

# Presenters



**Heydar Soudani**

PhD Candidate  
Radboud University  
heydar.soudani@ru.nl



**Evangelos Kanoulas**

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



**Roxana Petcu**

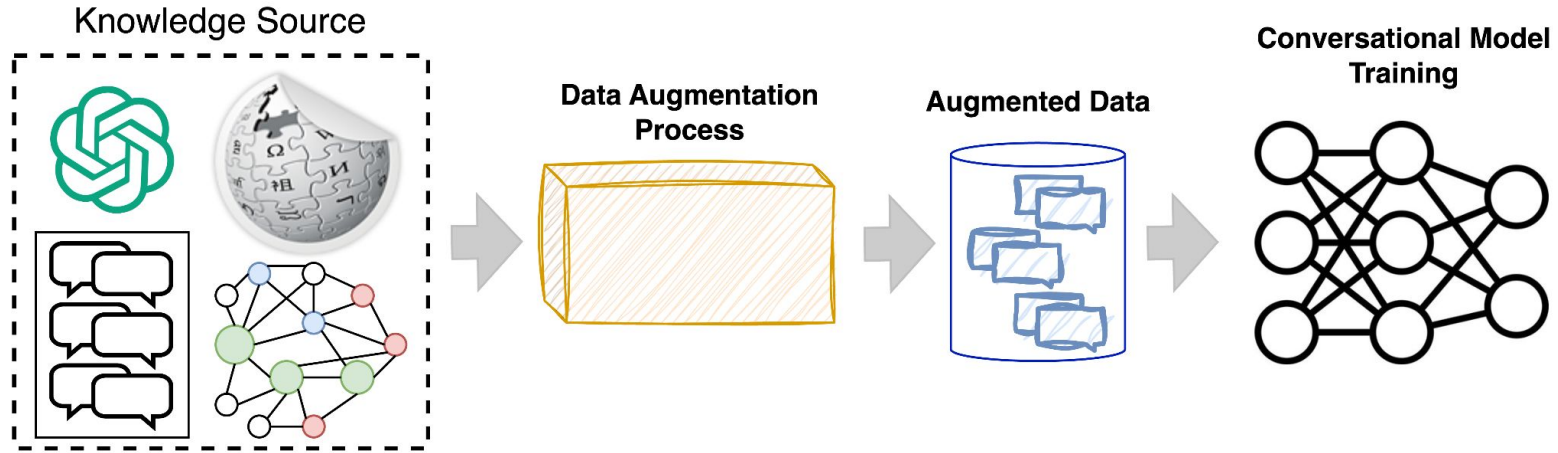
PhD Candidate  
University of Amsterdam  
—@uva.nl



**Faegheh Hasibi**

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

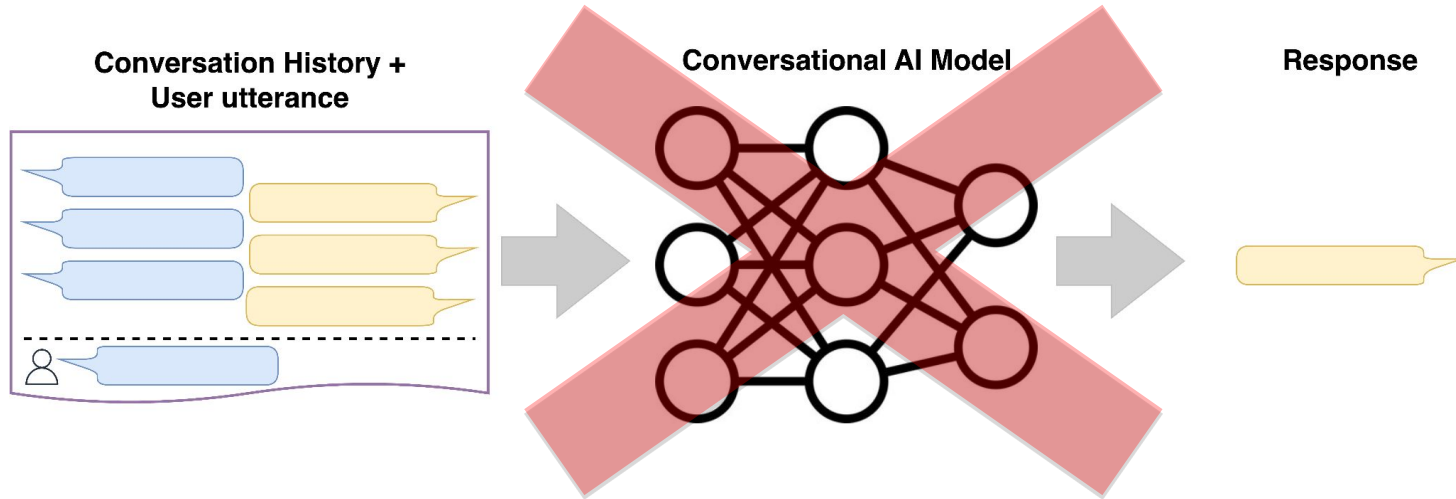
# What is This Tutorial About?



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

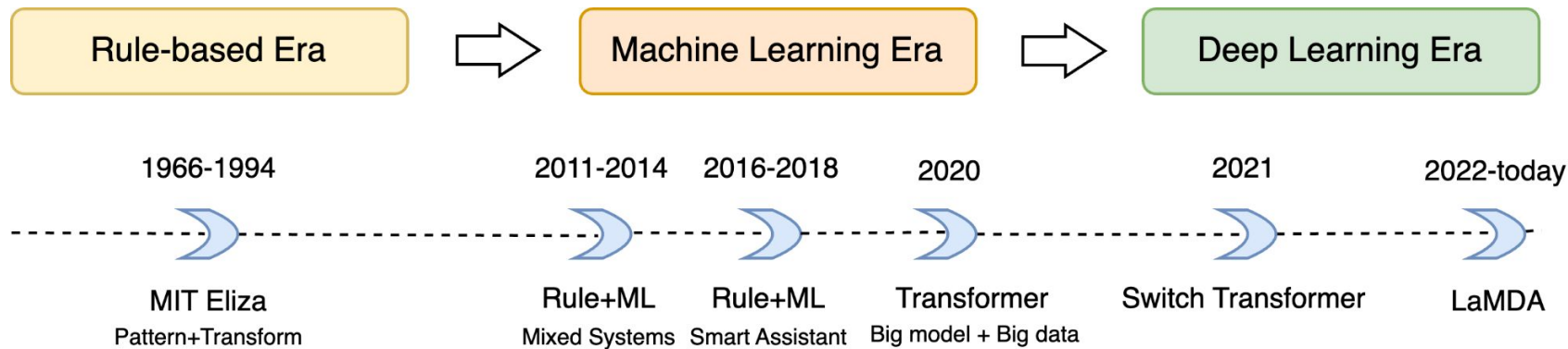
# What is This Tutorial **NOT** about?

Developing conversational AI systems



# 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)

# Two Classes of Conversational AI systems

## Task Oriented

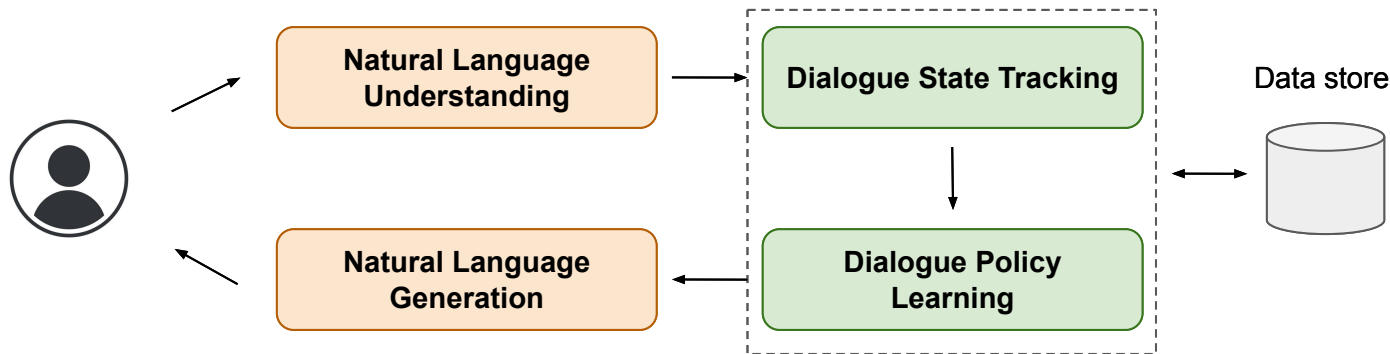
- Focuses on task completion
- Solves specific problems in a certain domain; e.g., restaurant reservation, movie ticket booking, etc.
- Developed using pipeline or end-2-end approaches

## 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

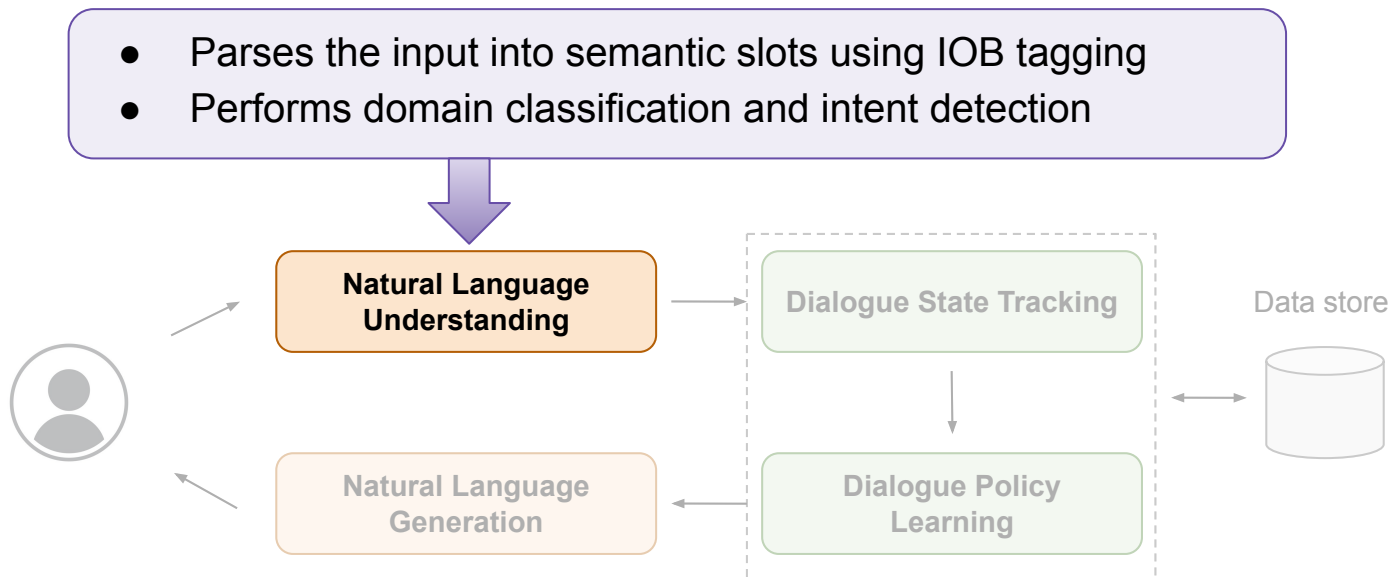
# Task Oriented Dialogue Systems

- Need to accurately handle users' requests
- Often developed using a modular pipeline approach



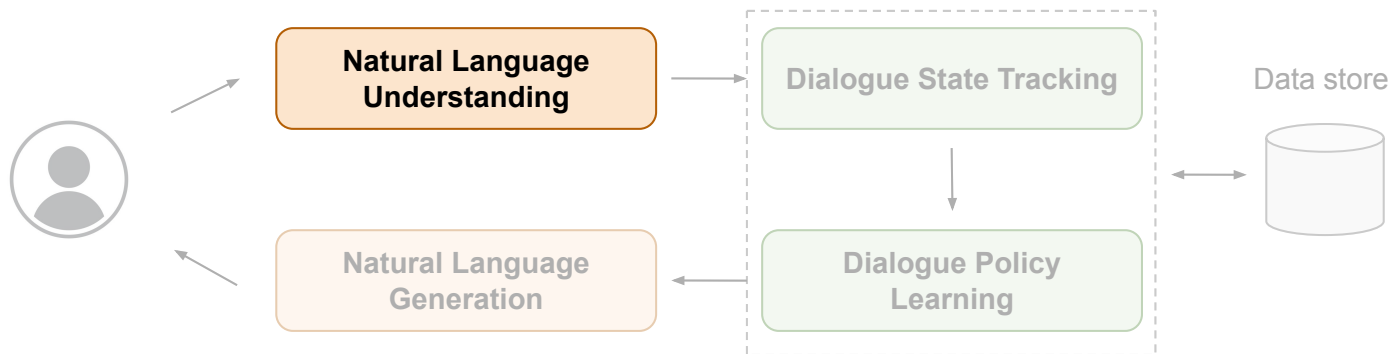
(Ni et al., 2023)

# The Pipeline Approach



# The Pipeline Approach

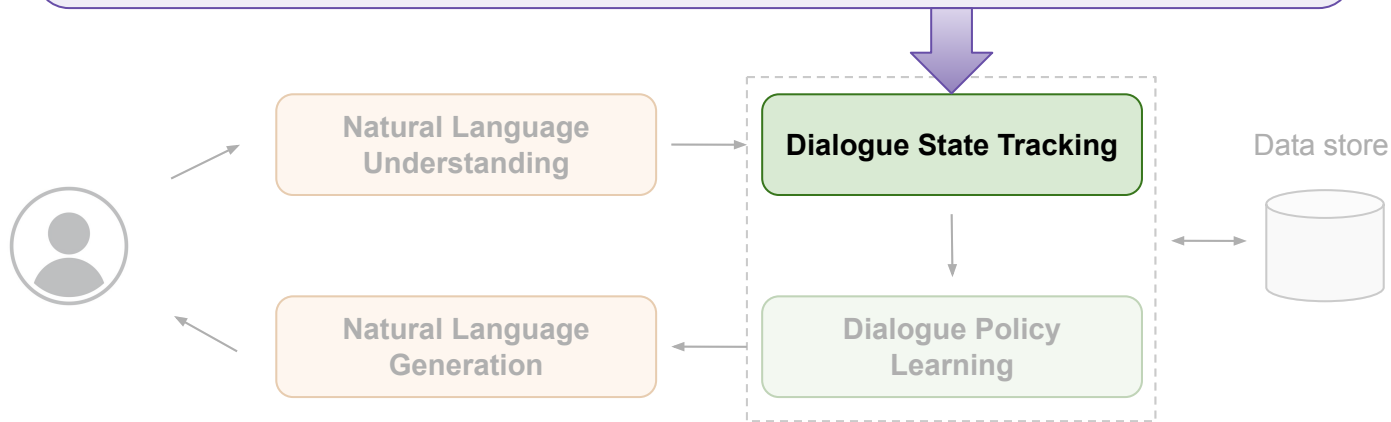
<b>Example</b>	Recommend	a	restaurant	at	New	York	today
<b>Slots</b>	○	○	○	○	B-desti	I-desti	B-time
<b>Intent</b>	inform			<b>Domain</b>		restaurant	





# The Pipeline Approach

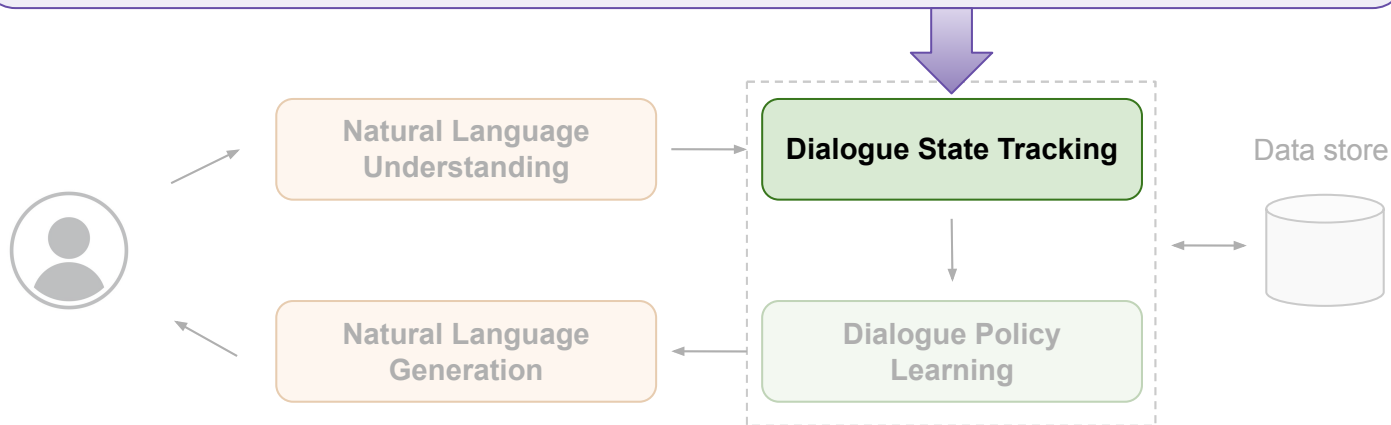
- Looks up the dialogue history and current turn and decides which slots can be filled
- Finds values from user utterances to fill in pre-existing slots list



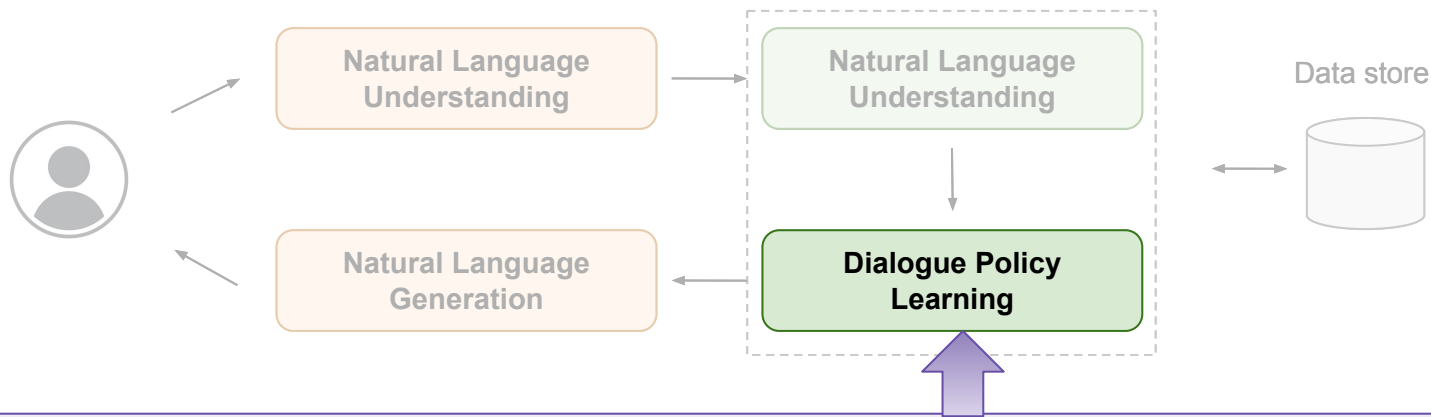
# The Pipeline Approach

**Example:** “Recommend a restaurant at New York today”

**intent:** inform; **domain:** restaurant; **name:** \_; **date:** today; **pricerange:** \_

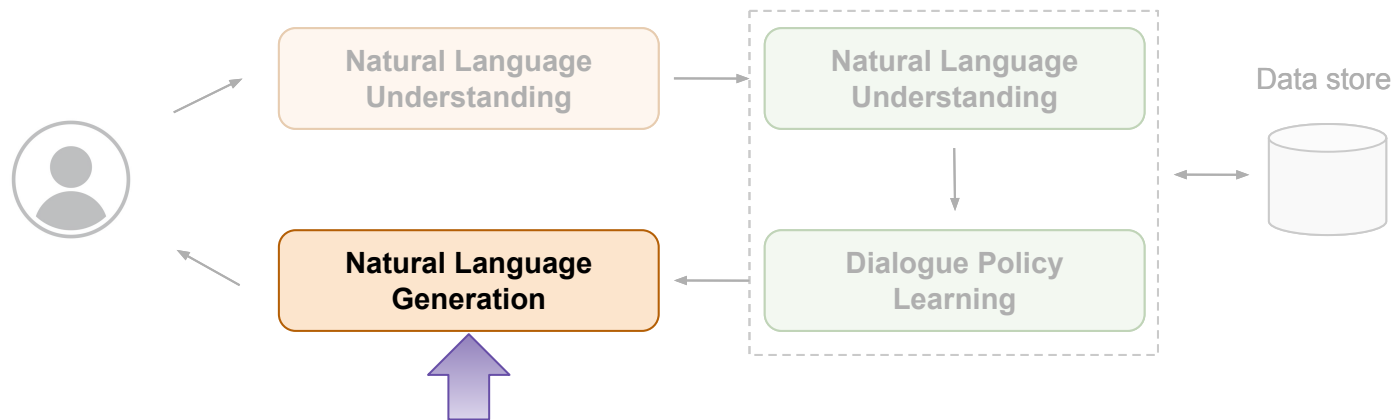


# The Pipeline Approach



- Learns a dialogue policy that generates the next satisfactory action based on the current dialogue state
- Often trained using supervised and reinforcement learning
- **E.g.**, *Inform (name=Kochi, desti=New York, cuisine=korean)*

# The Pipeline Approach



- Converts the dialogue action from the previous step to natural language representation
- E.g., *There is a Korean restaurant called Kochi in New York.*

# 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

- Context awareness
  - Response coherence
- Response diversity
  - Avoid making dull responses (Zhao et al., 2019)
- Controllable generation (Chen et al., 2023)
  - More natural and satisfying interactions with AI agents
  - Policy planning; eg., target-oriented system, ask clarifying questions, or make suggestions
- Informative responses
  - Pre-defined FAQs or knowledge bases
  - Knowledge graphs

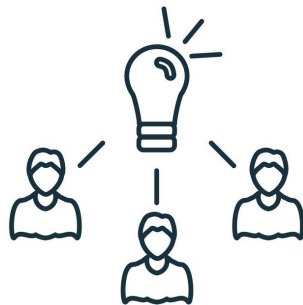
# Why data augmentation for conversational AI?

**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
- Human-authored data possess weaker performance on out of distribution samples (Yang et al., 2020)
- Do not necessarily consider the full diversity of question types (Gu et al., 2021)



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



**But is not prompt engineering/tuning of large language models (LLMs) a solution?**

# Yes, Some Success so far ...

## 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.

## But not Always ...

- Utilizing LLMs is not always possible
  - Time and compute budget constraints by small/medium scale enterprises
  - Although ChatGPT and its successor GPT-4 are successful as products, they remain proprietary and non-replicable (Kong et al., 2023)
- Generic LLMs are not aware of domain specific data (Deng et al., 2023); e.g.,
  - Product information in a retail domain
  - Sensitive data specific to police domain
- Increasing response quality for some subtasks of conversational systems
  - Availability of limited data for some complex tasks; proactive agents

# Tutorial Schedule

