Customized Arabic IT Translator

Project Report for IT PowerPoint Translation
Academic Version

1. Introduction

- This project implements an automated pipeline to convert English IT PowerPoint presentations (.pptx) into Arabic PDF reports. The system extracts textual content, translates it using a customized Large Language Model fine-tuned with LoRA, and preserves technical terminology and programming code.

2. Dataset Description

- The dataset consists of textual content dynamically extracted from PowerPoint slides. It includes English IT explanations, technical terms, acronyms, and code snippets. Text is extracted using Python libraries for automated preprocessing before translation.

3. Methodology

Workflow:

- Text Extraction: Extract slide content using python-pptx.

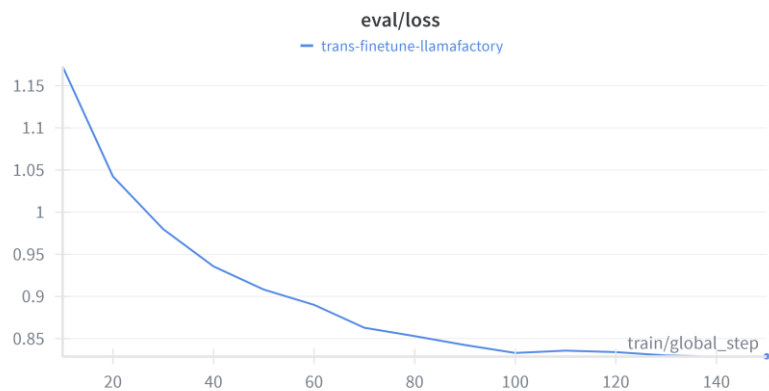- Translation: Custom LLM fine-tuned with LoRA preserves technical terms and code.

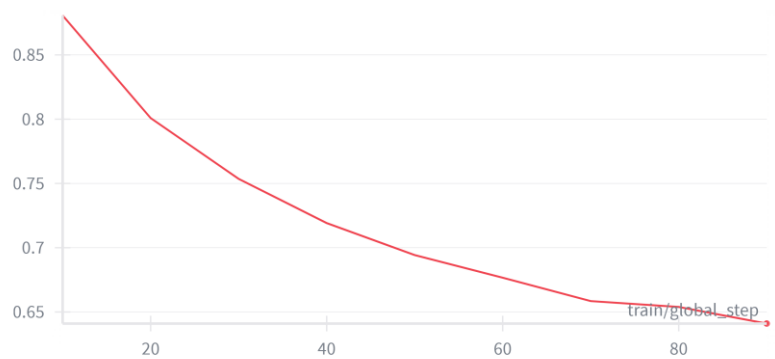- Structuring: Organized text into JSON/HTML format with glossary.

- PDF Generation: Convert structured content to professional PDF with RTL formatting.
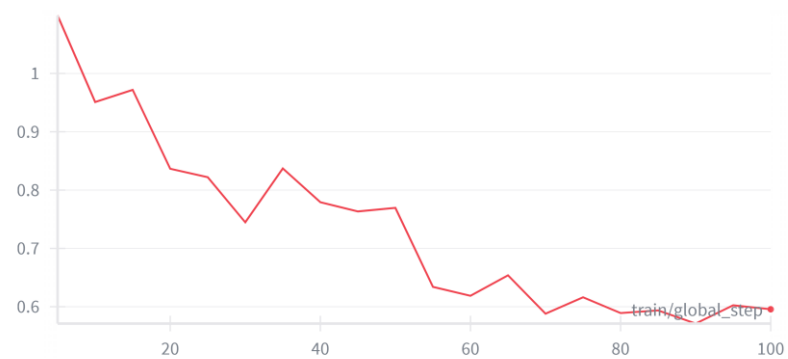
4.Result

eval/loss
— trans-finetune-llamafactory

train/loss
— trans-finetune-llamafactory

eval/loss

train/loss

eval/loss

train/loss

5. Interpretation & Conclusion

The project demonstrates effective domain-specific translation using LoRA.

It distinguishes natural language, technical terms, and code.

Future work includes GUI support, additional document types, and expanded glossary.