CoDial: Interpretable Task-Oriented Dialogue Systems Through Dialogue Flow Alignment

Anonymous ACL submission

Abstract

It is often challenging to teach specialized, unseen tasks to dialogue systems due to the high cost of expert knowledge, training data, and high technical difficulty. To support domain-specific applications—such as law, medicine, or finance—it is essential to build frameworks that enable non-technical experts to define, test, and refine system behaviour with minimal effort. Achieving this requires cross-disciplinary collaboration between developers and domain specialists. In this work, we introduce a novel framework, **CoDial** (Code for Dialogue), that converts expert knowledge, represented as a novel structured heterogeneous graph, into executable conversation logic. CoDial can be easily implemented in existing guardrailing languages, such as Colang, to enable interpretable, modifiable, and true zero-shot specification of task-oriented dialogue systems. Empirically, CoDial achieves state-of-the-art performance on the STAR dataset for inference-based models and is competitive with similar baselines on the well-known MultiWOZ dataset. We also demonstrate CoDial's iterative improvement via manual and LLM-aided feedback, making it a practical tool for expert-guided alignment of LLMs in high-stakes domains.[1]

1 Introduction

The recent emergence of Large Language Models (LLMs) has transformed the field of Natural Language Processing (NLP), achieving remarkable performance across a wide range of benchmarks. As language models become more widely adopted, experts in high-stakes domains such as law, medicine, and accounting are seeking ways to apply them responsibly to specialized tasks. Many domain experts lack coding skills, so it is important to provide tools that let them define, validate, and refine AI behaviour without writing code (Tian et al., 2024; Dahan et al., 2023). Cross-disciplinary collaboration often requires frameworks to possess interpretability and be generalizable to a wide variety of tasks, as domain experts may have limited expertise in programming and often need to understand step-by-step execution of the system. However, creating generalizable conversational systems is challenging due to the complexity of human conversation. Task-Oriented Dialogue (TOD) is an area of research dedicated to accomplishing real-world tasks (Qin et al., 2023; Jacqmin et al., 2022). Through these general frameworks, it becomes easier to build intelligent dialogue systems for specific, well-defined tasks across various domains.

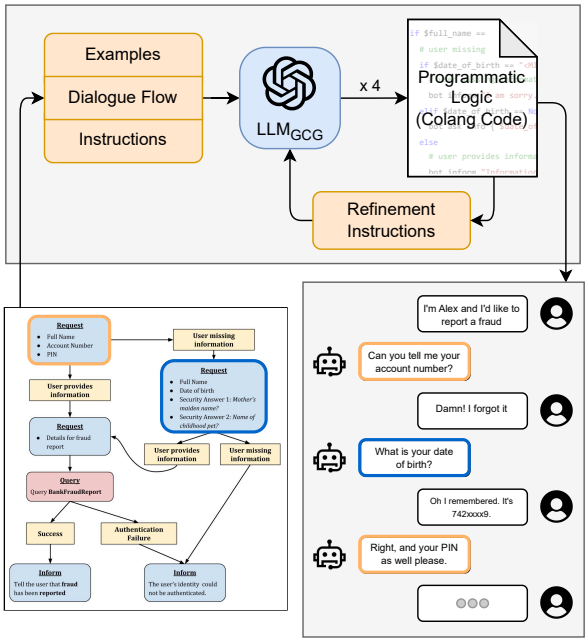


Figure 1: Overview of the proposed CoDial framework. An expert-curated dialogue flow (left) is transformed into executable programmatic logic using an LLM (top). The generated code is iteratively refined before producing the final program, which powers a conversational application (right), enabling the chatbot to follow the designer's requirements.

Many existing works utilize a schema-oriented framework to enforce complex task structures in

[1]Our code and data will be made publicly available at [GitHub Placeholder].

TODs (Zhang et al., 2023; Zhao et al., 2023; Mehri and Eskenazi, 2021). These systems convert tasks into a parsable *task schema* or *dialogue flow*, allowing experts to specify and integrate new tasks with minimal effort. However, these systems do not satisfy two critical properties: (1) **interpretability**, the ability to identify how the schema is being used by a language model to arrive at its outputs; and (2) the ability for a domain expert to easily **modify** the components, which is crucial for many real-world applications. For example, while programmatic representations like those in AnyTOD are technically interpretable, they remain difficult for non-experts to modify, as they lack an intermediary interface that would make the task schema accessible and modifiable. To advance real-world applications and support collaboration with domain experts, we explore a novel approach to create a framework that is easy for an expert to specify, validate, and modify. The contributions of our work are summarized as follows:

- We propose a novel approach for dialogue flow alignment that aims to *minimize human effort* requirement through high-level abstraction with strong *interpretability*. This allows domain experts without programming knowledge to effectively align dialogue systems at inference time.
- The proposed framework, CoDial, consists of three novel components. The heterogeneous dialogue flow representation allows domain experts to define rich task schemas. The guardrail-grounded code generation pipeline transforms dialogue flows into executable LLM guardrailing program, allowing for flexible control of LLMs in the inference stage. The CoDial human-feedback mechanism incorporates human and LLM feedback to refine and optimize the generated guardrailed conversational models.
- We demonstrate the effectiveness of our framework on publicly available benchmarks, STAR and MultiWOZ. We experiment with different code refinement strategies, by incorporating user feedback through manual and LLM-aided modification.

2 Related Work

Task-Oriented Dialogue Building generalizable conversational systems is challenging due to the complexity of human conversations, particularly when domain expertise is involved (Chen et al., 2017), leading to a focus on task-oriented systems for specific domains (Jacqmin et al., 2022). While LLMs have demonstrated impressive capability in a wide variety of domains, they struggled with TOD and fell behind if not used properly (Hudeček and Dusek, 2023). Some research (Zhang et al., 2023; Zhao et al., 2023; Mehri and Eskenazi, 2021) has used a schema-guided approach to generalize TOD systems to unseen tasks. Zhao et al. (2023) viewed the task schema as a program and adopted a neuro-symbolic approach to execute the policy program and control the dialogue flow.

Guardrails Guardrailing aims to enforce human-imposed constraints to control LLMs in the inference time (Dong et al., 2024; Rebedea et al., 2023; Guardrails AI). While it originated from AI safety, they can generally be used to define desired behaviour to constrain. Although traditional dialogue management systems, like Google Dialogflow[2], allow rigid modelling of dialogue states, they often lack flexibility to define complex task logic, and it is difficult for a user to further enhance the system. NVIDIA NeMo-Guardrails (Rebedea et al., 2023) is a toolkit that adds programmable guardrails to LLM-based conversational applications without fine-tuning. NeMo-Guardrails employs Colang (NVIDIA, 2024), a programming language, to establish highly flexible conversational flows and guide LLMs within them. Dong et al. (2024) suggested using neuro-symbolic approaches to guardrail LLMs, where a neural agent (e.g., an LLM) can deal with frequently seen cases, and a symbolic agent can embed human-like cognition through structured knowledge for the rare cases.

Code Generation and Prompt Optimization We use code generation strategies to convert structured graphs into programmatic guardrails. Code generation has made remarkable progress with the introduction of LLMs (Le et al., 2022). Although there are still challenges such as logical consistency and hallucinations (Liu et al., 2024). LLMs are proficient when in-context examples, documentations, or plans are provided (Jiang et al., 2024). There are many emerging methods to further optimize LLM generations (e.g., self-reflection, where LLMs are requested to update their own response), which have been shown to reduce hallucinations and improve problem solving (Ji et al., 2023). There has been research to improve output by rewriting the input prompt, referred to as prompt optimization (Yang et al., 2023; Yuksekgonul et al., 2024).

[2] https://dialogflow.cloud.google.com

3 Methodology

We introduce CoDial, a novel framework for constructing interpretable dialogue systems without requiring training data or programming expertise, as illustrated in Figure 1. We leverage programmatic LLM guardrailing, which allows flexible control of the behaviour of an LLM in the inference stage, and accordingly, ground TOD to graph-based dialogue flows that define the behaviour. CoDial is composed of three key components: (1) CoDial Heterogeneous Dialogue Flows (CHIEF) that allows domain experts to define the task schema (Section 3.1), (2) Guardrail-Grounded Code Generation (GCG) that automatically generates an executable guardrailing program based on the input dialogue flow (Section 3.2); In this paper, we use Colang (NVIDIA, 2024) guardrailing language, while any other programmatic guardrailing paradigm can be applied, and (3) CoDial Human Feedback (CHF) that incorporates human/LLM feedback to optimize the generated Colang conversational system (Section 3.3).

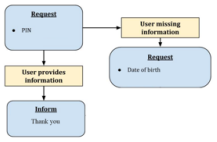
3.1 CoDial Dialogue Flow Representation

We design a structured framework that defines task schema as heterogeneous directed graphs, called CoDial Heterogeneous dIaloguE Flows (**CHIEF**) representation. Unlike prior work (Mehri and Eskenazi, 2021; Zhang et al., 2023) that define task schema as a homogeneous graph—where the only node type represents user intent, an API return value, or a dialogue state—CHIEF allows for different node or edge types in a heterogeneous manner, supporting more structured task definition. Specifically, CHIEF defines different node types that can define rich metadata and natural language logic to cover a wide range of tasks and domains, inspired by Mosig et al. (2020). To the best of our knowledge, we are the first to frame TOD task schema as a heterogeneous directed graph. We will show that CHIEF representation is compatible with, and can be converted to code for open-source guardrailing libraries. Below, we discuss the main node types and actions in CHIEF.

Request The request nodes define the variables, called slots, that CoDial tracks throughout the conversation (e.g. *the departure location in a taxi booking task*). When a conversation reaches this node, the system will request information specified by the slots. Each slot is assigned a data type (e.g. categorical) and accompanied by a few example

=== DST in Colang Example

[Simple Dialogue Flow]



Equivalent DST Code

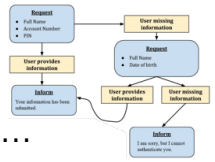
```
$pin = dst [...] $pin "extract user pin from conversation history ..."
$dob = dst [...] $dob "extract user date of birth from conversation history ..."
```

=== Colang Action Calling Definition

```
<ActionName>[(param=<value>[, param=<value>]…)]
```

=== NAP in Colang Example

[Simple Dialogue Flow]



Equivalent NAP Code

```
if $full_name == "<MISSING>" or ...
  # user missing information
  if $date_of_birth == "<MISSING>"
    # user missing information
    bot inform "I cannot authenticate you."
  elif ...
else
  # user provides information
  bot inform "Your information has been submitted"
```

=== Dialogue Flow

[x: Input CHIEF-Formatted Dialogue Flow as JSON]

=== Task Description

Define a variable for each request slot ...

Convert graph into nested if/else statements ...
For request nodes, ...
For external_action nodes, ...
For inform nodes, ...
For inform_and_confirm nodes, ...

Figure 2: An overview of $\text{prompt}_{\text{GCG}}(x)$, where a dialogue flow x is wrapped with system prompt template.

values. Additionally, CHIEF includes a free-form `rule` property to define the conditions under which a slot should be requested (e.g. in a taxi booking scenario, providing either a departure or arrival time is sufficient for booking). Since we leverage LLMs to build the TOD system, textual extensions can be easily incorporated.

External Action This node specifies a call to an external function within a dialogue flow. External actions enable the designer to execute complex logics through programming functions, interact with APIs, or invoke an LLM.

Inform (and Confirm) This node defines a template for providing information to the user (e.g. *Your taxi is booked with reference number [ref_no]*). The confirmation variant additionally allows the agent to ask a follow-up question (e.g. *Do you confirm the booking?*) and follow the appro-

Algorithm 1 An outline of CoDial's GCG output.

```
for each v in V^(H) do
    v ← NULL or FALSE                                  ▷ Initialize helpers
end for
while True do
    h_{2i-1} ← (h_{2i-2}; U_i)                         ▷ User input

    intent ← DetectIntent(h_{2i-1})                    ▷ Global action
    if intent ≠ NULL then
        B_i ← IntentResponse(intent)
        continue
    end if

    for each v_j^(s) in V^(S) do                       ▷ DST
        v_old ← v_j^(s)
        v_j^(s) = DST(h_{2i-1}, LLM_A, p_j^(s))
        if v_j^(s) ≠ v_old then
            for each hv in v_j^(s)'s dependents do
                hv ← NULL or FALSE
            end for
        end if
    end for

    state ← (V^(S), V^(H))                             ▷ NAP
    B_i, V^(H) ← NAP(state, LLM_A)
    if B_i = NULL then                                 ▷ Fallback action
        B_i ← LLM_A(V^(H))
    end if
end while
```

priate predefined dialogue path based on the user's response.

Global and Fallback Actions In addition to nodes, CHIEF supports representing global and fallback actions that are not tied to particular dialogue steps. Global actions can be triggered at any point in the dialogue flow. We also define fallback actions, general responses used when no other action is selected (e.g. *Sorry, I can't help with that*).

The defined nodes logically connect with **edges**. We add a textual `condition` property to edges to allow conditional branching in dialogue flows. We encode the graphs defined by CHIEF as text in JSON format. The JSON representation consists of a list of nodes and a list of edges (Figure 8). The JSON-encoded CHIEF represented dialogue flow is translated into guardrails code with our automatic code generation pipeline.

3.2 Guardrail-Grounded Code Generation

Guardrailing is a general paradigm to enable inference-stage control over LLMs' behaviour (Dong et al., 2024; Rebedea et al., 2023). Our work is the first to formulate TOD schema as programmatic guardrailing, which allows highly flexible model behaviour adjustment. Prior work (Zhang et al., 2023) use prompting, which requires prompt engineering and could be inefficient with long prompt contexts. In contrast, guardrailing acts as a proxy between the user and LLM, allowing for black-box LLM alignment and removing the need for model fine-tuning or data collection (Rebedea et al., 2023; Dong et al., 2024).

We propose CoDial Guardrail-Grounded Code Generation (GCG) that translates CHIEF-represented dialogue flows into Colang guardrailing code[3]. GCG is performed through prompting powerful LLMs (e.g., GPT-4o)[4]. Formally, the GCG process is denoted as $g = \text{LLM}_{\text{GCG}}(\text{prompt}_{\text{GCG}}(x))$, where $\text{prompt}_{\text{GCG}}(x)$ is a JSON-encoded CHIEF graph x wrapped with the prompt template instructions, and g is the output. Figure 2 shows an overview of $\text{prompt}_{\text{GCG}}$. The generated guardrailing code g is the *executable* TOD system.

Our $\text{prompt}_{\text{GCG}}$ outlines the output g, containing x's programmatic logic, as presented in Algorithm 1, with the definitions of the notations provided later in this section. g is a complete and executable guardrailing program powered by an LLM agent LLM_{A}. The conversation runs within an indefinite while loop, where the agent waits for user input, detects the user's intent for global actions, predicts the slot variables defined in request nodes (DST component), and selects an action and generates a response (NAP component). In this work, we leverage Colang's built-in intent detection feature for global actions. Note that DST and NAP are combined and generated by LLM_{GCG} as a single program (i.e., g). Finally, if the NAP component does not generate a response based on the defined logic (e.g. when user says *Goodbye!*), the LLM_{A} is prompted to choose an action, given the fallback actions and all actions defined in the dialogue flow. Figure 4 illustrates the execution life cycle of the agent.

We denote a conversation between user U and chatbot B as a history of messages, Equation 1, where U_i and B_i show user's and chatbot's i-th utterance, respectively. Therefore, the total number of utterances in h_{2i} is $2i$.

$$h_{2i} = (U_1, B_1, \ldots, U_i, B_i) \tag{1}$$

[3] https://docs.nvidia.com/nemo/guardrails/colang_2/overview.html

[4] We also experimented with (1) retrieval-augmented generation using the Colang Language Reference documentation and (2) fine-tuning GPT-4o-mini on generation pairs of (programming task, Colang code), but found that prompting with examples works best.

Moreover, we define a set of slot variables $V^{(S)}$ that track values for all of the slots defined in all request nodes, and helper variables $V^{(H)}$ that track the state for other (non-request) types of nodes. $V^{(S)}$ and $V^{(H)}$ together form the state of conversation $s = \left(V^{(S)}; V^{(H)}\right)$ at each turn, which is used to determine the next action. Please refer to Appendix A.1 for more details on code generation.

Dialogue State Tracking (DST) As suggested by Feng et al. (2023), LLM prompting shows promising performance in DST. Therefore, we use a simple prompting approach for DST. We leverage Colang's Natural Language Description (NLD) feature to extract the value of each slot from the conversation history. For each slot, $\text{prompt}_{\text{GCG}}$ instructs LLM_{GCG} to write instructions about extracting the value for that slot, given the history. The NLD instructions then prompt LLM_{A} when executed by Colang.

Formally, a slot variable $v_j^{(s)} \in V^{(S)}$ is predicted at bot's turn i as Equation 2, where $p_j^{(s)} \in P^{(s)}$ is the prompt generated by LLM_{GCG} to extract the value for $v_j^{(s)}$.

$$v_j^{(s)} = \text{DST}\left(h_{2i-1}, \text{LLM}_{\text{A}}, p_j^{(s)}\right) \quad (2)$$

Since updating a slot may affect the state (e.g. in a search task, modifying the search criteria requires re-executing the search), LLM_{GCG} needs to identify the helper variables that need to be invalidated when each slot is updated. We instruct LLM_{GCG} to list the helper variables of nodes that are reachable from the updated slot in the graph (i.e., nodes that are direct or indirect children of the slot's request node). These variables are then reset to null or false, depending on their type, when the slot is updated during execution.

Next Action Prediction (NAP) We instruct LLM_{GCG} to convert the dialogue flow tree into a conditional logic, consisting of nested if/else statements, to generate a response given the current state. For each node n_j in the dialogue flow, an if statement is generated to check whether the conversation is in that node, using $v_j^{(s)}$ or $v_j^{(h)}$, depending on the type of the node. If the condition holds, the corresponding action for that node is executed; otherwise, the logic proceeds to check its child nodes. For each outgoing edge from a node, the dialogue logic checks whether there is a condition associated with the edge and evaluates whether or not the condition is met. If there is no condition, the edge is followed automatically. Formally, next bot utterance is defined in Equation 3.

$$\left(B_i, V_{i+1}^{(H)}\right) = \text{NAP}(s_i, \text{LLM}_{\text{A}}) \quad (3)$$

3.3 CoDial Human Feedback Integration

CoDial's Human Feedback (CHF) mechanism incorporates feedback to refine or improve the generated dialogue logic. The code enhancement through human feedback comprises two broad approaches: i) manual modifications and ii) LLM-aided modifications.

CHF assist human feedback incorporation in the form of **refinement instructions (RIs)**, shown at the top in Figure 1. RIs allow the domain expert or developer of CoDial to refine the generated logic through text-to-code instructions. As the first step, we provide three instructions for refining certain aspects of the output code: correct logic (i.e., if statement) for each node, DST initialization, and request node checks. These RIs, presented in Table 6, are always applied to fix the problems in the output, if any. Moreover, we attempt to refine the generated DST prompts ($P^{(s)}$) through automatic prompt optimization. Please refer to Appendix A.3 for details on automatic DST prompt optimization. In addition, CHF allows for manual modifications on the dialogue flow (Appendix A.2) and manual DST prompt optimization (Appendix A.3).

4 Experimental Settings

Models We use GPT-4o[5], Claude 3.5 Sonnet[6], Gemini 2.0 Flash[7], and DeepSeek V3 (DSV3) (DeepSeek-AI et al., 2024) as LLM_{GCG}, and GPT-4o-mini and DSV3 as LLM_{A}[8] . Larger models are used for code generation—given the complexity of the task, we found that smaller models often fail to fully adhere to instructions. For further details, please refer to Appendix A.4.

4.1 Datasets

STAR The STAR dataset (Mosig et al., 2020), collected in a Wizard-of-Oz setup (human-human conversations), provides **explicit task schemas** (i.e., dialogue flows) to ensure consistent and deterministic system actions. We also use silver state

[5] https://openai.com/index/hello-gpt-4o/

[6] https://www.anthropic.com/news/claude-3-5-sonnet

[7] https://developers.googleblog.com/en/gemini-2-family-expands/

[8] All models were accessed from January to May 2025.

annotations created in STARv2 (Zhao et al., 2023) for ablation studies. Refer to Appendix A.2 for more implementation details on STAR.

MultiWOZ MultiWOZ (Budzianowski et al., 2018) is a large-scale, multi-domain TOD dataset consisting of human-human conversations, with most domains involving booking subtasks such as hotel reservations and taxi services. Given the impracticality of crafting dialogue flows for every possible domain combination (Zhang et al., 2023), we report results in a naive oracle domain setting. Please refer to Appendix A.3 for more details.

4.2 Metrics

For the STAR dataset, we compute BLEU-4 score (Papineni et al., 2002). We also follow Mosig et al. (2020) to compute F1 and accuracy. For the MultiWOZ dataset, we compute BLEU, Inform and Success rates, and Joint Goal Accuracy (JGA) using the official evaluation script (Nekvinda and Dušek, 2021). We report the mean result of three runs. Please refer to Appendix A.4 for more implementation details.

4.3 Baselines

For a complete list of compared methods, please refer to Appendix A.5. Our most comparable baselines are as follows:

- *AnyTOD* (Zhao et al., 2023) pretrains and finetunes T5-XXL for DST and response generation. It views task schema as a Python program to enforce the conversation logic.
- *IG-TOD* (Hudeček and Dusek, 2023), a few-shot prompting-based approach.
- *SGP-TOD* (Zhang et al., 2023) is a zero-shot prompting approach that employs graph-based dialogue flows and out-of-domain formatting examples. Refer to Appendix A.5 for details on fair comparison.

5 Experimental Results

5.1 Results and Analysis on STAR

SOTA Performance Without Training. Table 1 summarizes our results on the STAR dataset. CoDial achieves strong performance, surpassing all training-based approaches except AnyTOD PROG+SGD XXL, and sets the new SOTA among inference-based methods. Our framework improves F1 by **+5** and accuracy by **+6.9** points over the previous SOTA. While AnyTOD PROG+SGD XXL achieves higher scores, it requires manually written dialogue logic programs, making it less accessible to non-programmers, and extensive pretraining with task-specific data. In contrast, CoDial operates in a strict zero-shot setting, eliminating the need for manual programming and training. Without modifications (Appendix A.2), the original STAR dialogue flows result in lower performance (F1: 51.9). After manually modifying the dialogue flow and applying LLM-aided modifications to the generated code, we significantly enhance performance. We further explore the impact of LLM-aided corrections in Section 5.3.

Impact of Model Selection We experience with different model choices for the (LLM_{GCG}, LLM_{A}) pairing. Better instruction following and more robust code generation often translate to higher overall performance. Because most LLMs are unfamiliar with guardrailing languages such as Colang, they must accurately interpret the $\text{prompt}_{\text{GCG}}$ to produce syntactically correct code. When the chosen LLM struggles with instruction following, code generation can fail, leading to incorrect or incomplete programs. Among the tested configurations, CoDial (4O, 4O-MINI) achieves the highest performance in all metrics. We also report results in an oracle voting setting (Table 4) between GPT-4o-mini and DSV3 as LLM_{A}, where for each task, we take the best-performing LLM_{A} by F1. This results in an increase of $+1.7$ F1 and $+1.5$ accuracy.

State and Action Prediction and API Calls We find that NeMo Guardrails' intent detection performs strongly, achieving an F1 score of 96.3 on global actions (Table 3). Additionally, we observe that STAR's API calling precision—measured as the ratio of correct API calls to the total number of API calls—stands at 74.9. Table 3 also summarizes the performance of the actions that are generated by LLM_{A} (i.e., when NAP component does not generate an output). LLM-generated actions account for 25% of all predicted actions, with 70% of them belonging to three fallback actions: `goodbye`, `out_of_scope`, and `anything_else`. Excluding fallbacks, LLM-generated actions only account for 9.2% of predictions, indicating that our NAP logic is generally effective at generating outputs based on the predicted state. Since fallback actions are a simple 3-way classification, we would expect high performance. However, LLM_{A} achieves an F1 score of only 51.4. We attribute this to the lack of an explicit schema for fallback actions in the STAR dataset, leading to inconsistencies in wizard

Model	Approach	LLM_{GCG}	LLM_A	RI	F1	Acc.	BLEU
Training-based Approaches					*Domain Transfer*		
BERT + Schema	Fine-tuning	-	-	-	29.7	32.4	-
SAM	Fine-tuning	-	-	-	51.2	49.8	-
AnyTOD NOREC BASE	Fine-tuning	-	-	-	55.8	56.1	32.4
AnyTOD PROG+SGD XXL	Pretraining	-	-	-	70.7	70.8	44.2
Inference-based Approaches					*Strict Zero-shot*		
SGP-TOD-GPT3.5-E2E	Zero-shot	-	-	-	53.5	53.2	-
CoDial (Ours)							
CoDial ORIGINAL DFS	Zero-shot	4o	4o-mini	✓	51.9	51.1	38.9
CoDial − RI	Zero-shot	4o	4o-mini	✗	56.1	57.3	38.4
CoDial − RI	Zero-shot	Sonnet	4o-mini	✗	57.0	58.4	39.2
CoDial	Zero-shot	DSV3	4o-mini	✓	46.1	48.0	28.0
CoDial*	Zero-shot	Gem. 2 Fl.	4o-mini	✓	50.5	52.1	32.9
CoDial	Zero-shot	Sonnet	4o-mini	✓	57.7	58.5	39.3
CoDial	Zero-shot	4o	DSV3	✓	55.6	56.8	44.2
CoDial	Zero-shot	4o	4o-mini	✓	**58.5**	**60.1**	**45.2**

Table 1: Comparison of models on the STAR dataset. Our CoDial model achieves the SOTA in a "Strict Zero-Shot" setting, where we do not require any training samples or sample conversations. SAM results are cited from Zhao et al. (2023). The generated code for the model with an asterisk (*) has been manually fixed and is not directly comparable. DF stands for "dialogue flow."

Model	JGA	Inform	Success	BLEU	Combined
Training-based Approaches					
SOLOIST	35.9	81.7	67.1	13.6	88.0
MARS	35.5	88.9	78.0	19.6	103.0
AnyTOD XXL	30.8	76.9	47.6	3.4	65.6
Inference-based Approaches					
IG-TOD (fs)	27	-	44	6.8	-
SGP-TOD	-	82.0	72.5	9.2	86.5
CoDial	28.4	76.6	54.6	3.5	69.1

Table 2: Comparison of models on the MultiWOZ dataset. SOLOIST and MARS results are cited from Zhao et al. (2023). (*fs*) indicates few-shot.

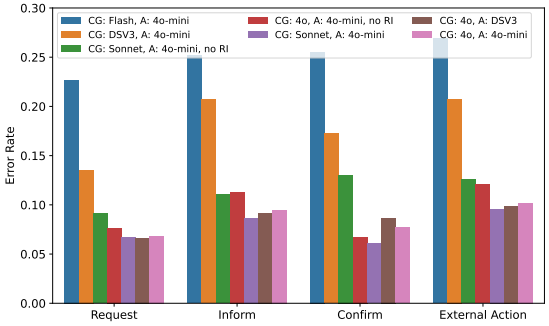


Figure 3: Error rate comparison of agents' predicted state on the STAR dataset across different node types, coloured by (LLM_{GCG}, LLM_A) pairs.

annotations. Additionally, we observe a significant performance drop from fallback to non-fallback actions in both F1 (51.4 → 38.7) and accuracy. This suggests that despite having an explicit schema, LLMs struggle to capture the more complex logic needed to predict non-fallback actions. Our findings align with Dong et al. (2024), reinforcing the need for a neuro-symbolic approach.

Figure 3 shows the error rate of the predicted conversation state across different node types for each model. To approximate the error, we compare the model's predicted state with the estimated ground-truth state (i.e., the wizard's state), as described in Appendix A.2. We find that the error rate generally inversely correlates with the overall performance in Table 1; higher-performing models tend to exhibit lower state prediction error.

5.2 Results on MultiWOZ

Our results on the MultiWOZ dataset are summarized in Table 2. Unlike STAR, where wizards were provided structured guidance for system responses, MultiWOZ lacks a predefined dialogue flow, making interactions less consistent. This variability in MultiWOZ poses additional challenges for heuristics-grounded and programmatic approaches like CoDial and our most comparable system, AnyTOD. AnyTOD trains LLMs to use a strict programmatic schema and generates output text with templates. However, we achieve comparable performance to AnyTOD *without pre-training*.

Most of the MultiWOZ test set consists of multi-domain conversations, where a user may, for example, book both a taxi and a restaurant in the same dialogue. Since CoDial is designed for single-domain interactions, we report its performance on single-domain dialogues in Table 5, where it performs well. However, with our naive oracle domain setting, CoDial performance drops significantly. This is likely due to compounded errors from DST to NAP, which we analyze further in Section 5.3.

5.3 Detailed Analysis

Oracle DST Performance To assess the impact of compounded DST error, we evaluate CoDial under an Oracle setting. Since STAR does not provide gold DST labels, we simulate an oracle setting by

Actions	F1	Acc.
Intent Detection (Global Actions)		
All	96.3	92.8
LLM Generated Actions		
Fallbacks	51.4	57.8
Excluding Fallbacks	38.7	39.0
All	49.1	52.1

Table 3: Individual action prediction performance of intent detection and LLM_A in CoDial. Fallback actions include `goodbye`, `out_of_scope`, and `anything_else`. All entries are micro-averaged.

Model	F1	Acc.	BLEU
CoDial	58.5	60.1	**45.2**
Refinement Instructions (RI)			
− RI 3	56.1	58.0	41.6
− RI 3 & 2	55.9	57.6	41.7
− RI 3 & 2 & 1	56.1	57.3	38.4
Generative Approach			
− NAP	47.4	47.0	25.8
− NAP & DST	42.7	43.0	23.8
Oracle Vote			
DST: 4o-mini + DSV3	60.2	61.6	46.8
Silver Label DST			
+ STARv2 States	**60.7**	**62.9**	44.3

Table 4: Ablations on the STAR dataset.

Model	JGA	Inform	Success	BLEU	Combined
Predicted Belief State					
IG-TOD (fs)	27	-	44	6.8	-
CoDial SINGLE*	46.2	91.5	77.6	3.2	87.7
CoDial	28.4	76.6	54.6	3.5	69.1
Oracle Belief State					
IG-TOD (fs)	-	-	68	6.8	-
CoDial SINGLE*	-	94.6	90.6	3.5	96.1
CoDial	-	93.1	75.3	4.0	88.2
DST Prompt Optimization					
CoDial AUTO	31.1	72.3	57.2	3.6	68.3
CoDial MANUAL	28.5	80.4	57.9	3.6	72.7

Table 5: Ablations on MultiWOZ. Settings with an asterisk (*) are not directly comparable due to a simpler task setup. (fs) indicates few-shot.

replacing CoDial's DST component with the silver-annotations from STARv2 (Table 4). This results in a performance gain of +2.2 F1 and +2.8 accuracy. We do the same for MultiWOZ, where we replace our predictions with the gold belief state. As shown in Table 5, this leads to a substantial performance improvement, making CoDial competitive with other baselines. These findings suggest that exploring more advanced DST approaches could be a promising direction to improve performance, regardless of the dataset. We experiment briefly with prompt optimization, described below, but we urge further exploration in this direction.

Code Optimization We use LLMs to perform iterative code refinement and automatic prompt optimization for the DST prompts. Refining the code with RIs consistently enhances CoDial's performance, demonstrating the benefits of integrating user feedback into the generation process. After prompting the LLM to iteratively refine its outputs, CoDial achieves better accuracy and fluency (compared to CoDial − RI in Table 1). We also conduct an ablation study to examine the effect of the individual RIs, summarized in Table 4. Although all RIs are beneficial, most of the performance improvements can be attributed to the third RI, which refines the conditional logic of request nodes.

After observing the results of the oracle DST setting, we also apply prompt optimization to improve DST accuracy. As shown in Table 5, automatic prompt optimization yields only marginal gains across metrics, with the exception of Inform, suggesting that improving DST remains a non-trivial challenge—potentially due to the limitations of our naive multi-domain evaluation setup. This is a surprising result, but it suggests that not all DST slots are equally valuable for the final performance. To explore the impact of human feedback, we also experiment with manual prompt optimization (Appendix A.3), making minor edits to the prompts for the "attraction" domain. This results in consistent improvements across all metrics, reinforcing that human-crafted prompts can still outperform automatic optimization.

Generative Approach To better understand the effectiveness of the proposed CoDial architecture, we experiment with a setting in which the NAP component is removed and all actions are predicted in a fully generative manner by the LLM_A, similar to Zhang et al. (2023). We prompt the LLM_A with simplified dialogue flows following their work and include predicted DST slots in the prompt. As shown in Table 4, this results in a substantial drop in performance, highlighting the importance of our structured NAP logical flow approach. We further ablate the model by removing the DST component, causing a larger drop as expected.

6 Conclusion

In this work, we introduced CoDial, a novel framework for building interpretable TOD systems by grounding structured dialogue flows to programmatic guardrails. CoDial introduces CHIEF, a heterogeneous graph representation of dialogue flows, and employs LLM-based code generation to automatically convert dialogue flows into executable guardrail specifications (e.g., in NVIDIA's Colang), enabling zero-shot creation of interpretable TOD systems with minimal expert effort. Moreover, through iterative manual and LLM-aided refinements, CoDial supports rapid incorporation of domain-expert feedback, further enhancing the generated code. Our empirical findings support CoDial's effectiveness, achieving SOTA performance among inference-based methods on STAR and competitive results on MultiWOZ.

Limitations

While CoDial offers an interpretable and modifiable approach to TOD systems, it has certain limitations. First, scalability remains a challenge. For large and complex dialogue flows, CoDial requeries all slots every turn, which may increase latency and computational cost. In general, the DST is the biggest challenge of the system and we leave improvements to future work. Second, CoDial is less effective for multi-domain dialogues, as it operates on a single dialogue flow at a time. Handling seamless transitions between multiple domains would require additional mechanisms beyond the current framework, and structured logic on how different domains flow into each other, which we leave to future work. In general, dialogue flows assume there is a logical progression to conversations, which isn't the case for all dialogue structures. Finally, reproducibility could be a concern when relied on API-based LLMs. Since these models are periodically updated, responses may vary across different API versions, making consistent evaluation challenging.

Ethics Statement

This work adheres to ethical research practices by ensuring that all models, codebases, and datasets used comply with their respective licenses and terms of use. The STAR and MultiWOZ datasets employed in our experiments do not contain personally identifiable information or offensive content.

As with any system leveraging LLMs, CoDial inherits potential risks related to bias and factually incorrect outputs. However, our framework mitigates these risks by enforcing structured dialogue flows, guardrailing based on user intent, and template-based responses, reducing the likelihood of hallucinated or biased content. Future work may integrate NeMo Guardrails' input and output rails to filter inappropriate inputs and outputs, enhancing system safety. Since our focus is on structured dialogue flows, we leave this for future exploration.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Samuel Dahan, Rohan Bhambhoria, David Liang, and Xiaodan Zhu. 2023. Lawyers should not trust ai: A call for an open-source legal language model. *Available at SSRN 4587092*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Building guardrails for large language models. *Preprint*, arXiv:2402.01822.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755, Singapore. Association for Computational Linguistics.

Guardrails AI. Guardrails: Adding guardrails to large language models. `https://github.com/guardrails-ai/guardrails`. Accessed: 2025-05-16.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. 2024. Self-planning code generation with large language models. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–30.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 21314–21328. Curran Associates, Inc.

Zekun Li, Zhiyu Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Dong, Adithya Sagar, Xifeng Yan, and Paul Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2024. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36.

Shikib Mehri and Maxine Eskenazi. 2021. Schema-guided paradigm for zero-shot dialog. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 499–508, Singapore and Online. Association for Computational Linguistics.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

NVIDIA. 2024. Nvidia nemo guardrails, docs.nvidia.com/nemo/guardrails/colang_2/overview.html. Accessed: 2025-05-19.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941, Singapore. Association for Computational Linguistics.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 431–445, Singapore. Association for Computational Linguistics.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2023. Mars: Modeling context & state representations with contrastive learning for end-to-end task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11139–11160, Toronto, Canada. Association for Computational Linguistics.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text.

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023. SGP-TOD: Building task bots effortlessly via schema-guided LLM prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13348–13369, Singapore. Association for Computational Linguistics.

Jeffrey Zhao, Yuan Cao, Raghav Gupta, Harrison Lee, Abhinav Rastogi, Mingqiu Wang, Hagen Soltau, Izhak Shafran, and Yonghui Wu. 2023. AnyTOD: A programmable task-oriented dialog system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16189–16204, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Details on Code Generation

We represent the graphs defined by CHIEF (Section 3.1) as text in JSON format. The JSON representation consists of a list of nodes and a list of edges. The node list defines the dialogue flow nodes, specifying their types and assigning each a unique identifier (node ID). The edge list specifies the connections between nodes using their IDs (Figure 8). The JSON nodes, global and fallback actions, and functional specifications for function

ID	Description	Instruction
RI 1	Revise if statements	Revise the 'if's to exactly reflect the nodes. Comment each 'if' to specify the corresponding node ID. Make sure the generated 'if' statement and its body reflect the instructions for that node type.
RI 2	Fix dst dependent vars	Fix dst's first input parameter. It should reflect which variables should be invalidated when the corresponding slot is updated.
RI 3	Fix request node checks	Fix 'if' checks for request nodes. Comment their rule, if available. The 'if' should reflect the rule for each node.

Table 6: Instructions for Code Refinement

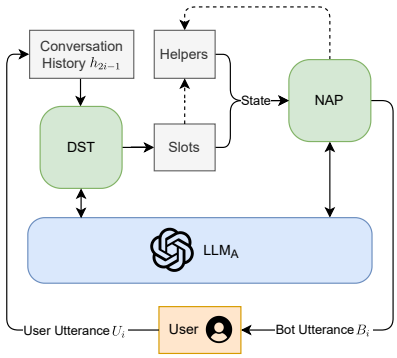


Figure 4: Execution life cycle of the generated agent.

calls are translated into Colang code with our automatic code generation pipeline. The external action node functions, referred to as Actions in Colang, are implemented in Python.

Post-processing Following code generation by the LLM, we apply rule-based post-processing to ensure proper execution. This includes adding helper flows (Colang's equivalent of functions) to support algorithm execution, enabling the loading of the STAR API function, and injecting additional code for evaluation purposes.

Helper Variables The algorithm designed in Colang (Algorithm 1) determines whether a request node should be executed (i.e., prompt the user for information) by checking the values of its associated slots. To track the state of other node types, we instruct LLM$_{GCG}$ to define helper variables following a structured naming pattern, where `<id>` represents the corresponding node's ID:

- `action_<id>`: Stores the return value of external actions.
- `inform_<id>`: Indicates whether the node has been executed and the user has been informed.
- `answered_<id>`: For inform and confirm nodes, stores the user's response.

Example Dialogue Flow and Generated Code Figures 7 and 8 present an example of the STAR task schema pictures, our JSON representation of the corresponding dialogue flow, and the generated Colang code.

A.2 STAR Implementation Details

The STAR dataset (Mosig et al., 2020), collected in a Wizard-of-Oz setup (human-human conversations), provides **explicit task schemas** (i.e., dialogue flows) to ensure consistent and deterministic system actions. It serves as a benchmark for TOD systems, enabling evaluation across 24 tasks and 13 domains. STAR's structured collection aligns well with our objectives and CoDial's design choices. We also use silver state annotations created in STARv2 (Zhao et al., 2023) for ablation studies.

API Calling While not the primary focus of this paper, we use prompting to generate Colang's Python action code for calling STAR's API and processing its outputs automatically, rather than directly feeding ground-truth API responses as input as done in other works. Every piece of code in our pipeline is automatically generated. Since STAR's API returns randomized outputs, we return the ground-truth API response object when it is available for the exact same turn, instead of the random sampling response.

Dialogue Flows We convert the STAR task schemas, originally provided as images, into CHIEF representation described in Section 3.1. We use one-shot prompting with GPT-4o to convert pictures into JSON. We convert yellow nodes in pictures into conditions for edges. However, we observed that GPT-4o occasionally misassigns edge connections, requiring manual corrections. Additionally, we enrich the JSON representations by adding more context, such as example values for each slot. We also define `hello` action as the only global action and `goodbye`, `out_of_scope`, and

Algorithm 2 Wizard state approximation

Require: Variable v, Graph G, Ground-truth action a_{gt}, Mapping ϕ
Ensure: Approximated value or NULL

$n_{tgt} \leftarrow \phi(a_{gt})$
$n_v \leftarrow v.node$
$P \leftarrow \text{DFSPATH}(G, G.start, n_{tgt})$
if $n_v \notin P$ **then**
    return NULL
end if
for each $e \in P$ **do**
    if $e.target = n_v$ **then**
        $e_v \leftarrow e$
        break
    end if
end for
return $\text{APPROXVALUE}(e_v.condition, v)$

`anything_else` as fallback actions for all tasks.

To better align the dialogue flows with the actual collected dialogues, we introduce minor modifications, such as adding the `inform_nothing_found` action for search tasks. We also identified small inconsistencies between the provided API schema and its implementation. To address this, we refine the API definitions and modify the sampling logic to prevent errors when no results match the given constraints. We will release these improvements, aiming to support future research.

Wizard State Approximation For evaluation, since we are working with offline conversations (i.e., the user is not interacting with the actual TOD system), we approximate the wizard's state at the end of each turn and adjust the program's state accordingly. This helps prevent the program's state from deviating from the ground-truth conversation. To achieve this, we first find the node in dialogue flow that the ground-truth conversation was in by mapping the ground-truth action label, if available, to a node in the dialogue flow. We manually create this mapping from action labels to the dialogue flow nodes. Next, we use depth-first search to trace the path from the start of the dialogue flow to the current conversation node. Finally, we adjust each state variable based on whether the corresponding node is part of the current conversation pathway, as described in Algorithm 2.

Prompt Context During evaluation, we incorporate the textual guidelines provided to wizards into LLM_A's context. This additional context helps the LLM infer some details, such as the time or location of the conversation. For example, a guideline might look like: *Some facts you should be aware of: Right now, it is Tuesday, 12 PM.*

A.3 MultiWOZ Implementation Details

MultiWOZ (Budzianowski et al., 2018) is a large-scale, multi-domain TOD dataset consisting of human-human conversations, with most domains involving booking subtasks such as hotel reservations and taxi services. Given the impracticality of crafting dialogue flows for every possible domain combination, as mentioned in previous work (Zhang et al., 2023), we report results in a naive oracle domain setting. We preprocess MultiWOZ 2.2 using the code from Li et al. (2024) to annotate each conversation turn with its active domains. For each turn i, we use the dialogue flow(s) of the corresponding domain(s) to predict the output and merge all turns at the end.

Manually Crafted Dialogue Flows Unlike STAR, MultiWOZ does not provide explicit dialogue flows for each domain, nor do its conversations adhere to a specific flow. To address this, we manually construct simple dialogue flows by analyzing a few example dialogues from each domain. We will release these crafted MultiWOZ dialogue flows. Additionally, for evaluation, we modify the prompts and instruct the LLM to generate delexicalized texts.[9]

Naive Multi-domain Rather than adding a separate domain detection step, we use the gold labels for the active domains at each conversation turn and directly apply the corresponding dialogue flows. We preprocess MultiWOZ 2.2 using the code from Li et al. (2024) to annotate each turn with its active domains. Since evaluation is offline, we separate turns in a conversation by domain, simulate the conversation with prior history, and use the corresponding Colang program(s). Finally, we merge all turns and treat slots from all domains as a single set, accumulating DST predictions during evaluation.

DST Prompt Optimization The NAP component's performance is largely dependent on DST, as the next action is determined by the values known to the dialogue system (Equation 3). However, we found in preliminary experiments that the DST performance can be poor with original $P^{(s)}$

[9] Refer to Nekvinda and Dušek (2021) for more details.

Algorithm 3 Our prompt optimization algorithm. We randomly sample a training and validation set of size 20 and 50 for every DST slot, respectively.

Require: Training set $\mathcal{D}_{\text{train}}$, Validation set $\mathcal{D}_{\text{val}}$, Instruction I, Agent LLM_A, Optimizer LLM M, Batch size B

$\hat{Y}_{\text{val}} \leftarrow \text{DST}(\mathcal{D}_{\text{val}}.H, \text{LLM}_\text{A}, I)$
Initialize $S \leftarrow \textsc{ComputeScore}(\hat{Y}_{\text{val}}, \mathcal{D}_{\text{val}}.Y)$
$I_{\text{best}} \leftarrow I$
Divide $\mathcal{D}_{\text{train}}$ into batches $\mathcal{B}_1, \ldots, \mathcal{B}_n$ of size B
for each batch $\mathcal{B}$ in $\mathcal{D}_{\text{train}}$ **do**
    $(H, Y) \leftarrow \mathcal{B}$
    $\hat{Y} \leftarrow \text{DST(H, LLM}_\text{A}\text{, } I_{\text{best}})$
    $I \leftarrow M.\textsc{Rewrite}(H, \hat{Y}, Y, I)$
    $\hat{Y}_{\text{val}} \leftarrow \text{DST}(\mathcal{D}_{\text{val}}.H, \text{LLM}_\text{A}, I)$
    $S \leftarrow \textsc{ComputeScore}(\hat{Y}_{\text{val}}, \mathcal{D}_{\text{val}}.Y)$
    if $S > S_{\text{best}}$ **then**
        $S_{\text{best}} \leftarrow S$
        $I_{\text{best}} \leftarrow I$
    end if
end for
return $I_{\text{best}}, S_{\text{best}}$

prompts, generated by general guidelines outlined in $\text{prompt}_{\text{GCG}}$. To this end, we further refine $P^{(s)}$ with automatic prompt optimization.

Our optimization algorithm is summarized in Algorithm 3. For each DST variable $v_j^{(s)}$, we randomly sample two mutually exclusive sets of conversation turns to serve as training and validation sets. The training examples are divided into batches of 5, and each batch is used to guide the optimizer GPT-4o model to rewrite the instruction $p_j^{(s)}$, resulting in a candidate prompt. If the revised instruction improves performance on the validation set, it is retained; otherwise, the original is kept, ensuring that modifications are only accepted when they lead to measurable improvements.

In addition, we manually refine the prompts for the worst-performing domain, "attraction." The edits include defining what an "attraction" is by listing all possible types, and propagating the predicted `type` value to other slot instructions to maintain consistency. We leave further investigation of this technique—passing key slot predictions across instructions within a domain—as future work.

A.4 Experimental Details

For the STAR dataset, we compute BLEU-4 score (Papineni et al., 2002). using SacreBLEU (Post, 2018). We also follow Mosig et al. (2020) to compute F1 and accuracy. Since we applied STAR's response templates for response generation, we use regex patterns to match generated responses with actual values to a template. For the MultiWOZ dataset, we compute BLEU, Inform and Success rates, and Joint Goal Accuracy (JGA) using the official evaluation script (Nekvinda and Dušek, 2021). We report the mean result of three runs.

If a generated program contains syntax or runtime errors, we regenerate the code to obtain a functional version. The only exception is Gemini 2.0 Flash, which struggles with calling our defined Colang helper flows. Since this issue is minor, we manually correct the syntax to assess the model's ability to generate programmatic logic for dialogue flows. We access OpenAI models through OpenAI and other models through OpenRouter[10] API.

NeMo-Guardrails To implement the Colang guardrails, we use a fork from NeMo-Guardrails version 0.11, modified to inject our evaluator class[11]. We use this class to evaluate on the ground truth user-wizard history, instead of the history of user-bot conversation, similar to Nekvinda and Dušek (2021).

We modify NeMo's default `value_from_instruction` prompt structure to begin with a system message, followed by the entire conversation history and instructions combined into a single user message (Figure 5). During our initial experiments, we suspected that NeMo's original prompt structure—where each message in the conversation history was passed as a separate user or assistant message—hindered LLM_A's ability to follow instructions effectively.

Additionally, we refine the post-processing of this action. We found that LLM_A was inconsistent in formatting return values, sometimes enclosing strings in quotation marks while omitting them for non-string types. To address this, we first check if both leading and trailing quotation marks are present and remove them if so. We then attempt to evaluate the return value as a Python literal. If this evaluation fails, we then enclose the value in quotation marks to ensure proper parsing as a string.

Moreover, we fixed an issue related to if-else statements in the Colang parser, which was later

[10] https://openrouter.ai/

[11] The modified NeMo-Guardrails version that we used for the experiments is available at `[GitHub Placeholder]`.

System Prompt

Below is a conversation between a helpful AI assistant and a user. The bot is designed to generate human-like text based on the input that it receives. The bot is talkative and provides lots of specific details. If the bot does not know the answer to a question, it truthfully says it does not know.

Your task is to generate value for the specified variable. The generated value should be a valid Python literal that is parsable by ast.literal_eval. Always put strings in quotes.

Do not provide any explanations, just output value.

User Prompt

This is some information that is given to the bot to answer to user:
Authenticate the user and tell them their bank balance

This is the current conversation between the user and the bot:
== User:
user action: Hi I would like to check my balance.
== Bot:
bot action: Could I get your full name, please?
== User:
user action: Katarina Miller
== Bot:
bot action: Can you tell me your account number, please?
== User:
user action: I can't remember it right now.

Follow the following instruction to generate a value that is assigned to: $val
Instruction: `this variable stores user's account number. examples of the variable value are "12345678", "87654321". the current variable value is None. given the last user and bot interaction in the current conversation, if the last user message has provided a new value for this variable, output it. if the last interaction is not about this variable, output the current value.`

Figure 5: Example of the modified NeMo `value_from_instruction` action prompt, which is used for DST. h_{2i-1} and $p_j^{(s)}$ are provided in each prompt to generate a value for that slot.

merged into the official NeMo repository[12].

A.5 Detailed Baselines

- *AnyTOD* (Zhao et al., 2023) pretrains and fine-tunes T5-XXL for dialogue state tracking and response generation. It uses a Python program to enforce the complicated logic defined by a dialogue flow to guide the LM decisions.

- *IG-TOD* (Hudeček and Dusek, 2023) is a prompting-based approach using ChatGPT to track dialogue states via slot descriptions, retrieve database entries, and generate responses without fine-tuning.

- *SGP-TOD* (Zhang et al., 2023) is a purely generative approach that uses two-stage prompting to track dialogue state and generate response. It employs graph-based dialogue flows to steer LLM actions, ensuring adherence to predefined task policies without requiring fine-tuning or training data. To ensure a fair comparison, we replicated their setup using the same newer LLM_A model as ours (Table 7). We ran their released code without modification, except for switching the API model to GPT-4o-mini. Surprisingly, performance dropped significantly. After contacting the authors, they advised adapting the prompt structure to the aligned LLMs—placing instructions in the system message and including examples and dialogue history in the user message. However, even with this adaptation, the performance did not match the results originally reported with GPT-3.5, suggesting that a generative approach could not be a trivial solution and requires careful prompt engineering. Figure 6 further illustrates differences between CoDial and SGP-TOD through two cherry-picked examples.

Model	F1	Acc.
SGP-TOD GPT3.5-E2E	53.5	53.2
SGP-TOD GPT4O-MINI-E2E	41.3	44.3
SGP-TOD GPT4O-MINI-E2E Adapted	40.3	43.8

Table 7: Comparison of SGP-TOD baselines.

- *BERT + Schema* and *Schema Attention Model (SAM)* (Mosig et al., 2020; Mehri and Eskenazi, 2021) incorporate task schemas by conditioning on the predefined schema graphs, enabling structured decision-making in TODs. SAM extends BERT + Schema approach with an improved schema representation and stronger attention mechanism, aligning dialogue history to the schema for more effective next-action prediction. Both models rely on fine-tuning to learn schema-based task policies and improve generalization across tasks.

- *SOLOIST* (Peng et al., 2021) is a Transformer-based model that unifies different dialogue modules into a single neural framework, leveraging transfer learning and machine teaching for TOD systems. It grounds response generation in user goals and database/knowledge, enabling effective adaptation to new tasks through fine-tuning with minimal task-specific data.

- *MARS* (Sun et al., 2023) is an end-to-end TOD system that models the relationship between dia-

[12] GitHub pull request at `[PLACEHOLDER]`.

Dialogue History	USER: Help there have been suspicious transfers over the past week. my account number is 351531510 and my PIN is 1596.
Wizard Action	ask_name
SGP-TOD Action	bank_ask_fraud_report
CoDial Action	ask_name

(a) *Bank Fraud Report* example dialogue. SGP-TOD fails to collect all necessary authentication details before requesting fraud report information, as its schema defines the next action after `user_bank_inform_pin` as `bank_ask_fraud_details`. In contrast, CoDial verifies that all required information is provided at each request node before proceeding, correctly identifying that the user's name is missing.

Dialogue History	USER: Hi, I am Ben. I would like to plan a party. WIZARD: On what day would you like your party organised? USER: Saturday at 10pm. WIZARD: At what venue would you like to have your party organised? USER: The North Heights Venue if it's available.
Wizard Action	party_ask_number_of_guests
SGP-TOD Action	party_venue_not_available
CoDial Action	party_ask_number_of_guests

(b) *Party Plan* example dialogue. SGP-TOD produces an incorrect and uninterpretable prediction. In contrast, CoDial follows a programmatic logic aligned with the dialogue flow, ensuring interpretability.

Figure 6: Cherry-picked comparison of CoDial and SGP-TOD performance. We use GPT-4o-mini to reproduce SGP-TOD results.

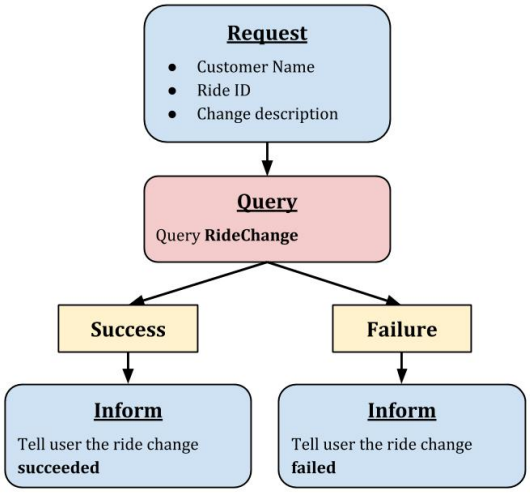
logue context and belief/action state representations using contrastive learning. By employing pair-aware and group-aware contrastive strategies, Mars strengthens the modelling of relationships between dialogue context and semantic state representations during end-to-end dialogue training, improving dialogue state tracking and response generation.

```
import core
import llm

flow main
  activate automating intent detection
  activate generating user intent for unhandled user utterance
  $action_2 = None
  $inform_3 = False
  $inform_4 = False

  global $generated_output
  while True
    $generated_output = None
    when user said hello
      bot say "Hello, how can I help?"
      continue
    or when unhandled user intent as $state
      $transcript = $state.event.final_transcript

    $customer_name = dst ["action_2", "inform_3", "inform_4"] $customer_name "this variable
stores the name of the customer requesting the ride change. examples of the variable value are
\"John\", \"Jane\". the current variable value is {$customer_name}. given the last user and bot
interaction in the current conversation, if the last user message has provided a new value for
this variable, output it. if the last interaction is not about this variable, output the current
value."
    $ride_id = dst ["action_2", "inform_3", "inform_4"] $ride_id "this variable stores the unique
identifier for the ride (ride id). examples of the variable value are \"102\", \"500\". the
current variable value is {$ride_id}. given the last user and bot interaction in the current
conversation, if the last user message has provided a new value for this variable, output it. if
the last interaction is not about this variable, output the current value."
    $change_description = dst ["action_2", "inform_3", "inform_4"] $change_description "this
variable stores the description of the requested change to the ride. examples of the variable
value are \"Change pickup time\", \"Change destination\", \"Update contact details\". the current
variable value is {$change_description}. given the last user and bot interaction in the current
conversation, if the last user message has provided a new value for this variable, output it. if
the last interaction is not about this variable, output the current value."

    if $customer_name == None or $ride_id == None or $change_description == None
      bot ask info {"$customer_name": $customer_name, "$ride_id": $ride_id,
"$change_description": $change_description}
    else
      if $action_2 == None
        $action_2 = await RideChangeAction(customer_name=$customer_name, ride_id=$ride_id,
change_description=$change_description)

      if $action_2["status"] == "Success"
        if not $inform_3
          bot inform "Alright, thats all changes done for you!"
          $inform_3 = True
      elif $action_2["status"] == "Failure"
        if not $inform_4
          bot inform "Unfortunately I wasn't able to update your booking, sorry."
          $inform_4 = True

    if $generated_output == None
      $generated_output = await LLMGenerateOutputAction()
      $generated_output = str($generated_output)
      await UtteranceBotAction(script=$generated_output)
```

Figure 7: Example of a generated code for STAR *Ride Change* task

```
{
  "nodes": [
    {
      "id": 1,
      "type": "request",
      "slots": {
        "customer_name": {
          "description": "Name of the customer
requesting the ride change",
          "type": "categorical",
          "examples": [ "John", "Jane" ]
        },
        "ride_id": {
          "description": "Unique identifier
(ride ID) for the ride",
          "type": "integer",
          "examples": [ "102", "500" ]
        },
        "change_description": {
          "description": "Description of the
requested change to the ride",
          "type": "string",
          "examples": [
            "Change pickup time",
            "Change destination",
            "Update contact details"
          ]
        }
      }
    },
    {
      "id": 2,
      "type": "external_action",
      "action": "query",
      "query": "RideChange"
    },
    {
      "id": 3,
      "type": "inform",
      "message": "Alright, thats all changes
done for you!"
    },
    {
      "id": 4,
      "type": "inform",
      "message": "Unfortunately I wasn't able
to update your booking, sorry."
    }
  ],
  "edges": [
    { "source": 1, "target": 2 },
    { "source": 2, "target": 3, "condition":
"Success" },
    { "source": 2, "target": 4, "condition":
"Failure" }
  ]
}
```


(a) Converted JSON representation for STAR *Ride Change* task



(b) STAR *Ride Change* task schema

Figure 8: Example of STAR task schema and converted JSON object