

# I know you are different! Towards Persona Driven Knowledge-infused Dialogue Assistant

Shifali Agrahari<sup>1\*</sup>† Moushumi Mahato<sup>2\*</sup> Abhisek Tiwari<sup>2</sup> Javaid Nabi<sup>2</sup>

<sup>1</sup>Indian Institute of Technology, Guwahati, India

<sup>2</sup>Samsung Research Institute, Bengaluru, India

a.shifali@iitg.ac.in {moushumi.m, abhisek.t, javaid.nabi}@samsung.com

## Abstract

Despite advances in large language models (LLMs), Task-Oriented Dialogue (TOD) systems often fall short in delivering personalized, context-rich responses, especially in low-resource, code-mixed, and multimodal settings like Hinglish (Hindi-English). To bridge this gap, we introduce *HiVisTask*, the first Hinglish multimodal, multidomain, persona-based TOD dataset that captures user-assistant interactions across text and visual modalities. We also propose *G<sup>3</sup>TOD*, a generalizable framework that enhances personalization using three structured knowledge graphs: entity context, user persona, and commonsense reasoning, all extracted from conversation history. Extensive experiments with LLMs (e.g., LLaMA3.2, Phi3, GPT4, Mistral7b, Qwen3, Gemma3) show that *G<sup>3</sup>TOD* consistently outperforms both standard and ablated baselines. We observe substantial gains across evaluation metrics (both quantitative: BLEU  $\uparrow$  and qualitative: Human Eval  $\uparrow$ ) over existing models. The observed improvements strongly underscore the value of structured and selective contextualization in generating personalized and engaging responses.

## 1 Introduction

Task-Oriented Dialogue (TOD) systems facilitate the completion of specific user tasks through natural language interaction, such as booking flights, setting reminders, or making restaurant reservations (Chen et al., 2017). These systems have become increasingly vital across domains such as customer service, personal assistants, e-commerce, and enterprise applications. However, building effective TOD systems requires large volumes of diverse and high-quality training data, which can be both time-consuming and expensive to collect

\*Equal contribution.

†Work done as an intern.



Figure 1: Demonstration of how *G<sup>3</sup>TOD* generates different responses based on user personas in the *HiVisTask* dataset.

(Zhang et al., 2020). The advent of Large Language Models (LLMs) such as LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), and ChatGPT (OpenAI et al., 2024) has transformed the landscape of NLU. These models perform well in open-domain dialogue but remain underutilized in TOD settings, as they are rarely optimized for structured, goal-driven interactions. Moreover, several open challenges remain such as: (a) LLMs handle casual conversations well, but personalization in TOD is underexplored. Persona-grounded agents enhance relevance and engagement, yet task-based systems lack this integration (Mendonça et al., 2023). (b) Most TOD models rely only on dialogue history, ignoring structured external knowledge like user goals, preferences, or commonsense. Incorporating such knowledge via graphs can improve response quality (Zamani et al., 2023). (c) Existing TOD datasets are primarily in English or Chinese. Hinglish (Hindi + English),

widely used in India, lacks dedicated TOD datasets with persona-aware and multimodal features (Khan et al., 2020; Kumar et al., 2023). Notably, Hindi is the fourth most spoken language worldwide, as reported by Wikipedia. (d) Real-world tasks often require visual inputs (e.g., hotel images, food menus), yet TOD systems are mostly unimodal. They fail to integrate multimodal cues with user preferences or commonsense (Ohashi et al., 2024).

To bridge these gaps, we propose a dataset and a framework:

(I) *HiVisTask*, a large-scale, multimodal, and persona-grounded Hinglish dataset for task-oriented dialogue. To the best of our knowledge, *HiVisTask* is the first dataset to combine code-mixed language, visual modality, and user persona traits in real-time user-assistant interactions. It enables research on personalization, code-mixed, multimodality, and LLM reasoning in low-resource TOD.

(II) *G<sup>3</sup>TOD (Graph<sup>3</sup> for TOD)*, a novel zero-shot framework that enhances LLM-driven TOD by injecting structured knowledge in the form of three distinct knowledge graphs: (a) *Entity Context Graph* (b) *Persona Graph* (c) *Commonsense Graph*. An example of how *G<sup>3</sup>TOD* responds to different user personas on same utterance is shown in Fig. 1. In this work, we aim to investigate the following two research questions: **(RQ1)** What is the impact of all three different types of graphs on LLM reasoning? Which provides better personalized grounding for response generation: the persona knowledge graph ( $\mathcal{G}_P$ ) or the textual persona description? **(RQ2)** Which of the existing LLMs are more flexible towards personalization with minimal adaptation? Do the existing LLMs show bias towards any particular persona type?

**Key Contributions.** The main contributions of this work are three-fold:

- We present *HiVisTask*, the first multimodal, persona-grounded, Hinglish task-oriented dialogue dataset, containing 2,700 dialogues and 42,000 utterances across 7 diverse personas.
- We propose *G<sup>3</sup>TOD*, a zero-shot prompting framework that enriches LLM reasoning with structured graphs combining entity context, persona, and commonsense knowledge.
- *G<sup>3</sup>TOD* outperforms LLM baselines in task success, persona alignment, and fluency, validating our graph-augmented design.

## 2 Related Work

In the following paragraphs, we have summarized the relevant works & highlighted the research gap.

**End to End TOD Dataset.** Dialogue (TOD) datasets are commonly built via user-system interactions (Lee et al., 2025; Zhao et al., 2025; Moradshahi et al., 2023; Williams et al., 2014; Raux, 2005) or Wizard-of-Oz setups (Kelley, 1984; Wen et al., 2016; Asri et al., 2017; Budzianowski et al., 2018; Byrne et al., 2019; Andreas et al., 2020), but face scalability and control issues. Low-resource languages further lack standardized NLU benchmarks (Adelani et al., 2021; Goyal et al., 2022; Moghe et al., 2022). Cross-lingual TOD methods explore code-mixed and zero/few-shot learning (Moghe et al., 2021; Li et al., 2018; Siddhant et al., 2020). Existing Hinglish and multimodal datasets are chat-oriented or domain-limited. *HiVisTask* introduces a persona-grounded, multidomain, multimodal code-mixed TOD dataset for diverse tasks, detailed in Table 1.

**Knowledge Graph and Triplet-based Persona Modeling.** Task-oriented dialogue (TOD) systems often model dialogue context as a sequence of past utterances (Kulhánek et al., 2021; Hosseini-Asl et al., 2020), using generative models like GPT-2 (Radford et al., 2019; Mohapatra et al., 2020) or prompting-based LLMs (Li et al., 2022; Zhang et al., 2023; Ulmer et al., 2024). Knowledge graphs (KGs) enhance context understanding via entity reasoning (Wan et al., 2023; Wu et al., 2022; Yang et al., 2020; Rony et al., 2022; He et al., 2020). Triplet embeddings (subject, relation, object) further support response generation (Ma et al., 2021b; Chen et al., 2025), while Subgraph-RAG (Li et al., 2024) enables LLM-based reasoning over subgraphs. Commonsense and persona modeling via KGs like ConceptNet (Havasi et al., 2007) is effective in English (Bosselut et al., 2019; Sun et al., 2025), but remains underexplored in code-mixed languages like Hinglish (Speer et al., 2017).

**Personalized Dialogue Generation.** Incorporating personality traits is key for personalized dialogue (Chen et al., 2024). While explicit persona data (e.g., roles, traits, styles) enhances interaction (Song et al., 2021; Zheng et al., 2020; Zhang et al., 2020; Liu et al., 2020; Lu et al., 2023), dialogue history provides a scalable alternative with implicit cues and fewer privacy concerns. Methods like DHAP (Ma et al., 2021a) and MSP (Tang et al., 2023) retrieve relevant history for genera-

Dataset	Turn	Task	Size	Lang	Image	CodeMix	Task	Chat	Persona
MultiWOZ 2.0 (Ohashi et al., 2024)	Multi	Hotel	10k+	EN	✗	✗	✓	✗	✗
SNIPS (Coucke et al., 2018)	Single	Booking	13k	EN	✗	✗	✓	✗	✗
CamRest676 (Quan and Xiong, 2019)	Multi	Restaurant	676	EN	✗	✗	✓	✗	✗
FRAMES (Asri et al., 2017)	Multi	Travel	1.3k	EN	✗	✗	✓	✗	✗
DSTC2 (Banerjee et al., 2018)	Multi	Restaurant	Varies	HI,EN	✗	✓	✓	✗	✗
MMDialog (Feng et al., 2022)	Multi	Open	1M	EN	✓	✗	✗	✓	✗
MASAC (Bedi et al., 2021)	Multi	TV serial	8.6k	Hinglish	✗	✓	✗	✓	✗
X-RiSAWOZ (Moradshahi et al., 2023)	Multi	Travel etc.	151k	EN,HI KO,FR	✗	✓	✓	✗	✗
PersonaLens (Zhao et al., 2025)	Multi	Restaurant	122k	EN	✗	✗	✓	✗	✓
PicPersona (Lee et al., 2025)	Multi	Map,Wiki	18k	EN	✗	✗	✓	✗	✓
<b>HiVisTask(Ours)</b>	Multi	Multi-domain	21k	Hinglish	✓	✓	✓	✓	✓

Table 1: Comparison of the *HiVisTask* dataset with existing task-oriented dialogue datasets. *HiVisTask* is a multi-turn Hinglish dataset featuring multimodal inputs, code-mixed language, and persona-grounded dialogues across multiple real-world domains. Here, EN = English, FR = French, HI = Hindi, and KO = Korean.

tion, but recent efforts (Wang et al., 2023a; Yang et al., 2025b,a) focus on chat-oriented data, with limited progress in TOD and challenges in persona extraction and retrieval efficiency.

### 3 Dataset

We present *HiVisTask* dataset, consisting of user–assistant conversations generated through LLMs. Comparison between *HiVisTask* and existing TOD datasets is shown in Table 1.

#### 3.1 Dataset Creation

To construct the *HiVisTask* dataset, we simulate realistic, persona-grounded, multimodal, and task-oriented dialogues in Hinglish that reflect diverse user needs across real-world tasks. The dataset creation pipeline comprises three key stages.

**(a) Persona Definition.** We select personas based on demographic factors such as age, occupation, and language fluency, which guide the assistant’s tone, response style, and decision-making strategies. These personas, including *Housewife*, *IT Professional*, *College Student*, *Senior Citizen*, *Small Business Owner*, *Freelancer*, and *School Teacher*, are crafted to capture unique preferences, daily tasks, and contextual constraints (e.g., budget, time, technical fluency). Persona details are mentioned in Appendix Fig.7.

**(b) Prompt Designing.** In this stage, we design few-shot prompts for each persona and task scenario to guide the LLM in generating context-aware, appropriate responses. The prompts are composed of the following components: a clear task description outlining the user’s intent, relevant contextual history, and multimodal cues such

as visual inputs when applicable. These carefully crafted prompts simulate realistic user inputs to ensure the assistant responds appropriately in a code-mixed Hinglish setting, considering the user’s persona and context. Some example prompts are shown in Appendix Sec. A.3, Table 9.

**(c) Dialogue Generation.** This stage leverages the prompts designed in the previous step to guide the LLM in generating dialogues turn-by-turn, where each assistant response is shaped by the persona, task scenario, and multimodal input. Each generated turn consists of a Hinglish utterance from a persona, followed by a response from the assistant that is contextually relevant and grounded in the persona’s goals. The LLM also incorporates multimodal cues, such as interpreting images or describing task-related actions. Sample images from the dialogues is shown in Appendix Sec. A.1, Fig. 6. The images are sourced from two primary sources: (i) publicly available platforms (e.g., Pinterest) for generic images such as decorative tables, and (ii) original captures and screenshots that depict real-world scenarios, such as restaurant bills. Dialogue data is generated using GPT-3.5-Turbo (Chen et al., 2023) and LLaMA-2-13B (Touvron et al., 2023). Examples from the dataset is shown in Appendix Table 10.

#### 3.2 Dataset Quality and Characteristics

The *HiVisTask* dataset undergoes manual evaluation by three annotators (Appendix Sec. A.2) to ensure it meets quality standards. Annotators review each dialogue turn-by-turn, making edits where necessary to ensure responses are grounded in the given background information, user persona,

Metrics	Count
# of Personas	7
# of Dialogues	2700
# of Turns	21000
# of Utterances	42000
Avg # of utterances per dialogue	15.50
Avg # of words in user utterance	18.58
Avg # of words in assistant utterance	23.36

Table 2: Overall statistics of the *HiVisTask* Dataset

intent, actions, and overall task flow. While most responses are accurate, some tend to be too generic or resemble casual chat rather than task-specific replies. Details on annotation and quality assurance is mentioned in Appendix Sec. A.2.

The *HiVisTask* dataset includes the following characteristics (i) **Task-Oriented Multimodal Dialogues**: The conversations are centered around common assistant tasks such as booking flight, scheduling meeting, setting reminders, etc. including visual inputs. (ii) **Code-Mixed Hinglish**: All conversations are in Hinglish, reflecting real-world scenarios where users communicate in a blend of Hindi and English. (iii) **Persona-Driven Interactions**: Each dialogue reflects one of the seven defined personality archetypes (Appendix Table 16, 17), influencing the assistant’s linguistic style and user’s communication tone. (iv) **Diverse User Behavior**: The dataset captures a wide range of user behaviors, including five behavioural tone types: polite, formal, neutral, impatient, and humorous. The proportion of each tone style ranges from 9% (minimum) to 22% (maximum) across personas, with an overall mean distribution of around 15% tone variation throughout the dataset. The overall statistics and the domain-wise distribution of the proposed dataset is shown in Table 2 and Table 3 respectively.

## 4 Methodology

The proposed  $G^3TOD$  zero-shot framework, illustrated in Fig. 2, has four prominent stages each described in the subsequent sub-sections.

### 4.1 Dialogue Encoding and Knowledge Graph Construction

The task-oriented dialogue is represented as a sequence  $D = \{u_1, a_1, \dots, u_t\}$ , where  $u_i$  denotes the user utterance at turn  $i$  for  $i = 1, \dots, t$ , and  $a_j$  denotes the assistant response at turn  $j$  for  $j = 1, \dots, t-1$ . We collect ~80% of user–assistant conversations, ordered from oldest to newest, to build

Domain	Count	Domain	Count
Food	2,763	Communication	1,090
Finance	2,846	Shopping	402
Home	915	Government	539
Health	1,207	Weather	2,381
Work	1,004	Education	1,204
Social	1,184	Emergency Support	579
Family	348	Travel	4,538

Table 3: Domain-wise Distribution of the Proposed Dataset.

an overall dialogue history  $\mathcal{H}_u$  for each user. From each dialogue history  $\mathcal{H}_u$ , we extract three different types of knowledge graphs that capture various aspects of the user’s interactions. The remaining 20% of the most recent conversations are reserved for testing and validation, and are excluded from graph construction.

**Entity Context Knowledge Graph.** Firstly, we extract named entities from the dialogue history  $\mathcal{H}_u$  using a Named Entity Recognition (NER) model (Singh et al., 2018; He et al., 2024). The corresponding relations between these entities are inferred using a large language model (LLM), and together they form the entity context graph  $\mathcal{G}_E$ , represented as triplets in equation 1.

$$\mathcal{G}_E = \{(e_i, r_{ij}, e_j) \mid i, j = 1, \dots, N\}, \quad (1)$$

where  $e_i$  and  $e_j$  are entities,  $r_{ij}$  is the relation between them and  $N$  is the total number of distinct entities extracted from  $\mathcal{H}_u$ .

**Persona Knowledge Graph.** We leverage a large language model (LLM) (Wang et al., 2023b) to extract entities and relations that would reflect information about the user’s lifestyle, preferences, and personal traits. These are structured into the persona graph  $\mathcal{G}_P$ , shown in equation 2:

$$\mathcal{G}_P = \{(s_i, p_i, o_i) \mid i = 1, \dots, N\}, \quad (2)$$

where each triplet consists of a subject  $s_i$ , a predicate  $p_i$ , and an object  $o_i$ , and  $N$  is the number of such triplets.

**Commonsense Knowledge Graph.** We also extract commonsense-level knowledge from the dialogue history  $\mathcal{H}_u$  by leveraging external resources such as ConceptNet (Havasi et al., 2007). This knowledge helps capture implicit and inferred relations between concepts mentioned in the dialogue. The resulting graph  $\mathcal{G}_C$ , shown in equation 3:

$$\mathcal{G}_C = \{(s_i, r_i, o_i) \mid i = 1, \dots, N\} \quad (3)$$

where  $s_i$  is the subject,  $r_i$  is the relation (as defined in ConceptNet (Krawczyk et al., 2015; Speer

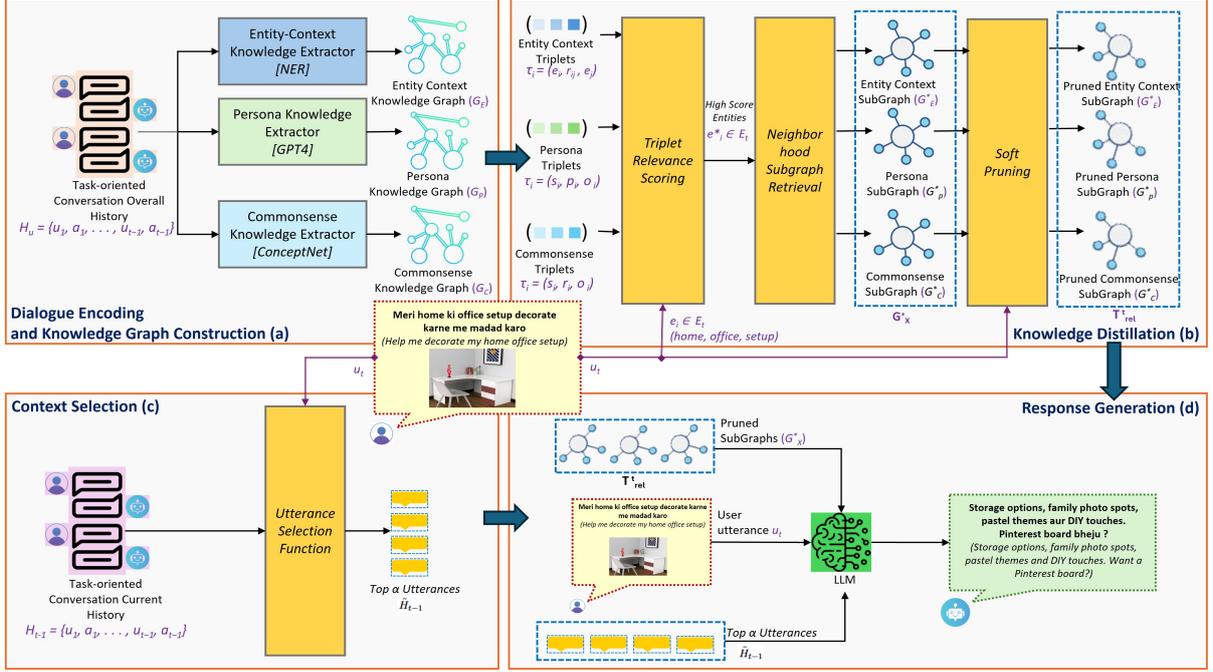


Figure 2:  $G^3TOD$  framework. (a) Dialogue Encoding and Knowledge Graph Construction of three graphs: entity context knowledge graph, persona knowledge graph and commonsense knowledge graph; (b) Knowledge Distillation, consisting of triplet relevance threshold scoring, neighborhood subgraph retrieval and soft pruning extraction; (c) Context Selection; and (d) Response Generation.

et al., 2017), and  $o_i$  is the object. Here,  $N$  is the total number of commonsense triplets.

Table 20 provides a detailed example highlighting the role of each knowledge graph. It illustrates how the Entity, Persona, and Commonsense graphs capture important relations, user preferences, and world knowledge, respectively, enabling more accurate and personalized dialogue responses.

For multimodal utterances, during the construction of triplets for both the Entity Context Knowledge Graph ( $\mathcal{G}_E$ ) and the Persona Knowledge Graph ( $\mathcal{G}_P$ ), we use LLM that is supplied with the corresponding visual input whenever an image is available, in addition to the textual dialogue context and the associated prompt instructions. This multimodal conditioning enables the LLM to extract information from the image that may not be explicitly mentioned in the text, allowing the model to jointly interpret textual and visual cues during entity extraction and persona inference. As a result, the generated triplets are more context-aware. In contrast, the Commonsense Knowledge Graph ( $\mathcal{G}_C$ ) is constructed exclusively from textual input.

## 4.2 Knowledge Distillation

To ensure that the LLM receives only the most relevant information, we perform a triplet-wise factor-

ization and filtering step for each of the knowledge graphs. This process retains only the triplets that are contextually relevant to the current user utterance  $u_t$  and its preceding dialogue history  $\mathcal{H}_{t-1}$ .

**Triplet Relevance Scoring.** For each extracted entity  $e_i \in \mathcal{E}_t$  in the current utterance, we define a relevance function  $Rel(e_i, \tau_i)$  that computes the semantic alignment between the entity  $e_i$  and a given triplet  $\tau_i \in \mathcal{G}_X$  where  $\mathcal{G}_X \in \{\mathcal{G}_E, \mathcal{G}_P, \mathcal{G}_C\}$ . Each triplet node  $\tau_i$  in the graph is scored for its contextual compatibility with the current dialogue.

**Neighborhood Subgraph Retrieval.** To focus on local and meaningful knowledge, we use a breadth-limited subgraph retrieval function  $Subgraph(\cdot, \beta)$ . For each graph  $\mathcal{G}_X \in \{\mathcal{G}_E, \mathcal{G}_P, \mathcal{G}_C\}$ , we extract a subgraph around each entity  $e_i \in \mathcal{E}_t$  using a depth parameter  $\beta$ , which controls the size of the contextual neighborhood. The retrieved subgraph is defined as:

$$\mathcal{G}_X^* = \bigcup_{e_i \in \mathcal{E}_t} Subgraph(\mathcal{G}_X, e_i, \beta) \quad (4)$$

Here,  $\mathcal{G}_X^* \in \{\mathcal{G}_E^*, \mathcal{G}_P^*, \mathcal{G}_C^*\}$  denotes the filtered subgraph corresponding to each knowledge graph.  $\beta$  determines contextual depth, and  $e_i$  represents entities identified in the current user utterance  $u_t$ .

**Soft Pruning Extraction.** To further refine the knowledge context, we use a Soft Pruning Extrac-

Model		Variant	PPL	BLEU	ROUGE-L	METEOR	BS-F1	ConSim	R-Len	PA	PC
LLama-3.2 (Touvron et al., 2023)	Baselines	Vanilla	13.55	2.94	37.89	66.12	0.67	0.34	10.28	-	-
		Self-Reflection	13.20	5.10	41.80	67.95	0.69	0.36	10.55	-	-
		Dynamic Self-Reflection	12.90	6.80	44.70	69.88	0.71	0.39	11.00	-	-
	Ablations	with utterance selection	12.80	5.32	42.37	68.34	0.70	0.37	10.65	12.89	18.12
		with all triplets	<b>12.15</b>	7.32	45.13	70.45	0.72	0.40	11.30	11.45	19.85
		$\mathcal{G}_E$	24.62	20.34	70.56	80.25	0.83	0.82	16.50	14.45	18.98
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	23.90	21.02	73.11	<b>82.24</b>	0.84	<b>0.81</b>	16.75	15.97	20.98
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		22.40	<b>21.93</b>	<b>75.12</b>	81.08	<b>0.85</b>	0.80	<b>17.08</b>	18.96	22.28	
Phi-3(Abdin et al., 2024)	Baselines	Vanilla	14.10	3.65	35.64	65.11	0.65	0.32	9.80	-	-
		Self-Reflection	13.80	4.50	37.80	66.85	0.67	0.34	10.10	-	-
		Dynamic Self-Reflection	13.20	5.90	40.50	68.80	0.69	0.36	10.60	-	-
	Ablations	with utterance selection	13.50	4.67	38.45	64.56	0.69	0.36	10.30	5.23	11.23
		with all triplets	<b>12.90</b>	6.57	43.33	69.12	0.70	0.38	10.80	3.24	9.96
		$\mathcal{G}_E$	25.10	19.00	<b>68.23</b>	79.22	0.81	0.81	15.90	11.23	13.45
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	24.50	<b>20.14</b>	68.13	0.80	<b>0.83</b>	<b>0.82</b>	16.30	14.35	15.12
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		24.00	19.23	68.00	<b>81.43</b>	0.81	0.81	<b>16.70</b>	18.35	14.56	
Gemma-3(Team et al., 2025)	Baselines	Vanilla	12.80	4.56	37.69	67.12	0.68	0.35	10.70	-	-
		Self-Reflection	12.40	5.60	39.80	68.90	0.70	0.37	11.00	-	-
		Dynamic Self-Reflection	11.90	6.80	42.60	70.40	0.72	0.39	11.40	-	-
	Ablations	with utterance selection	<b>11.10</b>	6.12	35.12	70.22	0.72	0.39	11.20	8.34	14.23
		with all triplets	12.50	7.12	44.12	71.23	0.74	0.72	11.80	5.87	12.23
		$\mathcal{G}_E$	23.20	9.12	53.08	72.12	0.78	0.80	17.00	9.23	11.98
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	22.60	13.56	66.34	<b>79.12</b>	0.83	0.79	17.30	14.67	17.23
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		22.10	<b>18.67</b>	<b>68.13</b>	79.03	<b>0.84</b>	<b>0.81</b>	<b>17.80</b>	21.34	22.35	
mistral-7b(Jiang et al., 2023)	Baselines	Vanilla	12.30	4.56	40.23	68.19	0.69	0.36	10.90	-	-
		Self-Reflection	11.90	5.40	42.00	70.10	0.71	0.38	11.20	-	-
		Dynamic Self-Reflection	11.60	6.70	44.50	71.90	0.73	0.40	11.60	-	-
	Ablations	with utterance selection	11.70	5.12	41.87	71.82	0.74	0.40	11.50	8.23	9.35
		with all triplet	<b>11.40</b>	7.13	50.12	72.87	0.75	0.42	11.80	6.99	13.24
		$\mathcal{G}_E$	22.50	12.01	64.98	75.98	0.77	<b>0.81</b>	17.50	11.24	11.98
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	22.00	16.98	<b>78.45</b>	78.23	<b>0.78</b>	0.75	17.90	17.24	21.45
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		21.50	<b>20.24</b>	75.33	<b>79.34</b>	0.71	0.79	<b>18.30</b>	19.24	23.98	
Qwen3(Yang et al., 2024)	Baselines	Vanilla	12.70	5.34	41.12	68.09	0.70	0.37	10.90	-	-
		Self-Reflection	12.40	6.10	43.00	69.20	0.72	0.39	11.20	-	-
		Dynamic Self-Reflection	12.10	7.20	45.30	70.90	0.74	0.41	11.50	-	-
	Ablations	with utterance selection	12.40	6.34	44.23	69.56	0.72	0.39	11.20	9.22	13.56
		with all triplets	<b>11.80</b>	8.12	47.34	71.56	0.74	0.42	11.70	8.24	9.23
		$\mathcal{G}_E$	23.00	18.21	72.98	81.45	0.83	0.84	16.80	13.98	15.35
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	22.50	20.13	75.45	82.67	0.85	0.86	17.20	17.09	22.13
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		22.00	<b>21.23</b>	<b>78.02</b>	<b>83.97</b>	<b>0.85</b>	<b>0.83</b>	<b>17.60</b>	22.12	18.35	
GPT4(OpenAI et al., 2024)	Baselines	Vanilla	11.00	5.45	44.21	70.23	0.72	0.38	11.50	-	-
		Self-Reflection	10.70	7.80	48.10	72.30	0.74	0.41	11.90	-	-
		Dynamic Self-Reflection	10.20	9.90	52.00	74.80	0.77	0.44	12.50	-	-
	Ablations	with utterance selection	10.30	9.12	52.45	75.22	0.78	0.44	10.40	11.23	11.22
		with all triplets	<b>9.90</b>	11.23	56.12	77.56	0.80	0.46	12.90	12.98	19.34
		$\mathcal{G}_E$	20.50	14.23	76.45	81.21	0.83	0.75	18.20	21.44	21.34
	Proposed	$\mathcal{G}_E + \mathcal{G}_P$	17.00	19.23	78.23	84.23	0.85	0.86	18.60	24.04	22.22
$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$		19.70	<b>23.45</b>	<b>79.96</b>	<b>86.99</b>	<b>0.87</b>	<b>0.90</b>	<b>29.10</b>	<b>24.22</b>	<b>23.45</b>	

Table 4: Comparison of different models and model variants on *HiVisTask* dataset across multiple evaluation metrics, refer to Sec. 5. More details at Appendix B.2.

tion (SPE) mechanism. SPE selectively identifies prior dialogue turns from  $\mathcal{H}_{t-1}$  and aligns them with most relevant subgraphs. This contextual alignment ensures the extracted triplets support both the current dialogue turn and the user’s intent. Each subgraph  $\mathcal{G}_X^*$  may contain many triplets, which can exceed the LLM’s input budget. To prevent overfitting and stay within the context limit, we apply soft pruning using relevance thresholding. Each triplet  $\tau_k \in \mathcal{G}_X^*$  is linearized into a sentence  $s_k$ , and the semantic similarity  $\delta_k \in [0, 1]$  between  $s_k$  and utterance  $u_t$  is computed using cosine similarity. We then apply a relevance threshold  $\delta$  to prune out low-signal triplets:

$$\mathcal{T}_X = \{\tau_k \in \mathcal{G}_X^* \mid \delta_k \geq \delta\} \quad (5)$$

The final set of relevant knowledge triplets is:  $\mathcal{T}_{\text{rel}}^t = \mathcal{G}_E^* \cup \mathcal{G}_P^* \cup \mathcal{G}_C^*$ . This filtered set  $\mathcal{T}_{\text{rel}}^t$  contains triplets from all three graphs (Entity, Persona, Commonsense) that are most semantically aligned with the current utterance  $u_t$ . (For details, see Appendix Subsection D.2, Table 18).

### 4.3 Context Selection

Not all previous utterances are equally relevant for generating a response. Therefore, we apply an utterance selection step that prunes the dialogue history  $\mathcal{H}_{t-1}$  to retain only those turns that contribute semantically or topically to the current interaction. This subset, denoted  $\mathcal{H}_{t-1}$ , is used alongside  $\mathcal{T}_{\text{rel}}^t$  during response generation. We retain the top- $\alpha$

scoring utterance to maintain the context.

$$\tilde{\mathcal{H}}_{t-1} = \{u_i \in \mathcal{H} \mid \text{Rel}(u_i, u_t) \text{ is among top-}\alpha\} \quad (6)$$

The optimal value of  $\alpha$  is selected via a greedy search over a predefined range of values, aiming to maximize a development objective (e.g., BLEU, F1, or response accuracy).

#### 4.4 Response Generation

We pass the selected utterance  $u_t$ , the filtered history  $\tilde{\mathcal{H}}_{t-1}$ , and the reduced triplet set  $\mathcal{T}_{\text{rel}}^t$  to a LLM via a structured prompt. The LLM is then tasked with computing  $\mathcal{R}_t$ ,

$$\mathcal{R}_t = \text{LLM}(u_t, \tilde{\mathcal{H}}_{t-1}, \mathcal{T}_{\text{rel}}^t) \quad (7)$$

where  $\mathcal{R}_t$  denotes the generated assistant response. This response is thus informed by dialogue context, personalized user knowledge, and commonsense reasoning.

## 5 Experimental Details

We conduct experiments for evaluating our proposed  $G^3TOD$  zero-shot framework using the *HiVisTask* dataset and an open-source out-of-domain dataset *MediTOD* (Saley et al.). Our dataset is partitioned into train and test sets with a standard 80:20 ratio. The train split is utilized exclusively for constructing the three knowledge graphs (Entity Context, Persona, and Commonsense), while model evaluation is carried out directly on the test set. For entity and relation extraction, we use IndicNER (Mhaske et al., 2022) to identify named entities, while GPT-4 (OpenAI et al., 2024) is employed as the LLM to infer semantic relations among entities in the Entity Context Knowledge Graph and to generate triplets for the Persona Knowledge Graph (Chen et al., 2023). It is also used to generate triplets from images where ever available. Commonsense knowledge is integrated from ConceptNet (Speer et al., 2017). To measure similarity between triplets and utterances, we apply the all-MiniLM-L6-v2 model (all-MiniLM-L6-v2) with cosine similarity. The relevant hyperparameter and configurations are summarized in Appendix Table 8 in Sec. B.1. Details of LLM Prompt mentioned in Appendix Fig. 8.

**Baseline.** We benchmark  $G^3TOD$  zero-shot framework against three baseline prompting strategies: *Vanilla*, *Self-Reflection* (Shinn et al., 2023), *Dynamic Self-Reflection* (Wang et al., 2023c). We measure the performance of various

LLMs Models such as LLama-3.2 (*Llama-3.2-3B-Instruct*) (Grattafiori et al., 2024), Phi-3 (*Phi-3-medium-4k-instruct-bnb-4bit*) (Abdin et al., 2024), Gemma-3 (*Gemma-3-4b-it-GGUF*) (Team et al., 2025), Mistral-7b (*mistral-7b-instruct-v0.2-bnb-4bit*) (Jiang et al., 2023), Qwen3 (*Qwen3-1.7B*) (Yang et al., 2024), and GPT4 (OpenAI et al., 2024), assessed in their default configurations to establish a reference for their core generation capabilities. Details of LLMs in Appendix Table 12.

**Evaluation Metrics.** The LLM models are evaluated using perplexity (PPL) (Brown et al., 1992) for fluency, BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for n-gram and semantic overlap, BERTScore-F1 (Zhang et al., 2019) (BS-F1) for embedding-based similarity, and context similarity (ConSim) (Biberman, 1994) to assess contextual coherence. We also explore dialogue-specific metrics that includes response length (R-Len), persona alignment (PA), which measures how well the response aligns with the user’s persona, and persona consistency (PC), which reflects how consistently the model maintains persona traits across dialogue turns (details of PA, PC Appendix Sec. B.2.1).

## 6 Results and Discussion

### 6.1 Results

Table 4 shows results on our proposed dataset *HiVisTask* and Table 5 shows results on out-of-domain dataset *MediTOD*.  $G^3TOD$  consistently outperforms all baseline models across evaluation metrics. Despite a minor increase in perplexity, it achieves the highest BLEU, ROUGE-L, METEOR, BERTScore-F1, and ConSim scores. Incorporating the full set of knowledge graphs entity context  $\mathcal{G}_E$ , persona  $\mathcal{G}_P$ , and commonsense  $\mathcal{G}_C$  leads to notable quality gains. BLEU improves substantially, reflecting better lexical overlap with reference responses. ROUGE-L nearly doubles, indicating stronger content coverage, while METEOR sees a 16–25% boost in semantic and grammatical quality. BERTScore-F1 and ConSim rise by 20–30%, showing improved meaning preservation and contextual grounding. Additionally, inclusion of persona knowledge  $\mathcal{G}_P$  notably enhances personalization, increasing Persona Alignment (PA) to 24.22% and Persona Consistency (PC) to 23.45%—surpassing GPT4 with only  $\mathcal{G}_E$  by 2.78% and 2.11%, respectively. Details about the metrics is given in

Model		Variant	PPL	BLEU	ROUGE-L	METEOR	BS-F1	ConSim	R-Len	PA	PC
GPT4(OpenAI et al., 2024)	Baselines	Vanilla	16.05	2.05	21.22	51.32	0.42	0.18	9.03	-	-
		Self-Reflection	17.23	3.04	29.23	60.21	0.51	0.25	11.01	-	-
		Dynamic Self-Reflection	15.67	4.10	32.45	70.12	0.60	0.31	12.60	-	-
	Ablations	with utterance selection	17.00	3.25	29.85	61.05	0.52	0.26	11.25	-	-
		with all triplets	16.85	3.30	30.10	61.42	0.53	0.27	11.40	-	-
		$\mathcal{G}_E$	15.20	4.50	34.80	72.55	0.64	0.41	13.40	-	-
		$\mathcal{G}_E + \mathcal{G}_P$	14.65	5.05	36.25	74.00	0.67	0.47	13.85	-	-
<b>G<sup>3</sup>TOD</b>	Proposed	$\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$	13.90	6.12	39.30	77.60	0.72	0.55	14.65	-	-

Table 5: Comparison of  $G^3TOD$  and other GPT4 model variants on out-of-domain *Medi*TOD (Saley et al.) dataset.

Metric	Fluency	Coherence	Engagingness	Correctness	Overall
Krippendorff’s ( $\alpha$ )	0.71	0.68	0.65	0.69	0.72
Fleiss’ ( $\kappa$ )	0.67	0.63	0.60	0.65	0.69

Table 6: Inter-annotator agreement scores using Krippendorff’s Alpha ( $\alpha$ ) and Fleiss’ Kappa ( $\kappa$ )

pha (Krippendorff, 2011) and Fleiss’ kappa (Fleiss, 1971), shown in Table 6. The inter-annotator agreement scores indicate substantial agreement for Fluency and Correctness of KG, while Coherence shows moderate consistency across annotators.

Appendix B.2 and Sec. D.1.

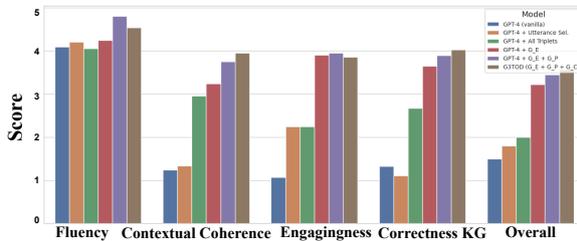


Figure 3: Human evaluation results across multiple metrics, including Fluency, Contextual Coherence, Engagingness, Correctness with Knowledge Grounding (KG), and Overall quality. Each bar colour corresponds to a specific model configuration: blue represents GPT-4 (vanilla), orange denotes GPT-4 + Utterance Selection, green corresponds to GPT-4 + All Triplets, red indicates GPT-4 +  $\mathcal{G}_E$ , purple represents GPT-4 +  $\mathcal{G}_E + \mathcal{G}_P$ , and brown denotes the proposed G<sup>3</sup>TOD ( $\mathcal{G}_E + \mathcal{G}_P + \mathcal{G}_C$ ) model.

**Human Evaluation Results.** To further assess  $G^3TOD$ ’s performance, we conduct a human evaluation on 150 test conversations from each persona-specific dataset. Three independent researchers (not involved as authors) evaluate 50 samples per model across different configurations, with model identities hidden to avoid bias. Each response is rated on five criteria: *Fluency* (grammatically correct and natural), *Contextual Coherence*, *Engagingness*, *Correctness KG*, and *Overall Performance*, using a 5-point Likert scale. Detailed descriptions of these metrics are provided in Appendix Sec. B.3. The evaluation results, shown in Fig. 3, reveal that the full  $G^3TOD$  setup—leveraging  $\mathcal{G}_E$ ,  $\mathcal{G}_P$ , and  $\mathcal{G}_C$  knowledge—achieves the highest overall scores, highlighting the effectiveness of multi-graph grounding in improving TOD quality. We measure inter-annotator agreement using Krippendorff’s al-

## 6.2 Findings

Based on the experimental findings, we report the following answers (with evidence) to our investigated research questions (RQs).

**(RQ1) What is the impact of all three different types of graphs on LLM reasoning? Which provides better personalized grounding for response generation: the persona knowledge graph ( $\mathcal{G}_P$ ) or the textual persona description?** As shown in Table 4, using the full set of graphs **entity context graph ( $\mathcal{G}_E$ )**, **persona graph ( $\mathcal{G}_P$ )**, and **commonsense graph ( $\mathcal{G}_C$ )** leads to consistent improvements across lexical, semantic, and personalization metrics.  $\mathcal{G}_E$  provides contextual grounding by capturing entities relevant to the conversation.  $\mathcal{G}_P$  offers personalized insights by modeling user-specific traits, preferences, and habits.  $\mathcal{G}_C$  enhances the model’s ability to reason about implicit knowledge, improving naturalness and coherence. The persona knowledge graph ( $\mathcal{G}_P$ ) provides stronger personalized grounding than textual persona description as mentioned in Fig. 4 and Table 13. Textual descriptions often lack specificity and are harder for LLMs to parse reliably, leading to less focused and generic outputs. **(RQ2) Which of the existing LLMs are more flexible towards personalization with minimal adaptation? Do the existing LLMs show bias towards any particular persona type?** In Table 4, larger models like GPT4 and Qwen3 consistently outperforms smaller models, even models such as LLaMA-3.2 and Mistral exhibit notable gains in persona accuracy (PA, PC). This indicates that structured personalization aids both high- and low-resource models. We evaluate  $G^3TOD$  across seven distinct personas shown in Table 7 and observe significant variations in perfor-

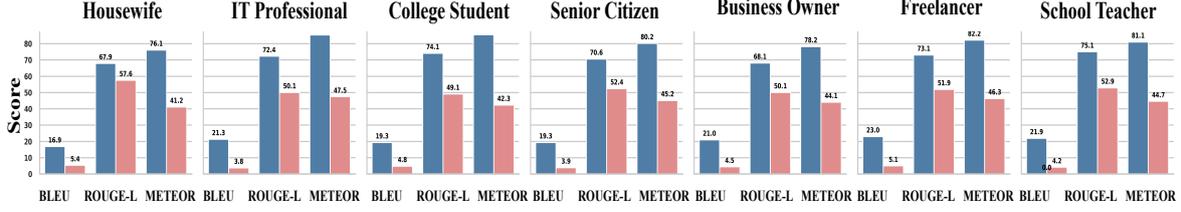


Figure 4: Comparison results on the *HiVisTask* dataset showing the impact of graph persona (Blue colour denote) and textual persona (Red colour denote) on various automatic evaluation metrics. For all metrics refer Appendix Table 13.

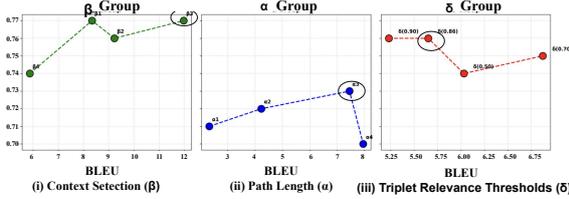


Figure 5: Results on the *HiVisTask* dataset showing the impact of hyperparameters (i) Context Selection ( $\beta$ ), (ii) Path Length ( $\alpha$ ), and (iii) Triplet Relevance Thresholds ( $\delta$ ) on BLEU scores.

mance across different metrics, indicating persona-specific behavior. Notably, the *IT Professional* and *Freelancer* personas achieve the highest ConSim scores (0.87), suggesting strong contextual alignment in responses. In contrast, personas such as *Senior Citizen*, *Housewife*, and *Business Owner* show comparatively lower performance across several metrics (Table 7). These disparities suggest that  $G^3TOD$  adapts differently to each persona, potentially due to implicit biases or imbalances in the LLMs. Examples are provided in Table 17.

## 7 Analysis

We further analyze the impact of various parameters such as context selection ( $\beta$ ), path length ( $\alpha$ ) and triplet relevance threshold ( $\delta$ ) on all personas. The results are presented in Fig. 5 and detailed in Appendix Table 15. Further, we also provide details of case study (Appendix Sec. C.3, Fig. 9) and the failure cases (Appendix Sec. C.2, Table 19). This detailed analysis leads to the following three key observations: (a) Optimal performance across all evaluation metrics is achieved with parameter settings  $\beta_3$  (context selection ( $\beta$ ) = 3),  $\alpha_3$  (path length ( $\alpha$ ) = 3), and  $\delta_{0.86}$  (triplet relevance threshold ( $\delta$ ) = 0.86). (b) The  $G^3TOD$  model exhibits varying performance across different user personas; details mentioned in Sec. 6.2.

## 8 Conclusion

In this work, we propose  $G^3TOD$ , a zero-shot framework that advances task-oriented dialogue (TOD) by leveraging pruned knowledge graphs and contextually aligned utterances. Alongside, we introduce *HiVisTask*, the first Hinglish multimodal, multidomain, persona-aware TOD dataset. Our findings show that representing structured knowledge as relational triplets significantly enhances dialogue quality over unstructured text. Experiments across diverse models demonstrate consistent gains in relevance, coherence, and personalization under both automatic and human evaluation. Moreover, structured knowledge facilitates stronger memory retention, richer context modeling, and more precise user adaptation. By unifying entity, persona, and commonsense knowledge,  $G^3TOD$  establishes a new benchmark for adaptive TOD systems in code-mixed, multimodal environments.

## Limitations

While our approach shows strong potential, it also presents a few minor limitations that invite future exploration: (i) Triplet Dependency: The approach relies on accurate and complete triplet extraction, which can be challenging in colloquial Hinglish with implicit meaning. For example, "*bhook lagi hai, kuch order karna hai*" may not map correctly to (*User, feels, hungry*). This can be improved with context-sensitive extraction models or paraphrase-based augmentation. (ii) Token Limitations: In some cases, even pruned prompts may approach LLM token limits. Advanced compression techniques can address this without major architectural changes. (iii) Static Triplet Weighting: Equal weighting across triplets may under-represent context. Lightweight solutions like adaptive attention mechanisms can make weighting more dynamic. Overall, these aspects represent refinements that open up promising avenues for future research in

graph reasoning and conversational adaptability.

## Ethics Statement

This work relies on a synthetically generated dataset, eliminating concerns related to user privacy or inadvertent exposure of personal data. To ensure ethical data creation, all prompts used during generation were carefully crafted to avoid eliciting sensitive, offensive, or culturally inappropriate content. These prompts explicitly excluded topics related to race, religion, politics, and other potentially controversial nuances. Additionally, persona traits were defined in a neutral and constructive manner to prevent the reinforcement of stereotypes. To ensure authenticity, three annotators manually reviewed and post-edited the generated dialogues, aiming to reflect realistic interactions between users and assistants. Annotators were instructed not to include any sensitive information during this process. For human evaluation, we employed three independent researchers (not involved in the paper’s authorship), unaware of model identities to reduce bias and ensure fairness in scoring. All contributors were fairly compensated. While the dataset and framework are intended for research and development of task-oriented dialogue systems, we advise caution when deploying such models in real-world or sensitive domains such as healthcare, legal assistance, mental health, or financial services where inaccuracies or lack of context awareness could lead to harmful consequences or ethical concerns.

## References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, and 1 others. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. *arXiv preprint arXiv:1806.05997*.
- Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.
- Yoram Biberian. 1994. A context similarity measure. In *Machine Learning: ECML-94: European Conference on Machine Learning Catania, Italy, April 6–8, 1994 Proceedings 7*, pages 49–63. Springer.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Leonardo 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*.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*.
- 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.
- Jiale Chen, Xuelian Dong, Wenxiu Xie, Tao Gong, Fu Lee Wang, and Tianyong Hao. 2025. Span attention for entity-consistent task-oriented dialogue response generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.
- Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *arXiv preprint arXiv:2303.00293*.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, and 1 others. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.
- Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. John Benjamins Philadelphia, PA.
- Li He, Hayilang Zhang, Jie Liu, Kang Sun, and Qing Zhang. 2024. Zero-shot relation triplet extraction via knowledge-driven llm synthetic data generation. In *International Conference on Intelligent Computing*, pages 329–340. Springer.
- Zhenhao He, Yuhong He, Qingyao Wu, and Jian Chen. 2020. Fg2seq: Effectively encoding knowledge for end-to-end task-oriented dialog. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8029–8033. IEEE.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.
- Shahid Nawaz Khan, Maitree Leekha, Jainendra Shukla, and Rajiv Ratn Shah. 2020. Vyaktivt: A multimodal peer-to-peer hindi conversations based dataset for personality assessment. In *2020 IEEE sixth international conference on multimedia big data (BigMM)*, pages 103–111. IEEE.
- Marek Krawczyk, Rafal Rzepka, and Kenji Araki. 2015. Extracting conceptnet knowledge triplets from japanese wikipedia. In *Proceedings of the 21st Annual Meeting of The Association for Natural Language Processing*, pages 1052–1055.
- Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.
- Jonáš Kulhánek, Vojtech Hudecek, Tomáš Nekvinda, and Ondrej Dušek. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*, 26:532–535.
- Shivani Kumar, Md Shad Akhtar, Tanmoy Chakraborty, and 1 others. 2023. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. *arXiv preprint arXiv:2310.13080*.
- Jihyun Lee, Yejin Jeon, Seungyeon Seo, and Gary Lee. 2025. Picpersona-tod: A dataset for personalizing utterance style in task-oriented dialogue with image persona. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7937–7958.
- Mufei Li, Siqi Miao, and Pan Li. 2024. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. *arXiv preprint arXiv:2410.20724*.
- Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

- Zekun Li, Wenhui Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. *arXiv preprint arXiv:2210.04185*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*.
- Zhenyi Lu, Wei Wei, Xiaoye Qu, XianLing Mao, Danyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. *arXiv preprint arXiv:2310.18342*.
- Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021a. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564.
- Zhiyuan Ma, Jianjun Li, Zezheng Zhang, Guohui Li, and Yongjing Cheng. 2021b. Intention reasoning network for multi-domain end-to-end task-oriented dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2273–2285.
- John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Dialogue quality and emotion annotations for customer support conversations. *arXiv preprint arXiv:2311.13910*.
- Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy V, and Anoop Kunchukuttan. 2022. Naamapadam: A large-scale named entity annotated data for indic languages. *arXiv preprint arXiv:2212.10168*.
- Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2022. Multi3nlu++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. *arXiv preprint arXiv:2212.10455*.
- Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. *arXiv preprint arXiv:2109.13620*.
- Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2020. Simulated chats for building dialog systems: learning to generate conversations from instructions. *arXiv preprint arXiv:2010.10216*.
- Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gaël de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, and 1 others. 2023. X-risawoz: High-quality end-to-end multilingual dialogue datasets and few-shot agents. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2773–2794.
- Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. 2024. Jmultiwoz: A large-scale japanese multi-domain task-oriented dialogue dataset. *arXiv preprint arXiv:2403.17319*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. *Gpt-4 technical report*. *Preprint*, arXiv:2303.08774.
- 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 of the Association for Computational Linguistics*, pages 311–318.
- Jun Quan and Deyi Xiong. 2019. Effective data augmentation approaches to end-to-end task-oriented dialogue. In *2019 International Conference on Asian Language Processing (IALP)*, pages 47–52. IEEE.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Antoine Raux. 2005. Let’s go public! taking a spoken dialog system to the real world. In *Proc. of Interspeech 2005*.
- Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialogk: Knowledge-structure aware task-oriented dialogue generation. *arXiv preprint arXiv:2204.09149*.
- Vishal Vivek Saley, Goonjan Saha, and Rocktim Jyoti Das. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.
- Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8854–8861.

- Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Named entity recognition for hindi-english code-mixed social media text. In *Proceedings of the seventh named entities workshop*, pages 27–35.
- Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2025. Comet: Dialog context fusion mechanism for end-to-end task-oriented dialog with multi-task learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10541–10553.
- Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv preprint arXiv:2305.11482*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*.
- Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. 2023. Multi-grained knowledge retrieval for end-to-end task-oriented dialog. *arXiv preprint arXiv:2305.10149*.
- Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *arXiv preprint arXiv:2305.11792*.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.
- Yu Wang, Zhiwei Liu, Jianguo Zhang, Weiran Yao, Shelby Heinecke, and Philip S Yu. 2023c. Drdt: Dynamic reflection with divergent thinking for llm-based sequential recommendation. *arXiv preprint arXiv:2312.11336*.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.
- Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.
- Jie Wu, Ian G Harris, and Hongzhi Zhao. 2022. Graphmemdialog: Optimizing end-to-end task-oriented dialog systems using graph memory networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11504–11512.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Guisong Yang, Decao Ma, Jiahao Yuan, Na Li, and Zied Bouraoui. 2025a. A parallel network encoding dialog history template for end-to-end task-oriented dialog. *CCF Transactions on Pervasive Computing and Interaction*, pages 1–11.
- Mingtao Yang, See-Kiong Ng, and Jinlan Fu. 2025b. Omnidialog: An omnipotent pre-training model for task-oriented dialogue system. *IEEE Transactions on Audio, Speech and Language Processing*.
- Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2010.01447*.
- Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, and 1 others. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. *arXiv preprint arXiv:2305.09067*.
- Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.
- Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. Personalens: A benchmark for personalization evaluation in conversational ai assistants. *arXiv preprint arXiv:2506.09902*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

## Appendix

This section presents complementary materials, including detailed dataset creation procedures, annotator information, experiment details, and a discussion on the conversation samples to enhance the reader’s understanding of the work.

### A Dataset Details

#### A.1 Image samples in dataset

We present screenshots illustrating real-world scenarios referenced in the dataset, shown in Fig. 6.

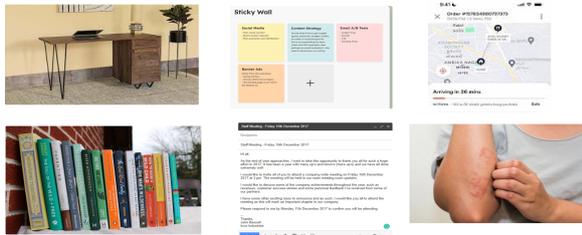


Figure 6: Example of Image and screenshot of events present in *HiVisTask* Dataset

#### A.2 Quality assurance and annotation for proposed dataset

To ensure the quality and realism of persona-consistent dialogues in the *HiVisTask* dataset, we employ three annotators, proficient in both Hindi and English, with prior experience in dialogue evaluation and linguistic annotation. These annotators were responsible for manually reviewing each LLM-generated dialogue to assess its fluency, coherence, task relevance, and alignment with the given persona and scenario. When issues were identified, such as unnatural phrasing, inconsistencies, or lack of task orientation, the annotators post-edited the responses to improve quality while preserving the original intent. In addition to review and editing, the annotators verified the assistant’s adherence to intent, task flow, and multimodal inputs. Each dialogue was annotated with structured metadata, including persona IDs, text and image. For multimodal turns, annotators ensured that referenced images were contextually grounded and accurately interpreted. A semi-automated annotation pipeline, supplemented by manual validation, was used to maintain consistency across dialogues.

Annotators were compensated at 2 rupees per utterance, with detailed guidelines and regular feedback to ensure alignment with the realism, personalization goal and task orientation goal of the dataset.

#### A.3 Prompts used to generate *HiVisTask* dataset

To construct the *HiVisTask* dataset, we curated a set of structured prompts that simulate realistic, persona-grounded, task-oriented dialogues in Hinglish across both textual and visual modalities. Each prompt was designed to reflect one of seven distinct personas—*College Student*, *IT Professional*, *Senior Citizen*, *Freelancer*, *Housewife*, *Small Business Owner*, and *School Teacher*—each with specific backgrounds, preferences, and communication styles (details in Fig. 7). The prompts incorporated a task goal (e.g. booking a cab, ordering food, scheduling an appointment), a concise persona description (e.g. age, profession, technology familiarity), and an optional visual context (e.g. menu image, catalog, or location map). Instructions were included to ensure that the responses were generated in Hinglish, maintaining the natural code-mixed patterns commonly observed in everyday interactions. These prompts were used to elicit multi-turn dialogues from large language models (LLMs), where the assistant adapted its responses based on both the task and the persona characteristics, enabling the generation of rich, personalized, and multimodal interactions across diverse real-world scenarios. Prompt examples for different personas is mentioned in Table 9, and details of prompt template is mentioned in Fig. 8.

## B Experiment

### B.1 Experimental Details

The  $G^3TOD$  zero-shot framework integrates text and multimodal signals into knowledge graphs for dialogue personalization. Entities were extracted using IndicNER, while persona and commonsense relations were inferred via GPT-3.5-Turbo, supplemented with ConceptNet 5.7. For the triplet ranking, the embeddings from all-MiniLM-L6-v2 were used with cosine similarity and only the most relevant triplets (up to 50) were retained. A neighborhood search with depth  $\beta = 3$  ensured contextual breadth, while the top- $\alpha = 3$  utterances from the dialogue history were selected for grounding.

For response generation, GPT-3.5-Turbo (and optionally LLaMA-2-13B) was used with struc-

Model	Persona	PPL	BLEU	ROUGE-L	METEOR	BS-F1	ConSim	R-Len
$G^3TOD$	Housewife	18.55	16.94	67.89	76.12	0.77	0.84	20.28
	IT Professional	22.80	21.32	72.37	85.34	0.80	<b>0.87</b>	28.65
	College Student	<b>12.15</b>	19.32	74.13	<b>85.45</b>	0.82	0.80	31.00
	Senior Citizen	24.62	19.34	70.56	80.25	0.83	0.82	29.50
	Business Owner	23.90	21.02	68.11	78.24	0.84	0.81	26.75
	Freelancer	23.90	<b>23.02</b>	73.11	82.24	<b>0.85</b>	<b>0.87</b>	26.75
	School Teacher	19.40	21.93	<b>75.12</b>	81.08	0.84	0.80	<b>31.08</b>

Table 7: Evaluation of  $G^3TOD$  across diverse user personas in the *HiVisTask* dataset, highlighting variations in response quality across standard metrics.



Figure 7: Overview of the seven user personas used in HiVisTask: *College Student*, *IT Professional*, *Senior Citizen*, *Freelancer*, *Housewife*, *Small Business Owner*, and *School Teacher*. Each persona captures distinct preferences, behavior patterns, and task priorities to enable personalized, grounded dialogue generation.

LLM Prompt for Prompt generation:
You are a helpful assistant responding to a user's latest query. Use the previous conversation context and the following knowledge triplets to generate a personalized, relevant, and natural response.
Previous user messages: {.....}
Knowledge triplets: {.....}
•Entity Graph {.....}
•Persona Graph : {.....}
•Commonsense Graph: {.....}
•User's latest query: {.....}
Instructions:
<ul style="list-style-type: none"> <li>Understand the user's preferences and past behavior from the persona and entity triplets.</li> <li>Use commonsense knowledge to make responses practical and sensible.</li> <li>Incorporate previous conversation context to keep the response coherent.</li> <li>Reply naturally in Hinglish.</li> <li>Keep the response helpful and friendly.</li> </ul>
•Expected output: (Assistant's response generated by the LLM)

Figure 8: Example Prompt for LLM to Generate User Utterance Response

tured prompts containing user history, current utterance, and relevant triplets. The models were run with a token limit of 512–1024, temperature = 0.7, and top- $p$  = 0.9.

## B.2 Automatic Evaluation Metrics

The LLM models are evaluated using both surface-level and semantic metrics including perplexity

(PPL) (Brown et al., 1992) for fluency, BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for n-gram and semantic overlap, BERTScore-F1 (Zhang et al., 2019) (BS-F1) for embedding-based similarity, and context similarity (ConSim) (Biberman, 1994) to assess contextual coherence. We also explore dialogue-specific metrics that includes response length (R-Len), persona alignment (PA), which measures how well the response aligns with the user's persona, and persona consistency (PC), which reflects how consistently the model maintains persona traits across dialogue turns.

### B.2.1 Evaluation Metrics: Persona Alignment and Consistency

**Persona Alignment (PA).** Persona Alignment measures how well the assistant's responses reflect the overall persona of the user. Let  $R = \{r_1, r_2, \dots, r_N\}$  denote the set of assistant re-

Component	Hyperparameter / Setting
Conversational Dataset Split	80% for Graph(s) construction, 20% for Testing and Validation
NER Tool	IndicNER
Persona/Common Sense Extractor	GPT-3.5-Turbo
Commonsense Source	ConceptNet 5.7
Embedding Model	all-MiniLM-L6-v2
Triplet Relevance Threshold ( $\delta$ )	0.86
Triplet Selection Strategy	Cosine similarity-based ranking
Max Triplets per Prompt	50
Context Retrieval (top- $\alpha$ )	3 utterances
Graph Neighborhood Depth ( $\beta$ )	3
LLM for Response Generation	GPT-3.5-Turbo / LLaMA-2-13B
Prompt Max Tokens	512 – 1024
LLM Temperature	0.7
Top-p (Nucleus Sampling)	0.9

Table 8: Experimental Hyperparameters and Settings

sponses, and let  $P$  be the complete persona description of the user. We compute sentence embeddings using a pretrained encoder  $\phi(\cdot)$ , and define cosine similarity as  $\text{sim}(\cdot, \cdot)$ . Persona Alignment (PA) is then calculated as:

$$\text{PA} = \frac{1}{N} \sum_{i=1}^N \text{sim}(\phi(r_i), \phi(P))$$

where a higher PA score indicates stronger alignment between responses and the overall persona.

**Persona Consistency (PC).** Persona Consistency evaluates how consistently the assistant’s responses align with the specific persona attributes that are relevant to each response. Let  $T_i = \{t_{i1}, t_{i2}, \dots, t_{iM_i}\}$  denote the set of persona triplets (attributes) relevant to response  $r_i$ . Persona Consistency (PC) is then defined as:

$$\text{PC} = \frac{1}{N} \sum_{i=1}^N \left( \frac{1}{M_i} \sum_{j=1}^{M_i} \text{sim}(\phi(r_i), \phi(t_{ij})) \right)$$

where  $M_i$  is the number of persona triplets associated with response  $r_i$ . A higher PC score reflects stronger consistency with individual persona attributes.

### B.3 Human Evaluation Metrics

Table 11 is designed to provide a systematic approach for evaluating AI-generated responses. It



Figure 9: Difference in GPT4 responses with various configurations

includes five key metrics: Fluency, Contextual Coherence, Engagingness, Correctness, and Overall performance. Each metric is rated on a scale from 1 to 5, where 1 indicates poor performance and 5 indicates excellent performance. Fluency assesses the grammatical accuracy and natural phrasing of the response, while Contextual Coherence measures its relevance and coherence within the given context. Engagingness evaluates how interesting and captivating the response is, and Correctness checks for factual accuracy. The Overall score offers a comprehensive view of the response’s quality across all metrics. This structured evaluation helps in identifying strengths and areas for improvement, ultimately aiding in the refinement of AI models to better meet user expectations and enhance conversational capabilities.

## C Result and analysis in detail

### C.1 How to choose optimal value of $\beta$ , $\alpha$ , and $\delta$

Table 8 shows hyperparameters and their optimal values across different settings. For context selection,  $\beta_3$  appears to be the most balanced, achieving a good BLEU score (11.98), high ROUGE-L (76.23), and METEOR (77.08), along with decent BS-F1 (0.77) and ConSim (0.81). For path length,  $\alpha_3$  provides a balanced performance with a moderate BLEU score (7.45), consistent ROUGE-L (65.34), and METEOR (65.33), along with strong BS-F1 (0.73) and ConSim (0.81). Regarding triplet relevance threshold,  $\delta_{0.86}$  offers a reasonable balance with BLEU, ROUGE-L, and METEOR, while maintaining a fair BS-F1 and ConSim. Over-

Persona & Task	Modality	Few-shot Prompt
<i>College Student</i> - Movie Booking	Text	Create a Hinglish multi-turn dialogue between a college student and a virtual assistant. The student wants to book a movie ticket for Friday evening with student discount. Include at least 5 dialogue turns. Use casual Hinglish tone and maintain task relevance. Format: User: ..., Assistant: ...
<i>Housewife</i> - Grocery Order via Image	Text + Image	Create a Hinglish multi-turn dialogue between a housewife and a virtual assistant. The user shares an image of a handwritten grocery list (atta, doodh, sabzi) and wants to place an order. Include 5-6 turns, where the assistant interprets the image, confirms items, and proceeds to checkout. Format: User: ..., Assistant: ...
<i>IT Professional</i> - Calendar Scheduling	Image + Text	Simulate a Hinglish dialogue between an IT professional and a smart assistant. The user uploads a calendar screenshot and asks to schedule a team meeting for Thursday afternoon. Include 5 turns. Use professional yet natural Hinglish. Format: User: ..., Assistant: ...
<i>Senior Citizen</i> - Medicine Refill	Image	Generate a Hinglish multi-turn dialogue where a senior citizen sends a photo of a prescription and asks for medicine refills. The assistant should confirm medicine names, dosage, and delivery time. Use a polite and supportive tone. Include at least 4 turns.
<i>Freelancer</i> - Payment Reminder Setup	Text	Create a Hinglish dialogue between a freelancer and assistant. The user wants to set reminders for upcoming client payments due on Monday. Use 5 turns. Ensure task-oriented assistant responses and informal Hinglish phrasing.

Table 9: Few-shot prompt examples for Hinglish multi-turn dialogue generation in the *HiVisTask* dataset.

all,  $\beta_3$ ,  $\alpha_3$ , and  $\delta_{0.86}$  are optimal values for achieving a balanced performance across all metrics. Results demonstrated in Table 15.

## C.2 Error Analysis

We performed a qualitative analysis of failure cases and identified the following key error categories, detailed in Table 19:

- **Context Misalignment:** Some models failed to correctly associate user intent across turns, especially when coreference or ellipsis was involved. Examples mentioned in Table 19 where failure cases are: *Lack of Clarification*, *Incomplete Triplet Coverage* and *Incorrect Entity Resolution*.
- **Overgeneralization:** Personalized models occasionally make incorrect assumptions about user preferences, such as suggesting early morning cabs for business users even when the dialogue history implied flexible timing. Examples mentioned in Table 19 where failure cases are: *Overemphasis on Triplets*, *Temporal Reasoning Failure*, *Excessive Formality*, *Ignoring Negative Preferences* and *Overfitting to Training Data*.

- **Response Redundancy:** In certain cases, the generated responses repeat information or rephrase the user utterance without adding value, especially in shorter multi-turn interactions. Examples mentioned in Table 19 where failure cases are: *Redundancy Context Misinterpretation*, and *Handling Mixed Language*.
- **Lack of Commonsense Reasoning:** Baseline models struggled in scenarios requiring implicit knowledge. Example mentioned in Table 19 where failure case is: *Commonsense Reasoning Gaps*.
- **Incorrect Persona Mapping:** Errors occur when user profiles are ambiguous or when the model hallucinate persona traits not grounded in history, affecting response quality and trustworthiness. Examples mentioned in Table 19 where failure cases are: *Inaccurate Personalization*, and *Ambiguous User Intent*.

## C.3 Case Study and Analysis

We perform comparative study for several top performing LLM models across common test cases. One such comparison, which presents the responses of GPT4 in different ablation configura-

Turn	Converation
User	Mujhe is tarah ka watch chahiye. 
Assistant	Oh! To kya aapko fitness tracker chahiye? Kya features aur budget bataye?
User	Mujhe heart rate monitor aur sleep tracking features chahiye. Budget thoda mid-range ho.
Assistant	Aap <b>Fitbit Charge 5</b> ya <b>Garmin Vivosmart 4</b> try kar sakte hain. Dono features aur budget mein acchi hain.
User	5 achha lag raha hai. Kahan se kharid sakte hain?
Assistant	Aap is link se <b>Fitbit Charge 5</b> kharid sakte hain: <i>Fitbit Store</i>
User	Mere phone ka screen black ho gaya hai, koi solution batao.
Assistant	Aapne phone restart karne ki koshish ki? 
User	Haan, lekin abhi bhi black screen aa raha hai.
Assistant	Aapko phone ko recovery mode mein restart karne ki zarurat ho sakti hai. Mai aapko steps bata deta hoon.

Table 10: Example from the *HiVisTask* dataset, how responses change according to the user’s input over 2 topics.

tions, is illustrated in Fig. 9. The detailed analysis leads to following three observations: (i) **Entity grounding alone helps bridge implicit intent gaps.** For example, where the user mentions "Momo", the  $G_E$  model correctly infers the need for a pet-friendly cab. (ii) **Incorporating persona context enhances temporal and planning awareness.** The model not only understands the user’s intent but also offers proactive suggestions (e.g., proposing a 2 PM cab for a 5 PM flight), which demonstrates an understanding of behavior. (iii)  **$G^3TOD$  excels at integrating commonsense reasoning with personalization.** Our model successfully generates responses that blend both factual and contextual cues, producing coherent and helpful dialogue turns.

## D Discussion

### D.1 Qualitative Analysis

A complete sample conversation illustrating each person’s interaction with the assistant and how the proposed system,  $G^3TOD$ , adapts responses based on persona is provided in Table 16 and Table 17. These tables demonstrate the assistant’s ability to tailor its communication style, content, and additional features according to the specific needs and preferences of different user personas.

In Table 16, the assistant responds to generic requests such as "Help me set up a budget" and "Set up a study timer." The responses differ notably between personas such as *Business Owner*, *Freelancer*, *College Student*, and *School Teacher*. For example, the *Business Owner*’s response focuses on organizing budget elements around operational costs and revenue forecasts, reflecting a professional and data-driven approach. In contrast, the

*Freelancer*’s response emphasizes variable project-based payments, savings tracking, and tax estimates, catering specifically to freelance work patterns. Similarly, study timer settings are personalized for students and teachers with varying session lengths and additional support features like social media notification pauses or focus session logs.

Table 17 presents personalized conversations for booking doctor appointments, where the assistant modifies its language, tone, and functionalities based on the persona of the user—ranging from *Housewife* and *IT Professional* to *Senior Citizen* and *Small Business Owner*. The assistant not only adjusts the formality and politeness but also incorporates persona-specific conveniences such as calendar integration for IT professionals, flexible timing adjustments for housewives, and extended reminders for senior citizens. These nuanced variations highlight the system’s ability to maintain contextual relevance while respecting individual user requirements and communication preferences.

Overall, the qualitative analysis shows that  $G^3TOD$  effectively personalizes dialogue management and response generation, improving user experience by providing tailored, context-aware assistance. This persona-aware approach enhances the perceived naturalness and utility of the assistant across diverse user groups.

### D.2 Structured knowledge drives the LLM reasoning

Table 18 provides Hinglish task-oriented dialogues along with extracted knowledge triplets that guide assistant responses. These conversations cover diverse scenarios like stress relief, food ordering, medical appointments, family dinner planning, medicine reminders, and daily commute. The

Metrics and Scoring	Description
<b>Fluency</b>	
1	The response is difficult to understand due to grammatical errors and unnatural phrasing.
2	The response has several grammatical errors and some unnatural phrasing.
3	The response is mostly grammatically correct but may have minor errors or slightly unnatural phrasing.
4	The response is grammatically correct with natural phrasing, but not perfect.
5	The response is grammatically flawless and sounds completely natural.
<b>Contextual Coherence</b>	
1	The response is off-topic or does not make sense in the given context.
2	The response is somewhat related to the context but has significant coherence issues.
3	The response is mostly coherent but may have minor relevance issues.
4	The response is coherent and relevant to the context with minor lapses.
5	The response is fully coherent and perfectly relevant to the context.
<b>Engagingness</b>	
1	The response is dull and unengaging.
2	The response is somewhat engaging but lacks interest.
3	The response is moderately engaging.
4	The response is engaging and holds interest well.
5	The response is highly engaging and very interesting.
<b>Correctness KG</b>	
1	The response contains incorrect information.
2	The response has several factual inaccuracies.
3	The response is mostly correct but may have minor inaccuracies.
4	The response is correct with minor lapses.
5	The response is completely accurate and correct.
<b>Overall</b>	
1	The response is poor in most aspects.
2	The response is below average.
3	The response is average.
4	The response is above average.
5	The response is excellent in all aspects.

Table 11: Human Evaluation Guide with each metric and their relevant score define to help annotators and researchers.

triplets capture persona traits, preferences, and commonsense relations, enabling personalized, coherent, and context-aware assistant replies. For example, a stressed user is advised to try meditation or music, and someone craving food receives mild vegetarian recommendations from Domino’s.

### D.3 LLMs details

The baseline evaluation measures the performance of various large language models (LLMs), including LLaMA-3.2 , Phi-3 , Gemma-3, Mistral-7B , Qwen3 , GPT-4, GPT-3.5-Turbo , and LLaMA-2-13B (all details in Table 12). These models are assessed in their default configurations to establish a reference point for their core text generation capabilities.

### D.4 Comparison of Textual vs. Textual+Image inputs

To highlight the importance of images in our dataset, we conducted an experiment comparing purely textual inputs with multimodal (tex-

tual+image) inputs. During knowledge graph construction, we process both textual and visual modalities through GPT-4 to extract entities and relations that capture information about the user’s lifestyle, preferences, and personal traits. The visual modality plays a critical role in providing contextual signals that are often missed when relying on text.

**Example.** User (Text + Image): “I love this look! Can you add it to my wishlist?” (User uploads an image of a pair of red sneakers.) From this interaction, we extract persona information in the form of knowledge graph triplets such as: (*user, likes item, red sneakers*).

This multimodal grounding enables us to build richer and more personalized knowledge graphs that reflect user preferences beyond explicit language. During testing, the LLM receives both the image and the utterance, along with distilled knowledge from the graphs, to generate more informative and relevant responses. Table 14 shows the comparison results. The addition of visual inputs significantly improves performance across BLEU

Model	Parameters (B)	Context Length (tokens)	Layers
LLama-3.2 ( <i>Llama-3.2-3B-Instruct</i> ) (Grattafiori et al., 2024)	3	8k–32k	~28
Phi-3 ( <i>Phi-3-medium-4k-instruct-bnb-4bit</i> ) (Abdin et al., 2024)	~3.8	4k	N/A
Gemma-3 ( <i>Gemma-3-4b-it-GGUF</i> ) (Team et al., 2025)	4	128k	28
Mistral-7B ( <i>mistral-7b-instruct-v0.2-bnb-4bit</i> ) (Jiang et al., 2023)	7	Up to 32k	32
Qwen3 ( <i>Qwen3-1.7B</i> ) (Yang et al., 2024)	1.7	N/A	N/A
GPT-4 (OpenAI et al., 2024)	~1.7 – 175	Up to 128k	N/A
GPT-3.5-Turbo (Chen et al., 2023)	6	4k–16k	N/A
LLaMA-2-13B (Touvron et al., 2023)	13	4k	40

Table 12: Comparison of selected large language models (LLMs) used in this paper.

Persona	Model	PPL	BLEU	ROUGE-L	METEOR	BS-F1	ConSim	R-Len
<i>Housewife</i>	$\mathcal{G}_P$	18.55	16.94	67.89	76.12	0.77	0.84	20.28
	Persona+	16.01	5.40	57.60	41.20	0.7191	0.3557	14.19
IT Professional	$\mathcal{G}_P$	22.80	21.32	72.37	85.34	0.80	0.87	28.65
	Persona+	14.67	3.80	50.10	47.50	0.7118	0.3491	15.09
<i>College Student</i>	$\mathcal{G}_P$	12.15	19.32	74.13	85.45	0.82	0.80	31.00
	Persona+	13.90	4.80	49.12	42.30	0.7052	0.3415	14.80
<i>Senior Citizen</i>	$\mathcal{G}_P$	24.62	19.34	70.56	80.25	0.83	0.82	29.50
	Persona+	16.75	3.90	52.42	45.18	0.7128	0.3524	15.22
<i>Business Owner</i>	$\mathcal{G}_P$	23.90	21.02	68.11	78.24	0.84	0.81	26.75
	Persona+	15.10	4.50	50.12	44.12	0.7085	0.3432	14.75
<i>Freelancer</i>	$\mathcal{G}_P$	23.90	23.02	73.11	82.24	0.85	0.87	26.75
	Persona+	14.80	5.10	51.92	46.30	0.7094	0.3501	14.95
<i>School Teacher</i>	$\mathcal{G}_P$	19.40	21.93	75.12	81.08	0.84	0.80	31.08
	Persona+	14.20	4.20	52.88	44.70	0.7140	0.3472	15.12

Table 13: Comparison of  $\mathcal{G}_P$  and *Persona+* with GPT4 results across different personas. *Persona+* considers only the textual persona description (refer to Fig. 7), while  $\mathcal{G}_P$  includes the persona graph.

Model	Variant	BLEU	ROUGE-L
G <sup>3</sup> TOD	Only Textual	13.45	55.67
G <sup>3</sup> TOD	Textual + Visual	23.45	79.96

Table 14: Comparison of only textual and textual + visual results in G<sup>3</sup>TOD.

and ROUGE-L metrics.

### D.5 Comparison of $\mathcal{G}_P$ vs. *Persona+* across distinct personas

To further evaluate the contribution of graph-based personalization, we compare the performance of  $\mathcal{G}_P$  (Persona graph with GPT4) against *Persona+* (static persona description only with GPT4) across seven distinct personas: *Housewife*, *IT Professional*, *College student*, *Senior Citizen*, *Business Owner*, *Freelancer*, and *School Teacher*. The results are summarized in Table 13. Overall, the  $\mathcal{G}_P$  variant demonstrates substantial gains over *Persona+* across all evaluation metrics, validating the role of graph-enhanced personalization in grounding the dialogue system to user-specific contexts.

### D.6 Example Conversation and Extracted Knowledge

Table 20 illustrates an example showcasing the importance of three knowledge graphs: Entity Con-

text Graph, Persona Graph, and Commonsense Graph. The Entity Context Graph  $\mathcal{G}_E$  captures object-level relations, the Persona Graph  $\mathcal{G}_P$  models user-specific traits and preferences, while the Commonsense Graph  $\mathcal{G}_C$  provides general world knowledge that supports reasoning. Together, these triplets demonstrate how structured knowledge can be derived from natural conversations to support downstream tasks such as personalization. Example in Table 20 illustrates how structured knowledge graphs derived from prior dialogue turns enable context-aware and personalized responses. The turn-level extraction of entity, persona, and commonsense triplets captures factual relations, user preferences, and general world knowledge, respectively. When a new query is issued, relevant triplets are retrieved from the knowledge graph and used to guide response generation, allowing the system to reason beyond surface-level dialogue history. This structured representation improves interpretability, supports preference persistence, and leads to more accurate task execution in multi-turn, task-oriented conversations.

Setting	Value	PPL	BLEU	ROUGE-L	METEOR	BS-F1	ConSim	R-Len
<b>Context Selection</b>	$\beta_1$	24.24	8.34	66.98	69.12	0.77	0.79	13.34
	$\beta_2$	34.25	9.23	71.23	70.12	0.76	0.79	17.13
	$\beta_3$	18.23	11.98	76.23	77.08	0.77	0.81	18.96
	$\beta_4$	11.24	5.87	66.98	75.77	0.74	0.81	32.97
<b>Path Length</b>	$\alpha_1$	22.34	2.33	63.24	64.68	0.71	0.77	15.34
	$\alpha_2$	25.42	4.23	68.23	62.33	0.72	0.79	20.23
	$\alpha_3$	21.24	7.45	65.34	65.33	0.73	0.81	24.56
	$\alpha_4$	28.45	7.95	64.98	62.98	0.70	0.83	32.97
<b>Triplet Relevance Threshold</b>	$\delta_{0.50}$	39.12	6.03	41.23	59.87	0.74	0.61	44.25
	$\delta_{0.70}$	37.54	6.87	46.11	61.22	0.75	0.58	47.34
	$\delta_{0.86}$	45.33	5.65	38.94	60.78	0.76	0.54	48.97
	$\delta_{0.90}$	43.24	5.23	33.45	62.34	0.76	0.52	49.45

Table 15: Analysis results on the *HiVisTask* dataset showing the impact of (i) context selection ( $\beta$ ), (ii) path length ( $\alpha$ ), and (iii) triplet relevance thresholds ( $\delta$ ) on various automatic evaluation metrics.

Utterance	Image Description	Assistant Response
<b>Help me set up a budget.</b>	<p>A photo of a bank statement and a wallet with some cash.</p> 	<p><b>Business Owner (Male):</b> "Let's organize your budget around monthly operational costs, revenue forecasts, and reinvestment goals. I've pulled in recent business expenses from your connected accounts for accuracy."</p> <p><b>Hinglish Translation:</b> "Chalo, aapka budget banate hain monthly operational costs, revenue forecasts, aur reinvestment goals ke aas-paas. Maine aapke connected accounts se recent business expenses ko bhi include kiya hai, taaki sab accurate ho."</p> <p><b>Freelancer (Female):</b> "I've drafted a monthly budget based on your average income and variable project-based payments. I also included a savings tracker and tax estimate section just for freelancers like you."</p> <p><b>Hinglish Translation:</b> "Maine aapke average income aur variable project-based payments ke hisaab se monthly budget draft kiya hai. Maine freelancers ke liye savings tracker aur tax estimate section bhi add kiya hai."</p>
<b>Set up a study timer.</b>	<p>A study desk with textbooks, a laptop, and a clock.</p> 	<p><b>College Student (Female):</b> "Timer set for 45 minutes with a 10-minute break after each session. I've also enabled Do Not Disturb on your phone and paused social media notifications."</p> <p><b>Hinglish Translation:</b> "Timer set kar diya hai 45 minutes ke liye aur har session ke baad 10-minute break hai. Maine aapke phone pe Do Not Disturb bhi enable kar diya hai aur social media notifications pause kar di hain."</p> <p><b>School Teacher (Male, for student use):</b> "I've set a 30-minute timer for your student with a 5-minute brain break in between. Would you like me to log their focus sessions for the week?"</p> <p><b>Hinglish Translation:</b> "Maine aapke student ke liye 30-minute timer set kiya hai, jisme 5-minute ka brain break bhi hai. Kya aap chahenge ki main unke focus sessions ko week ke liye log karoon?"</p>

Table 16: Utterance, Image, and Assistant Response Examples with Hinglish Translation for Different Personas.

Persona	Conversation (Assistant's Response)
<i>Housewife</i>	<p><b>User:</b> Mujhe doctor ke paas jaana hai, Thursday ko appointment chahiye.</p> <p><b>Assistant:</b> Ji madam, kaunsa doctor chahiye aapko? Aapka <b>last appointment pediatrician</b> ke saath tha, toh kya waisa hi doctor chahiye?</p> <p><b>User:</b> Haan, pediatrician ke paas jaana hai.</p> <p><b>Assistant:</b> Appointment <b>Thursday ko 10 baje</b> ke liye book ho gaya hai. Main aapko ek reminder bhi bhej dunga.</p> <p><b>User:</b> Thoda jaldi ya late kar sakte ho?</p> <p><b>Assistant:</b> Thursday ko appointment thoda flexible hai, main <b>timing adjust</b> kar sakta hoon agar aapko zarurat ho.</p>
<i>IT Professional</i>	<p><b>User:</b> Mujhe doctor ke paas jaana hai, ek appointment book kar do.</p> <p><b>Assistant:</b> Kaunsa doctor chahiye aapko? Aapka <b>last check-up cardiologist ke saath tha</b>, kya waise hi doctor chahiye?</p> <p><b>User:</b> Haan, cardiologist ke paas jaana hai.</p> <p><b>Assistant:</b> Appointment Thursday ko 10 baje ke liye book ho gaya hai. <b>Google Calendar mein bhi add</b> kar diya gaya hai.</p> <p><b>User:</b> Agar koi urgent kaam aa gaya toh timing change kar sakte hain?</p> <p><b>Assistant:</b> Aapka appointment thoda flexible hai. Main aapko ek link bhej dunga jahan se aap <b>easily reschedule</b> kar sakte hain.</p>
<i>College Student</i>	<p><b>User:</b> Kal doctor ke paas jaana hai, appointment book kar do.</p> <p><b>Assistant:</b> Kaunsa doctor chahiye? Aapka <b>last check-up dentist</b> ke saath tha. Kya wahi doctor chahiye?</p> <p><b>User:</b> Haan, dentist ke paas jaana hai.</p> <p><b>Assistant:</b> Appointment <b>Thursday ko 3 baje</b> ke liye book ho gaya hai. Main <b>reminder</b> bhi set kar raha hoon.</p> <p><b>User:</b> Agar main late ho gaya toh kya karna hoga?</p> <p><b>Assistant:</b> Agar late hote ho, toh main aapko <b>doctor ka contact number</b> bhej dunga taaki aap directly unse reschedule kar sakte ho.</p>
<i>Senior Citizen</i>	<p><b>User:</b> Doctor ke paas jaana hai. Appointment book kar do.</p> <p><b>Assistant:</b> Ji madam, kaunsa doctor chahiye? Aapka <b>last check-up dermatologist</b> ke saath tha, kya wahi doctor chahiye?</p> <p><b>User:</b> Haan, dermatologist ke paas jaana hai.</p> <p><b>Assistant:</b> Appointment <b>Thursday ko 11 baje</b> ke liye book ho gaya hai. Main aapko ek reminder bhi bhej dunga aur agar koi aur madad chahiye ho toh batayein.</p> <p><b>User:</b> Reminder thoda zyada time se bhejna, mujhe yaad rakhne mein thoda mushkil hota hai.</p> <p><b>Assistant:</b> Bilkul madam, main <b>aapko 1 din pehle reminder</b> bhej dunga taaki aapko yaad rahe.</p>
<i>Small Business Owner</i>	<p><b>User:</b> Mujhe doctor ke paas jaana hai, appointment book kar do.</p> <p><b>Assistant:</b> Ji, kaunsa doctor chahiye? Aapka last appointment <b>orthopedist ke saath</b> tha, kya wahi doctor chahiye?</p> <p><b>User:</b> Haan, orthopedist ke paas jaana hai.</p> <p><b>Assistant:</b> Appointment <b>Thursday ko 10:30 baje</b> ke liye book ho gaya hai. Google Meet link bhi add kar diya gaya hai agar virtual consultation chahiye ho.</p> <p><b>User:</b> Agar mujhe kaam ke beech mein reschedule karna ho toh?</p> <p><b>Assistant:</b> Aap <b>easily reschedule</b> kar sakte ho.</p>

Table 17: Personalized Conversations for Doctor's Appointment Booking with each persona, how response are different according to persona.

Turn	User Input	Important Triplets	Assistant Response (Hinglish)
1	Mujhe thoda stress ho raha hai, kuch relax karne ka tariqa batao.	- $\mathcal{G}_E$ : (User - feels - stressed) - $\mathcal{G}_C$ : (Stress - relieved by - relaxation)	Aap thoda stressed ho, samajh sakta hoon. Relax karne ke liye main madad karta hoon.
2	Kya main meditation ya music try kar sakta hoon?	- $\mathcal{G}_P$ : (User - likes - meditation) - $\mathcal{G}_P$ : (User - likes - music) - $\mathcal{G}_E$ : (Meditation - is - relaxation method) - $\mathcal{G}_E$ : (Music - calming - yes) - $\mathcal{G}_C$ : (Meditation/music - reduce - stress)	Ji haan! Aap meditation kar sakte ho ya soft music sun sakte ho jaise classical ya jazz. Yeh dono stress kam karte hain.
3	Kya fresh air lena bhi help karega?	- $\mathcal{G}_C$ : (Fresh air - helps - relaxation)	Bilkul, fresh hawa lene se mind relax hota hai aur stress kam hota hai. Aap thoda bahar ja ke fresh air le sakte hain.
1	Mujhe kuch khaane ka mann ho raha hai.	- $\mathcal{G}_P$ : (User - prefers - vegetarian food) - $\mathcal{G}_C$ : (When hungry - people - order food)	Aapko bhook lagi hai, toh chaliye kuch acha vegetarian khaana order karte hain.
2	Zyada spicy nahi chahiye.	- $\mathcal{G}_P$ : (User - dislikes - spicy food) - $\mathcal{G}_C$ : (Spicy food - may cause - discomfort)	Samajh gaya! Hum mild spice wale options dekhenge, jisse aapko problem na ho.
3	Domino's se kuch order kar sakte hain?	- $\mathcal{G}_E$ : (Domino's - offers - pizza) - $\mathcal{G}_P$ : (User - likes - cheese burst pizza) - $\mathcal{G}_C$ : (Pizza - is a - popular dinner option)	Haan, Domino's se cheese burst pizza order kar sakte hain. Aapko pehle bhi ye pasand aaya tha.
1	Kal se throat pain ho raha hai, doctor se milna hai.	- $\mathcal{G}_C$ : (Throat pain - requires - ENT specialist) - $\mathcal{G}_P$ : (User - prefers - nearby clinics)	Aapko ENT specialist ke paas le chalte hain. Kya main nearby clinic check karu?
2	Haan, aur evening ka time better hoga.	- $\mathcal{G}_P$ : (User - prefers - evening appointments) - $\mathcal{G}_C$ : (Working people - prefer - evening slots)	Thik hai, main evening slot ke liye appointments dekh raha hoon.
3	Apollo clinic mein kar do. Wahan main pehle bhi gaya hoon.	- $\mathcal{G}_E$ : (Apollo Clinic - offers - ENT services) - $\mathcal{G}_P$ : (User - visited - Apollo before)	Perfect! Apollo clinic mein aapke liye ENT specialist ka evening appointment book kar deta hoon.
1	Saturday ko family dinner plan hai, restaurant book karna hai.	- $\mathcal{G}_C$ : (Family dinner - requires - quiet place) - $\mathcal{G}_P$ : (User - dines - with family on weekends)	Family dinner ke liye quiet aur spacious restaurant dekh raha hoon. Kya koi preferred location hai?
2	Hauz Khas area mein kuch dekho.	- $\mathcal{G}_E$ : (Hauz Khas - has - multiple restaurants) - $\mathcal{G}_P$ : (User - prefers - Hauz Khas location)	Hauz Khas area mein 3 family-friendly restaurants mil gaye hain. Aapko Indian cuisine chahiye ya kuch aur?
3	Indian hi chahiye, kids bhi saath honge.	- $\mathcal{G}_C$ : (Kids - prefer - kid-friendly menu) - $\mathcal{G}_E$ : (Restaurant - serves - Indian cuisine)	XYZ restaurant perfect rahega – Indian food ke saath kids menu bhi available hai. Reserve kar du 8 PM ke liye?
1	Subah ki dawai ka reminder laga do.	- $\mathcal{G}_P$ : (User - is - senior citizen) - $\mathcal{G}_C$ : (Morning meds - should - reminded daily)	Sure! Subah 8 baje ka reminder laga diya hai aapki medicine ke liye. Aur koi dawai hai din mein?
2	Raat wali bhi yaad dila dena.	- $\mathcal{G}_E$ : (Medicine-schedule - night dose- 9 PM) - $\mathcal{G}_C$ : (Night meds - often - critical for senior)	Raat 9 baje ka bhi reminder set kar diya gaya hai. Kya medicine ka naam bhi add karna hai?
3	Haan, Ecosprin aur Metformin.	- $\mathcal{G}_E$ : (Ecosprin - is - blood thinner) - $\mathcal{G}_E$ : (Metformin - is - for diabetes) - $\mathcal{G}_C$ : (Seniors - may forget - medicine names)	Thik hai, Ecosprin aur Metformin ka reminder set ho gaya hai subah aur raat dono time. Health ka dhyan rakhein!
1	Office ke liye cab book karni hai 9 baje.	- $\mathcal{G}_P$ : (User - works - in IT company) - $\mathcal{G}_C$ : (Office commute - usually - at fixed time)	Zaroor! 9 baje ke liye cab book kar rahe hain aapke office location par. Pickup address confirm kar dein?
2	Haan wahi ghar se. Aur AC cab chahiye.	- $\mathcal{G}_E$ : (Preferred cab type - is - AC) - $\mathcal{G}_P$ : (User - prefers - comfort) - $\mathcal{G}_C$ : (AC cab - preferred - in summer)	AC cab book ho gayi hai. Estimated fare 220 rupess hoga. Aapka pickup address: 45, Saket Nagar. Confirm karoon?

Table 18: *HiVisTask* dataset with important triplets highlighting persona, entity, and commonsense knowledge used in assistant responses.

Failure Case	Example (User & Assistant Response)	Possible Explanation (in English)
<i>Inaccurate Personalization</i>	<b>User:</b> Main budget travel pasand karta hoon. <b>Assistant:</b> Kya aap luxury hotel book karna chahenge?	The assistant ignored the persona triplets and did not consider the user's budget preference.
<i>Context Misinterpretation</i>	<b>User:</b> Mujhe 6 baje subah ke liye airport ka cab chahiye. <b>Assistant:</b> Aapka cab 9 baje raat ko book kar diya hai.	The assistant failed to properly understand and use the timing information from previous utterances, resulting in an inconsistent response.
<i>Commonsense Reasoning Gaps</i>	<b>User:</b> Kya main flight se sirf 10 minute pehle airport pahunch sakta hoon? <b>Assistant:</b> Haan, bilkul theek hai.	The commonsense knowledge graph lacked information about typical airport arrival times, leading to incorrect advice.
<i>Overemphasis on Triplets</i>	<b>User:</b> Aaj mausam kaisa hai? <b>Assistant:</b> Aap aam tor par economy flight book karte hain, aur aapko early pickup pasand hai.	The assistant relied too much on triplets, causing the response to be off-topic and repetitive, reducing naturalness.
<i>Ambiguous User Intent</i>	<b>User:</b> Next weekend ka plan batao. <b>Assistant:</b> Kya aap flight booking karna chahte hain ya hotel?	Assistant failed to clarify the ambiguous user intent, resulting in a generic and unhelpful response.
<i>Incomplete Triplet Coverage</i>	<b>User:</b> Mujhe restaurant recommend karo Mumbai mein. <b>Assistant:</b> Yeh rahi flight options Delhi ke liye.	Relevant entity or commonsense triplets about restaurants were missing, leading to an irrelevant response about flights.
<i>Temporal Reasoning Failure</i>	<b>User:</b> Main kal subah 7 baje Delhi se Mumbai jaa raha hoon. <b>Assistant:</b> Aapki Mumbai flight aaj raat 10 baje hai.	Assistant confused dates/times in context, failing at temporal reasoning, causing incorrect scheduling info.
<i>Handling Mixed Language</i>	<b>User:</b> Kya aap mujhe ek achha 5-star hotel recommend kar sakte ho in Delhi? <b>Assistant:</b> Sure, here are some budget hotels.	Assistant did not correctly interpret the code-switching or priority in language, mixing English and Hindi preferences incorrectly.
<i>Excessive Formality</i>	<b>User:</b> Chalo movie dekhne chalte hain. <b>Assistant:</b> Apke manpasand manoranjn ke liye shubhkamnayein.	Response is too formal and unnatural for casual conversation, reducing user engagement.
<i>Redundancy</i>	<b>User:</b> Flight book kar do Delhi se Mumbai. <b>Assistant:</b> Main aapke liye flight book kar raha hoon Delhi se Mumbai flight booking kar raha hoon.	Repetition and redundancy in generated response affects fluency and user experience.
<i>Lack of Clarification</i>	<b>User:</b> Mujhe weekend mein trip plan karna hai. <b>Assistant:</b> Trip plan karna shuru karta hoon.	Assistant did not ask clarifying questions to understand destination, budget or preferences, leading to incomplete assistance.
<i>Ignoring Negative Preferences</i>	<b>User:</b> Main late night travel pasand nahi karta. <b>Assistant:</b> Aapka cab 11 baje raat ko book kar diya hai.	Assistant failed to incorporate negative user preferences from persona graph, leading to unsatisfactory suggestions.
<i>Incorrect Entity Resolution</i>	<b>User:</b> Main apni flight Mumbai se Goa ke liye cancel karna chahta hoon. <b>Assistant:</b> Aapki flight Delhi se Goa confirm ho gayi hai.	Assistant mixed up entity details, confusing origin cities due to inadequate entity graph processing.
<i>Overfitting to Training Data</i>	<b>User:</b> Mujhe nayi movies ke baare mein batao. <b>Assistant:</b> Main sirf travel related information de sakta hoon.	Assistant fails to generalize to diverse user queries outside training scope due to overfitting on travel domain.

Table 19: Error Analysis: Common Failure Cases with Hinglish Examples (Extended).

Turn	Utterance	Entity Context Graph ( $\mathcal{G}_E$ )	Persona Graph ( $\mathcal{G}_P$ )	Commonsense Graph ( $\mathcal{G}_C$ )
User	Mere dog ke liye appointment lo doctor ka.	(appointment, purpose, vet_doctor)	(user, owns, dog)	(dogs, may_need, veterinary_care)
Assistant	Dog ka kya naam hai?	—	—	—
User	Momo.	(dog, name, Momo)	—	—
Assistant	Okay noted.	—	—	—
User	Momo ko garmi lag rahi hai, AC room book karna.	(room, type, AC_room)	(user, prefers, AC_comfort), (user, cares_for, Momo)	(heat, causes, discomfort), (pet, requires, temperature_control_room)
Assistant	Done.	—	—	—

Table 20: Turn-Level Past Conversation with Entity, Persona, and Commonsense Triplets.

#### Using Retrieved Knowledge for the Next User Query

When the user asks a new question, the system retrieves relevant triplets from the knowledge graph (KG) and generates a context-aware response.

**User:** Book the cab for airport? Mai aur Momo ja rahe hain.

Relevant Triplets Retrieved from Past History:

- (dog, name, Momo),
- (pet, requires, temperature\_control\_room) ,
- (user, cares\_for, Momo)

**Response (Generated Using Retrieved Triplets):** Aap aur Momo airport ja rahe hain aur aapko AC comfort pasand hai. Main aapke liye pet-friendly AC cab arrange kar deta hoon. Kya main booking confirm kar du?

These examples motivate the need for structured relational knowledge rather than relying solely on raw dialogue history. The retrieved triplets enable reasoning about user preferences, pet-related constraints, and situational requirements.