PADERBORN UNIVERSITY

How to Train Your Large Language Model

Kickoff Meeting

Prof. Dr. Axel Ngonga

Tutor: Nikit Srivastava

Data Science Group
Paderborn University

Project Group - SoSe 2025

April 7, 2025

About DICE

What does DICE group do?

- Knowledge Graphs
- Natural Language Processing
- Machine Learning

For more information: https://dice-research.org/

Section 1

PG Organization

PG Structure and Deliverables

How is this project group structured?

- **Semesters:** SoSe 25 and WiSe 25/26
- **During Semesters:**
  - Regular Group Meetings – (Usually) Every Week
  - Weekly Report
- **Semester-end Requirements:**
  - Individual Report, Group Report, and Group Presentation
  - Grading and Feedback
- **Final Submission:** Source Code and Documentation

PG Setup

How is the project group setup and organized?

Module Handbook: *Project group is intended to be a preparation for industrial practice.*

- **Students**: Self-organized Dev Team
- **Supervisor**: Product Owner and Contact Point
- **Time Investment**:
  - 20 hours per week (with vacations)
  - 15 hours per week (without vacations)

Expectations

What do we expect from you?

→ Be a valuable member of your team!

- Manage your project
- Write code(!) and commit it
- Communicate with team members
- Support each other where possible
- ...

→ We won't keep students if they are not participating in their team

PG Schedule

- **Kickoff Meeting:** 07.04.25, FU.504
- **Weekly Meetings:** Date TBD, FU.504
- **Registration:** Studienleistung and Exam before 21.05.25—or leave the PG!
- **End SoSe 25:**
  - Individual Report, Group Report, and Group Presentation
  - Grading and Feedback
- **Between SoSe 25 and WiSe 25/26:** Vacations
- **Start WiSe 25/26:** Resume PG Work
- **Mid-January 2026:** End of the Implementation Phase
- **End WiSe 25/26:**
  - Individual Report, Group Report, and Group Presentation
  - Final Grading

Unforeseen Events

- **Vacations:**
  - Should be avoided during lecture time!
  - In very urgent cases, contact your supervisor
- **Project Issues:**
  - Problems within the team
  - Hardware issues
  - Personal issues

  → Get in touch with your supervisor
- **Illness:**
  - Get a **Doctor's Certificate**!
  - If you can foresee it, inform your teammates

Section 2

PG Objectives Recap

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

Section 3

Goals & Grading

PG Goals

Current Semester

What are the goals for the current semester?

1. Propose multiple (min. 4) approaches
2. Train and evaluate each approach at small scale
3. Provide in-depth analysis on what works better and what doesn't
4. Scale up the top 2 approaches

→ Considerations:

- **Mixture-of-Experts**: Possible architectural variations
- **Language Coverage**: How to deal with low-resource languages?
- **Resource Estimation**: Stay within the provided resource budget

Grading

How will the students be graded?

- Achieved Goals:
  - **Group**: What goals did the group achieve?
  - **Individual**: How much did you contribute towards this goal?
- Reports and Presentations (Group, Individual)
- Soft Skills (Group, Individual):
  - Commitment and engagement
  - Collaborative and learning skills
  - Motivation and volitional strategies
  - Scientific writing and self-management

Section 4

Icebreaker

First Task

"*Talk is cheap. Show me the code.*" – Linus Torvalds

1. Choose five distinct languages
2. Train a series of language models:
   - One model for each of the five languages
   - One covering all five languages
3. Evaluate all models on different tasks
4. Report findings

→ **Time Limit**: 1 Week

→ **Resources**: Access to noctua2 cluster, and a shared GPU VM (Upto 3 accounts)

Next Steps

"*A place for everything, everything in its place*" – Benjamin Franklin

→ Organize Yourselves

- Team Structure: Flat, Hierarchical, Holacratic, . . .
- Regular Meetings: `https://terminplaner6.dfn.de/`
- Communication: UPB Matrix
- Tasks Management: UPB KanBoard
- Code Repository: `https://github.com/dice-group/HTYLLM-PG`
- Other Tools?: Host them yourself @ `htyllm-pg.cs.uni-paderborn.de`

→ Try to use open-source tools wherever possible

Let's Go!

dice-research.org/teaching/HTYLLM-2025

Have questions?

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