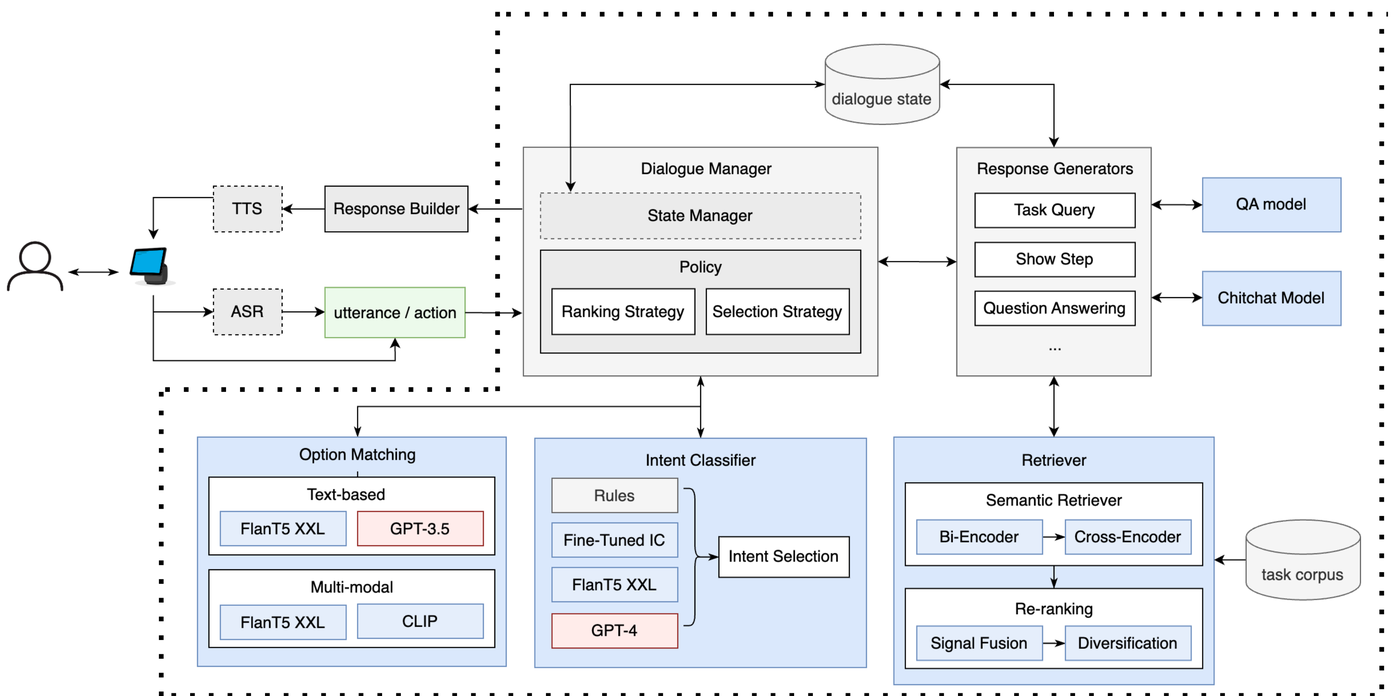
Fine-tuning Language Models for End-to-End Task-Oriented Dialogue

Chris Samarinas, Pracha Promthaw, Atharva Nijasure,
Rohan Lekwani, Hansi Zeng, Hamed Zamani

alexa prize

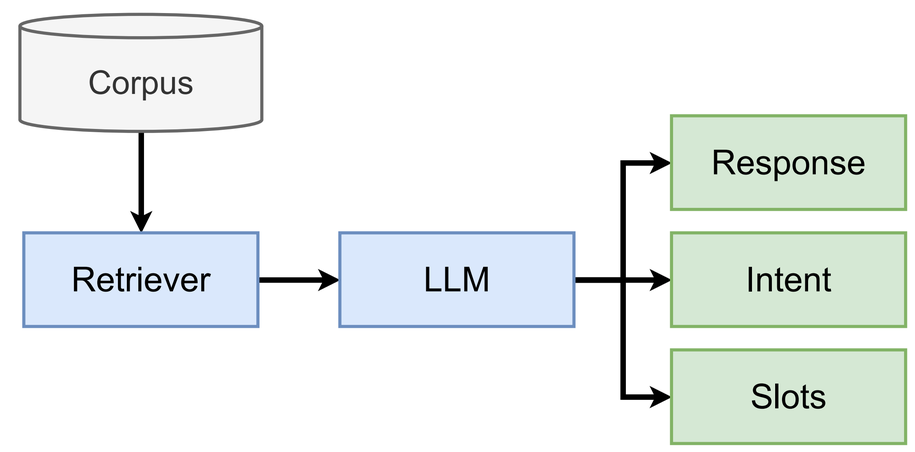
- We propose a framework for developing comprehensive TOD systems capable of performing intent classification, slot filling conversational question-answering, using external tools or sources, and even casual chitchat.
- We demonstrate the effectiveness of this framework on a real-world system developed for the Alexa Prize Taskbot Challenge using 4,000 conversations to fine-tune LLaMA.

Modular TOD



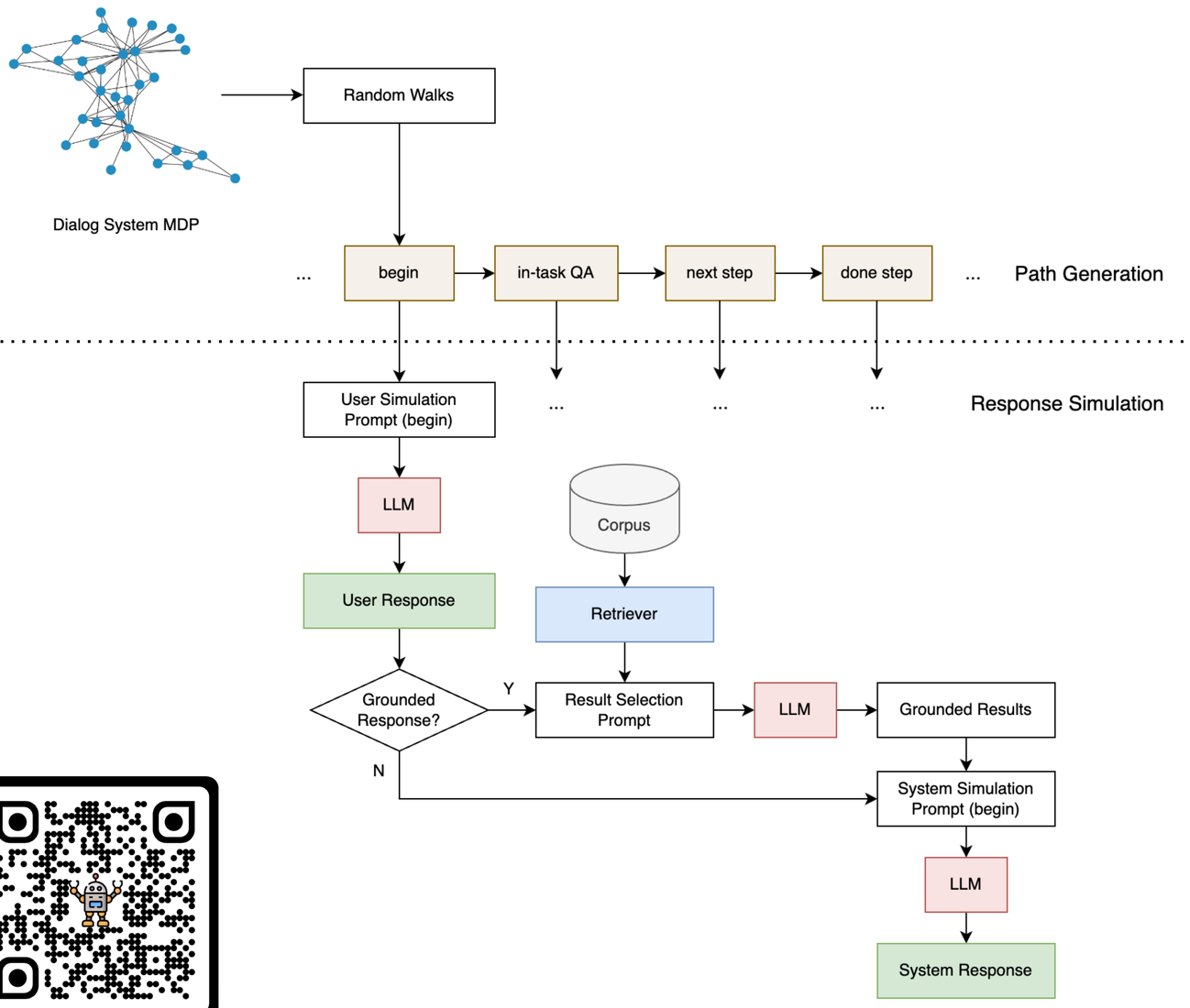
- ✓ Easier patching of issues
- ✓ Can be more reliable
- ✗ Repetitive responses
- ✗ Difficult to debug & maintain
- ✗ Limited contextual understanding

End-to-End TOD



- ✓ Easier to deploy & maintain
- ✓ Diverse and natural responses
- ✓ Better contextual understanding
- ✗ Can have halucinations
- ✗ Difficult to patch issues
- ✗ Difficult to make robust

Conversational Data Generation



DEMO

- The framework's core requirement is the definition of a **state transition graph**, encapsulating the desired behavior of the TOD system. In this graph, nodes represent system states, while edges symbolize user intents.
- Based on this graph, a set of random walks is generated. For each node and edge in these random walks, a large language model (LLM) with a custom prompt simulates a response from either the system or the user.
- In cases where the system involves retrieval, a corpus of documents serves as the seed to simulate search requests and QA based on specific documents. A retrieved document is appended to the LLM prompt to condition the response generation, ensuring the responses are contextually accurate with limited hallucinations.
- We demonstrate the effectiveness of this framework on a real-world system developed for the Alexa Prize Taskbot Challenge using **4,000** synthetic conversations to fine-tune **LLaMA 2 7B** with **QLoRA**.