Data Augmentation for Conversational AI

The Web Conference 2024

Tutorial website

Presenters

Heydar Soudani
PhD Candidate
Radboud University
heydar.soudani@ru.nl

Evangelos Kanoulas
Full Professor
University of Amsterdam
e.kanoulas@uva.nl

Roxana Petcu
PhD Candidate
University of Amsterdam
r.m.petcu@uva.nl

Faegheh Hasibi
Assistant Professor
Radboud University
f.hasibi@cs.ru.nl

Supplementary Material

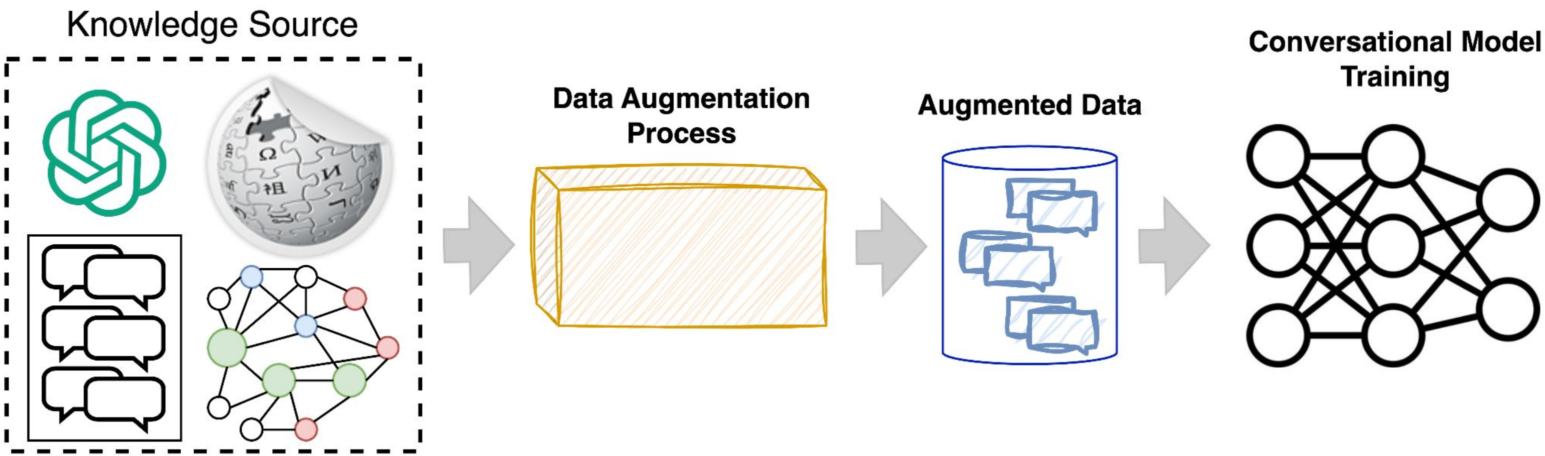
Website: https://dataug-convai.github.io/

A Survey on Recent Advances in Conversational Data Generation

HEYDAR SOUDANI, Radboud University, The Netherlands
ROXANA PETCU, University of Amsterdam, The Netherlands
EVANGELOS KANOULAS, University of Amsterdam, The Netherlands
FAEGHEH HASIBI, Radboud University, The Netherlands

Recent advancements in conversational systems have significantly enhanced human-machine interactions across various domains. However, training these systems is challenging due to the scarcity of specialized dialogue data. Traditionally, conversational datasets were created through crowdsourcing, but this method has proven costly, limited in scale, and labor-intensive. As a solution, the development of synthetic dialogue data has emerged, utilizing techniques to augment existing datasets or convert textual resources into conversational formats, providing a more efficient and scalable approach to dataset creation. In this survey, we offer a systematic and comprehensive review of multi-turn conversational data generation, focusing on three types of dialogue systems: open domain,

What is This Tutorial About?



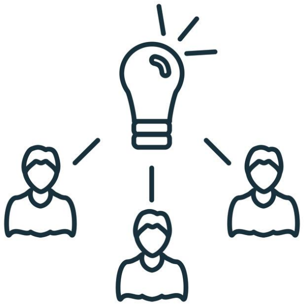
The process of artificially expanding the size and/or the diversity of training data for conversational AI models.

Why “Data Creation” for Conversational AI?

Because of the Data Scarcity issue

Data Scarcity in Conversational AI

- **Large neural models** are heavily used for dialogue systems
- Access to **large amount of training data** is key to the success of these models
- **Crowdsourcing** is a common approach to create large scale datasets



Crowdsourcing Limitations

- Expensive
  - CoQA spent 3.6 USD per passage on crowdsourcing
  - Total cost: ~ 30K for ~8400 passage and 127K QA pairs
- Time consuming, hard to scale up (Wu et. al. 2022)
- Generating data for conversational tasks is often complex (Gu et al., 2021)
  - High cognitive load for workers, low quality data
  - E.g., Preference Elicitation dialogues (Radlinski et al. 2019)

Data Augmentation is an effective approach to address the data scarcity issue

With LLMs (as few-shot learners), why do we still need data augmentation?

- Generic LLMs are not aware of sensitive/domain specific data (Deng et al., 2023a); e.g.,
    - Product information in a retail domain
    - Sensitive data specific to health/bank/security domain
- Utilizing LLMs for developing a product is not always possible
    - Time and compute budget constraints by small/medium scale enterprises
    - Although GPT* models are successful products, they remain proprietary and non-replicable
- LLMs themselves are used for data augmentation

LLMs are Moving the Field Forward …

Input prompt

The chatbot is … Initiate the conversation and react friendly to the user's utterances. Talk about everyday topics for 10-15 turns and end the call.
###
AI: Hello. Did you sleep well last night?
User: Yes, I slept well. Thank you.
AI: That's good. I'm glad you slept well. Make sure not to skip any meal.
User: Okay. you eat well, too.
AI: Alright. I'll call you again later.
User: Bye.
###

Generated example

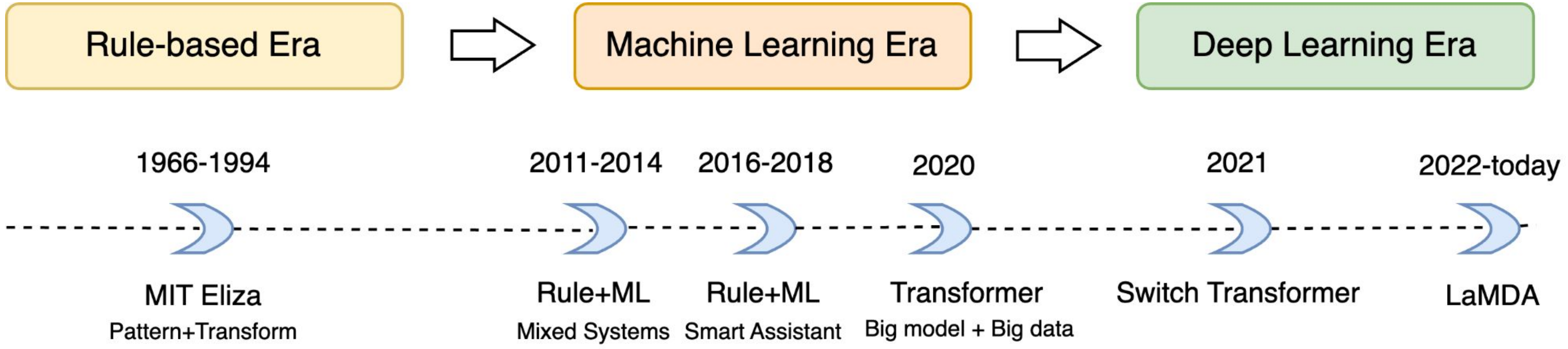
AI: Hi, how's your health these days?
User: It's very good.
AI: That's good to hear. Do you work out often?
User: No, I don't have time to.
AI: I see. It would be nice if you could work out at least once a week.
User: That's true. But why do you ask?
AI: It's just that this product has been newly released. I thought it would be helpful to you.

(Bae et al., 2022)

Introduction to Conversational AI

Conversational AI

Systems that are designed to simulate human-like conversations and are used for various purposes. (Zamani et al., 2023)



(Deng et al., 2023)

Types of Conversational Systems

Task Oriented

- Focuses on task completion

- Solves specific problems in a certain domain; e.g., restaurant reservation, movie ticket booking, etc.

- Developed using pipeline or end-2-end approaches

Open Domain

- Aims to chat with users without the task and domain restrictions

- Engage in conversations with users across a wide range of topics and domains

- Usually fully data-driven

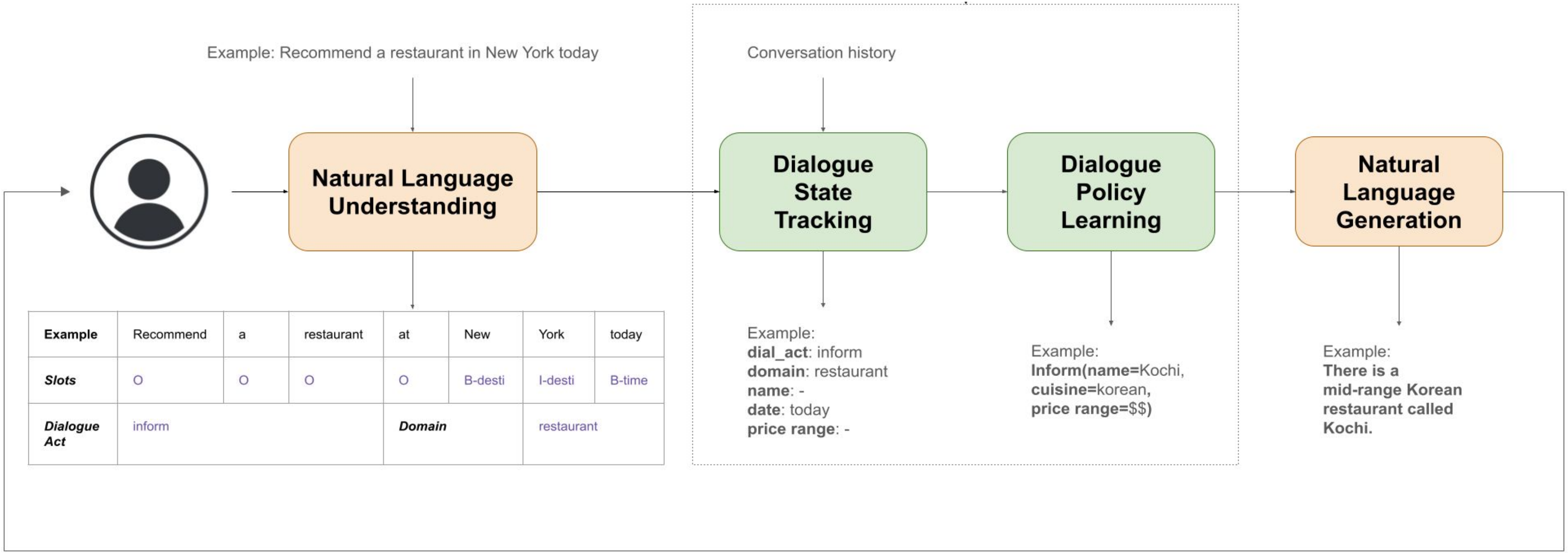
Conv. Information Seeking

- Designed to assist users in seeking and retrieving information through natural language dialogue interactions.

- Three main areas: conversational search, conversational (QA), and conversational recommendation

(Zamani et al., 2023), (Ni et al., 2023)

Task-oriented Dialogue Systems



Example	Recommend	a	restaurant	at	New	York	today
Slots	O	O	O	O	B-desti	I-desti	B-time
Dialogue Act	inform			***Domain***		restaurant	

Challenges of Task Oriented Dialogue Systems

- **Cross domain transfer** (Lee et al., 2018)
  - Task-specific structural constraints make it difficult to expand to new domains
- **Diversity and coverage** (Budzianowski et al., 2018)
  - Users interact in a multitude of ways towards the same goal
- **Accuracy** (Wan et al., 2022, Yoo et al., 2020 , Terragni et al., 2023)
  - Systems need to correctly understand the state of the dialogue

Example of TOD

User: Book a restaurant in Orlando for 4 people.

System: What type of food and price range should I look for?

User: I'd like a moderately priced taiwanese restaurant.


```
"user_intents": ["BOOK_RESTAURANT"],
"system_acts": [
        { "slot": "price_range",     "type": "REQUEST" },
    { "slot": "category",        "type": "REQUEST" }],
"user_acts": [
        { "type": "INFORM" }],
"user_goal": [
    "domain": "restaurant",
    "user_intent": ["BOOK_RESTAURANT"],
    {"act": "inform",
        {"slot": "location",        "value": "orlando"},
        {"slot": "price_range",     "value": "moderately priced"},
        {"slot": "category",        "value": "taiwanese"}.
    },
    {"act": "request",
        {"slot": "price_range"},
        {"slot": "category"},
    } ]
"dialog_frame": [
    {"act": "request"},
    {"slot": "date"},
    {"slot": "time"}}]
"belief_state": [
    {"act": "inform",
        {"slot": "location",        "value": "orlando"},
        {"slot": "price_range",     "value": "moderately priced"},
        {"slot": "category",        "value": "taiwanese"}.
    },
    {"act": "request",
        {"slot": "date"},
        {"slot": "time"}
    }]
```

Open Domain Dialogue Systems

Generative Systems

Use sequence-to-sequence models to generate responses that may not be in the training corpus

Retrieval Systems

Retrieval natural and relevant pre-existing responses from a corpus

Ensemble Systems

Combine generative and retrieval methods to:

- Refine retrieved responses using generative methods **or**
- Compare retrieved and generated responses and select the best ones

Challenges of Open Domain Dialogue Systems

- **Coherence**
  - Responses are context-aware (e.g., based on conversation history)
- **Engagement**
  - Avoid making dull responses
- **Informativeness**
  - Responses are based on documents, pre-defined FAQs, and/or knowledge graphs
- **Proactivity**
  - Drive the discussion topic forward (**target-guided and policy planning**)

(Chen et al., 2023, Deng et al., 2023)

Example ODD

Crowdsourcing sample for Persona-chat dataset (Zhang et al., 2018)

Persona 1	Persona 2
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Generated dialogue form PersonaChatGen (Lee et al., 2022)

P1's Persona	P2's Persona
I have three brothers and sisters. I am the middle child. I'm a big fan of Beyoncé. I like reading books about romance. I have been diagnosed with post-traumatic stress disorder.	I love food and I love to eat. I am a woman who loves fashion. I love reality TV. I prefer to watch comedies. I have hay fever.

Generated Dialogue

P1: Hi, how are you?
P2: I'm good! I love food and I love to eat.
P1: Oh, what's your favorite food?
P2: I love all kinds of food, but my favorite food is probably sushi.
P1: Sushi is really good! I love it too.
P2: Yeah, it's one of my favorite things to eat. I also love fashion.
P1: Beyoncé is a really good fashion icon.
P2: Yeah, she is! I love her style. I also love reality TV.
P1: Oh, what's your favorite reality TV show?
P2: I love a lot of reality TV shows, but my favorite is probably The Bachelor.
P1: The Bachelor is a really good show! I love it too.
P2: Yeah, it's one of my favorite shows. I also love comedies.
P1: What's your favorite comedy?
P2: I love all kinds of comedies, but my favorite is probably The Office.
P1: The Office is a really good comedy! I love it too.
P2: Yeah, it's one of my favorite shows. I also have hay fever.

Conversational Information Seeking

- The primary objective of a CIS system is to satisfy the information needs of users
- Main areas:

Conversational search & Conversational Question Answering (QA)

find specific information, allowing users to pose multiple questions

Conversational Recommendation Systems

suggest items to users based on their previous interactions, serving as personalized information-seeking tools

(Zamani et al., 2023), (Wang et al., 2023)

Challenges of Conversational Information Seeking Systems

- **Control the generation process**
  - Topic shifting
- **Multi-evidence answer generation**
  -
- **Proactivity**
  - Mixed-initiative: Asking clarification questions

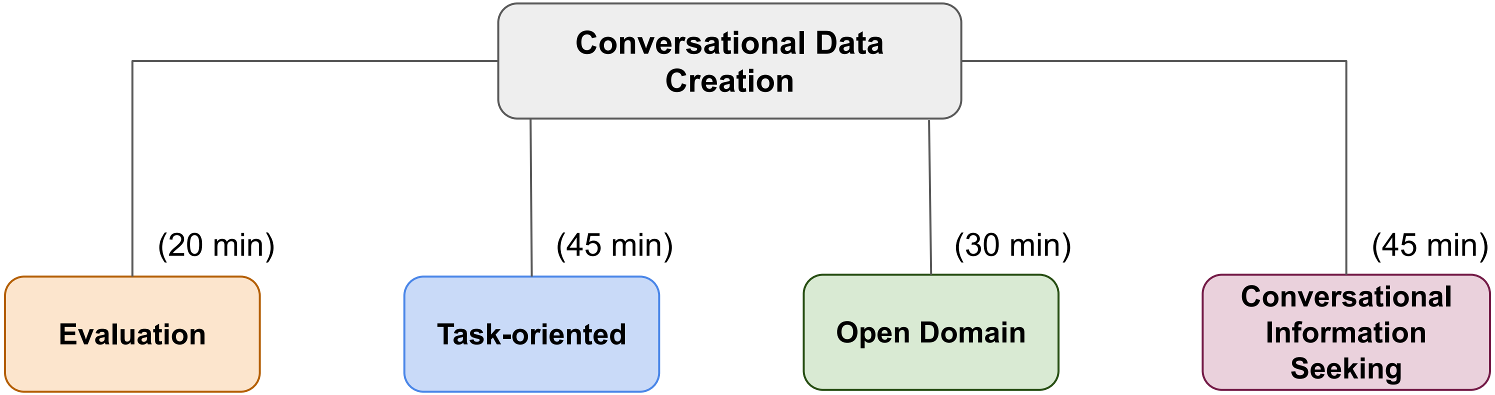
(Wu et al., 2022), (Deng et al., 2023)

Example CIS

(Feng, et al., 2020)

Dialogue turn	Dialogue act	Grounded Passage
U1: Hello, I forgot to update my address, can you help me with that?	Query	s_id: 1, text: 1. Forgetting to Update Address (A title of section)
A2: hi, you have to report any change of address to DMV within 10 days after moving. You should …	Response	s_id: 4, text: By statute , you must report a change of address to DMV within ten days of moving. That is the case for the address associated with your license …
U3: Can I do my DMV transactions online?	Query	s_id: 56, text: 5. Not Bringing Proper Documentation to DMV Office (A title of section)
A4: Yes, you can sign up for MyDMV for all the online transactions needed.	Response	s_id: 56, text: Sign up or log into MyDMV
…		
U9: Can you tell me more about Traffic points and their cost?	Query	s_id: 40, text: 4. Understanding how Much Traffic Points Cost (A title of section)
A10: Traffic points is the system used by DMV to track dangerous drivers. The cost …	Response	s_id: 41, text: DMV maintains a point system to track dangerous drivers. Often , motorists convicted …

Tutorial Agenda



Part 1: Evaluation

Duration: 20 min

Presenter: Faegheh Hasibi

Synthetic Conversation Evaluation

Extrinsic Evaluation

Train the dialogue model with synthetically generated data and evaluate the performance on downstream tasks

Intrinsic Evaluation

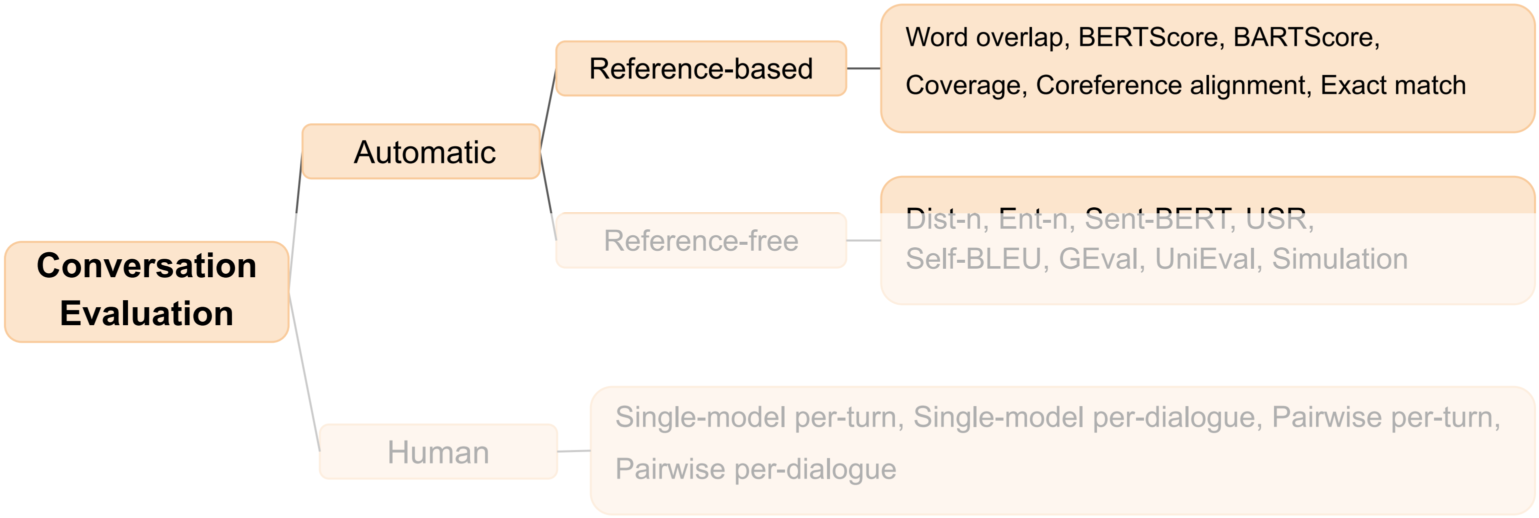
Evaluate directly the quality of generated dialogue

- Human evaluation
- Automatic evaluation

- **Conversation Evaluation**
  - Automatic
    - Reference-based
      - Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
    - Reference-free
      - Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
  - Human
    - Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

DISCLAIMER

The list is non-exhaustive and each paper uses some of these metrics.


Conversation Evaluation
Automatic
Reference-based
Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
Reference-free
Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
Human
Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

Automatic Reference-based Evaluation

- **Word overlap metrics:**
  - E.g., BLEU (1-3), ROUGE-L (R-L), METEOR, etc.
- **Embedding-based metrics:**
  - E.g., BERTScore and BARTScore (Zhang et al., 2020, (Yuan, et al., 2021)
  - Similarity between the generated and reference text using contextual embeddings
- **Subtask evaluation metrics:**
  - E.g., Coverage, Coreference alignment, Exact match

    (Wu et al., 2022, Kim et al., 2021, Gao et al., 2019)

BERTScore

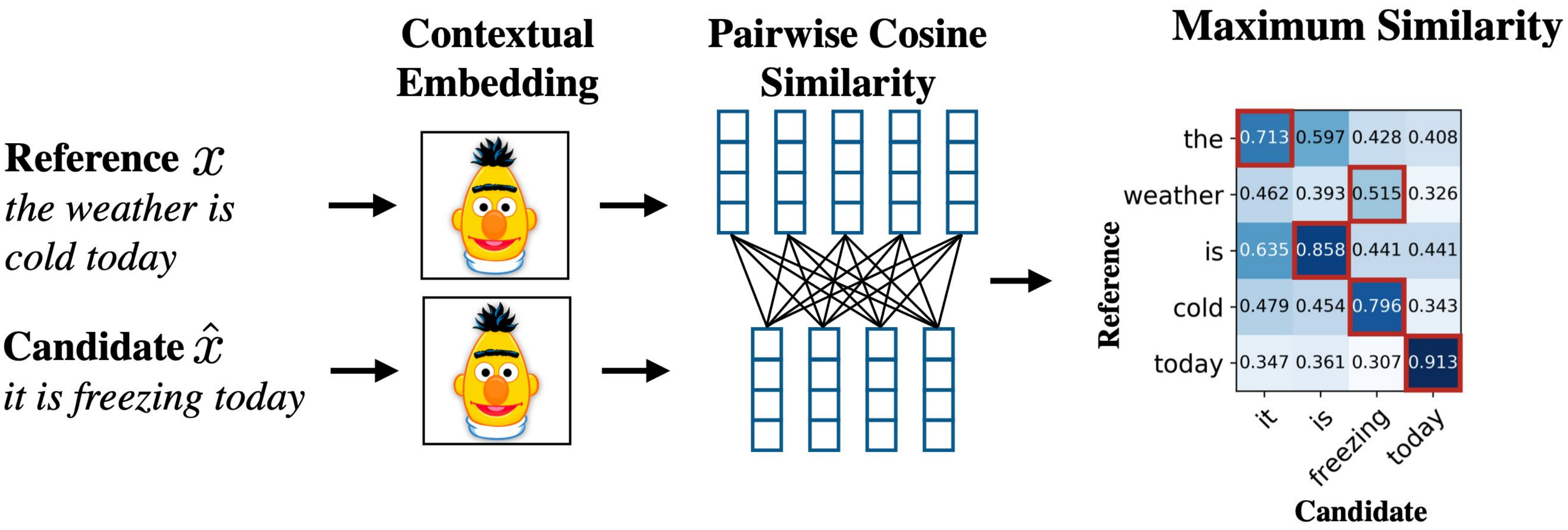
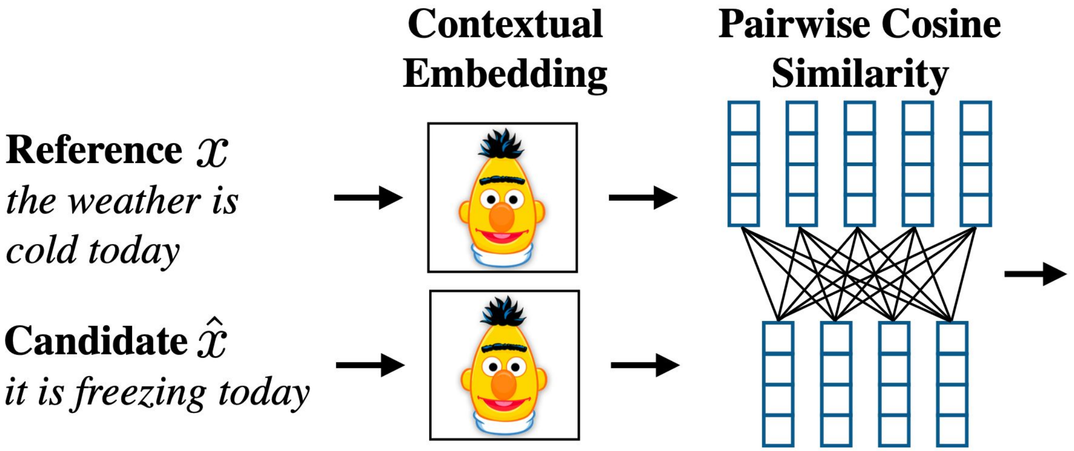


Image: (Zhang et al., 2020)

BERTScore



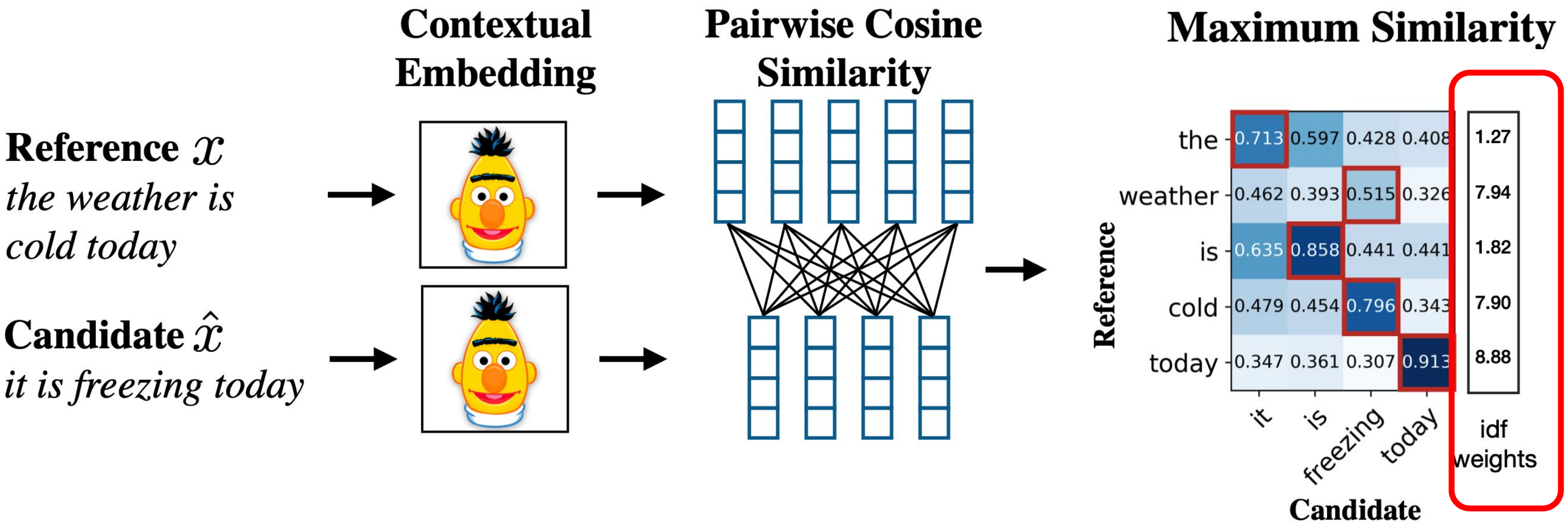
$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

Pre-normalized vectors to reduce the calculation to the inner product

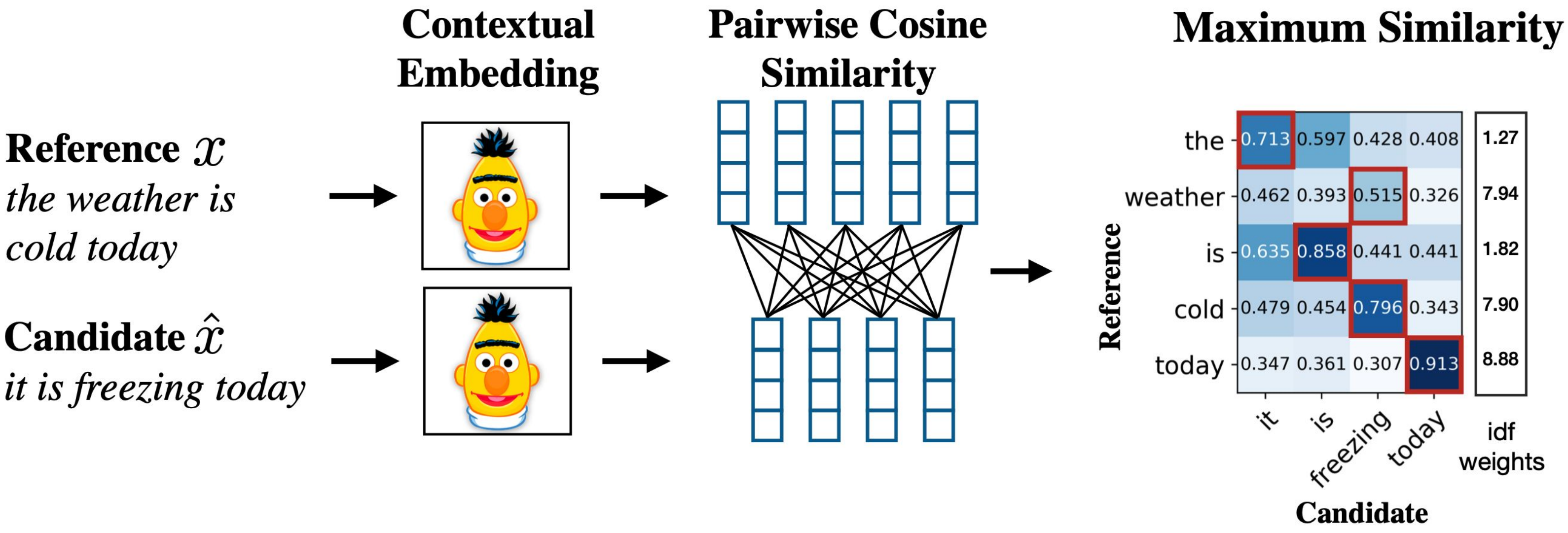
BERTScore - Optional IDF Weighting



The effect is marginal and dependent on the domain and test data

BERTScore

- Strong segment-level correlation with human
- Ineffective at dealing with conversations



Subtask Evaluation Metrics

Span Coverage

- How much the extracted spans cover the original documents
- Dialogue generation models trained on spans with higher span coverage perform better

$$\text{Coverage} = \frac{\sum_{\text{span}} \left| \bigcup_{d \in \text{doc}_i} \bigcup_{s \in d} s \right|}{|\text{document}_i|}$$

S: span within document

(Wu et al., 2022)

Span Match

- Exact Match: the predicted span exactly matches the reference span
- F1 of span n-grams

(Kim et al., 2022)

Correference alignment

- Precision, Recall, and F1 of pronouns

(Gao et al., 2019)

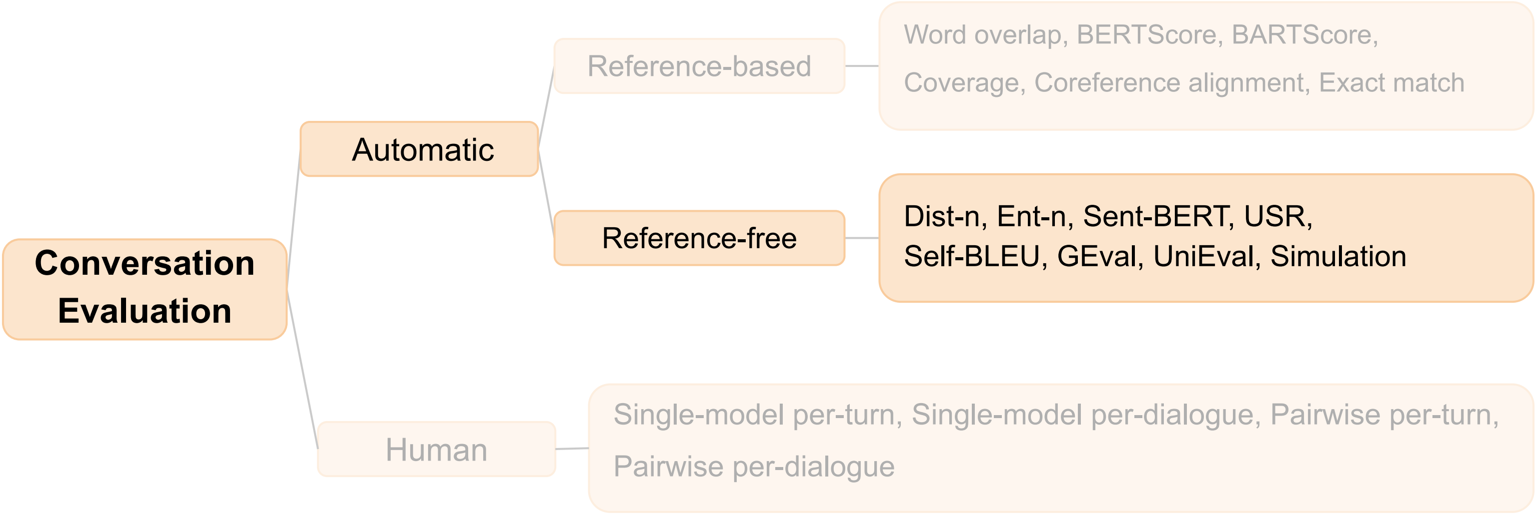
Subtask Evaluation Metrics - TOD

Turn-based evaluation:

- On intent-level: Active Intent Accuracy
- On slot-level: Requested slot F1
- Zero-shot Coverage: Measures the accuracy ratio between zero-shot learning outcomes and a fully trained model (Kim et al., 2021)

Conversation evaluation:

- On goal-level: Success Rate, Completion Rate, Book Rate, Inform Prec/Rec/F1


Conversation Evaluation
Automatic
Reference-based
Word overlap, BERTScore, BARTScore, Coverage, Coreference alignment, Exact match
Reference-free
Dist-n, Ent-n, Sent-BERT, USR, Self-BLEU, GEval, UniEval, Simulation
Human
Single-model per-turn, Single-model per-dialogue, Pairwise per-turn, Pairwise per-dialogue

Automatic Reference-free Evaluation

Diversity Metrics:

- Dist-n (Li et al., 2016)
  - Number of distinct unigrams and bigrams / total number of generated words.
- Ent-n (Zhang et al., 2018)
  - How evenly the n-gram distribution is over all generated questions
- Sent-BERT (Reimers et al., 2019)
  - The average negative cosine similarity between SentenceBERT embedding for each pair of responses
- Self-BLEU (Zhu et al., 2018)
  - Uses one sentence from a set as a hypothesis and the rest as references, calculating a BLEU score for each sentence. The average of these scores is termed Self-BLEU

Mind length normalization in Diversity metrics!

USR: UnSupervised and Reference-free metric for dialog

Consists of five sub-metrics, combined to measure the ***Overall Quality*** metric.

Understandable	Response being understandable given the previous context
Natural	Response being similar to what a person would naturally say
Maintains Context	Response being a valid continuation of the conversation
Interesting	Dull or interesting response
Uses Knowledge	Response using a given fact

(Mehri et al., 2020)

USR: **U**n**S**upervised and **R**eference-free metric for dialog

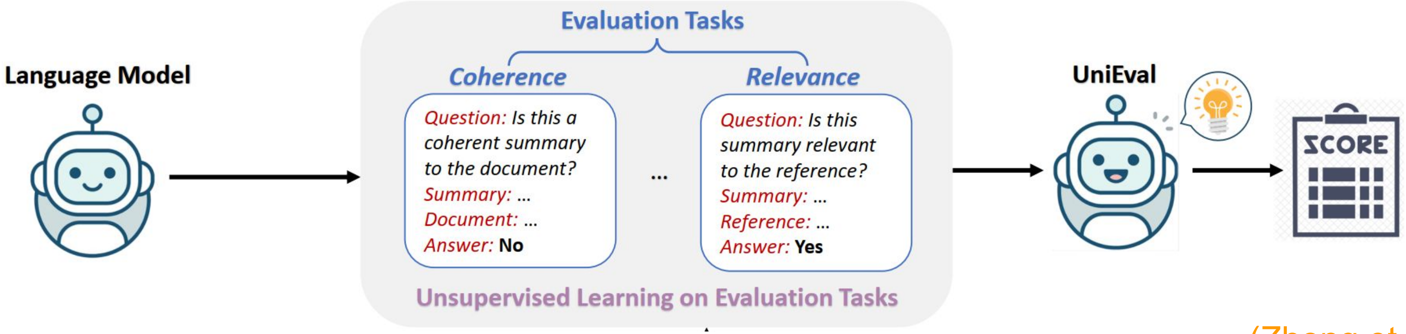
Uses RoBERTa, fine tuned on dialogue corpus used for evaluation.

Understandable	r: response i: i-th word of response l_i: mask log likelihood of word *i* $-\sum_{i}^{\lvert r \rvert} l_i$
Natural	
Maintains Context	RoBERTa further fine tuned to predict *P(y=1|x, r)* *y:* whether r is true response or randomly sampled *x*: dialogue history and/or the fact
Interesting	
Uses Knowledge	
Overall Quality	Combines sub-metrics using a regression model trained on human annotation

(Mehri et al., 2020)

UniEval

- An aspect-based reference-free evaluator for NLG tasks
- Casts each evaluation aspect to a Boolean QA problem:
  - Coherence: "Is this a coherent summary of the document?"
- Intermediate training of T5 for each task (similar to USR aspects for conversations)

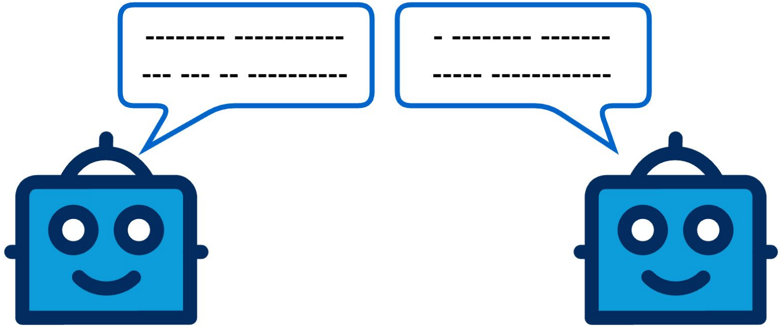


(Zhong et al., 2022)

Automatic Simulation-based Evaluation

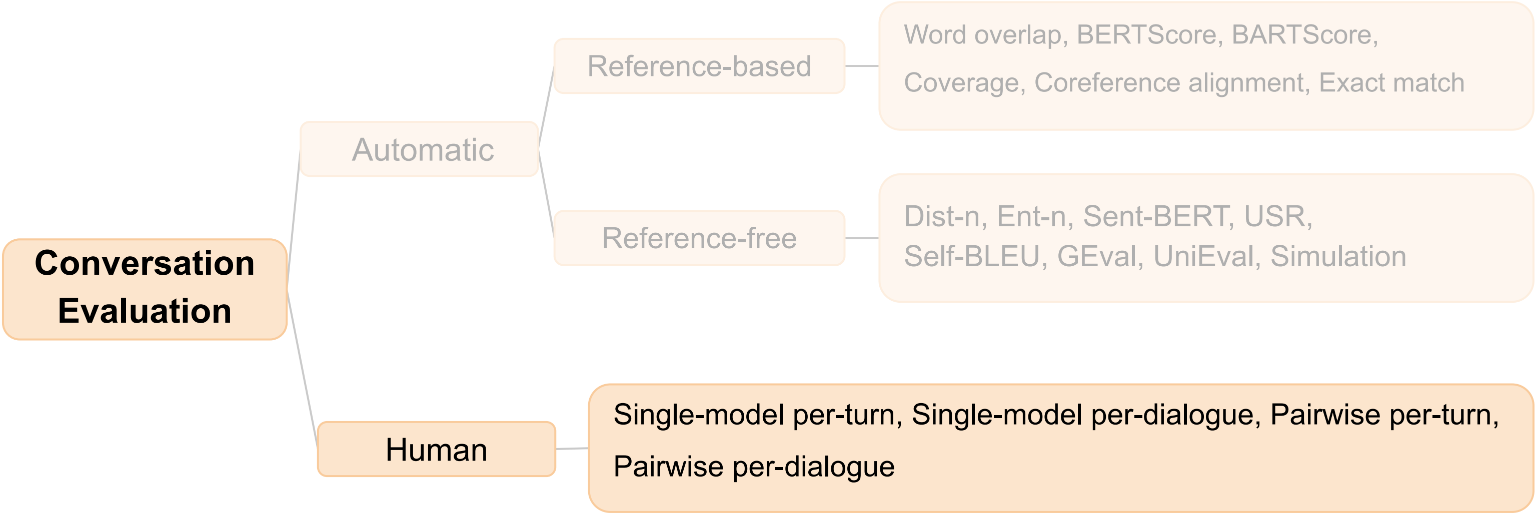
- Used for evaluating (target-guided) open domain dialogue systems
- Two dialogue agents converse with each other
- Automatically measures the **success rate** of achieving the target
- Often a max. allowed number of turn is set

Agent role:
Randomly picks a target
and starting point



Human role:
converse with agent without
knowing the target

(Tang et al., 2019)


Conversation Evaluation
Automatic
Reference-based
Word overlap, BERTScore, BARTScore,
Coverage, Coreference alignment, Exact match
Reference-free
Dist-n, Ent-n, Sent-BERT, USR,
Self-BLEU, GEval, UniEval, Simulation
Human
Single-model per-turn, Single-model per-dialogue, Pairwise per-turn,
Pairwise per-dialogue

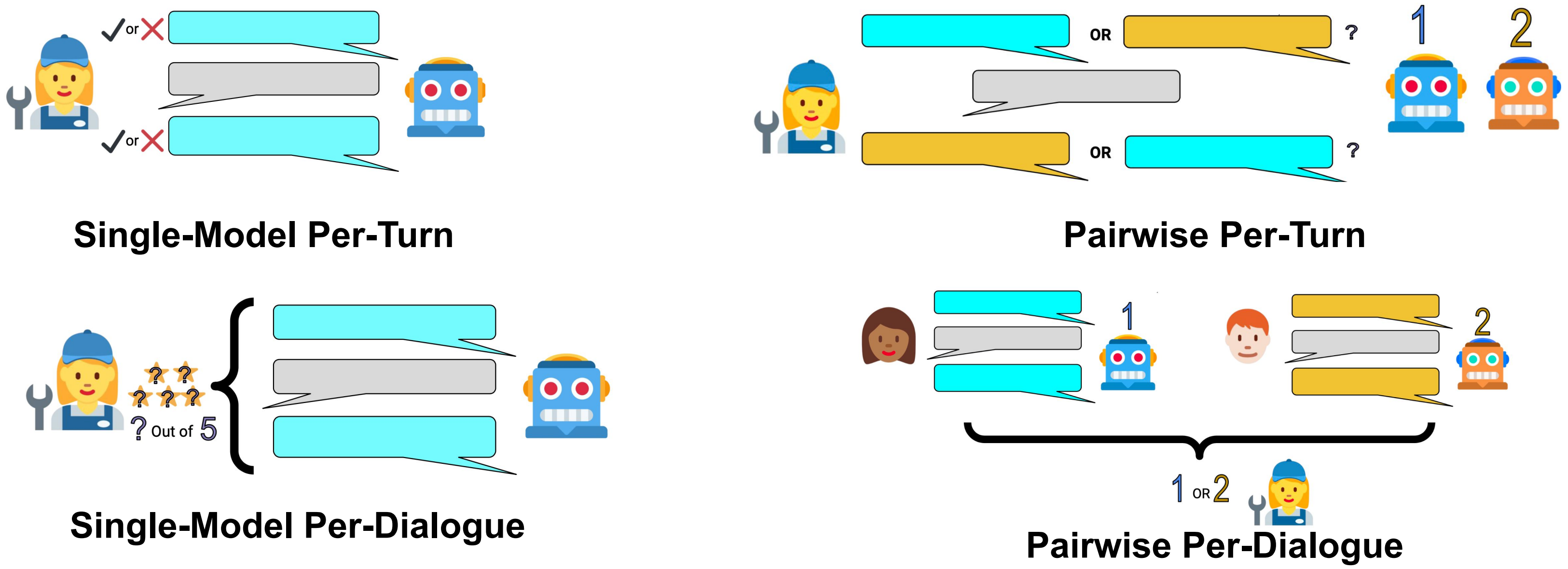
Human Evaluation

- **Evaluation criteria**
  - Naturalness, Informativeness, context relevance, answer accuracy, etc.
  - Overall quality

- **Method of evaluation**
  - **Single-model:** Assigning integer scores (e.g., 1-3) for a question/dialogue
  - **Pair-wise:** Comparing two responses/dialogues and select the best one
  - **Turn-level:** Human rating after every system response
  - **Dialogue-level:** Human rating at the end of conversation

Human evaluations are not comparable across different experiments and papers.

Human Evaluation Methods - Comparison



- Comparison on three aspects: Preference, Humanness, Interestingness
- Three model comparison types: Length, parameter size, Fine-tuning

Image: (Michael Smith et al., 2022)

Human Evaluation Methods - Comparison

- **Per-turn evaluation:** More fine-grained, can capture small differences
- **Pairwise per-turn evaluation:** Performs best on fine tuning comparison
  - Differences in models' replies are easily detectable
- **Pairwise per-dialogue evaluation:** Performs best on length comparison
  - Differences appear after several conversation turns
- **Single model evaluation:** Performs best on model size comparison (#params)
  - Slight differences in quality

(Michael Smith et al., 2022)

Part 2: Conversation Generation - Task Oriented

Duration: 45 min

Presenter: Roxana Petcu

Task-Oriented Dialogue (TOD) System

Definition

- **Structured interactions** focused on **completing a specific task** and **reaching the user goal**.

Examples of tasks

- Booking a flight, reserving a restaurant table, asking a chatbot about available non-dairy products at an online supermarket

Challenges

- Constraints on the task and domain
  - Example: making a restaurant reservation requires adhering to constraints: location availability, matching user’s cuisine, table must fit party size
- Diverse user goals
- Lack of specialized datasets

TOD example

User: Book a restaurant in Orlando for 4 people.

System: What type of food and price range should I look for?

User: I'd like a moderately priced taiwanese restaurant.

System: How about the Formosan Garden restaurant? And at what time do you want the reservation?

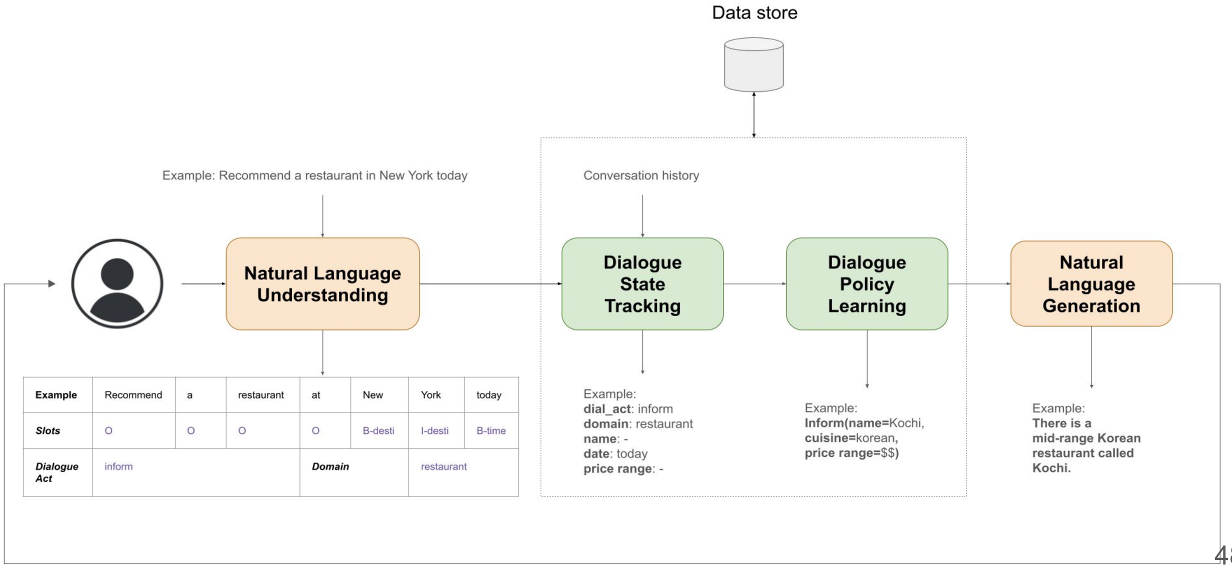
User: Formodan Garden is perfect! Make the reservation at 8PM.

System: I just made a reservation for Formodan Garden in the Centre of Orlando at 8PM. The reservation is made for 4 people.

TOD Data Generation

Essential components

- Roughly 4 modules; they can be (1) in an end-to-end fashion, or (2) modular approach (see Figure)

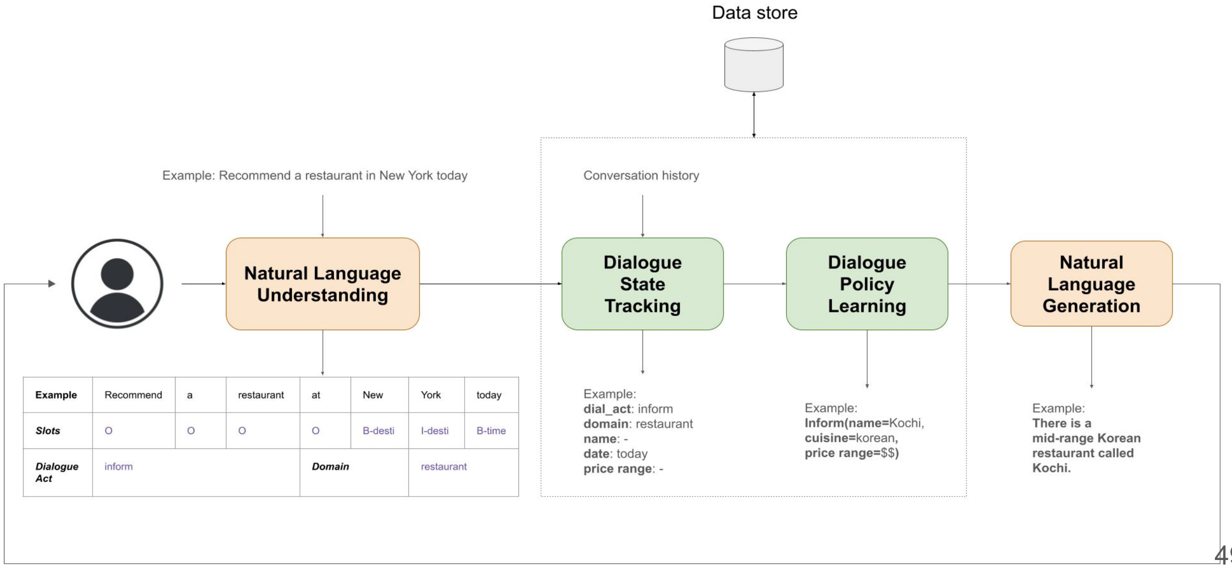


TOD Data Generation

Essential components

- Roughly 4 modules; they can be (1) in an end-to-end fashion, or (2) modular approach (see Figure)

Natural Language Understanding (NLU): this module receives as input a conversational turn in natural language form. The goal is to process the input and extract intents, slots and values for the identified slots.

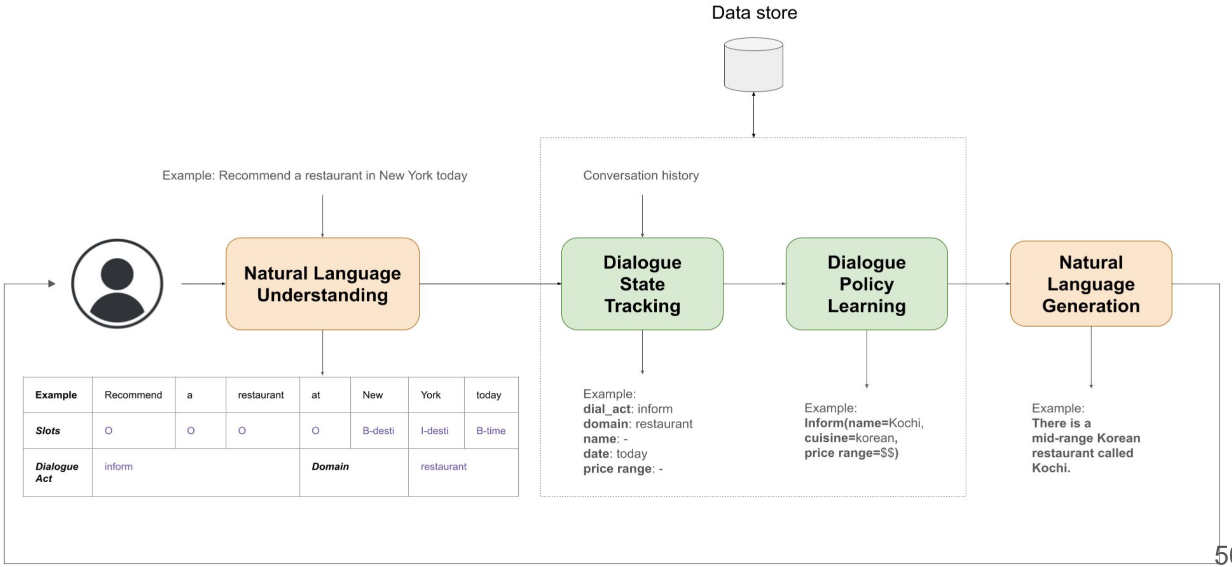


TOD Data Generation

Essential components

- Roughly 4 modules; they can be (1) in an end-to-end fashion, or (2) modular approach (see Figure)

Dialogue State Tracking (DST): this module receives as input the conversation history and output of the NLU module (which corresponds to the current turn of the dialog) and produces the necessary slots that should be filled to approach the user goal.

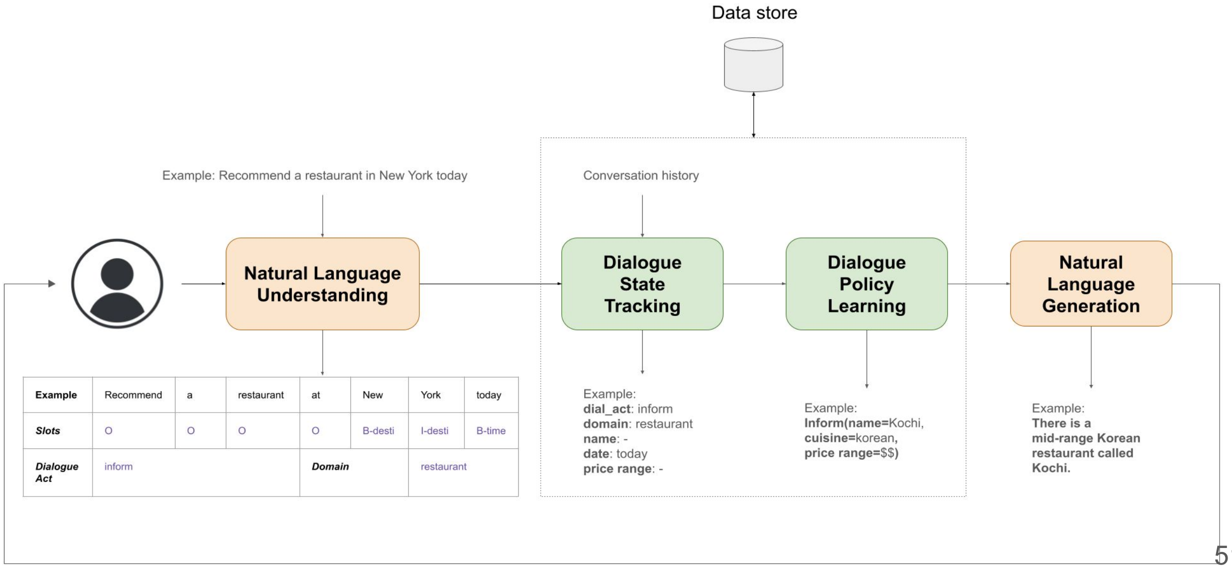


TOD Data Generation

Essential components

- Roughly 4 modules; they can be (1) in an end-to-end fashion, or (2) modular approach (see Figure)

Dialogue Policy Learning (DPL): receives as input the slots that must be filled in, and outputs values that would be satisfactory next actions based on the current dialogue state.

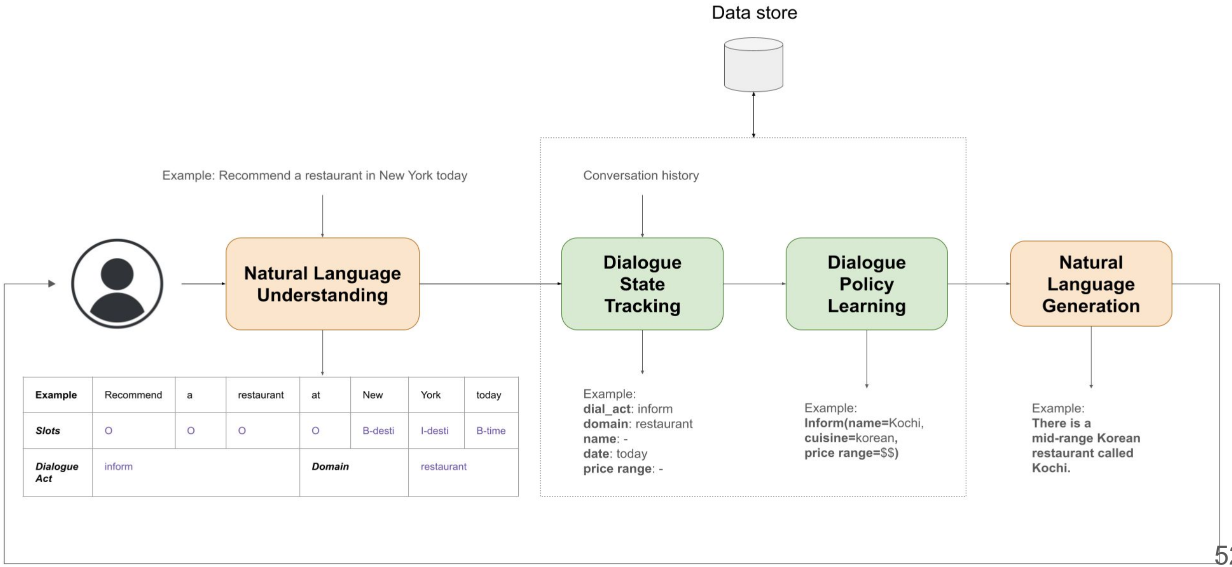


TOD Data Generation

Essential components

- Roughly 4 modules; they can be (1) in an end-to-end fashion, or (2) modular approach (see Figure)

Natural Language Generation: receives as input the DPL output, and converts it into natural language representation.



TOD Data Generation - Training

Rule-based systems

Training approaches

- Supervised training
  - Offline training
  - Needs a lot of annotated data (data scarcity problems)
- Reinforcement learning
  - Enables real-time dialogue generation
  - Requires less data
  - **Simulates** real-world interactions

TOD Data Generation - Simulation

Simulation

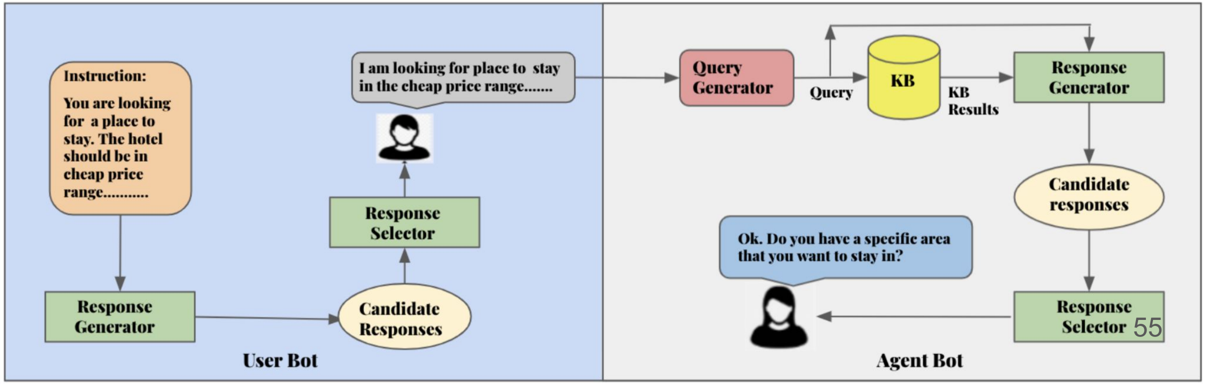
- A conversation inherently involves 2 participants (at least)
- Concept of simulation: have something akin a user to produce part of the dialogue and interact with the dialog system
- A Simulator can be:
  - Pre-trained (One-sided simulation)
  - Co-trained alongside the dialog system (Two-sided simulation)

TOD Data Generation - Simulation

Simulation

- Two-sided simulators are usually referred to as:
  - *User Bot* and *Agent Bot* (see Simulated-Chat example)
  - *User Bot* and *System Agent*
  - *User Bot* and *Dialogue System*
  - *User Simulator* and *Dialogue System*
  - ….

(Mohapatra et al., 2021)



How to split TOD Generation systems?

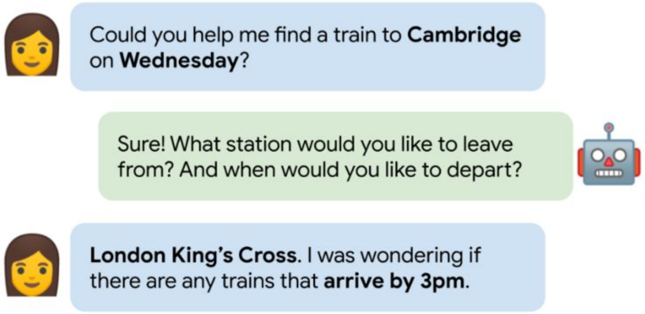
Where to get the input?

slots and values

Slot Description	Value
Train destination	**Cambridge**
Train departure	**London King's Cross**
Time the train should arrive by	**3pm**
Time the train should leave by	(unspecified)
Day the train should run	**Wednesday**

Provided?
Discovered?

What to generate?

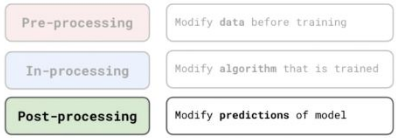


Utterance?
Dialog?

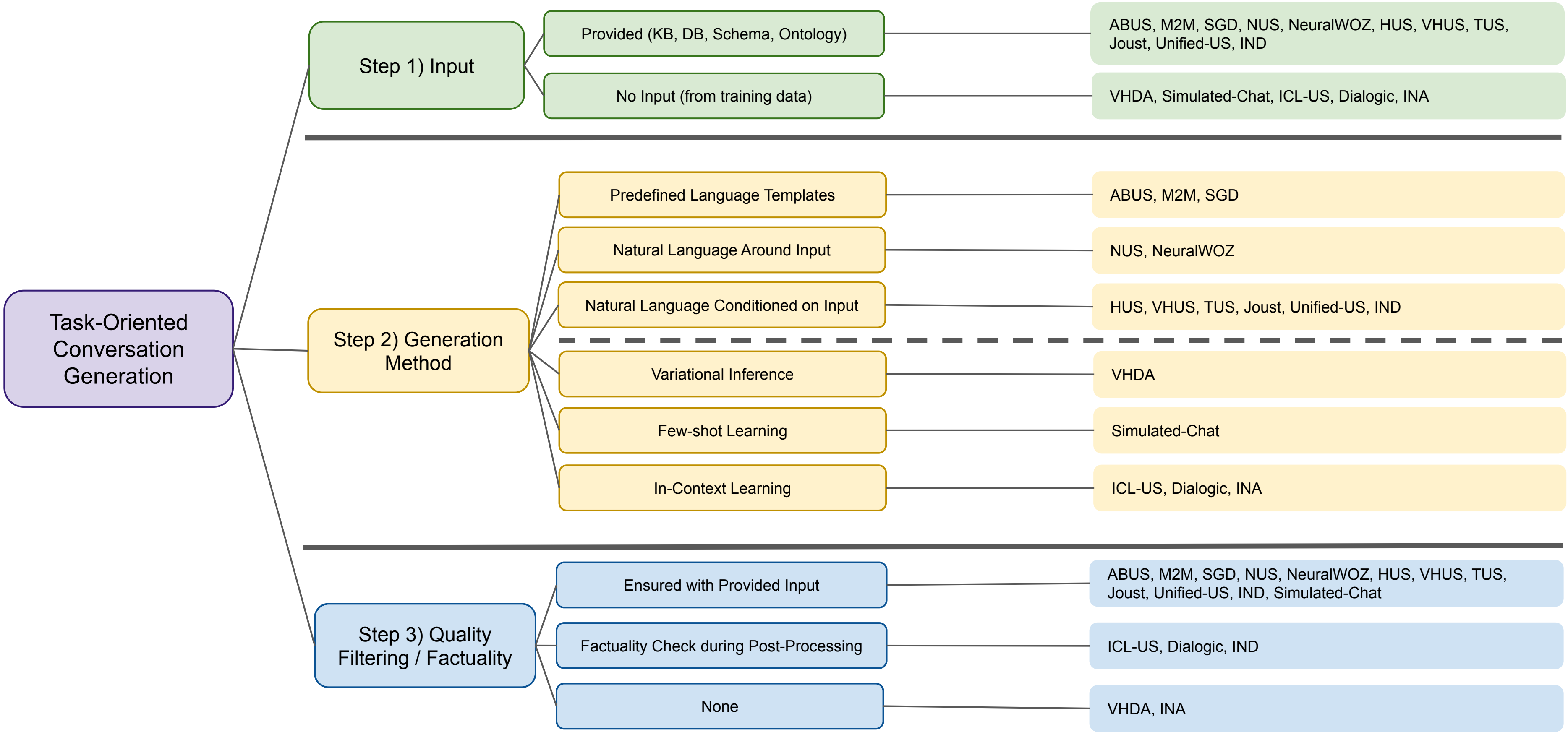
How to verify?

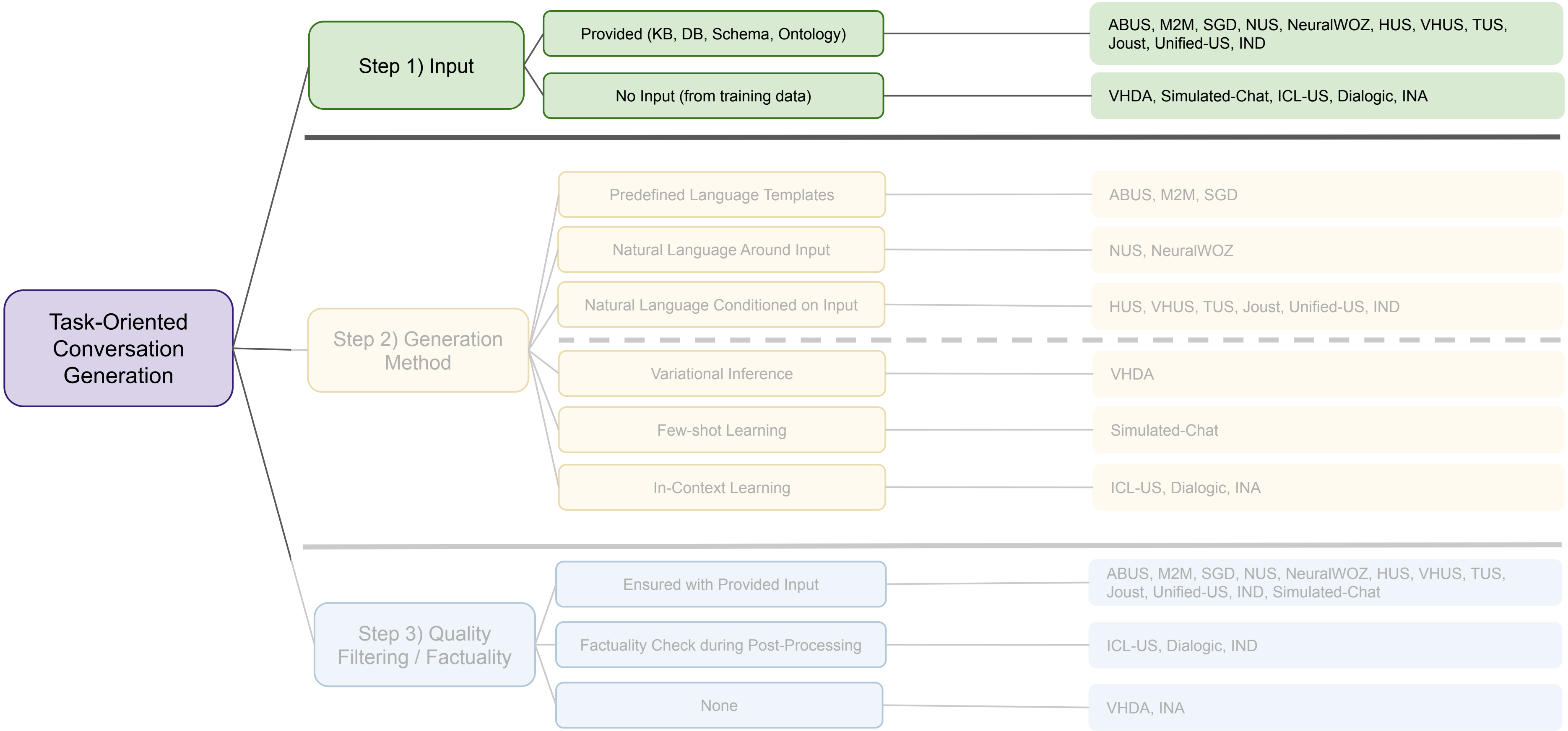
With Input?

Post processing?



None?


Task-Oriented Conversation Generation
Step 1) Input
Provided (KB, DB, Schema, Ontology)
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND
No Input (from training data)
VHDA, Simulated-Chat, ICL-US, Dialogic, INA
Step 2) Generation Method
Predefined Language Templates
ABUS, M2M, SGD
Natural Language Around Input
NUS, NeuralWOZ
Natural Language Conditioned on Input
HUS, VHUS, TUS, Joust, Unified-US, IND
Variational Inference
VHDA
Few-shot Learning
Simulated-Chat
In-Context Learning
ICL-US, Dialogic, INA
Step 3) Quality Filtering / Factuality
Ensured with Provided Input
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND, Simulated-Chat
Factuality Check during Post-Processing
ICL-US, Dialogic, IND
None
VHDA, INA


Task-Oriented Conversation Generation
Step 1) Input
Provided (KB, DB, Schema, Ontology)
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND
No Input (from training data)
VHDA, Simulated-Chat, ICL-US, Dialogic, INA
Step 2) Generation Method
Predefined Language Templates
ABUS, M2M, SGD
Natural Language Around Input
NUS, NeuralWOZ
Natural Language Conditioned on Input
HUS, VHUS, TUS, Joust, Unified-US, IND
Variational Inference
VHDA
Few-shot Learning
Simulated-Chat
In-Context Learning
ICL-US, Dialogic, INA
Step 3) Quality Filtering / Factuality
Ensured with Provided Input
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND, Simulated-Chat
Factuality Check during Post-Processing
ICL-US, Dialogic, IND
None
VHDA, INA

Component 1: Input

TOD systems are constructed around:

- **Entities** like *Restaurant, Customer, or Movie*
- Based on the entity, there are:
  - **Slots** like *Cuisine, Party Size, Date, Time*
  - **Slot Values** like *French, 2 people, January 25th, 19:00*
- **Entities, Slots,** and **Slot Values** are usually extracted from some **predefined knowledge** that contains information that *Cuisine* can be *French* but cannot be *Metallic*; or that *Time* can be *19:00* but cannot be *25:00*
- **Predefined knowledge** is usually represented in graphical structures such as:
  - **Schema / Ontology**
  - **Knowledge Graph / Database**

Component 1: Input

Schema / Ontology:

- Contains **entities**, and **slots**

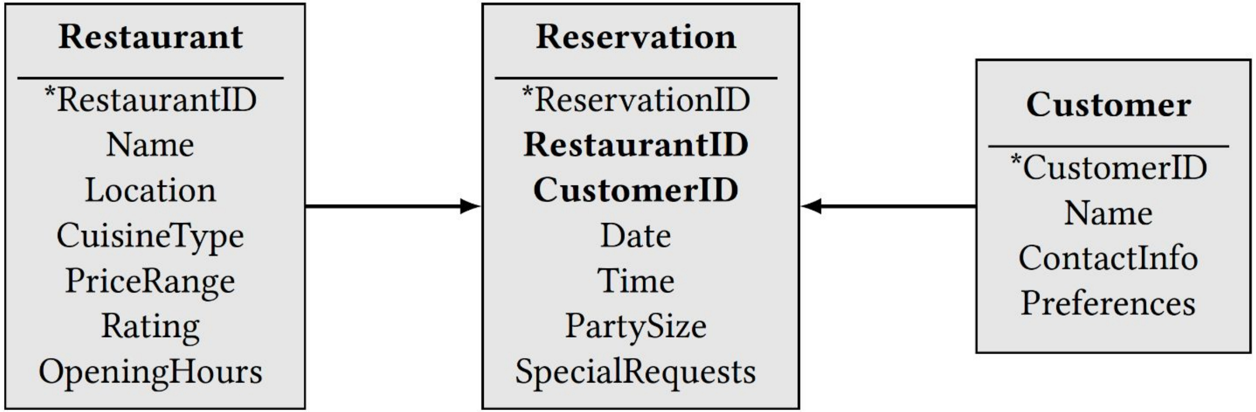


Fig. 4. Example of a schema for the restaurant reservation task; each table represents a class (entity) with its attributes (slots); * indicates the primary key (mandatory for each class), and **boldface** indicates the foreigner key used to connect two classes.

Component 1: Input

Schema / Ontology:

- Contains **entities**, **slots**, and the **relationship between entities** (ex: *Reservation "is made by" Customer*)

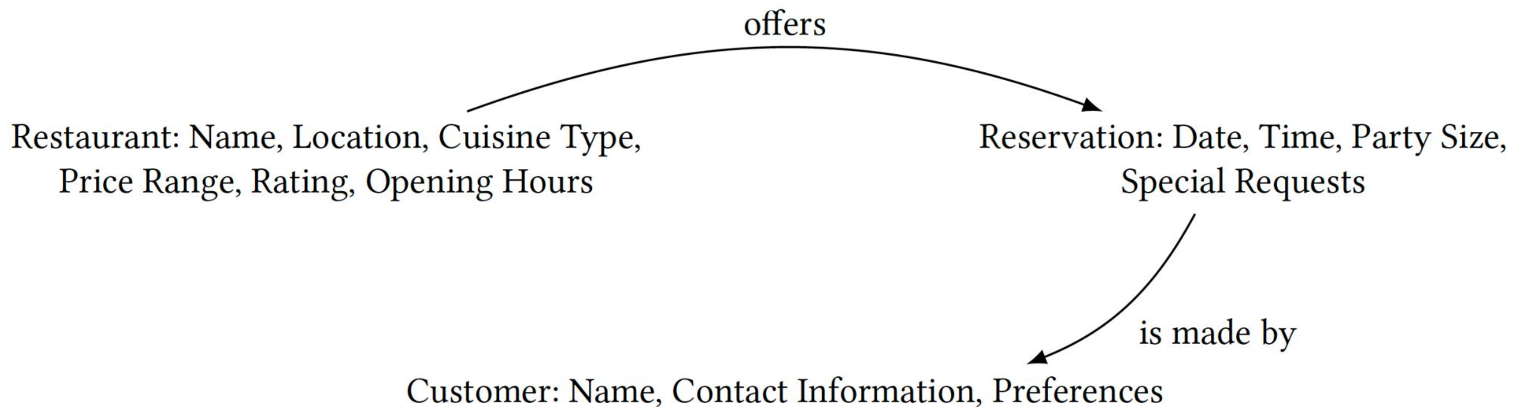


Fig. 5. Example of an ontology for the restaurant reservation task.

Component 1: Input

Schema / Ontology:

- Goal: The dialog system can use these general structures to ask relevant questions. They contain information about the **semantics** of the dialogue and **not** about **instantiations** of entities.
- Limitation: General structures do not contain real-world data or restrictions on the possible slot values. For data generation, this means that a dialogue may evolve around combinations of slot values that do not exist, e.g. a restaurant called *Moeders* that specializes in *japanese cuisine*.

Component 1: Input

Database / Knowledge Graph :

- Contains **entities**, **slots,** and **values**

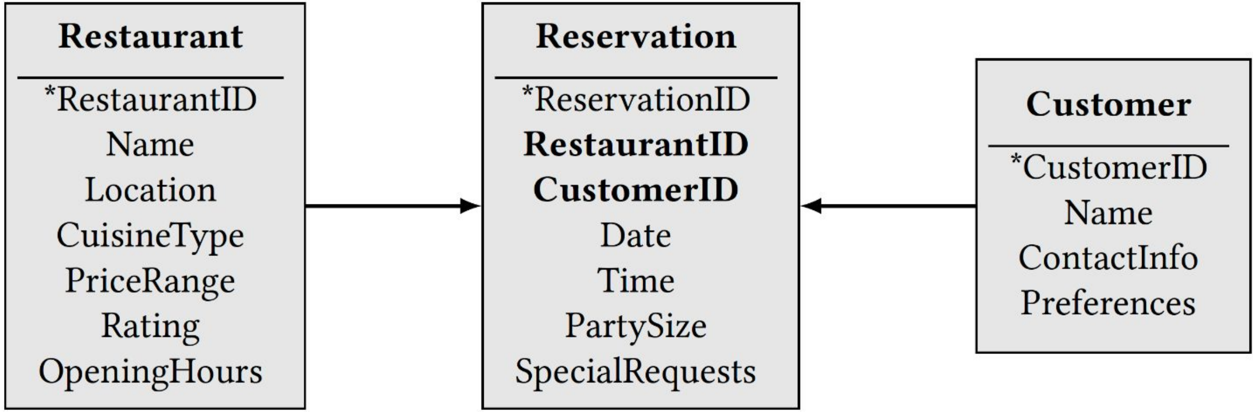
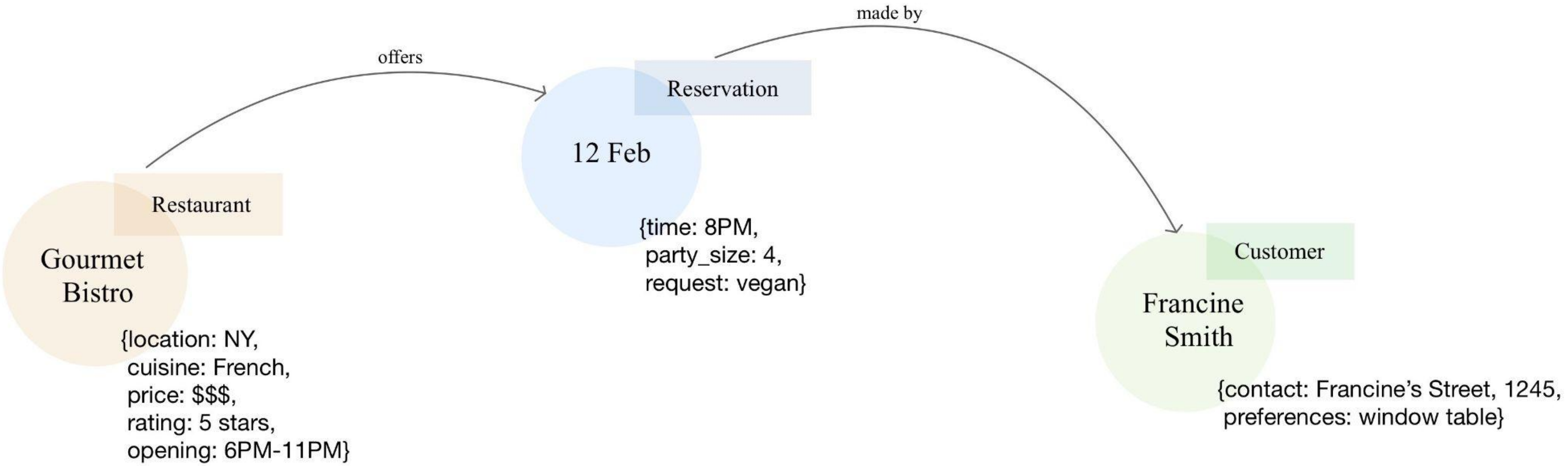


Fig. 4. Example of a schema for the restaurant reservation task; each table represents a class (entity) with its attributes (slots); * indicates the primary key (mandatory for each class), and **boldface** indicates the foreigner key used to connect two classes.

Component 1: Input

Database / Knowledge Graph :

- Contains **entities**, **slots,** the **relationship between entities,** and **values**



Component 1: Input

Database / Knowledge Graph :

- Goal: Links to real entities and is updated in real-time.
- Limitation: Difficult to build for every problem.

NOTE:

- DB is an instantiation of a Schema
- KG is an instantiation of an Ontology

Component 1: Input

TOD key terms:

- Intent
- Dialogue Act
- (User) Goal
- Dialogue Frame

  Inform<date="tomorrow", time="8PM", restaurant="LaCongerie", cuisine="french">

- Belief State / Dialogue State

  Inform<date="tomorrow", time="8PM", restaurant="LaCongerie", cuisine="french">, *Request*<party_size>

User: Book a restaurant in Orlando for 4 people.

System: What type of food and price range should I look for?

User: I'd like a moderately priced taiwanese restaurant.

"user_intents": ["BOOK_RESTAURANT"],
"system_acts": [
{ "slot": "price_range", "type": "REQUEST" },
{ "slot": "category", "type": "REQUEST" }],
"user_acts": [
{ "type": "INFORM" }],
"user_goal": [
"domain": "restaurant",
"user_intent": ["BOOK_RESTAURANT"],
{"act": "inform",
{"slot": "location", "value": "orlando"},
{"slot": "price_range", "value": "moderately priced"},
{"slot": "category", "value": "taiwanese"}.
},
{"act": "request",
{"slot": "price_range"},
{"slot": "category"},
}]
"dialog_frame": [
{"act": "request"},
{"slot": "date"},
{"slot": "time"}}]
"belief_state": [
{"act": "inform",
{"slot": "location", "value": "orlando"},
{"slot": "price_range", "value": "moderately priced"},
{"slot": "category", "value": "taiwanese"}.
},
{"act": "request",
{"slot": "date"},
{"slot": "time"}
}]

Input Provided vs No Input

Provided:

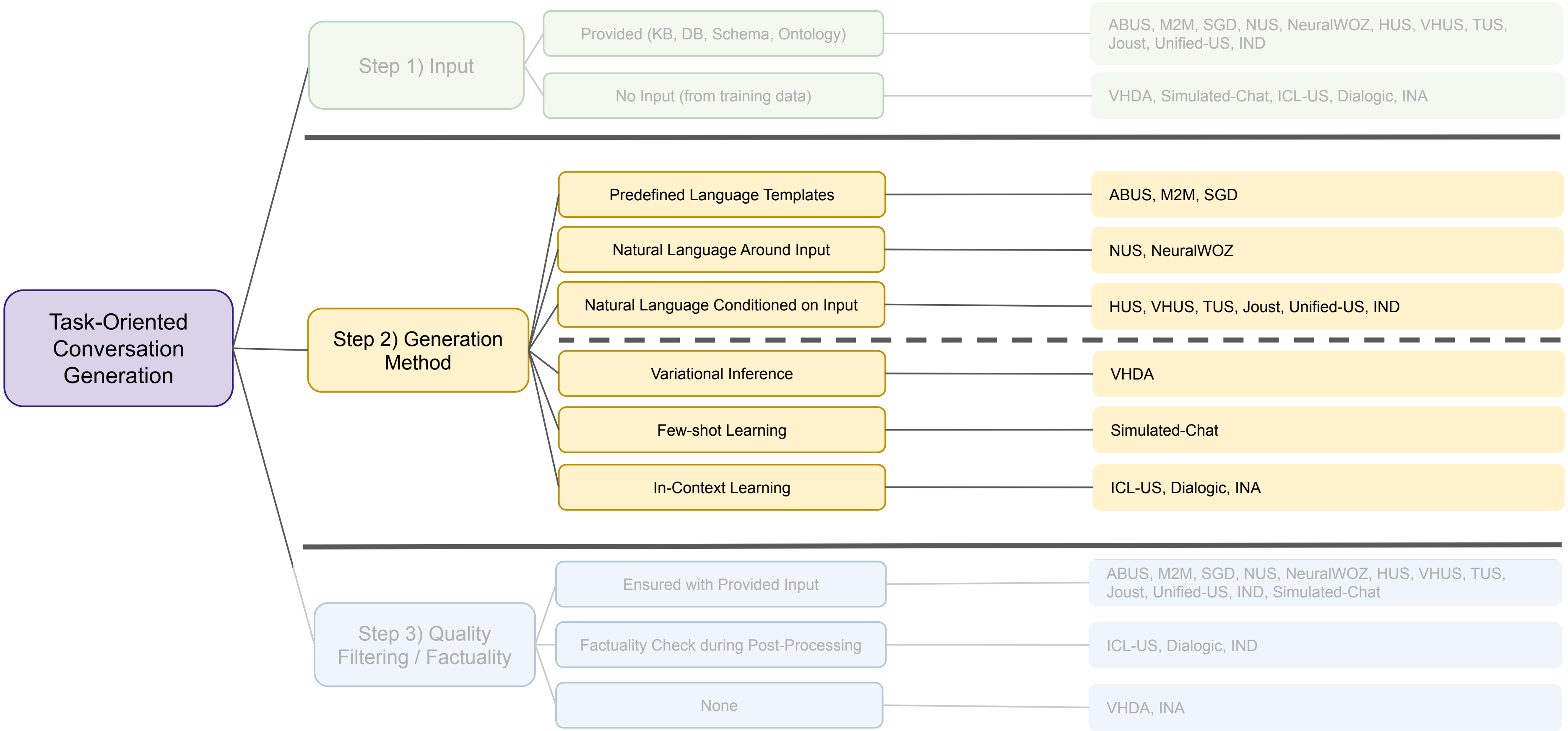
- If provided, slots and slot values are plugged into the dialogue system and natural language are generated around/conditioned on them
- Guarantees factuality

ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND

Not Provided:

- It not provided, the dialogue system must learn them through training
- Does NOT guarantee factuality

VHDA, Simulated-Chat, ICL-US, Dialogic, INA


Task-Oriented Conversation Generation
Step 1) Input
Provided (KB, DB, Schema, Ontology)
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND
No Input (from training data)
VHDA, Simulated-Chat, ICL-US, Dialogic, INA
Step 2) Generation Method
Predefined Language Templates
ABUS, M2M, SGD
Natural Language Around Input
NUS, NeuralWOZ
Natural Language Conditioned on Input
HUS, VHUS, TUS, Joust, Unified-US, IND
Variational Inference
VHDA
Few-shot Learning
Simulated-Chat
In-Context Learning
ICL-US, Dialogic, INA
Step 3) Quality Filtering / Factuality
Ensured with Provided Input
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND, Simulated-Chat
Factuality Check during Post-Processing
ICL-US, Dialogic, IND
None
VHDA, INA

Component 2: Generation Method

Step 2) Generation Method

Generation Method	User Simulators	Input
Predefined Language Templates	ABUS, M2M, SGD	Provided Input (intents, slots, slo values)
Natural Language Around Input	NUS, NeuralWOZ	
Natural Language Conditioned on Input	HUS, VHUS, TUS, Joust, Unified-US, IND	
Variational Inference	VHDA	No Input (training data in the form of dialogues)
Few-shot Learning	Simulated-Chat	
In-Context Learning	ICL-US, Dialogic, INA	

Generation Method - Predefined Language Templates

- Access to intents, slots and slot values that are plugged-in predefined templates
- Referred to as **agenda-based** approaches
- **Task-dependent!**
- Follow a predetermined set of templates (**outlines**) for generating dialog turns
- Example:
  - the **intent** *<book_movie>* can be associated with template *"Book movie with [**name**="value"] and [**date**="value"]”*
  - for Inform<**intent**=*book_movie*, **name**=*Inside Out*, **date**=*tomorrow*>
  - The template is filled and generates the turn *"Book movie with **name** Inside Out and **date** is tomorrow.”*
- Paraphrasing can be added to generate more diverse human-like turns:
  - *“I want to buy tickets for Inside Out for tomorrow”*

Generation Method - Predefined Language Templates

ABUS (Li et al., 2017)

- Input: agenda and example dialogues
  - agenda is used as a stack-like representation for user states
  - example dialogues are used for training a simulator
- Simulator: using RL policy
- Challenge addressed: training dialogue systems to respond accurately and in-real time

Generation Method - Predefined Language Templates

M2M (Shah et al., 2018)

- Input: multiple agendas and task specification (it has access to multiple APIs, each API has a task-dependent agenda)
- Simulator: using RL policy
- Challenge addressed: enhances generalizability by allowing to scale to new tasks and domains if provided a new API

Generation Method - Predefined Language Templates

SGD (Rastogi et al., 2020)

- Input: multiple agendas and task specification (it has access to multiple APIs, each API has a task-dependent agenda)
- Simulator: using RL policy
- Challenge addressed: in the real world, multiple services have overlapping functionality. The authors build a single unified model for all services by having dynamic APIs that allow for sharing knowledge between services.
- Spans over 26 services, 16 domains, resulting in a 16k dialogue dataset
- They use crowdsourcing for paraphrasing

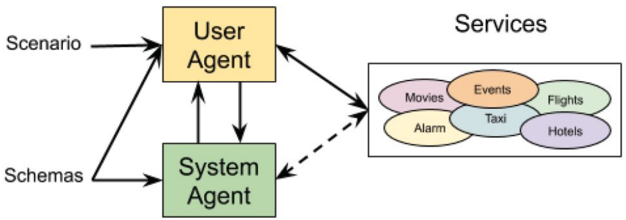


Figure 2: The overall architecture of the dialogue simulation framework for generating dialogue outlines.

Generation Method - Natural Language Around Input

- Access to intents, slots and slot values that are plugged-in **generated** natural language utterances
- No language templates
- Natural Language Generation -> generations are more versatile
- Requires less human-involvement

NUS / NeuralWOZ (Kreyssig et al., 2018, Kim et al., 2021)

- Eliminates **hand-crafted templates,** but still uses API calls
- Corpus-driven
- (NUS) Dynamic goal generation: the system can dynamically change the goal, assuming the user would want to shift their goal mid-conversation

Generation Method - Natural Language Conditioned on Input

- Access to intents, slots and slot values that are used as **input** to **generate** natural language utterances

HUS (Gür et al., 2021)

- Same family as ABUS and NUS
- Employs a multifaceted encoding scheme: it encodes different features in different vector representations (the user goal, the current dialogue turn, the dialogue history)

VHUS (Gür et al., 2021)

- HUS but created more human-like generations
- How? HUS is deterministic, while VHUS introduces variability through variational inference
- VHUS models the dialog latent space without affecting the slots and values extracted from a KB

Generation Method - Natural Language Conditioned on Input

TUS (Lin et al., 2021)

- Similarly to VHUS, TUS maps different inputs to different representations in the feature space
- BUT it is domain-agnostic
- By adding a *domain and slot index feature* representation that can be changed

JOUST (Tseng et al., 2021)

- Simulator: two pre-trained agents, fine-tuned using RL
- Novelty is added by fine-tuning on multi-domain dialogues

Generation Method - Natural Language Conditioned on Input

JOUST (Tseng et al., 2021)

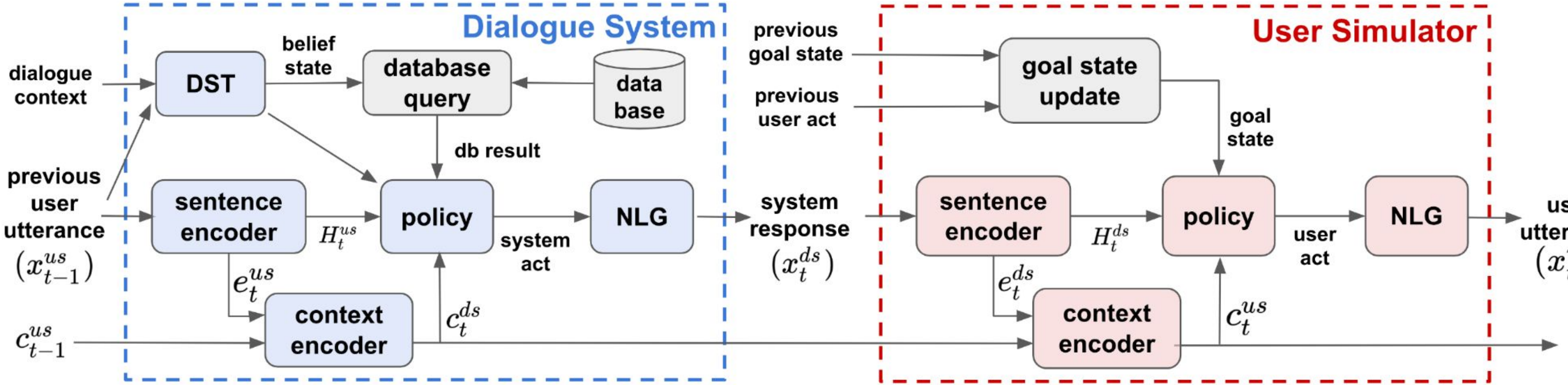


Figure 1: Overall architecture of the proposed framework, where the dialogue system (DS) and user simulator (US) discourse with each other. t denotes dialogue turn index. The context encoder is shared between the two agents.

Generation Method - Natural Language Conditioned on Input

INA (Ahmad et al., 2023)

- Simulator: two pre-trained agents, fine-tuned using RL
- Negotiation in a win-win manner, meaning each party must understand the other's needs and goal is mutual satisfaction
- Generates a Negotiation Dialogue dataset using negotiation-specific intents
- Novelty: adds negotiation-inteints such as *Negotiate-Price-Decrease*, *Add-X*, ..
- Data correction with human-in-the-loop for quality check
- Uses GPT-J for generation
- Challenge: negotiation strategies are highly context-dependent, so it adds a layer of complexity compared to the previous approaches

Generation Method - Variational Inference

VHDA (Yoo et al., 2020)

- **NO predefined knowledge**
- Input: human-generated dialogues
- Models latent variables over all dialogue aspects similar to VHUS, and TUS, but this time also for learning **intents**, **slots** and **slot values**
- Allows for the model to generate attributes beyond the training data
- However, there is no guarantee these generations are valid (we will discuss this more in part 3 of this section)

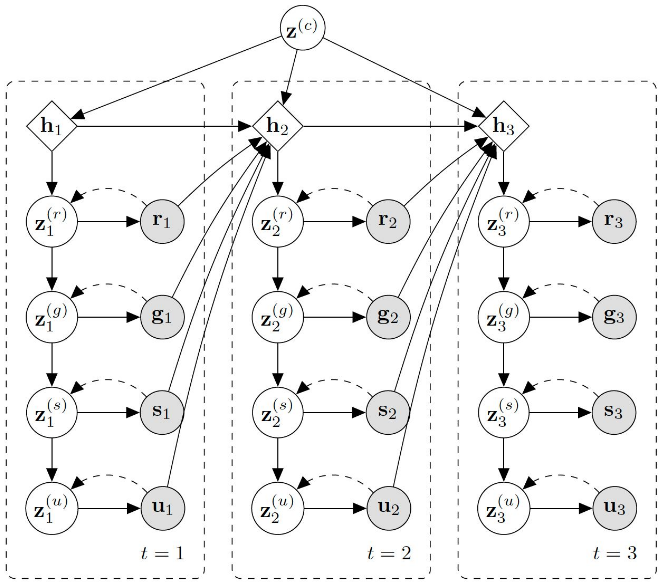


Figure 1: Graphical representation of VHDA. Solid and dashed arrows represent generation and recognition respectively.

Generation Method - Few-shot learning

Simulated-Chat (Mohapatra et al., 2021)

- **NO predefined knowledge**
- Few-shot learning: the ability of a model to generalize when provided a very small dataset for **training** or **fine-tuning**
- Input: set of instruction based on which an LLM can generate dialogues
- Uses GPT-2 and Longformer
- First receives human-generated dialogues, then self-generated simulated dialogues

Generation Method - In-context learning

ICL-US (Terragni et al., 2023)

- **NO predefined knowledge**
- In-context learning: the ability of a model to generalize when provided a very few examples in the input prompt without explicitly **training** or **fine-tuning**
- Input: set of instruction based on which an LLM can generate dialogues and example dialogues

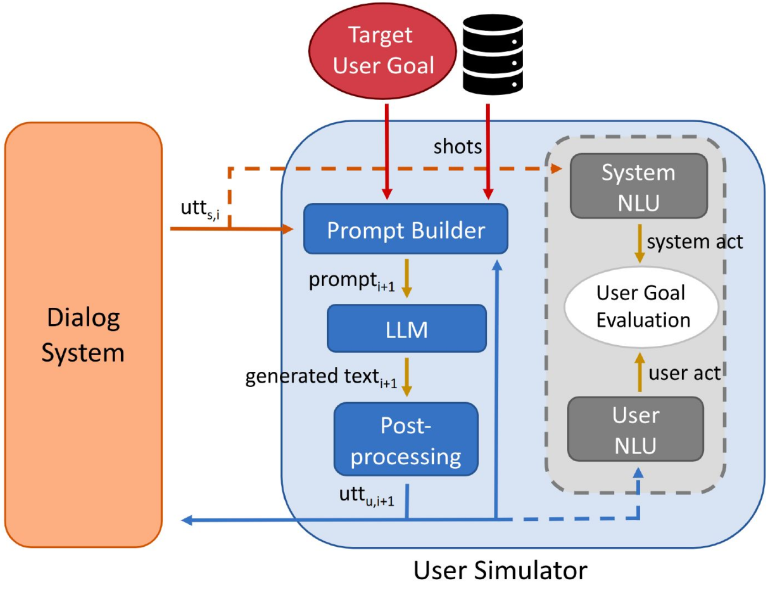


Figure 1: System and user simulator architecture sketch.

Generation Method - In-context learning

Dialogic (Li et al., 2023)

- **NO predefined knowledge**
- Input: set of instruction based on which an LLM can generate dialogues and example dialogues
- In-context learning: the ability of a model to generalize when provided a very few examples in the input prompt without explicitly **training** or **fine-tuning**

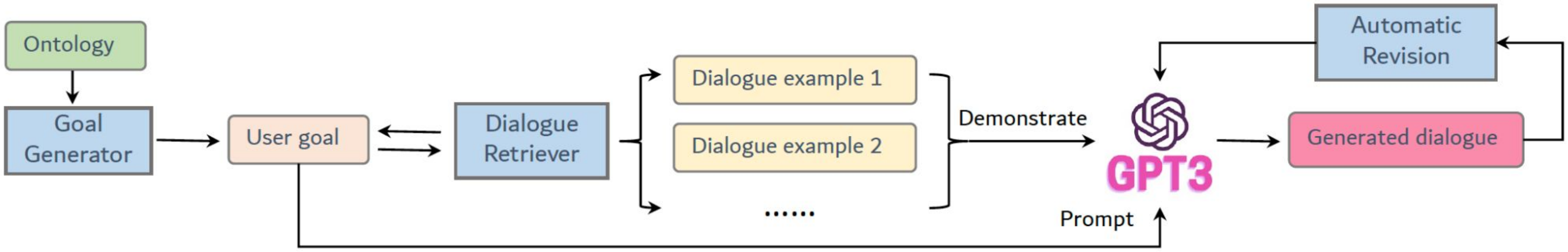
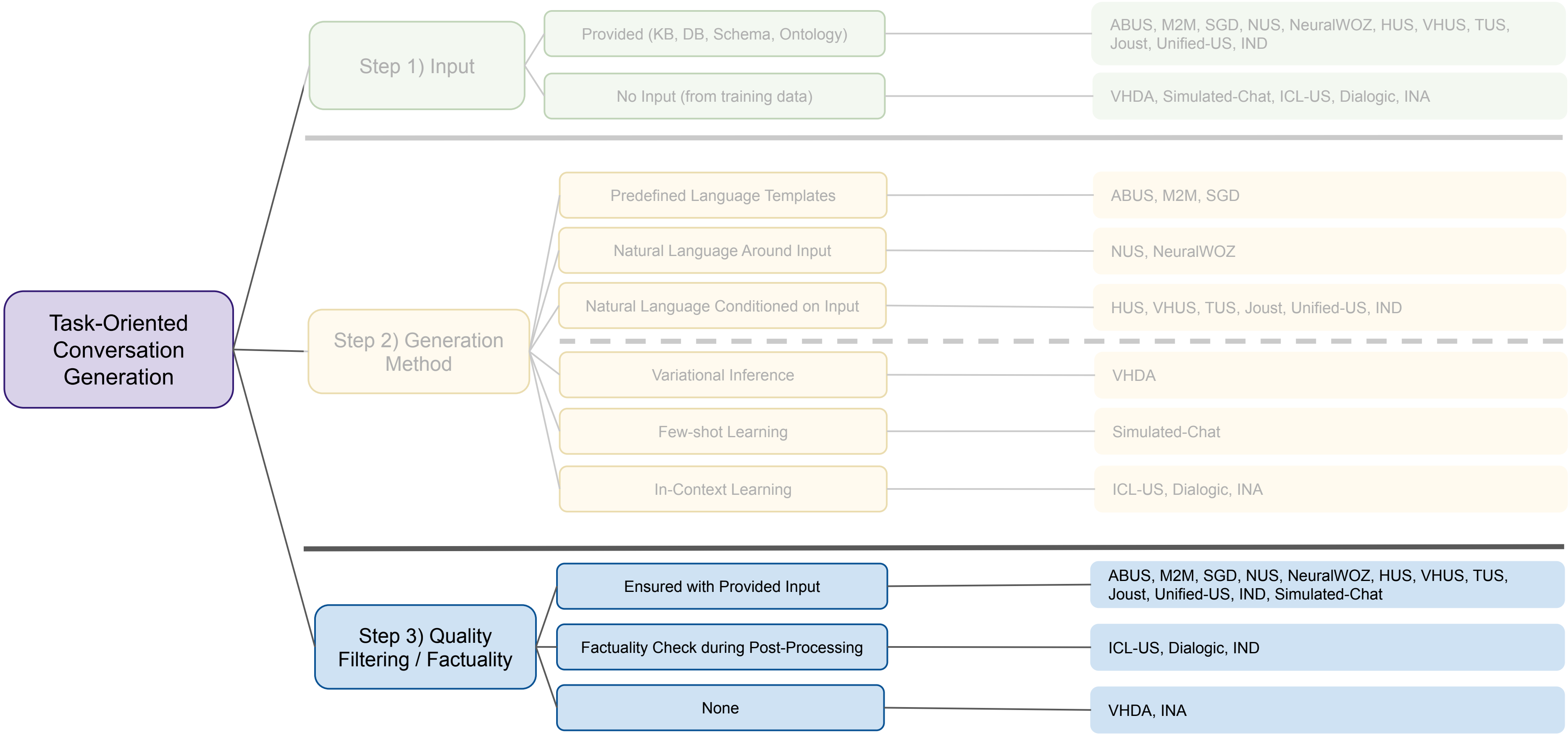


Figure 2: Overview of the proposed method.



Task-Oriented Conversation Generation
Step 1) Input
Provided (KB, DB, Schema, Ontology)
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND
No Input (from training data)
VHDA, Simulated-Chat, ICL-US, Dialogic, INA
Step 2) Generation Method
Predefined Language Templates
ABUS, M2M, SGD
Natural Language Around Input
NUS, NeuralWOZ
Natural Language Conditioned on Input
HUS, VHUS, TUS, Joust, Unified-US, IND
Variational Inference
VHDA
Few-shot Learning
Simulated-Chat
In-Context Learning
ICL-US, Dialogic, INA
Step 3) Quality Filtering / Factuality
Ensured with Provided Input
ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, TUS, Joust, Unified-US, IND, Simulated-Chat
Factuality Check during Post-Processing
ICL-US, Dialogic, IND
None
VHDA, INA

Component 3: Quality Filtering

Ensured with Provided Input

- When extracting slots and slot values from Ontology, Schema, KG and DB, factuality is granted
- ABUS, M2M, SGD, NUS, NeuralWOZ, HUS, VHUS, Joust, Unified-US

None

- Although uncommon, approaches such as VHDA ensure semantic logic of dialogue turn but does not constrict, or edit generations given lack of factuality or lack of plausible interactions.

Component 3: Quality Filtering

Factuality Check during post-processing

- Methods that discover slots and slot values in the latent space
- Dialogic has a step called automatic revision, where it corrects for potential errors by comparing GPT-3 generated belief states with the current utterance; The errors can be either due to de-generation or over-generation
- ICL-US adds an evaluation step by comparing all dialogue act extracted from the generated system and User NLU competent at each turn

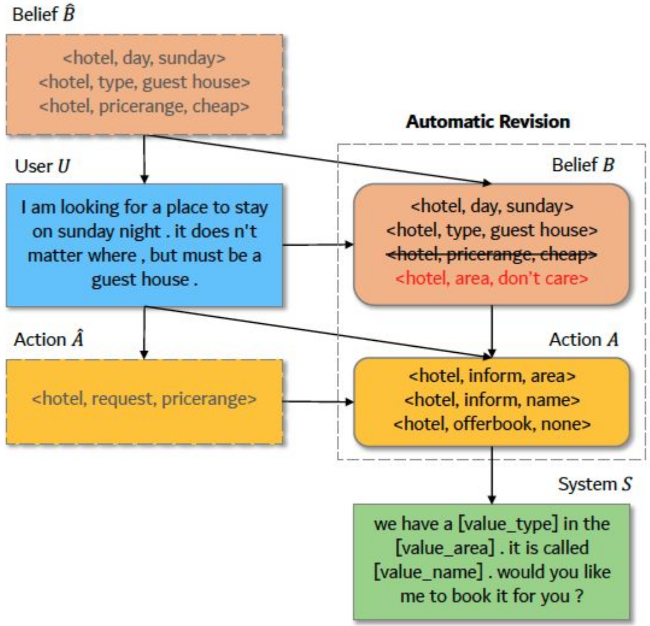


Figure 5: Illustration of the controllable generation process of a dialogue turn. An example of the generation process of a complete dialogue is shown in Appendix C.1 as Table 9.

Part 3: Conversation Generation - Open Domain

Duration: 30 mins

Presenter: Heydar Soudani

Open Domain Dialogue (ODD) System

Definition

- Engage users in conversations across a wide variety of topics
  - without being confined to specific tasks or domains

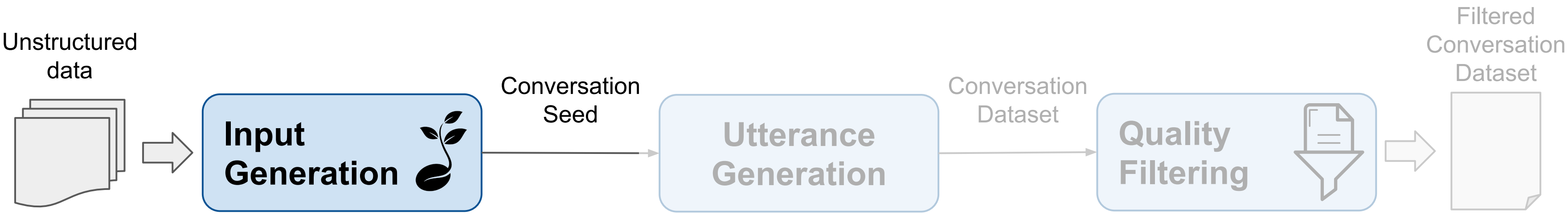
(Ni et al. 2023)

Key Features of ODD

- **Coherence:** Conversation's turns meaningfully connect to each other
- **Diversity:** Avoid bland and repetitive responses & encourage engaging interactions
- **Generality:** Encompass a broad spectrum of topics
- **Informativeness:** Elicit informative responses, knowledgeable and relevant conversations

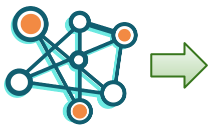
(Mehri et al., 2020) (Hwang et al., 2022) (Hwang and Lee, 2022)

ODD Data Generation



Example

(Kim et al., 2023)

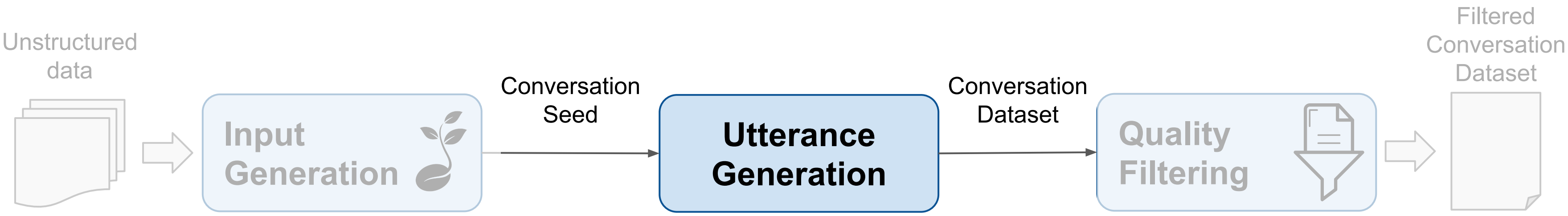


Conversation Seed

Participants:
Madeleine, Coach

Description:
Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.

ODD Data Generation



Example

(Kim et al., 2023)

Conversation Seed

Participants:
Madeleine, Coach

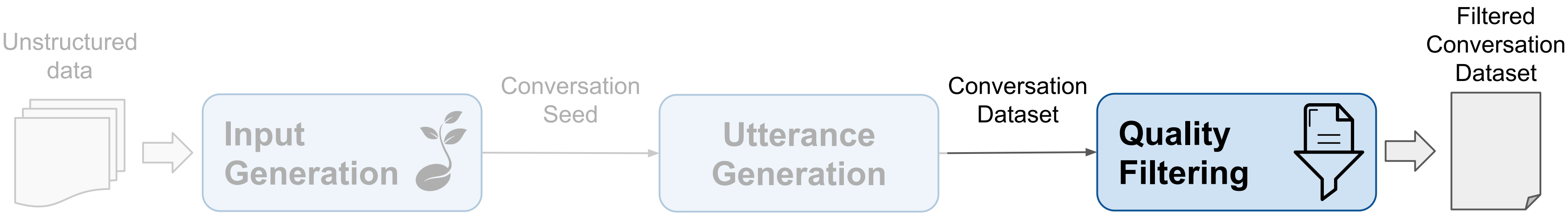
Description:
Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.

Prompt GPT-3.5 to generate a full, multi-turn conversation

Conversation Sample

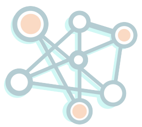
Madeleine: Hey coach, I wanted to talk to you about my performance today. I was really pushing …
Coach: Well Madeleine, you're progressing nicely. You've come a long way since we first started working together. But if you want to reach your full potential, …
Madeleine: I know that. And I'm willing to put in the work. It's just that …
Coach: It could be a number of things, Madeleine. But don't worry, we'll figure it …
Madeleine: Alright, coach. Thanks for the talk.
Coach: No problem. See you at practice tomorrow

ODD Data Generation



Example

(Kim et al., 2023)



Conversation Seed

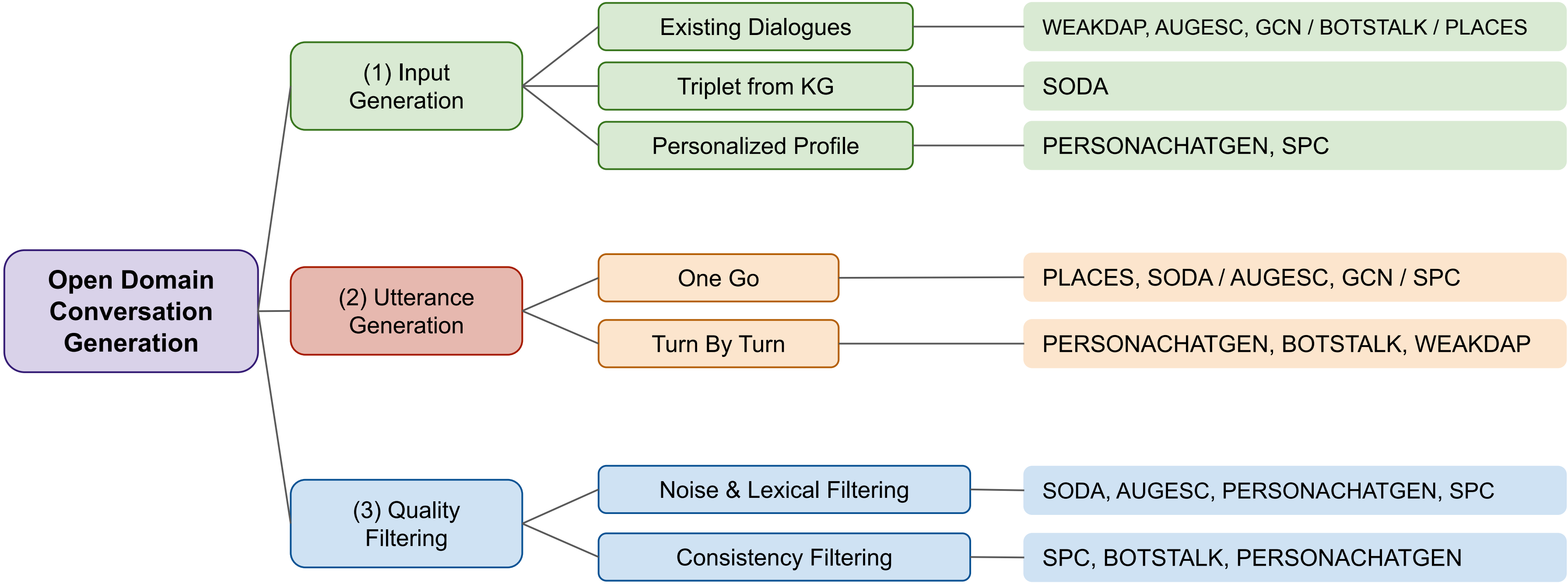
Participants:
Madeleine, Coach

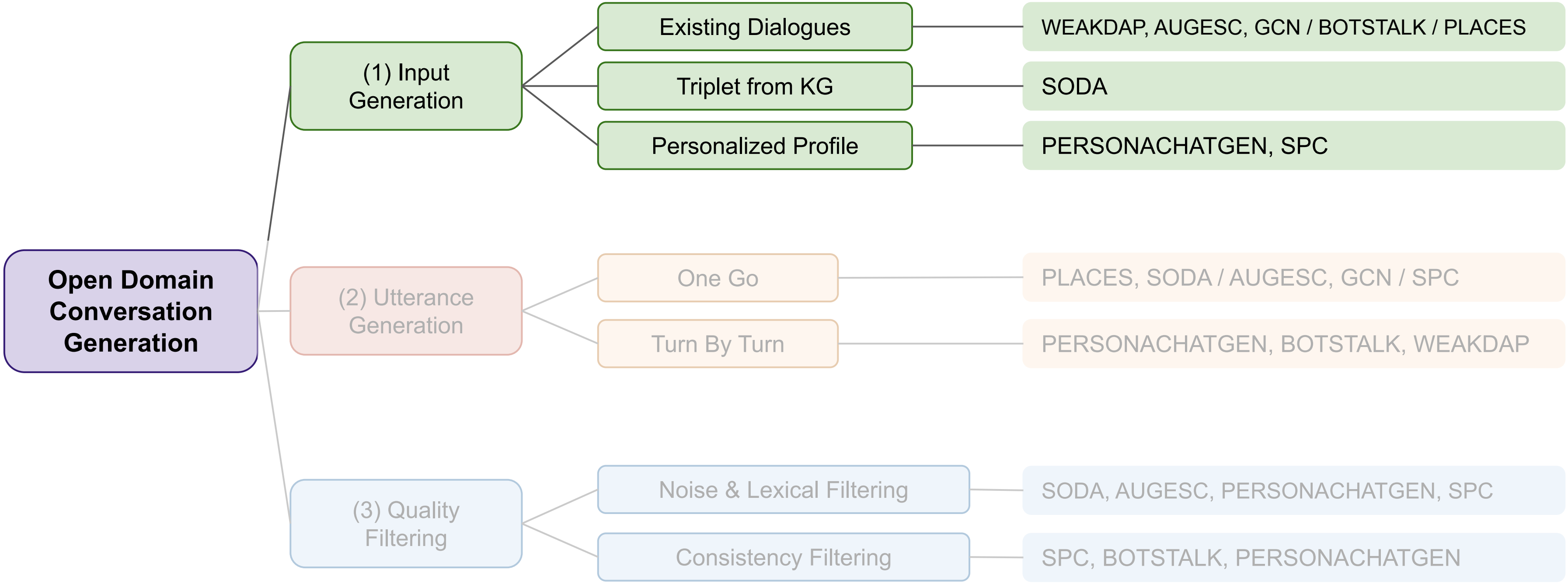
Description:
Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.

Prompt GPT-3.5 to generate a full, multi-turn conversation

Conversation Sample

Madeleine: Hey coach, I wanted to talk to you about my performance today. I was really pushing myself and I think I did pretty well. But I'm still not quite where I want to be.
Coach: Well Madeleine, you're progressing nicely. You've come a long way since we first started working together. But if you want to reach your full potential, there's still some work to be done.
Madeleine: I know that. And I'm willing to put in the work. It's just that sometimes I feel like I'm not making as much progress as I should be. Maybe I'm not training hard enough? Or maybe my technique is off?
Coach: It could be a number of things, Madeleine. But don't worry, we'll figure it out together. Let's just keep working hard and see how things go.
Madeleine: Alright, coach. Thanks for the talk.
Coach: No problem. See you at practice tomorrow


Open Domain Conversation Generation
(1) Input Generation
Existing Dialogues
WEAKDAP, AUGESC, GCN / BOTSTALK / PLACES
Triplet from KG
SODA
Personalized Profile
PERSONACHATGEN, SPC
(2) Utterance Generation
One Go
PLACES, SODA / AUGESC, GCN / SPC
Turn By Turn
PERSONACHATGEN, BOTSTALK, WEAKDAP
(3) Quality Filtering
Noise & Lexical Filtering
SODA, AUGESC, PERSONACHATGEN, SPC
Consistency Filtering
SPC, BOTSTALK, PERSONACHATGEN


Open Domain Conversation Generation
(1) Input Generation
Existing Dialogues
WEAKDAP, AUGESC, GCN / BOTSTALK / PLACES
Triplet from KG
SODA
Personalized Profile
PERSONACHATGEN, SPC
(2) Utterance Generation
One Go
PLACES, SODA / AUGESC, GCN / SPC
Turn By Turn
PERSONACHATGEN, BOTSTALK, WEAKDAP
(3) Quality Filtering
Noise & Lexical Filtering
SODA, AUGESC, PERSONACHATGEN, SPC
Consistency Filtering
SPC, BOTSTALK, PERSONACHATGEN

Input Generation

- What is the conversation seed and why has it been defined for the generation process
- **Conversation Seed:** An information card containing a main topic, subtopics, and key details about the topic
- The conversation about this topic is going to take place

(Kim et al., 2023), (Zheng et al., 2023)

Example from ESConv dataset
(Liu et al., 2021)

Conversation Seed

Experience type: Previous experience

Emotion type: Anxiety

Problem type: job crisis

Situation: I hate my job but I am scared to quit and seek a new career

Conversation Sample

Supporter: Hello, what would you like to talk about?

Seeker: I am having a lot of *anxiety* about *quitting my current job*. It is *too stressful* but pays well.

Supporter: What makes your *job stressful* for you?

Seeker: I have to deal with many *people in hard financial situations* and it is *upsetting*.

Supporter: Do you help your clients to make it to a better financial situation?

...

Input Generation - Existing Dialogues

Using background information directly

The task description and starting utterance are selected from existing dataset

AugESC
(Zheng et al., 2023)

The following is a conversation with an AI assistant. The assistant is helpful, empathetic, clever, and very friendly. It can use various support skills to provide emotional support to human.

Human: I moved into a new state recently, and there's a lot to do, but I don't have any friends in the new place I stay at.
AI: What's it like being away from family?
Human: Family is all I have here. They aren't exactly close, so I haven't gotten to see them in a while.
AI: That must be difficult. How do you feel about where you live?
Human: It's OK. I'm learning to like it a little bit. At least now I have someone who is usually around when I wake up.
AI: If only you were that lucky with people in general. People move for so many different reasons. I've found that often when I move, I just need to adjust my social circle a little, and I find that I then end up liking where I am.
Human: That's true. Maybe I should just find some people to hang out with.

Human generates background information

Given a list of topics and tasks, humans are asked to generate background info and some dialogue samples

PLACES(Chen et al., 2022)

Topic: Relationships
Background info: Bob got engaged

<Conversation 0>
The following is a conversation between Alice and Bob about relationships. Bob recently got engaged.
Alice: Congrats on your engagement! When do you think you will have your wedding?
Bob: Thank you!! We're thinking of having it in November.
Alice: That's amazing! Will you pick a fancy destination?
...

Input Generation - Triplet from KG

Example

- Input: A Knowledge Graph
- **SODA:** social dialogue
- Generation Technique:
  - Sample a socially relevant triplet
  - Define the conversation participants

(Kim et al., 2023)

Triplet from Atomic 10x:

- *Head:* PersonX moves a step closer to the goal
- *Relation:* xNeed
- *Tail:* to take the first step

Name participants:

Speakers: Madeleine, Coach

Input Generation - Triplet from KG

- Input: A Knowledge Graph
- Generation Technique:
  - Sample a triplet
  - Define the conversation participants
  - Convert the triplet to a sentence
  - Expand the sentence

(Kim et al., 2023)

Example

Triplet to Sentence:
Madeleine took the first step. Madeleine moves a step closer to the goal

Sentence to Description:
Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.

Conversation Seed

Input Generation - Personalized Profile

Personalized Dialogue Systems

- User Profile (UP)
- Profile Sentences (PS)
  - Contain personalized information about the user

Persona 1	**Persona 2**
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Fig. from PersonaChat dataset (Zhang et al., 2018)

Input Generation - Personalized Profile

Generation steps

1) Collect/Generate a pool of PS

- Given a list of topics, a LLM generates
- Using PSs from PersonaChat dataset
  - Prompt a LLM to generate more PS

2) Group number of PS to create a UP

- Contradiction score using NLI classifier

3) Filtering

- Heuristic: Do not follow the template

(Lee et al., 2022), (Jandaghi et al., 2023)

Example

User's persona: Want | Activity

Generate five profile sentences related to the given user's persona and the "activity" in each sentence:
1. I have always wanted to travel to Ireland or Puerto Rico. (activity: travel)
2. I hope to visit Quebec, Canada someday. (activity: travel)
3. One day I would really like to skydive. (activity: skydiving)
4. Before I die, I want to skydive. (activity: skydiving)
5. I hope to see the world with my husband. (activity: travel)

User's persona: Preference | Movie | Title

Generate five profile sentences related to the given user's persona and the "movie title" in each sentence:
1. I am a big fan of the Lord of the Rings movies. (movie title: Lord of the Rings)
2. I love all of the Harry Potter movies. (movie title: Harry Potter)
3. The Hobbit is one of my favorite movies. (movie title: The Hobbit)
4. I have seen all of the Star Wars movies. (movie title: Star Wars)
5. I enjoy watching Marvel movies. (movie title: Marvel)

Input Generation - Personalized Profile

Generation steps

1) Collect/Generate a pool of PS
   - Given a list of topics, a LLM generates
   - Using PSs from PersonaChat dataset
     - Prompt a LLM to generate more PS

2) Group number of PS to create a UP
   - Contradiction score using NLI classifier

3) Filtering
   - Heuristic: Do not follow the template

(Lee et al., 2022), (Jandaghi et al., 2023)

Example

I am studying at a community college.
I am a teacher at the high school.
"The Great Gatsby" is another book I enjoy.
I'm a big fan of the violin.
I love reading books that are full of adventure.

(a) An example of persona set containing contradiction between profile sentences

I am a very creative and imaginative person.
My older sister is a doctor.
I love to read books that are science fiction.
I enjoy watching suspenseful movies.
I have to be very careful in the springtime because of my allergies.

(b) An example of persona set containing no contradiction between profile sentences

Input Generation - Personalized Profile

Generation steps

1) Collect/Generate a pool of PS

- Given a list of topics, a LLM generates
- Using PSs from PersonaChat dataset
  - Prompt a LLM to generate more PS

2) Group number of PS to create a UP

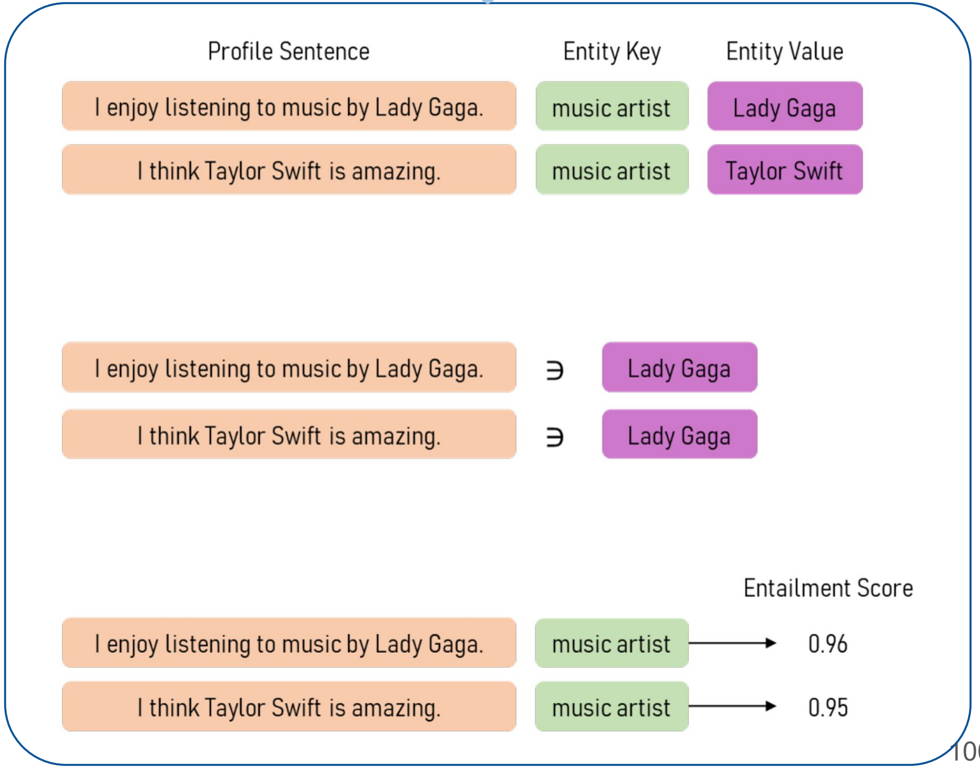
- Contradiction score using NLI classifier

3) Filtering

- Heuristic: Do not follow the template

(Lee et al., 2022), (Jandaghi et al., 2023)

Example



Component 2: Utterance Generation

Objective: Convert "Conversation Seed" to "Conversation Sample"

Example from PERSONACHATGEN
(Lee et al., 2022)

Conversation Seed

P1's Persona:

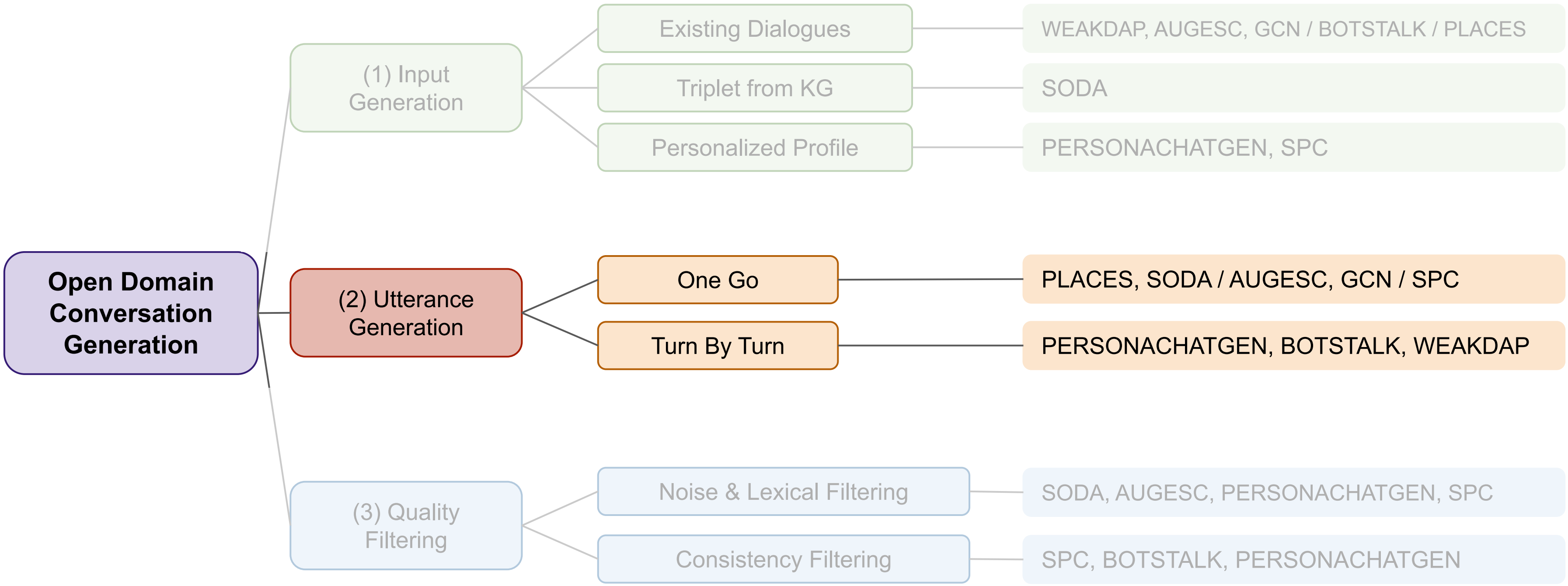
I love food and I love to eat.
I am a woman who loves fashion.
I love reality TV.
I prefer to watch comedies.
I have hay fever.

P2's Persona:

I've been to Italy three times.
I graduated from Yale.
I've read all of the books by Jodi Picoult.
I'm a big fan of books, and my favorite genre is fantasy.
I have asthma and it makes it hard to breathe sometimes.

Conversation Sample

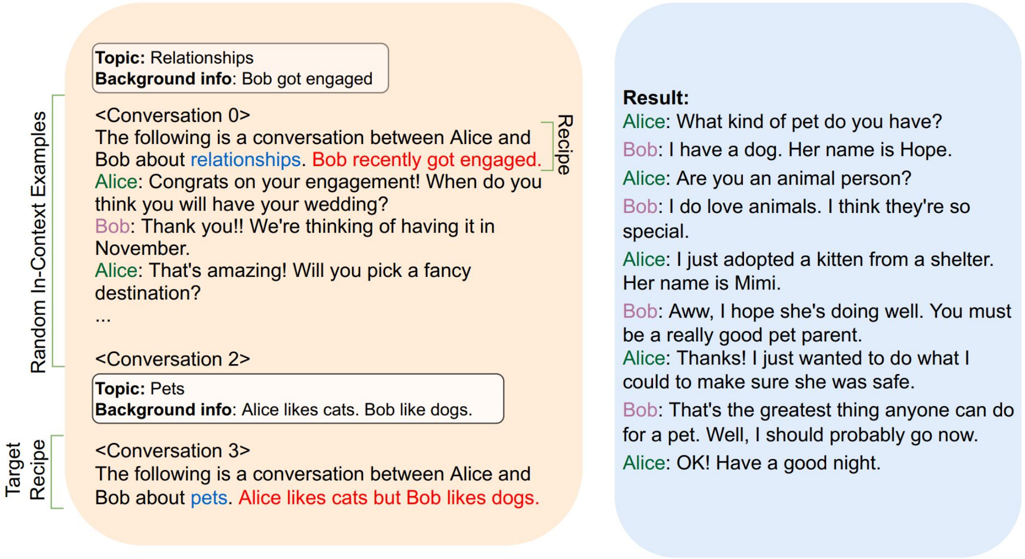
P1: Hey, how are you?
P2: I'm good, just got back from Italy.
P1: Oh, how was it?
P2: It was great, I love it there.
P1: What did you eat?
P2: Everything, I love Italian food.
P1: That sounds amazing, I need to go there sometime.
P2: Yeah, you should definitely go. It's a great place.
P1: I also love fashion, do you follow any trend?
P2: No, not really. I prefer classic styles.
P1: That's cool, I like to watch reality TV.
P2: Yeah, I don't really watch much TV. I prefer to read.
P1: That's great, what do you like to read?
P2: I like to read books, my favorite genre is fantasy.
P1: That's great, I have hay fever so I can't really read outside.
P2: Yeah, I have asthma so it's hard for me to breathe sometimes.


Open Domain Conversation Generation
(1) Input Generation
Existing Dialogues
WEAKDAP, AUGESC, GCN / BOTSTALK / PLACES
Triplet from KG
SODA
Personalized Profile
PERSONACHATGEN, SPC
(2) Utterance Generation
One Go
PLACES, SODA / AUGESC, GCN / SPC
Turn By Turn
PERSONACHATGEN, BOTSTALK, WEAKDAP
(3) Quality Filtering
Noise & Lexical Filtering
SODA, AUGESC, PERSONACHATGEN, SPC
Consistency Filtering
SPC, BOTSTALK, PERSONACHATGEN

Utterance Generation - One Go

1) ICL: Only prompting, wo. Fine tuning

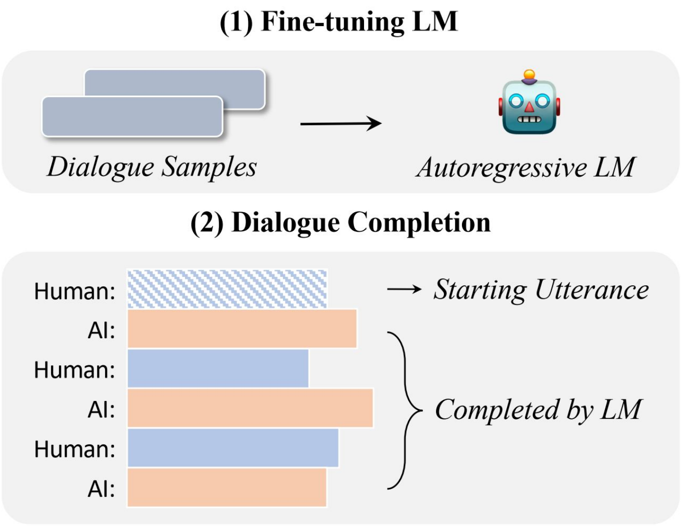
(Chen et al., 2022)



Utterance Generation - One Go

2) First Fine-tune on Dialogue Completion task, then prompt

(Zheng et al., 2023)



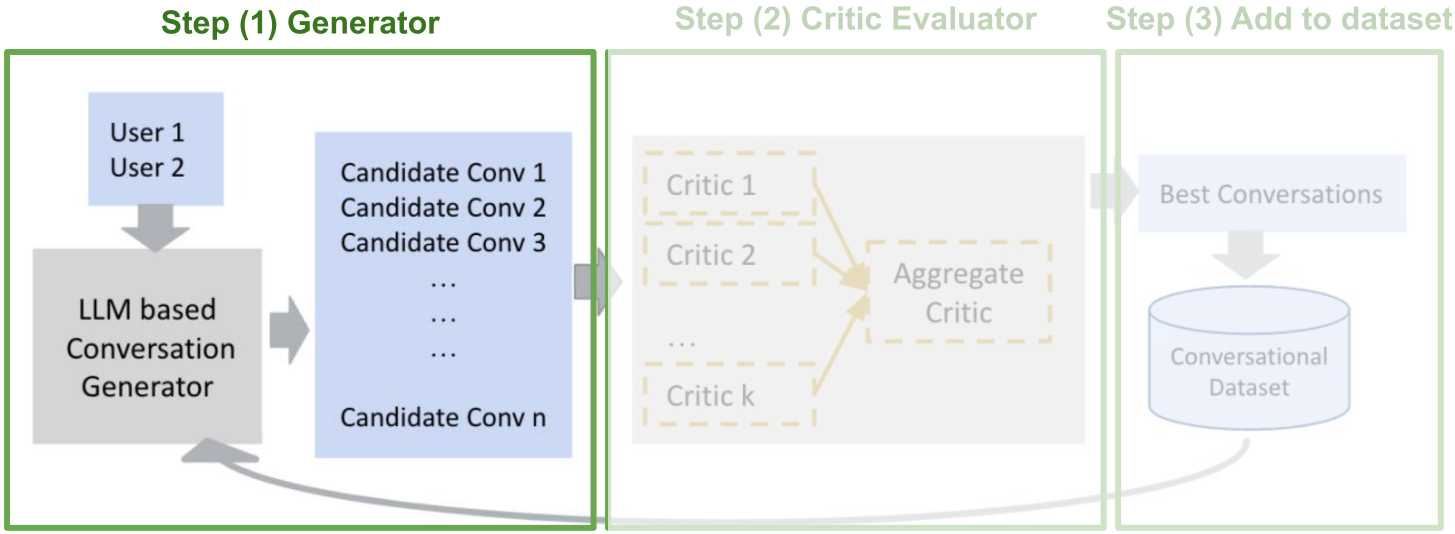
The following is a conversation with an AI assistant. The assistant is helpful, empathetic, clever, and very friendly. It can use various support skills to provide emotional support to human.

Human: I moved into a new state recently, and there's a lot to do, but I don't have any friends in the new place I stay at.
AI: What's it like being away from family?
Human: Family is all I have here. They aren't exactly close, so I haven't gotten to see them in a while.
AI: That must be difficult. How do you feel about where you live?
Human: It's OK. I'm learning to like it a little bit. At least now I have someone who is usually around when I wake up.
AI: If only you were that lucky with people in general. People move for so many different reasons. I've found that often when I move, I just need to adjust my social circle a little, and I find that I then end up liking where I am.
Human: That's true. Maybe I should just find some people to hang out with.

Utterance Generation - One Go

3) Prompting with Generator-Critic Architecture

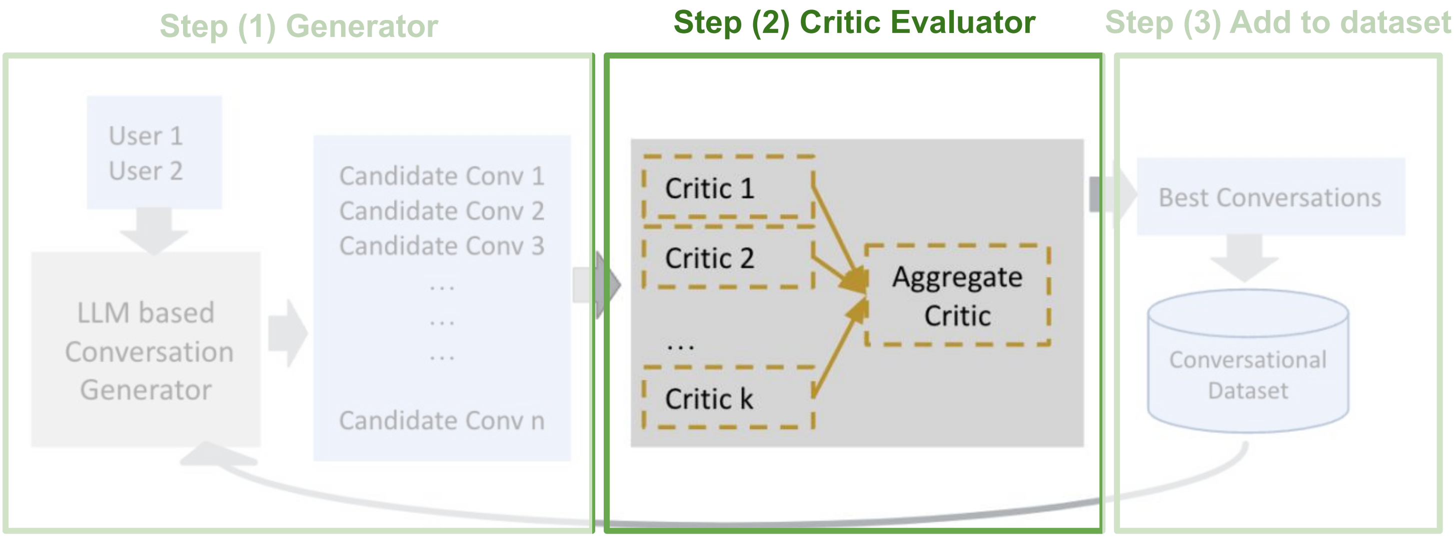
(Jandaghi et al., 2023)



Utterance Generation - One Go

3) Prompting with Generator-Critic Architecture

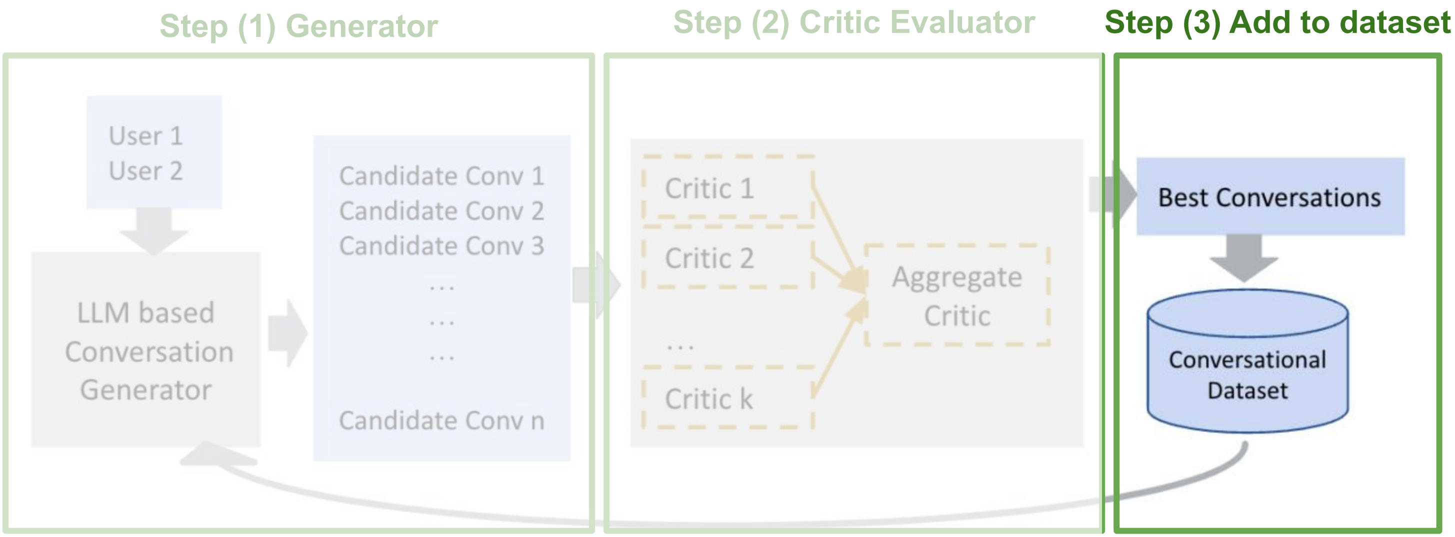
(Jandaghi et al., 2023)



Utterance Generation - One Go

3) Prompting with Generator-Critic Architecture

(Jandaghi et al., 2023)



Utterance Generation - Turn-by-Turn

(1) Two persona Profiles

- two LLMs, User simulation (Lee et al., 2022)

Reason 2: Merging multiple conversation datasets (Kim et al., 2022)

Person A	Person B
Skill context from ConvAI2 I like to play soccer; I like to read; ...	**Skill context from ConvAI2** I have 3 children; I am a karate black belt; ...
Skill context from WoW Nike Inc.	**Skill context from WoW** Nike Inc.; ... multinational corporation ... Air Jordan ...
Skill context from ED I really like this girl at my job, but I am ... ; Apprehensive	**Skill context from ED** None

Dialogue

A: Do you have much experience using the different types of cleats? Which do you like best? (P)
B: I have a little. I also know about Air jordans, a brand of footwear also popular with athletes. (K)
A: I enjoy Air jordans as well. I like to play soccer, and it's extremely hard to get good footwear. (P)
B: I agree. Air jordans are good for calves, and it's really easy to pull off. (K)
A: Air jordans are generally made with the most material so it makes sense they'd be easy to use. (K)
B: And now shoes can go together with clothes as well, like any other type of material. (K)
A: Do you know much about aeros then? They have excellent fit and beauty. (K)
B: I have a hard time finding it but they are great shoes. (P)
A: I hope you can find ones that are comfortable to you. (E)
B: Yes. I hope you can get those shoes too. (E)

Utterance Generation - Turn-by-Turn

Reason 3: More diversity and quantity

Trajectory Augmentation

All-turn Augmentation

Last-turn Augmentation

(Chen et al., 2022)

Output

Original Conversation

Turn 1: Alice in a happy mood: Oh, man. I had the best supper last night. My wife made a stir-fry and it was amazing!

Turn 2: Bob in a happy mood: I love stir fry crispy bitesize vegetables covered in a mixture of soy sauce and oyster sauce. Wilted greens and fresh bean sprouts. Throw in some onion and garlic and ginger! Mmm! Mmm! It's almost lunchtime. I would die for a plate of stir fry right now!

Turn 3: Alice in a neutral mood: Well, you can keep the vegetables, I'll take the meat. The stir fry my wife made was really hearty, with chunks of beef and slivers of bell peppers and onion...

Turn 4: Bob in a surprised mood: What? You call that a stir fry? More meat than vegetables? That's the worst insult you could throw at a Chinese stir fry. What disgrace to the wok she fried it in! What you had is equivalent to a fajita without the wrap!

Augmented Conversation

GT Turn 1: Alice in a happy mood: Oh, man. I had the best supper last night. My wife made a stir-fry and it was amazing!

GT Turn 2: Bob in a happy mood: I love stir fry crispy bitesize vegetables covered in a mixture of soy sauce and oyster sauce. Wilted greens and fresh bean sprouts. Throw in some onion and garlic and ginger! Mmm! Mmm! It's almost lunchtime. I would die for a plate of stir fry right now!

Generated Turn 3: Alice in a neutral mood: Stir fry is good. When my dad cooks, he makes stir fry for us.
Generated Turn 4: Bob in a surprised mood: Stir fry was was the only Asian food I had until I was ten. I've never known what it was called!

Utterance Generation - Turn-by-Turn

Reason 3: More diversity and quantity

Trajectory Augmentation

All-turn Augmentation

Last-turn Augmentation

(Chen et al., 2022)

Output

Original Conversation

Turn 1: Alice in a surprised mood: My goodness! She is thirty seconds faster than the world record in 5000 metres race.

Turn 2: Bob in a surprised mood: Excuse me, what did you say?

Turn 3: Alice in a happy mood: A Chinese girl has broken the world record in the Olympic Games.

Turn 4: Bob in a surprised mood: That's incredible. I can't believe it.

Turn 5: Alice in a happy mood: You have to. It is sure.

Turn 6: Bob in a surprised mood: So, what's her name?

Augmented Conversation 1

GT Turn 1: Alice in a surprised mood: My goodness! She is thirty seconds faster than the world record in 5000 metres race.
GT Turn 2: Bob in a surprised mood: Excuse me, what did you say?
Generated Turn 3: Alice in a happy mood: I said my goodness, that girl is fast.

Augmented Conversation 2

GT Turn 1: Alice in a surprised mood: …
GT Turn 2: Bob in a surprised mood: …
GT Turn 3: Alice in a happy mood: A Chinese girl has broken the world record in the Olympic Games.
Generated Turn 4: Bob in a surprised mood: I am dying to know who is the girl?

Augmented Conversation 3

GT Turn 1: Alice in a surprised mood: …
GT Turn 2: Bob in a surprised mood: …
GT Turn 3: Alice in a happy mood: A Chinese girl has broken the world record in the Olympic Games.
GT Turn 4: Bob in a surprised mood: That's incredible. I can't believe it.
Generated Turn 5: Alice in a happy mood: This is the first time in the history of the world

Augmented Conversation 4
…

Utterance Generation - Turn-by-Turn

Reason 3: More diversity and quantity

Trajectory Augmentation

All-turn Augmentation

Last-turn Augmentation

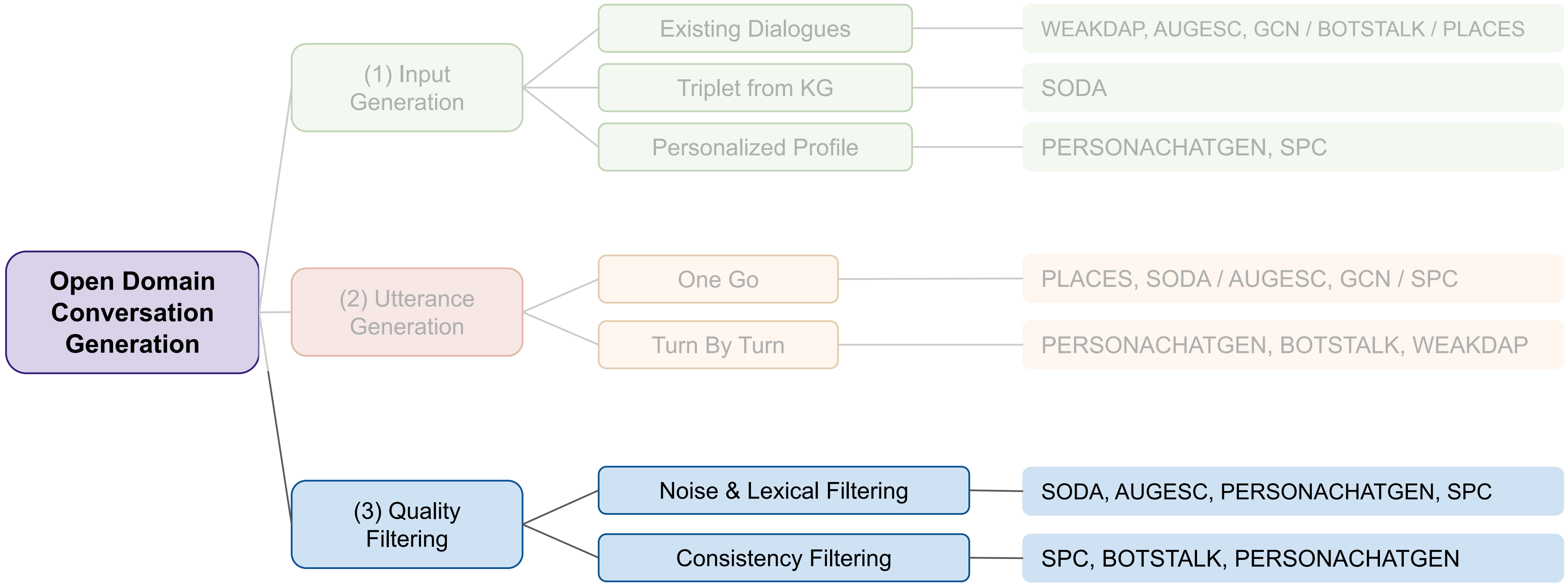
(Chen et al., 2022)

Output

Augmented Conversation

GT Turn 1: Alice informs Bob: Good morning.
GT Turn 2: Bob informs Alice: Er, good morning, yes, er...
GT Turn 3: Alice directs Bob: I'm phoning about the job that was in the paper last night.
GT Turn 4: Bob directs Alice: Oh, yes. Erm, well, could you tell me your name, please?
GT Turn 5: Alice informs Bob: Oh, Candida Fawcett.
GT Turn 6: Bob questions Alice: Oh, yes. Erm, well, what exactly is it that interests you about the job?
GT Turn 7: Alice informs Bob: Well, I just thought that it was right up my street you know.
GT Turn 8: Bob directs Alice: Really, hmmmm. Erm, well, could you perhaps tell me a little about yourself?
GT Turn 9: Alice informs Bob: Yes, arm, I'm 23. I've been working abroad, I'm um...

Generated Turn 10: Bob informs Alice: Well, perhaps you would like to think it over. Erm, and er...


Open Domain Conversation Generation
(1) Input Generation
Existing Dialogues
WEAKDAP, AUGESC, GCN / BOTSTALK / PLACES
Triplet from KG
SODA
Personalized Profile
PERSONACHATGEN, SPC
(2) Utterance Generation
One Go
PLACES, SODA / AUGESC, GCN / SPC
Turn By Turn
PERSONACHATGEN, BOTSTALK, WEAKDAP
(3) Quality Filtering
Noise & Lexical Filtering
SODA, AUGESC, PERSONACHATGEN, SPC
Consistency Filtering
SPC, BOTSTALK, PERSONACHATGEN

Component 3: Quality Filtering

Objective

- Remove conversions do not contain ODD features
  - (e.i. Correctness, Diversity, Consistency, Informativeness)

Noise & Lexical Filtering:

Checking the Correctness, Diversity

Approach: Heuristic rules

- Unfinished conversations
- Do not follow the wanted patterns
- Contain repetitive pattern
- Dangerous or Toxic context with social bias

(Kim et al., 2023), (Zheng et al., 2023), (Lee et al., 2022)

Consistency Filtering:

Checking the consistency

- Between the turns
- Between persona sentences in one user profile

Approach: NLI Classifier

(Kim et al., 2022), (Lee et al., 2022), (Jandaghi et al., 2023)

Part 4: Conversation Generation - Information Seeking

Duration: 45 min

Presenter: Heydar Soudani

Conversational Information Seeking (CIS)

- The main goal of CIS is fulfill users' information needs
- Allow users to search information using natural language dialogue, instead of traditional search queries

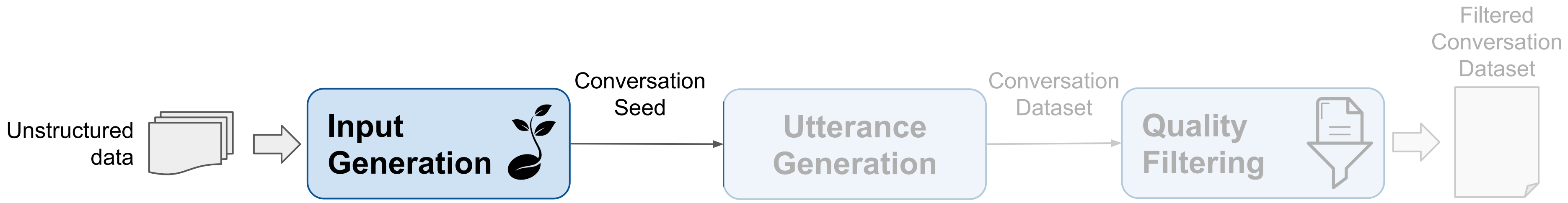
(Zamani et al., 2023)

Key Features of CIS

- Generation control, Topic shifting
- Multi-evidence answer generation
- Query ambiguity, asking clarification questions

(Wu et al., 2022), (Deng et al., 2023)

CIS Data Generation



Example

(Askari et al., 2024)

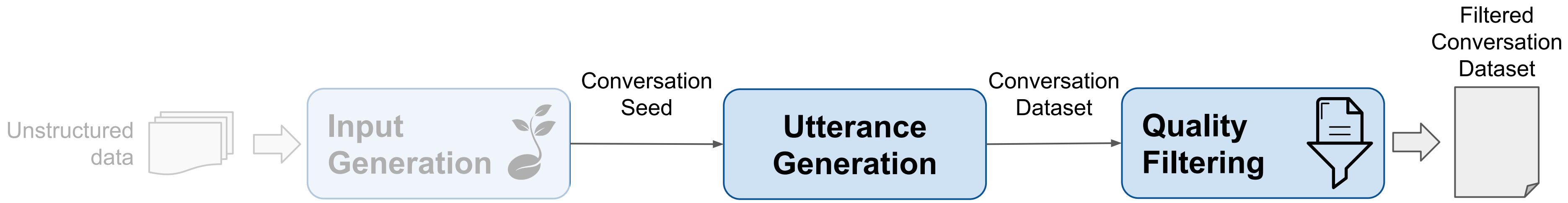
MSDialog Intents:
CQ
FD
GG
PA
IR
OQ

(Qu et al., 2018)

Conversation Seed

Entity type: Person
Entity type attribute: Occupation
Entity name: Albert Einstein
Entity background document: Albert Einstein was a German-born theoretical physicist who is ...
Conversation starter: Can you delve into the efforts and contributions of Albert Einstein in the field of physics?
Dialogue flow: [original question, clarifying question and information request, further details]

CIS Data Generation



Example

(Askari et al., 2024)

MSDialog
Intents:
CQ
FD
GG
PA
IR
OQ

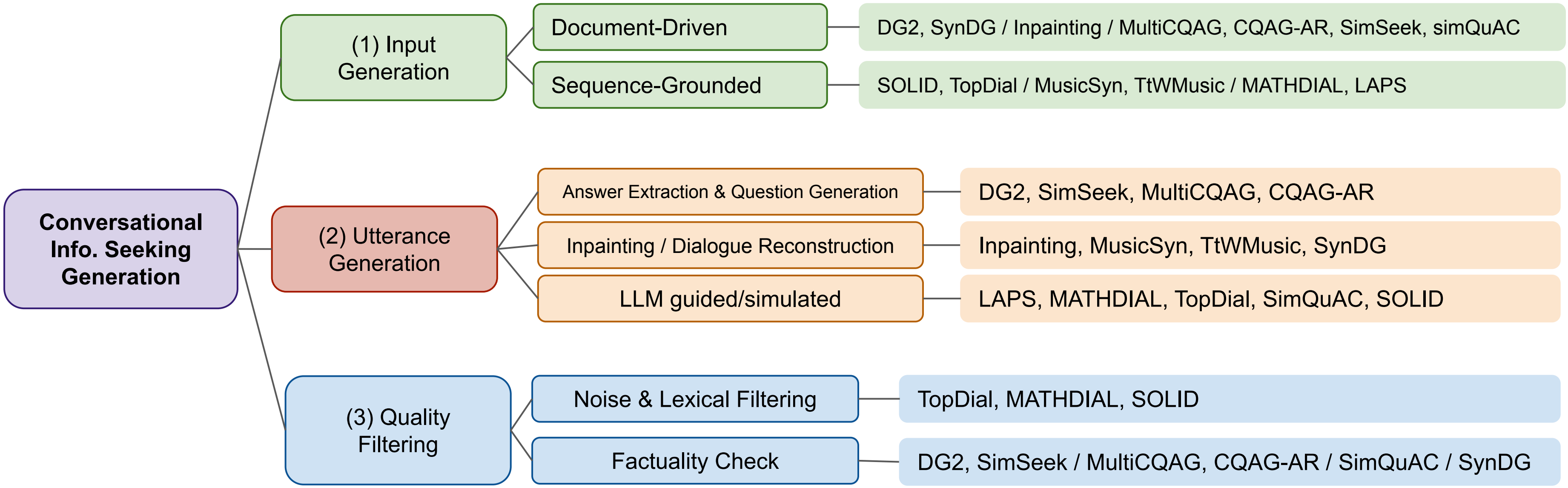
(Qu et al., 2018)

Conversation Seed

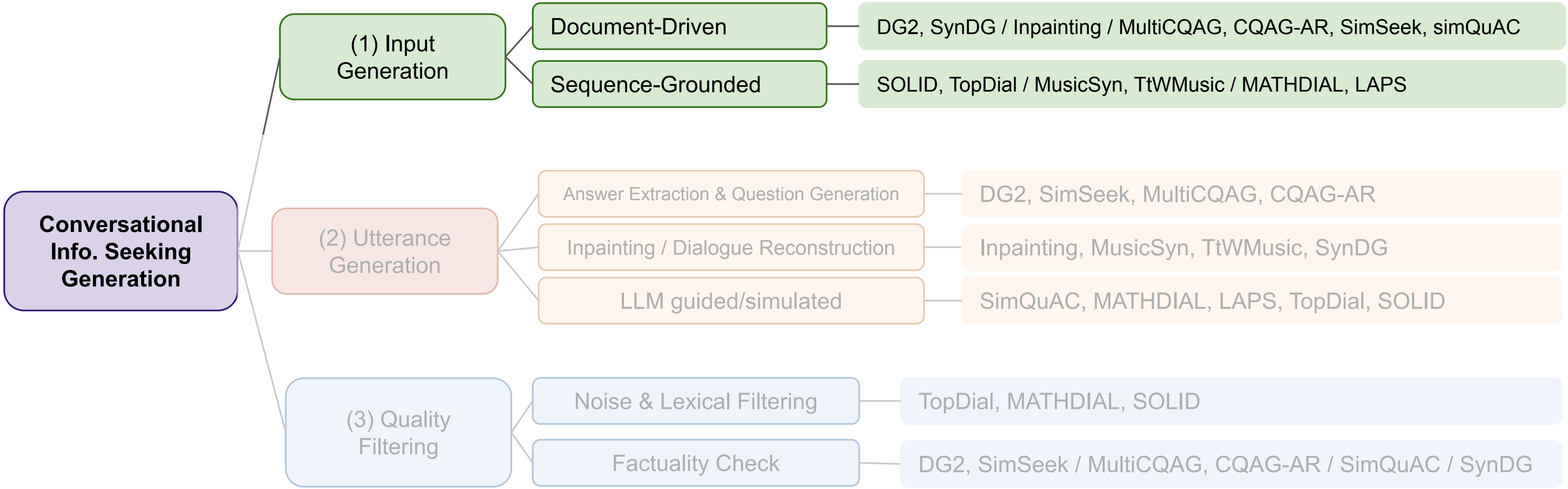
Entity type:
Entity type attribute:
Entity name:
Entity background document: ...
Conversation starter: …
Dialogue Flow: [original question, clarifying question and information request, further details]

Conversation Sample

User: Can you delve into the specific efforts and contributions made by Albert Einstein in the field of physics? (intent: original question)
Agent: Sure! Albert Einstein made groundbreaking contributions to physics, especially with his theory of relativity. What aspect would you like me to focus on or any specific topic you're interested in? (intent: clarifying question)
User: Could you provide more details about his theory of relativity and how it revolutionized our understanding of space and time? (intent: further details)
…



Conversational Info. Seeking Generation
(1) Input Generation
Document-Driven
DG2, SynDG / Inpainting / MultiCQAG, CQAG-AR, SimSeek, simQuAC
Sequence-Grounded
SOLID, TopDial / MusicSyn, TtWMusic / MATHDIAL, LAPS
(2) Utterance Generation
Answer Extraction & Question Generation
DG2, SimSeek, MultiCQAG, CQAG-AR
Inpainting / Dialogue Reconstruction
Inpainting, MusicSyn, TtWMusic, SynDG
LLM guided/simulated
LAPS, MATHDIAL, TopDial, SimQuAC, SOLID
(3) Quality Filtering
Noise & Lexical Filtering
TopDial, MATHDIAL, SOLID
Factuality Check
DG2, SimSeek / MultiCQAG, CQAG-AR / SimQuAC / SynDG


Conversational Info. Seeking Generation
(1) Input Generation
Document-Driven
DG2, SynDG / Inpainting / MultiCQAG, CQAG-AR, SimSeek, simQuAC
Sequence-Grounded
SOLID, TopDial / MusicSyn, TtWMusic / MATHDIAL, LAPS
(2) Utterance Generation
Answer Extraction & Question Generation
DG2, SimSeek, MultiCQAG, CQAG-AR
Inpainting / Dialogue Reconstruction
Inpainting, MusicSyn, TtWMusic, SynDG
LLM guided/simulated
SimQuAC, MATHDIAL, LAPS, TopDial, SOLID
(3) Quality Filtering
Noise & Lexical Filtering
TopDial, MATHDIAL, SOLID
Factuality Check
DG2, SimSeek / MultiCQAG, CQAG-AR / SimQuAC / SynDG

Input Generation

What does "Conversation Seed" contain?

- Information containing a main topic, subtopics, and key details about the topic
- ***Dialogue Flow:*** a comprehensive perspective of the conversation

Example from Doc2Dial dataset

(Feng, et al., 2020)

Conversation Seed

Title: Top 5 DMV Mistakes and How to Avoid Them

Document: Many DMV customers make easily avoidable mistakes that cause them significant problems, including …

Dialogue flow: [Query_condition, Respond_solution, Query_condition, Response_negative, query_condition, …]

Conversation Sample

Dialogue turns	Dialogue flow	Grounded Passage
I need to change my address?	Query Condition	Forgetting to Update Address (A title of section) …
You need to report it to the DMV.	Response Solution	By statute , you must report a change of address to DMV within ten days of moving. That is the case …
Is your license current?	Query Condition	Not Bringing Proper Documentation to DMV Office (A title of section) …
Yes, I just have to change the address.	Response Negative	Sign up or log into MyDMV …
What if my insurance had lapsed?	Query Condition	Understanding how Much Traffic Points Cost (A title of section) …
Your license and registration could be suspended.	Response Solution	DMV maintains a point system to track dangerous drivers. Often , motorists convicted …

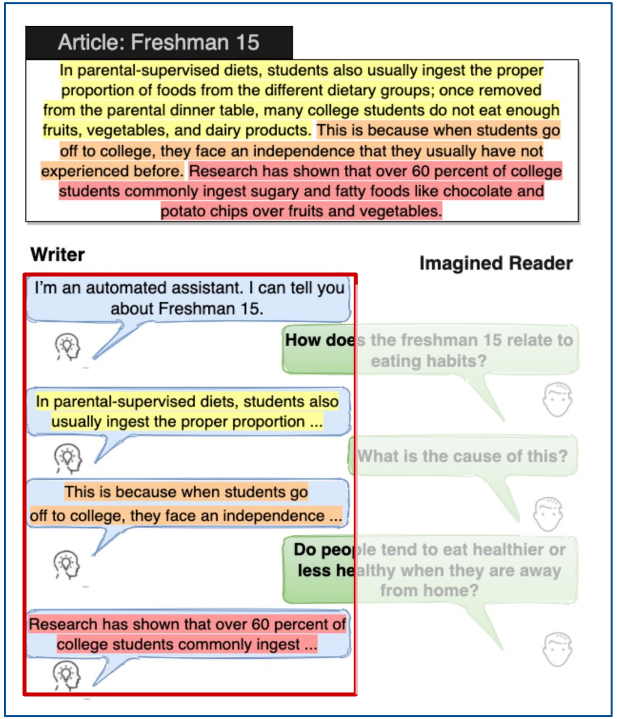
Input Generation - Document-Driven

- Why are documents used for CIS data generation?

Inpainting

- **Idea:** Documents are conceptualized as dialogues between the writer and an imaginary reader
- The dialogue flow consists directly of the document's sentences

(Dai et al., 2022)



Input Generation - Document-Driven

Document Segmentation

- A document is segmented into multiple passages
- Passage Ranker

$$\underbrace{p(c_t}_{\text{Selected passage in turn } t}|\underbrace{\{u_i, a_i\}_{i<t}}_{\text{Conversation History}}, \underbrace{C}_{\text{Document}})$$

- Not fixed and pre-defined
- Dialogue flow: a sequence of passages
- May not consist of sequential passages from a document

(Wu et al., 2022)

Top 5 DMV Mistakes and How to Avoid Them

<Passage 1> Many DMV customers make easily avoidable mistakes that cause them significant problems, …

<Passage 2>

<Passage 3> Not Bringing Proper Documentation to DMV Office. About ten percent of customers visiting a DMV office do not bring what they need to complete their transaction and see if your transaction can be …

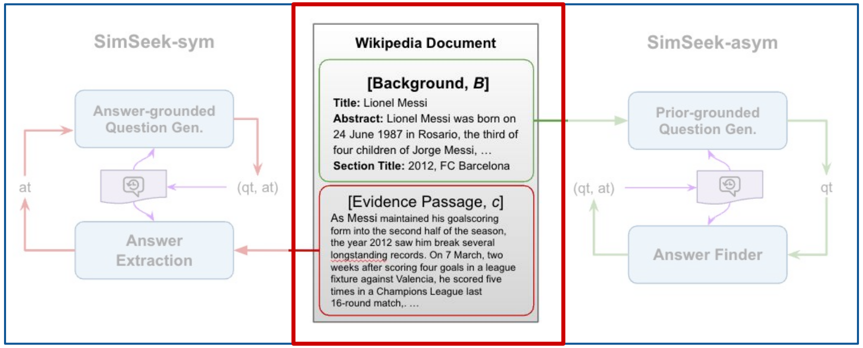
<Passage 4>

<Passage 5> We send 500,000 inquiry letters a year. If the inquiry letter does not resolve the problem, we must suspend the vehicle registration and, if it persists, your driver license!
We suspend 300,000 registrations a year for failure to maintain insurance. …

Input Generation - Document-Driven

Whole Document

- Input a document or provide background information, and leave it to the "utterance generation" component to decide which part of the document to use



Title : Esports **Section Title** : History Early history (1972–1989)

Document c

The earliest known video game competition took place on 19 October 1972 at Stanford Universit for the game "Spacewar". Stanford students were invited to an "Intergalactic spacewar olympics" whose grand prize was a years subscription for "Rolling Stone", with Bruce Baumgart winning the five-man-free-for-all tournament and Tovar and Robert E. Maas winning the Team Competition. The Space Invaders Championship held by Atari in 1980 was the earliest large scale video game competition, attracting more than 10,000 participants across the United States, establishing competitive gaming as a mainstream hobby. · · ·

Background $\mathcal{B}$

Esports (also known as electronic sports, e-sports, or eSports) is a form of competition using video games. Most commonly, esports takes the form of organized, multiplayer video game competitions, particularly between professional players, individually or as teams. Although organized online and offline competitions have long been a part of video game culture, these were largely between amateurs until the late 2000s, when participation by professional gamers and spectatorship in these events through live streaming saw a large surge in popularity. By the 2010s, esports was a significant factor in the video game industry, with many game developers actively designing toward a professional esports subculture.

Conversation in WIKI-SIMSEEK

q_1 : **What is the history of esports?**

a_1 : The earliest known video game competition took place on 19 October 1972 at Stanford University for the game "Spacewar".

q_2 : **What was the result of this competition?**

a_2 : Bruce Baumgart winning the five-man-free-for-all tournament and Tovar and Robert E. Maas winning the Team Competition.

q_3 : **Did esports grow from there?**

a_3 : The Space Invaders Championship held by Atari in 1980 was he earliest large scale video game competition, attracting more than 10,000 participants across the United States,

q_4 : **What happened after the Space Invaders Championship?**

· · ·

(Kim et al., 2022)

Input Generation - Sequence-Grounded

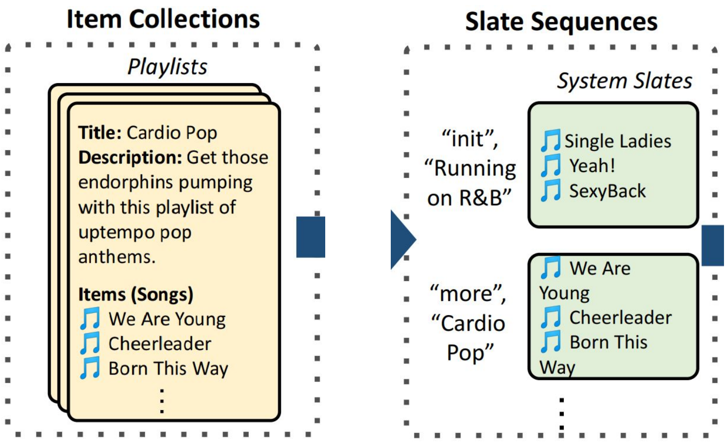
Fixed Sequence

1) Topics with their Background knowledges

- Select / Generate a topic

2) A sequence of dialogue acts

- Sampling a valid sequence



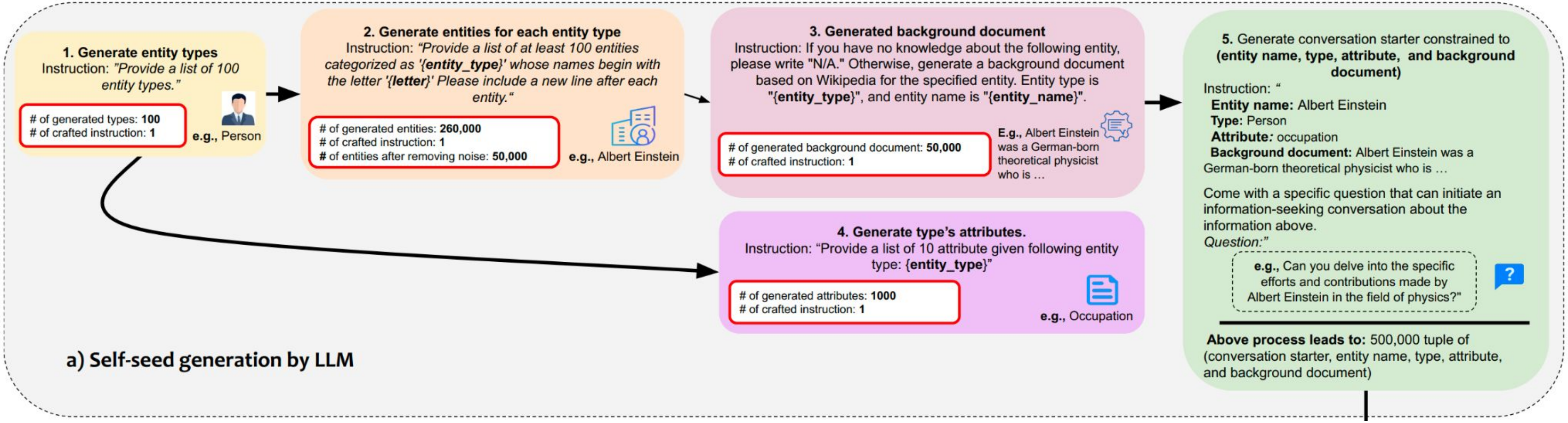
(Leszczynski et al., 2023)

Input Generation - Sequence-Grounded

Generate the background information

(Askari et al., 2024)

- Why Generation instead of Selection? Quality / Diversity
- Self-seeding approach, Prompt LLM to generate everything



Input Generation - Sequence-Grounded

Dialogue Acts - Fixed

- Main feature: validity
- Make conversation real, maintain the consistency
- How to ensure the validity? Using existing crowdsourcing dialogue datasets

(Askari et al., 2024)

SOLID: Full path
Used MSDialog-intent

Table 7: Intent taxonomy and distribution in MSDialog

Code	Label	Description	%
OQ	Original Question	The first question from the user to initiate the dialog.	13
RQ	Repeat Question	Other users repeat a previous question.	3
CQ	Clarifying Question	User or agent asks for clarifications.	4
FD	Further Details	User or agent provides more details.	14
FQ	Follow Up Question	User asks for follow up questions about relevant issues.	5
IR	Information Request	Agent asks for information from users.	6
PA	Potential Answer	A potential answer or solution provided by agents.	22
PF	Positive Feedback	User provides positive feedback for working solutions.	6
NF	Negative Feedback	User provides negative feedback for useless solutions.	4
GG	Greetings/Gratitude	Greetings or expressing gratitude.	22
JK	Junk	No useful information in the utterance.	1
O	Others	Utterances cannot be categorized using other classes.	1

(Wang et al., 2023)

TopDial: partial path
starting point and the target (act-topic)
Used DuRecDial 2.0

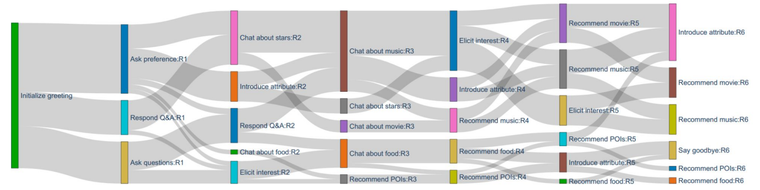


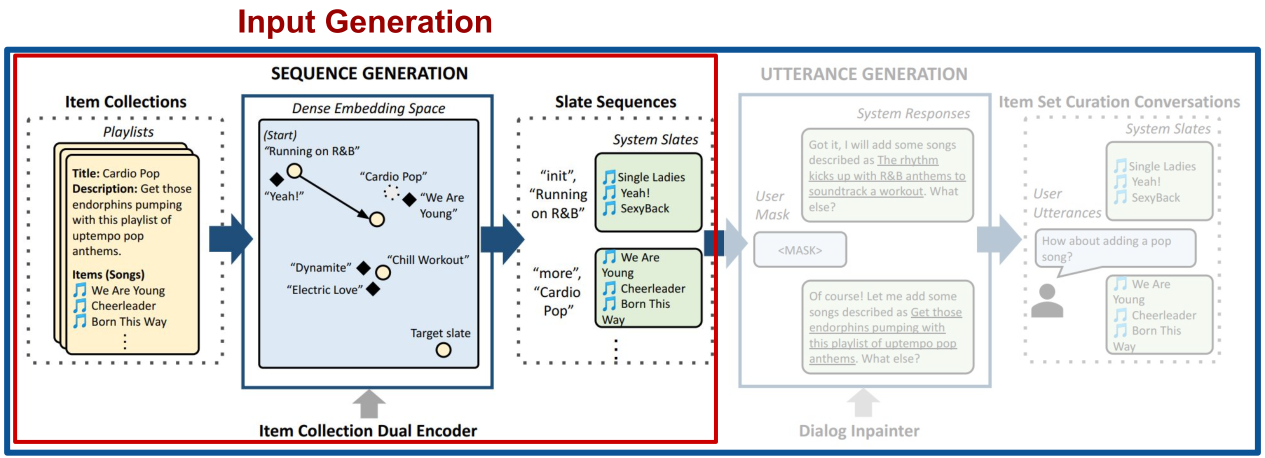
Figure 3: Transitions of dialogue acts of the system through the first six rounds.

Input Generation - Sequence-Grounded

Dialogue Acts - Fixed

(Leszczynski et al., 2022)
(Leszczynski et al., 2023)

- How to ensure the validity? Closeness in embedding space
- Example: Walk the Talk

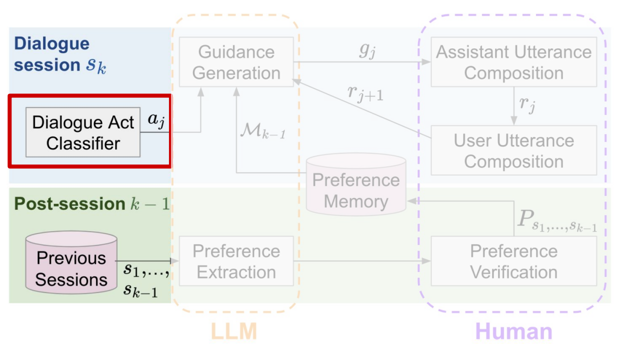


Input Generation - Sequence-Grounded

Dialogue Acts - Open

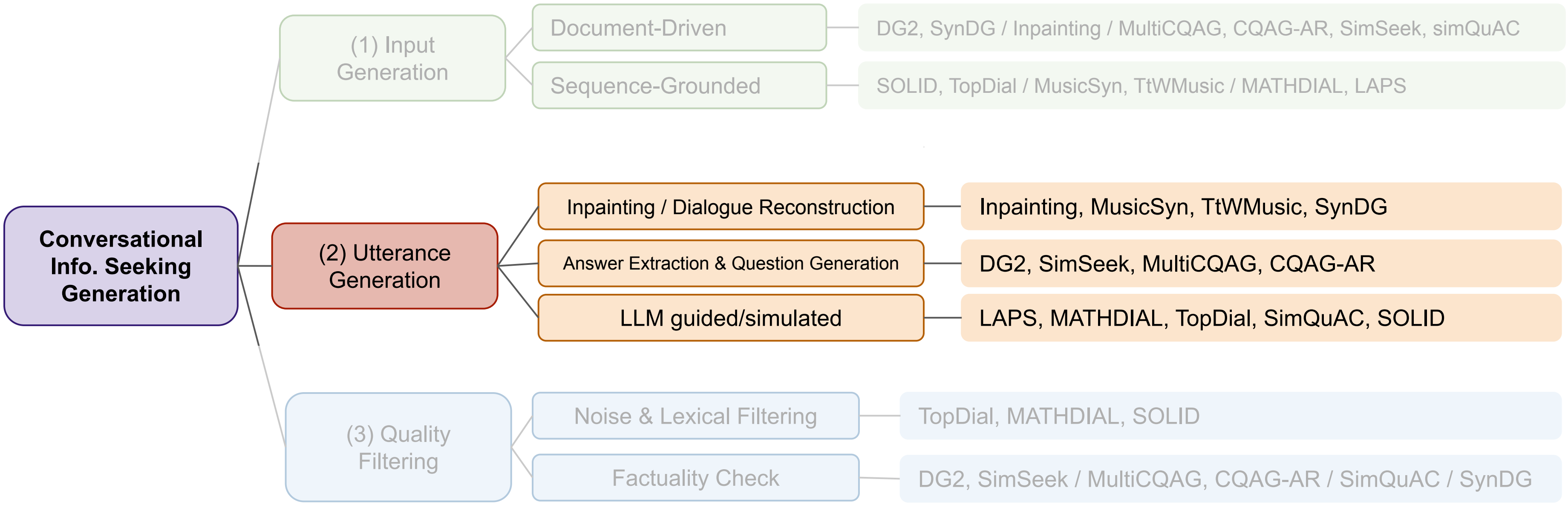
- Used in Human-AI collaboration based methods
- Dialogue act is predicted
  - Based on Dialogue history
  - Before the current tuen is generated

LAPS: LLM classifier (Joko et al., 2024)



MathDial: Human selects (Macina et al., 2023)

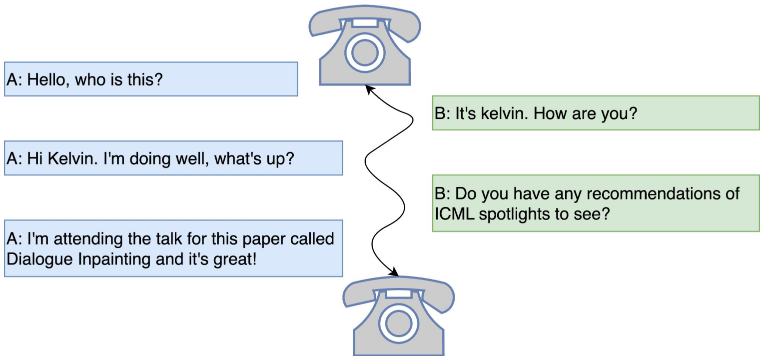
Category	Intent	Example
Focus	Seek Strategy	So what should you do next?
	Guiding Student Focus	Can you calculate . . . ?
	Recall Relevant Information	Can you reread the question and tell me what is . . . ?
Probing	Asking for Explanation	Why do you think you need to add these numbers?
	Seeking Self Correction	Are you sure you need to add here?
	Perturbing the Question	How would things change if they had . . . items instead?
	Seeking World Knowledge	How do you calculate the perimeter of a square?
Telling	Revealing Strategy	You need to add . . . to . . . to get your answer.
	Revealing Answer	No, he had . . . items.
Generic	Greeting/Fairwell	Hi . . . , how are you doing with the word problem? Good Job! Is there anything else I can help with?
	General inquiry	Can you go walk me through your solution?


Conversational Info. Seeking Generation
(1) Input Generation
Document-Driven
DG2, SynDG / Inpainting / MultiCQAG, CQAG-AR, SimSeek, simQuAC
Sequence-Grounded
SOLID, TopDial / MusicSyn, TtWMusic / MATHDIAL, LAPS
(2) Utterance Generation
Inpainting / Dialogue Reconstruction
Inpainting, MusicSyn, TtWMusic, SynDG
Answer Extraction & Question Generation
DG2, SimSeek, MultiCQAG, CQAG-AR
LLM guided/simulated
LAPS, MATHDIAL, TopDial, SimQuAC, SOLID
(3) Quality Filtering
Noise & Lexical Filtering
TopDial, MATHDIAL, SOLID
Factuality Check
DG2, SimSeek / MultiCQAG, CQAG-AR / SimQuAC / SynDG

Utterance Generation - Inpainting

(Dai et al., 2022)

- **Reminder:** Dialogue Flow -> Directly the document's sentences
- **Idea:** Fine-tuning a model to reconstruct a dialogue
- **Real world Motivation:**
    - Overhearing someone else's phone call
        - Hear on side, try to guess another side



- **Task:** Take a partial dialog $\Longrightarrow$ Generate a complete dialog

$$(u_1, u_2, \diamond, u_4, \diamond) \Longrightarrow d = (u_1, u_2, \ldots, u_t, \ldots, u_T)$$

Utterance Generation - Inpainting

(Dai et al., 2022)

Training: Dialog reconstruction

- Randomly mask one utterance (u_t)
- Train a generative model to predict the masked utterance
- Similar to the masked language modeling task used by BERT

$$d_{m(t)} = (u_1, \ldots, u_{t-1}, \diamond, u_{t+1}, \ldots, u_T)$$

$$p_\theta(u_t \mid d_{m(t)})$$

Inference: Transforming documents into dialogues

- Convert document into spans (e.g., sentences)
- Autoregressively generate utterances

$$(s_{\text{prompt}}, \diamond, s_1) \Longrightarrow \hat{u}_1$$

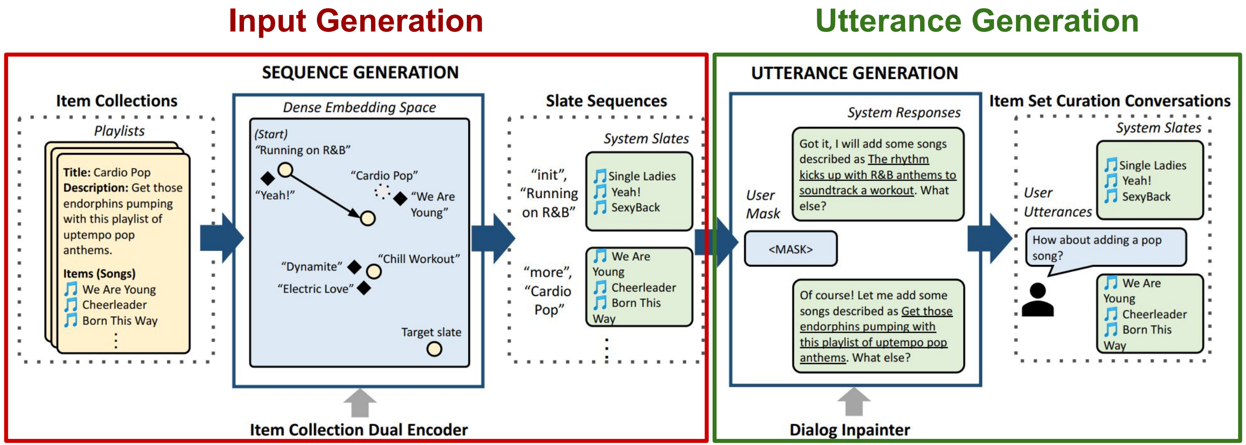
$$(s_{\text{prompt}}, \hat{u}_1, s_1, \diamond, s_2) \Longrightarrow \hat{u}_2$$

Utterance Generation - Inpainting

Another example of Inpainting

(Leszczynski et al., 2022)
(Leszczynski et al., 2023)

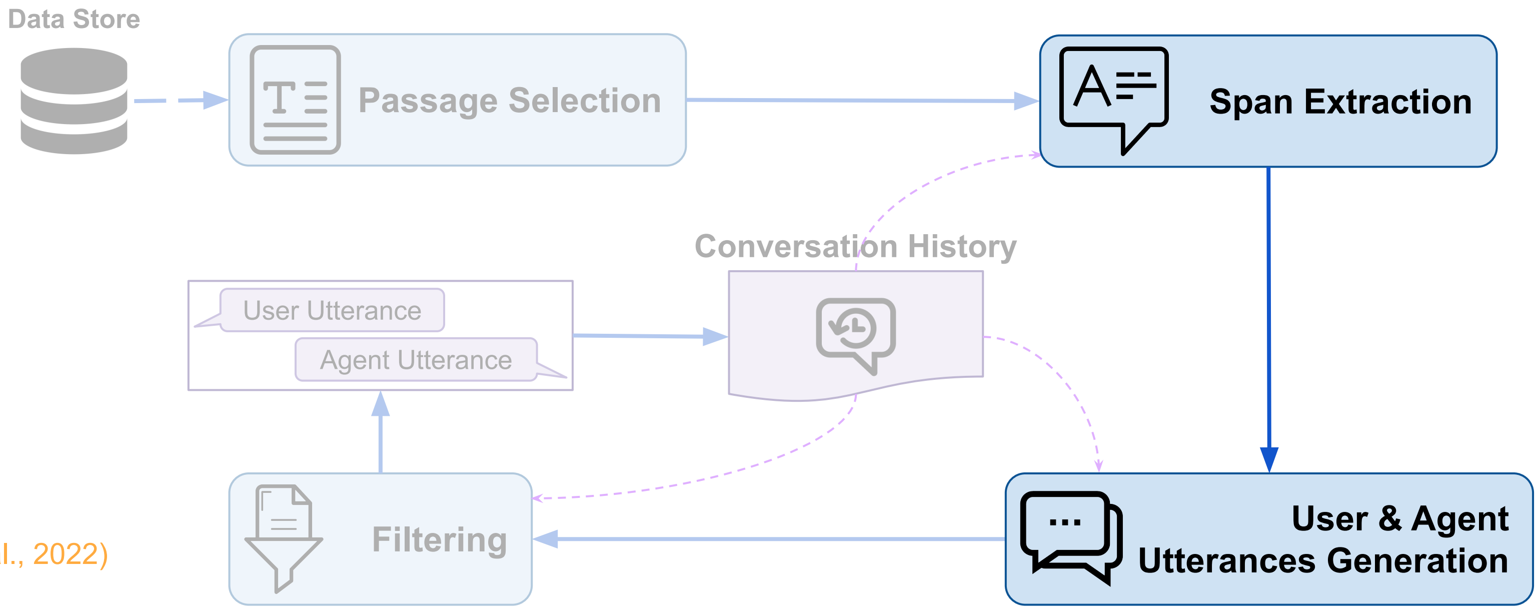
- **Reminder:** Dialogue Flow -> Slate (playlist) sequences



Utterance Generation - Answer Extraction & Question Generation

Reminder: Dialogue Flow -> not fixed, passages Passage Ranker

- The extended version of pipeline approach for "single-turn QA pair generation" (Alberti et al., 2019)



(Wu et al., 2022)

Utterance Generation - Answer Extraction & Question Generation

Answer/Span Extraction (Wu et al., 2022)

<Passage 3> Not Bringing Proper Documentation to DMV Office. About ten percent of customers visiting a DMV office do not bring what they need to complete their transaction and see if your transaction can be …

Conversation History

Highlights the rationale span used to generate the dialogue turn

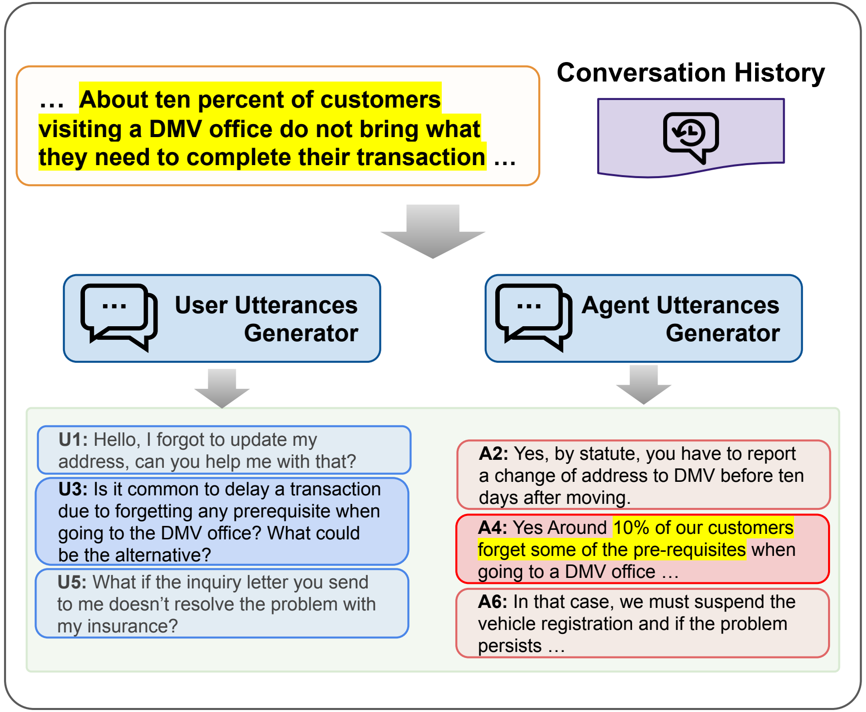
<Passage 3> Not Bringing Proper Documentation to DMV Office. About ten percent of customers visiting a DMV office do not bring what they need to complete their transaction and see if your transaction can be …

Extract a rationale span from the selected passage

$$p(r_t \mid \underbrace{\{u_i, a_i\}_{i<t}}_{\text{Conversation History}}, \underbrace{c_t}_{\text{Selected passage in turn } t})$$

Utterance Generation - Answer Extraction & Question Generation

User & Agent Utterance Generation (Wu et al., 2022)



User utterance generator

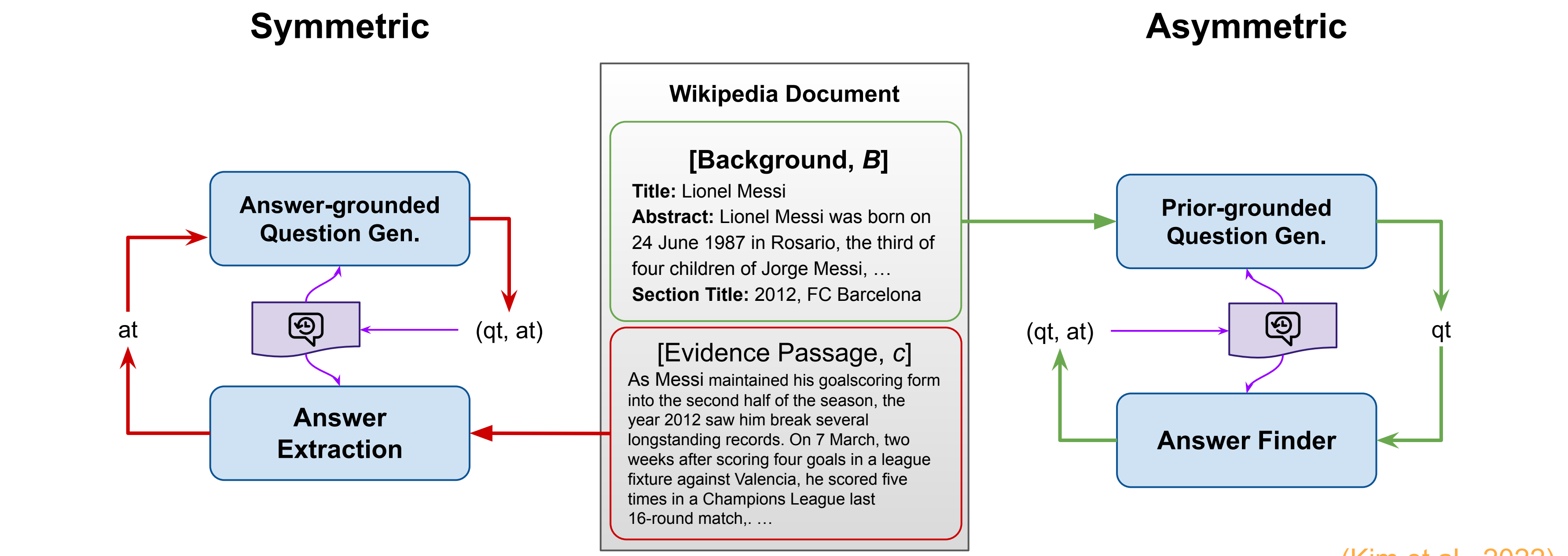
- Generates a question with the answer span
- Highlight the rationale span by wrapping its text

$$p(u_t) = p(u_t | \underbrace{\{u_i, a_i\}_{i<t}}_{\text{Conversation History}}, \underbrace{c'_t}_{\text{Selected passage in turn } t})$$

Agent utterance generator

- Generates the response with the answer span
- The dialogue history now includes the previous generated user utterance

Utterance Generation - Answer Extraction & Question Generation



(Kim et al., 2022)

Utterance Generation - Answer Extraction & Question Generation

Symmetric

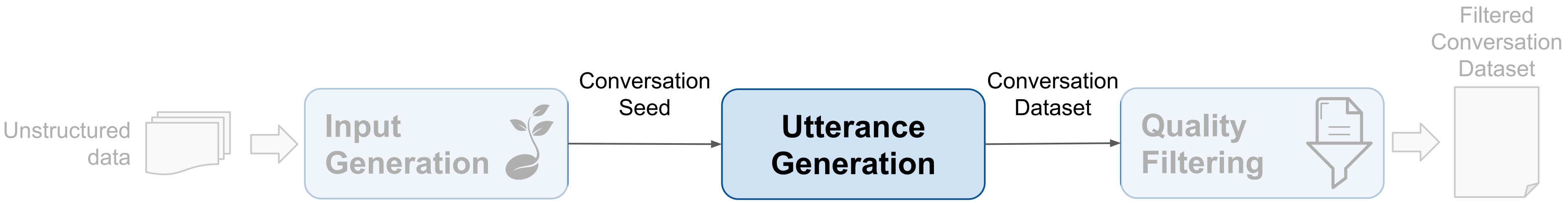
- First extracts an answer candidate from the passage
- Questioner can access all answer-relevant information
  - Pro: Coherency with answer
  - Con: Constraint to the predetermined answer

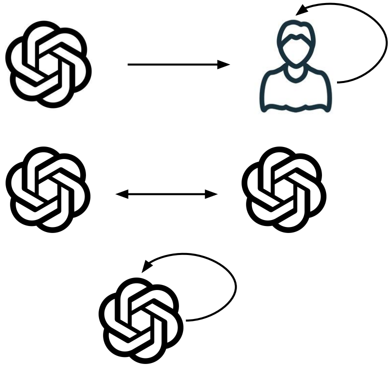
Asymmetric

- First asks a question without accessing an answer or passage
- Questioner asks any questions relevant to the topic, guessing inaccessible passage
  - Pro: encouraging information-seeking behaviour

(Kim et al., 2022)

Utterance Generation - LLM guided/simulated

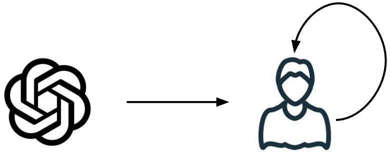


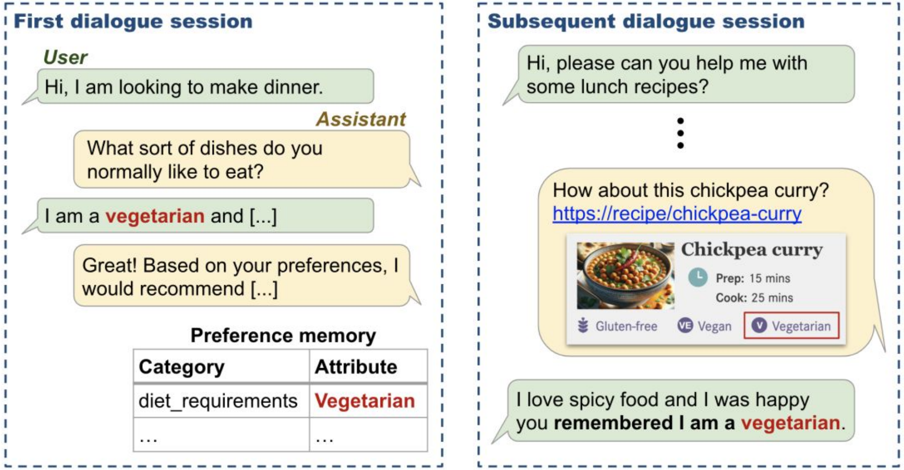


Utterance Generation - LLM guided/simulated

LLM-Guided Human Generation (LAPS) (Joko et al., 2024)

- Task: personalized multi-session dialogue





Utterance Generation - LLM guided/simulated

LLM-Guided Human Generation (LAPS**)** (Joko et al., 2024)

Main components:

1) Dialogue act classification

2) Guidance ge

3) Utterance co

4) Preference e

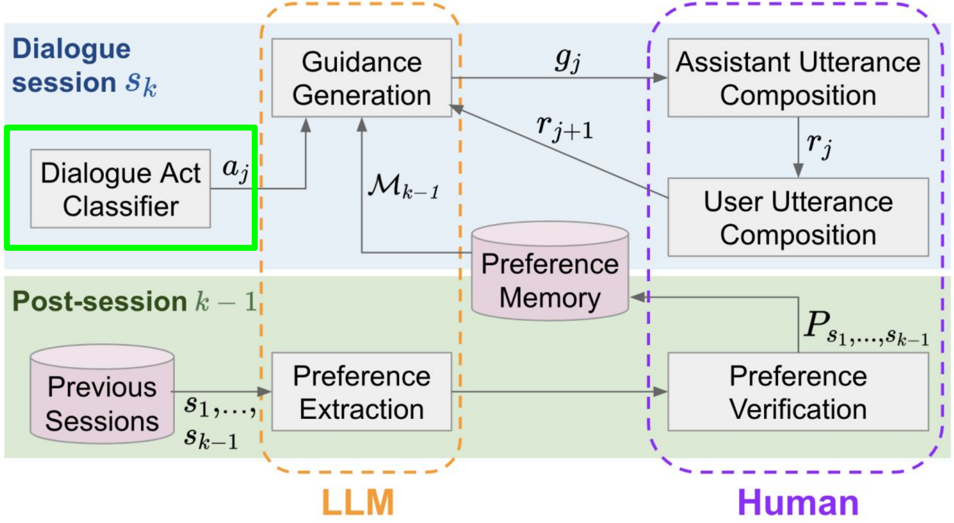
(1) Greeting

(2) Preference elicitation

(3) Recommendation

(4) Follow-up questions

(5) Goodbye



Utterance Generation - LLM guided/simulated

LLM-Guided Human Generation (LAPS) (Joko et al., 2024)

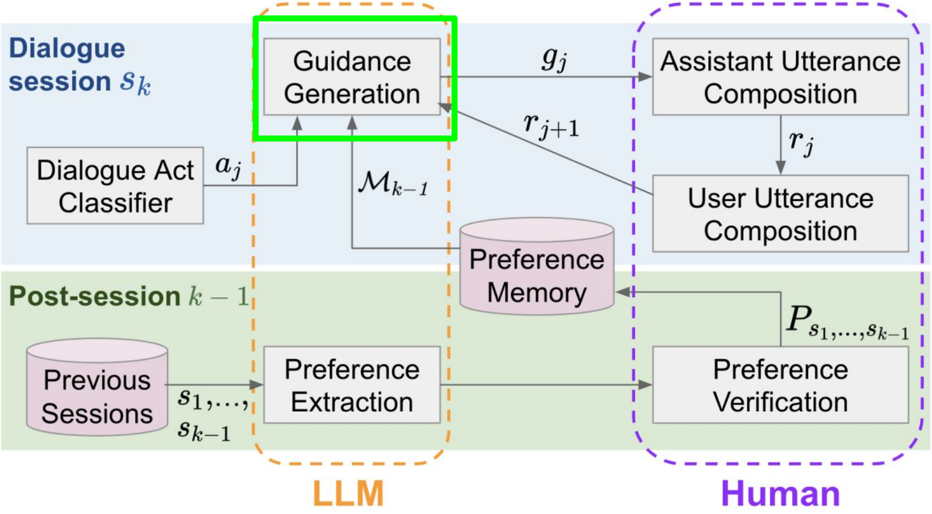
Main components:

1) Dialogue act classification

2) Guidance generation

3) Utterance composition

4) Preference extraction



Utterance Generation - LLM guided/simulated

LLM-Guided Human Generation (LAPS) (Joko et al., 2024)

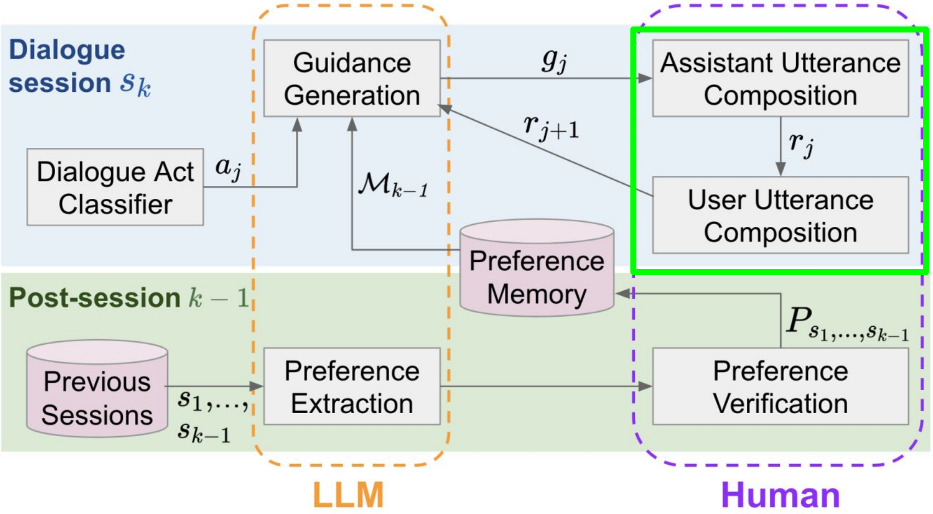
Main components:

1) Dialogue act classification

2) Guidance generation

3) Utterance composition

4) Preference extraction

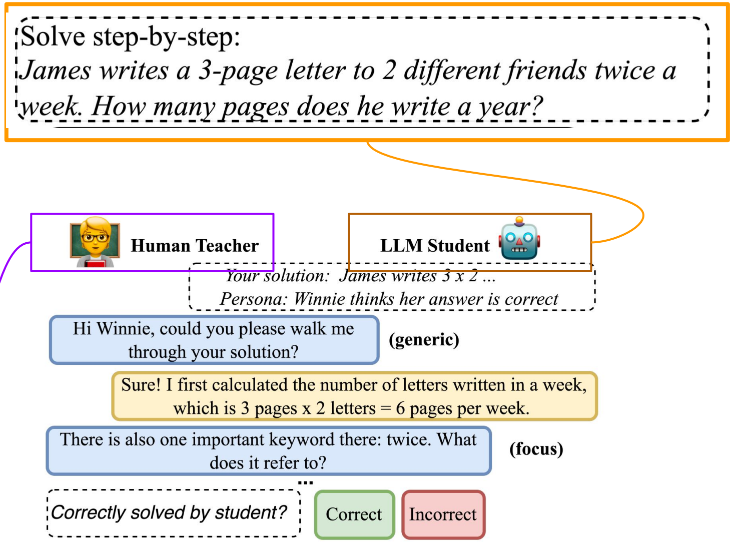


Utterance Generation - LLM guided/simulated

LLM-Human Collaboration (MathDial) (Macina et al., 2023)

- Task: Dialogue tutors
- Main components:
  - LLM as a student
  - Human as a teacher

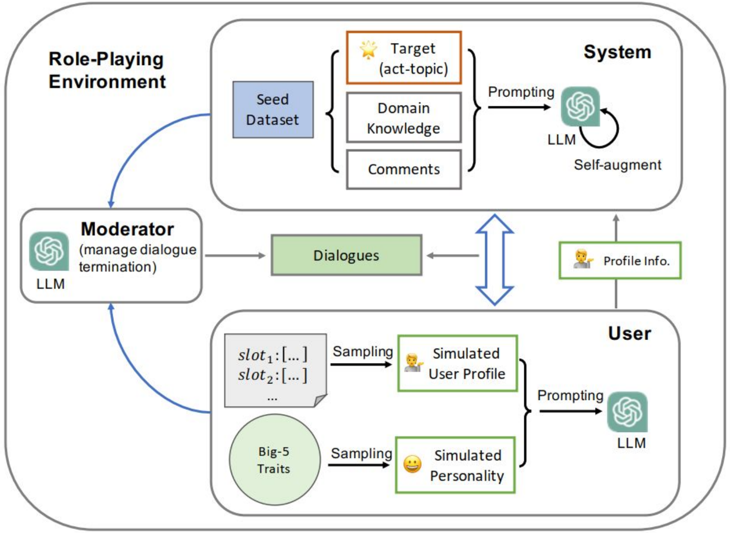
Category	Intent	Example
Focus	Seek Strategy	So what should you do next?
	Guiding Student Focus	Can you calculate . . . ?
	Recall Relevant Information	Can you reread the question and tell me what is . . . ?
Probing	Asking for Explanation	Why do you think you need to add these numbers?
	Seeking Self Correction	Are you sure you need to add here?
	Perturbing the Question	How would things change if they had . . . items instead?
	Seeking World Knowledge	How do you calculate the perimeter of a square?
Telling	Revealing Strategy	You need to add . . . to . . . to get your answer.
	Revealing Answer	No, he had . . . items.
Generic	Greeting/Fairwell	Hi . . . , how are you doing with the word problem? Good Job! Is there anything else I can help with?
	General inquiry	Can you go walk me through your solution?



Utterance Generation - LLM guided/simulated

Fully LLM Generation & Supervision (TopDial) (Wang et al., 2023)

- **Task:** Target-oriented Recommendation Syste
- 3 LLMs collaboration
  - LLM as a User
  - LLM as a System
  - LLM as a Moderator



Utterance Generation - LLM guided/simulated

Fully LLM Generation & Supervision (SimQuAC) (Abbasiantaeb et al., 2024)

- QuAC: information-seeking QA dialogs dataset (Choi et al., 2018)
    - Continuation dialogue act:
        - it allows teachers to guide the student's questioning towards aspects of the article that are especially important or interesting

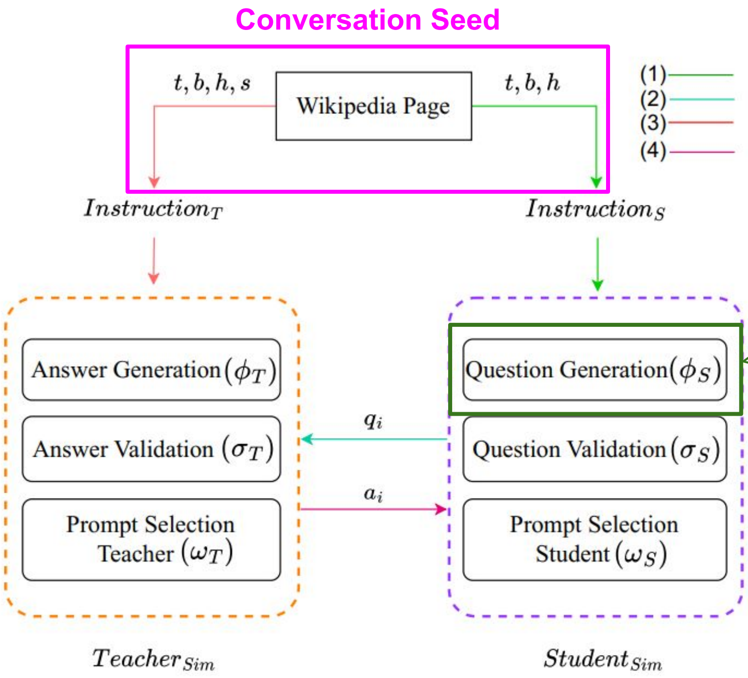
- replacing both human parties with LLMs
- Implement both the student and teacher by zero-shot prompting GPT-4

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↪̸ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

Utterance Generation - LLM guided/simulated

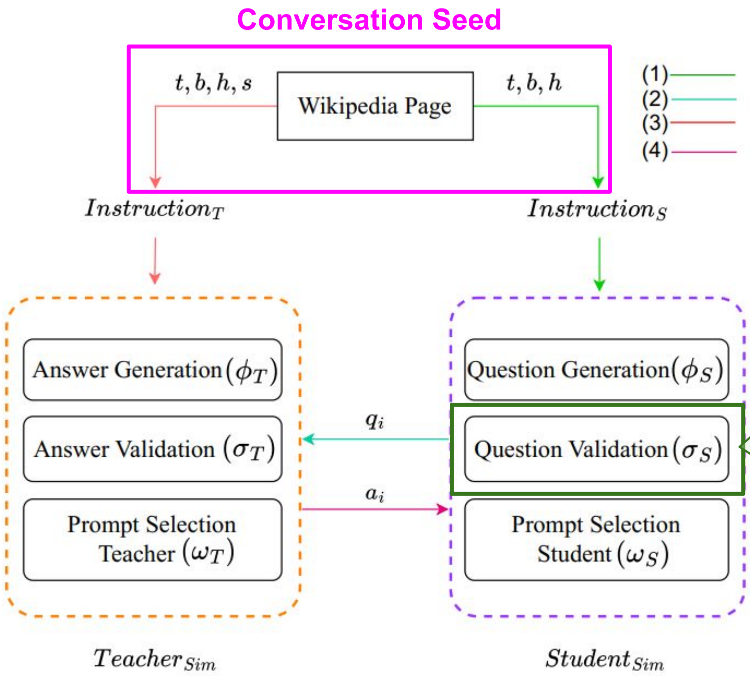
Fully LLM Generation & Supervision (SimQuAC) (Abbasiantaeb et al., 2024)



Student: Given information (h and b), pose questions

Utterance Generation - LLM guided/simulated

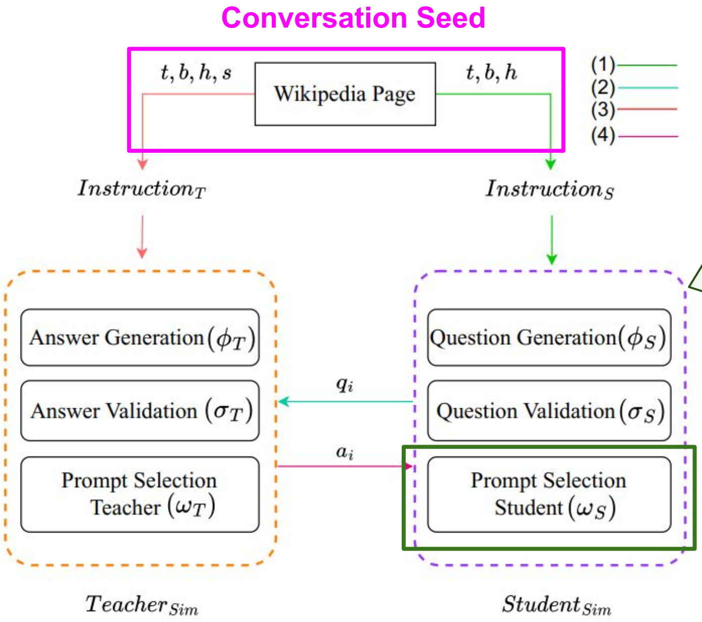
Fully LLM Generation & Supervision (SimQuAC) (Abbasiantaeb et al., 2024)



Problem: the LLM tends to generate multiple questions in one go
Solution: filter lengthy questions

Utterance Generation - LLM guided/simulated

Fully LLM Generation & Supervision (SimQuAC) (Abbasiantaeb et al., 2024)



Problem: unanswered from the given text, higher chance that the subsequent question might be overly specific and cannot be answered

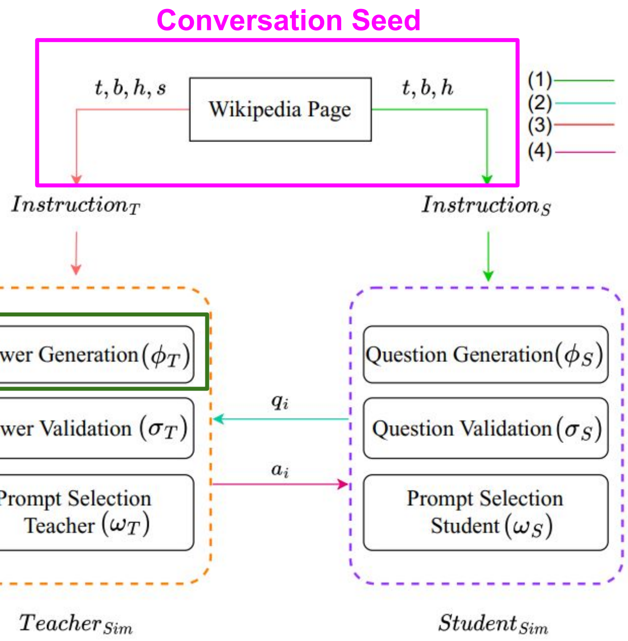
Solution: Randomly selects one of the following guiding prompts

(i) Ask a general question
(ii) Ask a question starting with where, when, or who
(iii) Ask a question about what is interesting in this article
(iv) Ask a question about another aspect of the topic

Utterance Generation - LLM guided/simulated

Fully LLM Generation & Supervision (SimQuAC**)** (Abbasiantaeb et al., 2024)

copy the exact spans from the passage to answer the given question



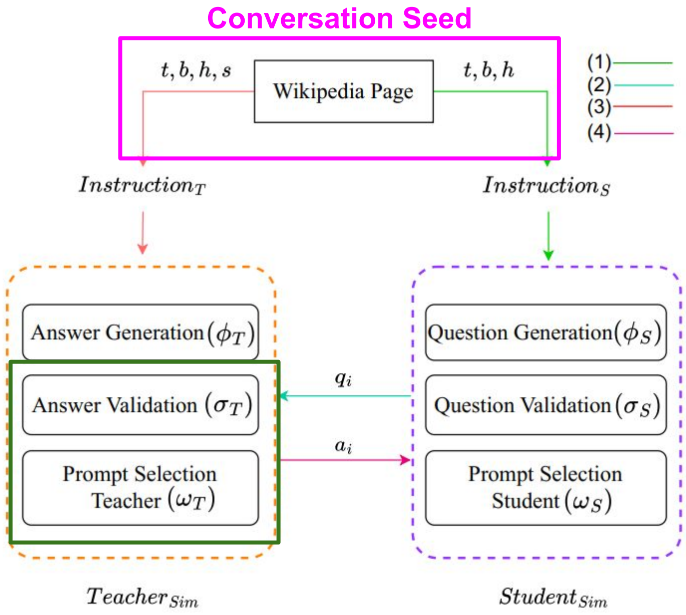
Utterance Generation - LLM guided/simulated

Fully LLM Generation & Supervision (SimQuAC) (Abbasiantaeb et al., 2024)

An iterative model to validate and refine the generated answers

- It checks whether the answer is copied from the text section or being "I cannot find the answer"

Solution: text search and multiple sequential prompts to generate other answers



Utterance Generation - LLM guided/simulated

One LLM plays all roles (SOLID**)** (Askari et al., 2024)

- **Reminder:** conversation seed: Generated background info + Sequence of intents
- How to apply intent in prompting?
  - Define Instruction

Table 10: The last part of the intent-based LLM-instruction. Actor type: Agent

Intent	Instruction
CQ	Reply with one follow-up response in conversation style.
FD	Reply with further details in conversation style.
GG	Continue the conversation by expressing gratitude for the user's questions.
PA	Provide a potential solution or answer in conversation style.
IR	Ask the user to provide relevant information needed for their previous question.
OQ	Formulate an original question posed by an agent.
FQ	Formulate a follow-up question from an agent, seeking further clarification or information.
RQ	Now you are talking from the point of view of a third participant in the conversation. Repeat Question:
PF	Express satisfaction and appreciation for the conversation.
NF	Convey dissatisfaction for the previous response.
JK	Reply with gibberish information. It can contain emojis.
O	Reply with a system error. Return N/A

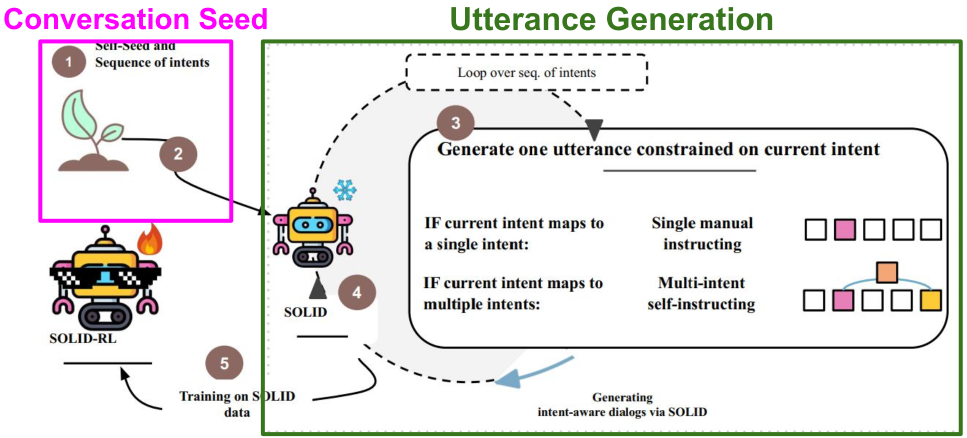
Table 11: The last part of the intent-based LLM-instruction. Actor type: User

Intent	Instruction
CQ	Reply with one question asking for clarification in conversation style.
FD	Reply with more details in conversation style.
GG	Continue the conversation by expressing gratitude for the agent's help.
PA	Provide a potential solution or answer in conversation style.
IR	Reply with relevant information.
OQ	Formulate the first question posed by a user that initiates a QA dialog.
FQ	Formulate a follow-up question from a user, seeking further clarification or information.
RQ	Now you are talking from the point of view of a third participant in the conversation. Repeat Question:
PF	Express satisfaction and appreciation for a working solution.
NF	Convey dissatisfaction for the previous response.
JK	Reply with gibberish information. It can contain emojis.
O	Reply with a system error. Return N/A

Utterance Generation - LLM guided/simulated

One LLM plays all roles (SOLID**)** (Askari et al., 2024)

- Generates utterances guided by a specific intent or intents
- Each utterance generation fits under one of two cases
  - Single intent
  - Multiple intent
    - Prompt LLM to generate one merged instruction

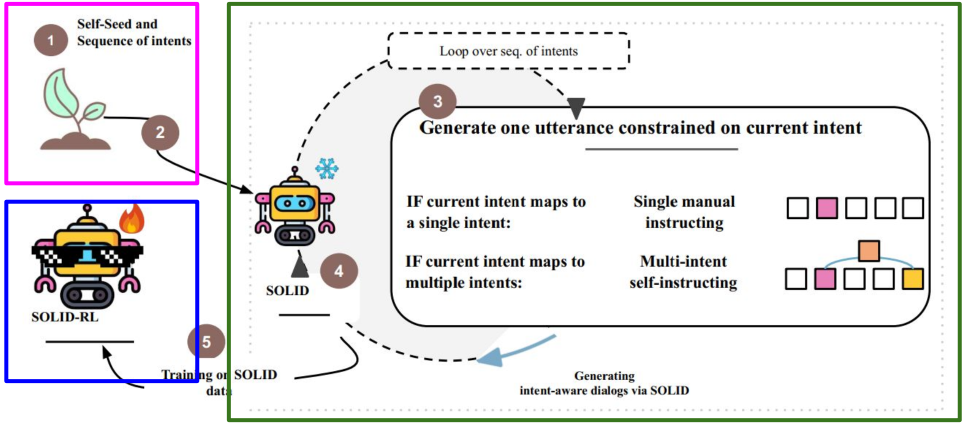


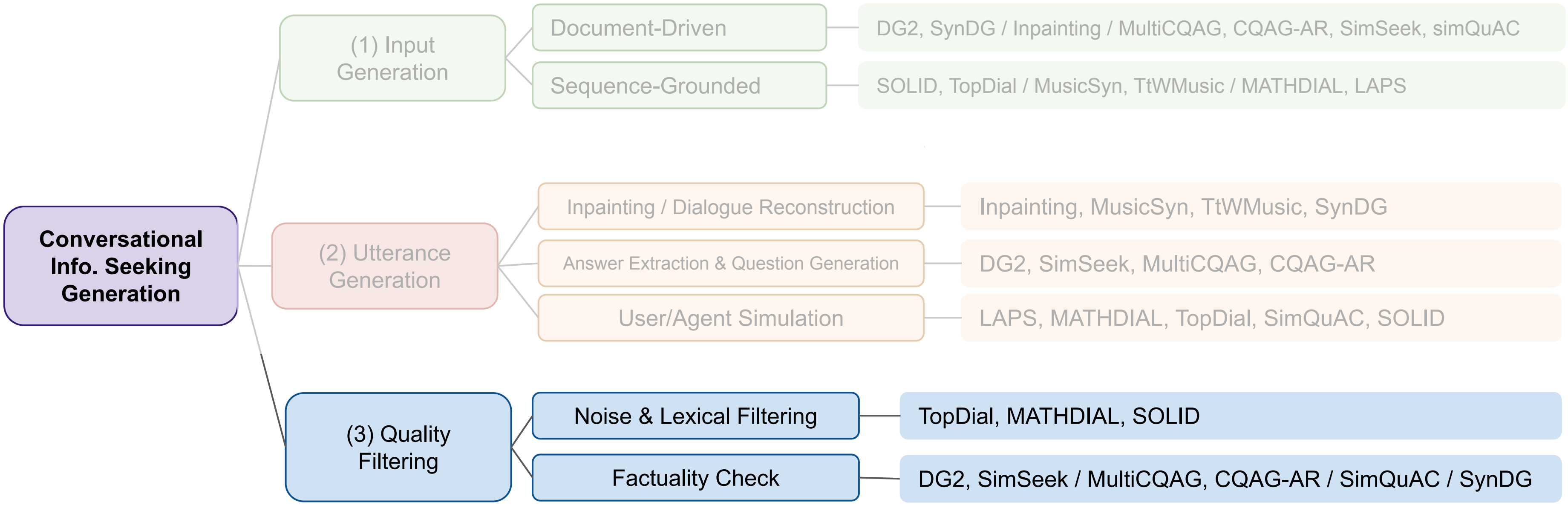
Utterance Generation - LLM guided/simulated

One LLM plays all roles - One Go Generation (SOLID-RL) (Askari et al., 2024)

- One Go generation advantages
  - Enhancing the naturalness
  - consistency of the conversation,
  - Increasing generation speed
- Approach
  - Fine-tuned on synthetic data

Fine-tuning for One-Go Generation

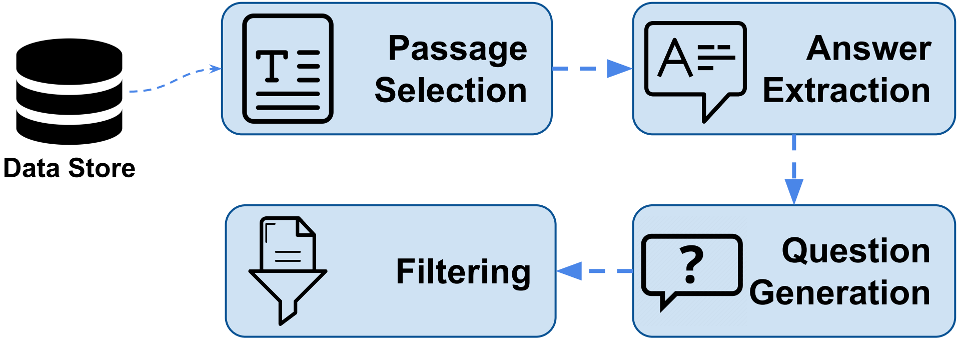



Conversational Info. Seeking Generation
(1) Input Generation
Document-Driven
DG2, SynDG / Inpainting / MultiCQAG, CQAG-AR, SimSeek, simQuAC
Sequence-Grounded
SOLID, TopDial / MusicSyn, TtWMusic / MATHDIAL, LAPS
(2) Utterance Generation
Inpainting / Dialogue Reconstruction
Inpainting, MusicSyn, TtWMusic, SynDG
Answer Extraction & Question Generation
DG2, SimSeek, MultiCQAG, CQAG-AR
User/Agent Simulation
LAPS, MATHDIAL, TopDial, SimQuAC, SOLID
(3) Quality Filtering
Noise & Lexical Filtering
TopDial, MATHDIAL, SOLID
Factuality Check
DG2, SimSeek / MultiCQAG, CQAG-AR / SimQuAC / SynDG

Quality Filtering - Factuality Check

Roundtrip Consistency

- For QA pair Generation



1) Passage Selection	**C**	… in 1903, boston participated in the first modern world series, going up against the pittsburgh pirates …
2) Answer Extraction	**C→A**	1903
3) Question Generation	**C, A→Q**	when did the red sox first go to the world series
4) Filtering	**C, Q→A′** **$A \stackrel{?}{=} A'$**	1903 Yes

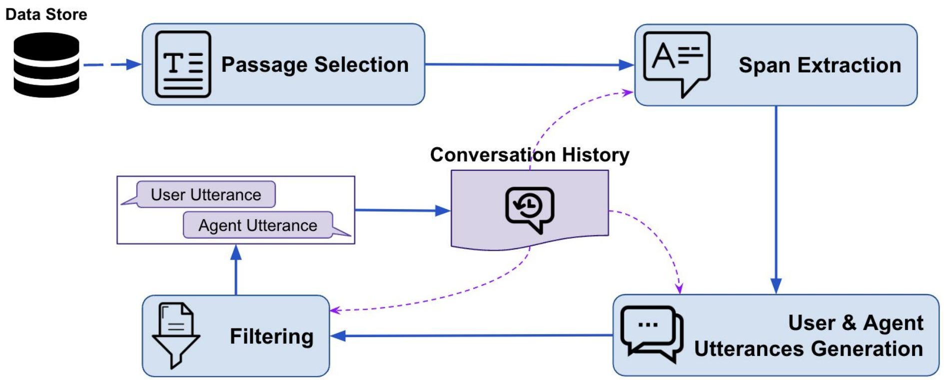
(Alberti et al., 2019)

Quality Filtering - Factuality Check

(Wu et al., 2022)

Roundtrip Consistency

- For Conversational Turn Generation



$$p(\underline{\hat{c}_t} | \underbrace{\{u_i, a_i\}_{i<t}}_{\text{Conversation History}}, , \underbrace{u_t}_{\text{User Utterance}}, \underbrace{C}_{\text{Document}})$$

$$p(\hat{r}_t | \underline{\{u_i, a_i\}_{i<t}}, \underline{u_t}, \underline{\hat{c}_t})$$

Conclusion and Future Directions

Duration: 10 min

Presenter: Evangelos Kanoulas

Your Conclusions

- Are zero-shot LLMs + prompting the ultimate dialogue system?
- Is there need for data generation?
- What is left to be done?

What we have so far - Task-oriented Dialogue

- Task-oriented dialogue systems require task-/domain-specific data
    - Strong dependence on individual task characteristics, constraints, etc.
- Task-specific data require modeling the task/domain through schemas, ontologies, etc.
    - In data augmentation there is a chance to make this data driven, but not in zero-shot
- LLMs are proven good UX towards consuming and producing text
    - Including generating dialogue goals
- … but passing task/domain constraints remains a challenge; even when leveraging LLMs, we need access to constraints such as schemas, or ontologies. They are mostly human-generated and not easily integrated in an e2e process

What we have so far - Open Domain Dialogue

- Data augmentation is proven effective for various types of open domain dialogue systems
- Methods have moved from Generative to Prompting based
  - Minimizes the need for human involvement
  - It is faster and more accessible
- General trend in LLM-based data augmentation:
  - Create Large-scale LLM-generated datasets; e.g., using GPT* models
  - (Parameter-efficient) Finetune another LLM (e.g., LLaMA) to generate a dialogue agent
  - E.g., for role-specified open domain dialogue systems, information seeking systems
- It still requires domain-specific knowledge (i.e., seed data, structural constraints)

What we have so far - Conversational information Seeking

- Single document grounding w/ simple flow management and answer extraction
- LLMs attempt to go beyond a single source of info and simulate/guide users behaviour
- Remaining challenges
  - Multi-source grounding
  - Conversation flow guidance
  - Mixed-initiative
  - Modeling of the CIS dialogues

Open Challenges

- There is less control over the generated data
  - Limited guards against unsafe and toxic content
  - Large-scale automatic evaluation and human evaluation is still an open problem
- LLM-generated dialogues lead to self-reinforcement of LLM-based dialogue systems
  - We already know LLM-based evaluation models prefer LLM-generated text
- Large scale data generation for complex and personalized tasks remains a challenge
  - E.g., tutoring tasks, modeling personas and preferences,

References

References: Introduction

- (Zamani et al., 2023) Conversational information seeking. Foundations and Trends® in Information Retrieval, 244-456.
- (Ni et al. 2023) Recent advances in deep learning based dialogue systems: A systematic survey. Artificial intelligence review, 3055-3155.
- (Deng et al., 2023) A survey on proactive dialogue systems: Problems, methods, and prospects. arXiv preprint arXiv:2305.02750.
- (Chen et al., 2023) An empirical survey of data augmentation for limited data learning in nlp. Transactions of the Association for Computational Linguistics, 191-211.
- (Liu et al., 2021) Towards Emotional Support Dialog Systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, 3469–3483.
- (Zhang et al., 2018) Personalizing Dialogue Agents: I have a dog, do you have pets too?, In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2204–2213.
- (Wu et al., 2022) INSCIT: Information-Seeking Conversations with Mixed-Initiative Interactions, arXiv:2207.00746.
- (Feng, et al., 2020) Doc2Dial: A Framework for Dialogue Composition Grounded in Documents, In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI, 13604–13605.
- (Choi et al., 2018) QuAC: Question Answering in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2174–2184.

References: Evaluation

- (Zhang et al., 2020) BERTScore: Evaluating Text Generation with BERT. ICLR
- (Yuan et al., 2021) BARTScore: Evaluating Generated Text as Text Generation, NeurIPS, 27263-27277.
- (Kim et al., 2021) NeuralWOZ: Learning to Collect Task-Oriented Dialogue via Model-Based Simulation, ACL, 3704–3717.
- (Wu et al., 2022) DG2: Data Augmentation Through Document Grounded Dialogue Generation, SIGDIAL, 204–216.
- (Gao et al., 2019) Interconnected Question Generation with Coreference Alignment and Conversation Flow Modeling, ACL, 4853–4862.
- (Kim et al., 2022) Generating Information-Seeking Conversations from Unlabeled Documents, Conference on Empirical Methods in Natural Language Processing, EMNLP, 362–2378.
- (Li et al., 2016) A Diversity-Promoting Objective Function for Neural Conversation Models, NAACL HLT, 110–119.
- (Zhang et al., 2018) Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization, NeurIPS, 1815–1825.
- (Reimers et al., 2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP-IJCNLP, 3980–3990.
- (Zhu et al., 2018) Texygen: A Benchmarking Platform for Text Generation Models, SIGIR, 1097–1100.
- (Mehri et al., 2020) USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation, ACL, 681–707.
- (Zhong et al., 2022) Towards a Unified Multi-Dimensional Evaluator for Text Generation, EMNLP, 2023–2038.
- (Liu et al.,2023) G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, EMNLP, 2511–2522.
- (Michael Smith et al.,2022) Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents, the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, 77–97.

References: Task-oriented

- (Li et al., 2016) A User Simulator for Task-Completion Dialogues, CoRR, volume 1612.05688, arXiv:1612.05688
- (Shah et al., 2018) Building a Conversational Agent Overnight with Dialogue Self-Play, CoRR, volume 1801.04871, arXiv:1801.04871
- (Rastogi et al., 2020) Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset, AAAI-20, pages 8689-8696, arXiv:1909.05855
- (Kreyssig et al., 2018) Neural User Simulation for Corpus-based Policy Optimisation for Spoken Dialogue Systems, SIGIDIA pages 60-69, arXiv:1805.06966
- (Kim et al., 2021) NeuralWOZ: Learning to Collect Task-Oriented Dialogue via Model-Based Simulation, ACL 2021, pages 3704-3717, arXiv:2105.14454
- (Gur et al., 2018) User Modeling for Task Oriented Dialogues, IEEE, pages 900-906, arXiv:1811.04369
- (Tseng et al., 2021) Transferable Dialogue Systems and User Simulators, ACL 2021, pages 152-166, arXiv:2107.11904
- (Wan et al., 2022) A Unified Dialogue User Simulator for Few-shot Data Augmentation, EMNLP pages 3788-3799
- (Yoo et al., 2020) Variational Hierarchical Dialog Autoencoder for Dialog State Tracking Data Augmentation, EMNLP 2020, pages 3406-3425, arXiv:2001.08604
- (Mohapatra et al., 2021) Simulated Chats for Building Dialog Systems: Learning to Generate Conversations from Instructions, ACL 2021, pages 1190-1203
- (Terragni et al., 2023) In-Context Learning User Simulators for Task-Oriented Dialog Systems, CoRR journal, volume abs/2306.00774, arXiv:2306.00774

References: Task-oriented

- (Li et al., 2023) Controllable Dialogue Simulation with In-Context Learning, ACL 2022, pages 4330-4347, arXiv:2210.04185
- (Lin et al., 2021) Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems, Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 445–456, arXiv:2106.08838
- (Ahmad et al., 2023) INA: An Integrative Approach for Enhancing Negotiation Strategies with Reward-Based Dialogue System, EMNLP 2023, pages 2536–2549, arXiv:2310.18207
- (Fellows et al., 2021) Task-oriented Dialogue Systems: performance vs. quality-optima, a review, CCSIT 2022 pp. 69-87, arXiv:2112.11176
- (Kwan et al., 2023) A Survey on Recent Advances and Challenges in Reinforcement Learning Methods for Task-oriented Dialogue Policy Learning, Springer Volume 20, pages 318–334
- (Budzianowski et al., 2018) MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, EMNLP 2018, pages 5016–5026, AAAI 2018, pages 6642--6649, arXiv:1810.00278
- (Lee et al., 2018) Zero-Shot Adaptive Transfer for Conversational Language Understanding, arXiv:1808.10059

References: Open Domain

- (Chen et al., 2022) Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding, arXiv:2210.14169.
- (Zheng et al., 2023) AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In Findings of the Association for Computational Linguistics: ACL 2023, 1552–1568.
- (Lin et al., 2022) Knowledge-Grounded Conversational Data Augmentation with Generative Conversational Networks, In Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, 26–38.
- (Kim et al., 2022) BotsTalk: Machine-sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics (EMNLP), 5149–5170.
- (Chen et al., 2023) PLACES: Prompting Language Models for Social Conversation Synthesis, EACL, 814–838.
- (Kim et al., 2023) SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. Empirical Methods in Natural Language Processing, EMNLP, 12930–12949.
- (Lee et al., 2022) PERSONACHATGEN: Generating Personalized Dialogues using GPT-3, the 1st Workshop on Customized Chat Grounding Persona and Knowledge, ACL, 29–48.
- (Jandaghi et al., 2023) Faithful Persona-based Conversational Dataset Generation with Large Language Models, arXiv:2312.10007.

References: Conversational Information Seeking

- (Wu et al., 2022) DG2: Data Augmentation Through Document Grounded Dialogue Generation, SIGDIAL, 204–216.
- (Bao et al., 2023) A Synthetic Data Generation Framework for Grounded Dialogues, ACL, 10866–10882.
- (Dai et al., 2022) Dialog Inpainting: Turning Documents into Dialogs, ICML, 4558–4586.
- (Hwang et al., 2022) Multi-Type Conversational Question-Answer Generation with Closed-ended and Unanswerable Questions, AACL, 169–177.
- (Hwang and Lee, 2022) Conversational QA Dataset Generation with Answer Revision, COLING, 1636–1644.
- (Kim et al., 2022) Generating Information-Seeking Conversations from Unlabeled Documents, Conference on Empirical Methods in Natural Language Processing, EMNLP, 362–2378.
- (Abbasiantaeb et al., 2024) Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions, In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), 8–17.
- (Askari et al., 2024) Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking dialogs, arXiv preprint arXiv:2402.11633.
- (Wang et al., 2023) Target-oriented Proactive Dialogue Systems with Personalization: Problem Formulation and Dataset Curation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, 1132–1143.

References: Conversational Information Seeking

- (Leszczynski et al., 2022) Conversational Music Retrieval with Synthetic Data. In Second Workshop on Interactive Learning for Natural Language Processing at NeurIPS.
- (Leszczynski et al., 2023) Talk the Walk: Synthetic Data Generation for Conversational Music Recommendation. arXiv preprint arXiv:2301.11489.
- (Macina et al., 2023) MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In Findings of the Association for Computational Linguistics: EMNLP 2023, 5602–5621.
- (Joko et al., 2024) Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search, SIGIR.
- (Qu et al., 2018) Analyzing and Characterizing User Intent in Information-seeking Conversations, SIGIR.
- (Alberti et al., 2019) Synthetic QA Corpora Generation with Roundtrip Consistency, ACL, 6168–6173.