# 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.

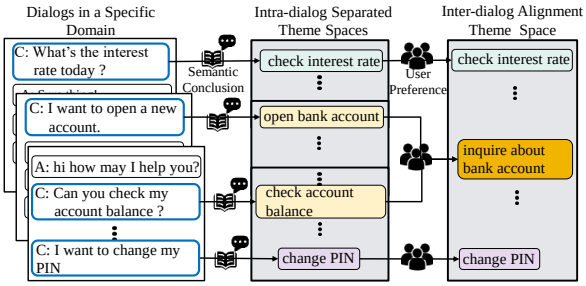


Figure 1: Illustration of the controllable theme detection task. Given a set of dialogues with unlabeled utterances, a theme is generated for each utterance. The theme granularity is influenced by auxiliary inputs such as user preferences, indicating whether a pair of utterances should be grouped under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

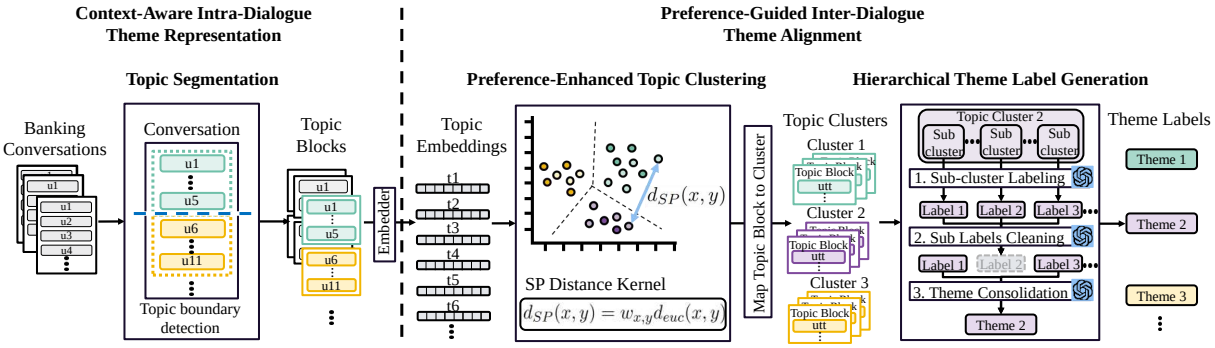


Figure 2: The overall architecture of CATCH.

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.

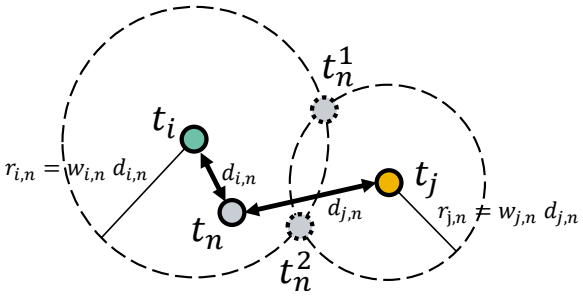


Figure 3: The illustration of positional conflict evoked by SP distance metric. Assume topic $t_n$'s position is defined by its semantic distance ($d_{i,n}$, $d_{j,n}$) to the topic $t_i$ and $t_j$. If the original distance is replaced by the SP distance ($r_{i,n}$, $r_{j,n}$), $t_n$ has two possible 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 $\theta_l$ to be 0.85 and the confidence threshold of splitting $\theta_s$ 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.

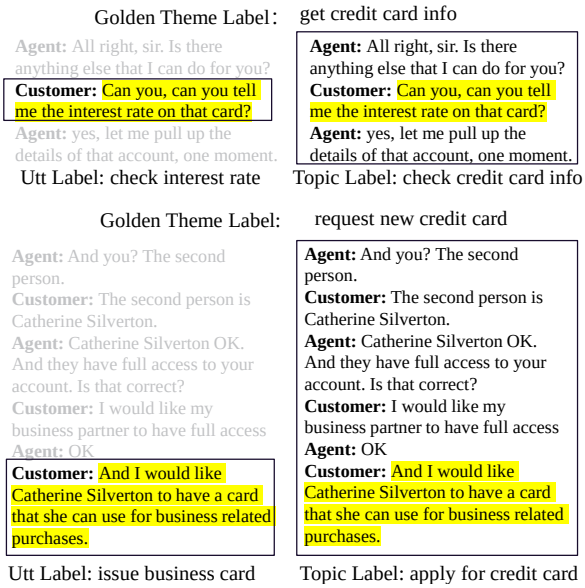


Figure 4: Two samples from finance dataset with utterance id: *"Finance_1e18a3a5_100410_SS01_A6-115"* and *"Finance_35138e33_100917_2464642A2-39"* respectively..

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.