banking, finance, travel, and insurance, accurately identifying the underlying theme of each utterance plays a pivotal role in enhancing service efficiency, understanding user intent, and retrieving relevant knowledge. Compared to intent classification, which typically maps utterances to a predefined label space (Pu et al., 2022; Costa et al., 2023), theme detection aims to uncover potentially novel and latent topics. Controllable theme detection requires not only accurate topic assignment within dialogues (Nguyen et al., 2022; Du et al., 2013a), but also consistency across dialogues and alignment with user preferences (Mendonça et al., 2025), as illustrated in Figure 1.

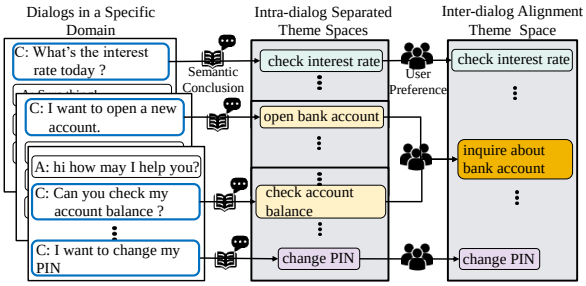


Figure 1: Illustration of the controllable theme detection task. Given a set of dialogues with unlabeled utterances, a theme is generated for each utterance. The theme granularity is influenced by auxiliary inputs such as user preferences, indicating whether a pair of utterances should be grouped under the same theme.

However, existing approaches such as topic modeling (Blei et al., 2003; Pham et al., 2024) fall short of these requirements. While such methods infer high-level themes using neural or probabilistic models, they often struggle to maintain consistency across dialogues due to the sparsity and fragmentation of utterances (Bach et al., 2021; Lin et al., 2024). Other works, like topic clustering, typically rely on semantic similarity between utterances but ignore how thematic consistency should reflect user-specific preferences (Gung et al., 2023; Chatterjee and Sengupta, 2020). Moreover, most previous methods lack an explicit theme generation, limiting their applicability in downstream tasks.

To address these challenges, we propose **CATCH** (**C**ontrollable **A**nd **T**hematic **C**lustering with **H**ierarchy), a controllable theme detection framework that integrates intra-dialogue context

*Feng Jiang is the corresponding author.

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 17–26

modeling with inter-dialogue user preference alignment. Specifically, CATCH consists of three key components: (1) a context-aware topic representation module that leverages dialogue-level topic segmentation to enrich semantic understanding; (2) a preference-guided topic clustering that jointly considers semantic similarity and user preferences for cross-dialogue thematic consistency; and (3) a hierarchical theme generation inspired by Chain-of-Thought prompting and refined through majority voting to produce robust, domain-adaptive outputs.

We evaluate CATCH on the DSTC-12 Controllable Conversational Theme Detection benchmark. Experimental results demonstrate that our framework outperforms competitive baselines in both in-domain and cross-domain settings, even under limited preference supervision. Our system ranks **second** in the official blind evaluation, achieving strong performance in both automatic and human assessments with a lightweight design. Extensive ablation and case studies further validate the robustness and generalizability of our approach. The main contributions of this work are as follows:

- We propose **CATCH**, a novel controllable theme detection framework that jointly models intra-dialogue contextual signals and inter-dialogue user preferences, effectively addressing the limitations of prior topic modeling and clustering methods.
- We design a hierarchical theme generation strategy that first generates topic candidates in small clusters and then refines them via majority voting, ensuring robustness and coherence.
- CATCH achieves 2nd place in the DSTC-12 Controllable Theme Detection task across both automatic and human evaluation settings.
- Detailed ablation studies and qualitative analysis demonstrate the effectiveness of each module and highlight the framework's generalizability in low-resource scenarios.

2 Related Works

The related task of theme detection in conversation can be broadly categorized into two levels based on granularity: **intra-dialogue** and **inter-dialogue** theme detection.

2.1 Intra-dialogue theme detection

Intra-dialogue theme detection focuses on identifying the topic affiliation of each utterance within a single dialogue, which typically includes two subtasks: *topic segmentation* and *topic generation*.

Topic segmentation. Dialogue Topic Segmentation (DTS) aims to divide a dialogue into coherent topical units by detecting boundaries between adjacent utterances. Hindered by scarce annotated dialogue data and dialogue fragmentation, which limits effective transfer from documents, most DTS approaches focus on unsupervised scenarios. Early methods use unsupervised signals such as word co-occurrence statistics (Hearst, 1997; Eisenstein and Barzilay, 2008) or topical distributions (Riedl and Biemann, 2012; Du et al., 2013b). Recent studies construct contrastive data sets through utterance-pair distances and fine-tuning models like BERT (Devlin et al., 2019; Xing and Carenini, 2021; Gao et al., 2023). However, these methods apply the same segmentation decoding algorithm uniformly across datasets with varying topic granularities, failing to account for dataset-specific differences and resulting in uneven performance.

Topic generation. The most direct way to generate a topic is through topic modeling, which trains a neural network or probabilistic model to infer abstract high-level themes of the input text (Blei et al., 2003; Pham et al., 2024). One main challenge to applying the topic model to theme detection is the sparsity of data, which is rendered by the brevity of short texts (Bach et al., 2021; Lin et al., 2024). Many topic models try to augment the short data into a long training signal to address the data sparsity problem (Lin et al., 2024; Nguyen et al., 2022; Tuan et al., 2020; Jiang et al., 2024). Although the topic model performs well in theme generation, existing work cannot maintain the consistency of the theme label within the same conversation scenario.

2.2 Inter-dialogue theme detection

Inter-dialogue theme detection, on the other hand, concerns the clustering and alignment of topics across multiple dialogues.

Topic clustering and alignment. The main approach to yield coherent topics between dialogues is topic clustering. Existing work generates topic groups by directly clustering the semantic representation of input text (Nguyen et al., 2024; Grootendorst, 2022; Zhang et al., 2022; Sia et al., 2020). These works are efficient in providing a coherent theme distribution. However, they assume that the theme is a fixed set and exclude theme discovery from the design (Perkins and Yang, 2019). Some methods are also proposed to explore the realistic

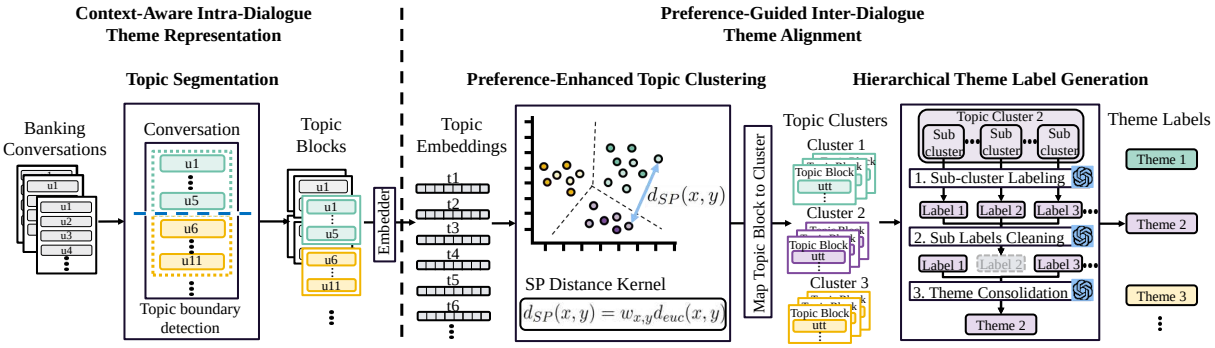


Figure 2: The overall architecture of CATCH.

complexity of the theme space (Perkins and Yang, 2019; Chatterjee and Sengupta, 2020; Gung et al., 2023). These methods use topic alignment to explain the topic space. Some works design the multi-view clustering method (Nguyen et al., 2024, 2025; Perkins and Yang, 2019), such as learning clustering representations by predicting cluster assignments of an alternative view of each input (Perkins and Yang, 2019) and iteratively breaking down the "noise" cluster from DBSCAN to address varying densities (Chatterjee and Sengupta, 2020). Others used intermediate structured prediction tasks, such as dependency parsing or abstract meaning representations, to aid intent induction (Liu et al., 2021; Zeng et al., 2021; Vedula et al., 2020). However, these works align the topic based on the semantic information without considering user preference.

3 Methodology

We define the controllable theme detection (TD) task as a structured theme generation problem over dialogue utterances. Given a set of utterances $U = \{u_1, \ldots, u_m\}$ extracted from dialogues of a specific domain, the goal is to assign each utterance $u_i \in U$ a theme label L_i that is both preference-aligned and contextually consistent across dialogues To achieve this goal, as illustrated in Figure 2, we propose **CATCH**, a controllable theme detection framework that incorporates both intra- and inter-dialogue modeling.

3.1 Context-Aware Intra-Dialogue Theme Representation

To address the semantic sparsity and ambiguity commonly observed in short utterances, we design a context-aware intra-dialogue theme representation module. It leverages a dual-branch topic segmentation framework to infer latent segment boundaries and construct thematically coherent spans by scoring relevance between adjacent utterance pairs with a two-stage adaption consisting of **unsupervised pre-training** and **preference-supervised fine-tuning**.

Inspired by DialSTART (Gao et al., 2023) the dual encoder evaluates topic similarity through a combination of semantic similarity and dialogue coherence in a dual-encoder framework: A SimCSE-based **topic encoder**, which produces an embedding for each individual utterance, capturing its semantic content; An NSP-BERT-based **coherence encoder**, which evaluates discourse continuity between intervals of utterance spans.

3.1.1 Unsupervised Pre-training

Concretely, given a dialogue $D = \{u_1, ..., u_n\}$, we define $n-1$ intervals v_i between u_i and u_{i+1} and assign each interval a topic relevance score r_i which is a calculated by topic representations h_i and h_{i+1}, and coherence score c_i. Higher r_i indicates higher topic continuity. To ensure both encoders learn topic-aware utterance representations from unlabeled dialogue data, we employ two auxiliary tasks:

Neighboring Utterance Matching, which focuses on utterance-level semantic similarity by encouraging closer alignment between adjacent utterance embeddings. Given an utterance u_i, its similar neighboring utterance index set U_i and dissimilar non-neighboring utterance index set $\bar{U}_i$ as:

$$U_i = \{j \in [1,n] \mid w \geq |i-j| \wedge j \neq i\}, \quad (1)$$

$$\bar{U}_i = \{j \in [1,n] \mid w < |i-j|\}, \quad (2)$$

where w specifies the number of neighboring utterances on each side of u_i. We encode each utterance using the topic encoder to obtain its vector representation. During training, the topic encoder maximizes a marginal ranking loss that pushes representations of $\{u_i, u_j\}$ pairs with $j \in U_i$ pairs

closer together than those with $j \in \bar{U}_i$.

Relevance Modeling, which leverages both semantic similarity and discourse coherence at the utterance-interval level to distinguish real contiguous fragments from synthetic ones. Given an utterance interval v_i, its real fragment F_i and synthetic fragment $\bar{F}_i$ are defined as:

$$F_i = \{[u_{i-1}, u_i], [u_{i+1}, u_{i+2}]\}, \quad (3)$$

$$\bar{F}_i = \{[u_{i-1}, u_i], [u_{rand}, u_{rand+1}]\}\}. \quad (4)$$

where u_{rand} is an utterance randomly selected from other dialogues. We then feed both interval pairs in F_i and $\bar{F}_i$ into two separate encoders: a topic encoder to compute the topic similarity and a coherence encoder to compute a coherence score. Summing these two values produces the relevance scores r_i^+ (for the real fragment) and r_i^- (for the synthetic fragment). During training, a margin-based ranking loss is applied to maximize the gap between r_i^+ and r_i^-, encouraging the model to assign higher relevance to genuine sequences.

3.1.2 Preference-supervised Fine-tuning

To encourage the model to better identify topical shifts and coherence patterns that align with human preferences, we refine the topic and coherence encoders by leveraging human-annotated preference utterance indices—each corresponding to a likely topic boundary—as supervision signals. Given a preference-labeled index set $L = \{l_1, l_2, ..., l_m\}$ corresponds to m annotated utterances in all dialogue set, we filter the original training data $[U_i, \bar{U}_i, F_i, \bar{F}_i]$ to construct new training sets $[U_p, \bar{U}_p, F_p, \bar{F}_p]$, where p belongs to L. Finally, we fine-tune both the topic and coherence encoders by continually optimizing the marginal ranking losses for the NUM and RM tasks over this filtered training set.

After the two-stage training process, we apply the TextTiling algorithm (Hearst, 1997) to the predicted relevance scores $R = \{r_1, r_2, ..., r_{n-1}\}$. A fixed threshold of 0.5 is used to identify topic boundaries. Based on the detected boundaries, we segment the entire dialogue into coherent topical blocks, each representing a contiguous span of utterances that share a common theme.

3.2 Preference-Guided Inter-Dialogue Theme Alignment

To align topic blocks across dialogues with user preference, we propose a preference-enhanced topic clustering that jointly considers semantic similarity and preference feedback. Then, we introduce a hierarchical LLM-based theme label generation method that effectively filters out noisy signals and ensures more robust and coherent theme generation for the preference-enhanced cluster.

3.2.1 Preference-Enhanced Topic Clustering

To dynamically fuse semantic similarity and user preference signals within the clustering process, we design a Preference-Enhanced Topic Clustering strategy with a new semantic-preference (SP) distance kernel to substitute the original distance metric. It measures the distance between a pair of utterances (x, y) in the semantic-preference union space:

$$d_{SP}(x, y) = w_{x,y} \cdot d_{sem}(x, y) \quad (5)$$

where d_{sem} is the Euclidean distance between topic embeddings, and $w_{x,y}$ is a preference scalar learned via a reward model trained on user preference data: should-link / cannot-link topic block pairs (the detail form of preference data is shown in Section 4.1). Notably, the generated preference scalar indicates the tendency of whether a pair of topics should belong to the same theme.

Because the true joint space combining semantic and preference information is latent and not explicitly constructed, the over-defined problem arises as shown in Figure 3. Therefore, we propose a two-stage algorithm grounded in semantic space but progressively incorporating preference signals, by first obtaining anchor semantic clusters as reference node, and then re-cluster the points with intense preference tendency using SP distance.

- **Semantic Clustering**. To acquire anchor clusters aligning the semantic similar topics, topic blocks are clustered solely based on semantic similarity to form initial anchor clusters.

- **SP Distance Clustering**. This stage first uses preference reward model to provide preference scalar (tendency). Two opposite kinds of preference-relevant topic pairs are obtained according to the linking and splitting tendency thresholds. Preference-relevant topic pairs are split from the anchor cluster, and then re-clustered to the nearest anchor node with minimum aggregated SP distance.

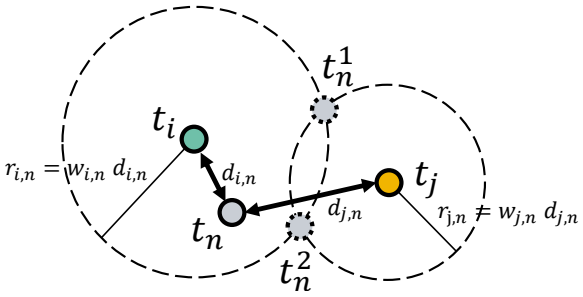


Figure 3: The illustration of positional conflict evoked by SP distance metric. Assume topic t_n's position is defined by its semantic distance ($d_{i,n}$, $d_{j,n}$) to the topic t_i and t_j. If the original distance is replaced by the SP distance ($r_{i,n}$, $r_{j,n}$), t_n has two possible positions (t_n^1, t_n^2) in the SP union space without observation of coordinates.

3.2.2 Hierarchical Theme Label Generation

We design a three-step prompting pipeline to generate a structurally coherent theme label for each preference-enhanced cluster. The hierarchical design endows the pipeline with the ability to effectively conclude key information by introducing a cleaning mechanism amid two theme generation processes, though the semantic inconsistency within a preference-enhanced cluster introduces noise that prohibits direct generation (Yang et al., 2025; Liu et al., 2024).

Sub-cluster Labeling. This step centers at a *divide and conquer* strategy which randomly divides a cluster into several smaller groups (e.g., 10 topic blocks) and prompts an LLM to separately generate fine-grained theme labels. Specifically, the prompt requires the theme label be a concise and actionable verb phrase.

Label Cleaning. In this step, we design a cleaning rule to reduce noise among the set of fine-grained labels, because these primary labels are highly inconsistent due to their fine granularity. Specifically we prompt the LLM to summarize and filter these labels into a consistent set by removing rare or irrelevant entries.

Theme Consolidation. The final theme label for each cluster is generated by prompting the LLM to unify the cleaned labels. This step ensures preference alignment and semantic coherence by summarizing on the key theme information rather than extracting the superficial semantic meaning.

This hierarchical label generation strategy not only enables global label consistency, but also mitigates the impact of clustering errors. If a preference cluster is mistakenly separated, the hierarchical design ensures that the same theme label will be generated for all these clusters, thereby merging them into the same cluster.

4 Experiments

4.1 Datasets

Datasets. We conduct experiments on the multi-domain customer support dialogue datasets (Banking, Finance, Insurance, and Travel) provided by DSTC-12 (Mendonça et al., 2025), as summarized in Table 1. Each dataset contains two key types of annotations of the themed utterances: (1) Theme Annotation: Each target utterance is annotated with its corresponding theme label. (2) Preference Annotation: A binary relation (*should-link* and *cannot-link*) of a pair of target utterances indicating whether they should be grouped under the same theme (should-link) or not (cannot-link).

In the offline evaluation, we use the banking dataset for training and the finance and insurance domains as the valid dataset. For online evaluation, we deploy our model (CATCH) to predict theme labels on the Travel dataset, which lacks golden annotations. The predicted results are submitted to the organizers of DSTC-12 for blind evaluation. Throughout training, CATCH is trained solely based on preference annotations without accessing the ground-truth theme labels.

Type	Domain	# Dialogues	# Utterance	# Preference
Offline	Banking	1634	58418 (980)	164/164
Offline	Finance	1725	196764 (3000)	173/173
Offline	Insurance	836	60352 (1333)	155/126
Online	Travel	765	72010 (999)	— / —

Table 1: Data Statistics of the DSTC-12 Dataset. The numbers in parentheses indicate the number of sampled utterances with annotated themes. In the *Preference* column, the values denote the number of should-link / cannot-link utterance pairs, respectively.

4.2 Metrics

Metrics. To comprehensively evaluate the effectiveness of CATCH, we follow the DSTC-12 (Mendonça et al., 2025) evaluation protocol, which assesses two core aspects: (1) the quality of theme segmentation (i.e., utterance clustering), and (2) the quality of generated theme labels.

Offline Evaluation. For theme segmentation quality, we use two standard clustering metrics: **Normalized Mutual Information (NMI)** (Vinh et al., 2010), which quantifies the mutual dependence between predicted and reference clusters nor-

malized by their entropies, and **Clustering Accuracy (Acc)**, computed via the Hungarian algorithm to align clusters optimally. For theme label quality, we evaluate the semantic and textual correspondence between predicted and reference labels using: **Cosine Similarity (CosSim)** based on Sentence-BERT embeddings, **ROUGE** (Lin, 2004) for n-gram overlap, and an **LLM-based score** that assesses label format and informativeness via vicuna-13B evaluation guided by human-crafted criteria.

Online Evaluation. For the held-out test set without golden labels, the DSTC-12 organizers perform additional evaluations including both automatic metrics and manual human judgments.

4.3 Baselines

We compare our framework with the following baselines:

GURP (generation on utterance by random preference assignment): The official baseline provided by DSTC-12, which directly generates a theme label for the utterance cluster after randomly linking or splitting the utterance pairs according to the preference data. **GTR** (generation on topic guided by reward model): An upgraded version of GURP, which directly generates themes for topic clusters, and uses a preference reward model to guide the random linking and splitting. **SPC** (semantic-preference clustering): A variation of GTR, which directly uses SP distance metric to cluster topics.

4.4 Implementation Details

In the intra-dialogue stage, we follow the previous work (Gao et al., 2023), using *bert-base-uncased* and *sup-simcse-bert-base-uncased* as our coherence encoder and topic encoder, respectively. During the pre-training and fine-tuning process, we both set the learning rate to 5e-6 and the epoch to be 3.

In the inter-dialogue stage, we employ *all-mpnet-base-v2* to obtain the sentence transformer embeddings and uses **UMAP** to reduce embedding dimension. For semantic clustering, we employ **Spectrum** clustering method with the default clusters number K being 30 following the common design. For the preference refinement, we use *bert-base-uncased* as default reward model with learning rate to be 2e-5 and epoch to be 3. During preference inference, we set the confidence threshold of linking θ_l to be 0.85 and the confidence threshold of splitting θ_s to be 0.15. For the theme label generation, we employ *LLaMA3-8B-Instruct* as the default LLM for label generation.

4.5 Offline Experimental Results

We train CATCH on the banking dataset and conduct the experiment in two data scenarios: in-domain data, out-of-domain data. In the in-domain task, we evaluate different methods by evaluating them on the same banking dataset. For the out-of-domain task, we evaluate on the finance and insurance datasets, respectively. Moreover, we provide the results of the blind evaluation of DSTC-12, which is tested on the travel dataset with extra metrics.

4.5.1 The Performance of the Models in the In-domain Dataset

Table 3 highlights the effectiveness of CATCH which outperforms all the baselines under both theme distribution and theme label quality. The proposed preference-enhanced topic clustering significantly improves the quality of topic distribution, as reflected in the superior ACC (55.8%) and NMI (67.1%) metrics comparing to GTR which achieves second best ACC (46.9%) and NMI (51.6%). Besides, CATCH significantly significantly enhances the theme label quality. The hierarchical generation paradigm is able to conclude a representative high-level theme from the diverse topics cluster as demonstrated by the superior ROUGE-1 (35.3%) and Cosine Similarity (58.5%) comparing to GTR's ROUGE-1 (22.0%) and Cosine Similarity (37.3%).

4.5.2 The Performance of the Models in the Out-of-domain Dataset

Since CATCH performs well on the in-domain task, we further validate its domain generalization ability on the out-of-domain task. The results are presented in Table 2. CATCH demonstrates its robustness and consistency, since it maintains the superior performance in both datasets across all metrics. Consequently, CATCH performs even better in the out-of-domain task (e.g. with 67.1% NMI for finance dataset) than in the in-domain task (e.g. with 65.4% NMI). For the theme label quality, CATCH achieves 42.4 % ROUGE-L in finance dataset and 41.8% ROUGE-L in insurance dataset, which both outperform the 35.3% ROUGE-L for in-domain task on banking dataset. This indicates the significant effectiveness and generalization ability of the hierarchical generation paradigm.

Notably, CATCH achieves better results on finance dataset (e.g. with 55.8% ACC and 24.5%

Method	Finance					Insurance				
	Clustering Metrics		Theme Label Quality			Clustering Metrics		Theme Label Quality		
	Acc	NMI	Rouge-1/2/L	CosSim	LLM-Score	Acc	NMI	Rouge-1/2/L	CosSim	LLM-Score
GURP	24.6	28.2	5.0 / 3.5 / 5.0	13.8	87.0	41.5	42.2	12.3 / 0.0 / 12.3	47.8	86.6
GTR	39.1	51.5	21.6 / 6.4 / 21.1	42.8	82.9	39.6	51.7	27.1 / 8.4 / 26.2	**57.5**	96.5
SPC	23.3	28.0	19.1 / 4.1 / 19.0	48.5	85.8	23.5	30.1	20.8 / 8.3 / 20.7	44.6	87.2
CATCH	**55.8**	**67.1**	**42.4 / 24.5 / 42.4**	**59.3**	**97.3**	**54.5**	**62.6**	**41.8 / 16.1 / 41.8**	57.0	**100.0**

Table 2: Out-of-domain Performance of the Model on Finance and Insurance Dataset.

Method	Acc	NMI	Rouge-1/2/L	Cos	LLM
GURP	36.8	33.4	11.1 / 2.9 / 11.1	30.8	82.0
GTR	46.9	51.6	22.0 / 3.8 / 20.4	37.3	86.8
SPC	15.4	4.4	6.9 / 0.6 / 6.7	52.8	90.7
CATCH	**56.7**	**65.4**	**35.3 / 10.0 / 35.3**	**58.5**	**95.9**

Table 3: In-domain Performance of the Model at Banking Dataset.

ROUGE-2) than on Insurance dataset (e.g. with 0.545 ACC and 0.161 ROUGE-2), being contrary to all the baselines. Since the input utterance in finance dataset is vague in theme (two utterances are shown in Section 4.7) comparing to other two datasets, the topic attribution representation is shown to be effective in improving the theme detection ability by augmenting the input utterance to a context block.

4.6 Online Official Blind Evaluation Results

Table 4 and Table 5 show the official blind evaluation results, covering both automatic and manual assessments. Our team (Team E) achieved **second place** in the overall ranking across all metrics. Notably, we achieved this result using a relatively lightweight model of only **8 billion parameters**, without leveraging any powerful proprietary models such as GPT-4 or GPT-4o at any stage of the pipeline. This demonstrates the effectiveness and efficiency of our approach under constrained computational budgets.

In the **automatic evaluation** (Table 4), Team E ranked second overall with a score of 67.48%, closely behind Team C (75.50%). Our system shows strong performance in both the clustering metrics and theme label generation metrics. For instance, our model achieved 42.28% in ROUGE-1 and over 93% in all BERTScore variants. Moreover, our results on the style alignment metrics (LLMAAJ) indicate consistent and well-formatted outputs.

In the **human evaluation** (Table 5), our model again achieved the **second-highest** overall average (71.83%). Particularly, we obtained 86.27% in semantic relevance and 91.11% in domain relevance, suggesting that our model excels at generating informative and contextually appropriate topic labels. These results validate that our model delivers high-quality and human-preferred outputs in real-world scenarios, reinforcing its applicability in practical theme detection systems.

4.7 Module Effectiveness via Ablation Study

We conduct ablation experiments on the finance dataset to assess the effectiveness of each core component in CATCH. As shown in Table 6, we evaluate the following variants: **w/o-PeC**: removes preference-enhanced clustering; falls back to baseline clustering. **w/o-TopSeg**: removes topic segmentation; uses only utterance-level representation. **w/o-HieGen**: removes hierarchical label generation; uses flat label generation.

All three modules contribute substantially to overall performance. Discarding **PeC** causes disalignment with user preferences, shown by -7.8% decrease in CosSim. Removing **TopSeg** significantly decreases clustering quality, with -14.2% Acc and -13.8% NMI, demonstrating its importance in capturing topical coherence across utterances. Simplifying **HieGen** in flat generation leads to the greatest loss in label generation quality, particularly in ROUGE-L (-2.8%), confirming the effectiveness of hierarchical modeling. Notably, w/o-HieGen also causes great backward in theme distribution quality with -19% Acc and -18.9% NMI, because HieGen is capable to assign correct label for majority topics with in a cluster, where flat label generation usually encounters malfunction thus provides meaningless label

4.8 Case Study

To demonstrate the effectiveness of topic attribution representation, Figure 4 shows two representative samples from finance dataset , each sample is a pair of input utterance (left) and the corresponding topic block (right) obtained by applying the topic segmentation (Section 3.1). The utt label is generated on the input utterance, while the topic label is generated on the topic block.

Team ID	LLM	Acc	NMI	Rouge-1/2/L	CosSim	BERTScore (P/R/F1)	Sec-1	Sec-2	Avg	Overall
Team C	API	**68.0**	**70.4**	**45.2 / 23.8 / 45.1**	**69.9**	**95.0 / 94.7 / 94.7**	**100.0**	**99.5**	**99.7**	**75.5**
Team E (ours)	<30B	35.8	47.7	42.3 / 16.5 / 41.2	62.5	93.9 / 92.8 / 93.3	93.5	95.7	94.6	67.5
Team D	<30B	51.8	47.7	34.6 / 21.3 / 34.3	55.9	92.5 / 91.5 / 91.9	80.4	76.6	78.5	63.1
Team A	API	48.4	42.0	32.7 / 4.6 / 29.8	59.5	89.8 / 91.2 / 90.4	46.0	56.5	51.2	53.5
Team F	<30B	26.7	9.1	23.1 / 0.8 / 21.1	46.0	85.7 / 89.3 / 87.2	4.1	3.5	3.8	33.4
Team B	API	17.9	2.0	5.0 / 0.0 / 5.0	37.1	85.2 / 88.0 / 86.5	12.0	0.1	6.1	28.8

Table 4: Automatic evaluation results on the blind test set (Travel). All values are percentages. LLM: API indicates usage of proprietary models via API; <30B denotes open models smaller than 30B.

Team ID	Per-Utterance Functional Metrics					Per-Cluster Structural Metrics		Per-Cluster Functional (TD)	Overall Avg.
	SR	AU	GR	ACT	DR	CWC	GS		
Team C	**89.67**	**82.75**	**47.84**	**74.77**	**98.82**	**100.00**	**100.00**	**91.11**	**85.62**
Team E (ours)	86.27	54.64	22.48	54.51	91.11	93.65	93.65	78.34	71.83
Team D	68.76	63.66	26.41	60.26	94.25	91.67	66.67	90.91	70.32
Team A	77.25	63.66	22.75	56.21	79.74	83.33	100.00	75.76	69.84
Team F	45.23	41.57	7.71	41.57	67.45	95.00	100.00	72.63	58.90
Team B	64.97	12.94	0.00	4.05	97.78	100.00	33.33	0.00	39.13

Table 5: Human evaluation results on the blind test set (Travel). All values are percentages. Metrics: Semantic Relevance (SR), Analytical Utility (AU), Granularity (GR), Actionability (ACT), Domain Relevance (DR), Conciseness & Word Choice (CWC), Grammatical Structure (GS), and Thematic Distinctiveness (TD).

Model	Acc	NMI	Rouge-1/2/L	Cos	LLM
CATCH	**55.8**	**67.1**	**42.4** / 24.5 / **42.4**	**59.3**	**97.3**
w/o-PeC	48.8	59.6	40.7 / **26.7** / 40.7	51.5	98.4
w/o-TopSeg	41.6	53.3	23.6 / 10.1 / 23.6	45.3	87.8
w/o-HieGen	36.8	48.2	19.6 / 9.1 / 19.6	30.3	82.3

Table 6: Ablation results on the finance dataset.

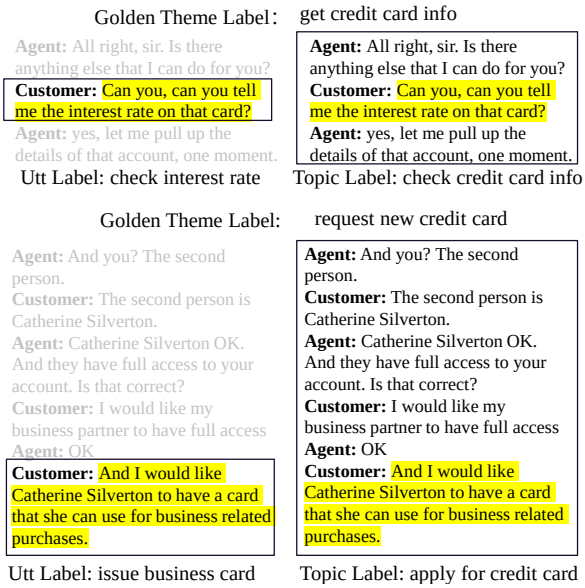


Figure 4: Two samples from finance dataset with utterance id: *"Finance_1e18a3a5_100410_SS01_A6-115"* and *"Finance_35138e33_100917_2464642A2-39"* respectively..

Using context topic block as theme representation provides more thematically precise label. In both examples, the thematic information of utterance is either miss-leading (i.e. "interest rate" for the first example) or vague (i.e. "business related purchases" for the second example) because of the data sparsity problem. The context topic block provides more hints of the theme which mitigates the miss-leading information, and clarifies the vague information by putting view on the complete topic context.

5 Conclusion

In this paper, we propose CATCH, a novel theme detection framework that significantly enhances the automatic discovery and consistency of themes within a latent topic space aligned with user preferences. By treating the entire architecture as a theme generation pipeline, CATCH jointly models intra-dialogue theme representation and inter-dialogue preference-aware alignment via preference-enhanced clustering, leading to coherent and user-aligned theme labels after hierarchical generation. Extensive experiments demonstrate the robustness and generalizability of CATCH across diverse tasks, while ablation studies further reveal the complementary roles and coordination of its three key modules. In future work, we plan to continuously improve the framework with cutting-edge techniques and make it more adaptive and dynamic, enabling its application to a broader range of downstream scenarios, such as proactive dialogue systems, dialogue control, and fine-grained dialogue analysis.

Limitation

Although our approach demonstrates promising for theme detection, there are a few limitations to acknowledge. Firstly, the number of clusters for semantic clustering requires manually predefined, which sets limitation on discovering new latent topics. Secondly, the preference-enhanced topic clustering of CATCH relies on the preference feedback typically being vacant in other dialogue dataset, which limits its direct application to other dataset. Thirdly, in the offline experiment, we only compare CATCH with three other baselines. More works should be included to provide more comprehensive comparing.

Acknowledgments

This research is supported by the project of Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002), Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing, SRIBD Innovation Fund (Grant No. K00120240006), and Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044.

References

Tran Xuan Bach, Nguyen Duc Anh, Ngo Van Linh, and Khoat Than. 2021. Dynamic transformation of prior knowledge into bayesian models for data streams. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3742–3750.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152.

Rita Costa, Bruno Martins, Sérgio Viana, and Luisa Coheur. 2023. Towards a fully unsupervised framework for intent induction in customer support dialogues. *arXiv preprint arXiv:2307.15410*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Lan Du, Wray Buntine, and Mark Johnson. 2013a. Topic segmentation with a structured topic model. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 190–200.

Lan Du, Wray Buntine, and Mark Johnson. 2013b. Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Atlanta, Georgia. Association for Computational Linguistics.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware contrastive learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2481–2485.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023. Intent induction from conversations for task-oriented dialogue track at dstc 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2024. Advancing topic segmentation and outline generation in chinese texts: The paragraph-level topic representation, corpus, and benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 495–506.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Lin, Xinyu Ma, Xin Gao, Ruiqing Li, Yasha Wang, and Xu Chu. 2024. Combating label sparsity in short text topic modeling via nearest neighbor augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13762–13774.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021. Open intent discovery through unsupervised semantic clustering and dependency parsing. *arXiv preprint arXiv:2104.12114*.

John Mendonça, Lining Zhang, Rahul Mallidi, Luis Fernando D'Haro, and João Sedoc. 2025. Dstc12: Dialogue system technology challenge 12.

Ha Nguyen, Hoang Pham, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022. Adaptive infinite dropout for noisy and sparse data streams. *Machine Learning*, 111(8):3025–3060.

Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2024. Glocom: A short text neural topic model via global clustering context. *arXiv preprint arXiv:2412.00525*.

Tung Nguyen, Tue Le, Hoang Tran Vuong, Quang Duc Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Sharpness-aware minimization for topic models with high-quality document representations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4507–4524.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. 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 4016–4025.

Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Ngo, Duc Nguyen, and Thien Nguyen. 2024. Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7758–7772.

Jiashu Pu, Guandan Chen, Yongzhu Chang, and Xiaoxi Mao. 2022. Dialog intent induction via density-based deep clustering ensemble. *arXiv preprint arXiv:2201.06731*.

Martin Riedl and Chris Biemann. 2012. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736.

Anh Phan Tuan, Bach Tran, Thien Huu Nguyen, Linh Ngo Van, and Khoat Than. 2020. Bag of biterms modeling for short texts. *Knowledge and Information Systems*, 62(10):4055–4090.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *Proceedings of the web conference 2020*, pages 2009–2020.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177.

Chenxiao Yang, Nathan Srebro, David McAllester, and Zhiyuan Li. 2025. Pencil: Long thoughts with short memory. *arXiv preprint arXiv:2503.14337*.

Zengfeng Zeng, Dan Ma, Haiqin Yang, Zhen Gou, and Jianping Shen. 2021. Automatic intent-slot induction for dialogue systems. In *Proceedings of the Web Conference 2021*, pages 2578–2589.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893.

Overview of Dialog System Evaluation Track: Dimensionality, Language, Culture and Safety at DSTC 12

John Mendonça[1,2], Lining Zhang[3], Rahul Mallidi [3], Alon Lavie [5,6], Isabel Trancoso [1,2], Luis Fernando D'Haro [4], João Sedoc [3]
[1]INESC-ID, Lisbon
[2]Instituto Superior Técnico - University of Lisbon
[3]Department of Technology, Operations, and Statistics, New York University
[4]Speech Technology and Machine Learning Group - Universidad Politécnica de Madrid
[5]Carnegie Mellon University
[6]Phrase, Pittsburgh
Correspondence: john.mendonca@inesc-id.pt

Abstract

The rapid advancement of Large Language Models (LLMs) has intensified the need for robust dialogue system evaluation, yet comprehensive assessment remains challenging. Traditional metrics often prove insufficient, and safety considerations are frequently narrowly defined or culturally biased. The DSTC12 Track 1, "Dialog System Evaluation: Dimensionality, Language, Culture and Safety," is part of the ongoing effort to address these critical gaps. The track comprised two subtasks: (1) Dialogue-level, Multi-dimensional Automatic Evaluation Metrics, and (2) Multilingual and Multicultural Safety Detection. For Task 1, focused on 10 dialogue dimensions, a Llama-3-8B baseline achieved the highest average Spearman's correlation (0.1681), indicating substantial room for improvement. In Task 2, while participating teams significantly outperformed a Llama-Guard-3-1B baseline on the multilingual safety subset (top ROC-AUC 0.9648), the baseline proved superior on the cultural subset (0.5126 ROC-AUC), highlighting critical needs in culturally-aware safety. This paper describes the datasets and baselines provided to participants, as well as submission evaluation results for each of the two proposed subtasks.

1 Introduction

The rapid advancements in Large Language Models (LLMs) have led to increasingly sophisticated conversational agents capable of engaging in complex and nuanced dialogues. As these models become more integrated into various applications, from customer service to personal assistants, ensuring their quality, reliability, and safety is paramount. However, evaluating dialogue systems comprehensively remains a significant challenge (Rodríguez-Cantelar et al., 2023; Mendonça et al., 2024a). Traditional metrics often fall short of capturing the multifaceted nature of human-like conversation, and safety considerations are frequently narrowly defined or culturally biased, failing to address the full spectrum of potential issues.

Addressing the first aspect of this challenge – the limitations of current evaluation metrics – previous challenges and works focus largely on turn-level dialogue evaluation (Zhang et al., 2021; Rodríguez-Cantelar et al., 2023; Mehri et al., 2022) and often lack further investigation of dialogue-level evaluation through automatic metrics. As LLMs advance, aspects of conversations beyond coherence, fluency, etc. should also be studied. Additionally, these aspects should provide a more fine-grained analysis of the levels of quality for the whole conversation, moving beyond simplistic turn-based scores.

Complementing the need for improved quality assessment, the safety dimension, highlighted as a critical concern from the outset, presents its own distinct set of urgent problems. Users are increasingly challenging current chatbots to generate harmful and/or unsafe answers. In addition, even without adversarial probing, generated responses may contain unhelpful and/or harmful content. Therefore, the automatic detection of this content is important in the deployment of these systems. Unfortunately, existing safety evaluation frameworks frequently narrow the notion of safety to strict definitions of bias and toxicity, discarding other safety aspects (Shuster et al., 2022; Ouyang et al., 2022). Furthermore, a significant limitation in current safety paradigms is their predominant focus on the English language. We attempt to mitigate this bias by expanding safety datasets to a diverse set of languages and cultures. Beyond facilitating the study of safety across cultures, this

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 27–35

Attribute	Description
Empathy	Do you think your conversational partner had genuine empathy?
Trust	Based on the conversation, your conversational partner seems trustworthy
Skill	Based on the conversation, your conversational partner seems skilled
Talent	Based on the conversation, your conversational partner seems talented
Capability	Based on the conversation, your conversational partner seems capable
Relevance	Responses address the given context or query well, ensuring that the information provided is pertinent and directly applicable.
Non-Repetition	How repetitive was this chatbot?
Proactivity	Responses actively and appropriately move the conversation along different topics
Curiosity	How much did the chatbot try to get to know you?
Overall	How was the conversation?

Table 1: Evaluation dimensions and definitions for Task 1.

also allows for the evaluation of the robustness of safety classifiers in terms of culture and language.

1.1 Track Details

To address these gaps, this paper presents Track 1 of DSTC12, entitled "Dialog System Evaluation: Dimensionality, Language, Culture and Safety." The shared task was divided into two tasks: Dialogue-level and Multi-dimensional Automatic Evaluation Metrics (§2), and Multilingual and Multicultural Safety Detection (§3). This year's iteration introduced two key novelties aimed at enhancing participation and streamlining the evaluation process: (1) a focus on model efficiency and (2) the adoption of an online competition platform.

Firstly, recognizing that current dialogue evaluation research (and the broader "LLM-as-a-judge" paradigm) often relies on extremely large, proprietary models such as GPT-4 or Claude accessed via APIs, we imposed a significant constraint on model size. Participants were restricted to utilizing open-source LLMs with fewer than 13 billion parameters. This decision was motivated to encourage innovative, efficient solutions that do not solely depend on prompting state-of-the-art models.

Secondly, we utilized the Codabench platform[1] for managing submissions and leaderboards. This facilitated a more dynamic and interactive participation experience. We also released the datasets via Huggingface Datasets to facilitate easy access[2]. On the one hand, it allowed participants to easily gauge their model's performance on the development set in real-time and compare their results against established baselines. On the other hand, for the test set, participants could receive immediate feedback on their system's performance upon submission. To maintain fairness and prevent over-fitting to the test set, submissions were limited to five attempts, and the test set leaderboard remained hidden until the conclusion of the competition.

2 Task 1: Dialogue-level and Multi-dimensional Automatic Evaluation Metrics

In this task, the goal was for participants to develop automatic evaluation metrics for open-domain dialogue. In particular, the submitted systems were expected to evaluate up to 10 different dimensions including previous common ones (Zhang et al., 2021; Rodríguez-Cantelar et al., 2023, i.e.), together with new ones like (Zhang et al., 2024). An overview of these dimensions are presented in Table 1. Similar to previous challenges and prior literature, we evaluated the systems using Spearman's rank correlation between human annotations and automatic metrics as our criterion.

2.1 Dataset

Our main dataset was separated into three collections: three bots (ChatGPT [2023], GPT-3, and BlenderBot-3) during Q1 2023 (TBD-Q1-2023), four bots (ChatGPT, Gemini, Claude, and Mixtral) during Q1 2024 (FBD-Q1-2024), and six bots (ChatGPT, Gemini, Claude, and through Hugging Chat (Mistral, Llama-3 instruct 70B, and Cohere)[3]

[1] We have opened the competitions as benchmarks for the broader community: Task 1; Task 2

[2] huggingface.co/dstc12

[3] The exact versions are `mistralai/Mistral-Nemo-Instruct-2407`, `meta-llama/Meta-Llama-3-70B-Instruct`, and `CohereForAI/c4ai-command-r-plus`.

during Q4 2024 (SBD-Q4-2024). The users in the conversations were undergraduate students. All conversations were read to verify that no personally identifiable information was present. For both FBD-Q1-2024 and SBD-Q4-2024 datasets, we controlled the topics present in the conversation:

T1 Talk about help for turning your homework in late.

T2 Finding an apartment.

T3 Finding something to do in the evening.

T4 Talk about something that is on your mind or bothering you.

T5 Learn about a topic that you are interested in.

T6 Talk about something silly with the chatbot.

Students were randomly assigned, without replacement, to both a chatbot and a conversation topic. They were instructed to interact for roughly 15 turns. After the conversation, they shared their conversation link and filled out the surveys. Subsequently, the conversation links were web scraped, and the conversational data were merged with the survey responses.

The dataset was split into development (TBD-Q1-2023 / FBD-Q1-2024) of 185 conversations and test set (SBD-Q4-2024) of 120 conversations. TBD-Q1-2023 included 8 participants, FBD-Q1-2024 had 4, and SBD-Q4-2024 had 6. TBD-Q1-2023 was used in the DSTC11 Track 4 challenge (Rodríguez-Cantelar et al., 2023) for both turn- and dialog-level evaluation, but only coarse-grained dimensions were used.

Following Zhang et al. (2024), we used a subset of dimensions for evaluation. Table 1 has the list of dimensions along with their definitions.

2.2 Baseline

As a baseline, we prompt Llama-3-8B-Instruct to provide an evaluation across all of the dimensions. The system prompt is presented in Table 3.

2.3 Participants

Team 1 Team 1 submitted four unique systems. System 1 employed a regression approach, training separate regression layers on top of a ModernBert encoder for each evaluation dimension using the DSTC-12, ConTurE (Ghazarian et al., 2022), and FED (Mehri and Eskenazi, 2020) datasets. System 2 utilized a prompting strategy, combining detailed dimension explanations and dialogue context with a selection of models (Deepseek Llama 8B, Deepseek Qwen 7B, Qwen 2.5 7B Instruct-1M), choosing the best-performing model per dimension based on validation set results. System 3 was a classification-based approach, training individual classifiers on an sBERT encoder for each dimension with normalized scores, also using the DSTC-12, ConTurE, and FED datasets. Finally, System 4, a hybrid model, selectively combined the outputs of System 1 (for dimensions like Talent and Relevance) and System 2 (for dimensions like Empathy and Overall) based on which system achieved the best correlation on the validation set for each specific dimension.

Team 2. This team adopted Qwen2.5-7B-Instruct as the base model and then utilized prompt engineering to enable the LLM to automatically output scores across various dimensions. Moreover, they included degree interpretations for different score levels within the prompt.

2.4 Results

The official results for Task 1 are provided in Table 2. The team score was computed as the mean absolute Spearman correlation across all dimensions. We can also see a per-dimension breakdown. Ideally, all correlations should be positive; however, in the absence of consistent definitions, some participants may have reverse-coded certain dimensions.

Quite surprisingly, the baseline outperformed the other teams, followed by Team 1 and then Team 2. This was largely due to the dimensions of relevance and non-repetition. While the baseline performed best overall, there was no clear winner across all dimensions, indicating substantial room for improvement. Most correlations were statistically significant, though relatively low for several dimensions. Unsurprisingly, all metrics correlated well with the overall conversational quality. The negative correlation for the 'skill' dimension highlights a clear area for further refinement.

We had one strange finding: on a relative basis, all automatic metrics favored Mistral, whereas human participants ranked it significantly lower. This is probably the reason for the large decrease in correlations between development and test sets.

Team	Avg.	Emp.	Trust	Skill	Talent	Capab.	Rel.	!Rep.	Proact.	Curio.	Overall
Baseline	.1681	.0647	-.1117	-.0955	.0962	.0677	.2337	*.3851*	-.0248	.2253	*.3766*
1	.1503	-.0788	.0067	-.2246	.0529	.1276	.0808	.1071	-.1457	*.3687*	*.3099*
2	.1360	-.1183	.2407	-.1094	.0873	.1717	.0738	.2220	.0185	.0064	*.3117*

Table 2: Spearman correlations for each team across all dimensions of Task 1. Correlations with $p < 0.05$ in *italic*.

3 Task 2: Multilingual and Multicultural Safety Detection

In this task, participants were tasked with developing automatic safety classifiers of responses generated by LLMs across different languages and cultures. To establish a clear and manageable benchmark, particularly given the diverse range of languages and the nuanced nature of safety, we simplified the core challenge to a binary classification problem: determining whether a given dialogue response is 'safe' or 'unsafe'. We consider a response to be "unsafe" if it violates the MLCommons AI Safety v0.5 Benchmark POC Taxonomy of Hazards (Vidgen et al., 2024)[4]. The taxonomy outlines seven key hazard categories within its scope for this version: Violent crimes, Non-violent crimes, Sex-related crimes, Child sexual exploitation, Indiscriminate weapons (CBRNE), Suicide & self-harm, and Hate. This approach, while a simplification of real-world complexities where harm can be multifaceted and context-dependent, allows for a more direct comparison of model capabilities in fundamental safety detection across varied linguistic and cultural contexts. It also provides a foundational step before tackling more granular multi-label or severity-level classifications.

3.1 Datasets

To allow participants to train and evaluate their models, we curated several datasets. These datasets were processed to conform to a shared conversational format, consisting of context and response pairs accompanied with additional metadata made available in the original datasets. These datasets were then translated to 7 languages (Arabic, German, Spanish, French, Japanese, Portuguese and Chinese) and made accessible to the participants on HuggingFace[5]. We present an overview of these datasets in Table 4.

[4] https://drive.google.com/file/d/1V8KFfk8awaAXc83nZZzDV2bHgPT8jbJY/view

[5] https://huggingface.co/dstc12

3.1.1 Development

Bot Adversarial Dialogue (Xu et al., 2021). This dataset was curated via a human-and-model-in-the-loop framework where crowdworkers were instructed to converse with various state-of-the-art dialogue models, actively probing the model to output unsafe or offensive responses. Each bot utterance within these interactions was annotated for safety, resulting in a corpus of approximately 5.8k dialogues (79k total utterances), with 40% of utterances being annotated as offensive.

Dialogue Safety (Dinan et al., 2019) was curated via a human-and-model-in-the-loop framework. Crowdworkers were presented with an existing dialogue context and were instructed to submit utterances they deemed offensive, specifically targeting those that an existing safety classifier would miss-classify as safe. This iterative process resulted in a corpus of approximately 6,000 "offensive" utterances, collected across both single-turn and multi-turn dialogue context settings. When combined with verified safe examples, these constitute a dataset totalling approximately 60,000 utterances, of which 10% are labelled offensive. For the purpose of this Task, we use the multi-turn subset.

ProsocialDialog (Kim et al., 2022) is a large-scale, multi-turn English dialogue dataset designed to teach conversational agents to respond prosocially to problematic user inputs. Generated via a human-AI collaborative framework, it contains 58,137 dialogues (331,362 utterances) covering diverse unethical, problematic, biased, and toxic situations. Prosocial responses are grounded in 160,295 commonsense social rules-of-thumb (RoTs), and dialogue turns are annotated with fine-grained safety labels accompanied by 497,043 free-form rationales.

3.1.2 Test

Soda-Eval (Mendonça et al., 2024b) is derived from the SODA dataset, and encompasses over 120,000 turn-level assessments across 10,000 dialogues. Each assessment, generated by GPT-4 and subsequently human-validated, includes identifica-

You are an impartial evaluator conducting a multidimensional assessment of text responses. Your role is to analyze and score all chatbot responses using the following criteria:

- Empathy: Based on the conversation, does the chatbot demonstrate understanding and compassion for the user's situation or emotions?
- Trust: Based on the conversation, does the chatbot seem trustworthy?
- Skill: Does the chatbot show competence in the subject matter, providing accurate and relevant information?
- Talent: Does the chatbot show talent in the subject matter, providing accurate and relevant information?
- Capability: Does the chatbot seem capable in interacting with the user?
- Relevance: Are all chatbot responses relevant given prior context?
- Non-Repetition: Does the chatbot avoid unnecessary repetition?
- Proactivity: Does the chatbot anticipate user needs?
- Curiosity: Does the chatbot demonstrate engagement by exploring the topic further or encouraging deeper discussion?
- Non-Repetition: Does the chatbot avoid unnecessary repetition?
- Overall: Overall assessment of the chatbot throughout the dialogue.

Scoring Guidelines:

- Focus only on the chatbot responses, not the user messages. - Assign a score between 1 and 5 for each relevant category based on the criteria above. - Do not output any other meta commentary or information.
Input: The input consists of a conversation between a user and a chatbot.

Output: [JSON format]

Table 3: Baseline evaluation prompt.

tion of fine-grained issues. We leverage the annotations for the quality dimensions "Anti-Social". However, since these annotations were automatically annotated using an LLM (GPT-4), we conduct a human validation step on the safety labels. From this validation step, we randomly select additional positive examples from Soda-Eval to derive a class-balanced set of size 325 examples.

CoSafe (Yu et al., 2024) is a benchmark designed to evaluate safety against multi-turn dialogue coreference attacks. The dataset was constructed by selecting 100 single-turn attack prompts for each of 14 harmful categories, originally defined by BeaverTails (Ji et al., 2023). These prompts were then automatically expanded into multi-turn dialogues using GPT-4, with the coreferentially-phrased attack query placed in the final turn to assess model vulnerabilities in resolving references within a harmful conversational context. We employed multiple LLMs to simulate diverse safety behaviors across model families and architectures. This diversity ensures that safety classifiers are not overfitted to idiosyncrasies of a single generation style and that evaluation generalizes across real-world deployment scenarios. The chosen models are aya-expanse-8b (Dang et al., 2024), EuroLLM-9B-instruct (Martins et al., 2024), LLama-3.2-Instruct (1B,3B) and LLama-3.1-8B-Instruct (Grattafiori et al., 2024), Ministral-8B-Instruct-2410 (MistralAI, 2024), and Qwen2.5-Instruct (3B,7B) (Qwen et al., 2025). We then conduct a human-validated automated annotation using GPT-4o (OpenAI et al., 2024) as an automated safety classifier. Then, all examples rated as unsafe are evaluated by a human annotator. A balanced safety-label subset is then sampled from these annotations.

Dataset	#Utterances (k)
BAD	69.3 / 7 / 2.6
Dialogue Safety	24 / 3 / 3
Prosocial Dialogue	120 / 20.4 / 25
Total	213.3 / 30.4 / 30.6
SODA-Eval	- / - / 325
CoSafe	- / - / 227
SafeWorld	- / - / 437

Table 4: Overview of datasets used in Task 2. For the development set, we provide train/validation/test sets.

SafeWorld (Yin et al., 2024) For the cultural subtask, we employ a curated version of the cultural-aware safety dataset of SafeWorld (Yin et al., 2024). SafeWorld is designed to assess alignment with geo-diverse cultural and legal safety standards by grounding queries on human-verified cultural norms and legal policies from 50 countries and 493 distinct regions/races. We focus on the "specific answer" and "comprehensive answer" query types. "Specific answer" queries (641 instances) require models to pinpoint a single, pre-defined cultural or legal guideline violated in a given scenario; "comprehensive answer" queries (577 instances) present situations where potential violations are ambiguous, tasking models to provide comprehensive responses covering relevant norms and policies across implicated regions. We prompt GPT-4o to determine if the policy or norm viola-

tion identified would elicit a safety violation. For the examples identified as unsafe, we then generate responses using several LLMs: aya-expanse-8B (Dang et al., 2024), EuroLLM-9B-instruct (Martins et al., 2024), gemma-3-9b-it (Gemma et al., 2025); gpt-4o-mini (OpenAI et al., 2024), LLama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.57B-Instruct (Qwen et al., 2025). Then, we ask GPT-4o (OpenAI et al., 2024) to determine if the the model response elicits the identified policy/norm violation.

Human validations to confirm the accuracy of the test set labels was conducted by a single annotator, a graduate student with a background of Computational Linguistics with extensive annotation experience.

3.1.3 Translations

We translate all of the context-response pairs for the development set as a whole, and for Soda-Eval and CoSafe. We selected a wide-range of languages that cover a diverse set of cultures: Arabic, German, English, Spanish, French, Japanese, Portuguese and Chinese.

Automated validation For translation model selection, we automatically validate the quality of translations in a smaller subset by employing an MT-QE model, in this case `wmt23-cometkiwi-da-xl`. We obtain a score of 0.7116 using GPT-4o and 0.7153 GPT-4o-mini. Given the small performance difference and the much lower price point for GPT-4o-mini, we opt with utilizing GPT-4o-mini for the translation of the full development set.

Human validation For the test set, we additionally conduct a human validation study to ensure the translations are of high quality and do not change the semantics of the response such that it changes the safety label. Human annotators – one per language, except for Japanese (2) – were recruited given their professional level of the target language and consisted of a mix of graduate students and post-docs with Linguistics and Computational Linguistics background with prior experience with annotations. Initial annotations showed that GPT-4o-mini was incurring in several translation errors that affect the safety of the response. As such, we re-translate the English set with GPT-4o, which, after subsequent human validation, confirmed the quality of its translations on the examples GPT-4o-mini failed.

3.2 Baseline

We adopted Llama-Guard-3-1B as the official baseline for Task 2. This model was aligned from the original Llama-3.2-1B LLM to safeguard against the MLCommons standardized hazards taxonomy. In particular, the English data used by Llama Guard was combined with additional multilingual human and synthetically generated Human-AI conversation data.

For the purposes of the task, baseline predictions for response classification are obtained by extracting the probability of the "unsafe" token appearing as the first decoded token to reflect the model's direct scoring of risk. This approach offers a deterministic, resource-efficient proxy for classification, aligning with recent work in zero-shot risk estimation and avoiding confounding artifacts from longer generation chains. Reproducible code can be found alongside the data on the HuggingFace dataset repository[6].

3.3 Participants

For this task, a total of 2 teams (Teams 2 and 3) participated with 6 submissions. Participants were asked to provide a description of their submissions. Team 2 submitted a similar system to the one presented in Task 1 (2, adapting the prompt for the safety task. Unfortunately, Team 3 did not provide an official description of their system. However, their submissions to the track platform suggest their approach consisted in the supervised finetuning of LLMs on the development data (`sft_500k_gemma-ck` and `llama3_sft_500k`) of gemma-2-9b-it and a LLama3 model respecting our model size restrictions (likely 8B).

Team	Average	Cultural	Multilingual
3	**.9046**	.4831	**.9648**
2	.8078	*.4830*	.8517
Baseline	*.7767*	**.5126**	*.8097*

Table 5: ROC-AUC results for Task 2. The first position is shown in **bold**, the second in underline and the third in *italic*.

3.4 Results

The official results for Task 2 are provided in Table 5. Team ranking is established by calculating the

[6] https://huggingface.co/datasets/dstc12/bot_adversarial_dialogue/blob/main/LlamaGuard.py

average ROC-AUC considering all languages and the cultural subset with equal weights. We also present ROC-AUC for the multilingual and cultural subsets separately. We employ ROC-AUC since it provides a threshold-independent assessment of a model's ability to distinguish between safe and unsafe content.

Team 3 ranked best in this Task, followed by Team 2. This is thanks to their strong performance on the multilingual subset, with Team 3 achieving a strong result of .9648, followed by Team 2 with .8517, which are significantly superior to the baseline results (.8097). However, when looking at the cultural subset, we note that the baseline was the best performing submission (.5126), with both Teams achieving similar results (around .4831). These results suggest that models finetuned for cultural agnostic safety concerns fail to account for cultural specificities. This behaviour may be an instance of catastrophic forgetting, since our baseline (LLama-Guard-3-1B) was able to outperform their stronger finetuned models.

4 Conclusions and Future Work

This paper presents the overview of Track 1 on "Dialog System Evaluation: Dimensionality, Language, Culture and Safety" organized as part of the 12th Dialogue System Technology Challenge (DSTC12). The track was organized in two tasks aimed at addressing two important problems of the state-of-the-art in Dialogue Systems: (1) Dialogue-level and Multi-dimensional Automatic Evaluation Metrics; (2) Multilingual and Multicultural Safety Detection.

While the track had 11 registered teams, only 3 participated. The first task drew two of these teams. We used Spearman's rank correlation coefficient absolute average value as the rank ordering for the teams. The baseline outperformed the best overall, but alone many different dimensions we see different methods performing better.

In the second task, two teams participated and comfortably outperformed the baseline on the multilingual subset, achieving very strong ROC-AUC. However, for the cultural subset, no team was able to outperform the baseline ROC-AUC, which sits at just .5126, indicating clear room for improvement.

As future work, Task 1, we plan to extend the analysis of fine-grained dimensions to understand the upper-bound of LLM-evaluation for dimensions of human quality assessment. Importantly, we plan to increase the diversity of participants to be more representative of larger populations. For Task 2, we plan on extending the safety classification task to include the full taxonomy, providing a more fine-grained assessment of risks.

Acknowledgments

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI), by Portuguese national funds through Fundação para a Ciência e Tecnologia (FCT) with references PRT/BD/152198/2021 and DOI:10.54499/UIDB/50021/2020.

This work is supported by the European Commission through Project ASTOUND (101071191 – HORIZON EIC-2021 – PATHFINDERCHALLENGES-01), and by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the European Union.

We also want to give thanks to MS Azure services (especially to Irving Kwong) for their sponsorship to continue processing new datasets that could be interesting for the dialogue community.

This research project is supported by the NYU ChatEval Team led by João Sedoc. He would like to thank NYU Stern for its funding.

References

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey

Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

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 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2022. Interactive evaluation of dialog track at DSTC9. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5731–5738, Marseille, France. European Language Resources Association.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2024a. On the benchmarking of LLMs for open-domain dialogue evaluation. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024b. Soda-eval: Open-domain dialogue evaluation in the age of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11687–11708, Miami, Florida, USA. Association for Computational Linguistics.

MistralAI. 2024. Un ministral, des ministraux | mistral ai.

OpenAI, 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, and 400 others. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander I. Rudnicky. 2023. Overview of robust and multilingual automatic evaluation metricsfor open-domain dialogue systems at DSTC 11 track 4. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 260–273, Prague, Czech Republic. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, and 81 others. 2024. Introducing v0.5 of the ai safety benchmark from mlcommons. *Preprint*, arXiv:2404.12241.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 2950–2968, Online. Association for Computational Linguistics.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Safeworld: Geo-diverse safety alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Gao Zuchen, Fei Mi, and Lanqing Hong. 2024. CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17494–17508, Miami, Florida, USA. Association for Computational Linguistics.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *Preprint*, arXiv:2111.02110.

Lining Zhang, João Sedoc, and Natalia Levina. 2024. Back to principles: Theory-driven evaluation of ai-based conversational agents. In *Forty-Fifth International Conference on Information Systems*.

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.

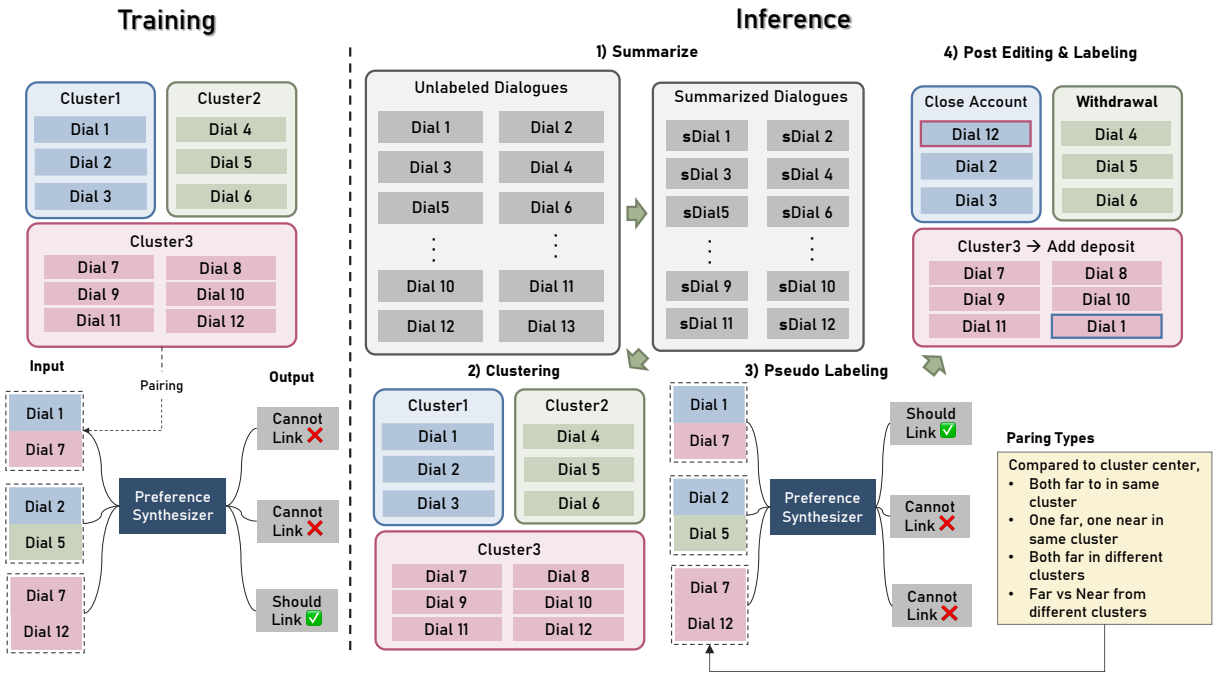


Figure 1: Overview of our pipeline. We train a preference model using cluster-based pairings, then apply it during inference to refine clustering results by predicting should-link/cannot-link pairs and adjusting utterance assignments accordingly.

> 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*.

KSTC: Keyphrase-driven Sentence Embedding and Task Independent Prompting for Filling Slot in the Generation of Theme Label

Sua Kim* **Taeyoung Jeong*** **Seokyoung Hong*** **Seongjun Kim***
Jeongpil Lee **Du-Seong Chang** **Myoung-Wan Koo**
Department of Artificial Intelligence, Sogang University, Korea
{lightwsrld, john428, hsy960925, ksj12035, jplee, dschang, mwkoo}@sogang.ac.kr

Abstract

Intent discovery in task-oriented dialogue is typically cast as single-turn intent classification, leaving systems brittle when user goals fall outside predefined inventories. We reformulate the task as multi-turn zero-shot intent discovery and present KSTC, a framework that (i) embeds dialogue contexts, (ii) performs coarse clustering, (iii) generates a predicted theme label for each cluster, (iv) refines clusters using the Large Language Model (LLM) with the predicted theme label, and (v) relocates utterances according to user's preferences. Because generating informative predicted theme label is crucial during the LLM-driven cluster refinement process, we propose the Task Independent Slots (TIS), which generates effective theme label by extracting verb and noun slot–value.

Evaluated on DSTC12 Track2 dataset, KSTC took first place, improving clustering and labeling quality without in-domain supervision. Results show that leveraging conversational context and slot-guided LLM labeling yields domain-agnostic theme clusters that remain consistent under distributional shift. KSTC thus offers a scalable, label-free solution for real-world dialogue systems that must continuously surface novel user intents. The code will be available at https://github.com/sogang-isds/KSTC.

1 Introduction

In task-oriented dialogue systems deployed in real-world services, it is essential to extract user intent from conversations (Ni et al., 2022). As customer needs diversify and business environments continue to evolve, the field of intent discovery has emerged, which aims to identify user intents from utterance collections that are either unlabeled or only partially labeled (Liu et al., 2021; Zhang et al., 2021; Liang and Liao, 2023). However, most prior work on intent discovery focuses on single-turn utterances, emphasizing the development of clustering algorithms designed to learn user utterance representations aligned with clustering objectives (Yin et al., 2021; Park et al., 2024).

Recent research has increasingly focused on intent classification in multi-turn dialogues, where users' intentions gradually become evident throughout a conversation. Such research highlights the need for robust intent discovery methods that can adapt to the dynamic characteristics of dialogues and diverse application environments (Liu et al., 2024a,b).

DSTC12 Track 2[1], formulates theme detection as a joint clustering and theme labeling for the input utterances. According to the task definition, intents are mapped to a fixed set of predefined labels, whereas themes are user-facing outputs, such as those presented to call center analysts, and thus require more flexible and expressive representations that can be tailored to user preferences. In theme detection, individual users may demand fine-grained analysis of specific themes or, conversely, prefer high-level overviews, depending on their business goals. Therefore, enabling personalized theme labeling based on user preferences is a crucial requirement in this task. Furthermore, the DSTC12 Track 2 task involves theme detection in a zero-shot, domain-agnostic environment, where themes emerge progressively through multi-turn dialogue.

To address this, we propose **KSTC** (Keyphrase-driven Sentence embedding and Task independent prompting for filling slot in the Generation of theme label), as shown in Figure 1, a novel framework that incorporates user preferences, refines clusters effectively, and generates theme labels that are both semantically coherent and practically useful.

In Stage 1, we generate keyphrases from each

[1]https://github.com/amazon-science/dstc12-controllable-conversational-theme-detection.git

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 44–73

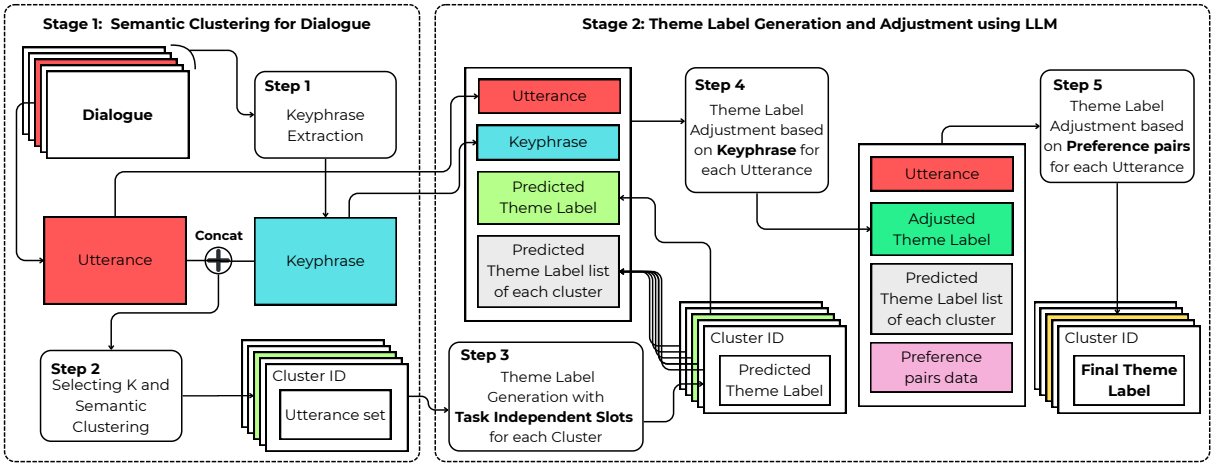


Figure 1: Overall Framework of KSTC. In Stage 1, we extracted keyphrases from each utterance and its surrounding context, and selected the most appropriate keyphrase using an LLM (Step 1). Each utterance was concatenated with its selected keyphrase and clustered accordingly (Step 2). In Stage 2, we generated theme labels for each cluster using Task Independent Slots (Step 3). We then refined the predicted theme labels by comparing them against the list of predicted labels within each cluster (Step 4). Finally, we generated the final theme label by adjusting the refined label based on preference pairs data (Step 5).

utterance and its context to extract conversational context in multi-turn dialogues. The keyphrases are embedded together with the corresponding utterance to perform semantic clustering. This process enables knowledge extension which cannot be fully represented by utterance-level embeddings, resulting in more semantically coherent and practical intent clusters.

In Stage 2, we utilize the Task Independent Slot methodology guided by LLM to extract key verbs and nouns associated with the intents of each utterance cluster. This enables the generation of effective predicted theme labels across diverse domains. We then use the predicted theme labels to refine the clusters through LLM based correction.

Our main contributions are as follows:

- **Semantic Clustering with Utterance and Keyphrase:** KSTC enhances semantic clustering by incorporating not only theme label annotated utterances but also up to three surrounding conversational turns. This allows for richer, more context aware clustering. To achieve this, we propose a semantic embedding method that leverages LLMs to extract keyphrases from the surrounding context of an utterance, which are then concatenated with the utterance prior to embedding.

- **Predicted Theme Label Generation for Initial Clusters:** To refine clusters effectively, we utilize the language understanding capabilities of LLMs to generate predicted theme labels for the initial clusters formed from semantic embeddings. These labels are created by filling Task Independent Slots using prompting technique, in which both fine-grained and broad semantic aspects of the cluster are captured.

- **Theme Label Adjustment Using Full Conversational Information:** We further improve the accuracy of the label by adjusting the theme label using not only the utterance's keyphrase context, but also the predicted theme labels of other groups and the current group's own label. This holistic use of conversational information enables more nuanced and accurate label refinement.

- **Incorporating Pre-defined Preferences:** Along with semantic information, KSTC also accounts for preference pairs data. These preferences allow for alternative clustering outcomes depending on user preference, even when the semantic content of the conversations is identical. This flexibility enables theme label adjustment to reflect both semantic structure and user's specific categorization needs.

2 Method

We use the NATCS (Gung et al., 2023) dataset introduced in DSTC12 Track 2, which consists of multiple dialogues, each composed of a sequence of utterances. A summary of the dataset statistics

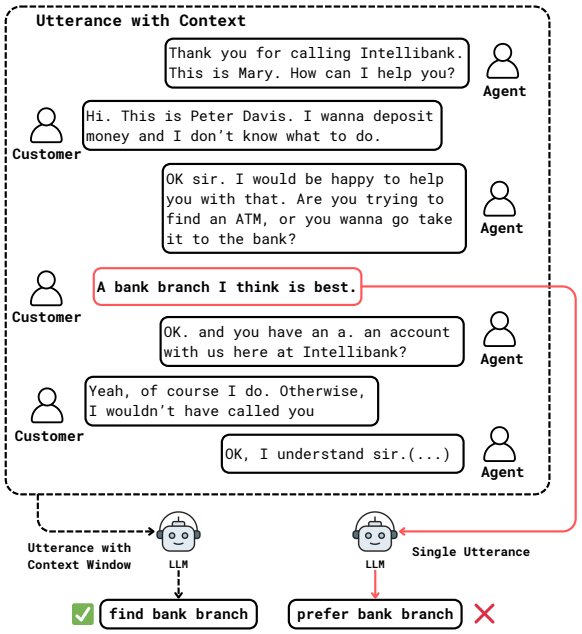


Figure 3: Comparison of keyphrases generated from single utterance and multi-turn context inputs

is presented in Appendix A. Among these, only a subset of utterances is annotated with theme labels. In this work, we focus exclusively on inferring theme labels for the annotated utterances.

Our overall framework consists of two stages, as illustrated in Figure 1. Here, Dialogue refers not to the entire conversation but to a localized context consisting of an annotated utterance and a few surrounding utterances within the same dialogue. Stage 1 performs semantic clustering by extracting keyphrases from each annotated utterance and its local context, followed by clustering based on these enriched semantic representations. Stage 2 generates theme labels for each cluster using an LLM, guided by the Task Independent Slot we designed, and then refines these labels at the utterance-level through additional LLM-based adjustments. We describe each step in detail in the following sections.

2.1 Step 1 : Keyphrase Extraction

Step 1 in Figure 1, illustrates the keyphrase extraction process. In natural language utterances, intent is often not explicitly stated but is instead contextually implicitly expressed. To address this, we generate keyphrases that aim to explicitly expose such contextually implicit intent, converting them into more interpretable and intuitive representations.

To improve the quality of the extracted keyphrases, we incorporate the surrounding dialogue context. Specifically, for each annotated utterance, we use its dialogue, which includes up to three preceding and three following utterances within the same dialogue. If the utterance appears near the beginning or end of a dialogue, fewer utterances may be included. To ensure thematic consistency, we exclude any surrounding utterances that are annotated with a different theme label.

We use an LLM to generate up to three candidate keyphrases for each dialogue. Subsequently, by verifying the candidate keyphrases, a single keyphrase that accurately reflects the core action of the utterance is selected. The prompt used for keyphrase generation is provided in Appendix D.

These context-aware keyphrases serve as intermediate semantic representations and are then used in both the clustering and labeling stages of our framework.

Figure 3 compares the keyphrases generated from single-utterance input versus those generated with multi-turn dialogue context (see Appendix B for additional examples). When the surrounding dialogue context (i.e., three preceding and following utterances) is provided (left), the LLM correctly generates "find bank branch", which accurately reflects the user's intent. In contrast, without the surrounding context (right), it generates "prefer bank branch", which fails to capture the intended meaning.

These results highlight the effectiveness of our context-aware approach. By leveraging additional conversational context, we are able to more accurately disambiguate intent and generate keyphrases that are both precise and semantically aligned.

2.2 Step 2 : Select K and Semantic Clustering

Step 2 in Figure 1 illustrates the semantic clustering process, which utilizes both the original utterances and the keyphrases extracted in the previous step.

To construct semantically rich representations for clustering, we first train an encoder following the ClusterLLM approach (Zhang et al., 2023), using the annotated utterances concatenated with the keyphrases as input. As suggested in their method, we repeat the training process for two iterations, which has been shown to improve the quality of semantic embeddings and enhance clustering performance.

During inference, we concatenate each annotated utterance with its corresponding keyphrase and encode the combined text using the trained encoder. This allows the resulting embedding to reflect both

the original utterance and the enriched semantic intent captured by the keyphrase.

Once all utterance embeddings are obtained, we apply a clustering algorithm to group similar utterances. To determine the optimal number of clusters K, we adopt a hybrid strategy that combines both intrinsic and extrinsic evaluation signals. Specifically, we use the Silhouette score (Rousseeuw, 1987) and preference pairs data, which reflect user perspectives. The preference pairs consist of pairwise annotations indicating whether two utterances should belong to the same theme (Should-Link) or different themes (Cannot-Link). These preference pairs constitute part of the ground truth data and are provided to enable user-customized control over clustering granularity. An example of this dataset is provided in Appendix A.

Dataset	# of Clusters (Ground Truth)	# of Clusters (Predicted)
Banking	26	30
Finance	34	38
Insurance	27	38

Table 1: Number of clusters across the three datasets.

As input to this clustering step, we use the final keyphrase selected for each utterance. Each keyphrase is embedded using the 'text-embedding-3-large'(OpenAI, 2024), and K-Means clustering is performed for values of K ranging from 2 to 40. For each value of K, we compute a Combined Score (CS), defined as:

$$CS = w_{sil} \cdot S + w_{sl} \cdot \text{Acc}_{sl} + w_{cl} \cdot \text{Acc}_{cl} \quad (1)$$

Here, S denotes the silhouette score, reflecting both the number of clusters and the degree of intra-cluster cohesion. Acc_{sl} denotes the proportion of Should-Link pairs that were assigned to the same cluster, while Acc_{cl} represents the proportion of Cannot-Link pairs that were assigned to different clusters. In our experiments, the weights were set to $w_{sil} = 0.5$, $w_{sl} = 0.25$ and $w_{cl} = 0.25$. We select the value of K that achieves the highest Combined Score as the optimal number of clusters. Table 1 presents the selected number of clusters for each dataset. By leveraging semantically enriched keyphrases and user-driven constraints, our method enhances both the internal coherence and external validity of the resulting clusters.

Finally, we apply clustering algorithms (e.g., K-means, Agglomerative) to the utterance embeddings, using the optimal number of clusters selected based on the Combined Score. This clustering benefits from both surface-level features and the additional semantic cues introduced by the keyphrases.

2.3 Step 3 : Theme Label Generation with Task Independent Slots for Each Cluster

Step 3 in Figure 1 illustrates the methodology for generating theme labels for each cluster. To support this, we employ Task Independent Slots (TIS) that facilitate the extraction of task-related keywords from utterances in the same cluster. We guide the LLM with prompts to generate these slots, aiming to decompose tasks independently at a general domain level. Specifically, the LLM was guided by prompts to produce high-level action and conceptual categories commonly observed in real-world service conversations. The prompts used for generating these slots are provided in Appendix E. The generated slots consist of two complementary components that target distinct linguistic elements essential for intent identification: **Task Independent Verb Slots** and **Task Independent Noun Slots**.

The **Task Independent Verb Slots** define key action categories frequently observed in customer-agent dialogues, such as *require*, *request_info*, *cancel*, *confirm*, *update*, *inquire_issue*, and *recommend*. These verbs represent common types of user requests and interactions.

In contrast, the **Task Independent Noun Slots** encompass relevant entities and concepts pertinent to the tasks, including *product*, *service*, *account*, *schedule*, *personal_info*, *payment*, *status*, *issue*, *location*, *document*, and *indicator*.

For each cluster, we independently apply the Verb and Noun Slots to the aggregated utterances. We first analyze the semantic content of the utterances and extract verbs and nouns that correspond to the predefined slot categories. This procedure enables the identification of frequently occurring, slot-consistent verbs and nouns, facilitating an accurate characterization of the core actions and entities associated with the cluster's shared intent. To automate this process, we design a zero-shot prompt that enables an LLM to perform slots application and theme labeling.

Finally, we input the clustered utterances, the corresponding Verb and Noun Slots, and their extracted entities into the LLM to generate the final theme label for the cluster.

2.4 Step 4 : Theme Label Adjustment based on Keyphrase for each Utterance

Step 4 in Figure 1 illustrates the process of refining the predicted theme label for each utterance using additional semantic cues. For each utterance, we integrate the following information to facilitate adjustment: the utterance itself, its associated keyphrase, the initially predicted theme label inherited from its cluster, and the full set of theme labels predicted across all clusters.

The appropriateness of the assigned theme label with respect to the utterance's content and task context was evaluated using an LLM. If the label was deemed semantically appropriate, it was retained; otherwise, a more suitable label was selected from the list of predicted theme labels, considering the semantic alignment between the utterance, its keyphrase, and the available theme labels.

This verification and adjustment process aims to enable fine-grained, utterance-level theme labeling by leveraging the keyphrase as additional contextual information. A zero-shot prompt is employed to guide the LLM in assessing the correctness of labels and revising misassigned ones.

2.5 Step 5 : Theme Label Adjustment based on Preference pairs for each Utterance

Once keyphrase-based adjustments have been applied to all utterances, an additional adjustment step is conducted for utterances specified in the preference pairs data, as shown in step 5 in Figure 1. In the case of Should-Link, we consider not only direct pairs but also transitive relations among them. For example, if utterance u_j is in a Should-Link relation with u_k, and u_k is also in a Should-Link relation with u_l, then all three utterances u_j, u_k, u_l are expected to belong to the same cluster. Based on these transitive relationships, all connected utterances are assigned to the same group. For each group, candidate theme labels are collected from the keyphrase-adjusted clusters to which its member utterances belong. Then, we choose semantically consistent and representative theme labels using the LLM from among the candidate set. In the case of Cannot-Link, if a given pair of utterances is assigned to the same cluster, one of them must be reassigned. For example, for a Cannot-Link pair u_l and u_k, we consider the theme labels of their current clusters (after keyphrase-based adjustment), as well as those of other clusters to which neither utterance is currently assigned. Therefore, we identify utterances whose semantics are less aligned with the current theme label and select a more appropriate label by LLM from the candidate set. This process enables fine-grained cluster adjustment that faithfully reflects users' actual preferences.

This two-step adjustment process enables more accurate grouping by explicitly exposing contextual intent, particularly in cases where the original utterance lacks sufficient standalone information. The prompts used for steps 4 and 5 are detailed in Appendix F.

3 Experiments

3.1 Datasets

We evaluated our proposed method using the three development datasets and one test dataset provided by the organizers of DSTC12 Track 2. All four datasets are designed based on NATCS and were collected from four distinct domains: Banking, Finance, Insurance, and Travel. These datasets consist of multi-domain customer support dialogues between customers and agents.

3.2 Implementation detail

To compare performance with the number of clusters we selected, we used the ground truth number of clusters. This follows the convention used in prior studies (Zhang et al., 2023; Viswanathan et al., 2023; Liang et al., 2024).
For fine-tuning the embedding model, we used the AdamW optimizer with a batch size of 16. We used GPT-4o to generate and filter keyphrases and to generate theme labels. For clustering adjustment, we employed GPT-4.1 due to its overwhelming long context performance[2]. The full prompts are available in Appendix D.

3.3 Evaluation Metric

We focus on both the quality of clustering and the accuracy of label generation.

To evaluate the clustering quality, we compare the accuracy (ACC) and normalized mutual information (NMI) scores of our method with baselines.

For each cluster, the reference labels of its utterances will be compared to the label predicted for the cluster. We evaluate both semantic similarity and the inclusion of key terms using cosine similarity, ROUGE scores, and BARTScore.

Cosine similarity is a metric for measuring semantic similarity between the Sentence-BERT

[2]https://openai.com/index/gpt-4-1/

Method	Dataset					
	Banking		Finance		Insurance	
	NMI	ACC	NMI	ACC	NMI	ACC
Instructor (w/ KMeans) (Su et al., 2023)	65.32	54.79	65.18	51.48	56.57	43.42
Instructor (w/ Agglom.) (Su et al., 2023)	61.75	52.51	63.59	51.59	56.88	44.19
Instructor + Keyph. Clust. (w/ KMeans) (Su et al., 2023)	**74.29**	63.15	**75.01**	60.42	62.76	47.49
Instructor + Keyph. Clust. (w/ Agglom.) (Su et al., 2023)	73.88	**66.89**	74.15	**61.16**	**62.85**	**47.79**
CLUSTERLLM-I-iter (w/ KMeans)	75.75	60.82	76.66	60.64	63.12	50.50
CLUSTERLLM-I-iter (w/ Agglom.)	77.25	62.06	74.86	61.39	64.87	**52.29**
CLUSTERLLM-I-iter + Keyph. Clust. (w/ KMeans)	75.54	62.77	77.74	60.85	65.80	50.44
CLUSTERLLM-I-iter + Keyph. Clust. (w/ Agglom.)	**77.32**	**67.38**	**79.23**	**62.49**	**65.84**	52.06
KSTC	**81.68**	**78.34**	**81.91**	**63.94**	**70.24**	**57.31**

Table 2: Comparison of NMI and ACC across clustering methods and datasets. The KSTC results are based on CLUSTERLLM-I-iter with keyphrases using Agglomerative clustering. Best results are bolded.

(Reimers and Gurevych, 2019) embeddings of the reference and predicted labels. ROUGE scores (Lin, 2004) are N-gram overlap metrics that are effective for comparing short and concise sequences between the reference and predicted labels. Specifically, we compute ROUGE-1, ROUGE-2 and ROUGE-L scores. BARTScore (Yuan et al., 2021) is a metric designed to measure semantic similarity between a generated text and a reference text and is known to have a high correlation with human judgment. We use the pretrained bart-large-cnn[3] model, where higher score indicates greater semantic consistency between the two texts.

4 Results

4.1 Analysis of Stage 1 results

Table 2 presents a comparative analysis of the performance of various clustering algorithms under different conditions, measured by NMI and ACC.

Comparison of Initial Clustering Algorithms K-means exhibits high performance variability depending on the initialization of cluster centroids, whereas Agglomerative Clustering adopts a deterministic merging approach. Using the encoder trained with CLUSTERLLM-I-iter, Agglomerative Clustering demonstrated superior performance in all three datasets. In this setting, we used the Instructor-large as the pre-trained embedder. This can be interpreted as the trained embedder enhancing the merging criteria of Agglomerative Clustering, thereby better capturing the similarities among data points.

[3]https://huggingface.co/facebook/bart-large-cnn

Performance Analysis of KSTC's Clustering We define the final KSTC method by applying keyphrase and preference adjustments after embedding the clustering method with the highest performance among existing approaches, CLUSTERLLM-I-iter+Keyph. Clust. (w/ Agglom.). KSTC achieves the highest performance in both NMI and ACC metrics. This improvement is attributed to the effective correction of ambiguous cluster boundaries when based solely on utterances and keyphrases, through the predicted theme labels generated by Task Independent Slots. In other words, our method integrates not only semantic information from the text but also information derived from external knowledge, enabling a more precise understanding of the intrinsic data structure and the formation of accurate clusters.

Effectiveness of Keyphrase Utilization Combining keyphrases extracted from conversational context with the previously introduced Agglomerative clustering, Table 2 demonstrates that the CLUSTERLLM-I-iter+Keyph. Clust. (w/ Agglom.) approach consistently achieves superior performance across various datasets. Specifically, compared to CLUSTERLLM-I-iter Clust. (w/ Agglom.), the keyphrase-enhanced model achieves an average improvement of 2.7% in ACC and 6.1% in NMI across all datasets. Furthermore, we employed t-SNE (Van der Maaten and Hinton, 2008) for visualization, as illustrated in Appendix G, our keyphrase-enhanced clustering method separates clusters more distinctly. Consequently, we propose CLUSTERLLM-I-iter+Keyph. Clust. (w/ Agglom.) as the initial clustering for KSTC. This indicates that keyphrases, which capture

DataSet	#Clusters	Clustering Algorithm	Theme Label Generation	Evaluation Metric										
				Clustering		Cosine	BART	Rouge-1		Rouge-2		Rouge-L		Avg.
				NMI	ACC	Similarity	Score	Recall	Precision	Recall	Precision	Recall	Precision	Cosine Rouge
Banking		Baseline		0.5984	0.5149	0.5579	-6.5217	0.4208	0.3959	0.1525	0.1629	0.4181	0.3946	0.3575
	K=26	Initial clustering of KSTC	Baseline	0.7732	0.6738	0.6324	-5.2352	0.5391	0.5446	0.2563	0.2229	0.5272	0.5268	0.4642
		Initial clustering of KSTC	TIS	0.7778	0.6916	0.7098	-4.7616	0.6081	0.6556	0.2948	0.2888	0.5912	0.6334	0.5403
		KSTC (TIS)		**0.8213**	0.7301	0.7289	-4.6786	0.6309	**0.6802**	0.3044	0.2971	0.6133	**0.6570**	0.5588
	K=30	Initial clustering of KSTC	Baseline	0.7692	0.645	0.6316	-5.4410	0.5090	0.5216	0.1950	0.1840	0.4967	0.5036	0.4345
		Initial clustering of KSTC	TIS	0.7906	0.7252	0.7280	-4.6786	0.6196	0.6570	0.3350	0.3447	0.6028	0.6348	0.5603
		KSTC (TIS)		0.8192	**0.7570**	**0.7452**	**-4.6070**	**0.6393**	0.6780	**0.3448**	**0.3555**	**0.6218**	0.6549	**0.5771**
Finance		Baseline		0.6218	0.4979	0.5398	-6.3143	0.4717	0.4286	0.2387	0.2084	0.4417	0.3895	0.3883
	K=34	Initial clustering of KSTC	Baseline	0.7923	0.6249	0.5986	-5.4484	0.4884	0.5202	0.2773	0.2877	0.4829	0.5129	0.4526
		Initial clustering of KSTC	TIS	0.7923	0.6249	0.6918	-4.2393	0.6716	0.6601	0.4623	0.4316	0.6716	0.6601	0.6070
		KSTC (TIS)		0.8222	0.6481	0.6997	-4.1861	0.6820	0.6699	0.4701	0.4387	0.6820	0.6699	0.6160
	K=38	Initial clustering of KSTC	Baseline	0.7914	0.6377	0.6091	-5.4215	0.4967	0.5182	0.2771	0.2841	0.4907	0.5099	0.4551
		Initial clustering of KSTC	TIS	0.7954	0.6441	0.6951	-4.1185	0.7043	0.6681	**0.4812**	0.4487	0.7043	0.6681	0.6243
		KSTC (TIS)		**0.8302**	**0.6771**	**0.7022**	**-4.0637**	**0.7109**	**0.6987**	**0.4812**	**0.4531**	**0.7109**	**0.6987**	**0.6365**
Insurance		Baseline		0.5173	0.3930	0.4221	-7.0239	0.2673	0.2294	0.1062	0.0703	0.2607	0.2223	0.2255
	K=27	Initial clustering of KSTC	Baseline	0.6564	0.5206	0.4433	-6.5038	0.3343	0.3235	0.1153	0.0946	0.3264	0.3156	0.2790
		Initial clustering of KSTC	TIS	0.6595	0.5206	0.4807	-5.4413	0.4248	0.3798	0.1341	0.0965	0.4258	0.3748	0.3309
		KSTC (TIS)		0.7123	**0.5882**	0.5042	-5.4325	0.4248	0.3863	0.1365	0.1002	0.4257	0.3818	0.3371
	K=38	Initial clustering of KSTC	Baseline	0.6733	0.5379	0.4592	-6.2450	0.3684	0.3251	0.1306	0.0985	0.3574	0.3142	0.2933
		Initial clustering of KSTC	TIS	0.6722	0.5349	0.5128	-5.3022	0.4574	0.4072	0.1658	0.1285	0.4456	0.4004	0.3597
		KSTC (TIS)		**0.7254**	0.5746	**0.5331**	**-5.2376**	**0.4780**	**0.4225**	**0.1717**	**0.1352**	**0.4595**	**0.4116**	**0.3731**

Table 3: Labeling performance comparison on the NATCS datasets. The best clustering result (ClusterLLM-I-iter with keyphrases and Agglomerative clustering) is used as the initial clustering for the KSTC.

the core information of a dialogue, serve as salient features that enhance cluster cohesion and contribute to improved clustering performance. This also suggests that keyphrases can further enhance clustering performance, even within an already optimized embedding space.

Method	Preference	Banking	Finance	Insurance
Initial clustering of KSTC	Should-Link	32.93%	41.62%	12.90%
	Cannot-Link	32.93%	41.62%	12.90%
keyphrase-based adjustment	Should-Link	52.44%	43.93%	19.35%
	Cannot-Link	84.15%	84.39%	87.3%
KSTC	Should-Link	99.39%	98.84%	96.13%
	Cannot-Link	98.78%	100%	99.21%

Table 4: Preference-satisfaction ratio

4.2 Analysis of Stage 2 results

Table 3 presents a comparative analysis of the KSTC's final clustering and label per on the Banking, Finance and Insurance datasets, following LLM-based adjustment.

- **Baseline**: It extracts utterance embeddings from Sentence-Transformers and performs K-Means clustering in the resulting embedding space. The number of clusters is set to the ground-truth value. Subsequently, the Mistral-7B-Instruct model is used to generate a single theme label for each cluster based on all utterances it contains.

Analysis of LLM Adjustment Performance and Label Generation Methods of KSTC

Following theme labeling on the initial clustering results, KSTC, which incorporates LLM-based adjustment and cluster refinement, achieves consistent performance improvements across all three datasets, significantly outperforming the baseline. These improvements are observed consistently across both clustering and labeling evaluation metrics.

As shown in Table 4, the proposed method enables fine-grained adjustments of complex and nuanced user intent representations through keyphrase-based contextual adjustment and preference-based adjustment. Table 4 reports the preference-satisfaction ratio, computed as the number of satisfied preference pairs divided by the total number of preference pairs (higher is better).

We attribute the improvement in clustering and labeling quality through LLM adjustment to more accurate predictions in theme label generation. As shown in Table 3, when using the value of K determined based on the Combined Score for initial clustering, followed by theme label generation using Task Independent Slots, performance improves over using the ground-truth K in terms of average cosine similarity and ROUGE scores, with improvements of 1.8%p in Banking, 1.7%p in Finance, and 3.6%p in Insurance datasets, respectively. To analyze the source of the performance gains, we examined, for each value of K, the degree to which utterances within a single cluster shared the same theme label (Cluster Purity). Using the proposed method, the proportion of perfectly pure clusters (100% Purity)

Dataset	Utterance	Predicted Theme Label	Theme Label (Ground Truth)
Finance	Also, what are your your hours at at at at the branch over there on on Baker Street?	get branch location/hours	get branch location/hours
	Yes, I'm trying to find out what I owe for my credit card.	check credit card balance	check credit card balance
	I need to find out what my net income is from January to June of this year.	get net income	get net income
	Thank you, I just, I'm looking for some. A line of credit, perhaps.	apply for line of credit	apply for line of credit
	Yes, so I was wondering if you could tell me the current CPI, please?	request consumer price index	get consumer price index

Table 5: Comparison between predicted theme labels and ground truth theme labels in Finance Dataset.

Method	Clustering		Theme Label accuracy						Theme Label Style			
							BertScore			LLM-as-a-Judge			
	ACC	NMI	Rouge-1	Rouge-2	Rouge-L	Cosine Similarity	Precision	Recall	F1	Section1	Section2	Average	Avg. Overall
(Team C) Ours	**0.6797**	**0.7039**	**0.4522**	**0.2381**	**0.4510**	**0.6991**	**0.9502**	**0.9469**	**0.9471**	**1.0000**	**0.9948**	**0.9974**	**0.7550**
(Team D)	0.5176	0.4771	0.3457	0.2131	0.3427	0.5593	0.9252	0.9148	0.9191	0.8039	0.7660	0.7850	0.6308
(Team E)	0.3582	0.4773	0.4228	0.1650	0.4122	0.6248	0.9385	0.9284	0.9327	0.9346	0.9569	0.9458	0.6748

Table 6: Official results for test submissions by DSTC12 Track2, Automatic evaluation

Method	Per-Utterance Functional				Per-Cluster Structural			Per-Cluster-Functional	
	Semantic Relevance	Analytical Utility	Granularity	Actionability	Domain Relevance	Conciseness Word Choice	Grammatical Structure	Thematic Distinctiveness	Avg. Overall
(Team C) Ours	**0.8967**	**0.8275**	**0.4784**	**0.7477**	**0.9882**	**1.0000**	**1.000**	**0.9111**	**0.8562**
(Team D)	0.6876	0.6366	0.2641	0.6026	0.9425	0.9167	0.6667	0.9091	0.7032
(Team E)	0.8627	0.5464	0.2248	0.5451	0.9111	0.9365	0.9365	0.7834	0.7183

Table 7: Official results for test submissions by DSTC12 Track2, Human evaluation

relative to the total number of clusters increased by 2.31%p in Banking, 0.8%p in Finance, and 8.38%p in Insurance. In terms of utterance counts, the number of utterances contained in perfectly pure clusters grew by 57 in Banking, 15 in Finance, and 58 in Insurance. The analysis is provided in Appendix H. These findings suggest that our approach improves overall clustering quality. Moreover, high-purity clusters with their strong topical coherence create favorable conditions for the subsequent LLM-based automatic labeling stage, leading to more accurate and reliable theme generation.

Our labeling method, Task Independent Slots, prioritizes the selection of core verbs and objects within the cluster and employs the LLM to generate more appropriate theme label expressions, thereby capturing finer details. This demonstrates that the high quality of initial labeling contributes to the overall improvement in final clustering and labeling performance. Table 5 substantiates these gains: each predicted label (i) removes superfluous words, (ii) appears as an event-centered verb phrase, (iii) strikes the right balance between being actionable and sufficiently general, and (iv) is informative enough to narrow downstream resolution steps—while almost matching the gold label for sampled utterance in the finance dataset. Additional examples for Banking and Insurance are provided in Appendix I.

Test Data Results

Tables 6 and 7 are the official results of the test submission by the participants. This includes both human evaluation and LLM-based evaluation. Our method, denoted as Team C, is the model ranking first.

5 Conclusion

We propose KSTC, a clustering and theme labeling framework that operates in unseen intent scenarios and exhibits robust domain adaptability. Our method enhances clustering performance by leveraging keyphrases extracted from conversational context, enabling the generation of semantically fine-grained theme labels using the Task Independent Slots. This approach facilitates high quality label creation even in practical datasets that require complex and nuanced intent understanding. Moreover, KSTC offers flexibility that reflects pre-defined user preferences. Experimental results demonstrate that LLM-based cluster refinement consistently improves both clustering and labeling performance across all three datasets. In addition, the effectiveness of our method was demonstrated by ranking first in both automatic and human evaluations in DSTC12 Track 2.

The domain independent performance of KSTC in this zero-shot setting is expected to significantly contribute to intent analysis in real-world industrial applications.

6 Limitations

KSTC generates informative predicted theme labels for each cluster using Task Independent Slots, and effectively performed clustering refinement based on this information, achieving significant performance improvements across multi-turn intent discovery datasets. However, our method is currently applicable only to datasets where each utterance is annotated with explicit intent labels. Future research should focus on developing an algorithm that can first determine whether an intent exists within a dialogue.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621, RS-2022-II220621, Development of artificial intelli- gence technology that provides dialog-based multi- modal explainability)

References

Grant Anderson, Emma Hart, Dimitra Gkatzia, and Ian Beaver. 2024. An open intent discovery evaluation framework. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 760–769, Kyoto, Japan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1373–1378. Association for Computational Linguistics (ACL).

Jinggui Liang and Lizi Liao. 2023. ClusterPrompt: Cluster semantic enhanced prompt learning for new intent discovery. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10468–10481, Singapore. Association for Computational Linguistics.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2024a. Intent-aware dialogue generation and multi-task contrastive learning for multi-turn intent classification. *Preprint*, arXiv:2411.14252.

Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2024b. Lara: Linguistic-adaptive retrieval-augmentation for multi-turn intent classification. *Preprint*, arXiv:2403.16504.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021. Open intent discovery through unsupervised semantic clustering and dependency parsing. *Preprint*, arXiv:2104.12114.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Preprint*, arXiv:2105.04387.

OpenAI. 2024. Chatgpt. `https://platform.openai.com/docs/guides/embeddings`. Accessed: 2025-06-02.

Jeiyoon Park, Yoonna Jang, Chanhee Lee, and Heuiseok Lim. 2024. Analysis of utterance embeddings and clustering methods related to intent induction for task-oriented dialogue. *Preprint*, arXiv:2212.02021.

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

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. *Preprint*, arXiv:2212.09741.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *Preprint*, arXiv:2307.00524.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Hui Yin, Xiangyu Song, Shuiqiao Yang, Guangyan Huang, and Jianxin Li. 2021. Representation learning for short text clustering. *Preprint*, arXiv:2109.09894.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *Preprint*, arXiv:2305.14871.

A Dataset Statistics

Dataset	# of Dialogues	Avg. Words per Turn	# of Intent-labeled Utterances	# of Intents	# of Domains
Banking	980	59.6 ± 23.1	1634	26	1
Finance	3000	65.6 ± 22.4	1723	34	1
Insurance	954	70.6 ± 19.2	1333	27	1

Table 8: Dataset statistics

A summary of NATCS dataset statistics is shown in Table 8. "Avg. Words per Turn" indicates the average number of words per dialogue turn (mean ± std.).

Dataset	should_link	cannot_link
Banking	164	164
Finance	173	173
Insurance	155	126

Table 9: Preference data statistics

The number of preference pairs dataset for each domain can be found in Table 9.
For example, in the case of Should-Link, if the user's preferences suggest that the utterances "I gotta get my grandma some money." and "We just transfer the first because I need to close the account..." should belong to the same theme, other similar utterances would be associated with a single theme that semantically unifies the meanings of those utterances "make external wire transfer" or a close paraphrase of it. On the other hand, in the case of Cannot-Link, if the preferences indicate that "I want to change my email" and "I want to update my personal information" should not belong to the same theme, then the corresponding themes, "update email" and "update personal info", along with their associated utterance clusters, should remain separate.

B Extract Keyphrase & Illustrative Examples

We conducted experiments to extract keyphrases from multi-turn dialogue contexts by setting the context window size to 1, 3, and 5, respectively. In each experiment, the context window determines how many utterances before and after the theme-labeded utterance are taken into account. Examples are shown below.

Context Window = 1

Input Data:
`"Theme_label: first, could you give me my balance perhaps? Maybe I can figure it out that way."`
Ground-truth theme label:
`"label_1": "check account balance",`
`"label_2": "check account balance"`
Output Data:
`keyphrase: "check account balance, view account balance, get balance information"`
`final_keyphrase: "get balance information"`

Context Window = 3

Input Data:
`"That's not a problem.",`
`"Take your time.",`
`"OK, Sundown. OK, that works. OK. Now, what you said there was a transaction you were concerned about?",`
`"Theme_label: first, could you give me my balance perhaps? Maybe I can figure it out that way.",`
`"OK, it looks like you've got two thousand six hundred forty-three dollars and twenty-eight cents.",`
`"OK. Oh, man. I'm not sure where that is actually what the problem is. could you give me the last date of my transaction and the dollar amount?"`
Ground-truth theme label:
`"label_1": "check account balance",`
`"label_2": "check account balance"`
Output Data:
`keyphrase: "check balance, recent transaction details, transaction date and amount"`
`final_keyphrase: "check balance"`

DataSet	#Clusters	Clustering Algorithm	Theme Label Generation	Evaluation Metric										
				Clustering		Cosine	BART	Rouge-1		Rouge-2		Rouge-L		Avg.
				NMI	ACC	Similarity	Score	Recall	Precision	Recall	Precision	Recall	Precision	Cosine Rouge
Banking	K=30	Initial clustering of KSTC	Action-Object Pairs	0.7688	0.6753	0.5219	-5.4862	0.3827	0.5529	0.115	0.115	0.3827	0.5529	0.3747
			CoT	0.7821	0.7001	0.6819	-5.6382	0.5907	0.6030	0.2267	0.2371	0.5722	0.5760	0.4982
			CoT + Few-shot	0.7791	0.6916	0.6799	-5.6073	0.5968	0.5777	0.2349	0.2363	0.5807	0.5563	0.4946
			TIS	**0.7906**	**0.7252**	**0.7280**	**-4.6786**	**0.6196**	**0.6570**	**0.3350**	**0.3447**	**0.6028**	**0.6348**	**0.5603**
Finance	K=38	Initial clustering of KSTC	Action-Object Pairs	0.7815	0.5994	0.4675	-5.6427	0.33	0.5533	0.0065	0.0197	0.33	0.5533	0.3229
			CoT	**0.7954**	**0.6441**	0.6631	-5.1056	0.6543	0.6150	0.4151	0.3672	0.6543	0.6150	0.5692
			CoT + Few-shot	**0.7954**	**0.6441**	0.6626	-5.0778	0.6520	0.6028	0.4276	0.3756	0.6520	0.6028	0.5679
			TIS	**0.7954**	**0.6441**	**0.6951**	**-4.1662**	**0.7043**	**0.6681**	**0.4812**	**0.4487**	**0.7043**	**0.6681**	**0.6243**
Insurance	K=38	Initial clustering of KSTC	Action-Object Pairs	0.6777	0.5611	0.3803	-6.2104	0.27	0.4017	0.1282	0.174	0.27	0.4017	0.2894
			CoT	0.6733	0.5379	0.5079	-6.0002	0.4276	0.3713	**0.1918**	**0.1499**	0.4256	0.3638	0.3483
			CoT + Few-shot	**0.6739**	**0.5386**	0.5066	-6.0164	0.4401	0.3595	0.1584	0.1139	0.4381	0.3520	0.3384
			TIS	0.6722	0.5349	**0.5128**	**-5.2990**	**0.4574**	**0.4072**	0.1658	0.1285	**0.4456**	**0.4004**	**0.3597**

Table 10: Labeling performance comparison on the NATCS datasets. The best clustering result (ClusterLLM-I-iter with keyphrases and Agglomerative clustering) is used as the initial clustering for the KSTC.

Context Window = 5

Input Data:

"OK, one more security question. what street did you grow up on?",
"Oh, dear now you're making me think. You know, if I remember it correctly, it was on. Oh, hell. See, I told you this has me all worked up. I don't know what ugh Gosh. It's five thirteen Sundown Avenue.",
"That's not a problem."
"Take your time."
"OK, Sundown. OK, that works. OK. Now, what you said there was a transaction you were concerned about?"
"Theme_label: first, could you give me my balance perhaps? Maybe I can figure it out that way."
"OK, it looks like you've got two thousand six hundred forty-three dollars and twenty-eight cents."
"OK. Oh, man. I'm not sure where that is actually what the problem is. could you give me the last date of my transaction and the dollar amount?"
"It looks it would've been forty-seven dollars eighty-three cents on September twenty-sixth."
"Hmm that doesn't ring any bells. OK."

Ground-truth theme label:

"label_1": "check account balance",
"label_2": "check account balance"

Output Data:

keyphrase: "inquire about recent transactions"
final_keyphrase: "inquire about recent transactions"

When the context window is set to 1, the model focuses solely on the target utterance. As a result, it successfully captures the general theme (e.g., balance inquiry) but fails to identify more detailed aspects of the user's request. With a context window of 3, the surrounding utterances are considered, allowing the model to extract keyphrases that better reflect the user's actual intent. These results are more aligned with the ground-truth theme labels. And when the context window is increased to 5, the broader context often includes utterance segments where the theme shifts. This can lead to keyphrases that diverge from the user's intended goal.

These results suggest that appropriate context window settings are crucial for extracting contextually aligned keyphrases in multi-turn dialogue. In particular, using only a single utterance may lead to information sparsity, while overly large context windows may harm topic consistency.

C Label Generation Prompt Ablation Study

Table 10 summarizes the performance of different label generation strategies using LLMs, specifically examining the effects of Action-Object Pairs, Chain-of-Thought (CoT), CoT + Few-shot, and the Task Independent Slots. As shown in the table, the Task Indepenent Slots consistently outperformed the other approaches across all datasets in terms of average Cosine Similarity and ROUGE scores for theme label generation. The BARTScore was also highest when using Task Independent Slots.

Importantly, the reported performance reflects the results *prior to applying any additional theme label reassignment using LLMs*. In other words, the evaluation is based solely on the labels initially generated by each prompting strategy, without any post-hoc refinement or correction.

Action-Object pairs. (Anderson et al., 2024; Liu

et al., 2021) extract ACTION-OBJECT pairs from utterances within each cluster using the direct object rule of the spaCy dependency parser (Honnibal and Johnson, 2015). They use the most frequent ACTION-OBJECT pair within each cluster as the cluster label.

Chain-of-Thought (CoT). Following the method proposed by Wei et al. (2023), this approach structures prompts such that the LLM performs step-by-step reasoning over the set of utterances to infer labels. This incremental reasoning process allows the model to generate appropriate labels even in zero-shot settings.

Few-shot. The few-shot setting, inspired by Brown et al. (2020), augments the CoT prompt with several example labels to guide the LLM in labeling clusters. While this approach tends to enhance labeling consistency, it is highly sensitive to the choice and composition of the examples, potentially introducing domain bias based on the examples provided.

D Prompt for Generating Keyphrases

Prompt for Generating Keyphrases in Banking

#Objective#
I am trying to cluster online banking-related queries based on whether they express the same intent.
For each dialogue, generate keyphrases ##that describe the utterance marked with a Theme_label's main intent or request##, with a maximum of 3 keyphrases.
Keyphrases must:
- Be highly relevant to online banking domain.
- Focus on a **single main intent** per phrase.
- Be closely related to each other within the utterance's context.
The output must be in the form of <Key phrase example>, not full sentences.

<Key phrase example>
- update phone/email/address
- request email
- find atm
- report notice
- update personal info
</Key phrase example>

#utterance#
{utterances}

Table 11: Prompt for generating keyphrases in the Banking Dataset.

Prompt for Generating Keyphrases in Finance

#Objective#
I am trying to cluster finance-related queries based on whether they express the same intent.
For each dialogue, generate keyphrases ##that describe the utterance marked with a Theme_label's main intent or request##, with a maximum of 3 keyphrases.
Keyphrases must:
- Be highly relevant to finance domain.
- Focus on a **single main intent** per phrase.
- Be closely related to each other within the utterance's context.
The output must be in the form of <Key phrase example>, not full sentences.

<Key phrase example>
- update phone/email/address
- request email
- get account info
- currency exchange rates
- update personal info
</Key phrase example>

#utterance#
{utterances}

Table 12: Prompt for generating keyphrases in the Finance Dataset.

Prompt for Generating Keyphrases in Insurance

#Objective#
I am trying to cluster insurance-related queries based on whether they express the same intent.
For each dialogue, generate keyphrases ##that describe the utterance marked with a Theme_label's main intent or request##, with a maximum of 3 keyphrases.
Keyphrases must:
- Be highly relevant to insurance domain.
- Focus on a **single main intent** per phrase.
- Be closely related to each other within the utterance's context.
The output must be in the form of <Key phrase example>, not full sentences.

<Key phrase example>
- update address
- create account
- change password/security question
- get pet insurance
- update personal info
</Key phrase example>

#utterance#
{utterances}

Table 13: Prompt for generating keyphrases in the Insurance Dataset.

Prompt for Filtering Keyphrases

#Objective#
Output one keyphrase that best describes ##the main request or intent from the utterances marked with a Theme_label##.
Must focus on the main action indicated by the Theme_label, not additional preferences or conditions. (ex: cuisine type, seating preferences, location)
Must select one keyphrase from the Keyphrases list.

#utterance#
{utterances}

#Keyphrases#
{keyphrases}
{format_instructions}

Table 14: Prompt for filtering keyphrases.

E Prompt for Generate Theme Label

E.1 Prompt for Generating Task Independent Slots

Prompt for Generating Task Independent Verb Slots

<task>
You're helping design a standardized **verb-based intent schema** for Dialogue State Tracking (DST) and intent classification across multiple domains.

Each slot name should represent a high-level **action or intention** that users commonly express during task-oriented conversations.

Please follow these guidelines:

1. Focus on **general categories of user actions or intentions**, not specific tasks. For example, use broad actions like "request" or "confirm", not specific activities like "book a flight" or "reset password".
2. Each slot name should be domain-agnostic and reusable across different sectors.
3. Cover a wide range of commonly expressed **user goals, requests, or dialogue functions** in real-world service conversations.

Now generate 10–15 such **generalized verb slot names** along with a **brief description** for each that explains its meaning and use case.

Format:

- slot_name_1: short description
- slot_name_2: short description
...
</task>

Table 15: Prompt for generating Task Independent Verb Slots

Prompt for Generating Task Independent Noun Slots

<task> You're helping design a standardized **entity-based slot schema** for Dialogue State Tracking (DST) and intent classification across multiple domains.

Each slot name should represent a high-level **conceptual category** of entities that users commonly refer to during task-oriented conversations.

Please follow these guidelines:

1. Focus on **abstract concepts or categories**, not specific instances. For example, use general terms like "document" or "location", not "passport" or "branch office".
2. Each slot name should be domain-agnostic and reusable across different sectors.
3. Cover a wide range of commonly referenced **objects, targets, or informational elements** in real-world dialogue tasks.

Now generate 10–15 such **generalized entity slot names** along with a **brief description** for each that explains its meaning and use case.

Format:

- slot_name_1: short description
- slot_name_2: short description
...
</task>

Table 16: Prompt for generating Task Independent Noun Slots

E.2 Prompt for Task Independent Verb Slots

Prompt for Task Independent Verb Slots

<Context>
You are assisting in building a Dialogue State Tracking (DST) system for the domain domain.
You are given utterances that express one intent enclosed in <Utterances> tags.
You are given a schema of generalized intent slots derived from verb groupings. The schema is enclosed in <Schema> tags.
</Context>

<Schema>
- require: The user is asking for a certain request or application.
- request_info: The user is asking for information or clarification about a product, service, or process.
- cancel: The user wants to cancel a service, request, or reservation.
- confirm: The user is verifying the correctness or status of a particular detail or action.
- update: The user wants to modify or refresh existing information or settings.
- inquire_issue: The user is reporting or inquiring about a problem, error, or complaint.
- recommend: The user is seeking advice or a suggestion for the best option.
</Schema>

<Objective>
Analyze the user utterances below and guess user's intent.
Then read <Schema> and determine which generalized intent slots from the <Schema> are relevant.
For each relevant slot, extract up to **three concise action verbs or verb phrases** that best represent the user's intent.
When you extract the verb, **you must follow both <Style> and <Caution> below**

Only extract **verbs or verb phrases** that meet all the following criteria:
- The verb must describe the **user's final goal**, NOT the object or topic.
- Use only the **base form** of the verb (e.g., "check", not "checking" or "checked").
- Avoid vague or speculative verbs unless they clearly reflect intent.

If a slot is **not relevant to the utterances or not useful for DST**, assign it a value of None.
However, **at least one slot must contain a valid verb or verb phrase** — do not return all None.
Always return **all five slots as keys in the JSON**, even if their value is None.
I will give you bunch of tip if you do great, let's think step by step.
</Objective>

<Style>
Use precise and concise verb phrases that clearly express intent.
If the user's action is directly stated, extract that exact verb or phrase.
If the intent is implicit or paraphrased, infer the most representative verb based on meaning.
</Style>

<Audience>
This output will be used by developers and researchers working on an LLM-based DST system.
They will use your output to evaluate whether the model correctly understands and generalizes user intent.
</Audience>

<Caution>
1. The verb have to make sense when the subject is 'user'.
Example:
utterance : Can you tell me about information?
correct verb : (user wants to) get (information)
incorrect verb : (You) tell (me about information)

2. The verb phrase must describe a class of EVENTS. **Do not** use states, entities properties, claims.
Example:
learn [event] vs. know [state]
redeem [event] vs. redemption[entity]
complain [event] vs. angry [property]
report defect [event] vs. product is defective [claim]
</Caution>

<Response Format>
Provide your answer strictly in the following JSON format:
{{
"request_info": [...],
"cancel": [...],
"require": [...],
"confirm": [...],
"update": [...],
"inquire_issue": [...],
"recommend": [...],
}}
</Response Format>

<Utterances>
utterances
</Utterances>

Now return the verb slot-value pairs as described above.

Table 17: Prompt for Task Independent Verb Slots

E.3 Prompt for Task Independent Noun Slots

Prompt for Task Independent Noun Slots

<Context>
You are assisting in building a Dialogue State Tracking (DST) system for the domain domain.
You are given utterances that express one intent enclosed in <Utterances> tags.
You are given a schema of generalized entity slots derived from semantic groupings. The schema is enclosed in <Schema> tags.
</Context>

<Schema>
- product: The product discussed or requested by the user.
- service: The service requested by the user.
- account: An account, subscription, or contract relevant to the user's service.
- schedule: Any time-based request or item such as a date, time, or appointment.
- personal_info: Personal identification details like name, contact number, or address.
- payment: Payment-related information such as method, status.
- status: The progress or result of a request, task, or application.
- issue: A technical or service-related problem the user is experiencing.
- location: A physical place relevant to the conversation (e.g., branch, region).
- document: An official document or form related to the user's intent.
- indicator: The indicator showing or measuring the condition or level of something.
</Schema>

<Objective>
Analyze the user utterances below and guess user's intent.
Then read <Schema> and determine which generalized intent slots from the <Schema> are relevant.
For each relevant slot, extract up to **three concise nouns or noun phrases** that **BEST REPRESENTS the user's INTENT**.
When you extract the nouns, **you must follow both <Style> and <Caution> below**

Only extract **nouns or noun phrases** that meet all the following criteria:
- The noun must describe the **user's final goal**.
- Extract **only noun phrases or named entities** — do not include verbs, adjectives, or statements.
- Avoid vague or overly generic terms like "thing".
- If you want to use verbal noun, do not use it, **use the noun which means same instead**.
- **Do not** include article, pronoun and possessive.
- Use expressions found in the utterances which represents intent.

If a slot is **not relevant to the utterances or not useful for DST**, assign it a value of None.
However, **at least one slot must contain a valid noun or noun phrase** — do not return all None.
Always return **all five slots as keys in the JSON**, even if their value is None.
I will give you bunch of tip if you do great, let's think step by step.
</Objective>

<Style>
Use clean, specific noun phrases.
Use lowercase unless the phrase is a proper noun.
Use real phrases from the utterances whenever possible.
</Style>

<Caution>
Do not extract exact noun for personal_info and location.
Example:
utterance : My name is Andy.
correct noun : name
incorrect noun : Andy
</Caution>

<Response Format>
Provide your answer strictly in the following **JSON format**:
{{
"product": [...],
"service": [...],
"account": [...],
"schedule": [...],
"personal_info": [...],
"payment": [...],
"status": [...],
"issue": [...],
"location": [...],
"document": [...],
"indicator": [...]
}}
</Response Format>

<Utterances>
utterances
</Utterances>

Now return the extracted entity slot-value pairs as described above.

Table 18: Prompt for Task Independent Noun Slots

E.4 Prompt for Generating Theme Label for each Cluster

E.4.1 Prompt for Generating Theme Label for each Cluster by Chain of Thought

Prompt for Generating Theme Label by Chain of Thought in NATCS

<task>
You are an expert call center assistant. You will be given a set of utterances in <utterances> </utterances> tags, each one on a new line.

The utterances are part of call center conversations between the customer and the support agent in the **{domain}** domain.

Your task is to generate a short label describing the theme of all the given utterances.
The label should capture the **customer's intended action** in the call and be written in a clear, standardized format.
The label should be a **verb phrase** starting with a base-form verb.

—

<guidance>
Output your response in the following way:
<theme_label_explanation>Your short step-by-step explanation behind the theme</theme_label_explanation>
<theme_label>Your final theme label</theme_label>
</guidance>
</task>

<utterances>
{utterances}
</utterances>

Table 19: Prompt for generating theme label by Chain of Thought in NATCS

E.4.2 Prompt for Generating Theme Label by Chain of Thought and Few Shot

Prompt for Generating Theme Label by Chain of Thought and Few Shot in NATCS

<task>
You are an expert call center assistant. You will be given a set of utterances in <utterances> </utterances> tags, each one on a new line.

The utterances are part of call center conversations between the customer and the support agent in the **{domain}** domain.

Your task is to generate a short label describing the theme of all the given utterances.
The label should capture the **customer's intended action** in the call and be written in a clear, standardized format.
The label should be a **verb phrase** starting with a base-form verb.

—

To help you understand the expected format, here are **example labels from a different domain (Travel)**:

- book flight ticket
- cancel hotel reservation
- change travel date
- request seat upgrade
- check baggage policy
- report lost luggage
- confirm airport pickup
- reschedule connecting flight
- apply travel insurance
- inquire visa requirement

—

<guidance>
Output your response in the following way:
<theme_label_explanation>Your short step-by-step explanation behind the theme</theme_label_explanation>
<theme_label>Your final theme label</theme_label>
</guidance>
</task>

<utterances>
{utterances}
</utterances>

Table 20: Prompt for generating theme labels by Chain of Thought and Few Shot in NATCS

E.4.3 Prompt for Generating Theme Labels by Task Independent Slots

Prompt for Generating Theme Labels by Task Independent Slots NATCS

<task>
You are an expert call center assistant. You will be given a set of utterances in <utterances> </utterances> tags, each one on a new line.
You will be given a set of verb candidates in <Verb Candidates> </Verb Candidates> tags, each one on a new line.
You will be given a set of entity candidates in <Entity Candidates> </Entity Candidates> tags, each one on a new line.

The utterances are part of call center conversations between the customer and the support agent in the **{domain}** domain.
Your task is to generate a short label describing the theme of all the given utterances.
The label should capture the **customer's intended action** in the call and be written in a clear, standardized format.
Use a set of verb and entity candidates if necessary.
The label should be a **verb phrase** starting with a base-form verb.

—

To help you understand the expected format, here are **example labels from a different domain (Travel)**:

- book flight ticket
- cancel hotel reservation
- change travel date
- request seat upgrade
- check baggage policy
- report lost luggage
- confirm airport pickup
- reschedule connecting flight
- apply travel insurance
- inquire visa

Strict Rules:
- The final theme label MUST NOT include any slot names such as "request_info", "inquire_issue", etc.
- You MUST select actual verbs and noun phrases that naturally appear in user language, not schema keys.
- Only use candidate expressions (e.g., "check") from the given sets — not their slot names.
- The theme label should be understandable to a human without knowing the underlying schema.

<guidance>
Output your response in the following way:
<theme_label_explanation>Your short step-by-step explanation behind the theme</theme_label$_e xplanation >$
<theme_label>Your final theme label</theme_label>
</guidance>
</task>

<utterances>
{utterances}
</utterances>

<Verb Candidates By Slot>
verb_dict
</Verb Candidates By Slot>

<Entity Candidates By Slot>
{entity_dict}
</Entity Candidates By Slot>

Table 21: Prompt for generating theme labels by Task Independent Slots in NATCS

F Prompt for Adujustment

F.1 Prompt for Theme Label Adjustment based on Keyphrase for each Utterance

Prompt for Theme Label Adjustment based on Keyphrase for each Utterance

You are tasked with determining the most appropriate label for a given Utterance.
When choose the label, focus on the require or intent of the utterance.
Keyphrase reflects the context of the dialogue and generated to capture the requir or intent, Use keyphrase as a reference to decide the label.
Instructions:

- Domain: "{domain}"
- Utterance: "{utterance}"
- Keyphrase: "{keyphrase}"
- Current label: "{predicted_label}"
- Candidates:
{candidates}

Choose the most appropriate label from the candidates.
Even if there are labels similar to the current lable, the Current label already captures the intent well, you must keep it.
Respond with the label only.
If none of the candidates are appropriate, respond with 'None'.
{format_instructions}

Table 22: Prompt for theme label adjustment based on Keyphrase for each utterance

F.2 Prompt for Theme Label Adjustment based on Preference pairs for each Utterance

Prompt for Theme Label Adjustment based on Should-Link for each Utterance

<task>
You will be given a set of utterances in <utterances> </utterances> tags, each one on a new line.
You will be given a set of Label candidates in <Label Candidates> </Label Candidates> tags, each one on a new line.

The utterances are part of call center conversations between the customer and the support agent in the **{domain}** domain.

Your task is to choose a label describing the theme of all the given utterances.
Use a set of set of utterances and Label Candidates.
Do not modify Label Candidates. Just choose a Label.

<guidance>
Output your response in the following way:
<theme_label_explanation>Your short step-by-step explanation behind the theme</theme_label_explanation>
<theme_label>Your final theme label</theme_label>
</guidance>
</task>

<utterances>
{utterances}
</utterances>

<Label Candidates>
{Current_Cluster_Labels}
</Label Candidates>

Table 23: Prompt for theme label adjustment based on Should-Link for each utterance

Prompt for Theme Label Adjustment based on Cannot-Link for each Utterance

<task>
You will be given a set of A utterances in <A_utterances> </A_utterances> tags, each one on a new line.
You will be given a set of B utterances in <B_utterances> </B_utterances> tags, each one on a new line.
You will be given a set of Cluster Label in <Cluster_Labels> </Cluster_Labels> tags, each one on a new line.
You will be given a set of Changed Cluster Label Candidates in <Label_Candidates> </Label_Candidates> tags, each one on a new line.

The utterances are part of call center conversations between the customer and the support agent in the **{domain}** domain.

Your task is to choose the utterance between A utterance and B utterance that is less aligned with the Cluster Label.
If you choose A utterance, you return just "A" else is "B".
Additionally, based on the selected utterances, you choose a group of candidate cluster labels that best match the current selected utterance among the Changed Cluster Label Candidates.

Use a set of set of utterances and Label Candidates.
Do not modify Changed Cluster Label Candidates. Just choose a Label.

<guidance>
Output your response in the following way:
<selected_utterance>Selected utterance</selected_utterance>
<theme_label_explanation>Your short step-by-step explanation behind the theme</theme_label_explanation>
<theme_label>Your final theme label</theme_label>
</guidance>
</task>

<A_utterances>
{A_utterances}
</A_utterances>

<B_utterances>
{B_utterances}
</B_utterances>

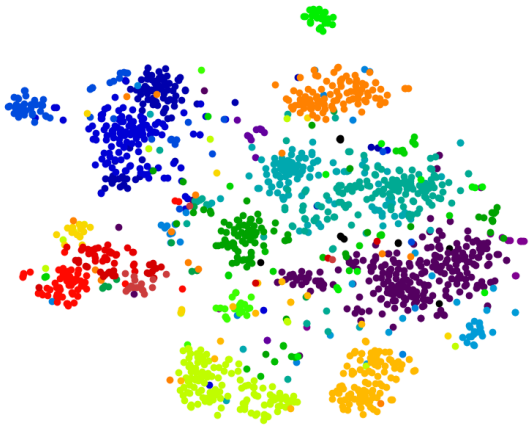
<Cluster_Labels>
{Cluster_Labels}
</Cluster_Labels>

<Label_Candidates>
{Label_Candidates}
</Label_Candidates>

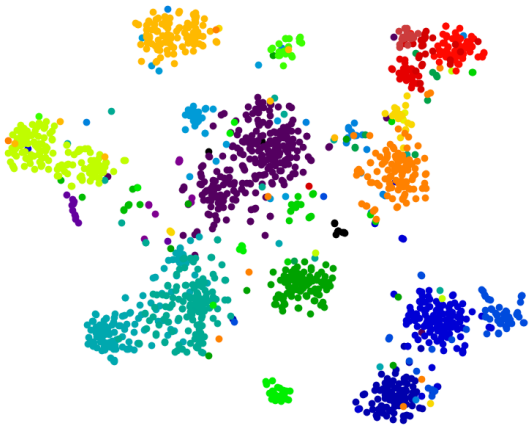
Table 24: Prompt for theme label adjustment based on Cannot-Link for each utterance

G Impact of Keyphrases on Embedding Structure (t-SNE)

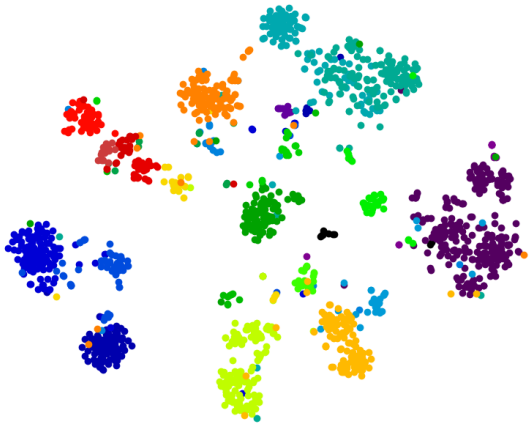
G.1 Banking Dataset



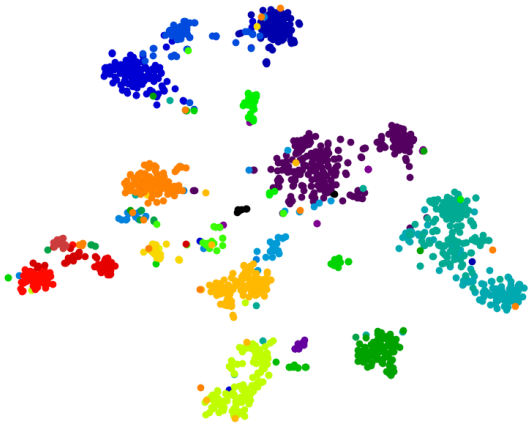
(a) Instructor (pre-training, no keyphrases)



(b) Instructor (pre-training, with keyphrases)



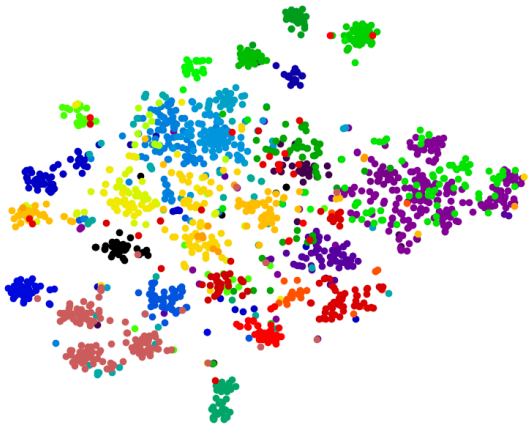
(c) CLUSTERLLM-I-iter (post-training, no keyphrases)



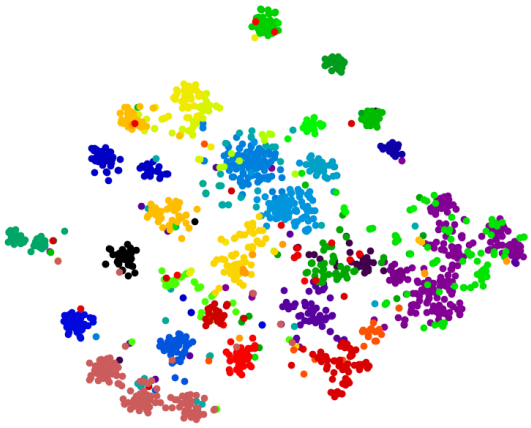
(d) CLUSTERLLM-I-iter (post-training, with keyphrases)

Figure 4: t-SNE Visualization of Embeddings on Banking Dataset

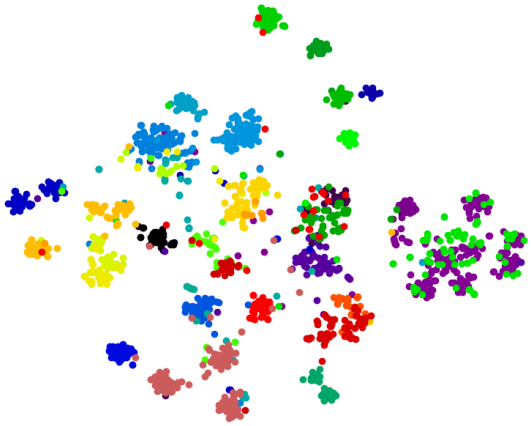
G.2 Finance Dataset



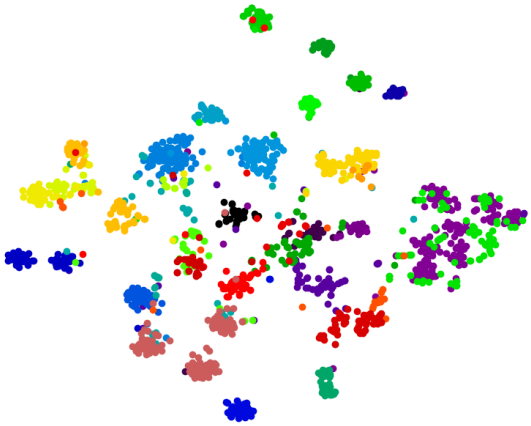
(a) Instructor (pre-training, no keyphrases)



(b) Instructor (pre-training, with keyphrases)



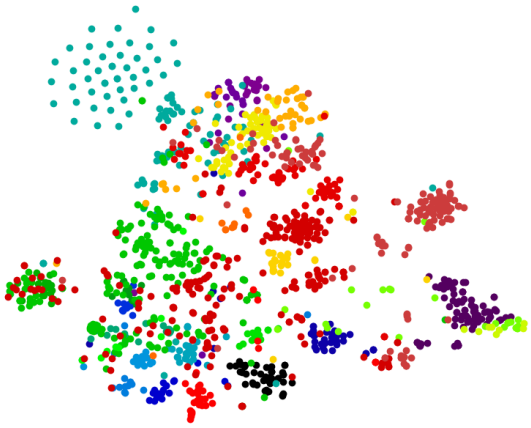
(c) CLUSTERLLM-I-iter (post-training, no keyphrases)



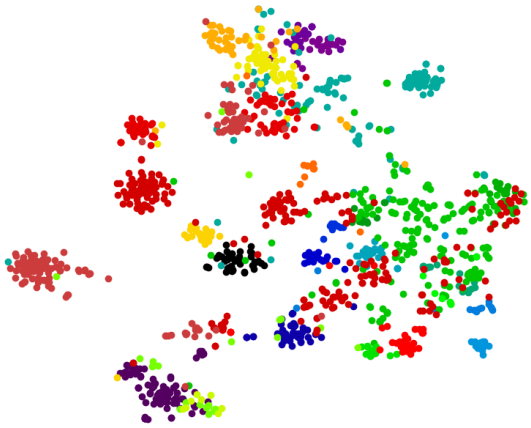
(d) CLUSTERLLM-I-iter (post-training, with keyphrases)

Figure 5: t-SNE Visualization of Embeddings on Finance Dataset

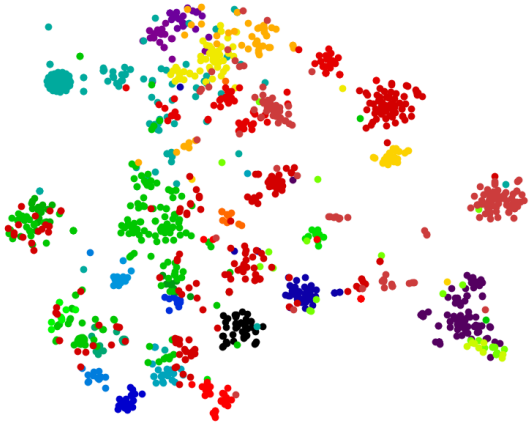
G.3 Insurance Dataset



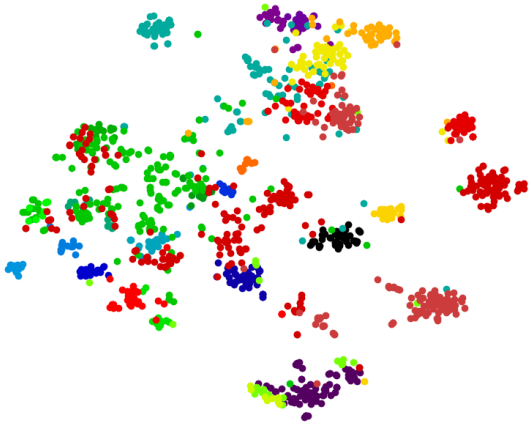
(a) Instructor (pre-training, no keyphrases)



(b) Instructor (pre-training, with keyphrases)



(c) CLUSTERLLM-I-iter (post-training, no keyphrases)



(d) CLUSTERLLM-I-iter (post-training, with keyphrases)

Figure 6: t-SNE Visualization of Embeddings on Insurance Dataset

H Cluster Purity and Num of Utterances

H.1 Banking Dataset

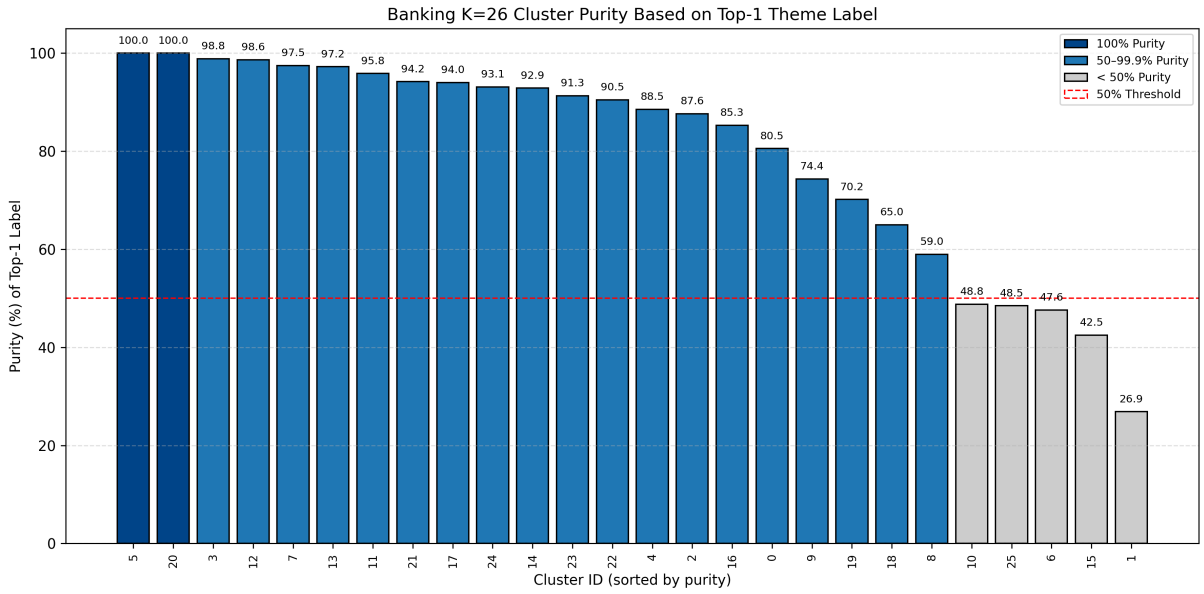


Figure 7: Histogram of cluster-purity on the Banking Dataset (K=26). Two clusters achieve 100% purity, whereas five clusters have purity below 50%.

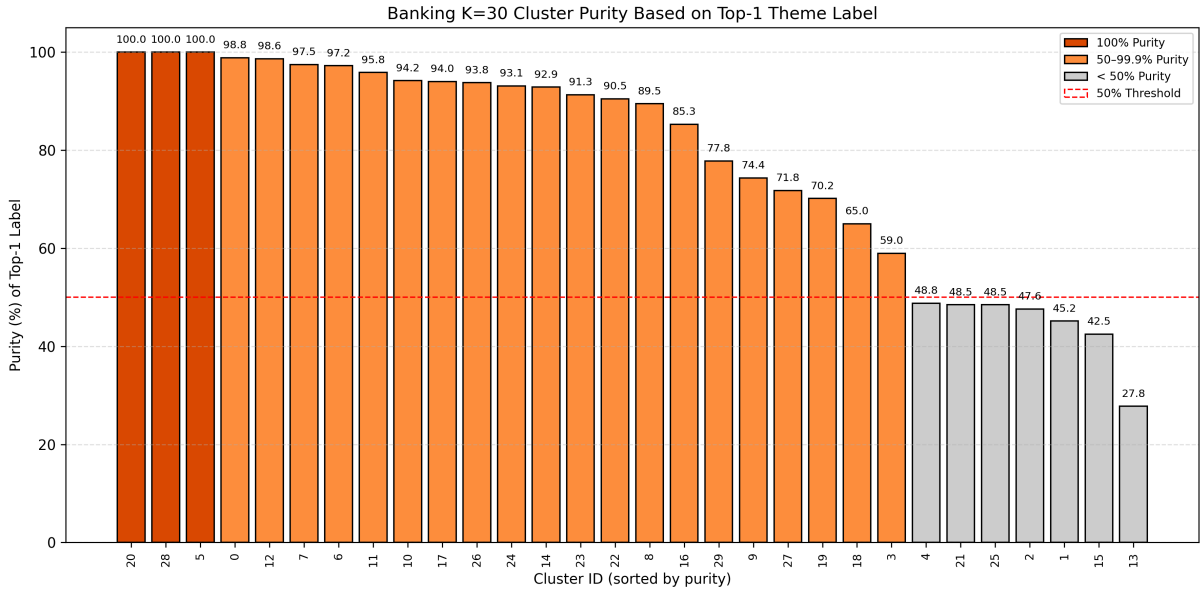


Figure 8: Histogram of cluster-purity on the Banking Dataset (K=30). Three clusters achieve 100% purity, whereas seven clusters have purity below 50%.

H.2 Finance Dataset

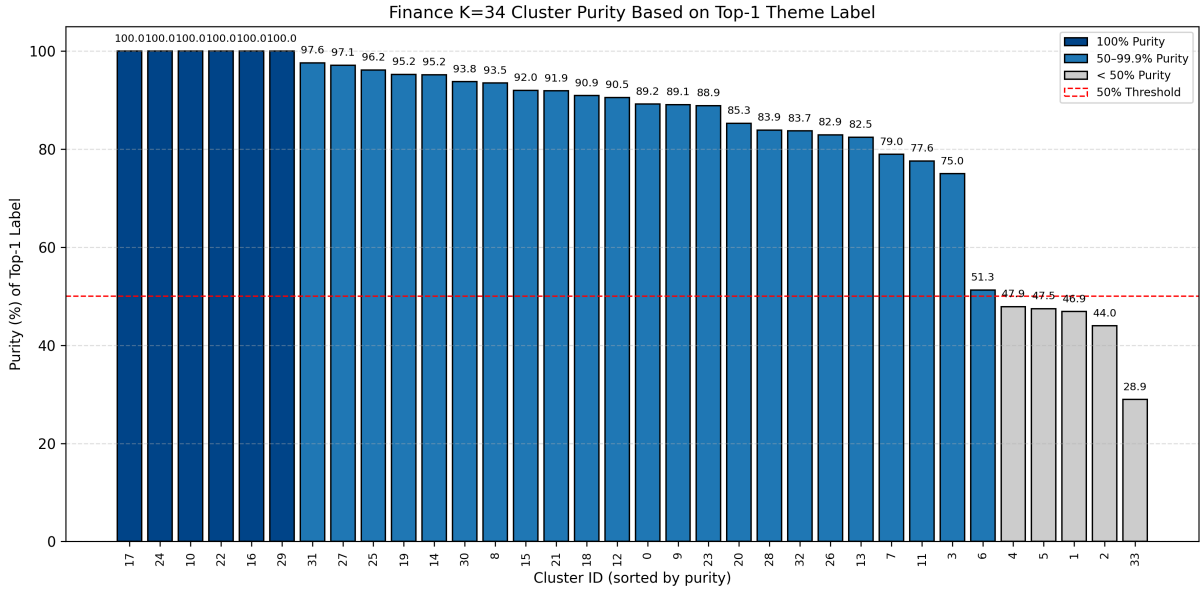


Figure 9: Histogram of cluster-purity on the Finance Dataset (K=34). Six clusters achieve 100% purity, whereas five clusters have purity below 50%.

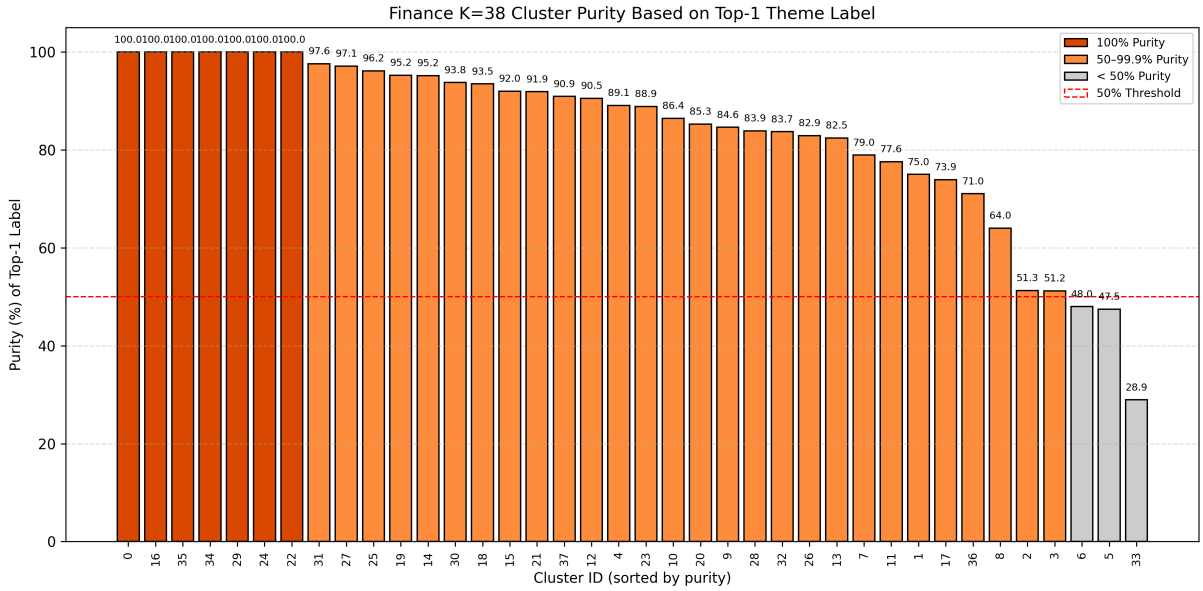


Figure 10: Histogram of cluster-purity on the Finance Dataset (K=38). Seven clusters achieve 100% purity, whereas three clusters have purity below 50%.

H.3 Insurance Dataset

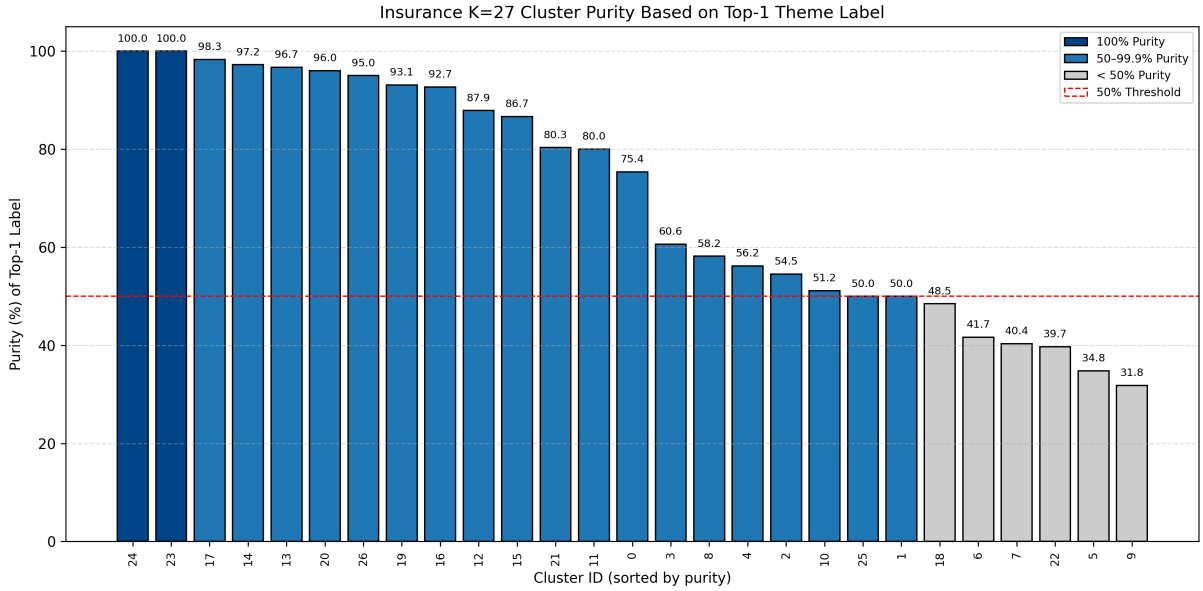


Figure 11: Histogram of cluster-purity on the Insurance Dataset (K=27). Two clusters achieve 100% purity, whereas six clusters have purity below 50%.

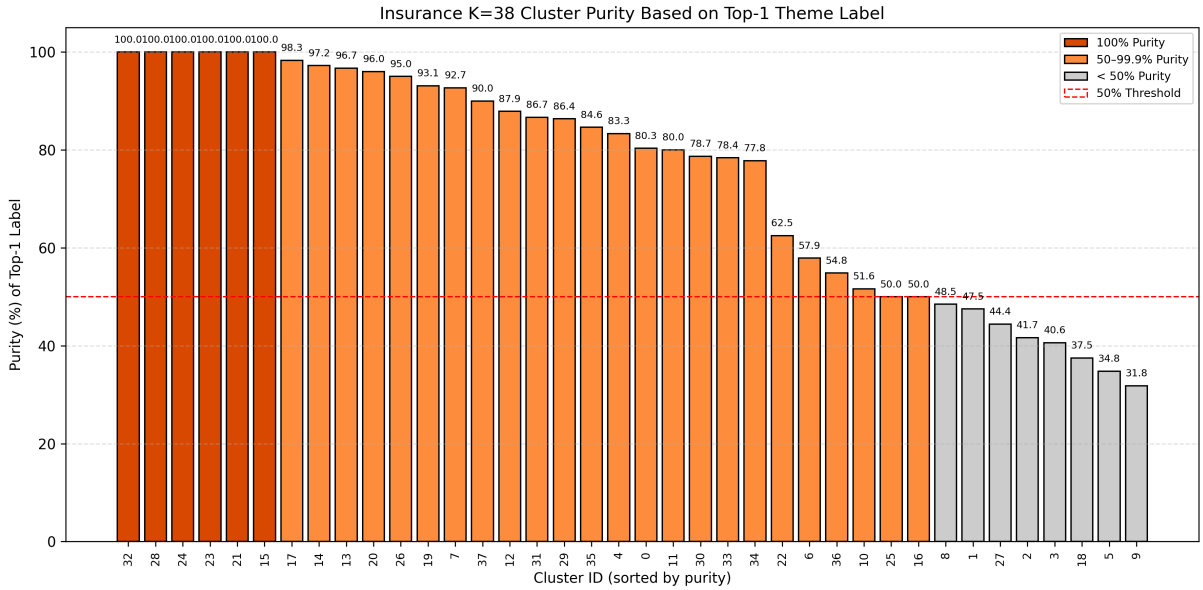


Figure 12: Histogram of cluster-purity on the Insurance Dataset (K=38). Six clusters achieve 100% purity, whereas eight clusters have purity below 50%.

H.4 Num of Utterances

Dataset	#Clusters	Cluster ID	Theme Label	#Utterances	Total Utterances
Banking	K=26	5	perform operations with limits	32	74
		20	request/check status of charge dispute	42	
	K=30	5	perform operations with limits	32	131
		20	request/check status of charge dispute	42	
		28	make external wire transfer	57	
Finance	K=34	10	get branch location/hours	42	223
		16	ask about correspondence	42	
		17	get consumer price index	28	
		22	change account or card pin	40	
		24	update phone/email/address	38	
		29	cancel/order check	33	
	K=38	0	get branch location/hours	42	238
		16	ask about correspondence	42	
		22	change account or card pin	40	
		24	update phone/email/address	38	
		29	cancel/order check	33	
		34	schedule appointment	15	
		35	get consumer price index	28	
Insurance	K=27	23	find agent	14	56
		24	get billing info	42	
	K=38	15	get branch location/hours	13	114
		21	file automobile claim/report accident	20	
		23	find agent	14	
		24	get billing info	42	
		28	get plan info	9	
		32	get plan info	16	

Table 25: Number of utterances contained in 100%-purity clusters.

I Predicted Theme Labels

Dataset	Utterance	Predicted Theme Label	Theme Label (Ground Truth)
Banking	first, could you give me my balance perhaps? Maybe I can figure it out that way.	check account balance	check account balance
	OK thanks. I really just need an ATM.	find atm	find atm
	yeah actually I was thinking of opening up a savings account.	open bank account	open bank account
	I need it transferred to my new checking account.	make wire transfer	make external wire transfer
	Oh I was wondering where your nearest branch location is?	find nearest branch	find branch
Finance	Also, what are your your hours at at at at the branch over there on on Baker Street?	get branch location/hours	get branch location/hours
	Yes, I'm trying to find out what I owe for my credit card.	check credit card balance	check credit card balance
	I need to find out what my net income is from January to June of this year.	get net income	get net income
	Thank you, I just, I'm looking for some. A line of credit, perhaps.	apply for line of credit	apply for line of credit
	Yes, so I was wondering if you could tell me the current CPI, please?	request consumer price index	get consumer price index
Insurance	Marian Wright here, Timothy. I was trying to pay my insurance online, and it did not confirm the submit.	check payment status	check payment status
	I have had an incident in my garage workshop.	file poperty claim	file poperty claim
	Yes, I was billed twice this month, and I need to see what's going on.	report billing issue	report billing issue
	Yes and my ex husband knows that so I would like to change it.	change security question	change password/security question
	Yes, I definitely need to speak to a supervisor. This is is craziness thing I have ever heard!	request call back	request callback

Table 26: Predicted theme labels in NATCS Dataset.

Controllable Conversational Theme Detection Track at DSTC 12

Igor Shalyminov, Hang Su, Jake Vincent, Siffi Singh, Jason Cai, James Gung, Raphael Shu, Saab Mansour
Amazon
{shalymin,shawnsu,jakevinc,siffis,cjinglun,gungj,zhongzhu,saabm}@amazon.com

Abstract

Conversational analytics has been on the forefront of transformation driven by the advances in Speech and Natural Language Processing techniques. Rapid adoption of Large Language Models (LLMs) in the analytics field has taken the problems that can be automated to a new level of complexity and scale.

In this paper, we introduce *Theme Detection* as a critical task in conversational analytics, aimed at automatically identifying and categorizing topics within conversations. This process can significantly reduce the manual effort involved in analyzing expansive dialogs, particularly in domains like customer support or sales. Unlike traditional dialog intent detection, which often relies on a fixed set of intents for downstream system logic, themes are intended as a direct, user-facing summary of the conversation's core inquiry. This distinction allows for greater flexibility in theme surface forms and user-specific customizations.

We pose *Controllable Conversational Theme Detection* problem as a public competition track at Dialog System Technology Challenge (DSTC) 12 — it is framed as joint clustering and theme labeling of dialog utterances, with the distinctive aspect being controllability of the resulting theme clusters' granularity achieved via the provided user preference data.

We give an overview of the problem, the associated dataset and the evaluation metrics, both automatic and human. Finally, we discuss the participant teams' submissions and provide insights from those. The track materials (data and code) are openly available in the GitHub repository.

1 Introduction

Conversational analytics — at the intersection of Speech and Natural Language Processing — has undergone rapid transformation due to advances in both fields. *Automatic Speech Recognition (ASR)* now enables accurate transcription of conversations across diverse domains and durations. Simultaneously, *Natural Language Processing* (especially *Information Retrieval*) has enabled large-scale analysis of conversational data, revealing patterns such as word usage, emotional tone, and discussed topics. More recently, *Large Language Models (LLMs)* have elevated the complexity and quality of analysis tasks. For instance, large-scale text embedding models (Wang et al., 2024) significantly enhance document similarity search by capturing semantic meaning beyond surface forms.

In this paper, we propose the task of *Theme Detection*, a key problem in conversational analytics. Themes reflect the high-level topics discussed in conversations and aid in categorizing them by function — e.g., customer support, sales, or marketing. Automatically identifying and labeling themes can greatly reduce the manual effort required to analyze long conversations.

While related to dialog intent detection, theme detection serves a different purpose. Intents are typically tied to a fixed schema and used for downstream system logic. In contrast, themes are final outputs for users (e.g., analysts), summarizing the customer's inquiry and supporting diverse surface forms and customizations.

We introduce the task of *Controllable Conversational Theme Detection* as a new track in the Dialog System Technology Challenge (DSTC) 12. Building on the DSTC 11 track on Open Intent Induction (Gung et al., 2023b), our challenge adds two major innovations: (1) joint theme detection and labeling, and (2) controllable theme granularity. The latter enables customization of theme clusters based on user preferences — motivated by real-world use cases where businesses may want finer or coarser thematic distinctions.

This task is designed for a zero-shot setting on unseen domains. Models will be guided by user preference data (detailed in Section 4) to align both labels and cluster granularity. While especially

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 74–90

compelling in the context of LLMs, the proposed setup does not require their use.

2 Related Work

In this section, we discuss prior work related to the distinctive aspects of our proposed task.

2.1 Unsupervised dialog theme / intent detection

The task of open conversational intent induction was introduced in a DSTC 11 track by Gung et al. (2023b), which focused on utterance clustering in two setups of varying complexity: (1) intent detection with pre-defined intentful utterances to be clustered, and (2) open intent induction, which required identifying and clustering such utterances.

In contrast, our task involves a single setup with pre-defined themed utterances, and the goal is to jointly cluster and label them according to specific evaluation metrics. Unlike intent induction, we do not restrict the surface form of theme labels. Instead, labels are assessed based on their structural quality and functional usefulness for analysis (see Section 6 and Appendix B).

2.2 Controllable clustering

Our goal of controllable theme granularity builds on the concept of *constrained clustering*. A comprehensive taxonomy of constraint-based clustering tasks is provided by González-Almagro et al. (2025). We adopt instance-level pairwise constraints ("should-link" / "cannot-link"), implementing a semi-supervised clustering approach where supervision comes from labeled utterance pairs. This setup has been well-studied, from early work by Basu et al. (2004) to more recent approaches by Zhang et al. (2019) and Viswanathan et al. (2024).

2.3 Clustering with LLMs

The use of LLMs for utterance clustering has gained traction. Zhang et al. (2023) propose using hard triplets ("does A match B better than C?") derived from a teacher LLM to fine-tune a smaller embedding model and refine clusters via a hierarchical method similar to HAC (Manning et al., 2008). While this method enables controllable clustering guided by LLMs, it focuses solely on clustering — cluster labeling remains out of scope. In contrast, our task requires *labeled theme clusters*, combining clustering with label generation to better reflect real-world needs.

Viswanathan et al. (2024) provide a thorough study on integrating LLMs into clustering workflows. They identify three points of intervention: (1) *pre-clustering*, using LLMs to generate keywords and enrich input texts; (2) *during clustering*, by expanding human-provided pairwise constraints; and (3) *post-clustering*, correcting uncertain assignments with LLM-based prompting. Although their framework aligns well with our goals, their focus remains on clustering rather than labeling.

3 Task Description

The task of *Controllable Conversational Theme Detection* is defined as follows. The input data are:

1. a dataset of conversations with some utterances within them labeled as "themed" (those conveying the customer's requests, possibly several per conversation)

2. a set of preference pairs covering a sample of all the themed utterances and representing what pairs should belong to the same theme and which should not. — which we refer to as "should-link" and "cannot-link" pairs, respectively

3. a theme label writing guideline outlining the requirements to a label as both a linguistic expression and an analytical tool.

The goal of the task is to:

- cluster the themed utterances so that each cluster represents a meaningful semantic / thematic group, is distinguishable from other theme clusters and satisfies the should-link / cannot-link requirements on its utterances (if it contains utterances included in the preference data)

- give each theme a short, concise and actionable natural language label (more detail on our evaluation criteria is given in Section 6).

3.1 Controlling theme granularity

In the way we intend to control theme granularity, we loosely follow the Stage 2 approach of Zhang et al. 2023. That work described a data-efficient approach with user preference data in the should-link / cannot-link form. As such, if user preferences indicate that the utterances *"I want to purchase pet*

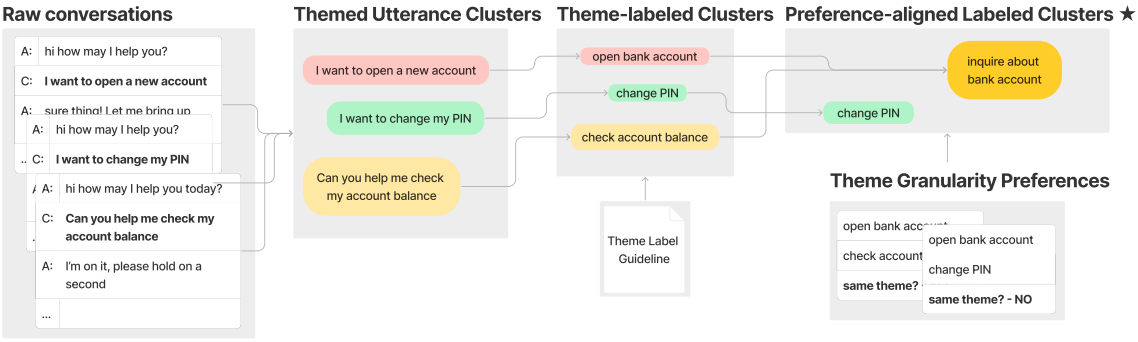


Figure 1: Diagram of the proposed task in the form of an example processing pipeline. The inputs to the "system" are raw conversations, user preferences on the theme granularity and theme label guidelines; the output is preference-aligned utterance clusters with the corresponding theme labels (marked with ★)

insurance" and *"I want to purchase travel insurance"* should belong to the same theme, all the utterances like these two would be associated to the single theme whose label semantically unifies both of the two utterances' meanings e.g. *"purchase insurance"* or some close paraphrase of it. On the other hand, if the preferences elicit that *"I want to find the closest branch"* and *"Give me the directions to the closest ATM"* should not belong to the same theme, the corresponding themes *"find branch"* and *"find ATM"* as well as the clusters of utterances belonging to them should be kept as separate. Some example usages of such data include contrastive fine-tuning of utterance representation as done by e.g. Chu et al. (2023) and Zhang et al. (2021) or adjusting the initial clusters/themes, as depicted in Figure 1.

3.2 Expected result

A successful completion of the task would assume assigning each utterance a theme label so that:

- theme labels are concise, exhaustively cover all the examples and are mutually exclusive,
- label wording conforms to the Theme label writing guideline (Appendix B),
- theme granularity matches the 'gold' held-out assignment which is supposed to be inferred from the provided user preference samples.

A visualization of the overall task is presented in Figure 1 where we depict a potential sequential pipeline as an example. The actual submissions can vary in architecture and the types of models used. We intend the problem to be solved in a zero-shot weakly supervised way, in the sense that all the training/development data provided to the participants has no domain overlap with the test data (more detail on the data in Section 4), and the only supervision signals provided are 1) user preference data covering a sample of the dataset and 2) theme label writing guideline.

While the input data suggests LLM-based solutions, we encourage the participants to use techniques from both LLM-based and traditional Machine Learning paradigms that adequately correspond to the problem specifics.

4 Data

We build our task on top of the NatCS (Gung et al., 2023a,b), a multi-domain dataset of human-human customer support conversations — the dataset statistics per domain are provided in Table 2.

We intend for the participants' submissions to work in a zero-shot setup naturally supported within the LLM-centered framework. As such, we provide the three original NatCS domains: **Banking**, **Finance** and **Insurance** — for the participants to use for the training/development purposes and assess the domain generalization of their approaches.

Our theme labels closely resemble the original intent annotations in NatCS, though those were altered in the following ways:

1. intent labels' surface form was rewritten where needed to conform with the theme label writing guideline (see Appendix B),
2. for each original intent label, we provide two theme labels, a more specific one and a more vague one, for the flexibility of evaluation,

Table 1: User Preference Data Statistics

Domain	# Should-link pairs	% data covered	# Cannot-link pairs	% data covered
Banking	164	10.04%	164	10.04%
Finance	173	10%	173	10%
Insurance	155	8.99%	126	7.30%
Travel (held out)	77	10.07%	76	9.93%

Table 2: Dialog Dataset Statistics

Domain	# Dialogs	# Themed utterances
Banking	980	1634
Finance	3000	1725
Insurance	836	1333
Travel (held out)	999	765

3. intent clustering itself was altered to reflect our task's custom theme granularity,

4. some noisy intent annotations were corrected or otherwise dropped.

The held out test domain, **Travel**, is publicly released for the first time in this challenge and has little to no overlap with the train/dev data.

Also introduced in this challenge is theme granularity preference data on top of NatCS dialogs, its statistics are shown in Table 1. We generated preference pairs in the following way.

Should-link pairs: we clustered themed utterances (we leave the specifics of the clustering algorithm behind to prevent evaluation metric hacking) and sampled pairs that belong to the same cluster in the gold assignment but to the different clusters as per the algorithm, with sampling weights set to the normalized cosine distances between the points in the pair (further points that should be in the same theme are more interesting). **Cannot-link pairs**: similarly, we sampled pairs of utterances that belong to different clusters in the gold assignment but to the same cluster as per the algorithm. Sample weights set to $1 - dist(utt_a, utt_b)$ normalized to make a probability distribution, where utt_a and utt_b are the utterances in the pair and $dist$ is cosine distance.

In each case, our target amount of pairs to generate corresponds to 10% of all the themed utterances in the dataset, and preference pairs cover no more than 30% of any given gold cluster's utterances.

5 Baseline and Experimental Setup

We provided the participants with a baseline solution that combines traditional machine learning approaches with LLM-based techniques. As such, the entire baseline workflow consists of 3 stages:

1. **Utterance clustering**. Each themed utterance is embedded with SentenceBERT (`all-mpnet-base-v2` model is used, Reimers and Gurevych 2019), then the embeddings are clustered using the K-means algorithm (Jin and Han, 2010) with 10 clusters by default and the `k-means++` initializer (Arthur and Vassilvitskii, 2007).

2. **Theme cluster adjustment to user preferences**. We apply a naïve algorithm that reassigns cluster labels for every utterance id containing in the should-link / cannot-link sets. For every $< utt_i, utt_j >$ pair in the should-link set, if they are assigned to different clusters, utt_j is re-assigned to utt_i's cluster. In turn, for every $< utt_m, utt_n >$ in the cannot-link set, utt_n is re-assigned to the cluster with the second closest centroid to it. Evidently, the baseline cluster adjustment algorithm doesn't have any generalization outside of the given preference sets.

3. **Theme label generation**. We used an LLM with the prompt as in Appendix B — the default model used in the baseline implementation is `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023). No limitation on the number of in-context utterances was set.

6 Evaluation

Theme assignment that is the result of our task's solution can be assessed from two perspectives: from the controlled clustering perspective and from the theme label generation perspective — our evaluation metrics reflect these two perspectives.

6.1 Automatic evaluation

Automatic evaluation metrics are mainly used for the development purposes and were provided to the participants as part of the starter code.

Table 3: Automatic Evaluation — Theme Label Metrics. Here and below, results in **bold** are the best, underlined are those above baseline.

Team ID	R-1	R-2	R-L	Cos sim	BERT P	BERT R	BERT F1	LLM s1	LLM s2	LLM avg
Team A	32.70%	4.60%	29.82%	59.51%	89.82%	91.20%	90.35%	46.01%	56.47%	51.24%
Team B	5.03%	0.00%	5.03%	37.08%	85.22%	88.02%	86.53%	12.03%	0.13%	6.08%
Team C	**45.22%**	23.81%	**45.10%**	**69.91%**	**95.02%**	**94.69%**	**94.71%**	**100.00%**	**99.48%**	**99.74%**
Team D	34.57%	21.31%	34.27%	55.93%	92.52%	91.48%	91.91%	80.39%	76.60%	78.50%
Team E	42.28%	16.50%	41.22%	62.48%	93.85%	92.84%	93.27%	93.46%	95.69%	94.58%
Team F	23.10%	0.79%	21.14%	46.02%	85.67%	89.29%	87.19%	4.05%	3.53%	3.79%
Baseline	43.74%	**24.56%**	42.87%	59.68%	89.25%	89.87%	89.52%	20.39%	39.48%	29.93%
BL-prefs	29.27%	4.21%	24.69%	48.79%	85.31%	87.77%	86.44%	12.81%	18.43%	15.62%

6.1.1 Clustering metrics

- **NMI** score (Vinh et al., 2010) — *Normalized Mutual Information* is a function that measures the agreement of the two cluster assignments, reference and predicted, ignoring permutations. Normalization is performed over the mean of the entropies of the two assignments

- **ACC** score (Huang et al., 2014) evaluates the optimal alignment between the reference cluster assignment and the predicted one, with the alignment obtained using the Hungarian algorithm.

6.1.2 Label generation metrics

We evaluate the predicted labels for theme clusters in two general ways: 1) similarity to the reference labels, 2) adherence to the theme label guideline.

Similarity of a predicted label to the references is calculated in the following way:

$$Score_i(Y_i, \hat{y}_i) = \max_j sim(Y_{i,j}, \hat{y}_i)$$

where Y_i are the reference labels for the i-th utterance (we provide two labels with a more specific and a more vague wording, respectively), y_i is the predicted label for the same utterance and sim is one of the similarity functions listed below.

- **Cosine similarity** — the semantic similarity measure over SentenceBERT embeddings (`all-mpnet-base-v2` model is used, Reimers and Gurevych 2019) of the reference and predicted labels,

- **ROUGE** score (Lin, 2004) — an token-level N-gram overlap metric useful for comparing short and concise word sequences,

- **BERTScore** (Zhang et al., 2020) combines the agility of embedding-based similarity and the interpretability of token-level overlap. The model tokenizes each utterance and generates a contextual embedding for each token. Then, a cosine similarity $sim_{i,j}$ is calculated between i-th token of the reference and j-th token of the prediction. We report BERTScore **Precision** (for each token in the prediction, finding the reference token with the highest similarity), **Recall** (for each token in the reference, finding the prediction token with the highest similarity) and **F1 score**.

Adherence to the guideline is evaluated with an LLM-as-a-Judge prompted with a version of the guideline attached in the Appendix B (it was provided for the participants). For the usage with the LLM, it was split into three sections spanning structural and functional criteria, i.e. how good the label is as a linguistic expression and how good it is as an analytical tool, respectively. For the sake of preventing evaluation metric hacking, we shared a different / condensed version of the guideline to the participants and kept the full version held out. For evaluation during the development phase, our provided code used a self-hosted solution with `vicuna-13b-v1.5` (Zheng et al., 2023) as the default LLMaaJ backbone. In the automatic evaluation of the final submissions, we used Claude 3.5 Sonnet (Anthropic, 2024). Our repository contains both the public version of the label style evaluation prompt (3 condensed sections optimized for usage with public self-hosted LLMs) and its held out version (2 expanded sections optimized for usage with Claude, uploaded after the end of the competition).

6.2 Human Evaluation

All submissions underwent expert human evaluation in order to verify automated evaluation results

and to expand the automated evaluation methodology to more precisely assess each solution's performance. The evaluation dimensions were divided into two broad categories covering formal and functional criteria, and each of these areas had additional subdimensions to be rated by evaluators in a binary fashion (*pass/fail*) using criteria distributed into into two broad categories: Structural/Functional. The structural criteria were based on the theme labeling guidelines provided to participants.

Structural Criteria (Theme Label as a Linguistic Expression): Conciseness & Word Choice, Grammatical Structure

Functional Criteria[1] (Theme Label as an Analytical Tool): Semantic Relevance, Analytical Utility, Granularity, Actionability, Domain Relevance, Thematic Distinctiveness.

The guidelines for each of these dimensions, along with the positive and negative examples provided to evaluators (with reasoning), are laid out in Appendix C. The theme labeling guidelines, upon which the structural criteria were based, are defined in Appendix B. The annotation task was completed in a single-pass way by two members of the track organizing team.

7 Results and Analysis

Table 6: Automatic Evaluation — Clustering Metrics

Team ID	ACC	NMI
Team A	48.37%	42.02%
Team B	17.91%	1.97%
Team C	**67.97%**	**70.39%**
Team D	51.76%	47.71%
Team E	35.82%	47.73%
Team F	26.67%	9.06%
Baseline	53.2%	50.59%
BL-prefs	47.97%	45.39%

We received submissions from 6 participant teams. During the development, the teams were free to use the provided public data across 3 domains for creating their own train / development setups and testing e.g. out-of-domain generalization of their approaches. The test domain was made public during the last week of the competition. When submitting the inference results via an online form, the participant teams were asked to provide a brief info about their approaches. Below are the questions and the summaries of the submitted answers:

What LLM type did you use? (*Open-source — self-hosted / Proprietary via API / No LLM / Other*)

Teams A, C and F used a proprietary API; teams B, D and E used an open-source self-hosted LLM.

How large of an LLM did you use? *(<30B / 30—100B / >100B / Unknown (proprietary API) / No LLM / Other*)

Team A, C and F's model size is unknown; teams B, D and E used a model with <30B parameters.

Did you use any conversational information (previous / past context of the utterance)? Please specify if yes

Team C used the context window of 5 turns; Team E used conversational context within the predicted topic segment.

What clustering algorithm did you use?

Team A used HDBScan (Campello et al., 2013); Teams B and D used K-Means (Jin and Han, 2010); Team C used ClusterLLM (Zhang et al., 2023); Team F experimented with K-Means, DBSCAN and HDBSCAN; Team E used Spectral Clustering (Shi and Malik, 2000).

What text embedding model did you use?

Teams A and C used Instructor model (Su et al., 2023); Teams B, D and E used SentenceBERT as per the baseline.

Did you use an embedding dimensionality reduction technique? (Please specify which one if yes)

Teams A and E used UMAP (McInnes et al., 2018).

Did you use a data augmentation technique (please specify what kind)?

Team A used Speech Acts as a data augmentation; Team B used SimCSE (Gao et al., 2021); Team E used contrastive learning to augment the limited unlabeled data.

How did you use the should-link / cannot-link pairs?

Teams A, B and D used the baseline approach. Team C used an LLM to re-assign the clusters for all the utterances from the should-link pairs. For

[1] All functional criteria dimensions were evaluated at the level of the utterance except for *Thematic Distinctiveness*, which was evaluated for each cluster label.

Table 4: Human Evaluation — Per-utterance Functional Metrics

Team ID	Semantic Relevance	Analytical Utility	Granularity	Actionability	Domain Relevance
Team A	77.25%	63.66%	22.75%	56.21%	79.74%
Team B	64.97%	12.94%	0.00%	4.05%	97.78%
Team C	**89.67%**	**82.75%**	47.84%	**74.77%**	**98.82%**
Team D	68.76%	63.66%	26.41%	60.26%	94.25%
Team E	86.27%	54.64%	22.48%	54.51%	91.11%
Team F	45.23%	41.57%	7.71%	41.57%	67.45%
Baseline	86.61%	66.84%	**47.98%**	66.84%	89.6%
BL-prefs	88.76%	42.09%	20.00%	42.09%	83.92%

Table 5: Human Evaluation — Per-cluster Metrics

	Structural		Functional
Team ID	Conciseness	Grammatical Structure	Thematic Distinctiveness
Team A	83.33%	**100.00%**	75.76%
Team B	**100.00%**	33.33%	0.00%
Team C	**100.00%**	**100.00%**	**91.11%**
Team D	91.67%	66.67%	90.91%
Team E	93.65%	93.65%	78.34%
Team F	95.00%	**100.00%**	72.63%
Baseline	80.00%	30.00%	**91.11%**
BL-prefs	80.00%	20.00%	66.67%

the cannot-link pairs, the LLM was used to identify the utterance of the pair not belonging to the cluster, and then to make the re-assignment. Team E trained a reward model from the should-link and cannot-link pairs that was later incorporated into the clustering algorithm to impose soft constraints.

Did you use the theme label styleguide — if yes, how?

Team C used General Schema to extract verbs and nouns for each utterance in the cluster, then using those, they generated theme labels. Theme D instructed the labeling LLM to generate Verb-Object pairs. Teams E and F used the provided styleguide itself. Team F added it directly into the labeling LLM's prompt, and Team E modified and simplified it first.

Short (1-2 paragraph) description of your approach

Team A proposed a cluster-then-label framework for thematic clustering of utterances. First, they compute utterance embeddings using either Sentence Transformers, InBedder, or Instructor models depending on the embedding type. they then apply clustering (KMeans or HDBSCAN with UMAP-based dimensionality reduction) to group thematically similar utterances. Clustering is refined using manually provided should-link and cannot-link preference pairs, ensuring better alignment with human notions of similarity. After clustering, each cluster is labeled automatically by prompting an LLM (ChatGPT or Gemini Flash) with a batch of utterances, extracting a theme label and brief explanation. The resulting predicted labels are assigned back to utterances, forming the final output for evaluation. This approach leverages both unsupervised structure discovery and lightweight LLM-based supervision for scalable and interpretable theme labeling.

Team B used SCCL (Zhang et al., 2021) and applied SimCSE for data augmentation. After training the SCCL, they clustered the utterances with K-Means. They performed hyperparameter search for the number of clusters based on the Silhouette score and set it to 7. User preference data was not used.

Team C first extracted keyphrases from conversations using an LLM. They also determined the numbers of clusters based on the Silhouette coefficient. Clustering was performed using ClusterLLM, and the embedder was fine-tuned on the clustered utterances. Subsequently, among the two candidates with the highest preference pair accuracy, the candidate with the greater number of clusters was selected as the final model. Utterances were then adjusted according to the preference pairs. Finally,

for the clustered utterances, a general schema was extracted in terms of verbs and nouns, and based on both the schema and the utterance content, the final theme labels were generated.

Team D explored two approaches. The first one involved designing a prompting strategy to generate concise labels in a Verb–Object format (e.g., *"update address"*, *"book flight"*), allowing for more structured and comparable cluster representations. The second approach used `LLaMA-3.1-8B-Instruct` to evaluate whether two utterances (with dialog history) belonged to the same cluster, based on their distance from the cluster center. The second method showed limited performance, and they submitted results using the first one, with a more robust prompting-based labeling strategy.

Team E propose PrefSegGen, a preference-aware topic segmentation and generation framework that addresses low-resource conversational theme understanding by integrating topical-structured context modeling with user-preference-aligned theme generation. First, they introduce a novel two-stage self-supervised contrastive learning topic segmentation framework to obtain the topic segment to which the target utterance belongs under low-resource conditions. It initially leverages the unlabeled dialogues to pretrain topic encoders (`bert-base-uncased` & `sup-simcse-bert-base-uncased`) on coherence and similarity patterns, followed by supervised fine-tuning with minimal labeled data to enhance segmentation precision. Subsequently, they incorporate a reward-guided clustering mechanism to guarantee that the generated themes are both contextually grounded and preference-aligned. A reward model, trained on should-link and cannot-link pairs, dynamically assigns linkage weights that reflect semantic proximity in line with user expectations. These weights guide spectral clustering after UMAP-based embedding reduction. Crucially, for each target utterance, they utilize its segmented topical context as input when prompting `LLaMA3-8B-Instruct`, coupled with the official style guide, to generate hierarchical theme labels. An ensemble refinement process further enhances topic consistency by filtering low-frequency labels, yielding final outputs that are structurally coherent, context-aware, and tailored to user preferences.

Team F employed a large language model (LLM) to annotate utterances based on preference signals, and subsequently attempted to merge clusters according to the LLM-based annotation.

Our evaluation results reveal that **Team C**'s approach achieved the highest accuracy across the board on both human and automatic metrics. It was tied with **Team B** on Label Conciseness, and with **Teams A** and **F** on Grammatical Structure. Although only **Team C**'s approach achieves **100%** on both, signifying that its label generation works in full accordance with the styleguide. **Team C** was also the only one to surpass the baseline on automatic clustering metrics. **Team E** achieved the second best overall performance in both automatic and human evaluation and **Team D** placed third.

It is noteworthy that for all the three winning places, the ranking induced by the automatic metrics matched that by the humans — indicating that 1) automatic similarity metrics are applicable for short text, and 2) automatic evaluation of higher-level concepts like our label guideline is sufficiently accurate with frontier LLMs-as-Judges.

8 Conclusions

In this paper, we introduced *Theme Detection* as a critical task in conversational analytics, and the associated *Controllable Conversational Theme Detection* competition track at Dialog System Technology Challenge (DSTC) 12 — where joint theme clustering and cluster label generation was further combined with the custom theme cluster granularity controllable via the provided preference data.

We gave an overview of the competition setup including the problem, the benchmark dataset and the details of evaluation, both automatic and human. We presented the participant team's submissions and gave an analysis of the insights from those.

We hope that this new problem, together with the dataset and the insights obtained from the competition will foster further research and advancements in Conversational AI.

9 Acknowledgements

We express our deep gratitude to Dr. Daniel Goodhue for his assistance at the final submission evaluation stage.

References

Anthropic. 2024. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-06-13.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA. Society for Industrial and Applied Mathematics.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004*, pages 333–344. SIAM.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Caiyuan Chu, Ya Li, Yifan Liu, Jia-Chen Gu, Quan Liu, Yongxin Ge, and Guoping Hu. 2023. Multi-stage coarse-to-fine contrastive learning for conversation intent induction. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 31–39, Prague, Czech Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Germán González-Almagro, Daniel Peralta, Eli De Poorter, José Ramón Cano, and Salvador García. 2025. Semi-supervised constrained clustering: an in-depth overview, ranked taxonomy and future research directions. *Artif. Intell. Rev.*, 58(5):157.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023a. NatCS: Eliciting natural customer support dialogues. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9652–9677, Toronto, Canada. Association for Computational Linguistics.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023b. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259, Prague, Czech Republic. Association for Computational Linguistics.

Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. 2014. Deep embedding network for clustering. In *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jianbo Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021.

Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Hongjing Zhang, Sugato Basu, and Ian Davidson. 2019. A framework for deep constrained clustering - algorithms and advances. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part I*, volume 11906 of *Lecture Notes in Computer Science*, pages 57–72. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Cluster Labeling Prompt

```
<task>
You are an expert call center assistant.
You will be given a set of utterances in
<utterances> </utterances> tags, each
one on a new line.
The utterances are part of call center
conversations between the customer and
the support agent.
Your task is to generate a short label
describing the theme of all the given
utterances. The theme label should be
under 5 words and describe the desired
customer's action in the call.

<guidance>
Output your response in the following
way.

<theme_label_explanation>
Your short step-by-step explanation
behind the theme
</theme_label_explanation>

<theme_label>
your theme label
</theme_label>

</guidance>
</task>

H:
<utterances>
{utterances}
</utterances>
```

B Theme Label Writing Guideline

An acceptable theme label is structurally and semantically well-formed according to the rules outlined in this appendix. *Structurally well-formed* means that the words and their arrangement in the theme label are acceptable. *Semantically well-formed* means that the meaning and usability of the theme label are acceptable.

B.1 Theme labels exclude unneeded and undesirable words.

Theme labels should be concise (2–5 words long). They should only include essential words (see B.2 and B.2.1 below). Essential words will primarily include content (open-class) words. Function (closed-class) words should be excluded. Prepositions may be included as needed but should be avoided when there is a synonymous alternative label without a preposition.

Theme labels should also exclude context-sensitive words like pronouns (*him*, *her*, *them*, *it*, *us*, etc.) and demonstratives (*this*, *that*, *those*, etc.).

B.2 Word types

- Content/open-class words:
  - nouns (*items*, *insurance*, *information*, *order*, etc.)
  - main verbs (*check*, *inquire*, *add*, *explore*, etc.)
  - adjectives (*new* patient, *missing* item, etc.)
  - other modifying words (*shipping* information, *product* options, etc.)
- Function/closed-class words:
  - articles/determiners (*the*, *a*, etc.)
  - auxiliary verbs (*have* or *be*, as in *I have eaten* or *I am eating*)
  - copulas
  - negation (*not* or *-n't*, as in *not on time* or *didn't arrive*)

  - conjunctions (*and*, *or*, *but*, etc.)
  - complementizers (clause-embedding uses of *that*, *for*, *if*, *whether*, *because*, etc.)
  - modals (*can*, *could*, *will*, *would*, *may*, *might*, *must*, *shall*)
  - question words (*who*, *what*, *where*, *when*, *how*, *why*)
- Context-sensitive words:
  - pronouns (*she*, *he*, *they*, *it*, *her*, *his*, etc.)
  - demonstratives (*this*, *these*, *that*, *those*, etc.)
  - temporal adverbs (*yesterday*, *tomorrow*, *next week*, etc.)
  - other context-sensitive language
    * *one*, as in *I'm looking for a nearby branch. Can you find* one*?*
    * deleted nouns (noun ellipsis), as in *I found his order, but not yours* __.

B.2.1 Examples

For a theme covering order tracking:

- **Good**: track order
- **Good**: track shipment
- **Bad**: track an order (includes an article)
- **Bad**: track their order (includes a pronoun)

For a theme covering finding the nearest branch of a chain:

- **Good**: find nearest branch
- **Good**: find closest branch
- **Bad**: find nearest one (includes context-sensitive *one*)
- **Bad**: check if there's a nearby branch (includes a complementizer *if*; includes a form of *be*)

B.3 Theme labels are verb phrases that classify events.

A verb phrase begins with a verb and may include arguments or modifiers of the verb (such as a direct object). The verb should be in its citation form, lacking any complex morphology such as tense or agreement suffixes. The citation form of a verb is what would normally follow the infinitive *to*, such as *sign up* in *I'd like to* sign up. Theme labels should not be other phrase types, such as noun phrases.

The verb phrase should describe a class of events. Events are things that can be said to **happen**, unlike states (e.g. *learn* [event] vs. *know* [state]), entities (e.g. *redeem* [event] vs. *redemption* [entity]), properties (e.g. *complain* [event] vs. *angry* [property]), and claims (*report defect* [event] vs. *product is defective* [claim]).

B.3.1 Examples

For a theme covering membership sign-ups:

- **Good**: sign up for membership (verb phrase; describes a kind of *signing up* event)
- **Bad**: signing up for membership (verb phrase, but verb is not in citation form)
- **Bad**: membership sign-up (noun phrase; describes a kind of entity)
- **Bad**: memberships (noun phrase; describes a kind of entity)

For a theme covering requests to check in early at a hotel:

- **Good**: request early check-in (verb phrase; describes a kind of requesting event)
- **Bad**: requested early check-in (verb phrase, but verb is not in citation form)
- **Bad**: request for early check-in (noun phrase; describes a kind of entity)
- **Bad**: customer wants early check-in (this is a claim)

For a theme covering reporting a defective product:

- **Good**: report defective product (verb phrase; describes events)
- **Bad**: reporting defective product (verb phrase, but verb is not in citation form)
- **Bad**: believe product is defective (verb phrase, but describes a state rather than an event)
- **Bad**: defective product (noun phrase; describes a kind of entity)

B.4 Theme labels are informative and actionable yet sufficiently general.

Theme labels should be informative enough to substantially narrow down the set of possible customer issue resolution steps (the steps to resolve the problem/need that drove the customer to make contact). For example, *check balance* is probably associated with a standard procedure for checking the balance of a range of customer account types, but *perform check* is so broad that it could be associated with an extremely diverse group of issue resolutions. Non-actionable theme labels may be excessively vague or uninformative, and hence not very useful.

B.4.1 Examples

For a theme covering appointment-scheduling themes:

- **Good**: schedule appointments
- **Bad**: ask about appointments (probably too general)
- **Bad**: schedule appointment for next week (too specific)
- **Bad**: schedule appointment for elderly parent (too specific)

For a theme covering adding a recognized user to an existing account or policy:

- **Good**: add user
- **Bad**: add one (too general)
- **Bad**: add oldest child (too specific)

For a theme covering user password issues:

- **Good**: reset password
- **Good**: troubleshoot password
- **Bad**: secure account (too general)
- **Bad**: reset password again (too specific)

For a theme covering credit or debit card charge disputes:

- **Good**: dispute charge
- **Bad**: complain about charge (too general)
- **Bad**: file card complaint (too general)
- **Bad**: dispute charge for defective blender (too specific)

C Human Evaluation Guidelines

C.1 Structural Dimensions

C.1.1 Conciseness & Word Choice

Options: Pass (1) / Fail (0)
Definition: The following criteria are consolidated by the evaluator into one Pass/Fail rating for Conciseness & Word Choice:

1. **Label length**: Is the label concise, containing only 2–5 words?
   - **Pass**: update billing address
     Fail: update customer's residential billing address for future statements
     Rationale: The good example uses 3 words, within the required 2-5 word range. The bad example uses 8, making it unnecessarily verbose when the core intent can be expressed more concisely.
   - **Pass**: access account statement
     Fail: statement
     Rationale: The good example uses 3 words, adhering to the 2-5 word guideline. The bad example uses only one word, which lacks sufficient specificity to be useful as a theme label.
2. **Function word exclusion**: Does the label exclude unnecessary function words (articles, auxiliary verbs, etc.)?
   - **Pass**: add dependent coverage
     Fail: add the dependent to coverage
     Rationale: The good example correctly excludes function words like articles ("the"), focusing only on essential content words. The bad example unnecessarily includes "the", which should be excluded according to guidelines.
   - **Pass**: troubleshoot internet connection
     Fail: troubleshoot why internet is not working
     Rationale: The good example properly excludes function words, while the bad example improperly includes function words "why," "is," and "not" which should be excluded for conciseness.
3. **Avoidance of context sensitivity**: Does the label exclude context-dependent words (pronouns, demonstratives, temporal adverbs, etc.)?

- **Pass**: return defective product
  Fail: return this item
  Rationale: The good example avoids context-sensitive words like "this" and uses the general term "product" that can apply across contexts. The bad example includes the context-sensitive demonstrative "this," which requires a specific context to understand its meaning.
- **Pass**: reschedule appointment
  Fail: reschedule it for tomorrow
  Rationale: The good example uses general terminology applicable to any appointment, while the bad example includes both the pronoun "it" and the temporal adverb "tomorrow," both of which are dependent on conversation context for their meaning.

4. **Preposition usage**: Are prepositions included only when necessary?
   - **Pass**: transfer funds
     Fail: transfer from account
     Rationale: The good example avoids unnecessary prepositions by using a concise verb-object structure. The bad example unnecessarily includes the preposition "from" when the more concise alternative without the preposition works just as well.
   - **Pass**: join rewards program
     Fail: sign up for rewards program
     Rationale: The good example avoids prepositions entirely, while the bad example unnecessarily includes the preposition "for" when alternatives without prepositions are available and equally clear.

C.1.2 Grammatical Structure

Options: Pass (1) / Fail (0)
Definition: The following criteria are consolidated by the evaluator into one Pass/Fail rating for Grammatical Structure:

1. **Verb phrase structure**: Is the label a verb phrase?
   - **Pass**: cancel flight
     Fail: flight cancellation
     Rationale: The good example correctly follows the verb phrase requirement by starting with a verb ("cancel") followed by a noun ("flight"). The bad example uses a noun phrase ("flight cancellation") instead.
   - **Pass**: redeem rewards
     Fail: rewards redemption process
     Rationale: The good example uses a verb phrase beginning with the verb "redeem". The bad example fails by using a noun phrase with "redemption" as the head noun rather than using a verb form.

2. **Citation form**: Does the verb appear in its citation form (without tense or agreement morphology)?
   - **Pass**: change delivery address
     Fail: changing delivery address
     Rationale: The good example correctly uses the citation form of the verb "change" without any tense or agreement morphology. The bad example fails by using the -ing form "changing" rather than the required base form.
   - **Pass**: cancel subscription
     Fail: canceled subscription
     Rationale: The good example properly uses the citation form of the verb "cancel" without inflectional endings. The bad example incorrectly uses the past tense form "cancelled" instead of the citation form.

3. **Event classification**: Does the verb phrase describe a class of events, rather than states, entities, properties, or claims?
   - **Pass**: verify warranty coverage
     Fail: warranty coverage
     Rationale: The good example describes an event (the act of verifying) rather than an entity. The bad example describes an entity (the warranty coverage itself) rather than an event, violating the requirement that theme labels classify events. Note: The bad example would also be ruled out by the verb phrase requirement.
   - **Pass**: express dissatisfaction
     Fail: customer is dissatisfied
     Fail: is dissatisfied
     Rationale: The good example describes an event (the act of expressing) rather than a state. The first bad example is

structured as a claim about the customer, rather than describing en event. The second bad example is a verb phrase but describes the wrong kind of situation: a state, rather than an event.

- **Pass**: complain about faulty product (event)
  Fail: angry about faulty product (property)
  Rationale: The good example describes an event (the act of complaining) rather than a property. The bad example describes a property or attribute of the customer, rather than an event describing the customer's intent.

C.2 Functional Dimensions

C.2.1 Semantic Relevance

Options: Pass (1) / Fail (0)
Definition: Does the label accurately capture the core intent/topic of the utterance it represents? Theme labels are expected to provide a gist of the dialogue from the customer's inquiry perspective.

- **Pass**: request card security support (For customer utterance: "I received a notification that my credit card might have been compromised. I need to know what steps I should take.")
  Rationale: This theme label demonstrates good semantic relevance by accurately capturing the core intent of the customer's inquiry—addressing a potential security issue—rather than focusing on peripheral aspects like the notification itself.

- **Fail**: express frustration (For customer utterance: "I've been on hold for 45 minutes trying to get help with activating my new debit card. This is ridiculous!")
  Rationale: This theme label fails the semantic relevance test because it focuses on the customer's emotional state rather than their actual intent, which is to activate their debit card. The frustration is secondary to the core purpose of the contact.

- **Pass**: book accommodation
  Fail: inquire about Chicago
  Rationale: The good example correctly identifies the core intent (booking a hotel room), while the bad example misidentifies the intent as seeking information about Chicago when the location is just a detail/slot related to the booking request.

C.2.2 Analytical Utility

Options: Pass (1) / Fail (0)
Definition: Does the label provide meaningful categorization that could directly support a reviewer or analyst's workflow when reviewing conversation data? Themes, which should be ready for presentation to the user/analyst, are supposed to highlight the topics discussed in the conversation that are useful for categorizing and further analyzing them according to the nature of the conversation.

- **Pass**: troubleshoot checkout error
  For customer utterance: "I'm getting error code E-503 when trying to complete my purchase on your website. I've tried three different browsers."
  Rationale: This theme label has good analytical utility because it categorizes the issue in a way that would allow analysts to, e.g., identify patterns in checkout problems, prioritize technical fixes, and track the frequency of specific error types.

- **Fail**: customer contact
  For customer utterance: "I ordered a blue shirt in size medium last week, but you sent me a red one instead. I'd like to exchange it."
  Rationale: This theme label lacks analytical utility because it's too broad to provide meaningful categorization. It fails to identify the specific issue (there's an order fulfillment error) in a way that could help improve operations or track problem patterns.

- **Pass**: downgrade service plan
  Fail: smart thermostat model TH8000 connection failure with iOS app version 3.2.1
  Rationale: The good example provides useful categorization at the right level of detail for business analysis. The bad example is too specific with technical details that would fragment similar issues into tiny categories, making pattern identification difficult.

C.2.3 Granularity

Options: Pass (1) / Fail (0)
Definition: Does the label maintain appropriate specificity, as determined by its closeness to the provided gold labels? (Submission authors are expected to infer ideal granularity from the provided user preference data.)

- **Pass**: update payment information
  Fail: manage account
  Rationale: The good example demonstrates appropriate granularity by categorizing the issue at a level that's neither too broad nor too specific. The bad example is too broad, grouping potentially diverse issues that would benefit from more specific categorization.

- **Pass**: troubleshoot device connectivity
  Fail: resolve Sony WH-1000XM4 headphones pairing failure with streaming app on Android 16 beta
  Rationale: The good example shows appropriate granularity by categorizing at a level that groups similar technical problems. The bad example has excessive granularity, including specific device models and OS versions that would create overly-fragmented categories.

C.2.4 Actionability

Options: Pass (1) / Fail (0)
Definition: Does the label provide sufficient information to categorize customer issues for resolution? Theme labels should be informative enough to substantially narrow down the set of possible customer issue resolution steps.

- **Pass**: dispute transaction
  Fail: seek assistance
  Rationale: The good example demonstrates good actionability by clearly identifying a specific process (transaction dispute) with established resolution procedures. The bad example is too vague to suggest any specific resolution path.

- **Pass**: trace missing shipment
  Fail: discuss app features
  Rationale: The good example shows good actionability by identifying a specific issue (shipment tracking problem) that points to clear resolution steps. The bad example has poor actionability because "discuss" doesn't point to a specific resolution-related action, and "app features" is too broad.

C.2.5 Domain Relevance

Options: Pass (1) / Fail (0)
Definition: Does the label reflect domain-specific terminology and concepts appropriate to the conversation context? Theme labels should reduce manual analysis by utilizing domain-relevant and context-relevant terminology.

- **Pass**: verify coverage details
  For customer utterance: "I need to know if my insurance policy covers damage from a burst pipe in my basement."
  Rationale: This theme label demonstrates good domain relevance by using terminology ("verify coverage") that's specific to the insurance industry and reflects how claims and policy questions are typically categorized in that domain.

- **Pass**: transfer prescription
  For customer utterance: "I want to transfer my prescription from my old pharmacy to your location. Can you help with that?"
  Rationale: This theme label shows good domain relevance by using standard pharmacy industry terminology ("transfer prescription") that accurately reflects how this process is categorized and handled within the healthcare/pharmacy domain.

- **Fail**: change money amount
  For customer utterance: "I need to increase my 401(k) contribution percentage starting with my next paycheck."
  Rationale: This theme label lacks domain relevance because it uses overly-generic terminology instead of financial industry-specific language. A more domain-relevant label would be "adjust retirement contribution" or "modify investment allocation."

- **Fail**: fix travel problem
  For customer utterance: "My flight was delayed and I missed my connection. I need to be rebooked on the next available flight.")
  Rationale: This theme label has poor domain relevance because it doesn't use airline industry terminology. A more domain-relevant label would be "rebook missed connection", "accommodate disrupted itinerary", etc.

C.2.6 Thematic Distinctiveness

Options: Pass (1) / Fail (0)
Definition: Does the label create a clear boundary that differentiates one theme from the other themes in the dataset? Theme labels should exhaustively cover all the examples AND be mutually exclusive.

- **Pass**: report stolen card
  In this context: Dataset already contains theme labels "report lost card" and "report fraudulent transaction"
  For customer utterance: "Someone stole my wallet and I need to block my credit card immediately."
  Rationale: This theme label demonstrates good thematic distinctiveness by creating a clear boundary between related but distinct issues: lost cards (misplaced by owner), stolen cards (taken by someone else), and fraudulent transactions (unauthorized use).'

- **Fail**: inquire about refund
  In this context: Dataset already contains theme label "request refund"
  For customer utterance: "I returned my purchase last week but haven't seen the money back in my account yet."
  Rationale: This theme label fails the thematic distinctiveness test because it doesn't create a clear boundary between refund requests and refund status checks. The new utterance is about tracking a refund in progress, which should be a distinct category (e.g. "check refund status". Instead, this category could be compatible with utterances that are already covered by "request refund".

- **Pass**: change delivery location
  Fail: reset account
  In this context: Dataset already contains theme labels "schedule delivery", "reschedule delivery", "reset password", and "update account information"
  Rationale: The good example shows appropriate thematic distinctiveness by creating a clear boundary between different delivery modification types. The bad example blurs the boundary between password resets and other profile updates, creating confusion about categorization.

D Input/Output Data Examples

Below is an input datapoint for a dialogue with one utterance marked as themed. For the train/dev domains, the theme labels will be available as in the example below. For the test domain, only the flag that an utterance is themed will be provided.


```
{
  "conversation_id": "Banking_123",
  "turns": [
      {
        "speaker": "Agent",
        "utterance": "Thank you for
            calling Intellibank. This is
             Melanie. How can I help you
            ?"
      },
      {
        "speaker": "Customer",
        "utterance": "Yeah, hey. This is
             John Smith. I've got a
            quick question."
      },
      {
        "speaker": "Agent",
        "utterance": "OK, John. What can
             I help you with?"
      },
      {
        "speaker": "Customer",
        "utterance": "Yeah I need to
            know what your ATM
            withdrawal limits are for
            the day.",
        "theme_label": "get daily
            withdrawal limit",
      },
      {
        "speaker": "Agent",
        "utterance": "Certainly. Our ATM
             withdrawal limit is on a
            per day basis and it is up
            to two hundred dollars."
      },
      {
        "speaker": "Customer",
        "utterance": "Oh perfect,
            perfect. Yeah, I think I'll
            just see if I can head down
            to the ATM now. Thank you."
      },
      {
        "speaker": "Agent",
        "utterance": "OK, thank you. You
             have a great day."
      },
      {
        "speaker": "Customer",
        "utterance": "You too."
      }
    ]
}
```


Below is an input datapoint with the example user preference on clustering granularity:


```
{
  "utterance_a": {
    "utterance": "Yeah, so I need to
        change the account number thing
        that I put in whenever I go to
        the ATM."
    "conversation_id": "Banking_123",
    "turn_id": 4
  },
  "utterance_b": {
    "utterance": "OK. Excellent. Thank
        you Ms. Crystal. And while I got
         you on the phone I see it's
```

```
        been a little bit since you've
        authenticated your account here.
         Would you like to add a PIN
        number to your account for
        security reasons?"
    "conversation_id": "Banking_345",
    "turn_id": 10
  },
  "belong_to_same_theme": "yes"
}
```

Author Index