Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

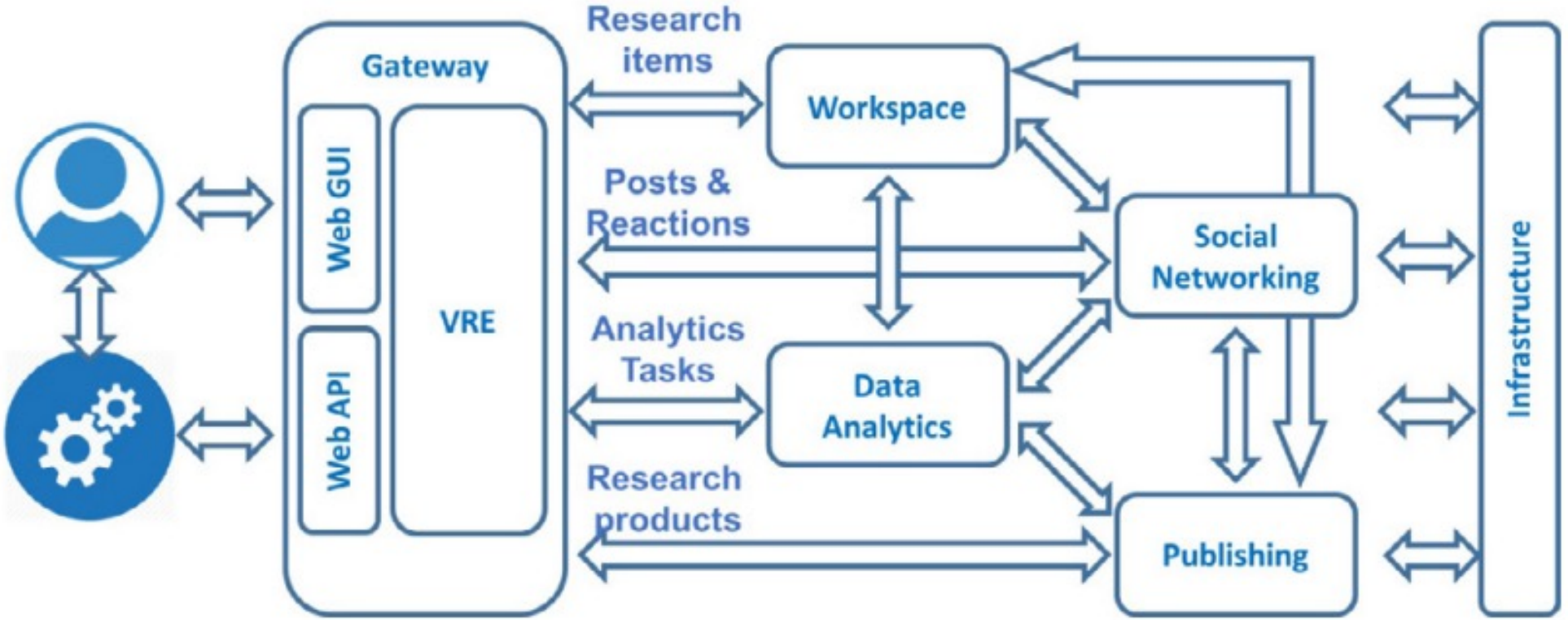
Introducing Janet: Early Findings on a Conversational Agent for Virtual Research Environments

Ahmed Salah Tawfik Ibrahim, Leonardo Candela
Istituto di Scienza e Tecnologie dell'Informazione - Italian National Research Council

16th International Workshop on Science Gateways (IWSG2024)
18-20th June 2024, Toulouse, France

Motivations

- Science gateways are data-rich
  - Example: D4Science



Motivations

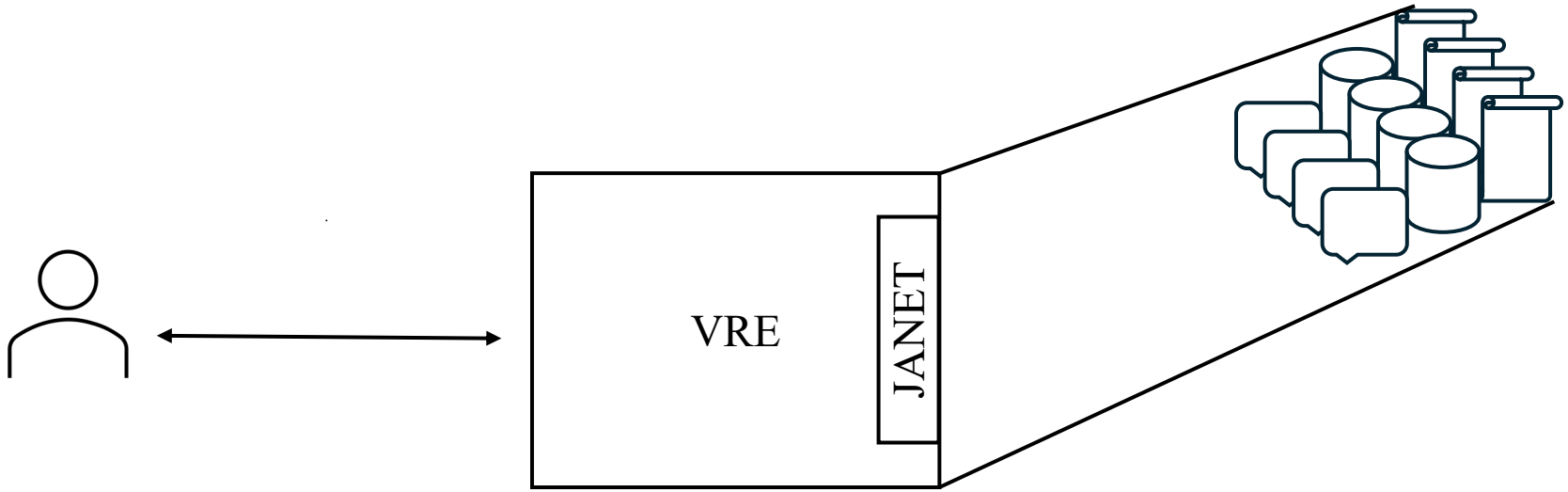
- Need to facilitate access to the data in VREs:
  - The community using the VRE has the potential to grow
  - Navigating the existing content could become a challenge
  - Time can be saved

- The rise of large language models (LLMs):
  - Hallucination: limitation or advantage?
  - The potential of retrieval-augmented generation (RAG)
  - The power of prompting

Goal

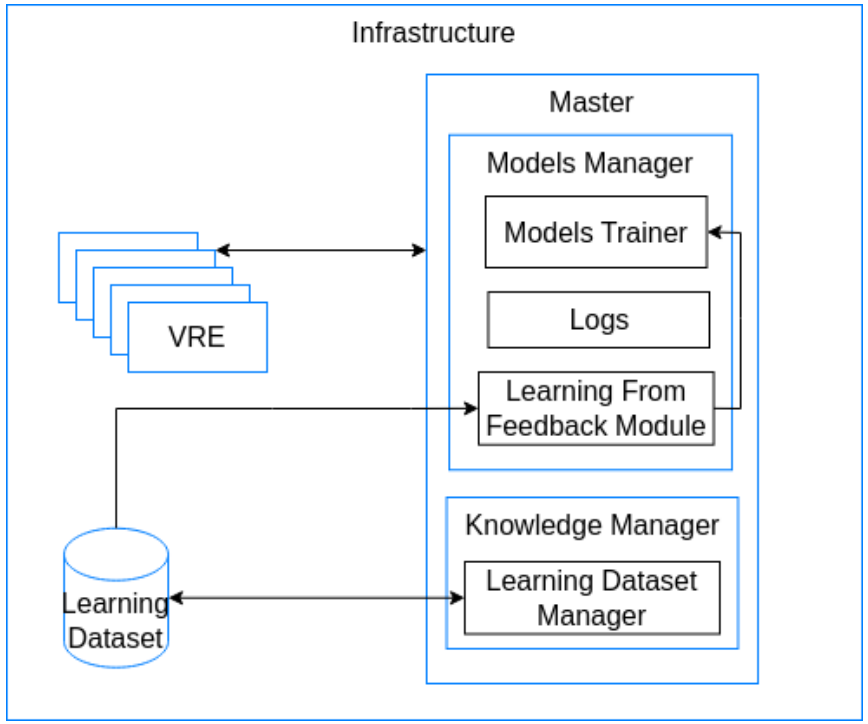
- Build an assistant for researchers in VREs
  - Leverages recent AI technologies
  - Facilitates the utilization of VRE resources
  - Expandable
  - Self-improving

Framework

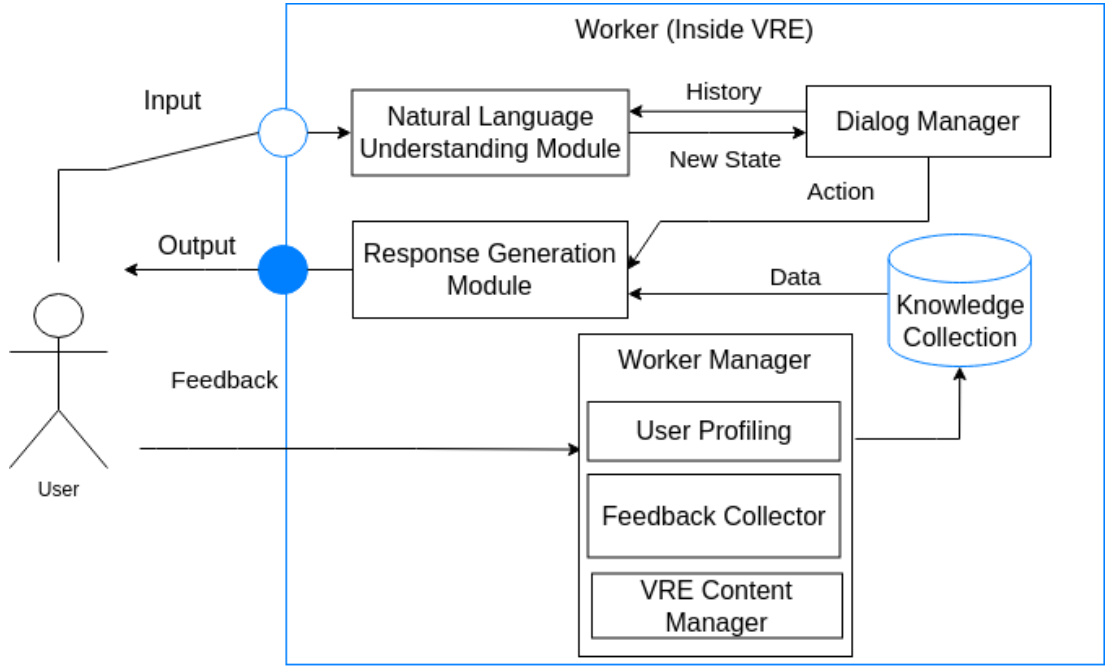


System Design

- Master
  - Models Trainers
  - Learning Dataset Storage

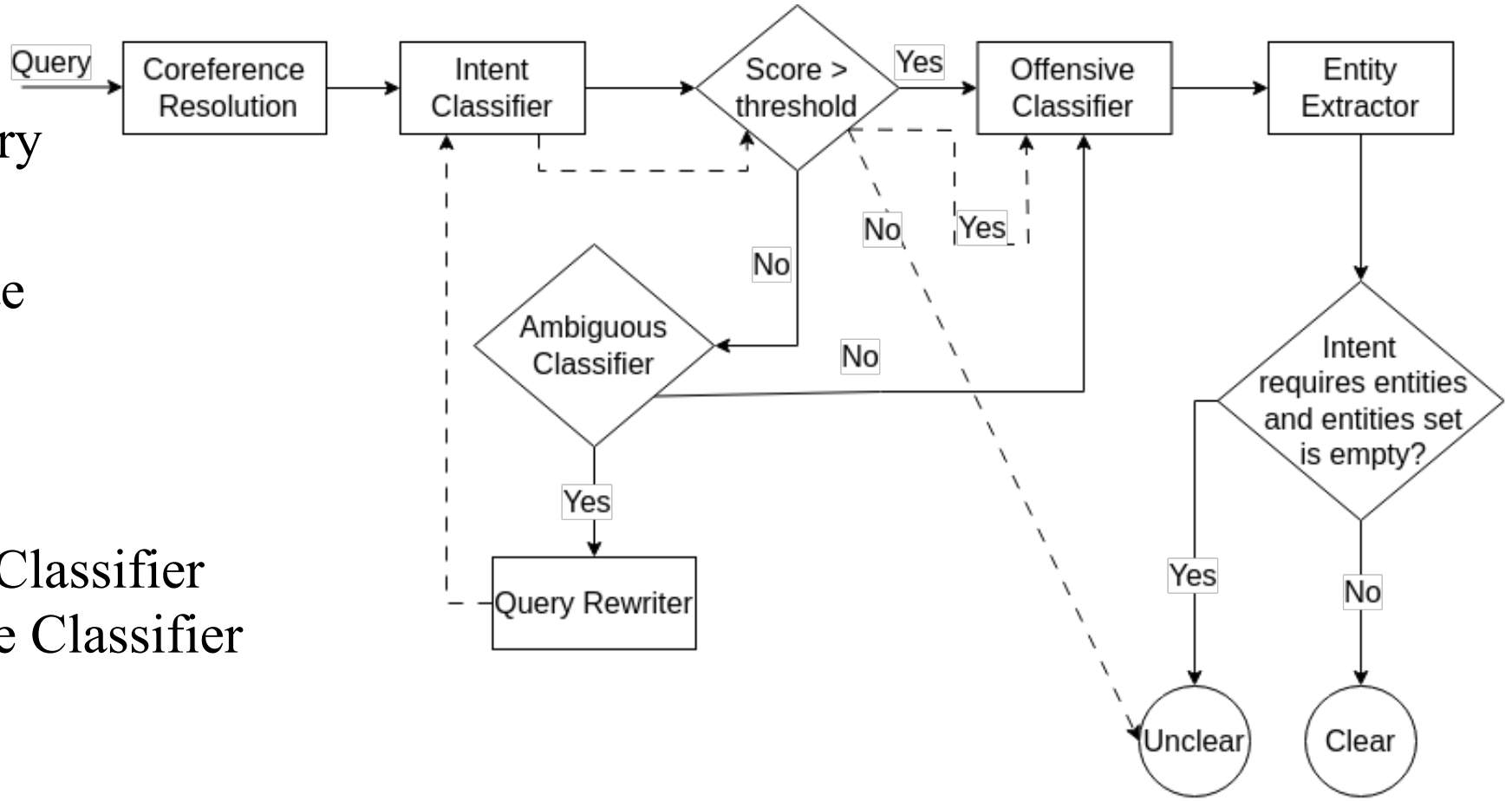


- Worker
  - Natural Language Understanding Module
  - Dialog Manager
  - Response Generator
  - Worker Manager



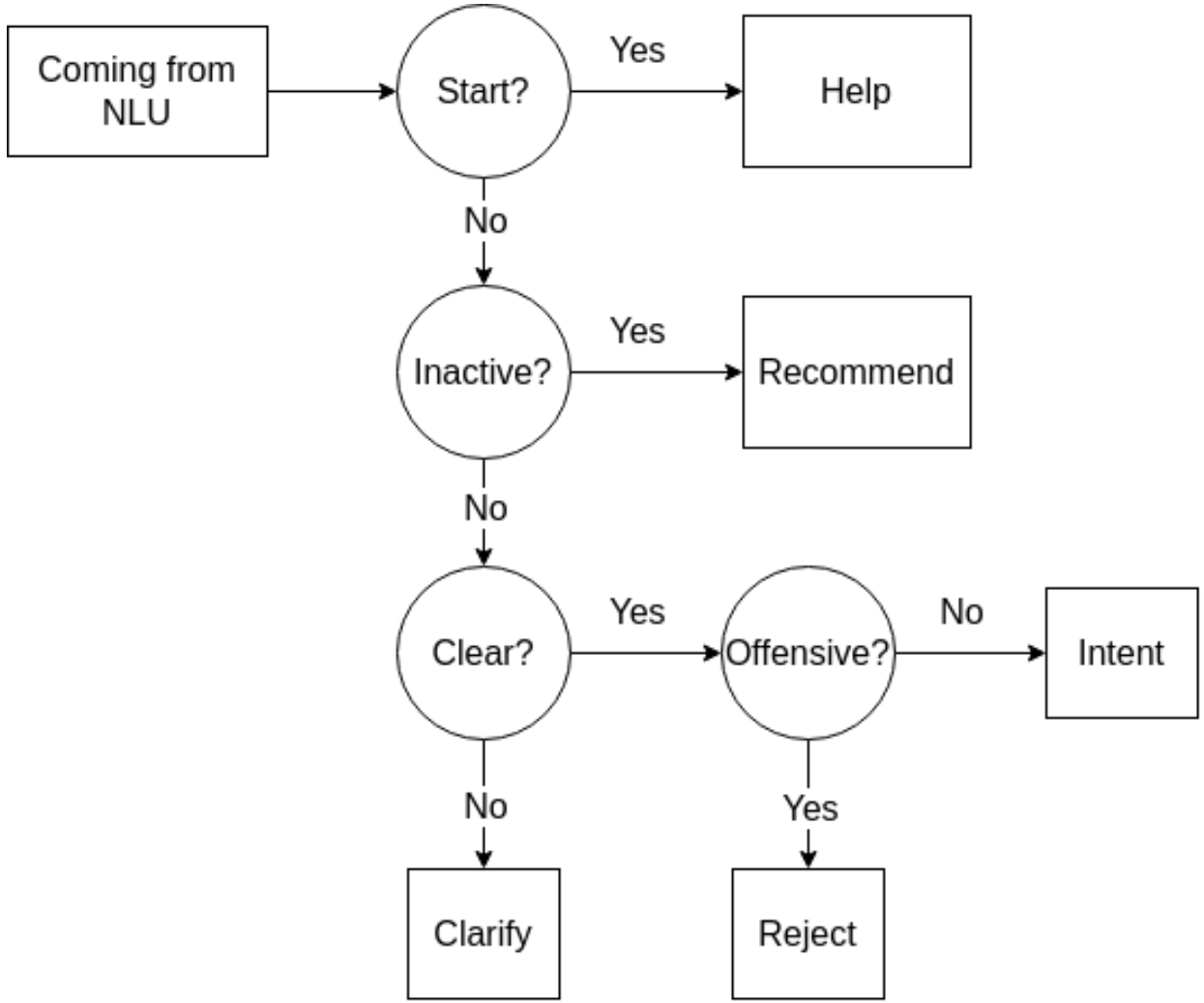
Natural Language Understanding

- Input
  - User Query
  - Conversation History
- Output
  - Current Dialog State
- Models:
  - Intent Classifier
  - Entities Extractor
  - Ambiguous Query Classifier
  - Offensive Language Classifier
  - Query Rewriter



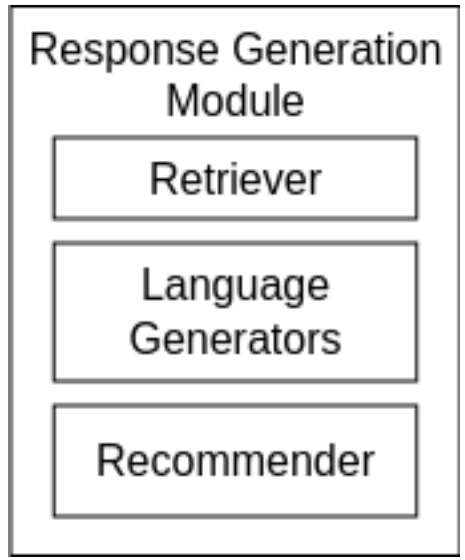
Dialog Manager

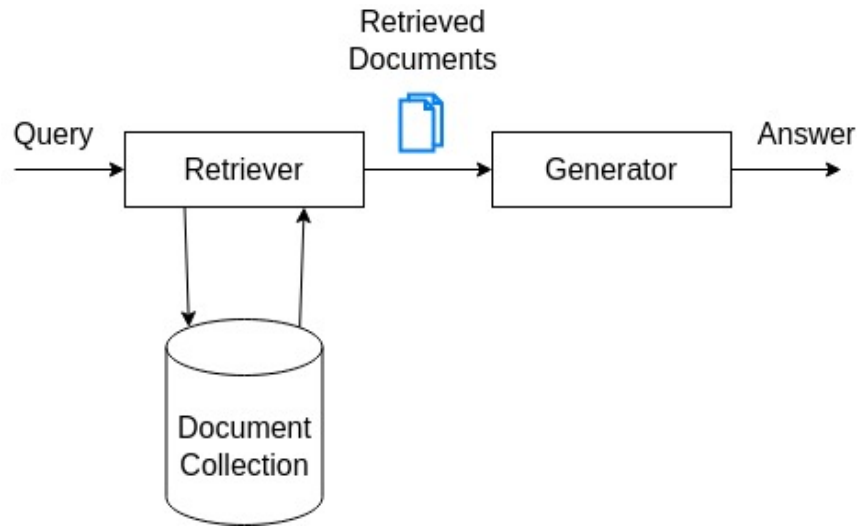
- Input
  - New State
- Model
  - Finite State Machine
- Function:
  - Update Chat History
  - Select Response Generation Mode



Response Generator

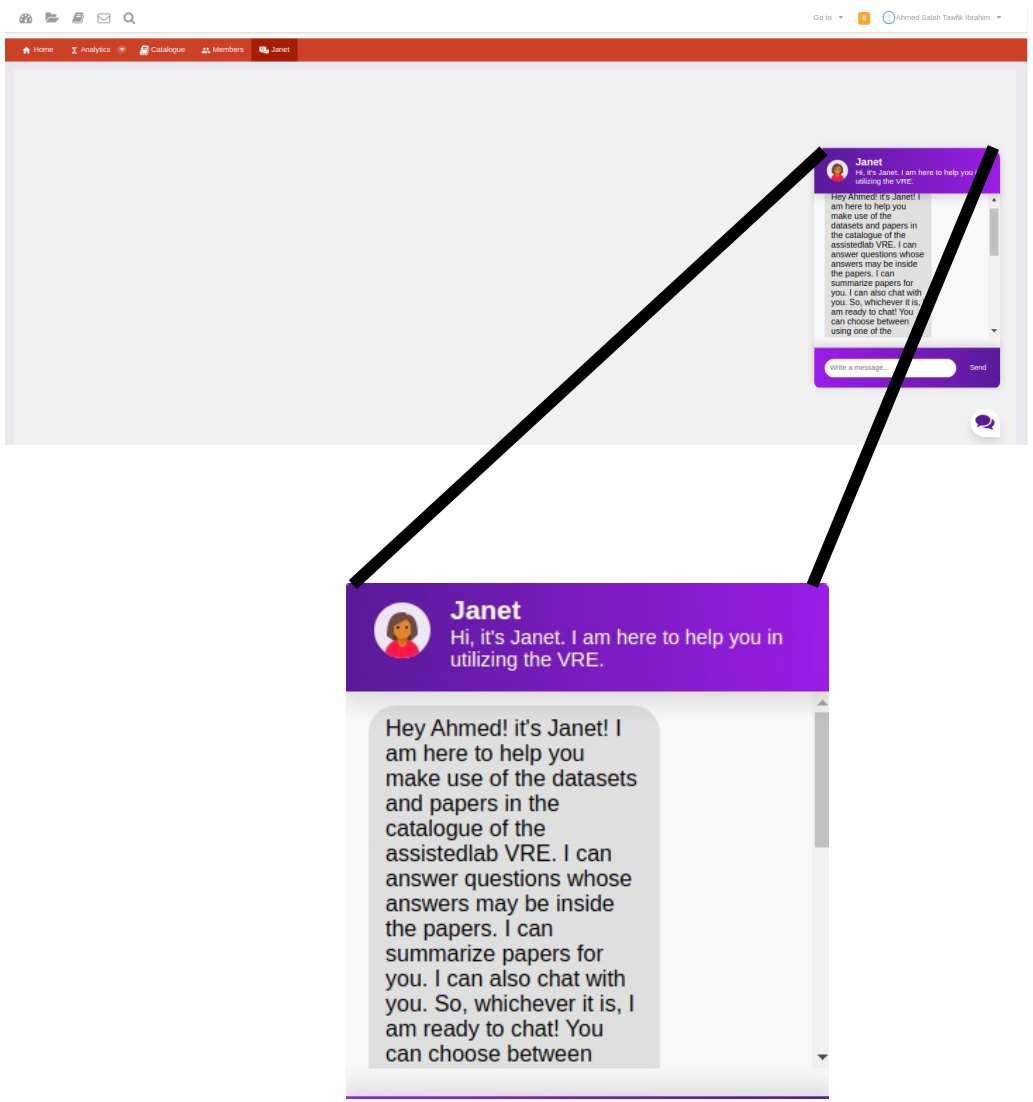
- Input:
  - Current Dialog State
  - Response Generation Mode
- Response Generation Modes:
  - Recommend
  - Help
  - QA (Retriever and Generate)
  - Retrieve
  - Generate





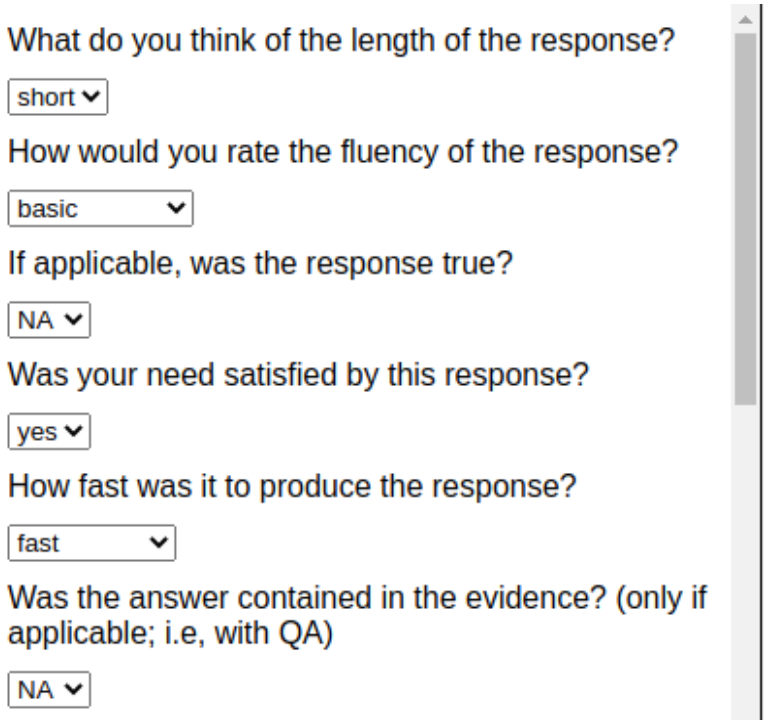
Deployment

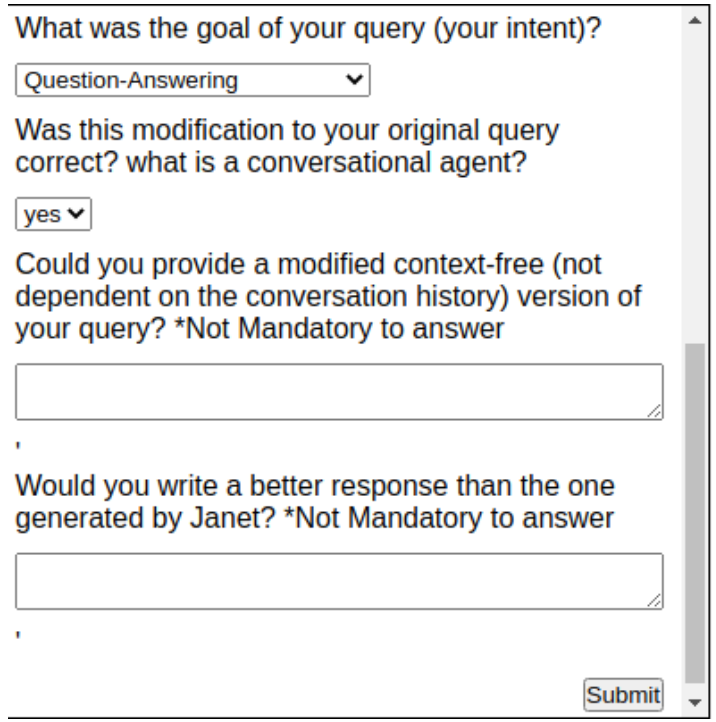
- Docker Containers
  - Frontend – React JavaScript
  - Backend – Python
  - Database – PostgreSQL
- Orchestration:
  - Docker-Compose



Testing

- Feedback Form





- Real users

Evaluation

TABLE I: Number of Responses evaluated by Length and Fluency

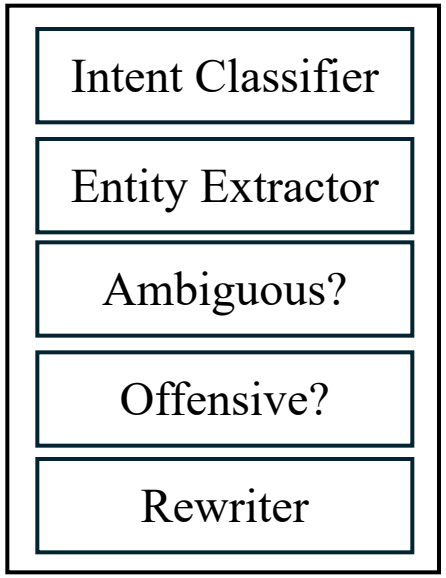
	Short	Appropriate	Long
Response Length	16	78	6
	Basic	**Intermediate**	**Fluent**
Response Fluency	19	6	75

TABLE II: Number of Responses evaluated by Factuality and Usefulness

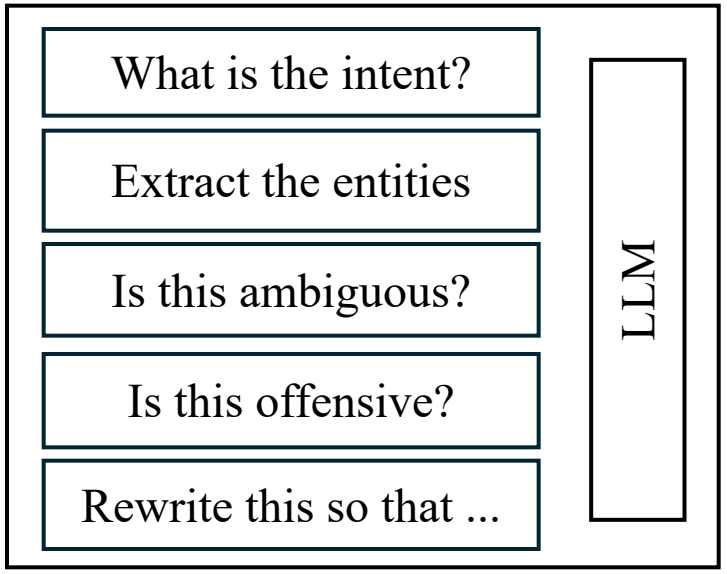
	True	False
Response Factuality	94	6
	Useful	**Useless**
Response Usefulness	69	31
Evidence Usefulness	13	6

Ongoing Modifications

- NLU
  - Does it really need one model per functionality?



NLU with models

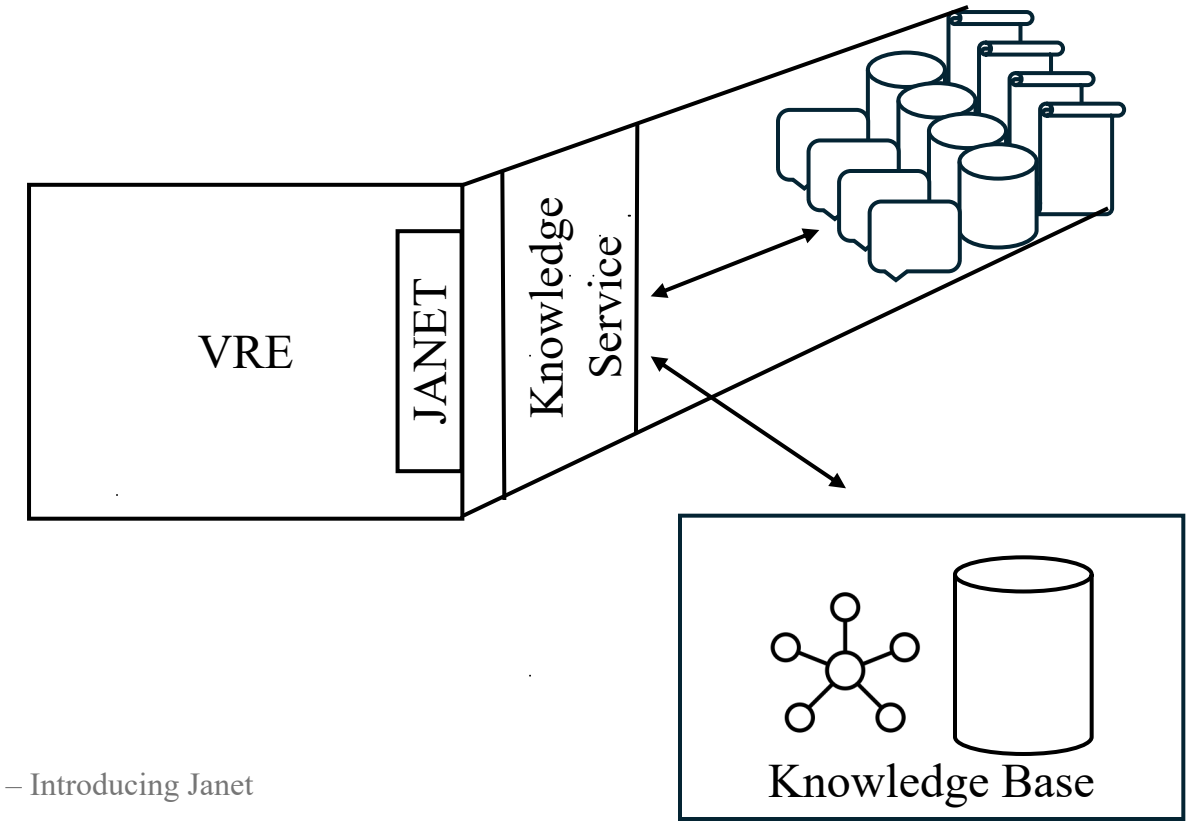


NLU with LLM+Prompts

Ongoing Modifications

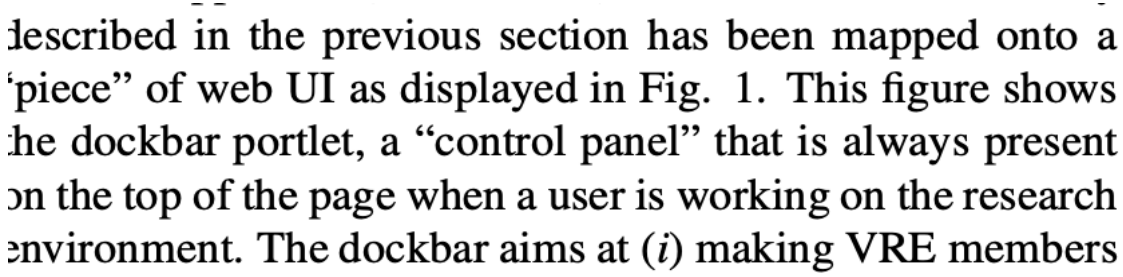
- Generator
  - Currently, finetuned T5 seq2seq model
  - Google Gemma Instruct model

- Knowledge Base
  - Currently, FAISS index
  - ChromaDB vector store
  - Knowledge service



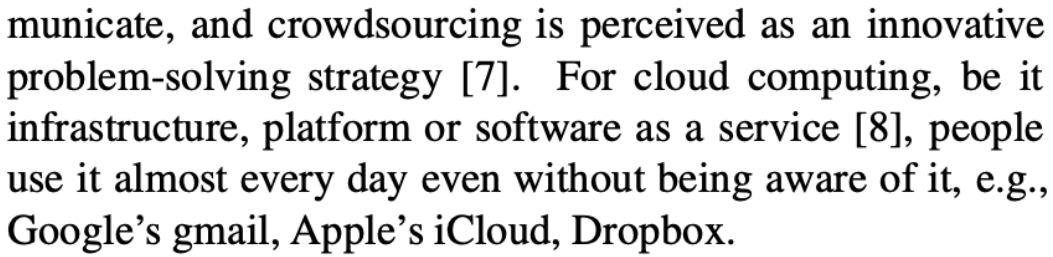
Limitations

- Data Extraction
  - Multimodality
  - "Unclean" text



described in the previous section has been mapped onto a 'piece" of web UI as displayed in Fig. 1. This figure shows the dockbar portlet, a "control panel" that is always present on the top of the page when a user is working on the research environment. The dockbar aims at (*i*) making VRE members

Figure 1: Multimodality



municate, and crowdsourcing is perceived as an innovative problem-solving strategy [7]. For cloud computing, be it infrastructure, platform or software as a service [8], people use it almost every day even without being aware of it, e.g., Google's gmail, Apple's iCloud, Dropbox.

Figure 2: Unclean Text

Limitations

- Training Data
  - Lack of publicly available datasets
  - Expensive to construct datasets
- Performance
  - Large models require extensive resources
- Testing
  - Internal testing
  - Limited time

Thank You

Contact the authors

Leonardo Candela

Ahmed Ibrahim