QuizBee — Local CLI Chatbot using Hugging Face LLM

A lightweight, modular chatbot that runs locally using TinyLlama-1.1B-Chat. Designed to handle factual, multi-turn conversations via a command-line interface.

Overview

QuizBee is a command-line chatbot developed as part of a technical ML assignment. It uses a Hugging Face small-scale language model to simulate intelligent multi-turn conversations. The focus is on local deployment, modular code organization, and short-term memory using a sliding window buffer.

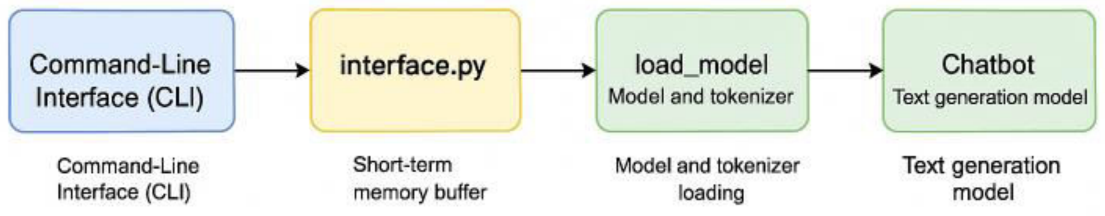
Objective

Title: Objective & Requirements

- Build a CLI chatbot using a small HF model
- Maintain recent conversation history (sliding memory)
- Use modular Python code
- Deliver source code, README, and a demo video

Architecture Overview

Title: Architecture Breakdown



- model_loader.py: Loads model & tokenizer
- chat_memory.py: Handles memory buffer (last 5 turns)
- interface.py: CLI loop, prompt building, threading

Features

Local inference — no internet/GPU required after initial download

Hugging Face pipeline for quick integration

Modular code (interface.py, chat_memory.py, model_loader.py)

Sliding memory window (last 3–5 turns) for context coherence

System prompt-driven behaviour control

Clean CLI interface with graceful /exit

Tested on: Windows 10 / Ubuntu 22.04, Python 3.10+

Model Choice

Title: Model Used: TinyLlama-1.1B-Chat

- Small footprint, 1.1B parameters
- Trained on 3T tokens (web, Wikipedia, books)
- Chat-tuned for factual Q&A, math, trivia

1. Pretraining Data (~3T tokens)

- Mixture of:
    - Web text (Common Crawl, Wikipedia, Books)
    - Technical content (e.g., StackExchange, arXiv)
    - Code snippets (possibly from The Stack or CodeParrot-style sets)

2. Chat Tuning Data

- Open-source chat datasets (comparable to OpenAssistant, ShareGPT)
- GPT-4-aligned outputs or distilled prompts from user-assistant exchanges
- Examples range across:
    - World capitals
    - Science facts
    - Basic math
    - Definitions
    - Trivia

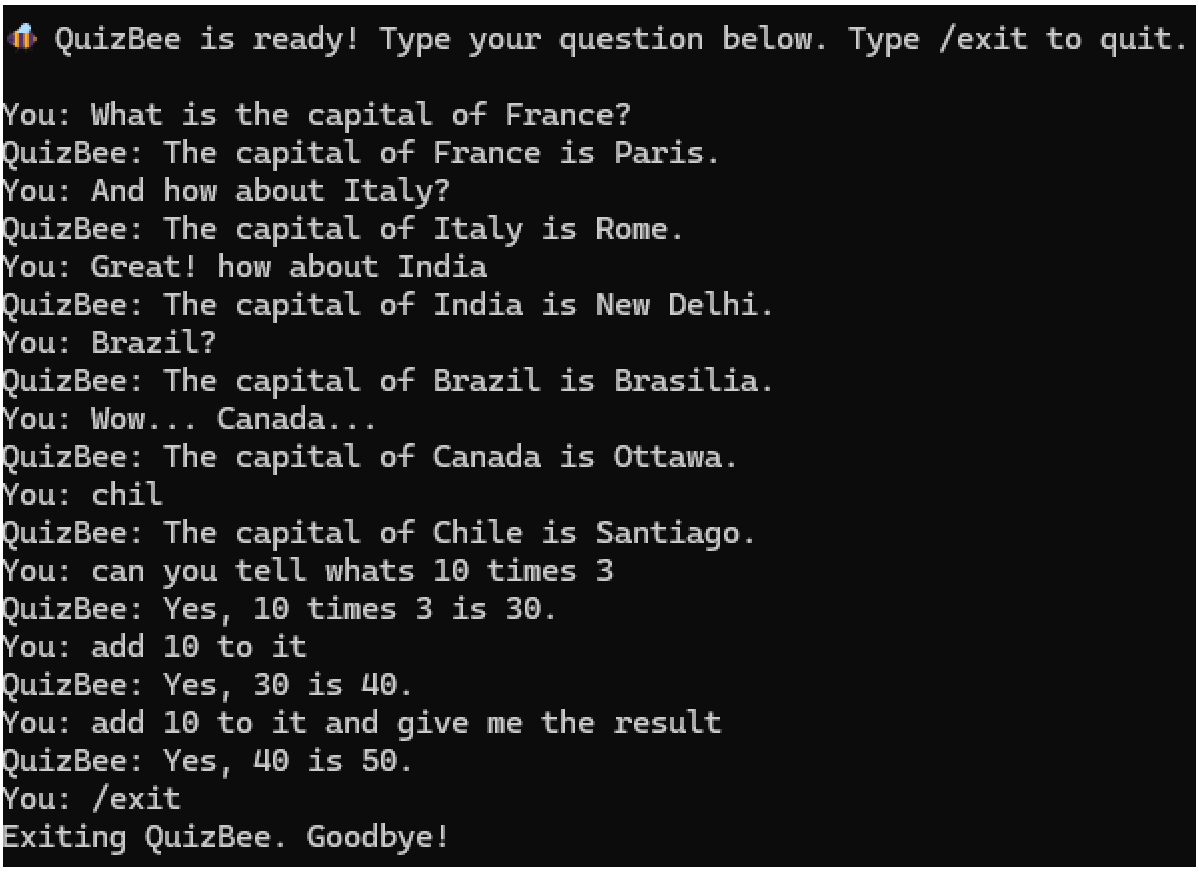
Sliding Memory Handling

Title: Short-Term Memory: 5 Turns

- Implemented using deque in chat_memory.py
- Maintains last 5 user+bot turns
- Enables multi-turn coherence

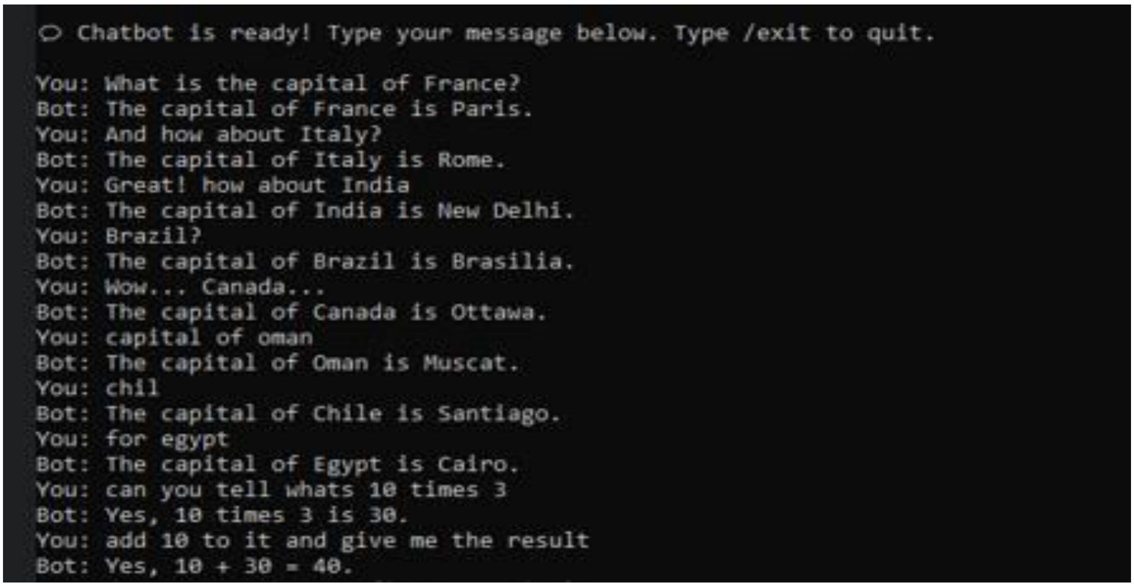
Sample Interactions

Title: Sample Interaction Flow



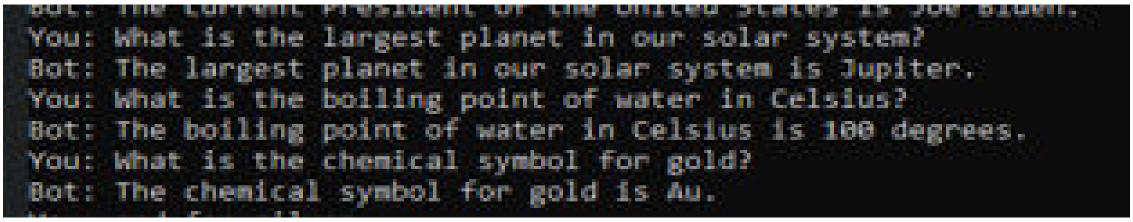
```
🐝 QuizBee is ready! Type your question below. Type /exit to quit.

You: What is the capital of France?
QuizBee: The capital of France is Paris.
You: And how about Italy?
QuizBee: The capital of Italy is Rome.
You: Great! how about India
QuizBee: The capital of India is New Delhi.
You: Brazil?
QuizBee: The capital of Brazil is Brasilia.
You: Wow... Canada...
QuizBee: The capital of Canada is Ottawa.
You: chil
QuizBee: The capital of Chile is Santiago.
You: can you tell whats 10 times 3
QuizBee: Yes, 10 times 3 is 30.
You: add 10 to it
QuizBee: Yes, 30 is 40.
You: add 10 to it and give me the result
QuizBee: Yes, 40 is 50.
You: /exit
Exiting QuizBee. Goodbye!
```

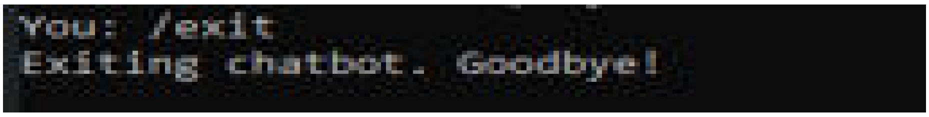


```
💬 Chatbot is ready! Type your message below. Type /exit to quit.

You: What is the capital of France?
Bot: The capital of France is Paris.
You: And how about Italy?
Bot: The capital of Italy is Rome.
You: Great! how about India
Bot: The capital of India is New Delhi.
You: Brazil?
Bot: The capital of Brazil is Brasilia.
You: Wow... Canada...
Bot: The capital of Canada is Ottawa.
You: capital of oman
Bot: The capital of Oman is Muscat.
You: chil
Bot: The capital of Chile is Santiago.
You: for egypt
Bot: The capital of Egypt is Cairo.
You: can you tell whats 10 times 3
Bot: Yes, 10 times 3 is 30.
You: add 10 to it and give me the result
Bot: Yes, 10 + 30 = 40.
```



```
You: What is the largest planet in our solar system?
Bot: The largest planet in our solar system is Jupiter.
You: What is the boiling point of water in Celsius?
Bot: The boiling point of water in Celsius is 100 degrees.
You: What is the chemical symbol for gold?
Bot: The chemical symbol for gold is Au.
```

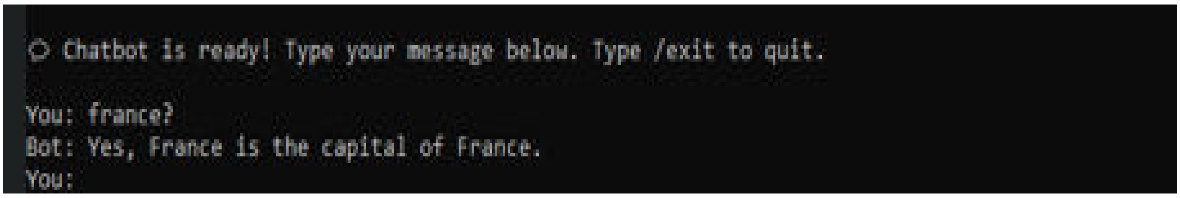


```
You: /exit
Exiting chatbot. Goodbye!
```

Here’s an example of the bot handling context-based follow-up questions about world capitals and more.

Edge Case:

One-word Input



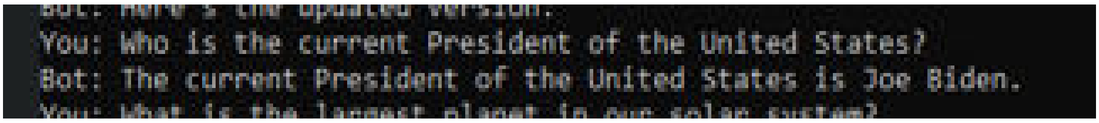
```
○ Chatbot is ready! Type your message below. Type /exit to quit.

You: france?
Bot: Yes, France is the capital of France.
You:
```

"France" → "France is the capital of France"

- Ambiguous user input
- Model infers a question
- Acceptable behavior under no context

Training Window



```
You: Who is the current President of the United States?
Bot: The current President of the United States is Joe Biden.
```

Step-by-Step: How to Run

Project Execution Flow

1. Clone the Repo / Unzip Folder

```
git clone [repo_link] && cd quizbee_chatbot
```

2. Set Up Virtual Environment

```
python -m venv venv
```

```
source venv/bin/activate  # Windows: venv\Scripts\activate
```

3. Install Dependencies

```
pip install -r requirements.txt
```

4. Run the Chatbot

```
python interface.py
```

Demo Starts:

```
💬 Chatbot is ready! Type your message below. Type /exit to quit.

You: What is the capital of Canada?

Bot: The capital of Canada is Ottawa.
```

Final Thoughts & What Next

Title: Reflection & Improvements

- Add fallback for vague/echo inputs
- Support dynamic memory sizing
- Switch to Mistral or quantized models for more speed

Thank You

Title: Thank You!

- GitHub: https://github.com/alekya0221/quizbee_chatbot
- Video: https://drive.google.com/file/d/1BZR-JByNsVYF_nfZVWq28eUQfV2OdVLA/view?usp=drive_link
- Developed by: Alekya Rani Seerapu