Tribhuvan University
Institute of Engineering
Purwanchal Campus
पूर्वाञ्चल क्याम्पस

Major Project Final Defense Presentation

A Comprehensive Study on Development of a Roman Nepali Chatbot for E-commerce Customer Care

Project Members:
Aakash Kumar Thakur (PUR077BCT003)
Kshitiz Gajurel (PUR077BCT042)
Manish Kathet (PUR077BCT044)
Manoj Kumar Baniya (PUR077BCT046)

Supervisor: Er. Pukar Karki

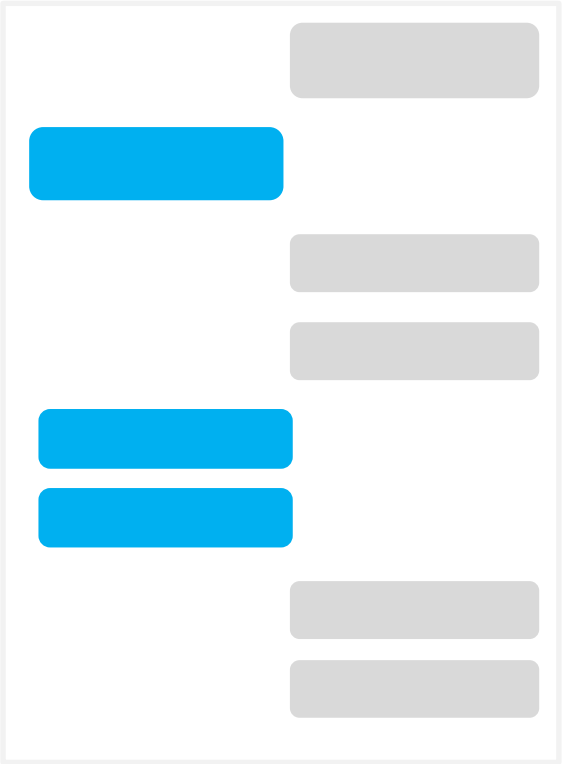
INTRODUCTION

Roman Nepali Chatbot

A chatbot designed to understand and respond to user queries in the Roman Nepali language.

Ecommerce Chatbot

An e-commerce customer bot is an automated tool designed to interact with customers on an e-commerce platform to simulate human-like conversations and assist users with various tasks.



PROBLEM STATEMENT

1. Manual Customer Support

2. Slow Response Time

3. Limited Language Support
(English/Nepali)

OBJECTIVES

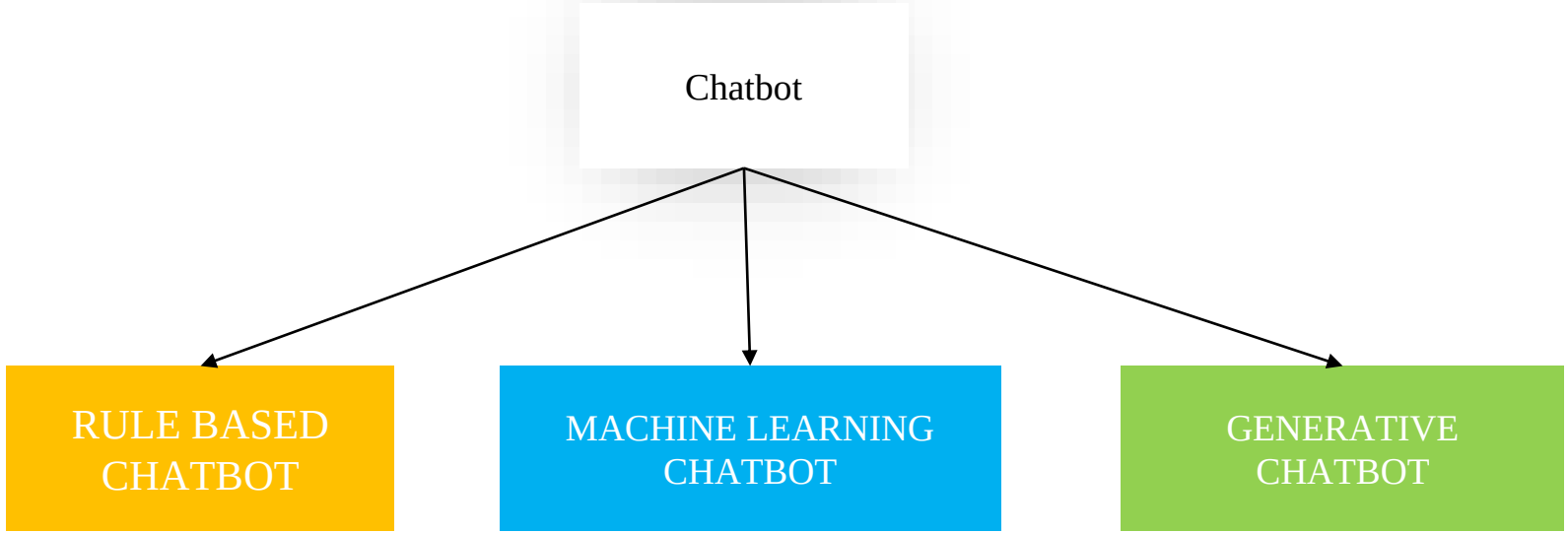
1. Romanized Nepali E-commerce Chatbot Development
2. Chatbot Architecture Analysis & Optimization
3. Enhanced Response Quality with Dynamic Retrieval

LITERATURE REVIEW

- **GUS (Genial Understander System) [1]** – A **frame-driven dialogue system (1977)** using **predefined templates** for structured conversations. It enabled **goal-directed interactions** and influenced modern **intent recognition**. Our chatbot applies **ML-based intent classification** for better **Romanized Nepali** support in e-commerce.

- **[2]** Optimizes **LLMs** for **low-resource languages** like Veitnamese. Enhances **chatbot accuracy** despite **limited data**. We apply this to improve **Romanized Nepali chatbot performance**, ensuring **better responses** for e-commerce customer care.

METHODOLOGY


Chatbot
RULE BASED CHATBOT
MACHINE LEARNING CHATBOT
GENERATIVE CHATBOT

PREPARING DATASET

STEPS ON PREPARING DATASET

- Synthetic Dataset Generation

  Ecommerce Q/A pair dataset (Roman Nepali)

- Data Cleaning and Filtering
- Labeling Intent and Entity Dataset
- Pre-Processing and Adding Prompt Template

Fine Tuning Dataset: 17886 (Q/A pairs)

We used **GPT-4o** and **DeepSeek AI** to generate Romanized Nepali dialogues through diverse prompts, ensuring natural chatbot conversations with a balanced mix of:

1. Common user intents *(browsing, ordering, tracking)* with suitable responses.
2. Varied tones *(formal, informal, friendly, professional).*
3. Multiple topics *(greetings, inquiries).*
4. To improve the diversity and quality of the dataset, we used various prompt engineering techniques *(Zero Shot, Few Shot and Dynamic Scenario Prompt)*
5. We further validate and filter our dataset final *(17886)*

Intent and Entity Dataset

- We aimed for only 4 different e-commerce intents *(product-inquiry, order-tracking, order-product, payment-method).*
- Manually Prepared dataset on each intents.
- Same dataset was used for training entity recognition model.

Query	Intent
ma saman order garna chahanxu	ORDER_PRODUCT
laptop ko price kati ho?	PRODUCT_INQUIRY
mero order ko status ke ho jankari dinu?	ORDER_TRACKING

Intent Classification: Dataset

WORD REPLACEMENT

chha:{cha,chhaa,xa,xha,xaa,x}

malaai:{malai,malaai,malae,molai,maalai,maalaai}

maile:{mailee,mailey,mailei,maileey}

k:{ke}

paryo:{paro,pariyo,pario,parryo,parryoo}

For Example:

Malae store location bare bataidinu paro.

Malai store location bareey bataidinu paryo.

Entity Recognition: Data Labeling

Labeling Entity using Custom Labeling App

Token	Store	ma	macbook	ko	laptop	xa	ki	xaina
Label	O	O	PRODUCT	O	CATEGORY	O	O	O

Token	malai	2	ta	IPhone	11	pro	order	garne
Label	O	QUANTITY	O	PRODUCT			O	O

Formatting Fine Tuning Dataset

Prompt Template for Fine Tuning Dataset

<start_of_turn>system *{SYSTEM_PROMPT}*<end_of_turn>

<start_of_turn>user *{USER_QUESTION} {CONTEXT}*<end_of_turn>

<start_of_turn>model *{MODEL_RESPONSE}*<EOS>

Example Training Dataset using Prompt Template

<start_of_turn>system *Timi auta e-commerce ko AI Assistant hou user le gareko question lai context herera ramro response deu.*<end_of_turn>

<start_of_turn>user *Samsung Galaxy ko price kati ho?*

Context: product_name Samsung Galaxy, price: 1,10,000, RAM: 16GB<end_of_turn>

<start_of_turn>model *Hajur Samsung Galaxy ko price Rs. 1,10,000 parxa.*<EOS>

Three Types of Chatbot

- Rule Based Chatbot *(Menu Chatbot)*
- Machine Learning Chatbot *(Dialogue System)*
- Generative Chat *(RAG Chatbot)*

Rule Based Chatbot

Rule Based Chatbot

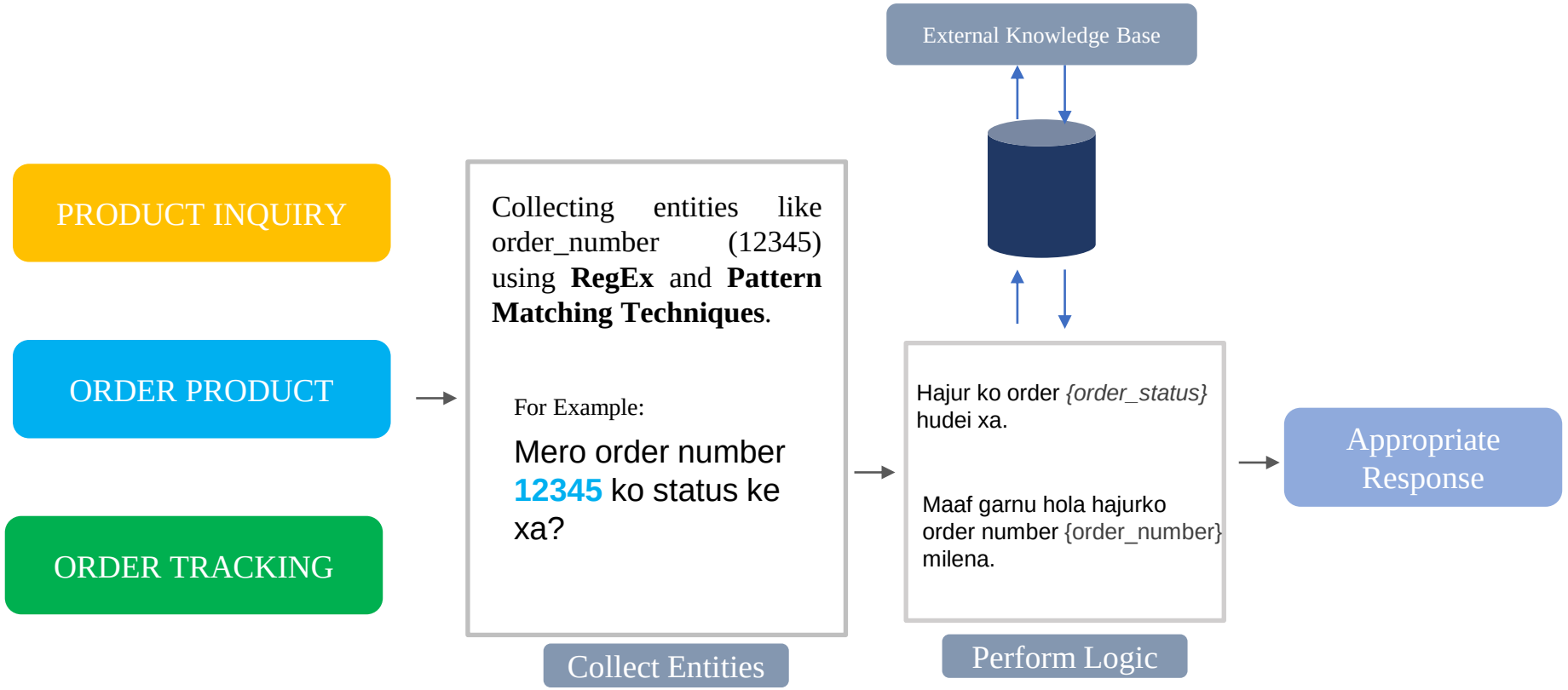
- The chatbot operates using predefined rules and structured menu-driven interactions to ensure accurate and efficient responses.
- It processes user queries based on structured dialogue states and retrieves relevant data using predefined logic and generates appropriate responses for user interactions.

PRODUCT INQUIRY

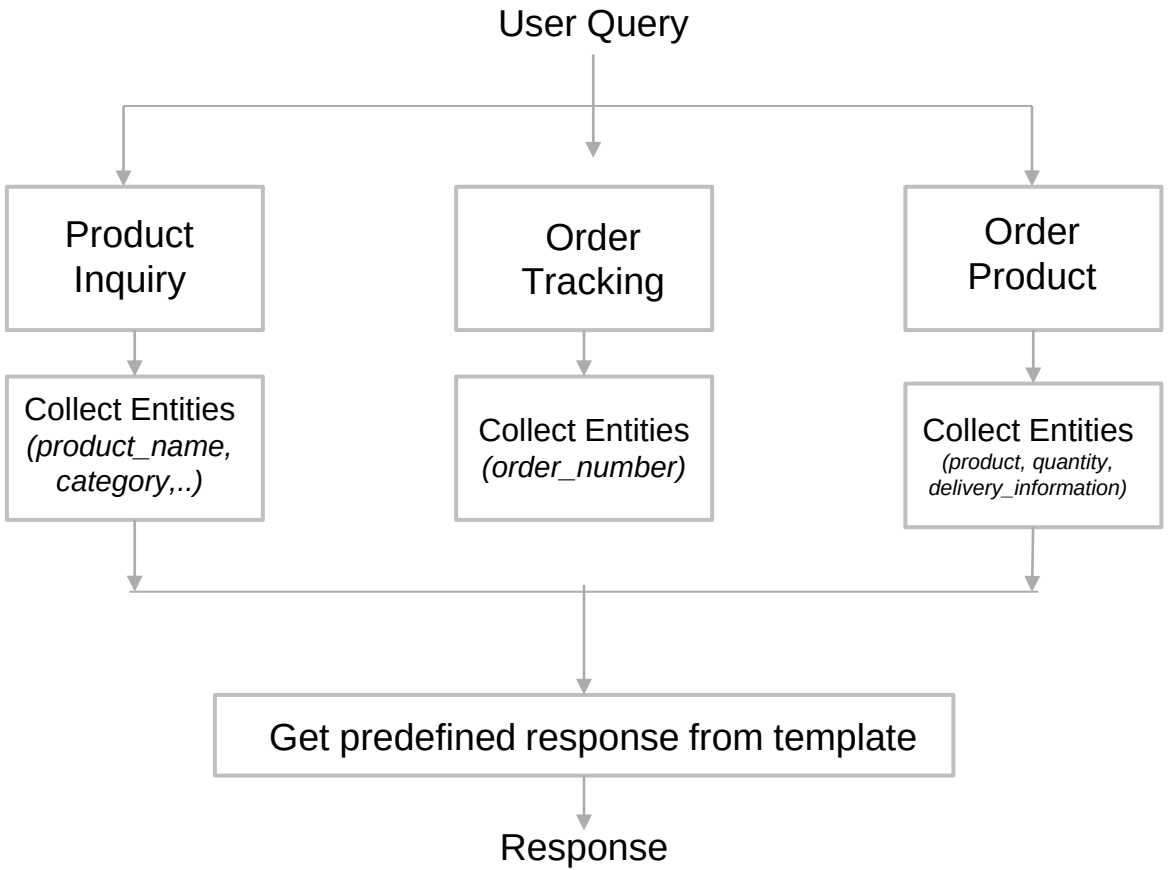
ORDER PRODUCT

ORDER TRACKING

Rule Based Chatbot



System Diagram: (Rule Based Chatbot)



ML Chatbot

ML Chatbot: Introduction

The chatbot is developed using a **machine learning-based approach** that integrates multiple natural language processing (NLP) techniques to ensure accurate intent classification, entity recognition, and dialogue management through slot filling.

INTENT CLASSIFICATION

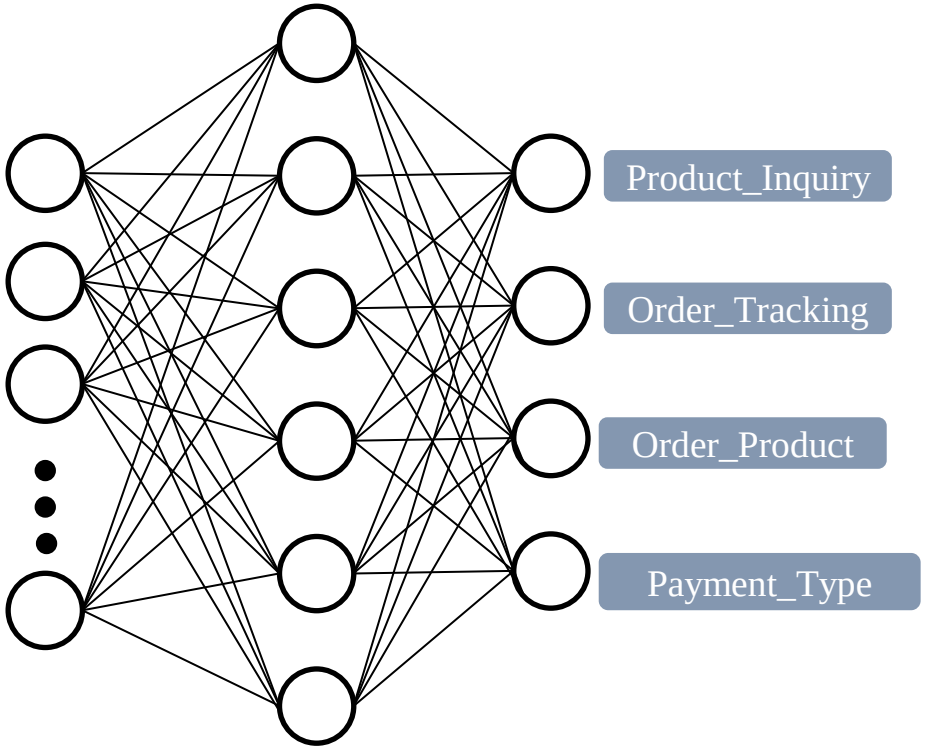
ENTITY RECOGNITION

ML Chatbot: Intent Classification

Query	Intent
ma saman order garna chahanxu	ORDER_PRODUCT
laptop ko price kati ho?	PRODUCT_INQUIRY
mero order ko status ke ho jankari dinu?	ORDER_TRACKING

Bag of Word
(Feature Vector)

[1 2 1 0 0 0 .. 0]
[1 0 1 0 0 0 .. 0]
[0 0 1 0 0 0 .. 0]



ML Chatbot: Entity Recognition

Conditional Random Field (CRF):

CRF is a machine learning model for sequence prediction, used in entity recognition. It considers word relationships and context for better accuracy.

Use in Entity Recognition:

CRF tags words as entities (e.g., names, locations) by analyzing surrounding words, ensuring meaningful label patterns (e.g., "New" → "York" as a location).

CRF Formula:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t}\sum_{k} \lambda_k f_k(y_t, y_{t-1}, x, t)\right)$$

x = Input sequence (words in a sentence)

y = Output sequence (entity labels)

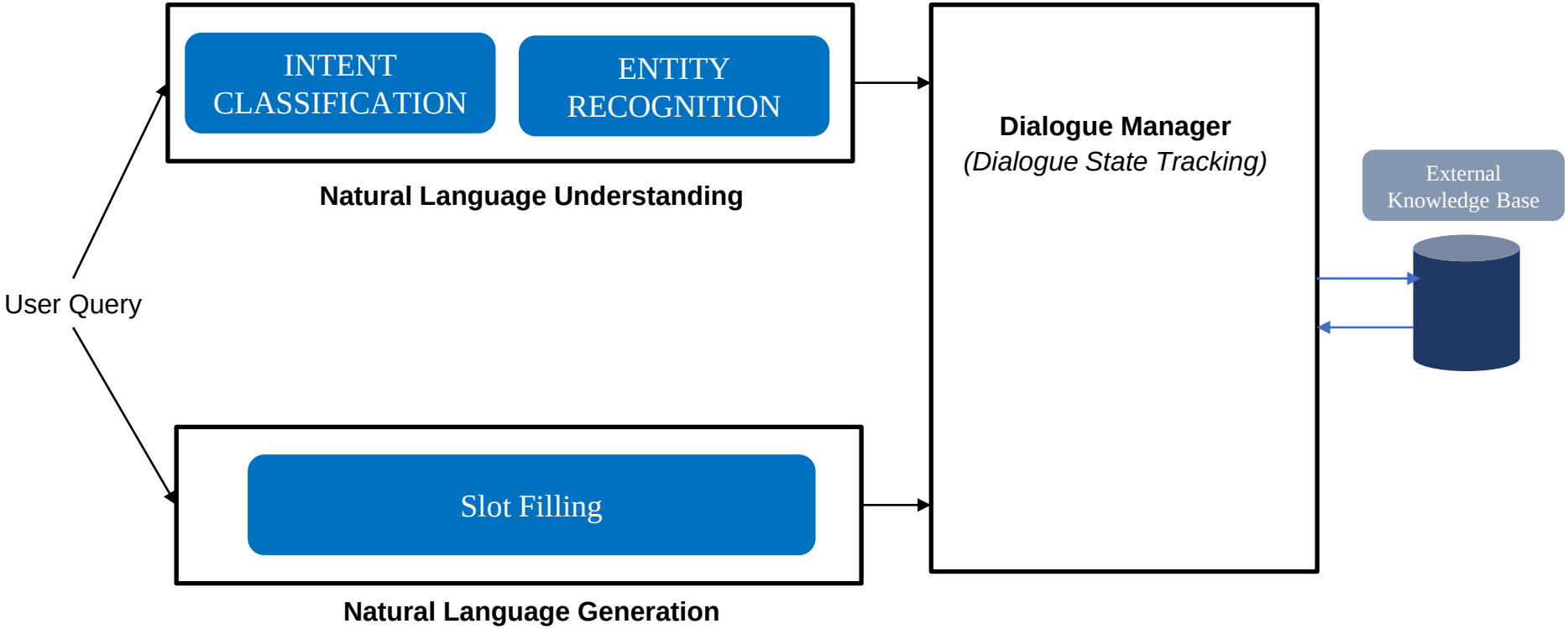
fk = Feature functions capturing word properties and dependencies

λk = Weights for feature functions

Z(x) = Normalization factor ensuring a valid probability distribution

Example:

"Buy a red Nike running shoe in size 10."


INTENT CLASSIFICATION
ENTITY RECOGNITION
Natural Language Understanding
User Query
Dialogue Manager
(Dialogue State Tracking)
External Knowledge Base
Slot Filling
Natural Language Generation

Generative Chatbot

Generative Chatbot: Introduction

- Uses **Deep Learning Techniques** to create responses from scratch so generative.

- Unlike **Rule-Based and Dialogue System Chatbot**, which select responses from a predefined set based on user input, generative chatbots generate new responses dynamically, tailoring them to the specific context and content of the conversation.

- Trained on large amount on text data.

- Much more complex than rule based and retrieval based approach.

RNN based
- LSTM (Long Short Term Memory)
- GRU (Gated Recurrent Unit)

Transformer based
- GPT (Open AI)
- Llama (Meta AI)
- **Gemma (Google Research)**

Fine Tuning: Gemma2 (9 Billion Parameter Model)

Gemma 2 9B

- **Model Architecture:** A **9-billion parameter decoder-only** transformer model.
- **Key Features:** Uses **optimized attention mechanisms** and **rotary positional embeddings** for better efficiency
- **Advantages:** Supports **fast inference, efficient fine tuning** and **domain adaptation**.

Why Gemma 2 9B?

- Its **pretrained model** understands **Romanized Nepali**, making **text tokenization easier**.

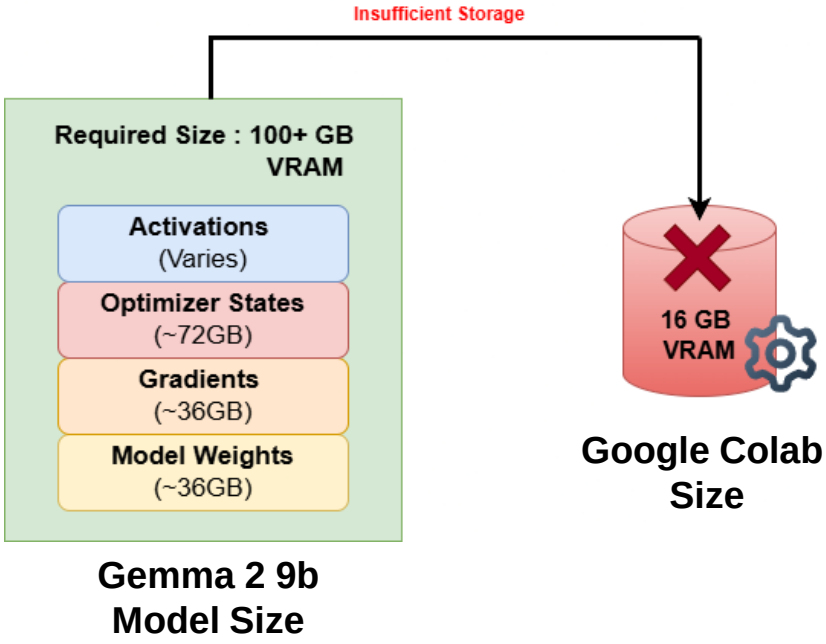
Training of Gemma 2 Model

- **17886** Instructional datasets used.
- Fine-tuned on **Google Colab** which a free **Tesla T4 GPU** (16GB VRAM).
- **Model Storage** on **Hugging Face** for easy access.
- Deployed using Google Colab with **API tunneling.**
- **Pinecone vector store** utilize for dynamic data retrieval.

Fine Tuning: Computation Required for Fine Tuning

- **Model Weights Storage**: Loads pre-trained parameters.
- **Gradient Storage**: Stores weight updates during training.
- **Optimizer States**: Maintains momentum and scaling factors.
- **Activations Memory**: Holds intermediate outputs for backpropagation.

Total Requirement: Exceeds standard GPU limits, making full fine-tuning infeasible.



SOLUTION:
Used PEFT techniques for memory-efficient fine-tuning.

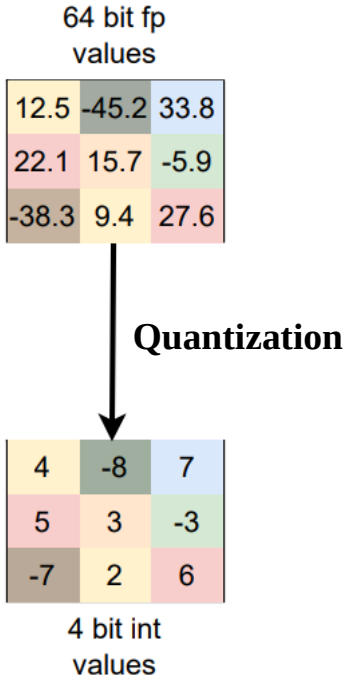
Fine Tuning: PEFT Technique

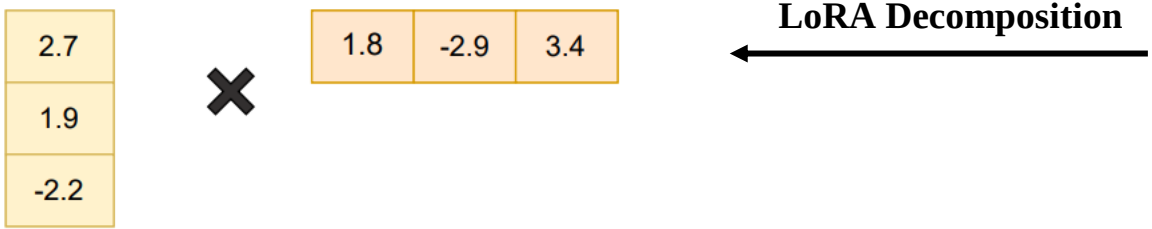
1. Quantization

- Reduces model weight precision from **64-bit/32-bit to 4-bit**, lowering memory use.

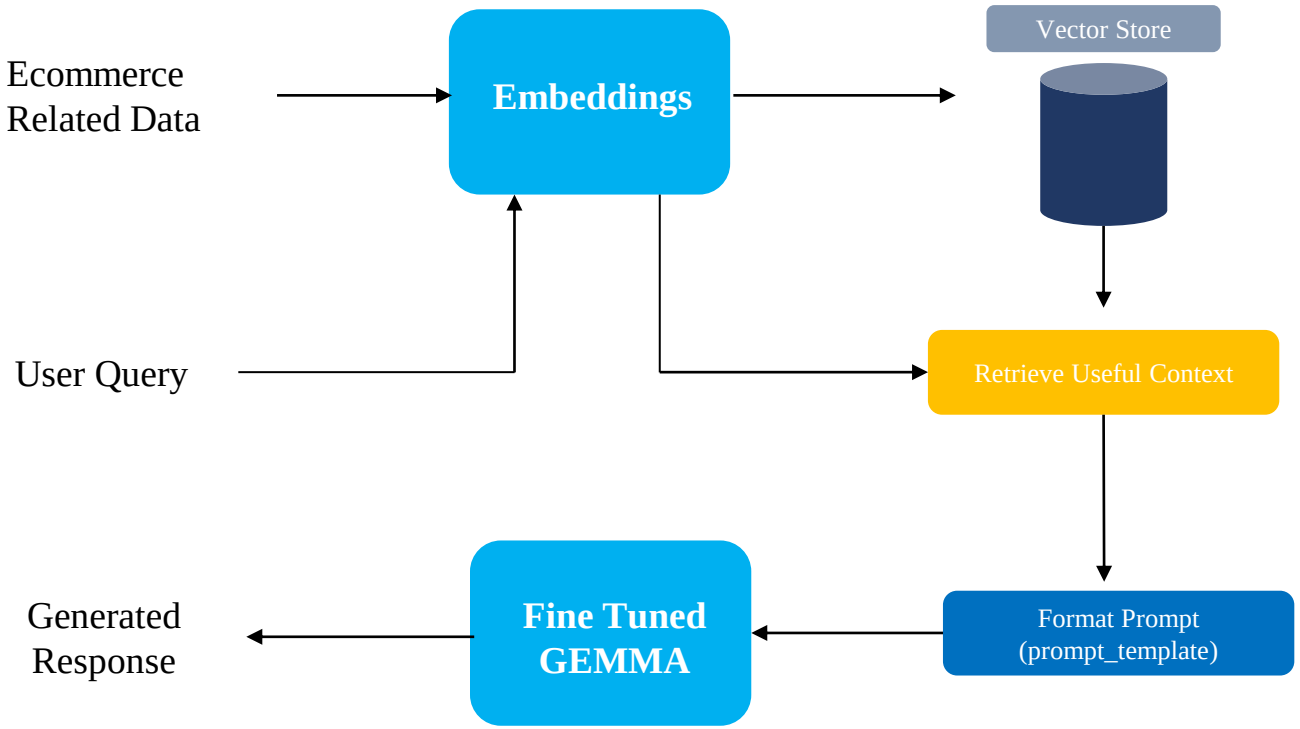
2. LoRA & QLoRA

- Freezes most model weights, trains only **small adapter layers** for efficiency.
- Applies **4-bit quantization** before using LoRA, further reducing memory needs.





System Diagram (RAG Chatbot)



Result and Discussion

Criteria	Rule Based Chatbot	ML Based Chatbot	LLM Based Chatbot
1. Query Adaptability	Predefined flows, struggles with deviation	Handles some variations, misclassifies unseen phrasing	Handles diverse queries, occasional hallucinations
2. Response Quality	Static responses, no personalization	Template-based, misclassifies tokens	Context-aware, BLEU score: **36.13**, human ratings: **8-10**, better entity recognition
3. Resource Efficiency	Low cost, limited functionality	Moderate cost, balanced performance	High cost, advanced capabilities

Challenges

1. Lack of Roman **Nepali Dataset** publicly available to experiment.

2. Fine-tuning large language models for Romanized Nepali remains challenging due to **limited pre-trained resources.**

3. Hardware **resource constraint** to train/fine tune large language model.

4. **Hallucinations** of LLM (for ecommerce customer care we need accurate responses).

Conclusion

- We explored chatbot techniques for e-commerce in Roman Nepali language focusing on rule-based, retrieval-based and generation-based approach.

- We Collected and prepared a diverse dataset to train machine learning model.

- Fine tuned small model [**gemma2 9B**] to learn the techniques for efficiently fine tuning on low resource.

References

[1] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, “Gus, a frame-driven dialog system,” *Artificial Intelligence*, vol. 8, no. 2, pp. 155–173, Apr. 1977. [Online]. Available: https://doi.org/10.1016/0004-3702(77)90018-2

[2] V.-T. Doan, Q.-T. Truong, D.-V. Nguyen, V.-T. Nguyen, and T.-N. N. Luu, “Efficient finetuning large language models for Vietnamese chatbot,” *arXiv preprint arXiv:2309.04646*, 2023. [Online]. Available: https://arxiv.org/abs/2309.04646

Thank You