PADERBORN UNIVERSITY

How to Train Your Large Language Model

Prof. Dr. Axel Ngonga

Tutor: Nikit Srivastava

Data Science Group
Paderborn University

Project Group - SoSe 2025

February 3, 2025

Language Models

Introduction

- Widespread Adoption
- Application Diversity
- AI-Driven Efficiency
- Continual Advancements

Image sources: vecteezy.com, flaticon.com, iconscout.com

Language Models

Shortcomings

Open Source Limitations

- Pay-to-use or hidden behind APIs (*e.g.*, GPT4, Gemini, Claude)
- Personal information requirements (*e.g.*, Llama)
- Not very "open" models (*e.g.*, Mistral, Grok, Llama)

Multilingual Gaps

- English centric (Üstün *et al.*, 2024)
- Limited multilingual coverage (Liu *et al.*, 2024)
- *The curse of multilinguality* (Conneau *et al.*, 2020)

Project Objective

Train a large and open-source multilingual language model and address the challenges posed by *the curse of multilinguality*.

- Support 500+ Languages
- Ensure Computational Efficiency
- Enable Multimodal Capabilities
- Maintain Linguistic Extensibility

Project Tasks

What types of tasks will the project group be responsible for?

- Study SOTA Models
- Gather Training Data
- Assess Frameworks
- Implement Custom Models
- Create Training/Evaluation Pipelines
- Document Findings

Learning Expectations

What knowledge and skills can we expect to gain by participating in this project group?

- Advanced ML Techniques
- LLM Inner Workings
- Distributed Computing
- Research and Literature Review
- Project Management
- Collaborative Work

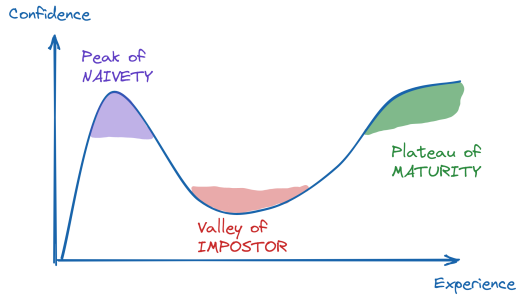

Image source: https://newsletter.techworld-with-milan.com/

Candidate Requirements

What are the requirements for joining this project group?

- Basic NLP and ML Knowledge
- Python and Shell Programming
- Linux Proficiency
- Adapt to Steep Learning Curve

We Offer

- Expert Tutors
- Training Compute Resources
- Follow-up Thesis Opportunities
- Publication Support

That's all Folks!

dice-research.org/teaching/HTYLLM-2025

Have questions?

Email: nikit.srivastava@uni-paderborn.de
Matrix: @nikit:chat.dice-research.org