Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

data science

CAIDAS

ML NLP Praktikum SS 2026

Jan Pfister, Julia Wunderle, Tom Völker

University of Würzburg

Organisation

- 5 (Prjak) or 10 (PIS 1/2) ECTS
- Work in small groups
- **Submission:**
  - Code: incl. instructions on how to run it; Python as programming language, PyTorch as Deep Learning framework
  - Project report: 1 per group for NLP, 10-15 pages; German or English; Marking which group member contributed to what parts (of implementation, design and writing)
  - Intermediate status reports in presentation form
  - Presentation at the end of the semester

Task

- **Generally:**
  - Independent work on a research task (advised by us)
  - Training to use current techniques from NLP/Machine Learning
  - Practical application of the content of the MLNLP lecture (no formal requirement)
- **This Semester:**
  - Several different topics: SuperGLEBer (German Benchmark), Reinforcement Learning, Data Synthesis, Citation Analysis, Model Safety

Project A: Data Synthesis

- Generate synthetic training data using LLMs for domain-specific tasks
- Compare different synthesis strategies (prompting, few-shot, self-instruct)
- Train downstream models on synthetic vs. real data
- Quality evaluation? (llm as a judge, quality filter, small ablations…)
- Compute available on project cluster
- (Focus on German-language)
- **Interests:** Evaluation methodology, prompt engineering, data curation, model training,

Project B: Reinforcement Learning for LLMs

- Systematic comparison of RL algorithms (PPO, DPO, GRPO, etc.) on the same data
- Design & evaluate domain-specific reward functions/ environments for reasoning tasks (e.g. Cyber Security)
- Train & benchmark verifier models for output validation
- Hard vs. easy sample analysis: difficulty-based training curricula
- **Interests:** Reinforcement learning, LLM training, reward modeling, distributed compute

Project C: Safety Training & Benchmarking

- Survey & benchmark existing safety datasets and evaluation methods
- Compare safety training approaches: SFT, RL, RLHF,…
- Evaluate problematic content detection across multiple LLMs
- Train safety-aligned models and measure trade-offs with task performance
- (Focus on German-language safety considerations (sovereign AI context))
- **Interests:** AI safety, evaluation design, posttraining

Project D: Citation Analysis Pipeline

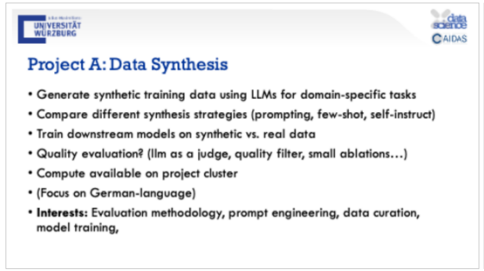
- Large-scale scraping & classification of citation intents across ArXiv (CS domain)
- AI-generated text detection at scale (e.g. Pangram, open-source classifiers)
- Track individual researchers over time: AI usage trends & citation style evolution
- Track individual papers: how citation intent changes as a paper ages
- Modular pipeline: scraping → classification → longitudinal analysis
- **Interests:** Data engineering, text classification, bibliometrics

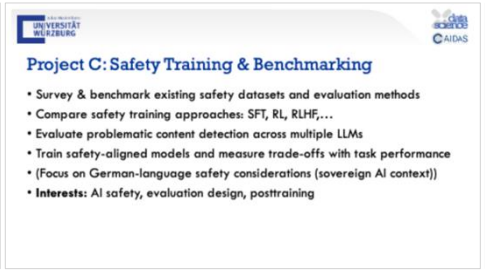
Project E: Error Analysis & Evaluation Framework

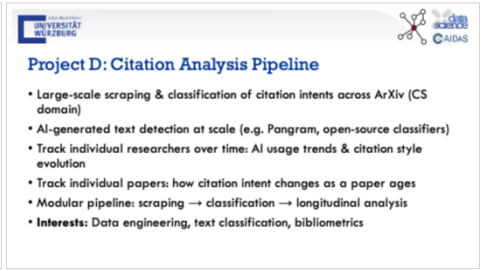
- Replace ad-hoc Jupyter notebooks with proper CLI evaluation pipeline
- Confusion matrices, per-category breakdowns, span analysis,…
- Statistical significance testing for model comparisons
- Automated error categorization
- Seed-variance analysis across multi-seed runs
- **Interests:** Applied NLP evaluation, data engineering

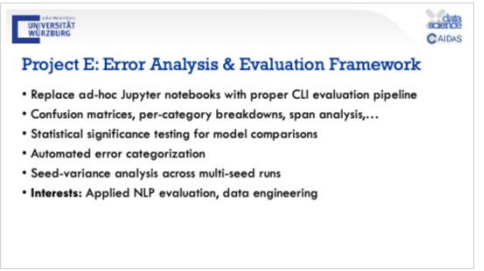
Next Steps

- Registration:
  - Register via email to mlnlprjak@informatik.uni-wuerzburg.de **until Sunday 19th**
    - preferred project
    - desired number of ECTS (5 or 10)
    - group members
- Meetings every two weeks to discuss progress









- A Data Synthesis
- B Reinforcement Learning for LLMs
- C Safety Training & Benchmarking
- D Citation Analysis Pipeline
- E Error Analysis