EMNLP2025 Main Conference | Long paper | Session 2 : Poster A

Beyond Task-Oriented and Chitchat Dialogues: Proactive and Transition-Aware Conversational Agents

Yejin Yoon Yuri Son Namyeong So Minseo Kim
Minsoo Cho Chanhee Park Seungshin Lee Taeuk Kim

Hanyang University, Hyundai Motor Company | Republic of Korea

Presented by Yejin Yoon

HYU 한양대학교 HANYANG UNIVERSITY HYUNDAI

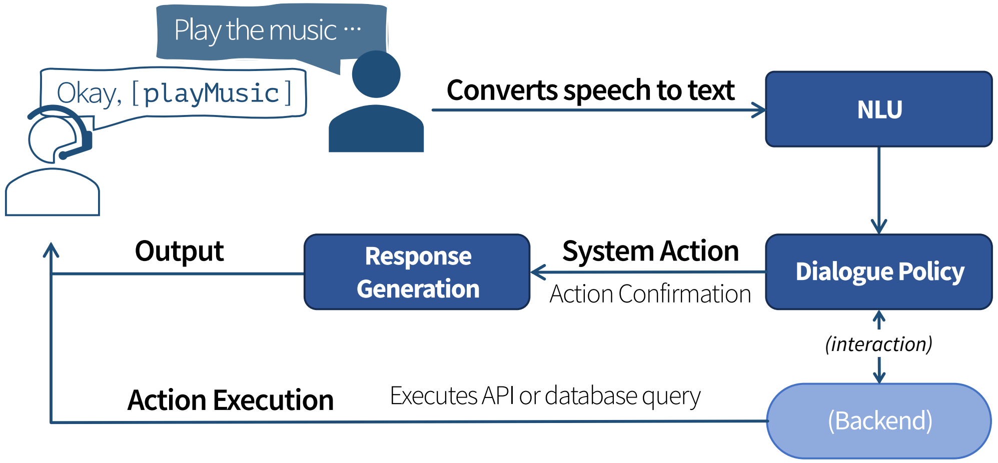
EMNLP 2025 Suzhou, China 中国苏州

Motivation

ToD vs. Chitchat → need for an integrated system

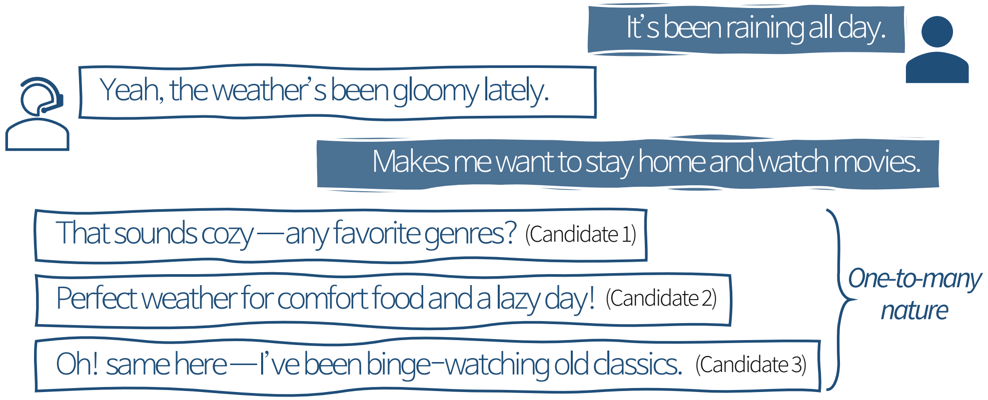
Real conversations blend both **ToD** and **Chitchat** naturally, yet current systems remain mode-bound and reactive.

Task-Oriented Dialogue (ToD)



- *Goal-driven* ; having exact goals
- *Structured* ; finite space of input and output state
- *Task-driven* ; finishing a dialogue as soon as possible

Chitchat



- *Open-ended ;* No exact goal
- *Flexible* ; infinite space of input and output state
- *Engaging* ; Sustaining a dialogue as long as possible

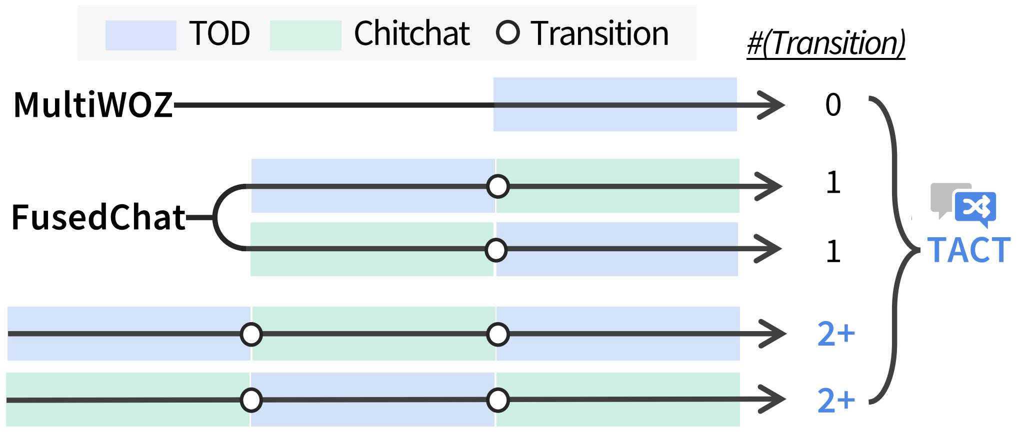
⛳ **Our Goal** : enable agents to understand and manage transitions between modes

Motivation

Benchmark Gap

- **MultiWOZ**: ToD only → no mode transition
- **FusedChat**: only one switch (ToD ↔ Chitchat) → limited dynamics
- **InterfereChat**: insert a single Chitchat turn into ToD → forced immediate return, no recoverable transitions
- **TACT**: supports multiple transitions (2+) and recoverable flows
  ; 12 distinct flow types, balanced initiators (user/agent)

[Flow diversity]



[Dataset comparison]

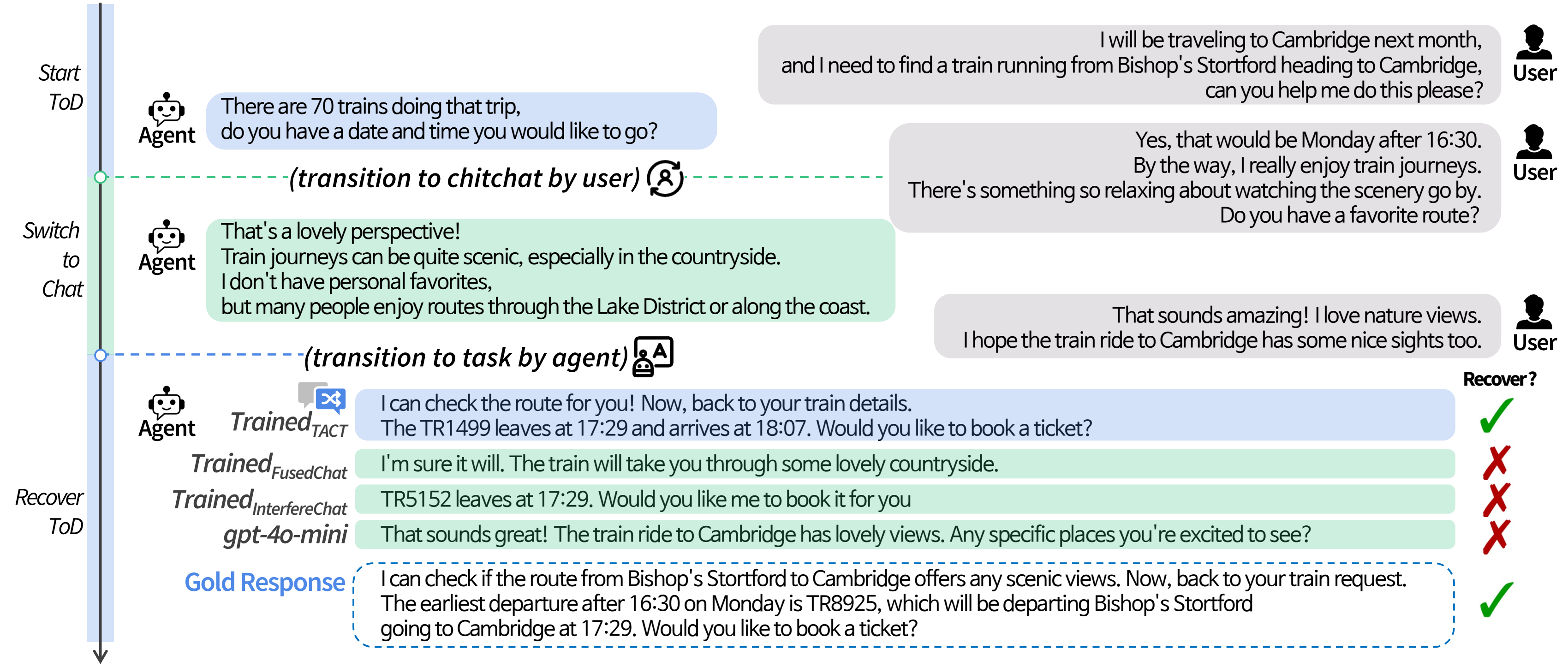
Dataset	SalesBot2.0	FusedChat	InterfereChat	TACT	
Seed	SalesBot1.0	MultiWOZ2.4	FusedChat	MultiWOZ2.2	SLURP
# Intents	6	11	11	11	50*
# Dialog	5,453	10,436	4,475	7,199	9,936
# Avg. Turn	7.71	18.36	13.58	15.04	16.42
# Avg. Switch	0.96	1	0*	1.93	2.06
# Avg. Recov.	-	0	0*	0.93	1.07
# Uniq. Flow	2	2	1	11	12
Flow Types	CT	TC, CT	T*	TCT, CTC, TCTCT, etc.	

Existing datasets can’t train agents to **return to task or handle complex flows**

Motivation

Our Goal : From reactive responders → From reactive responders to proactive flow managers

- Baseline agents (trained on other datasets or GPT-4o-mini) fail to recover once the dialogue digresses
- ***TACT*** enables transition-aware and proactive behavior across ToD–Chitchat flows.



TACT bridges the gap between ***structured*** goal completion and ***open-ended*** human dialogue

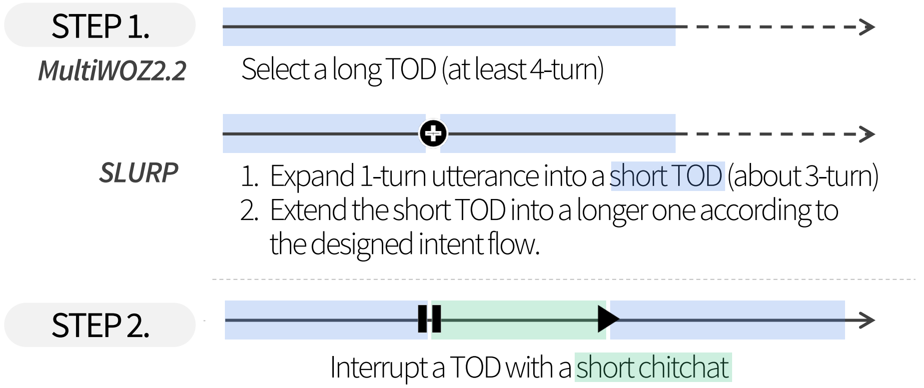
TACT Construction

Design : Generate diverse and realistic mode transitions simulating user - agent dynamics

TCT Flow *Task → Chat → Task*

Recovering from off-topic turns

1) Extract long ToD segment (≥ 4 turns)
2) Insert short chitchat block mid-task
3) Mimics user diversion and agent recovery



Enables richer **flow diversity** and **smoother context continuity** across dialogue modes

TACT Construction

Design : Generate diverse and realistic mode transitions simulating user - agent dynamics

TCT Flow

Task → Chat → Task

Recovering from off-topic turns

1) Extract long ToD segment (≥ 4 turns)
2) Insert short chitchat block mid-task
3) Mimics user diversion and agent recovery

STEP 1.

MultiWOZ2.2 — Select a long TOD (at least 4-turn)

SLURP —
1. Expand 1-turn utterance into a short TOD (about 3-turn)
2. Extend the short TOD into a longer one according to the designed intent flow.

STEP 2.

Interrupt a TOD with a short chitchat

CTC Flow

Chat → Task → Chat

Light conversation before and after the task

1) Start with short ToD (2–3 turns)
2) Add light conversation before and after
3) Simulates casual small-talk wrapping a brief task

STEP 1.

MultiWOZ2.2 — Select a short TOD (fewer than 4-turn)

SLURP — Expand 1-turn utterance into a short TOD (about 3-turn)

STEP 2.

Attach short chitchat before and after a short TOD

Enables richer **flow diversity** and **smoother context continuity** across dialogue modes

Dataset Validation

What we validate : Ensure accurate intent, coherent transitions, and natural flow

Filters low-quality dialogues and verifies transition plausibility using **3 key criteria**

Intent Accuracy

- Measures whether each user utterance's annotated intent accurately reflects the user's goal at that moment.
- Detects semantic drift or labeling errors — the intent must align with what the user is actually doing or asking.
- *Does the labeled intent truly capture what the user is trying to achieve in this turn?*

Transition Quality

- Evaluates how naturally and coherently the dialogue shifts between task-oriented and chitchat modes.
- Assesses logical consistency and context, avoiding artificial or over-assumed intent in mode switching.
- *Is the transition contextually justified and free from forced or premature intent changes?*

Dialogue Naturalness

- Captures overall conversational quality — how human-like, coherent, and satisfying the dialogue feels.
- Integrates fluency, recovery, politeness, and user satisfaction as markers of natural interaction.
- *Would this dialogue feel fluent, polite, and engaging to a real user while completing the task effectively?*

🧩 Defines human-aligned criteria for reliable dialogue validation

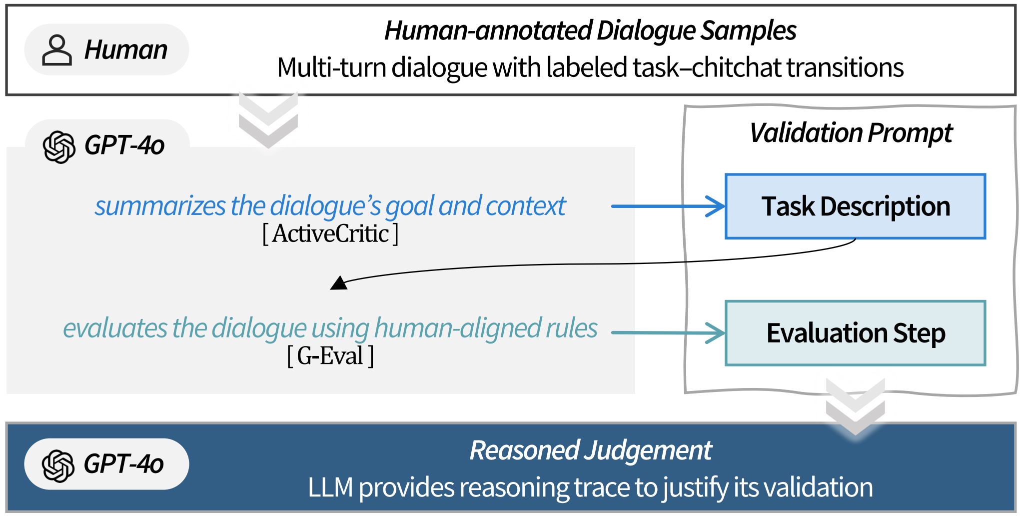
Data Validation

How we validate : Human–aligned criteria meet scalable LLM reasoning

Hybrid validation - combines **rule–based consistency checks** with **LLM–based reasoning** (Active Critic + G-Eval)

- Interpretable & scalable validation : integrates human criteria with autonomous evaluation steps
- Reasoning flow: dialogue → task description → criteria → reasoned judgement

TACT Validation Workflow



Validation Methods Comparison

Validation Approaches	Human-Annotated Criteria	LLM-Generated Task Description	LLM-Generated Evaluation Steps
Active Critic	✗	✓	✗
G-Eval	✓	✗	✓
Ours	✓	✓	✓

* *Comparison of validation methods*
: only ours integrates both human-aligned criteria and LLM reasoning.

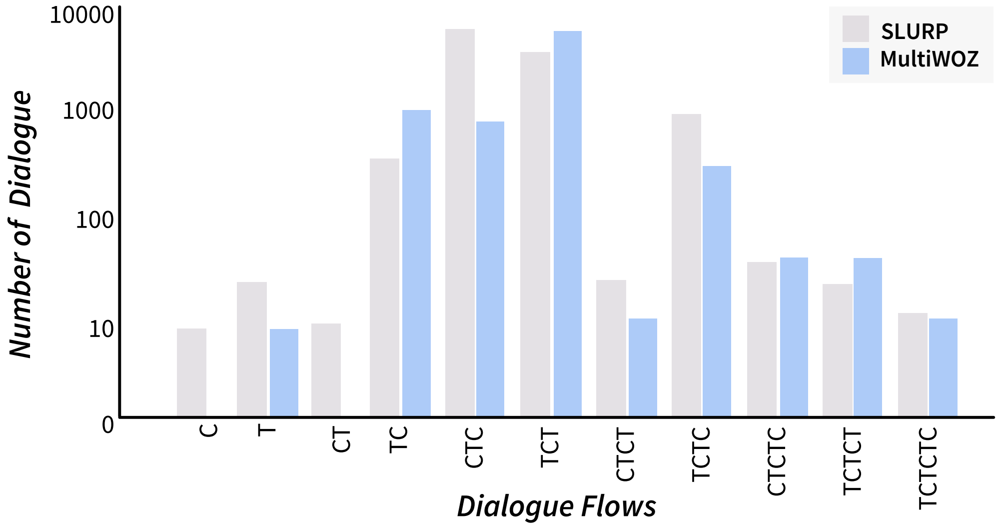
Establishes a ***unified validation pipeline*** combining human criteria and LLM reasoning

TACT Characteristics

Rich dialogue flows for realistic mode transitions

- Multiple flow types $_{\text{TCT, CTC, etc.}}$, allowing agents to experience realistic, recoverable transitions
- Recoverable structures: return to suspended mode
- Balanced initiators: both user- and agent-driven switches

[Distribution of dialogue Flow Types]



[Dataset comparison]

Dataset	SalesBot2.0	FusedChat	InterfereChat	TACT	
Seed	SalesBot1.0	MultiWOZ2.4	FusedChat	MultiWOZ2.2	SLURP
# Intents	6	11	11	11	50*
# Dialog	5,453	10,436	4,475	7,199	9,936
# Avg. Turn	7.71	18.36	13.58	15.04	16.42
# Avg. Switch	0.96	1	0*	1.93	2.06
# Avg. Recov.	-	0	0*	0.93	1.07
# Uniq. Flow	2	2	1	11	12
Flow Types	CT	TC, CT	T*	TCT, CTC, TCTCT, etc.	

TACT enables robust learning of multi−turn, recoverable transitions

Methodology

Modeling approaches for handling ToD - Chitchat transitions

1. **ICL** (GPT-4o) – zero/few-shot prompting for rapid adaptation *without training*
2. **SFT** (FnCTOD) – *supervised fine-tuning baseline* trained on full data (***FusedChat***, ***InterfereChat***, ***TACT***)
3. **Pipeline** – *separates* mode control and response generation *modules* for modular evaluation

🎯 Comparing modeling paradigms for agents to handle ToD + Chitchat + Transitions

Evaluation Metrics

Standard metrics - accuracy for tasks, quality for chitchat

Assesses both **goal completion** and **conversation quality** using standard dialogue evaluation schemes

ToD metrics

* ***Mode Selection (Acc.)***

: Checks whether the model selects the correct dialogue mode (ToD or Chitchat).

* ***Intent Detection (Acc.)***

- **Turn-level**: counts a prediction as correct for each turn.
- **Dialogue-level**: counts a dialogue as correct only if all intent predictions within it are correct.

* ***Joint Accuracy***

: Counts a turn as correct only when both the dialogue mode and intent predictions are correct.

Chitchat metrics

* ***SSI + Transition Naturalness (Win Rate)***

- Follow the **SSI framework** to measure chat quality.
  - **Sensibleness** – logical consistency of the reply
  - **Specificity** – informativeness and contextual grounding
  - **Interestingness** – engagement and diversity of tone
- Additionally, we introduce **Transition Naturalness**
  - **Transition Naturalness** – smoothness when entering or exiting the mode.

* Gemini-2.5-Pro and human annotators results are aggregated as win rates.

✅ Measures both ***what agents do*** and ***how they talk***

Evaluation Metrics

Flow-aware metrics — measuring proactive and recoverable transitions

Evaluates whether the agent can **proactively** move between modes and recover the original mode afterwards

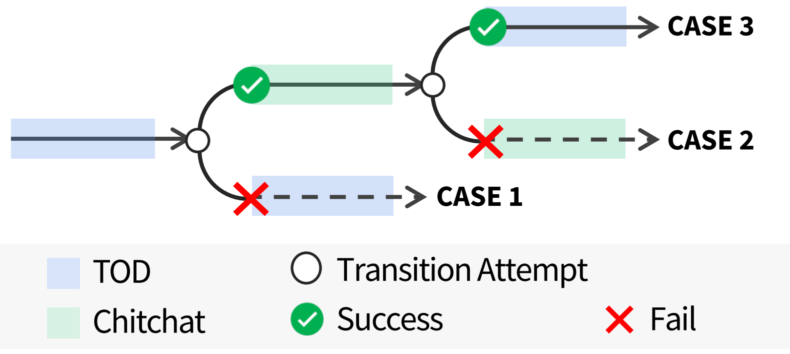
Switch – whether the ***agent*** initiates a mode change (e.g., T → C, C → T)

Recovery – whether the ***agent*** returns to the suspended task (e.g., T → C → T, C → T → C)

In **dialogue-level**, ***Switch*** and ***Recovery*** are measured by:

- **Attempt** : average number of transition trials per dialogue
  - *Switch* Attempt : count of initiated transitions (○)
  - *Recovery* Attempt : number of *Switch* Attempt - 1
- **Success** : proportion of user-accepted transitions among attempts
  - *Switch* Success : count of successful transitions (✓)
  - *Recovery* Success : number of *Switch* Success - 1

[Visualization of *Switch* & *Recovery*]



CASE 1 | 2 | 3	Attempt			Success		
Switch	+1	+2	+2	X	+1	+2
Recover	-	+1	+1	-	X	+1

🧐 "Can the ***agent*** switch and come back naturally?"

Evaluation

Quantitative Evaluation - Across Datasets

ToD Performance

- Exposure to transition-rich dialogue improves task robustness and generalization.
- The *TACT*-trained model achieves strong ToD performance and the **best** cross-dataset generalization, indicating that transition-rich supervision also strengthens task-level understanding.

		TOD						Flow			
		Mode Selection		Intent Detection		Joint Accuracy		Switch		Recovery	
Training Set	Test Set	Acc.	F1	Acc./turn	Acc./dialogue	Acc./turn	Acc./dialogue	Attempt	Success	Attempt	Success
FusedChat	MultiWOZ	98.44	76.71	93.79	69.80	93.57	68.90	0.000	0.000	-	-
	FusedChat	97.00	96.70	92.85	66.90	94.20	60.50	0.000	0.000	-	-
	InterfereChat	97.04	94.68	93.21	67.62	93.09	61.89	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	91.79	87.27	94.46	72.65	88.13	33.24	0.000	0.000	-	-
	Average	96.07	88.84	93.58	69.24	**92.25**	56.13				
InterfereChat	MultiWOZ	98.27	75.97	93.92	70.90	93.74	70.10	0.000	0.000	-	-
	FusedChat	79.92	73.43	92.85	67.10	76.19	7.10	0.000	0.000	-	-
	InterfereChat	97.63	95.72	93.28	68.44	93.34	64.34	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	79.26	58.41	93.95	71.34	84.33	34.89	0.000	0.000	-	-
	Average	88.77	75.88	**93.84**	**70.62**	84.82	35.39				
$\text{TACT}_{\text{MultiWOZ}}$	MultiWOZ	98.06	74.91	92.70	66.20	92.57	65.50	0.000	0.000	1.000	< 0.001
	FusedChat	90.63	89.05	92.57	65.70	87.08	34.10	0.160	0.008	1.000	< 0.001
	InterfereChat	97.32	95.20	92.13	64.14	92.38	59.22	0.619	0.309	0.013	0.104
	$\text{TACT}_{\text{MultiWOZ}}$	98.95	98.50	96.35	80.94	96.41	75.59	1.322	1.300	0.977	0.856
	Average	**96.24**	**89.42**	93.44	69.25	92.11	**58.60**				

Evaluation

Quantitative Evaluation - Across Datasets

Flow Performance

- Only the *TACT*-trained model records non-zero ***Switch*** and ***Recovery*** scores, showing that it can proactively transition and resume tasks across dialogue modes.
- Other models trained on single-mode datasets (***FusedChat***, ***InterfereChat***) show no recoverable flow behavior.

Training Set	Test Set	TOD: Mode Selection Acc.	TOD: Mode Selection F1	TOD: Intent Detection Acc./turn	TOD: Intent Detection Acc./dialogue	TOD: Joint Accuracy Acc./turn	TOD: Joint Accuracy Acc./dialogue	Flow: Switch Attempt	Flow: Switch Success	Flow: Recovery Attempt	Flow: Recovery Success
FusedChat	MultiWOZ	98.44	76.71	93.79	69.80	93.57	68.90	0.000	0.000	-	-
	FusedChat	97.00	96.70	92.85	66.90	94.20	60.50	0.000	0.000	-	-
	InterfereChat	97.04	94.68	93.21	67.62	93.09	61.89	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	91.79	87.27	94.46	72.65	88.13	33.24	0.000	0.000	-	-
	Average	96.07	88.84	93.58	69.24	92.25	56.13				
InterfereChat	MultiWOZ	98.27	75.97	93.92	70.90	93.74	70.10	0.000	0.000	-	-
	FusedChat	79.92	73.43	92.85	67.10	76.19	7.10	0.000	0.000	-	-
	InterfereChat	97.63	95.72	93.28	68.44	93.34	64.34	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	79.26	58.41	93.95	71.34	84.33	34.89	0.000	0.000	-	-
	Average	88.77	75.88	93.84	70.62	84.82	35.39				
$\text{TACT}_{\text{MultiWOZ}}$	MultiWOZ	98.06	74.91	92.70	66.20	92.57	65.50	0.000	0.000	1.000	< 0.001
	FusedChat	90.63	89.05	92.57	65.70	87.08	34.10	0.160	0.008	1.000	< 0.001
	InterfereChat	97.32	95.20	92.13	64.14	92.38	59.22	0.619	0.309	0.013	0.104
	$\text{TACT}_{\text{MultiWOZ}}$	98.95	98.50	96.35	80.94	96.41	75.59	1.322	1.300	0.977	0.856
	Average	96.24	89.42	93.44	69.25	92.11	58.60				

Evaluation

Quantitative evaluation - across MultiWOZ-derived datasets

When evaluated on different transition formulations (*MultiWOZ*, *FusedChat*, *InterfereChat*),
TACT-trained model exhibits **robust performance and transition-awareness**.

		TOD						Flow			
		Mode Selection		Intent Detection		Joint Accuracy		Switch		Recovery	
Training Set	**Test Set**	**Acc.**	**F1**	**Acc./turn**	**Acc./dialogue**	**Acc./turn**	**Acc./dialogue**	**Attempt**	**Success**	**Attempt**	**Success**
FusedChat	MultiWOZ	98.44	76.71	93.79	69.80	93.57	68.90	0.000	0.000	-	-
	FusedChat	97.00	96.70	92.85	66.90	94.20	60.50	0.000	0.000	-	-
	InterfereChat	97.04	94.68	93.21	67.62	93.09	61.89	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	91.79	87.27	94.46	72.65	88.13	33.24	0.000	0.000	-	-
	Average	96.07	88.84	93.58	69.24	**92.25**	56.13				
InterfereChat	MultiWOZ	98.27	75.97	93.92	70.90	93.74	70.10	0.000	0.000	-	-
	FusedChat	79.92	73.43	92.85	67.10	76.19	7.10	0.000	0.000	-	-
	InterfereChat	97.63	95.72	93.28	68.44	93.34	64.34	0.000	0.000	-	-
	$\text{TACT}_{\text{MultiWOZ}}$	79.26	58.41	93.95	71.34	84.33	34.89	0.000	0.000	-	-
	Average	88.77	75.88	**93.84**	**70.62**	84.82	35.39				
$\text{TACT}_{\text{MultiWOZ}}$	MultiWOZ	98.06	74.91	92.70	66.20	92.57	65.50	0.000	0.000	1.000	< 0.001
	FusedChat	90.63	89.05	92.57	65.70	87.08	34.10	0.160	0.008	1.000	< 0.001
	InterfereChat	97.32	95.20	92.13	64.14	92.38	59.22	0.619	0.309	0.013	0.104
	$\text{TACT}_{\text{MultiWOZ}}$	98.95	98.50	96.35	80.94	96.41	75.59	1.322	1.300	0.977	0.856
	Average	**96.24**	**89.42**	93.44	69.25	92.11	**58.60**				

Transition-aware training
→ consistent performance
across related datasets

$\text{TACT}_{\text{MultiWOZ}}$

TACT is the **only** dataset enabling proactive, transition-aware, and recoverable dialogues

Evaluation

Quantitative evaluation - across modeling methods

Performance and behavior differences across baseline modeling strategies: **ICL**, **SFT**, **Pipeline**

	TOD						Flow				Chitchat
	Mode Selection		Intent Detection		Joint Accuracy		Switch		Recovery		Overall
Method	Acc.	F1-score	Acc./turn	Acc./dialogue	Acc./turn	Acc./dialogue	Attempt	Success	Attempt	Success	Win-Rate
ICL-ZS	90.46	86.21	87.57	50.44	85.01	30.00	0.879	0.374	0.880	0.099	-
ICL-FS	91.45	88.98	84.09	40.00	86.89	36.76	**1.577**	0.865	**1.571**	0.652	-
SFT	**98.95**	**98.50**	**96.35**	**80.94**	**96.41**	75.59	1.322	1.300	0.977	0.856	23.16
Pipeline	**98.95**	**98.50**	**96.35**	**80.94**	**96.41**	75.59	1.322	1.300	0.977	0.856	24.32

1. **ICL** (Few-shot) → frequent but unstable transitions (low control, high variance)
2. **SFT** → highest ToD accuracy but weak chitchat engagement
3. **Pipeline** → stable but rigid; lacks flexibility across dialogue modes

🤷 "How to improve the naturalness of chat responses?"

Methodology

Modeling approaches for handling ToD - Chitchat transitions

1. **ICL** (GPT-4o) – zero/few-shot prompting for rapid adaptation *without training*
2. **SFT** (FnCTOD) – *supervised fine-tuning baseline* trained on full data (***FusedChat***, ***InterfereChat***, ***TACT***)
3. **Pipeline** – *separates* mode control and response generation *modules* for modular evaluation
4. **DPO** – *aligns model with human preferences* for natural tone and coherent mode recovery

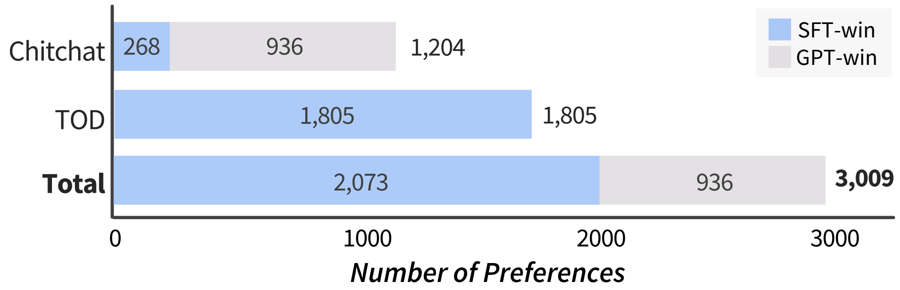
Preference Pair Construction ; Chosen / Rejected response pairs

3,009 pairs (SFT vs. GPT-4O-mini, judged by Gemini-2.5-Pro)

[Preference pair sample]

[Dialogue History] **User**: I'm looking for a hotel to stay in that's expensive and on the east side. **Agent**: Express by Holiday Inn Cambridge is on the east side and expensive. **User**: That sounds good, but can you tell me if it includes free Wi-Fi?	
Chosen Response	**Agent**: <function_call>{"function" : "find_hotel"}</function_call> Yes, it does include free Wi-Fi. Would you like me to book a room for you?
Rejected Response	**Agent**: <function_call>{"function" : "find_hotel"}</function_call> Yes, Express by Holiday Inn Cambridge includes free Wi-Fi.

[Preference Distribution]



🎯 Comparing modeling paradigms for agents to handle ToD + Chitchat + Transitions

Evaluation

Quantitative evaluation - with DPO preference alignment

Compares task, flow, and chitchat performance before and after **preference alignment**

Method	TOD						Flow				Chitchat
	Mode Selection		Intent Detection		Joint Accuracy		Switch		Recovery		Overall
	Acc.	F1-score	Acc./turn	Acc./dialogue	Acc./turn	Acc./dialogue	Attempt	Success	Attempt	Success	Win-Rate
ICL-ZS	90.46	86.21	87.57	50.44	85.01	30.00	0.879	0.374	0.880	0.099	-
ICL-FS	91.45	88.98	84.09	40.00	86.89	36.76	**1.577**	0.865	**1.571**	0.652	-
SFT	**98.95**	**98.50**	**96.35**	**80.94**	**96.41**	75.59	1.322	1.300	0.977	0.856	23.16
SFT-DPO	98.82	98.32	96.03	80.00	96.21	**75.74**	1.343	**1.322**	0.977	**0.859**	**40.86**
Pipeline	**98.95**	**98.50**	**96.35**	**80.94**	**96.41**	75.59	1.322	1.300	0.977	0.856	24.32

4. **DPO**
   - Maintains ToD accuracy (~96%) while improving response quality
   - +17pp improvement in chitchat win rate (40.9%)
   - Strongest flow-awareness: highest switch & recovery success (>1 per dialogue)

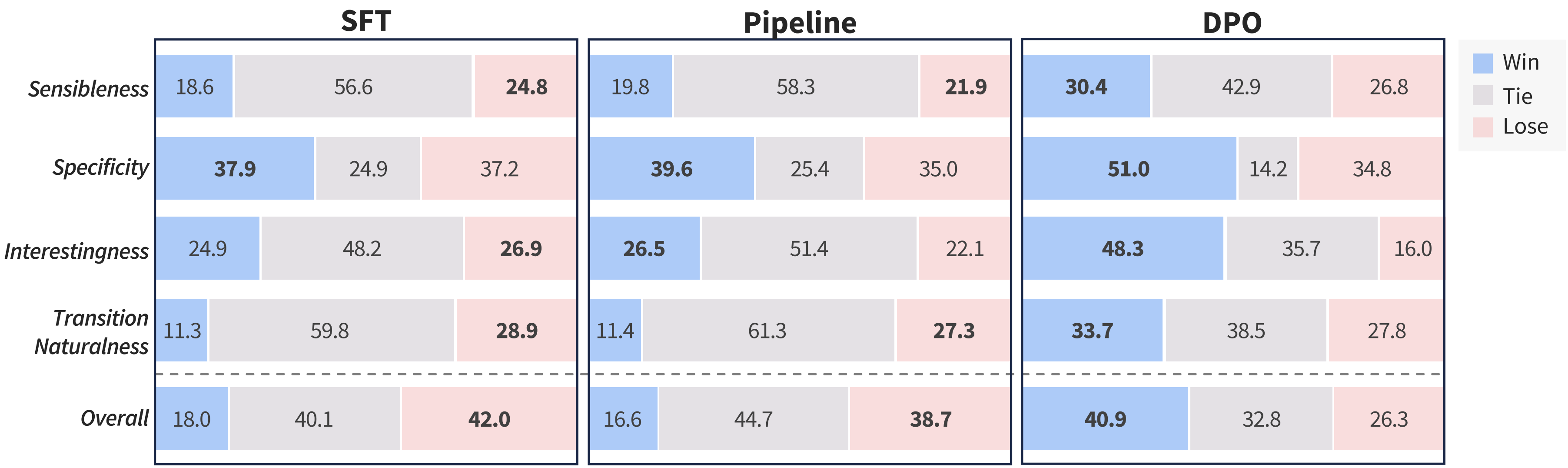
👍 DPO achieves **best** dialogue-level accuracy (75.7%) with **strongest** flow-awareness metrics.

Evaluation

Preference-Based Evaluation - automated LLM judge results

Chat Performance

- Evaluated with an LLM-based judge (Gemini-2.5-Pro) over the full test set.
- **DPO** consistently outperforms **SFT** and **Pipeline** across all qualitative dimensions.
- Gains +20 pp in Interestingness and +22 pp in Transition Naturalness, indicating that alignment training improves conversation flow and engagement.



👀 Even an LLM can tell — DPO makes dialogue flow naturally.

Evaluation

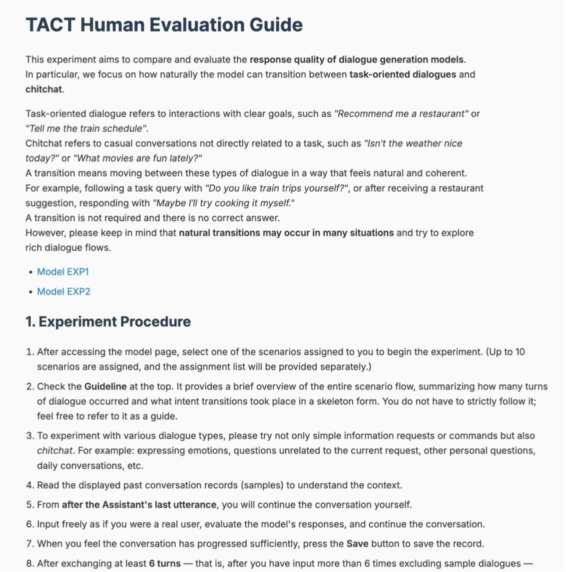
Preference-Based Evaluation - validating conversational naturalness beyond automated LLM

Human evaluation setup — assessing dialogue quality with human annotators

[Human Evaluation Setup]

- Conducted to evaluate the overall quality and naturalness of dialogues, including mode transitions.
- Evaluated 77 randomly sampled dialogues by 10 *human annotators.
  * including both NLP practitioners and general users
- Used the same 4 qualitative criteria as LLM-based evaluation (**SSI** + **Transition Naturalness**).
- Results are aggregated as pairwise win rates for each dimension.

[Annotation Guide]



TACT Human Evaluation Guide

This experiment aims to compare and evaluate the **response quality of dialogue generation models**.
In particular, we focus on how naturally the model can transition between **task-oriented dialogues** and **chitchat**.

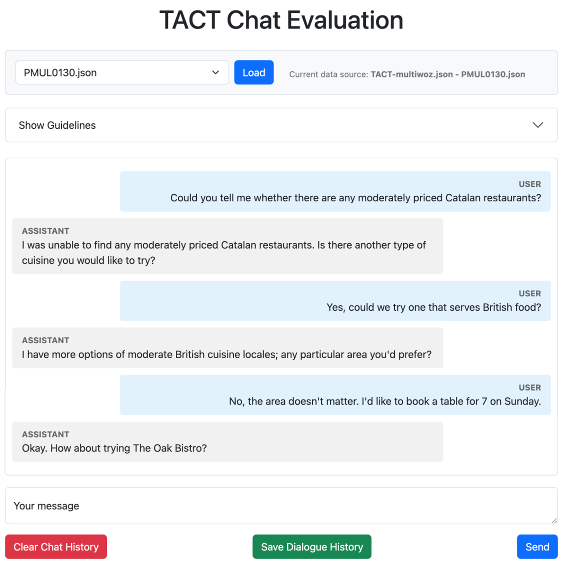
Task-oriented dialogue refers to interactions with clear goals, such as *"Recommend me a restaurant"* or *"Tell me the train schedule"*.
Chitchat refers to casual conversations not directly related to a task, such as *"Isn't the weather nice today?"* or *"What movies are fun lately?"*
A transition means moving between these types of dialogue in a way that feels natural and coherent.
For example, following a task query with *"Do you like train trips yourself?"*, or after receiving a restaurant suggestion, responding with *"Maybe I'll try cooking it myself."*
A transition is not required and there is no correct answer.
However, please keep in mind that **natural transitions may occur in many situations** and try to explore rich dialogue flows.

- Model EXP1
- Model EXP2

1. Experiment Procedure

1. After accessing the model page, select one of the scenarios assigned to you to begin the experiment. (Up to 10 scenarios are assigned, and the assignment list will be provided separately.)
2. Check the **Guideline** at the top. It provides a brief overview of the entire scenario flow, summarizing how many turns of dialogue occurred and what intent transitions took place in a skeleton form. You do not have to strictly follow it; feel free to refer to it as a guide.
3. To experiment with various dialogue types, please try not only simple information requests or commands but also *chitchat*. For example: expressing emotions, questions unrelated to the current request, other personal questions, daily conversations, etc.
4. Read the displayed past conversation records (samples) to understand the context.
5. From **after the Assistant's last utterance**, you will continue the conversation yourself.
6. Input freely as if you were a real user, evaluate the model's responses, and continue the conversation.
7. When you feel the conversation has progressed sufficiently, press the **Save** button to save the record.
8. After exchanging at least **6 turns** — that is, after you have input more than 6 times excluding sample dialogues —

[Evaluation Interface]



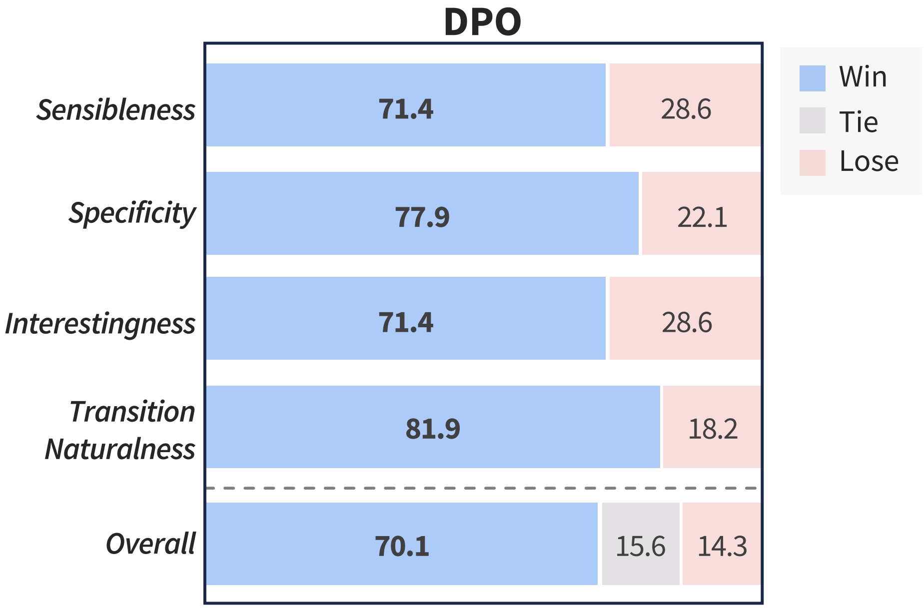
🤝 Complementing automated LLM judging with human preference validation

Evaluation

Preference-Based Evaluation - human preference confirms DPO's conversational naturalness

Chat Performance

- *Human Evaluation* : Humans perceive DPO as most natural and engaging
- *Overall* : DPO wins over 70 % of pairwise comparisons across 77 dialogues



[Response Comparison Example]

[Dialogue History]	
User: Can you help me find a train? I'll be traveling on Wednesday. **Agent**: What is your departure point and what time would you like to depart? **User**: Oh, of course. Sorry. I'm leaving out of Leicester and going into Cambridge. **Agent**: Is there a specific time you want to leave or arrive by? I have 2,828 trains to pick from. I need to narrow this down. **User**: I want to arrive by 16:15. **Agent**: That narrowed it down to 10 options. What time would you like to leave? **User**: I don't have a preference on leaving time. By the way, I really enjoy train journeys. There's something so relaxing about watching the scenery go by. Do you have a favorite route? **Agent**: Oh, I completely agree! Train journeys can be quite scenic. The route from Leicester to Cambridge has some lovely views, especially as you approach the countryside. **User**: That sounds nice! I love seeing fields and little villages. It makes the trip feel more like an adventure.	
SFT Response	**Agent**: How about TR8076? It leaves at 05:09. (Ignores the chitchat and returns directly to task.)
SFT-DPO Response	**Agent**: Absolutely! It adds a nice touch to the journey. Now, back to your train options. I would recommend TR 7853. It leaves at 13:09 and arrives at 14:54. (Most natural transition—acknowledges the chitchat and resumes the task smoothly.)
ICL-FS Response	**Agent**: Absolutely, the countryside views are charming! Now, let's finalize your train booking. Would you like me to proceed with one of the options that arrive by 16:15? (Acknowledges the chitchat but lacks concrete task grounding.)

🏆 Human evaluators favor DPO for its fluent, engaging, and context-sensitive transitions.

Ablation

Ablation Study — understanding how agents recover and adapt across dialogue flows

Recovery Success Analysis

suggests contextual flexibility rather than repetition

- Only ~34 % of recoveries return to the **same intent**
- 66 % start a **new but relevant** task

Method	Success	intent ✓	intent ✗
ICL-FS	0.652	33.89	66.11
SFT	0.856	34.58	65.42
SFT-DPO	0.859	34.23	65.77

Flow Type Comparison

reveals flow-dependent robustness

- Strongest on TCT flows
- Weak on CTC
  - * less agent-initiated data
- Over-triggering in TC flows → mode confusion

Flow type (# Dialogues)	TOD: Mode Selection Acc.	TOD: Mode Selection F1	TOD: Intent Detection Acc./turn	TOD: Intent Detection Acc./dialogue	TOD: Joint Accuracy Acc./turn	TOD: Joint Accuracy Acc./dialogue	Flow: Switch Attempt	Flow: Switch Success	Flow: Recovery Attempt	Flow: Recovery Success
$\text{TACT}_{\text{MutliWOZ}}$ (680)										
TCT (533)	99.35	99.04	96.64	82.18	97.18	81.24	1.390	1.372	1.006	0.977
CTC (38)	97.06	96.51	96.30	92.11	97.06	81.58	0.421	0.342	0.313	0.000
TC (74)	95.35	97.99	95.35	74.32	95.18	64.86	1.419	1.378	1.000	0.069
TCTC (27)	97.35	96.94	97.39	85.19	96.46	70.37	1.222	1.185	0.958	0.958
Others (7)	98.36	98.29	97.22	87.50	96.72	75.00	0.500	0.500	1.000	1.000
$\text{TACT}_{\text{SLURP}}$ (1,790)										
TCT (618)	99.41	99.12	94.18	72.98	95.28	72.01	0.974	0.964	1.019	0.985
CTC (907)	97.61	97.00	91.69	88.20	96.10	79.82	1.756	0.821	1.041	0.037
TC (60)	91.19	91.01	76.68	31.67	85.32	30.00	1.450	0.550	1.304	0.071
TCTC (174)	96.06	96.03	91.17	71.26	92.80	60.34	1.793	1.167	1.616	0.169
Others (31)	85.13	84.87	82.28	45.16	84.39	16.13	1.645	0.677	1.393	0.107

👀 Recovery ≠ repetition → contextual flexibility matters
🔧 Need balanced supervision for ambiguous transitions

Conclusion

1 Transition-Aware Dataset

- ***TACT*** enables recoverable, multi-turn dialogues
- Unifies ToD and Chitchat into a single transition-aware framework

2 Flow-Aware Metrics

- Introduce **Switch** and **Recovery** to measure mode adaptability

3 Preference-Aligned Model

- **DPO** improves conversational smoothness and engagement
- Achieves more natural and context-aware transitions

4 Reproducibility & Open Source

- All datasets (with scripts) and validation templates available on GitHub
- https://github.com/HYU-NLP/TACT

Thank You

Yejin Yoon

HYU NLP Lab.

Dept. of Computer Science
Hanyang University, South Korea

stillwithyou@hanyang.ac.kr