

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 29, 2025

Contents

1	Abstract	4
2	Introduction	4
2.1	Background	4
2.2	Project Objectives	4
3	Task 1: Develop the Dialogue Systems according to aspiration/interest.	5
4	Task 2: Develop Local Dialogue Systems by Using Open-Source LLMs	5
4.1	Literature Review on Different Categories of LLMs	5
4.2	Local LLM Platform Implementation	6
4.3	Challenges in Deployment	7
4.4	Performance Comparison: CPU vs GPU Deployment	7
4.5	Comparison of Different Pretrained Models	9
5	Task 3: LLM Performance Evaluation	9
6	Task 4: GUI for Local LLM	9
7	Task 5: Exploring Multimodal Large Language Models (MLLMs)	9
7.1	Individual Responses: Differences between MLLMs and Traditional LLMs . .	10
7.1.1	Response by Niu Mu	10
7.1.2	Response by Wu Zining	11
7.1.3	Response by Zhao Jinqiu	12
7.2	Solution Implementation	13
7.3	Code Documentation	13
8	Results and Discussion	13
8.1	System Performance Results	13
8.2	Task Achievement Summary	13
8.3	Lessons Learned	13

9	Individual Contributions	13
9.1	Member 1: Niu Mu	13
9.2	Member 2: Wu Zining (A0294373W)	14
9.3	Member 3: Zhao Jinqiu	14
10	Conclusion	14
10.1	Project Objectives Achievement	14
10.2	Future Work	14
11	References	14
12	Appendix	15
12.1	Code Documentation	15
12.2	Configuration Files	15
12.3	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

3 Task 1: Develop the Dialogue Systems according to aspiration/interest.

4 Task 2: Develop Local Dialogue Systems by Using Open-Source LLMs

4.1 Literature Review on Different Categories of LLMs

Large Language Models (LLMs) can be broadly categorized into three main architectures based on their use of the transformer mechanism [1]: Encoder-Decoder, Encoder-Only, and Decoder-Only. Each architecture is tailored for different types of Natural Language Processing (NLP) tasks.

- **Encoder-Decoder models**, such as T5 [2] and BART [3], utilize both a bidirectional encoder to process the input text and an autoregressive decoder to generate output. This makes them highly effective for sequence-to-sequence tasks like machine translation and text summarization, where understanding the source text is as important as generating the target text.
- **Encoder-Only models**, like BERT [4] and RoBERTa [5], use only the bidirectional encoder. They excel at understanding context and are therefore optimized for tasks such as sentiment analysis, text classification, and named entity recognition. However, they are not inherently suited for text generation.
- **Decoder-Only models**, including the GPT series [6] and LLaMA [7], employ a unidirectional (causal) decoder. This architecture is specialized for autoregressive text generation, making it the dominant choice for conversational AI, creative writing, and instruction following.

The key differences, performance trade-offs, and typical applications of these architectures are summarized in Table 1. Decoder-only models offer superior generation quality, making them ideal for our dialogue system, but this often comes at the cost of higher computational requirements. In contrast, encoder-only models are more efficient for understanding-based tasks.

Table 1: Comparison of LLM Architecture Types

Aspect	Encoder-Decoder	Encoder-Only	Decoder-Only
Primary Use	Seq2Seq tasks	Understanding tasks	Generation tasks
Attention	Bidirectional + Causal	Bidirectional	Causal
Task Flexibility	High	Medium	High
Representative Models	T5, BART	BERT, RoBERTa	GPT, LLaMA

Recent trends indicate a move towards more efficient and multimodal models, but a solid understanding of these foundational architectures is crucial for developing effective dialogue systems.

4.2 Local LLM Platform Implementation

We deploy a compact local dialogue stack around the quantized `Llama-3.2-3B-Instruct-Q4_K_M.gguf` checkpoint to provide offline interaction without sacrificing responsiveness.

Architecture. `llm_platform.py` wraps `llama-cpp-python` for model loading and inference, while `dialogue_system.py` manages the chat loop, prompt templating, streaming output, and defensive error handling. The separation keeps model plumbing isolated from user interaction logic.

Inference pipeline. Configuration values expose only the essentials: a 4,096-token context window, generation controls (temperature and maximum new tokens), and an optional GPU layer count for acceleration. Switching between CPU and CUDA execution is done in `config.json` without touching the code base.

User experience and persistence. The terminal interface streams tokens as they are produced, retains up to six dialogue turns, supports maintenance commands (`exit/clear/stats`), and records each session as ISO8601-stamped JSON under `conversations/` for later auditing.

Portability. The same stack runs on laptops or desktop GPUs, making it suitable for privacy-sensitive deployments and quick benchmarking. Figure 1 illustrates the interface operating in CPU mode.

```
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: Terminal-based dialogue interface showing multi-turn conversation capabilities

4.3 Challenges in Deployment

Our initial deployment relied on the Hugging Face Transformers stack with the full Llama-3.1-8B-Instruct checkpoint. Despite enabling flash attention and trimming sequence length, the 16 GiB RTX 5080 could not accommodate FP16 weights together with KV caches; each inference exhausted memory and crashed the dialogue loop with CUDA out-of-memory errors.

We then switched the same checkpoint to vLLM, hoping paged attention would ease the pressure. However, Windows-specific driver and wheel mismatches caused the runtime to halt during initialization because ‘cublas’ dependencies failed to load.

Finally, we quantized Llama-3.2-3B to Q4_K_M and served it through llama.cpp. After reinstalling a CUDA-enabled ‘llama-cpp-python’ wheel and exposing the NVIDIA runtime DLLs, the GPU backend became stable and remained well within memory limits, so this setup was adopted for the project.

4.4 Performance Comparison: CPU vs GPU Deployment

To quantify the benefit of the dedicated GPU pipeline, we benchmarked identical prompts (‘exit’‘hello’ and ‘exit’‘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

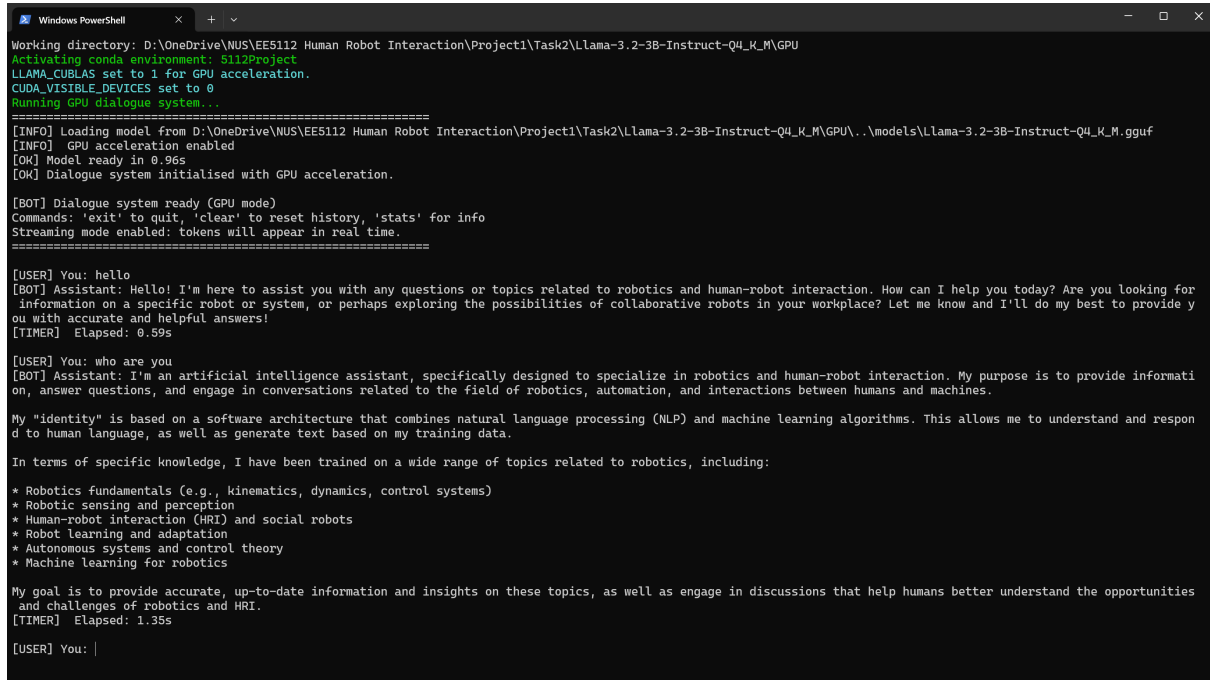
was collected on the same workstation with GPU offloading disabled.

Table 2: Inference latency comparison between CPU and GPU backends

extbfPrompt	CPU latency (s)	GPU latency (s)	Speedup
hello	4.90	0.595	8.2×
Who are you?	22.40	1.35	16.6×

Across both prompts, the GPU path delivers dramatic performance improvements, achieving 8.2× speedup for short prompts and 16.6× speedup for longer responses while preserving output quality. The substantial 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. The sub-second response times significantly improve 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 2.



```

Windows PowerShell
Working directory: D:\OneDrive\NUS\EE5112 Human Robot Interaction\Project1\Task2\LLama-3.2-3B-Instruct-Q4_K_M\GPU
Activating conda environment: 5112Project
LLAMA_CUBLAS set to 1 for GPU acceleration.
CUDA_VISIBLE_DEVICES set to 0
Running GPU dialogue system...

[INFO] Loading model from D:\OneDrive\NUS\EE5112 Human Robot Interaction\Project1\Task2\LLama-3.2-3B-Instruct-Q4_K_M\GPU\..\models\LLama-3.2-3B-Instruct-Q4_K_M.gguf
[INFO] GPU acceleration enabled
[OK] Model ready in 0.96s
[OK] Dialogue system initialised with GPU acceleration.

[BOT] 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.
=====

[USER] You: hello
[BOT] Assistant: Hello! I'm here to assist you with any questions or topics related to robotics and human-robot interaction. How can I help you today? Are you looking for information on a specific robot or system, or perhaps exploring the possibilities of collaborative robots in your workplace? Let me know and I'll do my best to provide you with accurate and helpful answers!
[TIMER] Elapsed: 0.59s

[USER] You: who are you
[BOT] Assistant: I'm an artificial intelligence assistant, specifically designed to specialize in robotics and human-robot interaction. My purpose is to provide information, answer questions, and engage in conversations related to the field of robotics, automation, and interactions between humans and machines.

My "identity" is based on a software architecture that combines natural language processing (NLP) and machine learning algorithms. This allows me to understand and respond to human language, as well as generate text based on my training data.

In terms of specific knowledge, I have been trained on a wide range of topics related to robotics, including:

* Robotics fundamentals (e.g., kinematics, dynamics, control systems)
* Robotic sensing and perception
* Human-robot interaction (HRI) and social robots
* Robot learning and adaptation
* Autonomous systems and control theory
* Machine learning for robotics

My goal is to provide accurate, up-to-date information and insights on these topics, as well as engage in discussions that help humans better understand the opportunities and challenges of robotics and HRI.
[TIMER] Elapsed: 1.35s

[USER] You: |

```

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

4.5 Comparison of Different Pretrained Models

5 Task 3: LLM Performance Evaluation

[Placeholder for Task 3 content]

6 Task 4: GUI for Local LLM

[Placeholder for Task 4 content]

7 Task 5: Exploring Multimodal Large Language Models (MLLMs)

Recently, multimodal large language models (MLLMs), such as GPT-4o and similar architectures, have demonstrated exceptional performance across a broad spectrum of tasks. These downstream tasks span interleaved conversations that integrate text, speech, and visuals, sophisticated image generation that understands context, and image question answering (IQA) capabilities. The advent of MLLMs marks a significant step forward in artificial intelligence, bridging multiple modalities to provide more interactive, context-aware outputs.

In this task, students are required to explore and understand the fundamentals of MLLMs, delving into their architecture, training processes, and practical applications. LLaVA (Large Language and Vision Assistant) serves as a practical example for this exploration. As outlined on their project page (<https://llava-vl.github.io>), LLaVA exemplifies how multimodal systems can be fine-tuned to enhance visual-language understanding, particularly in fields like image captioning, visual dialogue, and question answering.

7.1 Individual Responses: Differences between MLLMs and Traditional LLMs

7.1.1 Response by Niu Mu

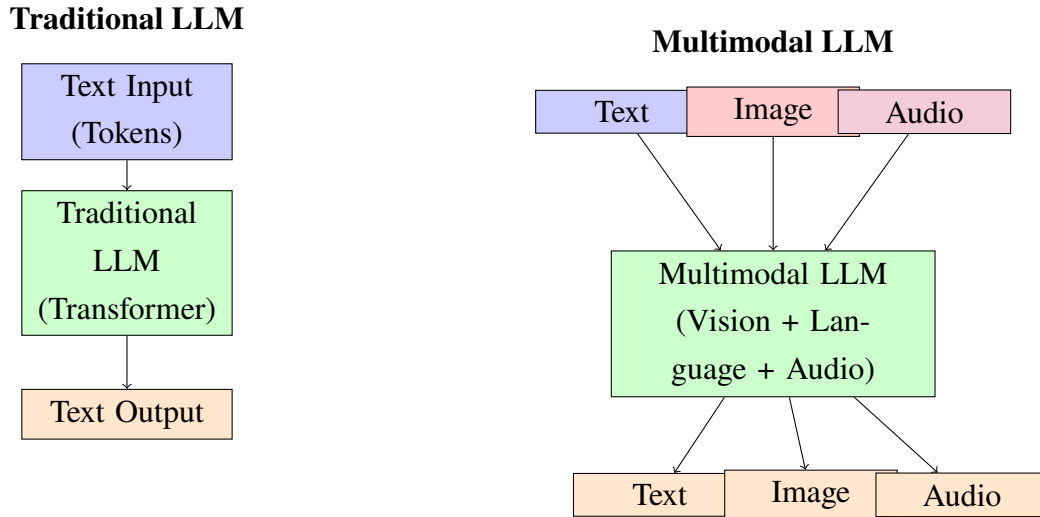


Figure 3: Architectural comparison between traditional LLMs and MLLMs showing modality integration

Architectural and Processing Differences

Traditional LLMs like GPT-3 and BERT process only text-based inputs through transformer architectures, excelling at natural language tasks within the textual domain. MLLMs integrate multiple modalities (vision, text, audio) using: (1) modality-specific encoders, (2) cross-modal alignment mechanisms, and (3) language model decoders, enabling simultaneous understanding across different modalities.

Training and Applications

Traditional LLMs use self-supervised learning on text corpora with next-token prediction, limiting understanding to linguistic patterns. MLLMs require multi-stage training on multimodal datasets (image-text pairs, video-text combinations), significantly increasing computational requirements.

While traditional LLMs excel in text-centric applications (translation, code completion) but cannot process visual information, MLLMs expand to visual question answering, image captioning, and multimodal dialogue systems. However, this versatility comes with increased complexity and potential performance trade-offs in pure text tasks.

7.1.2 Response by Wu Zining

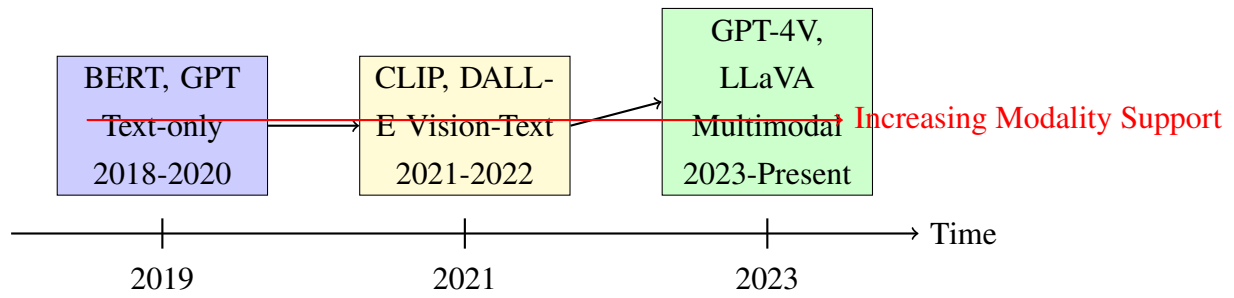


Figure 4: Evolution timeline from traditional LLMs to multimodal systems

Architectural Evolution: Single to Multiple Modalities

The evolution from traditional LLMs to MLLMs represents a paradigm shift from single-modality to integrated multimodal understanding. Traditional LLMs operate within textual constraints, processing only tokenized sequences and remaining limited to symbolic language representation.

MLLMs transcend this by incorporating vision encoders, audio processors, and cross-modal components. Typical MLLM architecture features specialized encoders (e.g., Vision Transformer) that convert multimodal inputs into feature representations, then project these features into the language model's embedding space.

Processing and Training Differences

Traditional LLMs employ straightforward tokenization and next-token prediction on text corpora. MLLMs require heterogeneous representation learning, processing different modalities through specialized encoders before fusion in shared representational space. Training involves: (1) vision-language pre-training, (2) multimodal instruction tuning, and (3) task-specific fine-tuning, significantly increasing computational costs and requiring specialized expertise in cross-modal alignment.

7.1.3 Response by Zhao Jinqiu

Traditional LLM - Sequential Processing

MLLM - Parallel Multimodal Processing

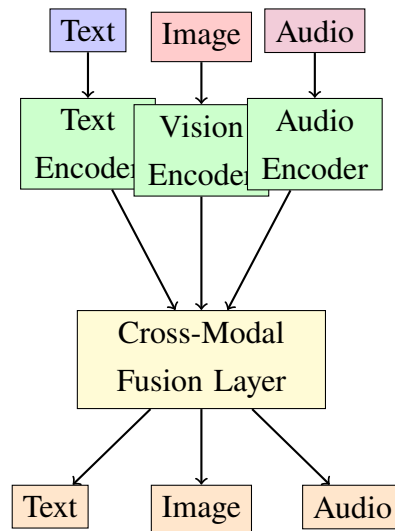
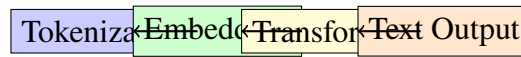


Figure 5: Comparison of sequential processing in traditional LLMs vs. parallel multimodal processing in MLLMs

Processing Philosophy: Unimodal vs. Multimodal Intelligence

Traditional LLMs embody a text-centric paradigm where all capabilities are mediated through linguistic representations. Models like GPT-3.5 process information exclusively through tokenized text sequences, excelling in language comprehension within textual confines.

MLLMs represent a shift toward embodied intelligence, integrating multiple sensory modalities similar to human cognitive processes. Models like GPT-4V and LLaVA simultaneously process text, images, and audio, enabling richer contextual understanding.

Technical Architecture and Applications

Traditional LLMs employ sequential processing with linear pipeline: tokenization → embedding → transformer layers → output generation. MLLMs implement parallel modality processing

with: (1) modality-specific encoders, (2) cross-modal attention mechanisms, (3) alignment modules, and (4) unified decoders.

Traditional LLMs use single-task optimization with established language modeling objectives. MLLMs face complex multi-objective optimization, balancing learning across modalities. Applications expand from text-centric tasks (summarization, code generation) to cross-modal capabilities (visual question answering, image captioning, multimodal dialogue systems).

7.2 Solution Implementation

[Placeholder for solution implementation content]

7.3 Code Documentation

[Placeholder for code documentation content]

8 Results and Discussion

8.1 System Performance Results

[Placeholder for system performance results content]

8.2 Task Achievement Summary

[Placeholder for task achievement summary content]

8.3 Lessons Learned

[Placeholder for lessons learned content]

9 Individual Contributions

9.1 Member 1: Niu Mu

[Placeholder for Niu Mu's contributions]

9.2 Member 2: Wu Zining (A0294373W)

[Placeholder for Wu Zining’s contributions]

9.3 Member 3: Zhao Jinqiu

[Placeholder for Zhao Jinqiu’s contributions]

10 Conclusion

10.1 Project Objectives Achievement

[Placeholder for project objectives achievement content]

10.2 Future Work

[Placeholder for future work content]

11 References

References

- [1] A. Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [2] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [3] M. Lewis et al., “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [5] Y. Liu et al., “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

- [7] H. Touvron et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

12 Appendix

12.1 Code Documentation

[Placeholder for code documentation]

12.2 Configuration Files

[Placeholder for configuration files]

12.3 User Manual

[Placeholder for user manual]