HEINZ NIXDORF INSTITUT
UNIVERSITÄT PADERBORN

How to Train Your Large Language Model 2

(HTYLLM2)

Prof. Dr. Axel Ngonga

Tutor: Nikit Srivastava

DICE

Data Science Group
Paderborn University

Project Group - SoSe 2025

February 9, 2026

Language Models

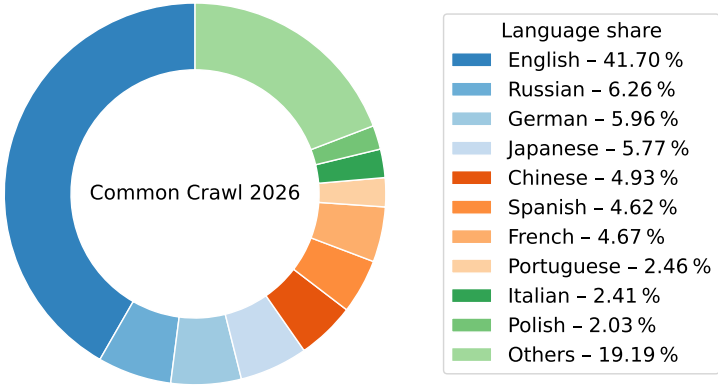
Introduction

- Widespread adoption
- Application siversity
- AI-driven efficiency
- Continual advancements

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

Language Models

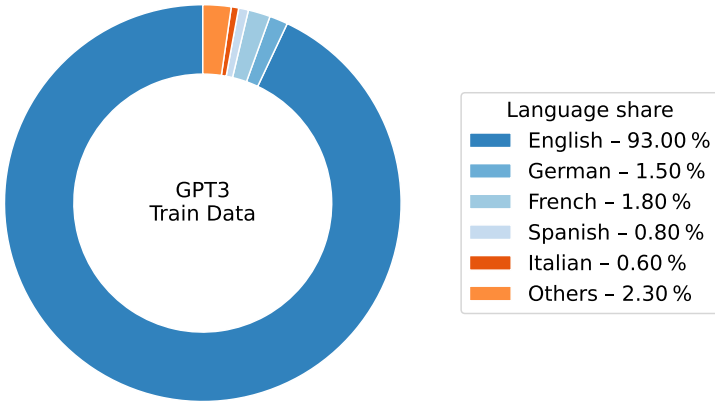
Training Corpus



Ref: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

Language Models

Training Subsample



Ref: Johnson et al., 2022

Language Models

Shortcomings

Open Source Limitations

- Pay-to-use or hidden behind APIs (*e.g.*, GPT5, Gemini, Claude)
- Personal information requirements (*e.g.*, Llama)
- Not very "open" models (*e.g.*, Mistral, Grok, GPT-OSS)

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

Image source: https://de.freepik.com

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

Image source: https://de.freepik.com

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 and collaboration

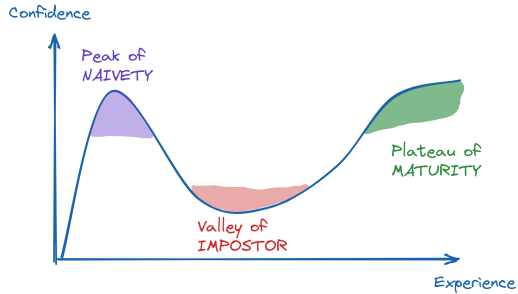


Image source: https://newsletter.techworld-with-milan.com/

Candidate Requirements

What is expected of the candidates applying to join this project group?

- Basic NLP and ML knowledge
- Python and shell programming
- Adapt to steep learning curve
- Strong problem-solving attitude

Image source: https://de.freepik.com

We Offer

- Expert tutors
- Training compute resources
- Follow-up thesis opportunities
- Publication support

That's all Folks!

dice-research.org/teaching/HTYLLM2-2026

Have questions?

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