Def-DTS: Deductive Reasoning for Open-domain Dialogue Topic Segmentation

Anonymous ACL submission

Abstract

Dialogue Topic Segmentation (DTS) aims to divide dialogues into coherent segments. DTS plays a crucial role in various NLP downstream tasks, but suffers from chronic problems: data shortage, labeling ambiguity, and incremental complexity of recently proposed solutions. On the other hand, Despite advances in Large Language Models (LLMs) and reasoning strategies, these have rarely been applied to DTS. This paper introduces Def-DTS: Deductive Reasoning for Open-domain Dialogue Topic Segmentation, which utilizes LLM-based multi-step deductive reasoning to enhance DTS performance and enable case study using intermediate result. Our method employs a structured prompting approach for bidirectional context summarization, utterance intent classification, and deductive topic shift detection. In the intent classification process, we propose the generalizable intent list for domain-agnostic dialogue intent classification. Experiments in various dialogue settings demonstrate that Def-DTS consistently outperforms traditional and state-of-the-art approaches, with each subtask contributing to improved performance, particularly in reducing type 2 error. We also explore the potential for autolabeling, emphasizing the importance of LLM reasoning techniques in DTS.

1 Introduction

Dialogue Topic Segmentation (DTS) is a task that aims to divide a dialogue into segments where each segment focuses on a coherent topic. Figure 1 shows an example of topic shift within a single dialogue. DTS is crucial for various natural language processing (NLP) tasks, including response prediction (Lin et al., 2020; Xu et al., 2021b; He et al., 2022), response generation (Li et al., 2016; Xu et al., 2021a; Liu et al., 2022), dialogue state tracking (Das et al., 2024), summarization (Bokaei et al., 2016; Chen and Yang, 2020; Qi et al., 2021; Zhong et al., 2022), question answering (Yoon et al., 2018; Zhang et al., 2022), and machine reading comprehension (Ma et al., 2024).

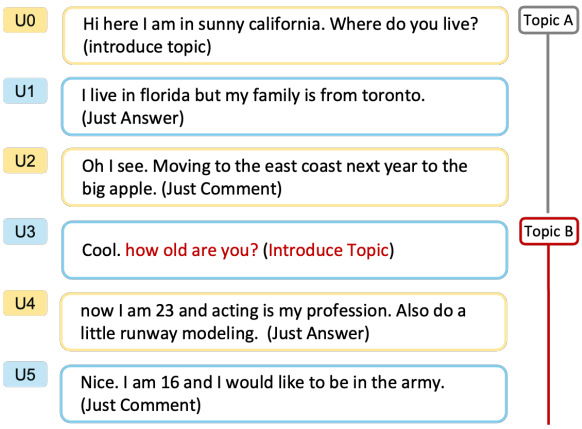

Figure 1: An example of a topic shift in a conversation. The cues for a topic shift are highlighted in red.

Despite growing interest, DTS suffers from several chronic challenges. First, the shortage of annotated data has led most recent DTS studies to an unsupervised way, which generally yields suboptimal performance. Second, ambiguity in segment labeling has hindered the development of effective approaches. Lastly, recent studies: DialSTART (Gao et al., 2023), UR-DTS (Hou et al., 2024) have proposed incremental approaches that require more parameter and complexity, enhancing prior studies: CSM (Xing and Carenini, 2021) and DialSTART (Gao et al., 2023), respectively. This progression indicates that DTS is a challenging and often underestimated problem.

While DTS struggles with its complexities, NLP has witnessed significant advancements with the rise of Large Language Models (LLMs) and reasoning methodologies. However, even considering the robust problem-solving skills of these LLMs and the challenges posed by DTS, reasoning strategies are rarely applied in the DTS area. This is because DTS has largely been treated as a lightweight subtask in NLP. Nonetheless, with the rise of AI-driven

chat services, the demand for more advanced DTS modules is growing. LLMs with reasoning capabilities are well-suited to meet this need, making LLM-based DTS a viable solution.

To establish clear criteria for topic shifts and simplify complex subtasks, we propose Def-DTS: a deductive reasoning approach for open-domain dialogue topic segmentation using LLM-based multi-step reasoning. Def-DTS employs structured prompting to guide LLMs through bidirectional context summarization, utterance intent classification, and deductive topic shift detection at the utterance level, with an emphasis on domain-agnostic intent classification.

To evaluate Def-DTS, we test it on three dialogue datasets spanning open-domain and task-oriented settings across three key metrics. Our method consistently outperforms traditional and state-of-the-art unsupervised, supervised, and prompt-based techniques by a significant margin. Ablation studies confirm that each subtask enhances overall performance, with intermediate intent classification particularly improving true-positive detection. Finally, we explore LLM-based auto-labeling for DTS. Our contributions are fourfold:

- We introduce LLM reasoning techniques to DTS for the first time, consolidating insights from previous methodologies into a deductive and cohesive prompt design.
- We propose a reformulation of DTS as an utterance-level intent classification, enabling flexible and task-agnostic prompting.
- Our method empirically demonstrates superior performance across nearly all comparative baselines, underscoring the efficacy of prompt engineering in DTS.
- Through an in-depth analysis of our approach's reasoning results, we shed light on the challenges LLM reasoning faces in DTS and discuss the possibility of using LLM as a DTS auto-labeler.

2 Related Works

2.1 Dialogue Topic Segmentation

Dialogue topic segmentation divides dialogues into coherent topic units. Due to limited annotated datasets, researchers have largely relied on unsupervised approaches despite their complexity (Xing and Carenini, 2021). Early methods like TextTiling (Hearst, 1997) detected topic shifts via lexical similarity, later improved with embedding-based methods (Song et al., 2016).

Recent research emphasizes topical coherence and similarity scoring. CSM (Xing and Carenini, 2021) leverages BERT-based coherence, while Dial-START (Gao et al., 2023) incorporates SimCSE (Gao et al., 2021) for topic similarity. SumSeg (Artemiev et al., 2024) extracts key information via summaries and applies smoothing to handle topic variations. UR-DTS (Hou et al., 2024) enhances segmentation by rewriting utterances to recover missing references. Despite these advances, data scarcity and performance limitations persist.

To address this, SuperDialSeg (Jiang et al., 2023) introduces a supervised approach using large-scale DGDS datasets (Feng et al., 2020, 2021). However, the available datasets remain limited for open-domain conversations.

LLMs are also influencing DTS. S3-DST (Das et al., 2024) applies structured prompting for dialogue state tracking and segmentation, but lacks general applicability to diverse DTS settings.

2.2 Reasoning Strategy at LLM Inference

The advancement of LLMs (Brown et al., 2020) has led to research on the integration of System 2 reasoning, including in-context learning (Brown et al., 2020) and chain-of-thought prompting (Wei et al., 2022). These techniques enable LLMs to tackle complex tasks, such as symbolic mathematics (Yang et al., 2024), retrieval-augmented generation (Lewis et al., 2020), and data generation (Adler et al., 2024). Studies show that intermediate reasoning steps significantly enhance performance in areas like multi-hop reasoning (Wang et al., 2023) and math problem-solving (Imani et al., 2023). Building on this, we integrate LLM-based reasoning into DTS, leveraging its potential to improve segmentation accuracy in this inherently complex task.

3 Def-DTS

3.1 Overall Flow

Our method (Figure 2) applies multiple subtasks to each utterance using a structured prompt format. The prompt template (Figure 2b) includes four parts: valid label list, task description, output format, and input. As shown in Figure 2c, it comprises three main subtasks: (i) bidirectional context ex-

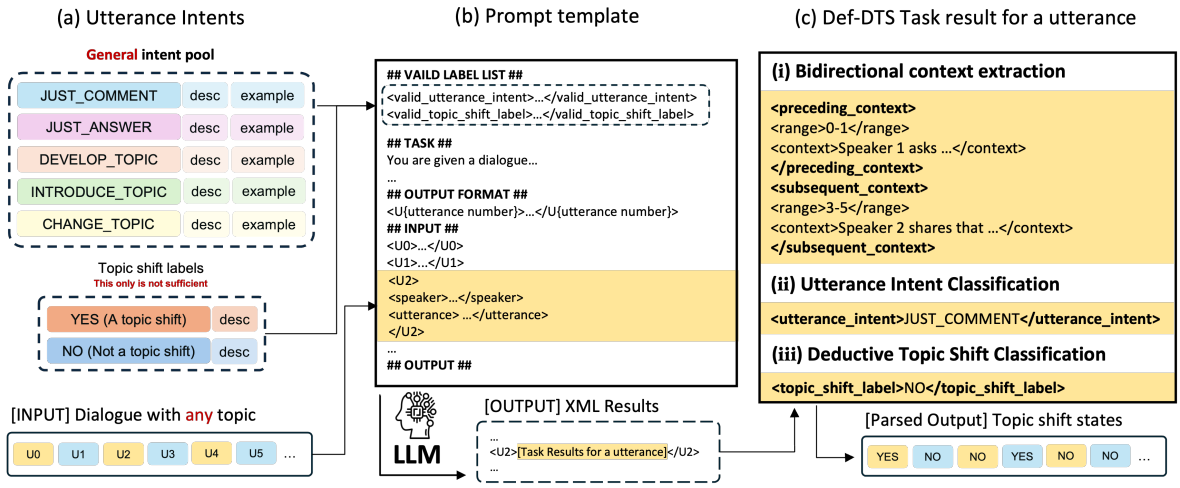

Figure 2: Prompt configuration and overall flow of our method: Def-DTS. (a) We utilize general intent list including the intent-specific examples to enable domain-agnostic categorization. (b) We employ the xml structured input-output format to stably provide the dialogue. (c) We instruct LLM to process the multi-step reasoning for each utterance in a inference.

Algorithm 1 Def-DTS

Require: IntentPool $X = \{x_1, x_2, \ldots, x_m\}$
Require: StructuredDialogue $D = \{d_i\}_{i=1}^{N}$, where $d_i = \{u_i, s_i\}$ (utterance u_i and speaker s_i)
Ensure: Results $R = \{r_i\}_{i=1}^{N}$, where $r_i = \{P_i, Q_i, X_i, T_i\}$ (bidirectional context (P_i, Q_i), intent X_i, topic shift T_i)

$R \leftarrow \emptyset$, $S \leftarrow$ StructureDialogue(D)
for $i \leftarrow 1$ to N **do**
  $P_i \leftarrow$ ExtractContext$(D, max(1, i-2), i)$
  $Q_i \leftarrow$ ExtractContext$(D, i+1, min(i+3, N))$
  $X_i \leftarrow$ ClassifyIntent$(D[i], P_i, Q_i, X)$
  $T_i \leftarrow$ ClassifyTopicShift(X_i)
  $r_i \leftarrow r_i \cup \{(P_i, Q_i, X_i, T_i)\}$
end for
return R

function ExtractContext$(D, start, end)$
  $context \leftarrow$ Summarize$(D[start : end])$
  return "U_{start}-U_{end}", $context$
end function

function ClassifyIntent$(utterance, P, Q, X)$
  return $classified\_intent$
end function

function ClassifyTopicShift(x)
  return $x \in$ {"introduce_topic", "change_topic"} ? "YES" : "NO"
end function

function Summarize$(context)$
  return $summarized\_context$
end function

traction, (ii) utterance intent classification, and (iii) deductive topic shift classification. Each subtask is executed deductively, with further details in Algorithm 1.

3.2 Structured Format

Inspired by previous DST research (Das et al., 2024), we use an XML-based structured prompt to standardize LLM output, enhancing parsing efficiency and minimizing post-processing. The input template (Figure 2b) organizes utterances in <Ux> elements, each containing speaker and utterance content. The output (Figure 2c) follows the same structured format, ensuring consistent labeling. A complete template is available in Appendix A.1.

3.3 Bidirectional Context Extraction

In the first stage of Def-DTS, we instruct the LLM to summarize both the preceding and subsequent dialogues for each utterance. Considering bidirectional context is commonly used in many methods such as the BERT architecture (Devlin et al., 2019) and frequently employed in unsupervised settings (Gao et al., 2023; Hou et al., 2024), this strategy has proven effective for understanding context in dialogue. Although Das et al., 2024 only extracts the preceding context to prevent contextual forgetting, we improve upon this by extracting both the preceding and subsequent contexts to enable context-aware dialogue topic segmentation and prevent contextual forgetting.

We opted for a fixed window size to ensure ap-

plicability in unsupervised and real-world environments without predefined segments, as it offers a robust approach when segment boundaries are not available. As shown in Figure 2c(i), we instruct the model to summarize no more than two preceding turns using the <preceding_context> tag and no more than three subsequent turns using the <subsequent_context> tag. This window size of -2:-1 and 1:3 was chosen to balance context informativeness and token efficiency, an approach that is similar to the method used in (Gao et al., 2023). By summarizing each context range, we aim to conserve tokens while maintaining nuanced topic relationships.

Each element is composed of <range> . . . </range> for defining the summary scope and <context>. . . </context> for the actual summary. Summarizing in this manner ensures that we capture relevant dialogue while keeping the context concise. Our experiments reveal that this approach effectively facilitates the observation of nuanced topic relationships.

3.4 Utterance Intent Classification

Intent	Description
JUST COMMENT	Commenting on the preceding context without any asking. Not a topic shift
JUST ANSWER	Answering preceding utterance. Not a topic shift
DEVELOP TOPIC	Developing the conversation to similar and inclusive sub-topics. Not a topic shift
INTRODUCE TOPIC	Introducing a relevant but different topic. A topic shift
CHANGE TOPIC	Completely changing the topic. A topic shift

Table 1: Utterance intent list for open-domain dialogue. Using this list, we categorize the utterance and deductively classify the topic shift label.

As there are issues with the ambiguity of the DTS datasets, simply providing descriptions for topic classification is insufficient to convey a precise definition of topic changes. Therefore, we suggest that classifying the utterance using a well-defined and distinct list of labels, similar to intent classification, would be beneficial for DTS. Although most of intent classification tasks have been conducted within the context of Task Oriented Dialogue (TOD) (Liu and Lane, 2016; Chen and Luo, 2023), TOD typically involves datasets that are specific to a particular domain, making generalization to other domains or more open-ended conversations challenging.

As a way to address this issue, we find that Xie et al., 2021 identified five patterns of conversational responses that are utilized in their annotation guidelines. These patterns reflect the natural characteristics of the utterances, enabling intent classification for various forms of dialogue. Detailed intent patterns and descriptions are in Table 1.

Inspired by this research, we instruct model to detect topic change through utterance intent classification. Specifically, after the bidirectional context extraction, as shown in Figure 2c(ii), model classifies the utterance into an intent of the predefined general intent pool (upper box of Figure 2a), considering previously generated bidirectional context.

Additionally, we enhance the model's understanding of the intents by providing example dialogue for each intent, as shown in the General intent pool in Figure 2a. The intent-specific examples provide the model with a helpful guideline detecting topic changes, allowing it to derive results in the subsequent subtask, deductive topic shift. Also in order to prove the significance of our intent labels through statistical frequency analysis from the traditional text segmentation method, which can be found in Section 5.4. We observed significant performance improvements across various dialogue settings using this technique, enabling a detailed analysis of each intent. The process of selecting examples in Appendix A.4.

3.5 Deductive Topic Shift Classification

Finally, the model predicts whether the topic is changing based on the deductive guidelines from the previous intent classification result, as shown in Figure 2c(iii). This task is processed in an enforced manner, and the model deductively outputs the predetermined label based on the intent classification results of the previous step, according to the tail of description (i.e. Not a topic shift or A topic shift) for each intent described in Table 1.

We explicitly instruct the model to process this step for two reasons. First, it simplifies the parsing process, making the task easy to handle. Second, it ensures that the model explicitly outputs the result of primary goal, allowing it to stay focused on the main objective(DTS), while working through the subtasks.

Dataset	# Sample	Utterance per Dial.			Segment per Dial.	
		avg	min	max	avg	avg.len
TIAGE	100	15.6	14	16	4.2	3.8
SuperDialseg	1322	12.1	7	19	4.0	3.0
Dialseg711	711	26.2	7	47	4.9	5.4

Table 2: Statistics of datasets for dialogue topic segmentation.

4 Experiments

4.1 Datasets

We evaluated our method on three datasets: TIAGE, SuperDialseg, Dialseg711 to verify the performance of our method in both open-domain and task-oriented settings. Dataset statistics are presented in Table 2.

TIAGE (Xie et al., 2021) is the only publicly available dataset with topically segmented daily conversations, derived from PersonaChat (Zhang et al., 2018) and designed to model topic shifts in open-domain dialogue. SuperDialseg (Jiang et al., 2023) is a large-scale dialogue segmentation dataset based on document-grounded corpora, offering a framework for identifying segmentation points in document-based dialogues. Dialseg711 (Xu et al., 2021b) is a real-world dialogue dataset auto-labeled from MultiWOZ (Budzianowski et al., 2018) and Stanford Dialog Dataset (Eric et al., 2017), created by joining dialogues with distinct topics, resulting in clear topical differences and low coherency at segment boundaries due to its synthetic nature.

4.2 Evaluation Metrics

As early studies (Xing and Carenini, 2021; Jiang et al., 2023; Artemiev et al., 2024) did, we leverage P_k error (Beeferman et al., 1997), WindowDiff (WD) error (Pevzner and Hearst, 2002), and the f1 score. The P_k error is calculated by counting the existence of a misallocated segment with a sliding window of predictions. The WD error is calculated by comparing the number of boundaries within the sliding window of gold labels and predictions. Note that the lower P_k or WD means the higher performance.

4.3 Comparison Methods

We propose a sophisticated DTS method based on prompt engineering way and compare its performance with various unsupervised, supervised, and LLM-based methodologies. First, we compare our method with a random baseline that arbitrarily assigns segment boundaries based on a randomly chosen number of segments. Next, we compare it to notable unsupervised learning methods based on TextTiling like Coherence Scoring Model (CSM) (Xing and Carenini, 2021), DialSTART (Gao et al., 2023), SumSeg (Artemiev et al., 2024). We also compare supervised learning methodologies. We selected the basic BERT model(Devlin et al., 2019), the advanced and high-performing RoBERTa model(Liu, 2019), RetroTS-T5(Xie et al., 2021) system for our comparative analysis. Finally, we compare our approach with recently introduced LLM-based methods. We applied these methods using gpt-4o.[1] We selected the PlainText prompts performed in SuperDialseg(Jiang et al., 2023) and the prompt of S3-DST(Das et al., 2024). The details of the methodology utilized for each of the actual comparisons are discussed in the Appendix D.1.

4.4 Experimental Results

The experimental results are presented in Table 3. Def-DTS consistently demonstrated superior performance among LLM-based methods and achieved state-of-the-art results on the TIAGE and Dialseg711 datasets, which are closely aligned with our objective of analyzing general open-domain dialogues. In contrast, other LLM-based methods showed lower performance not only compared to supervised learning and some unsupervised methods, indicating that simply using an LLM is not a guarantee of success.

TIAGE Compared to S3-DST_{uttr}, the recent LLM-based method, our method achieved reductions of more than 0.2 in both P_k and WD errors, along with an impressive increase of more than 0.4 in the F1 score. Furthermore, Def-DTS outperformed even the supervised approaches in TIAGE, surpassing them in all metrics by over 10%, thus highlighting the effectiveness of our approach in achieving high performance across various dialogue environments without additional training, even in domain-agnostic settings.

SuperDialseg Our method outperformed all unsupervised methods. Although it showed lower performance compared to models trained using supervised learning, it achieved the best results among unsupervised methods that use prompt-based techniques.

[1] https://platform.openai.com/docs/models/gpt-4o

Method	TIAGE			SuperDialseg			Dialseg711		
	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑
Unsupervised Learning Methods									
Random	0.526	0.664	0.237	0.494	0.649	0.266	0.533	0.714	0.204
TextTiling	0.469	0.488	0.204	0.441	0.453	0.388	0.470	0.493	0.245
TextTiling+Glove	0.486	0.511	0.236	0.519	0.524	0.353	0.399	0.438	0.436
CSM	**0.400**	**0.420**	**0.427**	0.462	0.467	0.381	0.278	0.302	0.610
DialSTART	0.482	0.528	0.378	**0.373**	**0.412**	**0.627**	**0.179**	**0.198**	**0.733**
SumSeg	0.482	0.496	0.075	0.479	0.485	0.119	0.477	0.483	0.070
Supervised Learning Methods									
BERT	0.418	0.435	0.124	0.214	0.225	0.725	-	-	-
RoBERTa	**0.265**	**0.287**	0.572	**0.185**	**0.192**	**0.784**	-	-	-
RetroTS-T5	0.280	0.317	**0.576**	0.227	0.237	0.733	-	-	-
LLM-based Methods									
Plain Text	0.445	0.485	0.185	0.412	0.427	0.048	0.333	0.353	0.010
S3-DST_{uttr}	0.439	0.498	0.265	0.442	0.469	0.404	0.087	0.109	0.790
Def-DTS (Ours)	**0.232**	**0.256**	**0.699**	**0.315**	**0.324**	**0.686**	**0.015**	**0.018**	**0.979**

Table 3: Performances on three datasets. Due to absence of train and validation split for Dialseg711 dataset, There are no report in dialseg711's supervised learning part. The best results for each method group are highlighted in **bold**. The best performances around all method are indicated as **red** colored text.

Dialseg711 Our approach also delivered superior performance. Notably, it surpassed the strongest LLM-based approach, S3-DST_{uttr}, underscoring the general applicability and robustness of our method.

Overall, these consistent improvements across all tested datasets confirm the robust effectiveness of our approach for diverse open-domain dialogue scenarios, demonstrating that it not only excels in unsupervised settings, but also surpasses previously leading LLM-based methods. This further establishes potential of our method for delivering high performance even under challenging, domain-agnostic conditions.

5 Analysis and Discussion

5.1 Ablation Study

To assess the contributions of each part of our approach, we performed an ablation study. The results are shown in Table 4. In the **w/o all** components case, the model is instructed to detect topic shifts without context extraction or intent classification. In the **w/o intent** case, the model detects topic shifts after context extraction for each utterance. We observe that w/o intent performs worse than w/o all components, indicating that relying solely on dialogue context for topic shift prediction does not yield optimal performance. In the **w/o examples** case, this is essentially Def-DTS but without examples for intent. w/o examples performed better than w/o intent, showing that processing context into intent before using it for topic shift prediction provides a significant advantage. In the **w/o context** case, the model is instructed to detect topic shifts after intent classification for each utterance, which is the opposite of the w/o intent case. This result demonstrates that intent classification, when supported by appropriate examples, has a significant impact on topic shift prediction for individual utterances. In the **w/o bidirectional** context case, the subsequent context is not considered at the context

Method	TIAGE		
	P_k↓	WD↓	F1↑
w/o all	0.295	0.333	0.605
w/o intent	0.316	0.342	0.524
w/o examples	0.287	0.308	0.617
w/o context	0.263	0.296	0.682
w/o bidirectional	0.269	0.301	0.659
Def-DTS	**0.232**	**0.256**	**0.699**

Table 4: Ablation study.

extraction step. Compared to both the Def-DTS and w/o context case, this case showed lower performance, highlighting that considering bidirectional context is crucial for intent classification and topic shift detection. In summary, each module of **Def-DTS** contributes to performance improvement, and when all modules are applied, they work synergistically to yield a substantial increase in performance.

5.2 Intent Classification Accuracy

Intent	TP	FP	TN	FN	Acc
JUST_COMMENT	0	1	498	35	0.93
JUST_ANSWER	0	1	456	23	0.95
DEVELOP_TOPIC	0	0	119	47	0.71
INTRODUCE_TOPIC	189	68	0	0	0.73
CHANGE_TOPIC	21	6	0	0	0.78

Table 5: Intent-level confusion matrix for TIAGE benchmark.

As our method deduces topic shifts directly from utterance intent, analyzing the intent classification results is crucial. We examine the confusion matrix (Table 5) to identify the utterance types that the model struggles with. Since there is no ground truth about intent and only topic shift labels are provided, correctness is determined by whether the predicted intent aligns with the topic shift label. For instance, if an utterance is a topic shift but classified as JUST_COMMENT for intent and NO for topic shift, it counts as a False Negative (FN).

Most results performed well in topic shift classification, except for two cases: positives in JUST_COMMENT and JUST_ANSWER. These findings indicate that the model's primary challenge lies in distinguishing subtle topic differences as actual shifts, a more significant factor in performance degradation than other utterance types. Analysis of additional datasets (SuperDialseg, Dialseg711) is in Appendix B.

5.3 Intent Level Comparison

We compared the performance of various approaches across different intent categories predicted by our method, as shown in Figure 3. In (a) MATCHED INTENT, for utterances without a topic shift, other methodologies achieved approximately 80-85% accuracy when our method was also correct. However, for utterances that involve a topic shift, the accuracy of other methods dropped to around 20-50% for the correct case of our method. In (b) MISMATCHED CASE, for

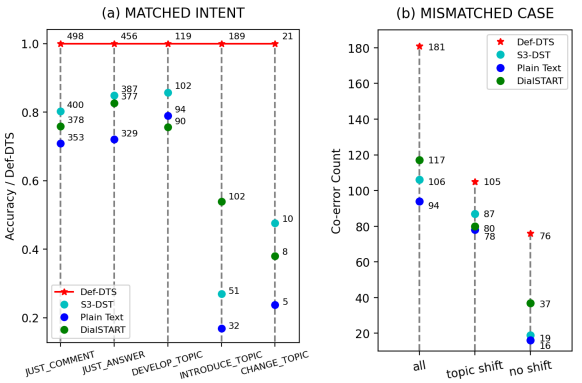

Figure 3: (a) MATCHED INTENT indicates the accuracy of the other methodologies for grouped utterances by our intent classification process only in the true cases of our method. (b) MISMATCHED CASE indicates the co-error count of the other methodologies with our methods for only in the false cases of our method.

utterances without a topic shift, other methods correctly classified 50% of the cases where our method was incorrect. However, for utterances with a topic shift, other methods failed to classify 80% of the cases that our method also missed. This demonstrates that detecting utterances with actual topic shifts is considerably more challenging than detecting those without topic shifts. Our method outperforms the other methodologies by roughly 20% in cases without topic shifts, and by over 40% in cases with topic shifts. In summary, while our approach improves accuracy across all cases, it shows even greater improvement when handling utterances with topic shifts.

5.4 Linguistic Test for Intent Labels

To demonstrate the impact of intent labels on topic shifts, we adopted methods from statistical linguistics. Traditional text segmentation uses pauses, cue words, and referential noun phrases to identify boundaries (Passonneau and Litman, 1997). Galley et al., 2003 found a significant correlation between cue phrases and topic segmentation. Building on this, we hypothesized that cue words in labels like "introduce topic" and "change topic" correlate with their overall frequency in the data. A χ^2 test yielded $\chi^2(32) = 76.2263$, $p < 0.001$, confirming a significant relationship. This validates our labels as linguistically rich markers of discourse boundaries and provides a criterion for selecting topic-shift data in new datasets.

Model	TIAGE			SuperDialseg			Dialseg711		
	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑
Plain Text + Llama	0.472	0.515	0.215	0.492	0.495	0.026	0.350	0.373	0.032
Plain Text + Qwen	0.495	0.533	0.162	0.485	0.487	0.059	0.422	0.434	0.012
S3-DST$_{\text{uttr}}$ + Llama	0.456	0.474	0.143	0.490	0.512	0.072	0.158	0.190	0.553
S3-DST$_{\text{uttr}}$ + Qwen	-	-	-	-	-	-	-	-	-
Def-DTS + Llama	**0.307**	**0.339**	**0.552**	**0.384**	**0.385**	**0.432**	**0.029**	**0.039**	**0.941**
Def-DTS + Qwen	0.327	0.345	0.530	0.433	0.434	0.171	0.102	0.208	0.729

Table 6: Performances on the local LLMs. We employed Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct for Llama and Qwen, respectively. The case of best performance across all method are highlighted in **bold**.

5.5 Performance Comparison for Local LLMs

We conducted experiments using local LLMs instead of GPT-4, specifically using Lama 3.1 and Qwen 2.5, with the exact model names listed in Table 6. For the experiments, we tested three prompts: Plain Text, S3-DST$_{\text{uttr}}$, and Def-DTS. To ensure efficiency, we randomly sampled 100 examples from each dataset. Def-DTS achieved the highest performance in all datasets in this experiment. Experiments with LLMs of various sizes are presented in the Appendix C. In the case of Qwen, formatting errors were observed in all datasets. Although plain is an unstructured method, it did not have errors, but its performance remained comparatively lower. These results demonstrate that Def-DTS maintains high accuracy in different LLMs.

5.6 Discussion for Possibility of Auto-Labeling

As various auto-labeling methodologies have been proposed to date, we assess prompt engineering could potentially serve as a viable auto-labeling methodology. We conducted a preliminary experiment to assess the feasibility of using prompt engineering for DTS. We compared the segment labels generated by GPT-4 for Def-DTS with the correct labels using Cohen's Kappa score. The results showed Kappa scores of 0.485 for TIAGE, 0.429 for SuperDialseg, and 0.975 for Dialseg711, indicating moderate agreement for TIAGE and SuperDialseg, and almost perfect agreement for Dialseg711. Notably, our labeling result for TIAGE exceeded the 0.479 agreement score observed between actual human annotators. While improvements are needed given the moderate agreement, these findings suggest that our approach can still function as a minimal annotator.

6 Conclusion

Previous approaches to DTS have been constrained by several challenges, including data shortage, ambiguity of segment labeling, and increasingly complex model architectures. Concurrently, The promising approach of reasoning with LLMs has yet to be explored in the context of DTS. To address these issues, we propose Def-DTS, that leverages LLMs in conjunction with sophisticated reasoning strategies. Def-DTS incorporates bidirectional context extraction, a crucial component in previous research, along with the novel task of utterance intent classification. This approach demonstrates significant performance improvements in both the open-domain dialogue setting and the task-oriented dialogue setting. Its efficacy across diverse datasets is enhanced through the provision of dataset-specific examples in the utterance intent classification task, enabling adaptable performance in varied dialogue contexts. Through its primary findings and diverse analysis, we demonstrate the efficacy of LLM-reasoning as a promising approach to DTS. It not only highlights the potential of our method, but it also statistically delineates the challenges to be addressed in future research. In subsequent investigations, we intend to explore the feasibility of automated labeling for DTS and examine the potential integration of DTS with other NLP downstream tasks through LLM reasoning.

7 Limitations

Firstly, though we have demonstrated the significance of the current intent labels by statistical linguistic experiments, we cannot entirely rule out the possibility that more suitable intent labels ex-

ist. Additionally, as we provide a first approach for selecting representative examples, a more in-depth exploration of methodologies for selecting optimal examples remains a future step of this research. In order to improve the quality of intents and examples in a variety of dialogue settings, the fundamental problem must be addressed first, namely the provision of quality datasets for DTS. Dialogue should include thorough labeling criteria and realistic dialogue domains to address our limitations. However, human labeling is not only still expensive but also carries the risk of inconsistent or ambiguous labeling. We believe that automated labeling using LLMs with a sophisticated guideline will play a crucial role in creating a more sustainable and reliable DTS environment.

References

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. 2024. Leveraging summarization for unsupervised dialogue topic segmentation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4697–4704, Mexico City, Mexico. Association for Computational Linguistics.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *Second Conference on Empirical Methods in Natural Language Processing*.

Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering*, 22(1):41–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Yan Chen and Zhenghang Luo. 2023. Pre-trained joint model for intent classification and slot filling with semantic feature fusion. *Sensors*, 23(5).

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2024. S3-DST: Structured open-domain dialogue segmentation and state tracking in the era of LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14996–15014, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. 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*, SIGIR '23, page 2481–2485, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10749–10757.

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

Xia Hou, Qifeng Li, and Tongliang Li. 2024. An unsupervised dialogue topic segmentation model based on utterance rewriting. *arXiv preprint arXiv:2409.07672*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. SuperDialseg: A large-scale dataset for supervised dialogue segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4086–4101, Singapore. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8360–8367.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7272–7282, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xinbei Ma, Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2024. Multi-turn dialogue comprehension from a topic-aware perspective. *Neurocomputing*, 578:127385.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1121–1130, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue session segmentation by embedding-enhanced texttiling. *Interspeech 2016*, pages 2706–2710.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2717–2731, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Singapore and Online. Association for Computational Linguistics.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021a. Discovering dialog structure graph for coherent dialog generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1726–1739, Online. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021b. Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14176–14184.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2018. Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1575–1584, New Orleans, Louisiana. Association for Computational Linguistics.

Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022. A slot is not built in one utterance: Spoken language dialogs with sub-slots. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 309–321, Dublin, Ireland. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

A Prompts

A.1 Prompt Template

Prompt Template
<valid_utterance_intent> <item> <name>JUST_COMMENT</name> <desc>Commenting on the preceding context without any asking. Not a topic shift</desc> <example> <speaker1>My dad works for the New York Times.</speaker1> <speaker2>Oh wow! You know, I dabble in photography; maybe you can introduce us sometime.</speaker2> <speaker1>Photography is the greatest art out there. (not a topic shift)</speaker1> </example> </item> <item> <name>JUST_ANSWER</name> <desc>Answering preceding utterance. Not a topic shift</desc> <example> <speaker1>Do you teach cooking? </speaker1> <speaker2>No, since I'm a native of Mexico, I teach Spanish. (not a topic shift)</speaker2> </example> </item> <item> <name>DEVELOP_TOPIC</name> <desc>Developing the conversation to similar and inclusive sub-topics. Not a topic shift</desc> <example> <speaker1>Pets are cute!</speaker1> <speaker2>I heard that Huskies are difficult dogs to take care of. (not a topic shift)</speaker2> </example> </item> <item> <name>INTRODUCE_TOPIC</name> <desc>Introducing a relevant but different topic. A topic shift</desc> <example> <speaker1>You are an artist? What kind of art, I do American Indian stuff.</speaker1> <speaker2> I love to eat too, sometimes too much. (a topic shift)</speaker2> </example> </item> <item> <name>CHANGE_TOPIC</name> <desc>Completely changing the topic. A topic shift</desc> <example> <speaker1>What do you do for fun?</speaker1> <speaker2>I drive trucks so me and my buds go truckin in the mud.</speaker2> <speaker1>Must be fun! My version of that's running around a library!</speaker1> <speaker2>That's cool! I love that too. Do you have a favourite animal? Chickens are my favourite. I love them. (topic shift)</speaker2> </example> </item> </valid_utterance_intent> <valid_topic_shift_label>

Prompt Template

<item>
<name>YES</name>
<desc>The current utterance has **weak OR no topical** relation to the preceding conversation context OR is the first utterance in the conversation, marking the beginning of a new dialogue segment.</desc>
</item>
<item>
<name>NO</name>
<desc>The current utterance has **relevant OR equal** topic to the preceding conversation context.</desc>
</item>
</valid_topic_shift_label>

TASK
You are given a dialogue starting with U. From utterance number 0, you have to answer the following sub-tasks for each utterance.
1. Summarize the preceding and subsequent context in <=3 sentences seperately
The range of the context should be previous or next 1-3 utterances except for the case of the first or last utterance.
For example, given current utterance number is 2, preceding range is 0-1, subsequent range is 3-5.
2. Output the utterance_intent
Use the list <valid_utterance_intent> ... </valid_utterance_intent> to categorize utterance.
Consider topical difference between preceding and subsequent context.
3. Output the topic_shift_label
Use the list <valid_topic_shift_label> ... </valid_topic_shift_label>.

OUTPUT FORMAT
<U{utterance number}>
<preceding_context>
<range>{range of utterances referred in context}</range>
<context>{context of the previous 1-3 utterances}</context>
</preceding_context>
<subsequent_context>
<range>{range of utterances referred in context}</range>
<context>{context of the next 1-3 utterances}</context>
</subsequent_context>
<utterance_intent>{valid utterance intent}</utterance_intent>
<topic_shift_label>{valid topic shift label}</topic_shift_label>
</U{utterance number}>

INPUT
{XML-structured dialogue}

OUTPUT

Table 7: We provide prompt template for main dataset: TIAGE. Each dataset has different characteristics to other datasets, so we modified intent pool from original template for each dataset.

A.2 Intent Labels for other datasets

Intent	Description
DIFFERENT QUESTION	Questioning about something that is not similar or topically different to preciding context. A topic shift
RELEVANT QUESTION	Questioning about something that is similar or topically coherent to preceding context. Not a topic shift
ANSWERING	Answering preceding utterance. Not a topic shift
ADDITIONAL COMMENT	An additional comment from the same speaker in addition to a previous utterance. Not a topic shift

Table 8: Utterance intent list for SuperDialseg dataset.

For the Dialseg711 dataset, we delete an intent named INTRODUCE_TOPIC from the original intent list. For the SuperDialseg dataset, the topic shift occurs when the utterance refers to a different document to previous utterance. As shown in Table 8, we completely change the intent from the original to fit to document grounded dialogue setting based topic transition.

A.3 Modification of TIAGE Response Patterns

Scenario	Topic Shift
Commenting on the previous context	No
Question answering	No
Developing the conversation to sub-topics	No
Introducing a relevant but different topic	Yes
Completely changing the topic	Yes

Table 9: Intent labels in different dialog scenarios.

TIAGE (Xie et al., 2021) was originally designed for real-time topic shift detection without using a subsequent context. Consequently, its conversation response pattern list cannot be applied as is for our full-dialogue segmentation task. To address this, we propose two methods. First, instead of classifying each utterance with only its immediate context, we retrieve both preceding and subsequent context to inform our decisions. This bidirectional view captures how an utterance relates to what was said before and what follows, enabling more precise intent classification. Second, we adapt TIAGE's original response pattern list to our classification objectives by reorganizing patterns (e.g., Asking) and refining the description of each intent (e.g., Relevant, Inclusive). This transformation ensures a more granular detection of whether an utterance continues the topic or shifts in a subtle way. By employing these enhancements, we achieve significantly better performance in segmenting topics across entire dialogues, surpassing the results obtained by using TIAGE's list unaltered.

A.4 Construction of Intent Examples

For the TIAGE dataset, we used the examples directly from their paper. For other datasets, we randomly selected parts of the conversation from the Train split that adhere to the following rules:

- Select 2–3 consecutive utterances for each example.
- Ensure that the final utterance in the example corresponds to the target utterance intent.
- Extract all the examples from a single dialogue.
- Keep the utterance lengths concise (within 100 characters).

This domain-independent guideline can be an initial pathfinder to tailor the best examples in dialogue, though it may not be perfect.

B Analysis on other datasets

B.1 Ablation Study

Method	SuperDialseg			Dialseg711		
	$P_k\downarrow$	WD↓	F1↑	$P_k\downarrow$	WD↓	F1↑
w/o all	0.378	0.382	0.467	0.007210	0.009416	0.987245
w/o intent	0.363	0.364	0.448	0.093486	0.127211	0.701330
w/o examples	0.338	0.341	0.646	0.005826	0.012322	0.984733
w/o context	0.327	0.331	0.635	**0.005800**	**0.008464**	**0.989770**
Def-DTS	**0.317**	**0.322**	**0.674**	0.009024	0.013738	0.982143

Table 10: Ablation study for identifying effectiveness of each subtask within our method.

As shown in Table 10, we found that the absence of some module leads to performance degradation. For efficient evaluation, we used 100 randomly sampled dialogue for the ablation study for SuperDialseg and Dialseg711 each. This data is the same as we used in Section 5.5. Especially on the SuperDialseg dataset, we observed a consistent improvement by adding any subtask. However, with respect to Dialseg711 dataset, the existence of context extraction module is crucial to performance improvement. Even the w/o case surpasses our full-attached method. we conjecture that the issue is due to over-concentration for local context, same as the result of ablation study for TIAGE, moreover, on the case of dialseg711 having clear topic shift signal, just predicting label is enough to solve the problem. After all, our intent classification module elevates the performance across all datasets.

B.2 Intent Classification Accuracy

SuperDialseg					
Intent	TP	FP	TN	FN	Acc
DIFFERENT_QUESTION	2456	688	83	192	0.74
RELEVANT_QUESTION	1	0	811	989	0.45
ANSWERING	0	2	7819	168	0.98
ADDITIONAL_COMMENT	0	0	1264	211	0.86
Dialseg711					
Intent	TP	FP	TN	FN	Acc
JUST_COMMENT	0	6	5067	14	0.996
JUST_ANSWER	0	6	7675	8	0.998
DEVELOP_TOPIC	0	3	2359	13	0.993
CHANGE_TOPIC	2708	66	0	0	0.976

Table 11: Intent-level confusion matrix for other datasets.

We conducted a detailed accuracy analysis for the other datasets and the results are presented in Table 11.

For SuperDialseg, as shown in Table 8, four new intent pools were applied. ADDITIONAL_COMMENT, ANSWERING, and RELEVANT_QUESTION are classified as non-topic shift cases, whereas the case of DIFFERENT_QUESTION is classified as a topic shift case. However, the instruction following was not well executed for the case of DIFFERENT_QUESTION. For the case of RELEVANT_QUESTION, the following instruction was well executed with one exception, but its accuracy was relatively low. The difference in explanations between the case of RELEVANT_QUESTION and DIFFERENT_QUESTION could be linked to the actual dataset characteristics and topic changes. In contrast, the cases of ANSWERING and ADDITIONAL_COMMENT showed significantly high classification accuracy. This comparison suggest that improving Question-type intents will lead to an overall improvement in performance.

For Dialseg711, overall accuracy was higher compared to other datasets where the following deductive instruction not executed for less than 1% of utterances. For the results of the three intents, excluding the case of CHANGE_TOPIC, 18% of DEVELOP_TOPIC, 30% of JUST_COMMENT, and 42% of JUST_ANSWER cases among all false cases were misclassified due to errors in instruction following. It is believed that this issue can be resolved through additional instructions or prompt modifications for instruction following.

C Additional experiments on local LLMs

Def-DTS leverages LLM reasoning capabilities, so model size significantly affects performance. Table 6 shows that S3-DST on Qwen 70B had formatting errors, which discouraged us from testing smaller LLMs initially. However, considering the growing capabilities and applications of sLLMs, we conducted additional experiments on: Llama 8B, Qwen 7B, Qwen 32B.

The experimental result is shown in table 12. Although Def-DTS struggled with smaller models and dialseg711, it showed greater improvements with larger models by leveraging LLM reasoning capability. However, we acknowledge that applying Def-DTS to smaller LLMs would require adjustments such as additional parameter modification.

D Details for Experiment

D.1 Details for Implementation

The model used for our experiments is gpt-4o for closed LLM, Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct for open-source LLM. At first, we considered two closed models: gpt-4o and Claude-3.5-sonnet. But Claude was excluded due to poor accuracy compared to gpt-4o at a preliminary evaluations and there are no prior studies applied their method to the claude family. For the inference of open-source LLMs, we utilized a computational infrastructure consisting of 4*NVIDIA A100 80GB GPU. We conducted our experiments without employing any model-specific tuning or quantization techniques, thus maintaining the original model architecture and parameters. We kept the hyperparameters initially stated, except for the temperature that we set to 0 for reproducibility of our experiments.

D.2 Details for Reproduce

For the SuperDialseg paper(Jiang et al., 2023), we obtained experimental results for the Random, TextTiling, TextTiling+Glove, CSM, BERT, RoBERTa, and RetroTS-T5 methods.

We reproduced the experimental result for PlainText(Jiang et al., 2023), S3-DST(Das et al., 2024), DialSTART(Gao et al., 2023), and SumSeg(Artemiev et al., 2024), which either lacked F1 scores or did not perform experiments on certain datasets.

During reproduction, we maintained all settings, including seeds, without any parameter modifications. However, We observed results that differed

Method	Model	TIAGE				SuperDialseg				Dialseg711			
		P_k↓	WD↓	F1↑	Error	P_k↓	WD↓	F1↑	Error	P_k↓	WD↓	F1↑	Error
Plain Text	Llama 8B	0.529	0.604	0.303	1	0.497	0.504	0.036	0	0.350	0.373	0.032	0
	Qwen 7B	0.509	0.563	0.249	2	0.517	0.522	0.132	1	0.486	0.513	0.069	0
	Qwen 32B	0.476	0.515	0.221	0	0.466	0.471	0.083	0	0.391	0.415	0.015	0
S3-DST$_{\text{uttr}}$	Llama 8B	0.460	0.460	0.018	0	0.472	0.494	0.076	0	0.188	0.196	0.705	0
	Qwen 7B	0.563	0.860	0.299	57	0.578	0.952	0.351	47	0.582	0.759	0.093	71
	Qwen 32B	0.430	0.455	0.211	22	0.431	0.443	0.106	57	0.237	0.270	0.497	83
Def-DTS	Llama 8B	0.474	0.525	0.218	1	0.473	0.527	0.398	0	0.246	0.268	0.464	10
	Qwen 7B	0.462	0.465	0.022	20	0.473	0.474	0.071	14	0.162	0.170	0.715	44
	Qwen 32B	0.338	0.374	0.501	38	0.392	0.400	0.429	1	0.221	0.343	0.435	18

Table 12: Performances on the local LLMs. We employed Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct for Llama 8B, Qwen 7b and Qwen 32B, respectively. P_k, WD, F1 were calculated only for correctly formatted outputs.

Model	TIAGE			SuperDialseg			Dialseg711		
	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑	P_k↓	WD↓	F1↑
Unsupervised Learning Methods									
DialSTART	-	-	-	-	-	-	0.179	0.198	-
SumSeg	0.438	0.455	-	0.469	0.480	-	-	-	-
LLM-based Methods									
Plain Text (GPT-3.5)	0.496	0.560	0.362	0.318	0.347	0.658	0.290	0.355	0.690
S3-DST$_{turn}$	-	-	-	-	-	-	0.009	0.008	-

Table 13: Performances reported in their original papers. Denoted as DialSTART indicates main result of Gao et al., 2023, SumSeg indicates main result of Artemiev et al., 2024, Plain Text indicates ChatGPT variant of Jiang et al., 2023's main result and S3-DST$_{turn}$ indicates main result of Das et al., 2024.

from the original experiments. Their original experimental results are presented in Table 13.

For the reproduction of LLM-based methods, we made the necessary modifications in the following cases.

Plain Text (Jiang et al., 2023): As Plain Text was the only methodology that disclosed the system prompt, we equitably refrained from using system prompts. We compared results with and without system prompts, finding nearly identical performance aside from parsing inconveniences.

S3-DST (Das et al., 2024): S3-DST constructs prompts on a turn basis, while we performed on an utterance basis. Their approach is not suitable for our approach when dialogue has consecutive utterance or odd numbers of utterances. Therefore, we modified turn-based inference to utterance-based.

All final prompts used are attached to our repository (prompts directory).

E Licenses for artifacts

Datasets TIAGE, SuperDialseg: MIT License, Dialseg711: We were unable to find a license for this dataset. However, you can find details of the dataset in the original paper(Xu et al., 2021b).

Models Qwen2.5-70B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-7B-Instruct: Qwen License, Llama3.1-72B-Instruct, Llama3.1-8B-Instruct: Llama License.