

G31

Small Scale Financial Language Model

Group

Kasuni Hansachapa(E/19/131)

Supervisors

Dr.Asitha Bandaranayake

Prof.Roshan G.Ragel

Problem

Accuracy

Hallucinations

Resource Intensivity

High memory and computational
power consumption

Security

Confidential Data



Problem Solution

Hallucinations

Fine Tuned SLM with
RAG Architecture

Resource Intensitivity

QLoRA Optimization

Security

Local Server

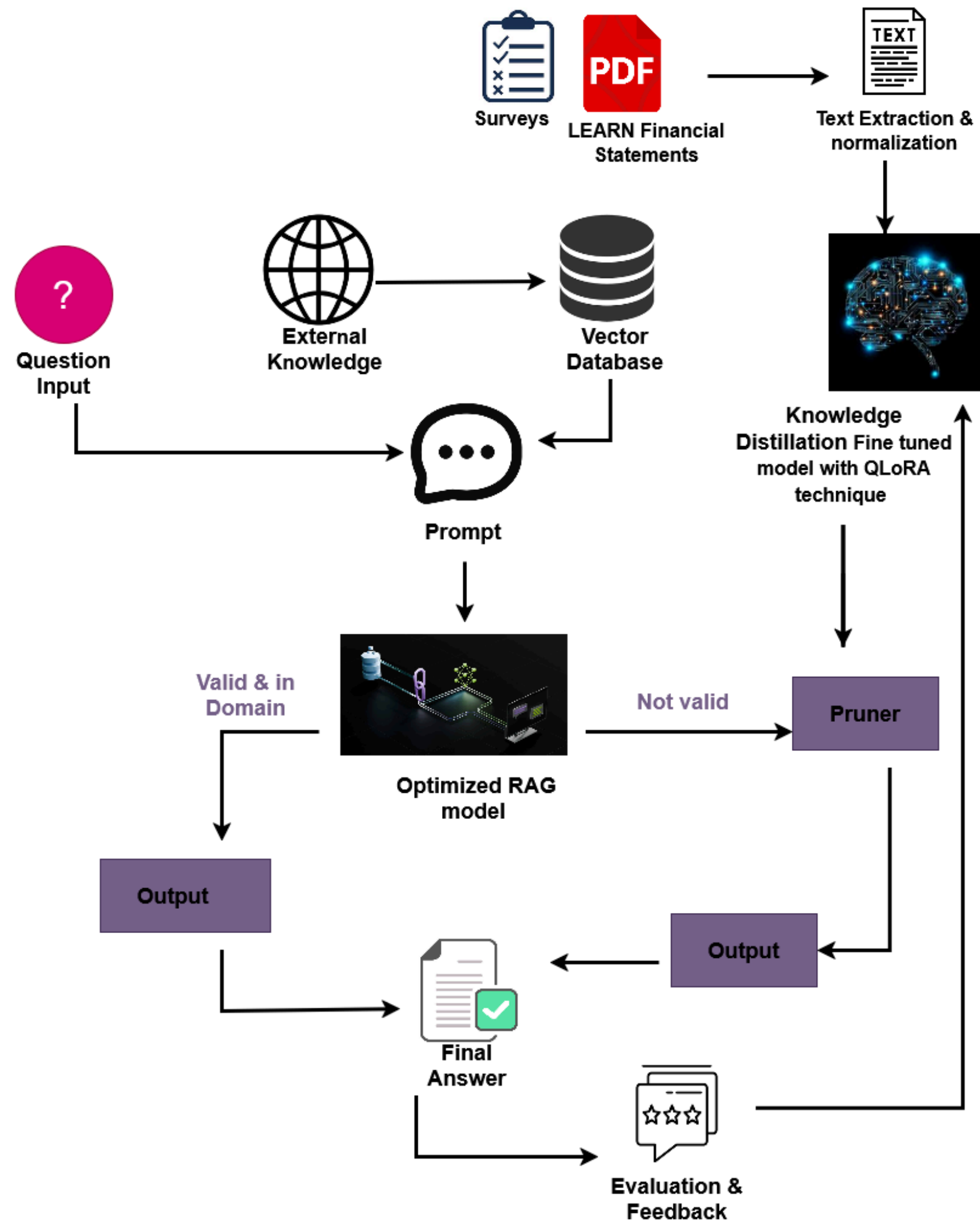


Proposed Solution

Developing a small language model fine-tuned on LEARN's financial dataset using QLoRA optimization and integrated into a Retrieval-Augmented Generation (RAG) pipeline

For Setiment Analysis, Question & Answering ,Text Summarization

Methodology



Implementation

Component	Technology
FrontEnd	React.js
BackEnd	FastAPI(Python)
Model Serving	HuggingFace/GitHub
Database	PostgreSQL(Structured data),MongoDB(Unstructured data),ChromaDB(VectorDB for RAG)
Authentication	Docker,nginx
CI/CD for model updates	Github Actions
Server & Infra	Ubuntu

Findings & Results

Data Set

**Auditor's reports & Financial Statements PDFs
for three financial years at LEARN**



- Auditor's report
- Statements of Financial Position
- Statement of Profit or Loss
- Statement of Changes in Equity
- Detailed notes on above metrics

Results & Findings

Model Selection

Specifications	FinGPT	FinBERT
Pre trained Dataset	Reddit, SEC Filings, Yahoo Finance, News APIs, Tweets	Financial PhraseBank (from Finland), SEC filings, Earnings Calls
GPU,memory usage for fine tuning	[0] NVIDIA A100-SXM4-40GB : 30°C, 0 % 3299 / 40960 MB CPU: 12.5 % Memory: 15 %	[0] NVIDIA A100-SXM4-40GB : 31°C, 0 % 3299 / 40960 MB CPU: 5.2 % Memory: 7.9 %
Training loss	0.453631	0.262100
Accuracy	0.74544	0.84433
Model availability	(GitHub + weights)	Hugging Face
Model Developers	AI4Finance Foundation	ProsusAI, HKUST

Challenges Encountered

Query: What is the financial position of LEARN as at 31 March 2022?

Answer: a corporation

Query: What are the total assets reported in 2022?

Answer: amounted to a total of amounted to a total of amounted to

Query: What is the total equity and liabilities as per the latest balance sheet?

Answer: a total of .2

Query: How has LEARN's financial position changed compared to the previous year?

Answer: a year

Query: What was the total income for LEARN in 2021 and 2022?

Answer: \$45

Query: What are the main sources of income for LEARN?

Answer: money

Query: What is the total expenditure for the year ended 31 March 2022?

Answer: £150,000

Query: What are the administrative expenses recorded in the financial year?

Answer: \$500,000

Query: What are the current assets listed in the balance sheet?

Answer: ad-hoc assets

Query: How much cash and cash equivalents does LEARN have?

Answer: \$

Query: What are the non-current liabilities for the year 2022?

Answer: a repurchase

Hallucinations from RAG



Using High-Quality Embeddings &
Chain of Thought

Demonstration

```
from transformers import BertTokenizer, BertForSequenceClassification, Trainer, TrainingArguments
from datasets import load_dataset

# Load FinBERT (you can use 'ProsusAI/finbert' or 'yiyanghkust/finbert-tone')
model_name = "yiyanghkust/finbert-tone"
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertForSequenceClassification.from_pretrained(model_name)
```

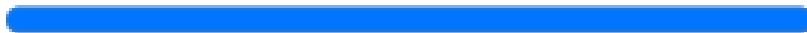
```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens).
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets
warnings.warn(
```

```
vocab.txt: 100%  226k/226k [00:00<00:00, 4.87MB/s]
```

```
config.json: 100%  533/533 [00:00<00:00, 52.1kB/s]
```

```
pytorch_model.bin: 100%  439M/439M [00:01<00:00, 261MB/s]
```

```
results = trainer.evaluate()
print(results)
```

```
 [122/122 00:02]
{'eval_loss': 0.856381893157959, 'eval_f1': 0.8436240365963721, 'eval_precision': 0.8442147825295664, 'eval_recall': 0.8443298969072165, 'eval_accuracy': 0.8443298969072165,
```

Demonstration

Fine Tuning FinBert

Map: 100% 4846/4846 [00:15<00:00, 349.88 ex

```
wandb: WARNING The `run_name` is currently set to the same value as `TrainingAr
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server locally: https:
wandb: You can find your API key in your browser here: https://wandb.ai/authori
wandb: Paste an API key from your profile and hit enter:wandb: WARNING If you'r
wandb: WARNING Consider setting the WANDB_API_KEY environment variable, or runr
wandb: No netrc file found, creating one.
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
wandb: Currently logged in as: hansachapakasuni (hansachapakasuni-university-of
Tracking run with wandb version 0.19.11
Run data is saved locally in /content/wandb/run-20250602_115519-ebi6x8jd
Syncing run ./finbert-financial to Weights & Biases \(docs\)
View project at https://wandb.ai/hansachapakasuni-university-of-peradeniya/huggingface
View run at https://wandb.ai/hansachapakasuni-university-of-peradeniya/huggingface/runs/ebi6x8j
```

[1303/1455 4:25:08 < 30:58, 0.08 it/s, Epoch 2.6]

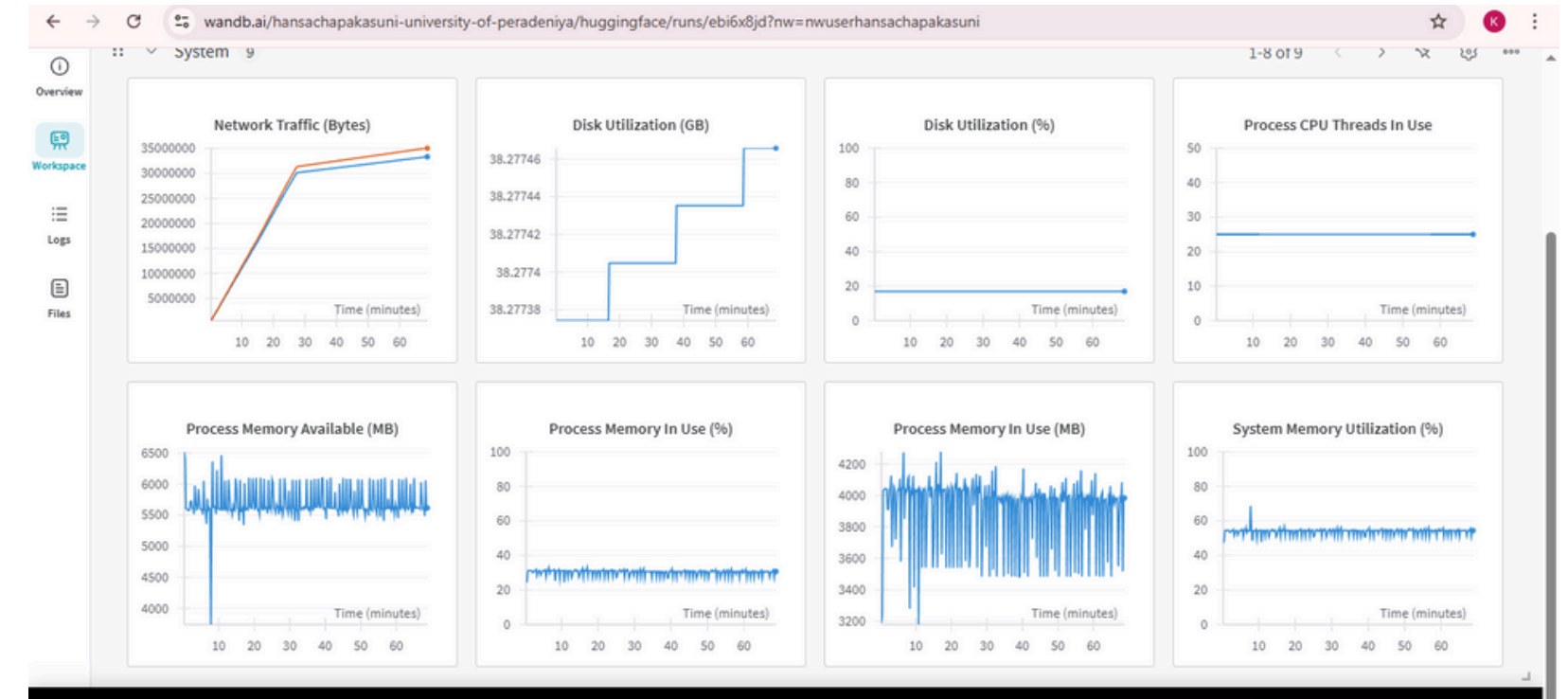
Step Training Loss

500	0.688300
1000	0.278400

[1455/1455 4:56:01, Epoch 3/3]

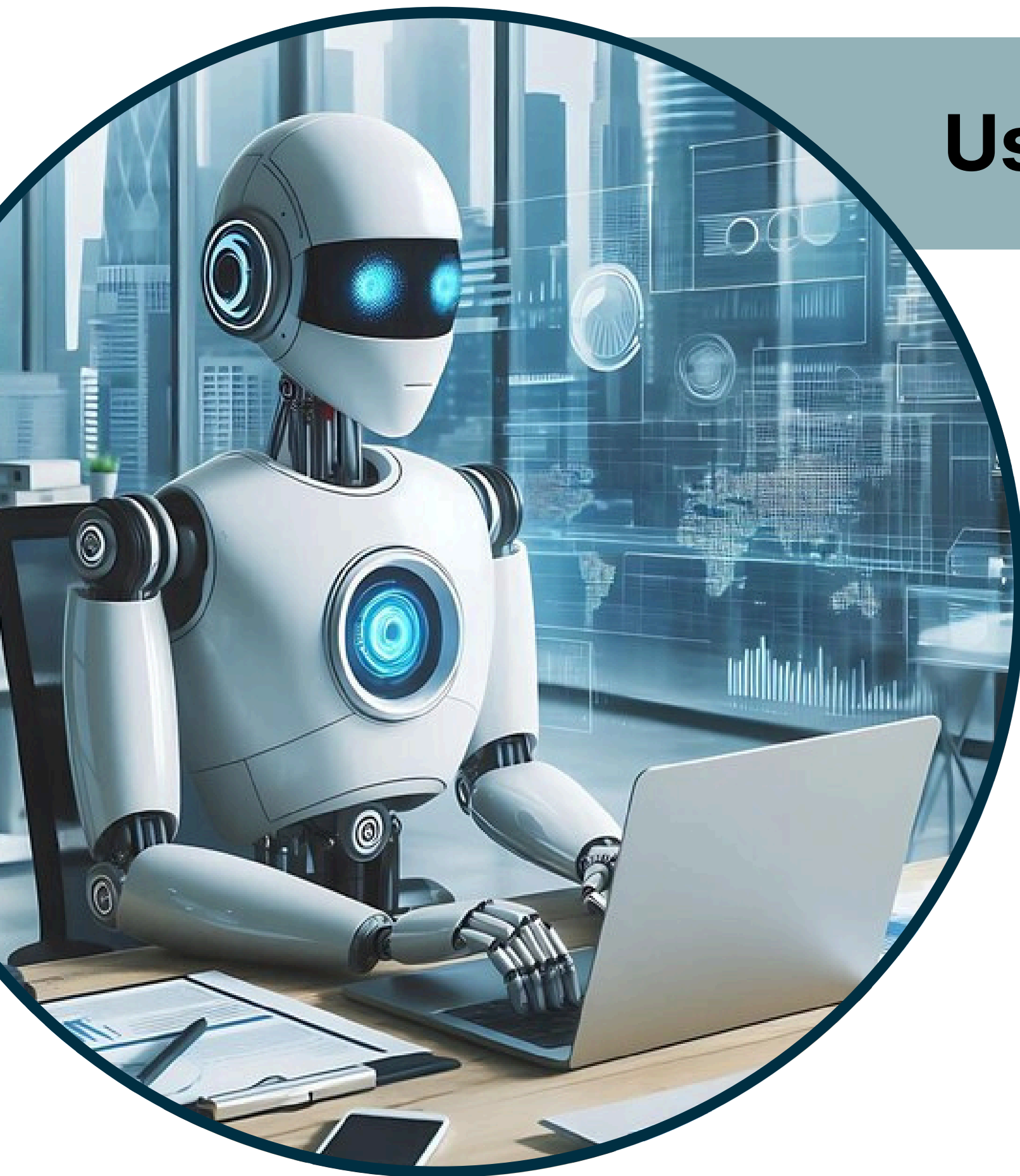
Step Training Loss

500	0.688300
1000	0.278400



TimeLine

Task/Month	Jan	Feb	Mar	Apr	May	June	July
Market Research							
Data Collection							
Model Selection							
UI Creation							
Testing & Evaluation							
Deployment & Review							
Thesis Writing							



Use of AI Tools

For Literature Review: Connected Papers, NotebookLM, Consensus, Grammarly

For Debugging: Github Copilot, Chatgpt

Use of AI Tools

Tool	Effectiveness	Bias & Fairness Issues
Connected Papers	Strong in mapping related research	Overemphasis on highly cited papers
NotebookLM	Effective for comprehension & synthesis	Data quality-dependent, limited domain
Consensus	Quick, evidence-based answers	Domain-specific bias
Grammarly	Enhances clarity and correctness	Biased toward Western norms
GitHub Copilot	Speeds up coding	May propagate insecure patterns
ChatGPT	Versatile and interactive	Can hallucinate, may give overconfident outputs

A circular graphic on the left side of the slide, featuring a low-angle view of several modern glass skyscrapers reaching towards a clear, light blue sky. The buildings are reflected in each other, creating a symmetrical, geometric pattern. The circular frame is partially overlaid by a dark blue horizontal band.

Thank You





Q&A

Project page & Github Repo

