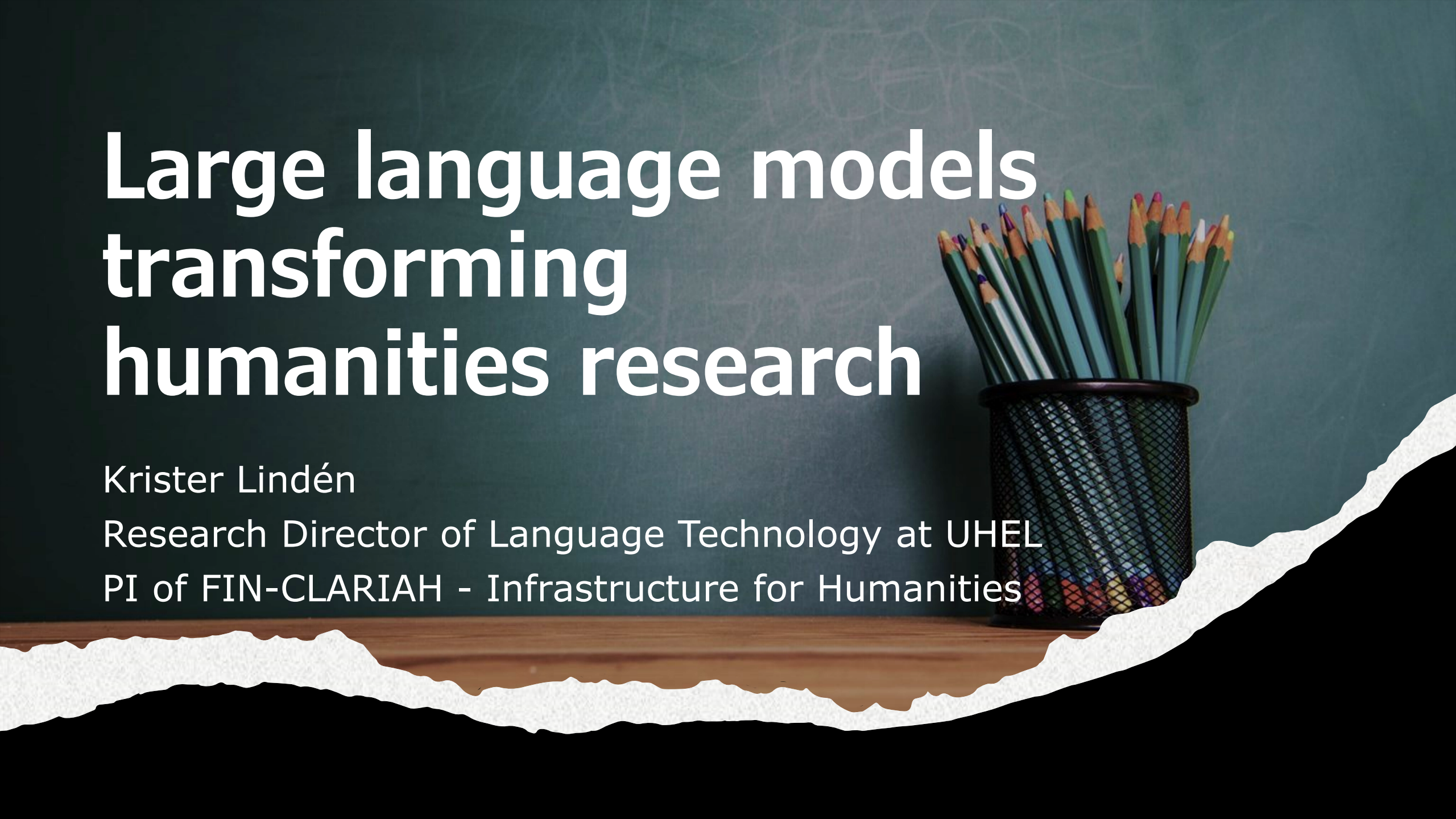
Large language models transforming humanities research



Krister Lindén
Research Director of Language Technology at UHEL
PI of FIN-CLARIAH - Infrastructure for Humanities

Large Language Models (LLM) and AI

- LLMs are trained to predict the next word, sentence and paragraph based on some given input
- Special training data are needed for
  - Structured answers with intro, analysis and summary
  - Each language, dialect, domain, …
  - Structured reasoning

LLM development in Europe and beyond

- Language-specific LLMs:
  - Bulgarian, Czech, Dutch, Greek, Finnish, Flemish, German, Latvian, Norwegian, Portuguese, Slovenian, Swedish, …
- Fine tuning of LLMs:
  - Icelandic, Sámi languages, South African languages, …
- Multilingual models and projects:
  - HPLT for EU+ languages based on Common Crawl,
  - DeepR3: (Spanish, Catalan, Basque, Galician, English) for new domains, genres and languages,
  - PULI GPT (Hungarian, English, Chinese),
  - Swiss LLM (French, German, Italian, Romansh),
  - YugoGPT (Bosnian, Croatian, Slovenian), …

Humanities Research

- Critically study and review sources to synthesise results
  - Get an overview of the sources deciding on their relevance
  - Answer specific research questions finding and checking facts
  - Create a synthesis and communicate the results
- The final frontier of manual labour to be automated
  - Still need to ask relevant questions
  - Still need to ask for the sources and check facts

LLMs as an Infrastructure for Research

- Modality conversion:
  - OCR, HTR, ASR, STT, sign language processing, video and picture narration and generation, …
- Annotation:
  - NER, description generation, metadata, keywords, classifications, …

LLMs as an Infrastructure for Research

- Data analysis
  - Perform analyses that previously were made by data scientists, computational linguists or statisticians
- Knowledge discovery
  - New forms of exploratory distant reading with an LLM as an interface

LLMs as an Infrastructure for Research

- Output for specific audiences:
  - Literature survey
  - Summaries
  - Scholarly paper
  - Review
  - Translation
  - Keyword generation
  - Simplification
  - News items
- https://curre.helsinki.fi/chat

Areas of Research related to LLMs

- Language acquisition modelling
- Knowledge representation
- Model creation and transformation
- Benchmarking LMMs
- Ethics of AI
- Responsible AI
  - Transparent, fair, explainable, robust, privacy and security observing, organisationally well-governed

Societal Consequences of LLMs

- Will we become dependent on the developers and their biases?
  - Cf. computer operating systems
- Will AI replace part of the work force?
  - Yes, but the expectations increase, and the focus will shift to tasks we now consider too time consuming
  - Cf. text processing and calculators

Economic Consequences of LLMs

- Long-term productivity
  - LLMs represent a platform shift like mobile phones and the internet over the next 10-30 years
  - LLMs accelerate global automation and productivity in 10 years
  - LLMs may double the productivity of higher education professionals
- Short-term effects
  - Increased use of multimodal data
  - Potentially reduced access to human-produced text as content-producers begin holding on to their manually verified data for monetary reasons

Human Consequences of LLMs

- Is there a danger in outsourcing our thinking?
  - Not if we train ourselves to use critical thinking and common sense. In routine tasks, this is difficult to maintain.
- If we use LLMs to produce more scientific papers, who is going to read all of them?
  - Most likely other LLMs as the human capacity to learn more quickly is unlikely to increase.
  - It is still important to create more verified well-researched data!