Code Assistant

with Tokenization & Evaluation Pipeline

Midterm Project, Generative AI, 11.Dec.2025

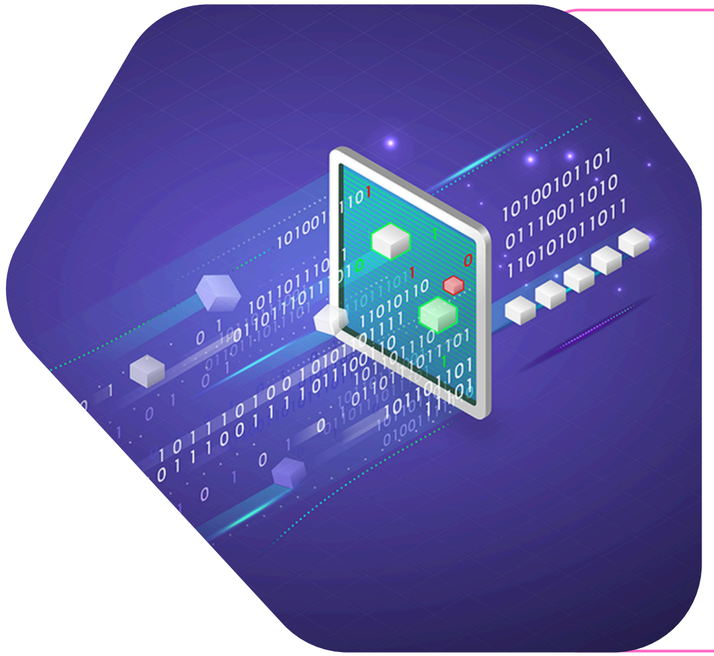
Group 2: Sondos Ibrahim, Sara Khirfan, Sadeel Alhaleeq, Tasneem Alassaf

Project Overview

This project showcases an **AI-powered code assistant** that explains, debugs, enhances, and tests code effectively.

Project Objectives

- *Analyze tokenization impact on efficiency and cost*
- *Build an optimized, multi-feature code assistant*
- *Compare automated vs. human evaluation*
- *Reduce hallucinations using structured prompting*



Tokenization Experiment

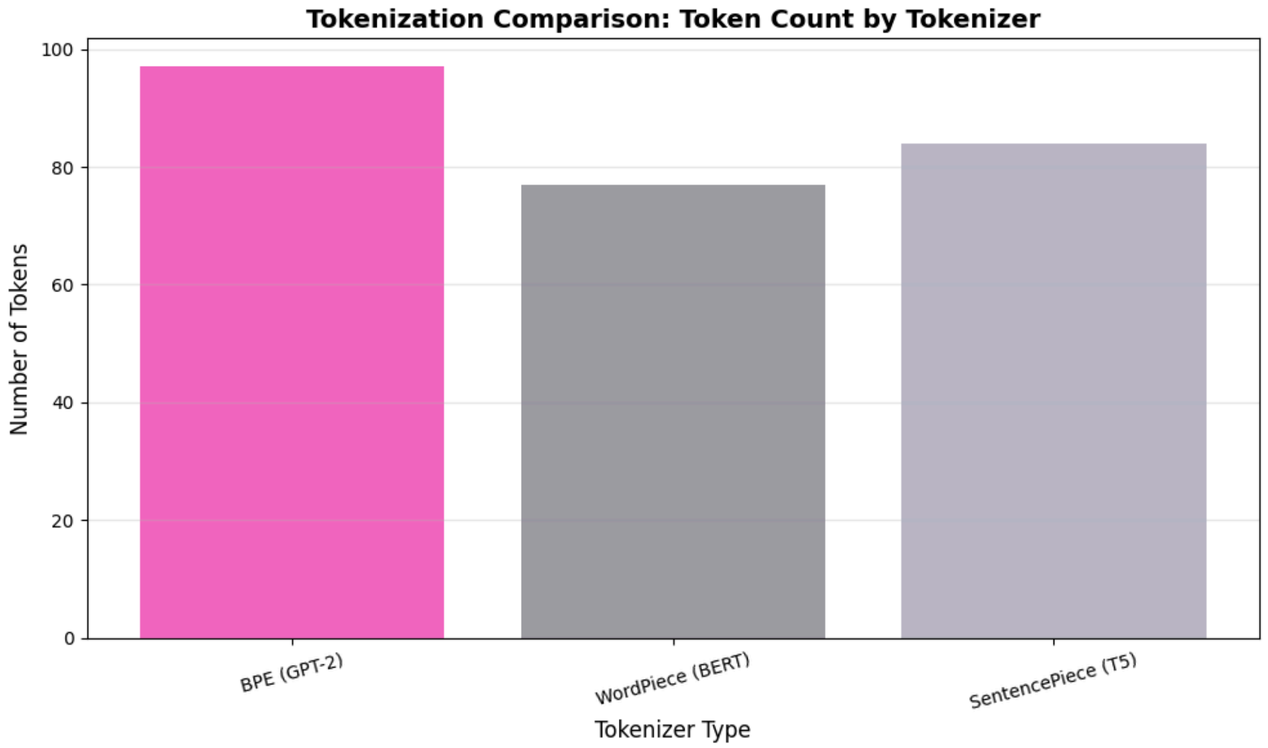
- Tested three tokenizers: BPE, WordPiece, SentencePiece
- Same Python sample used for comparison
- Measured: **token count, compression ratio, and semantic preservation**

Methods part 1

Tokenization Results

- **WordPiece:** *77 tokens*
- **SentencePiece:** *84 tokens*
- **BPE:** *97 tokens*

WordPiece proved unexpectedly the most efficient for Python code.



Token count comparison bar chart

Evaluation Pipeline

- Evaluated using **ROUGE-1 and BLEU**
- Human scoring rubric included clarity, correctness, coherence
- Compared basic vs detailed summarization prompts

Methods part 2

Automated Evaluation Results

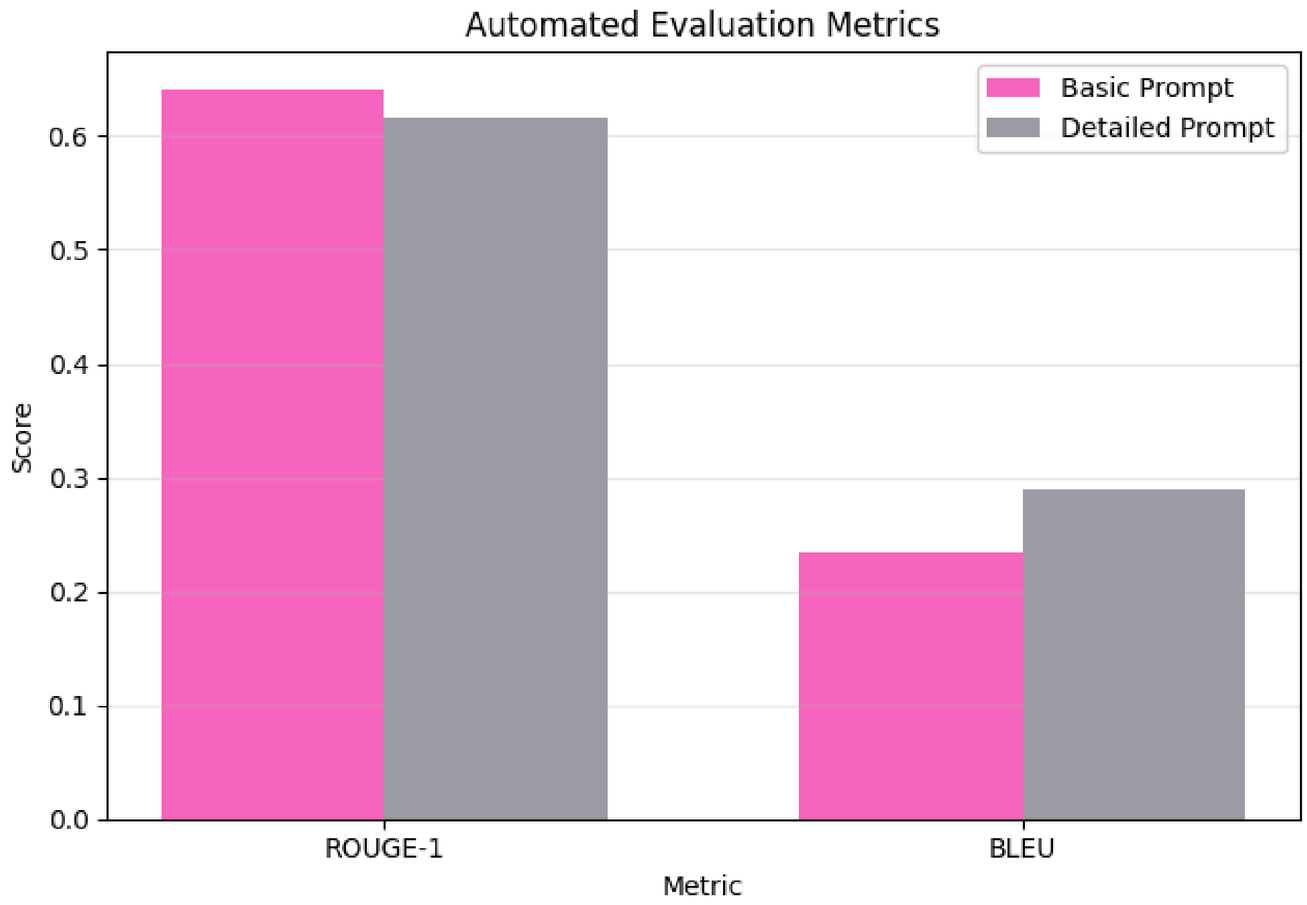
- **ROUGE-1:**

Basic = *0.641*, Detailed = *0.615*

- **BLEU:**

Basic = *23.35*, Detailed = *29.01*

Results were close, showing automated metrics don't fully capture quality.



ROUGE/BLEU comparison table

Human Evaluation Results

Human reviewers favored the **detailed prompts, giving them a perfect 15/15 versus 14/15 for the basic ones.**

Hallucination Detection

- ***Weak prompts caused the model to invent details about the non-existent "FastML" library.***

- ***Adding verification and anti-speculation rules produced accurate, factual responses.***

- ***Context grounding removed hallucinations entirely.***

Key insight: *Strong prompts and context control are the most effective ways to prevent hallucinations.*

Summary & Conclusion

1. WordPiece cut token usage and cost by ~26%.
2. Structured prompts gave the best results.
3. Automated metrics missed key quality differences.
4. Verification and context grounding reduced hallucinations.
5. Sequence length tuning balanced detail and cost.
6. The system is efficient, reliable, and practical.

Thank You!

Done By

Sondos Ibrahim, Sara Khirfan, Sadeel Alhaleeq, Tasneem Alassaf

Midterm Project, Generative AI 11.Dec.2025