Optimizing LLMs for Italian: Reducing Token Fertility and Enhancing Efficiency Through Vocabulary Adaptation

Moroni et al., 2025

Reproducibility Study

CS 421 — Fall 2025

Diana Gontero – Margherita Boanini

Why Optimize for Italian?

LLMs are mostly trained on English

Morphologically rich languages (like Italian) → subtoken fragmentation

Consequences:

more tokens	generation cost	zero-shot language quality

Paper's objective: semantically adapt vocabulary to improve Italian representations and efficiency.

Project Goal & Method

Goal:	Adapt LLaMA 3.1-8B for Italian
	Reducing token fertility
	Improving efficiency (tokenization speed, inference)
	Preserve generative quality (translation, QA)
Method:	Quantitative analysis of pre-trained models
	Mini-Continual training (Mini-CT) reproduction

Our Reproducibility Project

Quantitative Check-Up	Token Fertility
	Token Count
	Tokenization Speed
	Light Inference Time
	Generative evaluation (BLEU/ROUGE)
	Using official models: Base Llama, LAPT, SAVA, FVT
Mini-Continual Training	Simplified SAVA implementation
	Model initialization
	Mini-Continual Training with LoRA, 1 epoch

Models and Tokenizers

BASE MODEL:

Llama-3.1-8B

LAPT – Continual Training without vocabulary adaptation

FVT – Fast Vocabulary Transfer

SAVA – Semantic-Aligned Vocabulary Adaptation

- Align embeddings from helper model (Minerva-3B)
- Initialize new tokens in the target vocabulary

SAVA M-CT Reproduced

Our mini-CT reproduction of SAVA

Dataset

Corpus:

- 2,000 Italian Wikipedia sentences
- 2,000 English translations

Mini-CT Training Split:

- 75% Italian
- 25% English

Purpose:

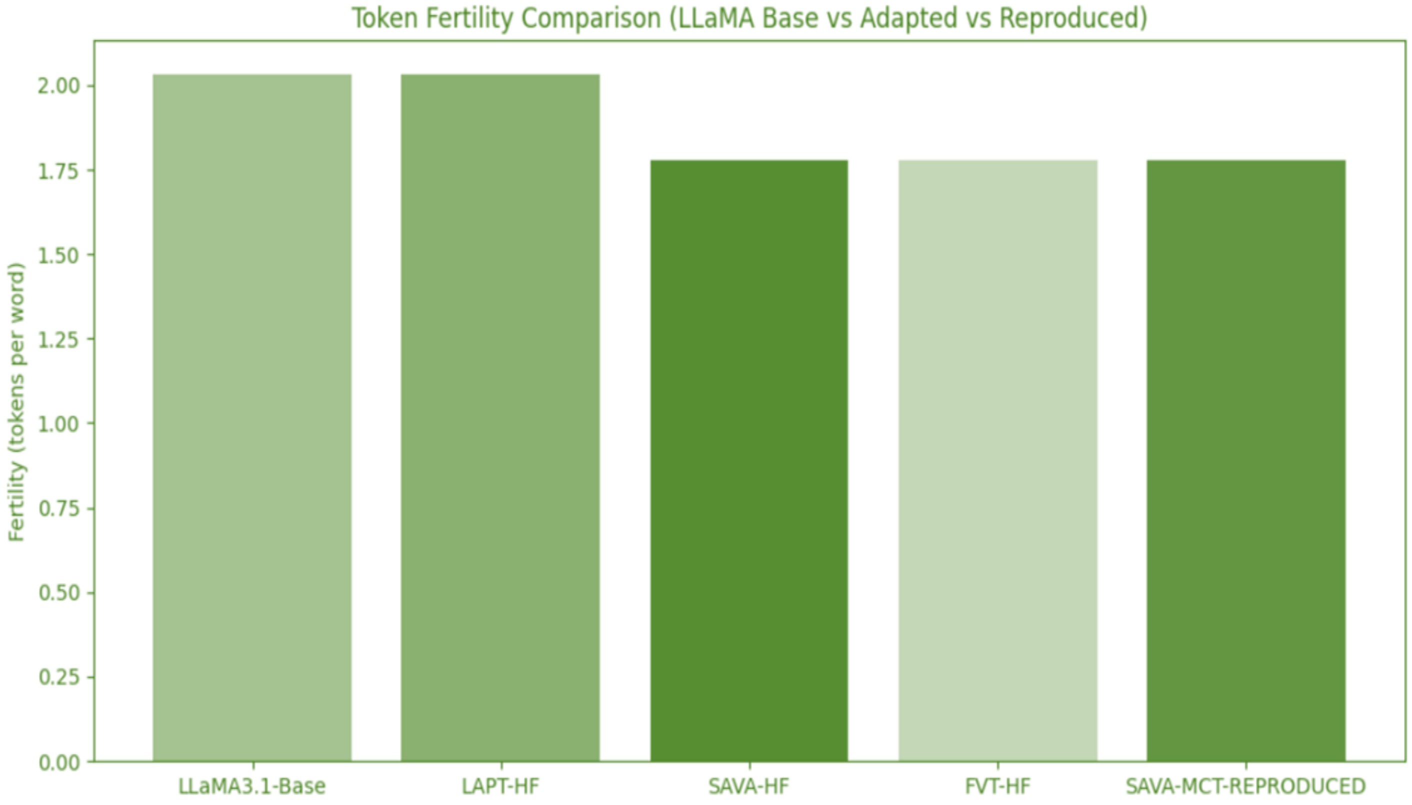
- Preserve English knowledge
- Teach Italian efficiently

Limitations

Element	Paper	Our Project
Continual Training	12B tokens, 16×A100 GPUs	~500 sentences, 1 epoch (LoRA)
Dataset	CulturaX	Sampled Wikipedia IT
Models	FP16 on multi-node	4-bit + single GPU
Metrics	ITA-Bench suite	small MT/QA test set

Reproducibility focus: maintain methodology, validate trends, document differences.

G1: Fertility

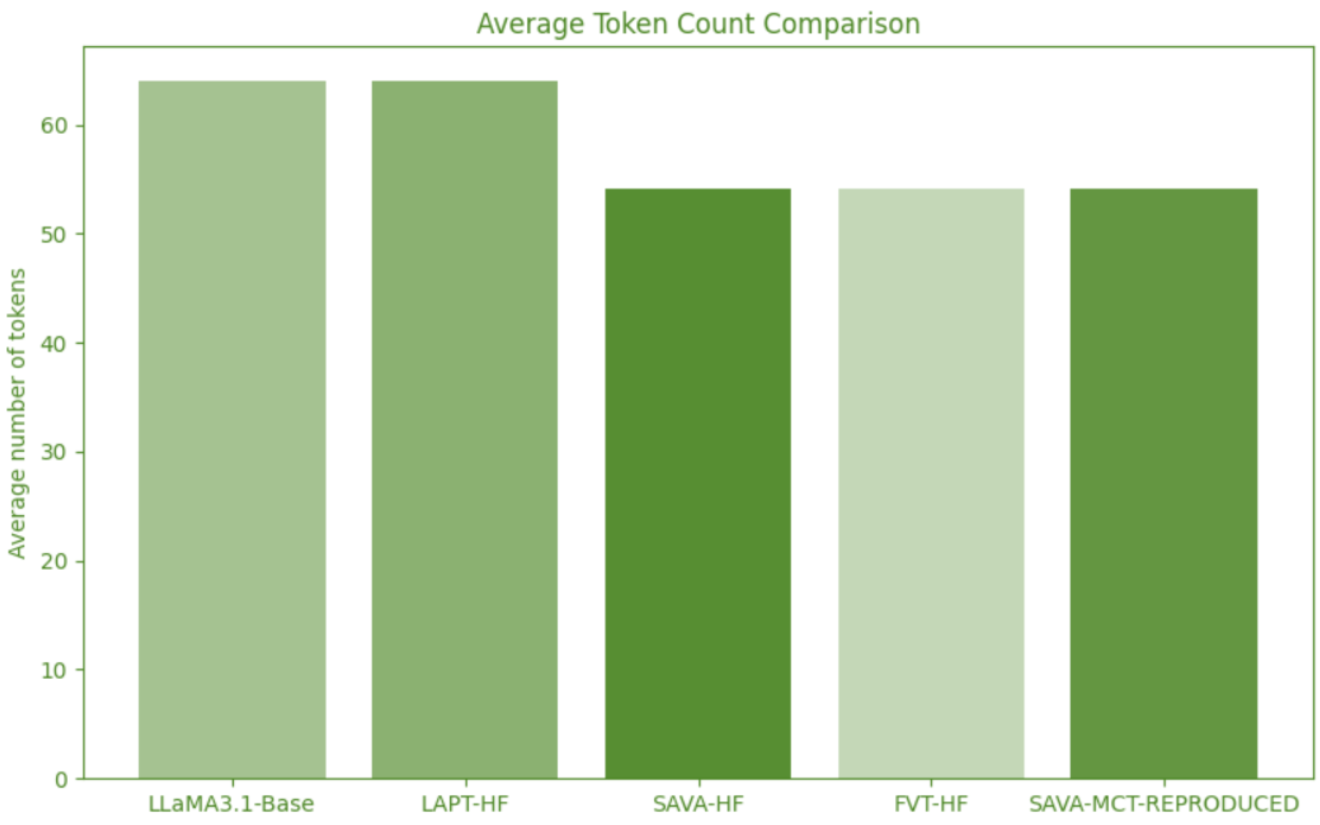


Vocabulary adaptation reduces token fertility

Mini-CT preserves gains:

2.03 → 1.78 (~12%)

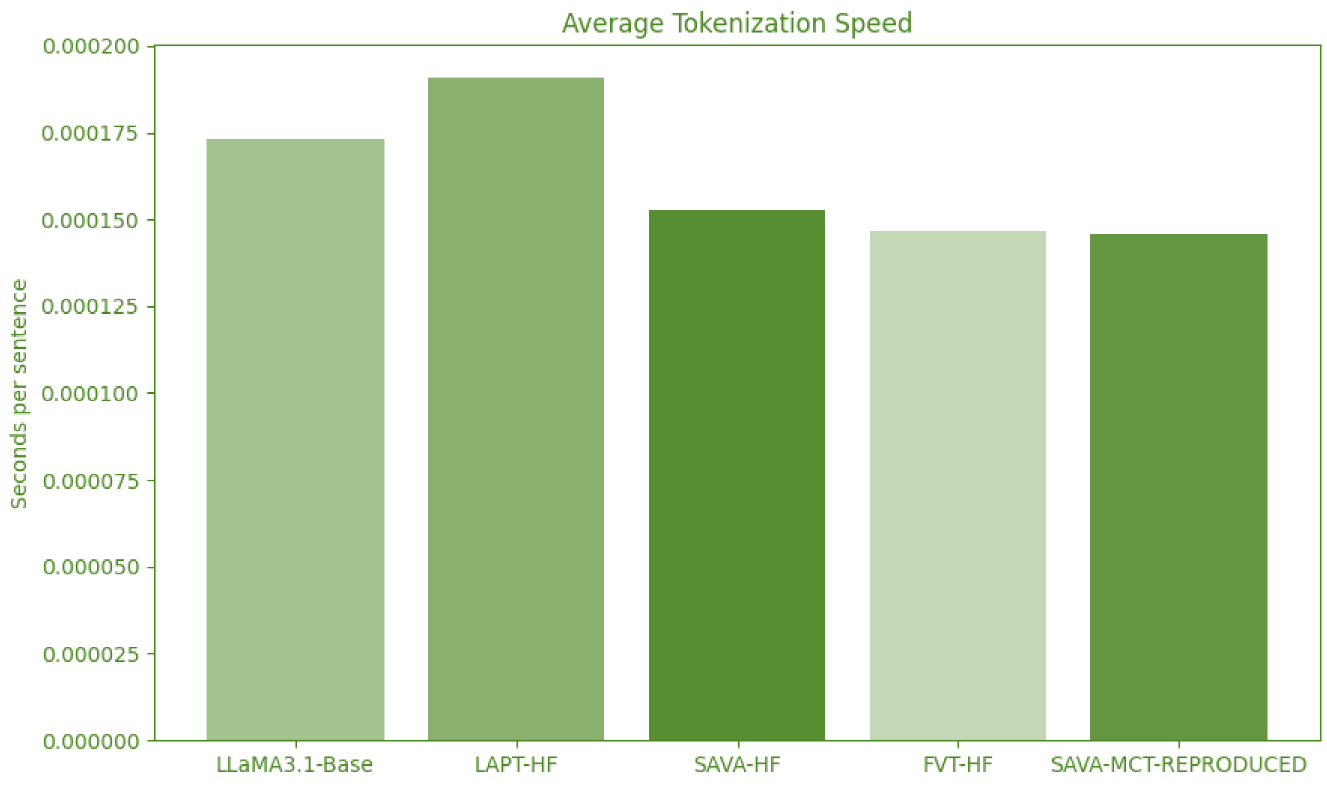
G2: Token count



SAVA & FVT reduce token count

Slight speed improvement in tokenization

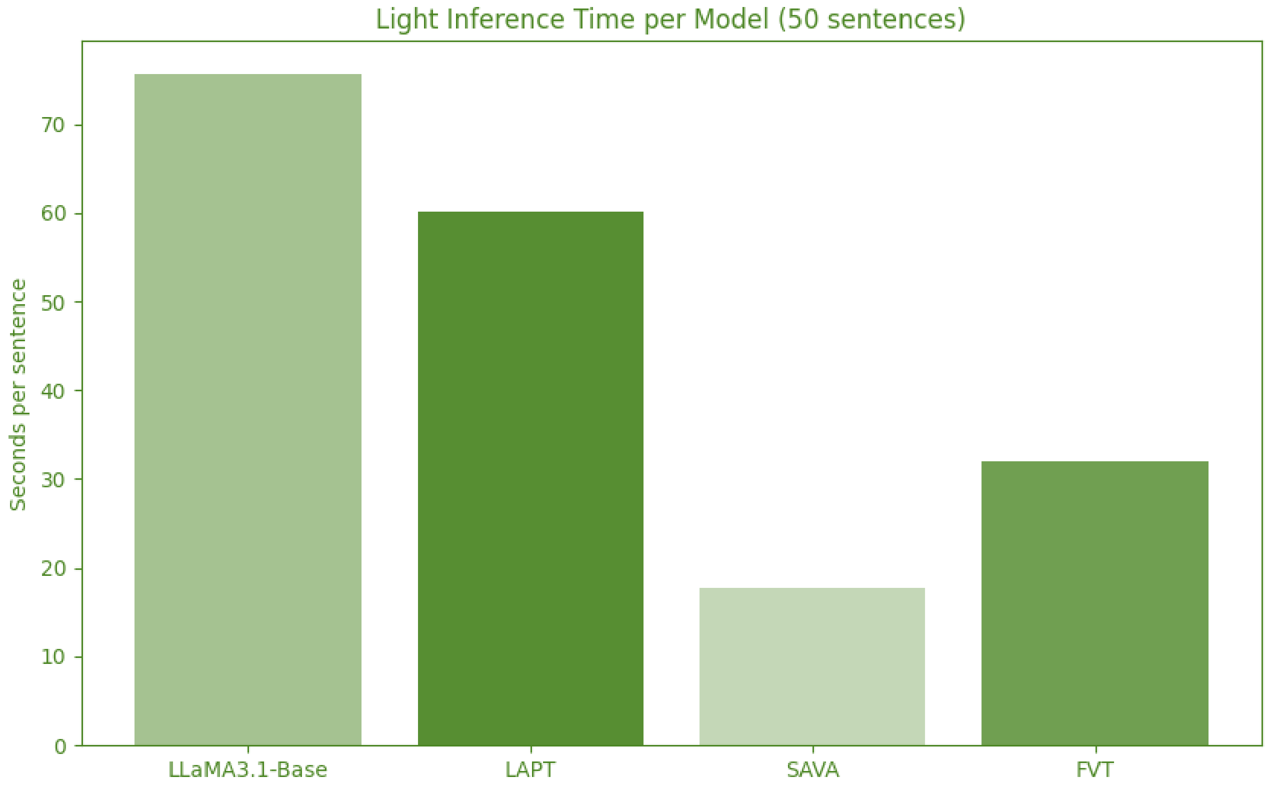
G2: Tokenization speed



SAVA & FVT are faster although the differences are small.

However, also smalling improvements matter.

G3: Inference Time



SAVA achieves fastest inference (~76% faster than Base)

Reduced token fertility directly improves generation speed

G4: Generative evaluation (BLEU/ROUGE)

Tokenizer	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Avg_latency_MT	Avg_Latency_QA
LLaMA3.1-Base	0.25	0.08	0.06	0.08	1.83	0.90
LAPT	0.26	0.21	0.14	0.20	1.81	1.09
SAVA	0.20	0.27	0.21	0.25	1.79	1.09
FVT	0.20	0.22	0.16	0.20	1.79	1.07

CONCLUSIONS

- Methodology Validated
- Efficiency Claims Proved
- Fast Convergence Confirmed
- Project Success Despite Technical Fragility

Thank you for your attention!

Diana Gontero

Margherita Boanini