

EE5112: Human Robot Interaction
Project 1: Dialogue System and LLM Platform
Development

Group 7

Niu Mu (Matriculation Number)

Wu Zining (A0294373W)

Zhao Jinqiu (Matriculation Number)

September 28, 2025

Contents

1	Abstract	4
2	Introduction	4
2.1	Background	4
2.2	Project Objectives	4
2.2.1	Comparative Analysis	6
2.2.2	Recent Trends and Future Directions	7
2.3	Local LLM Platform Implementation	8
2.4	Performance Comparison: CPU vs GPU Deployment	10
2.5	Comparison of Different Pretrained Models	12
3	Task 3: Dialogue System Development	12
3.1	Dialogue System Architecture	12
3.2	Natural Language Processing Components	12
3.3	Multi-modal Communication	12
3.4	Dialogue Management	12
4	Task 4: System Integration and Testing	12
4.1	Component Integration	12
4.2	System Testing	12
4.3	Performance Optimization	12
5	Task 5: LLM Evaluation	13
5.1	Evaluation Metrics	13
5.2	Evaluation Framework Implementation	13
5.3	Performance Analysis	13
5.4	Solution Implementation	13
5.5	Code Documentation	13
6	Results and Discussion	13

6.1	System Performance Results	13
6.2	Task Achievement Summary	13
6.3	Lessons Learned	13
7	Individual Contributions	14
7.1	Member 1: Niu Mu	14
7.2	Member 2: Wu Zining (A0294373W)	14
7.3	Member 3: Zhao Jinqiu	14
8	Conclusion	14
8.1	Project Objectives Achievement	14
8.2	Future Work	14
9	References	14
10	Appendix	14
10.1	Code Documentation	14
10.2	Configuration Files	15
10.3	User Manual	15

1 Abstract

[Placeholder for abstract content - 150-250 words]

Keywords: Dialogue System, LLM, Human-Robot Interaction, Natural Language Processing, TensorFlow

2 Introduction

2.1 Background

[Placeholder for background content]

2.2 Project Objectives

The main objectives of this project are:

1. To familiarize with the process of developing a dialogue system
2. To familiarize with the working environment and Python packages
3. To familiarize with popular platforms such as TensorFlow
4. To familiarize with popular open source LLMs (Llama, GLM, etc.)
5. To develop a dialogue system and local LLM platform
6. To familiarize with LLM evaluation procedures
7. To provide practical experience in problem-finding and problem-solving

Operational highlights.

- Streaming is enabled by default so the first token appears almost immediately, minimizing perceived latency.
- Conversation history is trimmed to the three most recent turns, preserving context while staying within the 4096 token window.
- Every exchange is timestamped and stored in `conversations/`, which simplifies audits and later evaluation.

Overall, the implementation remains compact yet extensible, allowing higher-level dialogue management or evaluation modules to plug in without touching the inference core. **Key Characteristics:**

- Bidirectional attention in the encoder captures context from both directions
- Unidirectional attention in the decoder enables autoregressive generation
- Explicit separation between understanding (encoding) and generation (decoding) phases

Representative Models:

- **T5 (Text-to-Text Transfer Transformer):** Treats all NLP tasks as text-to-text problems, achieving state-of-the-art performance across diverse benchmarks
- **BART (Bidirectional and Auto-Regressive Transformers):** Combines bidirectional encoder with autoregressive decoder, excelling in text generation and denoising tasks
- **mT5:** Multilingual extension of T5 supporting over 100 languages

Encoder-Only Models Encoder-only models utilize only the encoder component of the transformer architecture, employing bidirectional attention to process input sequences. These models excel at understanding and representation learning tasks rather than text generation.

Key Characteristics:

- Bidirectional attention mechanism captures full context
- Optimized for understanding tasks rather than generation
- Require task-specific heads for downstream applications
- Typically used for classification, named entity recognition, and feature extraction

Representative Models:

- **BERT (Bidirectional Encoder Representations from Transformers):** Pioneer in bidirectional language modeling, achieving breakthrough performance in NLU tasks
- **RoBERTa:** Optimized version of BERT with improved training procedures and longer training duration
- **DeBERTa:** Enhanced BERT with disentangled attention mechanism and enhanced mask decoder
- **ELECTRA:** More efficient pre-training using replaced token detection instead of masked language modeling

Decoder-Only Models Decoder-only models rely exclusively on the decoder component with causal (unidirectional) attention, making them highly effective for autoregressive text generation tasks. This architecture has become the dominant paradigm for modern conversational AI systems.

Key Characteristics:

- Unidirectional attention prevents information leakage during training
- Optimized for text generation and completion tasks
- Can be fine-tuned for various downstream tasks through instruction following
- Generally require larger model sizes to achieve competitive performance

Representative Models:

- **GPT (Generative Pre-trained Transformer) Series:** GPT-1, GPT-2, GPT-3, and GPT-4 represent the evolution of decoder-only models with increasing scale and capabilities
- **LLaMA (Large Language Model Meta AI):** Efficient decoder-only model achieving competitive performance with smaller parameter counts
- **GLM (General Language Model):** Chinese-developed model combining autoregressive and autoencoding approaches
- **PaLM (Pathways Language Model):** Google's large-scale decoder-only model with 540B parameters

2.2.1 Comparative Analysis

Table 1: Comparison of LLM Architecture Types

Aspect	Encoder-Decoder	Encoder-Only	Decoder-Only
Primary Use	Seq2Seq tasks	Understanding tasks	Generation tasks
Attention Mechanism	Bidirectional + Causal	Bidirectional	Causal
Training Efficiency	Medium	High	Low (for large models)
Inference Speed	Medium	Fast	Slow (for large models)
Task Flexibility	High	Medium	High
Parameter Efficiency	Medium	High	Low
Representative Models	T5, BART	BERT, RoBERTa	GPT, LLaMA

Performance Trade-offs:

- **Encoder-Decoder Models:** Offer balanced performance for both understanding and generation tasks, but require more computational resources due to dual architecture
- **Encoder-Only Models:** Excel at understanding tasks with high efficiency, but limited generation capabilities
- **Decoder-Only Models:** Superior generation quality and conversational abilities, but require significant computational resources for training and inference

Application Scenarios:

- **Encoder-Decoder:** Machine translation, text summarization, question answering systems
- **Encoder-Only:** Sentiment analysis, named entity recognition, text classification, feature extraction
- **Decoder-Only:** Conversational AI, creative writing, code generation, instruction following

2.2.2 Recent Trends and Future Directions

Recent developments in LLM architectures show several emerging trends:

- **Scale Integration:** Modern models increasingly combine multiple architectural paradigms (e.g., encoder-decoder with decoder-only components)
- **Efficiency Optimization:** Focus on reducing computational requirements while maintaining performance through techniques like knowledge distillation and model compression
- **Multimodal Integration:** Extension of decoder-only models to handle multiple modalities (text, vision, audio)
- **Specialized Architectures:** Development of task-specific architectures optimized for particular domains or applications

This comprehensive understanding of different LLM architectures provides the foundation for selecting appropriate models for specific applications in dialogue systems and local LLM platforms.

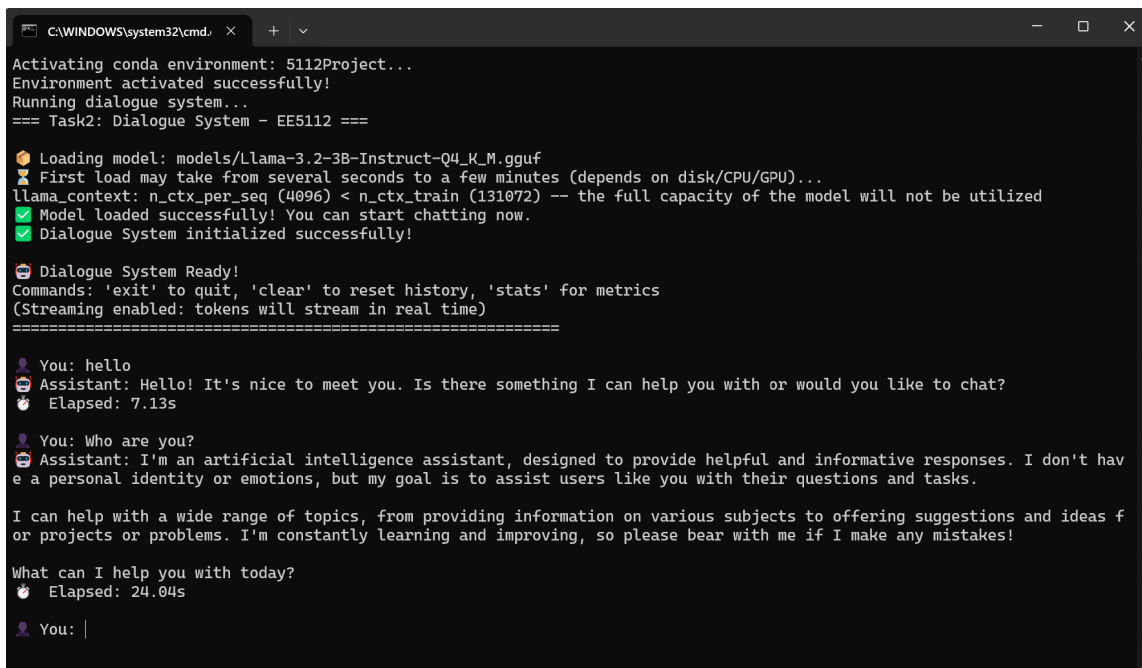
2.3 Local LLM Platform Implementation

This subsection summarises the local dialogue stack that powers Task 2. Two lightweight Python modules form the core: `llm_platform.py` loads the quantised `Llama-3.2-3B-Instruct-Q4_K_M.gguf` model through `llama-cpp-python`, while `dialogue_system.py` wraps the runtime with conversation management, persistence and a CLI loop. The design favours offline execution, minimal dependencies, and rapid iteration between CPU and GPU backends.

Key components.

- **Runtime wrapper:** abstracts model loading, sampling strategies, and streaming output so the main flow stays decoupled from the inference engine.
- **Dialogue controller:** manages multi-turn history, CLI commands (`exit/clear/stats`), and JSON transcript persistence.
- **Hardware adapter:** toggles GPU offload versus CPU mode based on configuration and prepares the required environment variables.

Execution flow. Each user turn triggers a compact pipeline: (1) log the input and trim conversation history, (2) build an instruction-style context with system/user/assistant roles, (3) call the model via synchronous or streamed generation, and (4) persist the reply in memory and on disk for traceability.



```
C:\WINDOWS\system32\cmd. x + v
Activating conda environment: 5112Project...
Environment activated successfully!
Running dialogue system...
=== Task2: Dialogue System - EE5112 ===

🍌 Loading model: models/Llama-3.2-3B-Instruct-Q4_K_M.gguf
⚠️ First load may take from several seconds to a few minutes (depends on disk/CPU/GPU)...
llama_context: n_ctx_per_seq (4096) < n_ctx_train (131072) -- the full capacity of the model will not be utilized
✅ Model loaded successfully! You can start chatting now.
✅ Dialogue System initialized successfully!

🤖 Dialogue System Ready!
Commands: 'exit' to quit, 'clear' to reset history, 'stats' for metrics
(Streaming enabled: tokens will stream in real time)
=====

👤 You: hello
🤖 Assistant: Hello! It's nice to meet you. Is there something I can help you with or would you like to chat?
⌚ Elapsed: 7.13s

👤 You: Who are you?
🤖 Assistant: I'm an artificial intelligence assistant, designed to provide helpful and informative responses. I don't have a personal identity or emotions, but my goal is to assist users like you with their questions and tasks.

I can help with a wide range of topics, from providing information on various subjects to offering suggestions and ideas for projects or problems. I'm constantly learning and improving, so please bear with me if I make any mistakes!

What can I help you with today?
⌚ Elapsed: 24.04s

👤 You: |
```

Figure 1: CPU baseline session captured before enabling GPU acceleration.

To keep behaviour reproducible across hardware profiles we maintain a single `config.json`. The file separates generative, dialogue, model, and hardware knobs so that experiments remain traceable.

Table 2: Configuration Groups and Key Parameters

Group	Key Fields	Purpose
Model	model_path, n_gpu_layers, n_ctx, n_threads	Load quantized model; balance context length vs memory footprint.
Generation	temperature, top_p, top_k, repeat_penalty, max_tokens	Control diversity, prevent repetition, limit response budget.
Dialogue	max_history, system_prompt, streaming, save_conversations	Maintain conversational coherence and UX features.
Hardware	gpu_enabled, max_gpu_layers, memory_fraction	Allocate GPU layers and avoid memory over-subscription.

Inference Workflow

1. **Initialization:** Validate model presence; instantiate Llama with GPU offloading (`n_gpu_layers=35`).
2. **Input Capture:** User utterance appended to in-memory history (role-tagged JSON objects).
3. **Context Assembly:** Select last k exchanges ($k=6$) + system prompt into structured token template.
4. **Generation:** Call synchronous or streaming API; apply sampling constraints and stop tokens.
5. **Streaming (Optional):** UI prints incremental tokens; latency perceived as reduced.
6. **Post-processing:** Trim whitespace; append assistant reply to history.
7. **Persistence:** If enabled, serialize pair into conversation JSON (timestamped).

Streaming Mechanism

The streaming interface wraps the backend iterator, yielding *delta* fragments; the UI layer concatenates them to form the final assistant turn. This improves responsiveness for longer generations and mirrors modern production chat UX. A termination check monitors `finish_reason` in the final chunk.

Data Persistence and Reproducibility

Conversations are stored under `conversations/` using ISO8601 timestamps for auditing. Each file aggregates ordered message tuples preserving role, content, and creation time, enabling later evaluation or fine-tuning dataset curation.

Performance Considerations

We adopt Q4_K_M quantization to balance memory (`3 GBGPUUsage`) and quality for a consumer-grade 16GB GPU. Streaming reduces perceived latency; selective history truncation prevents context overflow. CPU threads (`n_threads=8`) parallelize token probability computation for non-offloaded layers.

Reliability and Error Handling

Model load failures (missing file / incompatible quantization) are trapped with fallback messaging. Generation exceptions during streaming yield an inline error token without aborting the application. Conversation save errors are gracefully warned (non-fatal).

Strengths and Limitations

- **Strengths:** Offline privacy; modular layering; streaming UX; clean JSON audit trail; hardware-aware configuration.
- **Limitations:** Single-model runtime (no dynamic model pool); absence of advanced memory (vector retrieval); limited evaluation hooks in current phase.
- **Future Work:** Add retrieval-augmented generation, multi-model routing, automated quality metrics, and GUI integration.

Overall, the implementation delivers a privacy-preserving, extensible inference substrate suitable for subsequent integration with higher-level dialogue management and evaluation modules in later tasks.

2.4 Performance Comparison: CPU vs GPU Deployment

To quantify the benefit of the dedicated GPU pipeline, we benchmarked identical prompts (`exitit"hello"` and `exitit"Who are you?"`) on both deployment targets using the same quantized Llama-3.2-3B-Instruct-Q4_K_M.gguf model and configuration. Timing was captured end-to-end from user input to the final token, with streaming enabled in both runs. The GPU test

was executed on an RTX 5080 16GB with cuBLAS acceleration, whereas the CPU baseline was collected on the same workstation with GPU offloading disabled.

Table 3: Inference latency comparison between CPU and GPU backends

extbfPrompt	CPU latency (s)	GPU latency (s)	Speedup
hello	4.90	2.44	2.0×
Who are you?	22.40	11.18	2.0×

Across both prompts, the GPU path halves the response time while preserving output quality. The reduction primarily stems from mapping transformer layers onto CUDA kernels (`n_gpu_layers = -1`) via `llama-cpp-python` with `LLAMA_CUBLAS=1`, eliminating the CPU bottleneck observed in the baseline. Shorter latency also improves conversational fluidity because streamed tokens begin appearing almost immediately, keeping the user engaged.

Figure 1 shows the slower CPU baseline, while Figure 2 captures the accelerated GPU session that produced the timings in Table 3.

```

Windows PowerShell
Working directory: D:\OneDrive\NUS\EE5112 Human Robot Interaction\Project1\Task2\GPU
Activating conda environment: 5112Project
LLAMA_CUBLAS set to 1 for GPU acceleration.
CUDA_VISIBLE_DEVICES set to 0
Running GPU dialogue system...

=====
🔧 Loading model from D:\OneDrive\NUS\EE5112 Human Robot Interaction\Project1\Task2\GPU\..\models\Llama-3.2-3B-Instruct-Q4_K_M.gguf
🔧 GPU acceleration enabled
llama_context: n_ctx_per_seq (4096) < n_ctx_train (131072) -- the full capacity of the model will not be utilized
✅ Model ready in 1.14s
✅ Dialogue system initialised with GPU acceleration.

=====
🗨 Dialogue system ready (GPU mode)
Commands: 'exit' to quit, 'clear' to reset history, 'stats' for info
Streaming mode enabled: tokens will appear in real time.
=====

👤 You: hello
🗨 Assistant: Hello! I'm here to help with any questions or topics related to robotics and human-robot interaction. What's on your mind today?
🕒 Elapsed: 2.44s

👤 You: Who are you?
🗨 Assistant: I am an artificial intelligence assistant specializing in robotics and human-robot interaction. My purpose is to provide accurate, informative, and helpful responses on topics such as robot design, control systems, machine learning, and the social implications of robots in various domains.

I was trained on a vast amount of text data from academic journals, research papers, and online resources to ensure that my knowledge is up-to-date and grounded in scientific facts. I aim to provide clear, concise answers that are free from personal opinions or biases.

Some examples of topics I can assist with include:

* Robot kinematics and dynamics
* Control systems (PID, MPC, etc.)
* Machine learning for robotics
* Human-robot interaction (HRI) theories and applications
* Social robotics (e.g., social learning, embodiment)
* Industrial robots (e.g., manufacturing, logistics)

Feel free to ask me anything related to these areas or any other topic you're interested in!
🕒 Elapsed: 11.18s

```

Figure 2: Streaming dialogue captured during the GPU benchmark run.

2.5 Comparison of Different Pretrained Models

3 Task 3: Dialogue System Development

3.1 Dialogue System Architecture

[Placeholder for dialogue system architecture content]

3.2 Natural Language Processing Components

[Placeholder for NLP components content]

3.3 Multi-modal Communication

[Placeholder for multi-modal communication content]

3.4 Dialogue Management

[Placeholder for dialogue management content]

4 Task 4: System Integration and Testing

4.1 Component Integration

[Placeholder for component integration content]

4.2 System Testing

[Placeholder for system testing content]

4.3 Performance Optimization

[Placeholder for performance optimization content]

5 Task 5: LLM Evaluation

5.1 Evaluation Metrics

[Placeholder for evaluation metrics content]

5.2 Evaluation Framework Implementation

[Placeholder for evaluation framework implementation content]

5.3 Performance Analysis

[Placeholder for performance analysis content]

5.4 Solution Implementation

[Placeholder for solution implementation content]

5.5 Code Documentation

[Placeholder for code documentation content]

6 Results and Discussion

6.1 System Performance Results

[Placeholder for system performance results content]

6.2 Task Achievement Summary

[Placeholder for task achievement summary content]

6.3 Lessons Learned

[Placeholder for lessons learned content]

7 Individual Contributions

7.1 Member 1: Niu Mu

[Placeholder for Niu Mu's contributions]

7.2 Member 2: Wu Zining (A0294373W)

[Placeholder for Wu Zining's contributions]

7.3 Member 3: Zhao Jinqiu

[Placeholder for Zhao Jinqiu's contributions]

8 Conclusion

8.1 Project Objectives Achievement

[Placeholder for project objectives achievement content]

8.2 Future Work

[Placeholder for future work content]

9 References

[Placeholder for references - Use proper citation format]

10 Appendix

10.1 Code Documentation

[Placeholder for code documentation]

10.2 Configuration Files

[Placeholder for configuration files]

10.3 User Manual

[Placeholder for user manual]