Text+

Collections
Lexical
Resources
Editions
Infrastructure/
Operations

Posterslam

3. Text+ Plenary in Mannheim

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Förderungsnummer 460033370

Teil der
nfdi Nationale Forschungsdaten Infrastruktur

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Anwendungen und Methoden der Abgeleiteten Textformate (ATF) im Kontext von LLMs

  Florian Barth, José Calvo Tello, Keli Du, Philippe Genêt, Peter Leinen, Jörg Knappen, Thorsten Trippel, Andreas Witt

- ~~Decoding Discourse: Gender Dynamics in German Bundestag Debates (1949-2021)~~

  ~~Teresa Hailer~~

- Empowering AI Knowledge Management: A Community-Organizing Approach to Enhance Fidelity and Quality through Authority File Use in Metadata

  Barbara Fischer

- Entwicklung von Transformer-basierten Modellen für historische Textnormalisierung

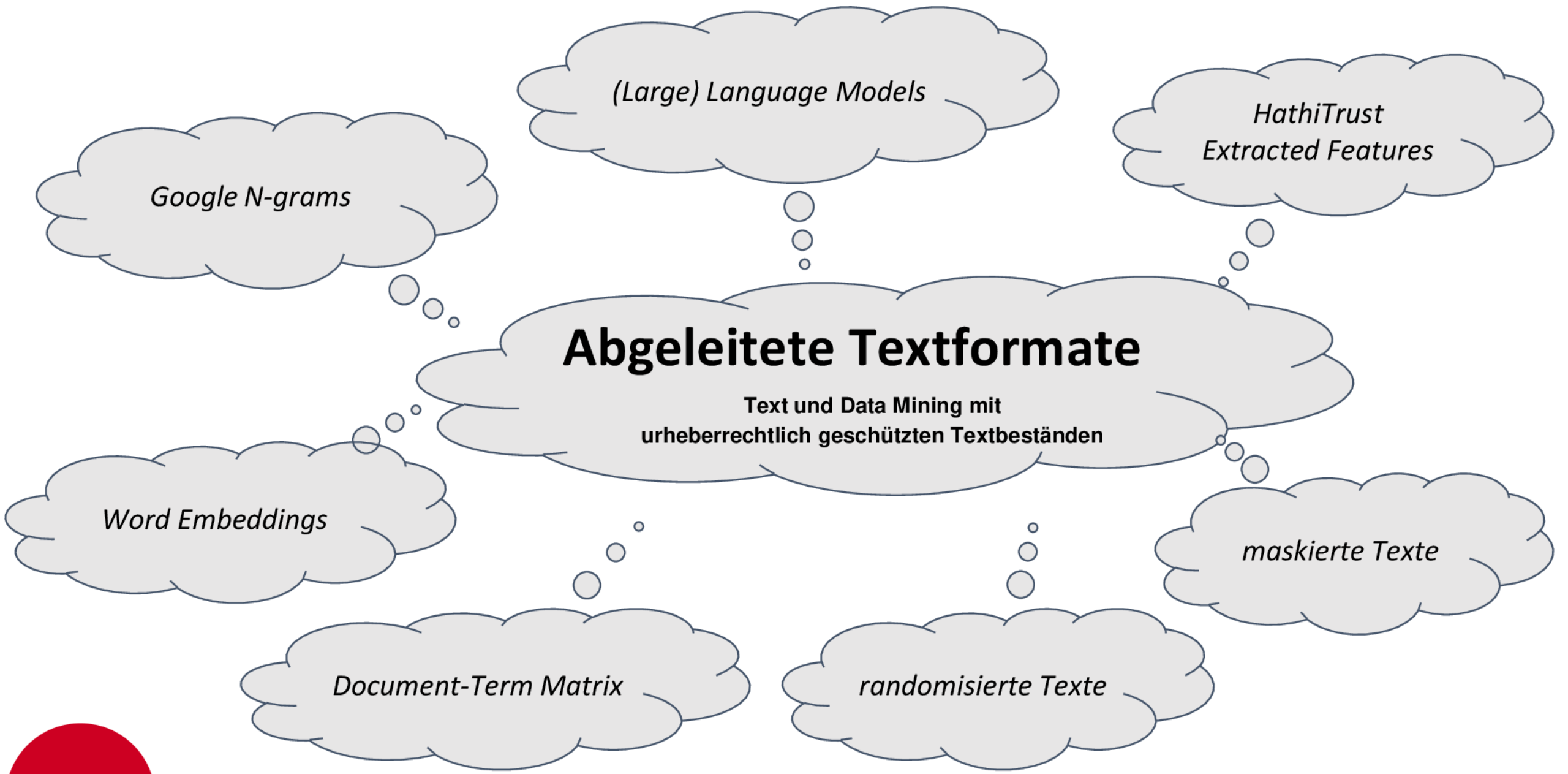
  Yannic Bracke, Gregor Middell, Alexander Geyken

- Evaluation of LLMs to Support the Development of GermaNet

  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn

- HERMES – Humanities Education in Research, Data, and Methods

  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak



Text+

Anwendungen und Methoden der Abgeleiteten Textformate (ATF) im Kontext von LLMs

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Anwendungen und Methoden der Abgeleiteten Textformate (ATF) im Kontext von LLMs

  Florian Barth, José Calvo Tello, Keli Du, Philippe Genêt, Peter Leinen, Jörg Knappen, Thorsten Trippel, Andreas Witt

- ~~Decoding Discourse: Gender Dynamics in German Bundestag Debates (1949-2021)~~

  ~~Teresa Hailer~~

- Empowering AI Knowledge Management: A Community-Organizing Approach to Enhance Fidelity and Quality through Authority File Use in Metadata

  Barbara Fischer

- Entwicklung von Transformer-basierten Modellen für historische Textnormalisierung

  Yannic Bracke, Gregor Middell, Alexander Geyken

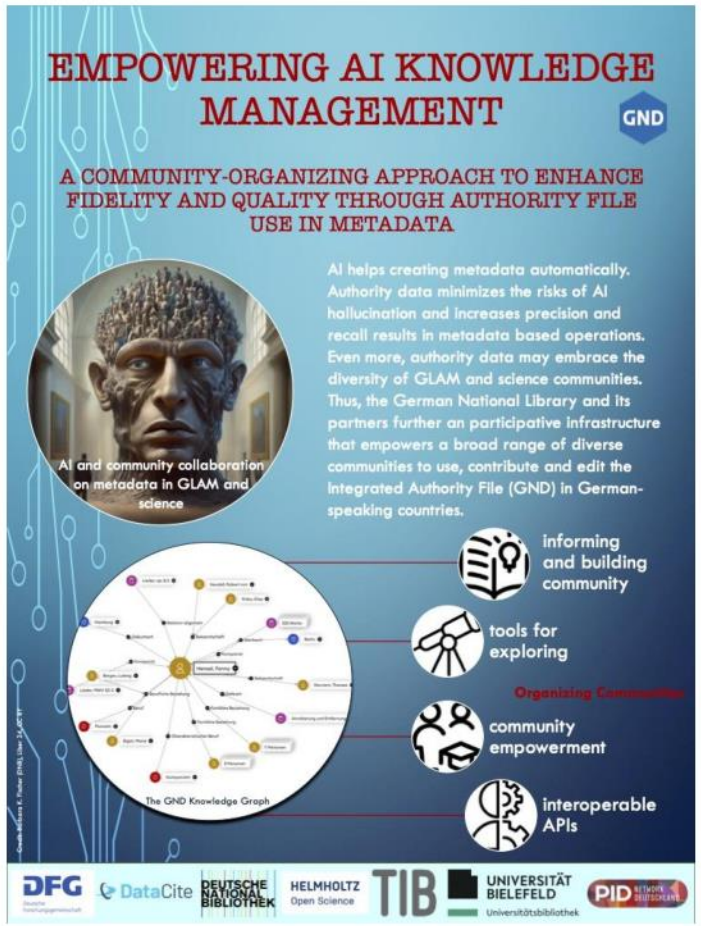
- Evaluation of LLMs to Support the Development of GermaNet

  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn

- HERMES – Humanities Education in Research, Data, and Methods

  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak

KI-basierte Wissensgraphen besser machen



1. GND-Normdaten integrieren
2. Precision & Recall verbessern
3. weniger Halluzinationen
4. mehr demokratische Kontrolle bei der Wissensorganisation durch partizipative Gestaltung der Normdaten
5. Integration von mehr Communities ermöglichen

GND GEMEINSAME NORMDATEI

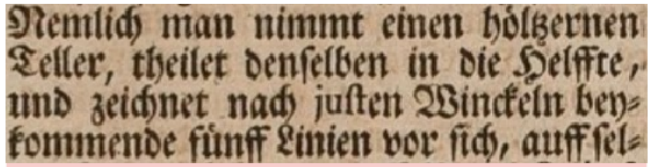
Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- ~~Decoding Discourse: Gender Dynamics in German Bundestag Debates (1949-2021)~~
  ~~Teresa Hailer~~
- Empowering AI Knowledge Management: A Community-Organizing Approach to Enhance Fidelity and Quality through Authority File Use in Metadata
  Barbara Fischer
- Entwicklung von Transformer-basierten Modellen für historische Textnormalisierung
  Yannic Bracke, Gregor Middell, Alexander Geyken
- Evaluation of LLMs to Support the Development of GermaNet
  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn
- HERMES – Humanities Education in Research, Data, and Methods
  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak
- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer

Transnormer – ein Transformer-basierter Ansatz für die Normalisierung historischer Texte

Yannic Bracke, Gregor Middell, Alexander Geyken (BBAW)


Nemlich man nimmt einen höltzernen Teller, theilet denſelben in die Helffte, und zeichnet nach juſten Winckeln beykommende fünff Linien vor ſich, auff ſel-

Nemlich man nimmt einen **höltzernen** Teller, theilet denſelben in die **Helffte**, und zeichnet nach juſten Winckeln **beykommende fünff** Linien vor ſich, auff ...

Nämlich man nimmt einen **hölzernen** Teller, teilt denselben in die **Hälfte**, und zeichnet nach justen Winkeln **beikommende fünf** Linien vor sich, auf ...

Quelle: Fleming, Hans Friedrich von: Der Vollkommene Teutsche Jäger. Bd. 1. Leipzig, 1719. In: Deutsches Textarchiv, https://www.deutschestextarchiv.de/fleming_jaeger01_1719/425

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Empowering AI Knowledge Management: A Community-Organizing Approach to Enhance Fidelity and Quality through Authority File Use in Metadata
  Barbara Fischer
- Entwicklung von Transformer-basierten Modellen für historische Textnormalisierung
  Yannic Bracke, Gregor Middell, Alexander Geyken
- Evaluation of LLMs to Support the Development of GermaNet
  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn
- HERMES – Humanities Education in Research, Data, and Methods
  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak
- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer
- Korpusproduktion in Zeiten großer Sprachmodelle
  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder

LLMs to support the development of GermaNet

R. Barkey, E. Hinrichs, M. Hinrichs, K. Sharp, & C. Zinn

- GermaNet mit 215,000 lexikalischen Einträgen, organisiert in 167,163 synsets
- Für Verben gibt es immer Beispielsätze
  - für Nomen und Adjektive aber selten

Respekt n. (Gefühl)
Paraphrase: *leichte Angst vor etwas*
Beispiel: *Er hatte vor Schlangen großen Respekt.*
Hyperonym: *Angst, Bammel, Muffe, Schiss…*
Hyponym: *Heidenrespekt*

➢Verwende LLMs zur Anreicherung

➢sind auf großen Datenmengen trainiert

➢generierte Sätze spiegeln die Wahrscheinlichkeit eines Satzes im Korpus wider

- Beispielsätze häufig typisch
- bessere Datengrundlage als Sprachexperten

➢Evaluation der von ChatGPT erzeugten Beispielsätze

➢Stärken und Schwächen (Spoiler: Frequenzeffekte und Polysemie)

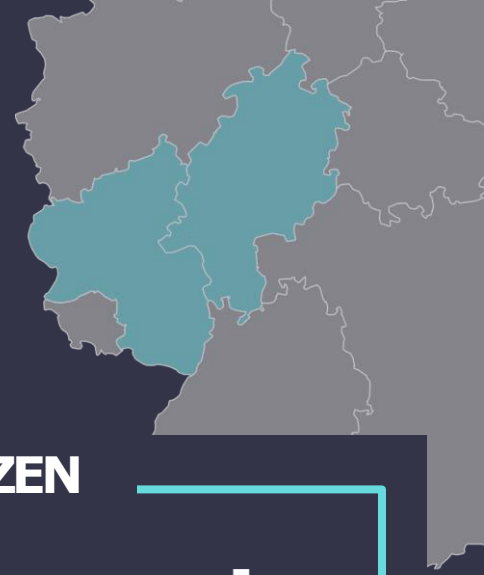
Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Entwicklung von Transformer-basierten Modellen für historische Textnormalisierung
  Yannic Bracke, Gregor Middell, Alexander Geyken
- Evaluation of LLMs to Support the Development of GermaNet
  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn
- HERMES – Humanities Education in Research, Data, and Methods
  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak
- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer
- Korpusproduktion in Zeiten großer Sprachmodelle
  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder
- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht
  Margret Mundorf

HERMES

Humanities Education in Research, Data, and Methods



FORSCHEN

- Bring-your-own-data-Lab
- Data Challenges
- Forschungsstudienprogramm

LERNEN

- Data Carpentries
- Transferwerkstatt
- Open Educational Resources

VERNETZEN

- Kommunikationsformate
- HERMES-Hub
- WissKomm Academy – HERMES Edition

Philipps Universität Marburg · Akademie der Wissenschaften und der Literatur Mainz · Deutsche National Bibliothek · Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft · h_da hochschule darmstadt · eutop · Hochschule Mainz University of Applied Sciences · IEG Leibniz-Institut für Europäische Geschichte · Technische Universität Darmstadt · Universität Trier

Finanziert von der Europäischen Union NextGenerationEU

Gefördert vom Bundesministerium für Bildung und Forschung

Agenda

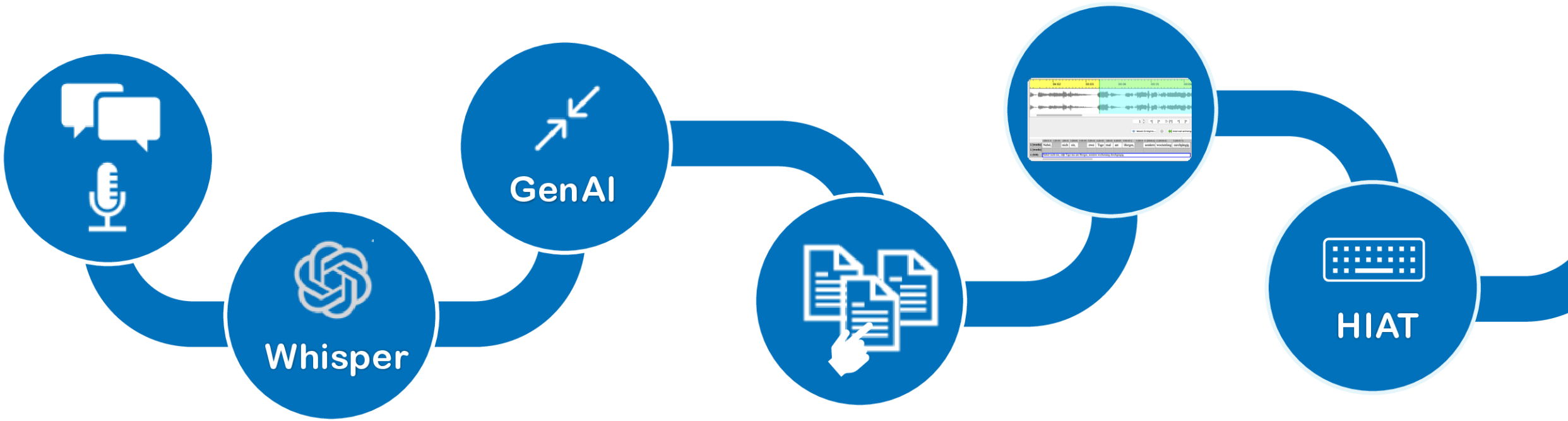
Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Evaluation of LLMs to Support the Development of GermaNet
  Reinhild Barkey, Erhard Hinrichs, Marie Hinrichs, Kimberly Sharp, Claus Zinn
- HERMES – Humanities Education in Research, Data, and Methods
  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak
- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer
- Korpusproduktion in Zeiten großer Sprachmodelle
  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder
- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht
  Margret Mundorf
- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data
  Liudmila Olalde, Thomas Malten, Frank Grieshaber

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hzsk hamburger zentrum für sprachkorpora

Text+



KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten

Kristin Bührig
Marcel Fladrich
Alina Hemmer

Text+ Plenary, 10.10.2024

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- HERMES – Humanities Education in Research, Data, and Methods
  Ruth Reiche, Andrea Rapp, Anna Schlander, Ksenia Stanicka-Brzezicka, Johanna Konstanciak
- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer
- Korpusproduktion in Zeiten großer Sprachmodelle
  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder
- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht
  Margret Mundorf
- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data
  Liudmila Olalde, Thomas Malten, Frank Grieshaber
- More uniformity and more diversity at the same time: LLMs and a 21st century standardisation paradox
  Christian Mair

Korpusproduktion in Zeiten großer Sprachmodelle

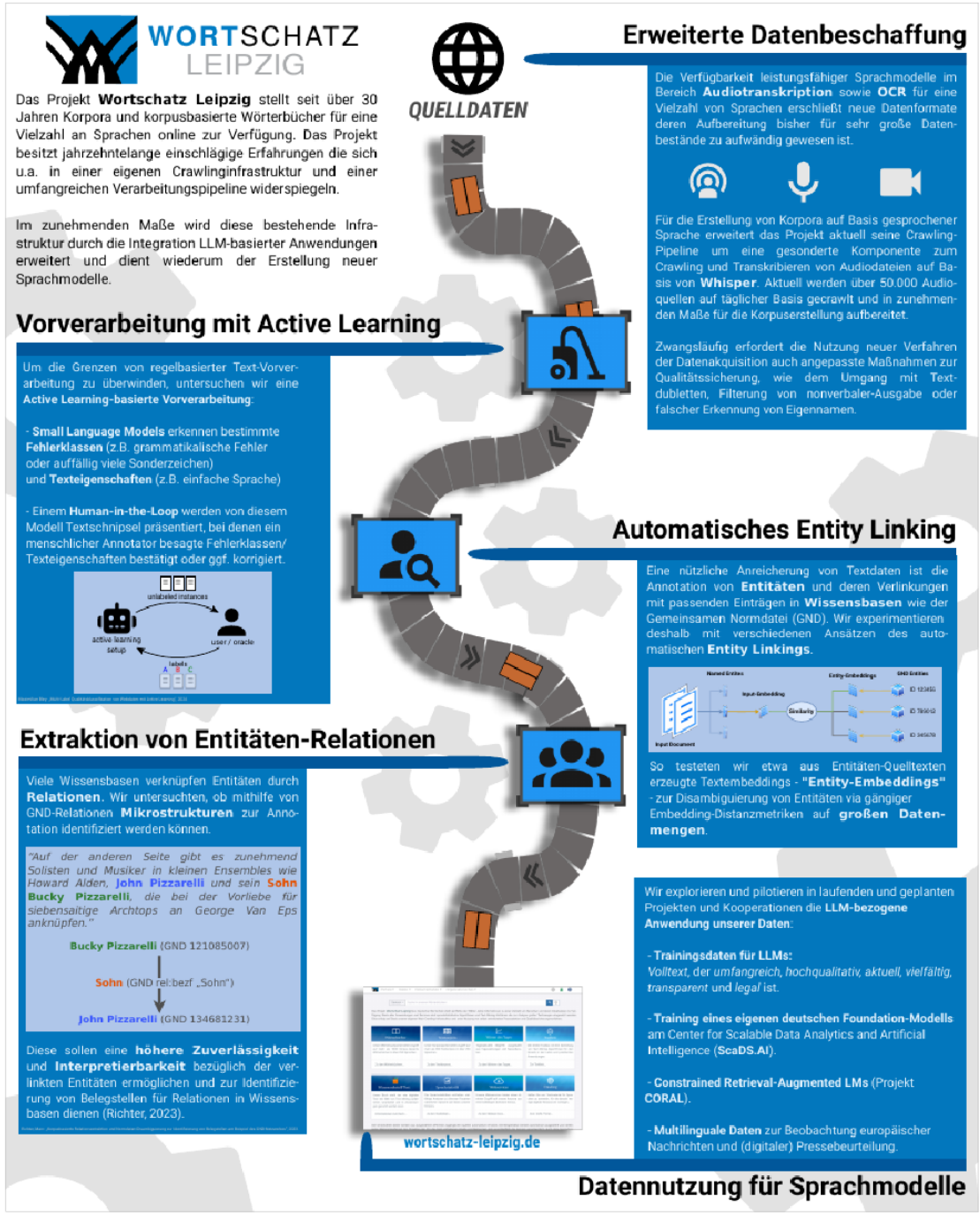
Integration von LLMs in die Verarbeitungspipeline des Wortschatz Leipzig

Thomas Eckart, Felix Helfer, Erik Körner, Frank Binder, Christopher Schröder

Sächsische Akademie der Wissenschaften zu Leipzig

UNIVERSITÄT LEIPZIG

InfAI Institut für Angewandte Informatik



Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- KI-gestützte Workflows im Umgang mit gesprochensprachlichen Daten
  Alina Hemmer
- Korpusproduktion in Zeiten großer Sprachmodelle
  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder
- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht
  Margret Mundorf
- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data
  Liudmila Olalde, Thomas Malten, Frank Grieshaber
- More uniformity and more diversity at the same time: LLMs and a 21st century standardisation paradox
  Christian Mair
- Perspektiven des Einsatzes von LLM in Text+
  Florian Barth, Philippe Genêt, Erik Körner, Peter Leinen, Pia Schwarz, Claus Zinn


Pitch

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Korpusproduktion in Zeiten großer Sprachmodelle

  Thomas Eckart, Christopher Schröder, Erik Körner, Felix Helfer, Frank Binder

- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht

  Margret Mundorf

- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data

  Liudmila Olalde, Thomas Malten, Frank Grieshaber

- More uniformity and more diversity at the same time: LLMs and a 21st century standardisation paradox

  Christian Mair

- Perspektiven des Einsatzes von LLM in Text+

  Florian Barth, Philippe Genêt, Erik Körner, Peter Leinen, Pia Schwarz, Claus Zinn

- ~~Reaching for the stars: Integration von LLMs in komplexe automatisierte Workflows~~

  ~~Daniela Schneider~~

LLOD-isierung des Madras Tamil Lexicon

Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data

Worum geht es? Um ein Tamil-Englisch Wörterbuch.

Was ist Tamil? Eine südindische Sprache nicht indoeuropäischen Ursprungs.

Wo wird Tamil gesprochen? Hauptsächlich in Südindien und Sri Lanka.

Wie viele Leute sprechen Tamil? Erstsprache von ca. 70-80 Millionen Menschen.

Was ist das Tamil Lexicon? Ein von 1924-1939 in sechs Bänden und einem Nachtragsband erschienenes historisch-literarisches Wörterbuch (Madras University).

Ist es nicht bereits online? Ja, aber in eingeschränkter Funktionalität in proprietärem Format.

Was machen wir? Eine Neumodellierung der digital vorliegenden Daten als LLOD.

Ist das alles? Nein, wir machen noch eine API.

Wer hat was davon? Alle.

Wo erfahre ich mehr davon? Am Posterstand.

தமிழ்

Liudmila Olalde - Thomas Malten - Frank Grieshaber

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Legal Linguistic Memos mit Large Language Models: Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht

  Margret Mundorf

- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data

  Liudmila Olalde, Thomas Malten, Frank Grieshaber

- More uniformity and more diversity at the same time: LLMs and a 21st century standardisation paradox

  Christian Mair

- Perspektiven des Einsatzes von LLM in Text+

  Florian Barth, Philippe Genêt, Erik Körner, Peter Leinen, Pia Schwarz, Claus Zinn

- ~~Reaching for the stars: Integration von LLMs in komplexe automatisierte Workflows~~

  ~~Daniela Schneider~~

- SwineBad: Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger

  Steffen Steiner, Frank Krüger

Christian Mair

Uniformity and Diversity at the Same Time: LLMs and the 21st Century Standardisation Paradox

- Die Geisteswissenschaften sollen aufhören, große Sprachmodelle nur mit Sorgenfalten zu betrachten.
- Sie schenken uns fantastische neue Forschungsthemen:
  - LLMs und Sprachplanung für intelligente Mehrsprachigkeit
  - LLMs und Diskriminierung
  - LLMS und die kritische Ethnographie der *clickworker sweatshops* von den Philippinen bis Westafrika
  - ...
- **Und, Thema des Posters:**
  - Wird die britische Schriftnorm die LLMs überleben?

https://www.etsy.com/listing/1230349109/hanging-union-jack-united-kingdom-draped -->

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- LLOD-isierung des Madras Tamil Lexicon: Modellierung eines Wörterbuchs einer in der IT-basierten Linguistik unterrepräsentierten außereuropäischen Sprache als Linguistic Linked Open Data

  Liudmila Olalde, Thomas Malten, Frank Grieshaber

- More uniformity and more diversity at the same time: LLMs and a 21st century standardisation paradox

  Christian Mair

- Perspektiven des Einsatzes von LLM in Text+

  Florian Barth, Philippe Genêt, Erik Körner, Peter Leinen, Pia Schwarz, Claus Zinn

- ~~Reaching for the stars: Integration von LLMs in komplexe automatisierte Workflows~~

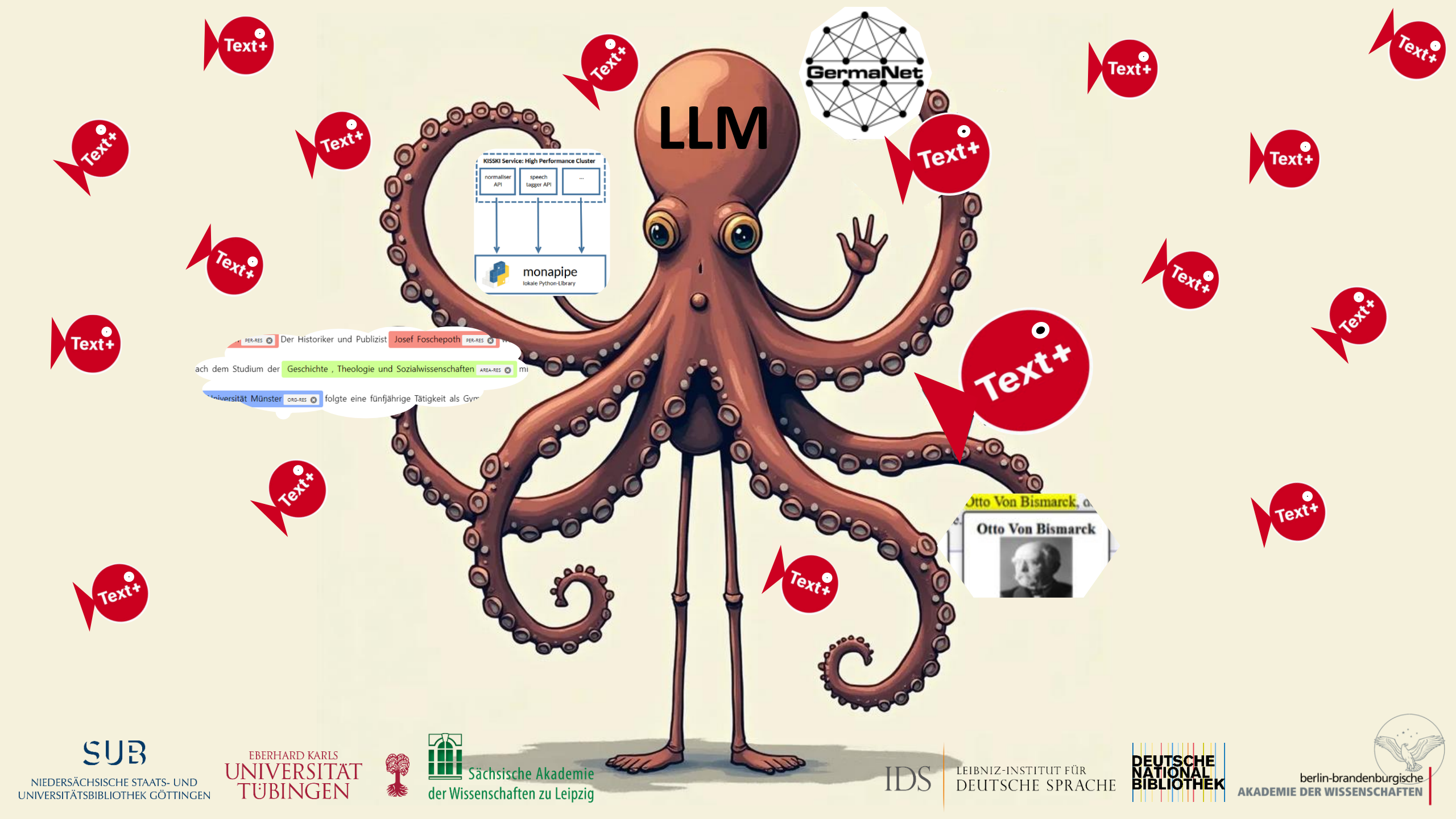
  ~~Daniela Schneider~~

- SwineBad: Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger

  Steffen Steiner, Frank Krüger

- SwissGB4Science - ein Volltext Korpus für die Forschung

  Eric Dubey, Matteo Lorenzini, Martin Reisacher, Tim Rüdiger



SUB
NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sächsische Akademie
der Wissenschaften zu Leipzig

IDS
LEIBNIZ-INSTITUT FÜR
DEUTSCHE SPRACHE

DEUTSCHE
NATIONAL
BIBLIOTHEK

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Perspektiven des Einsatzes von LLM in Text+

  Florian Barth, Philippe Genêt, Erik Körner, Peter Leinen, Pia Schwarz, Claus Zinn

- ~~Reaching for the stars: Integration von LLMs in komplexe automatisierte Workflows~~

  ~~Daniela Schneider~~

- SwineBad: Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger

  Steffen Steiner, Frank Krüger

- SwissGB4Science - ein Volltext Korpus für die Forschung

  Eric Dubey, Matteo Lorenzini, Martin Reisacher, Tim Rüdiger

- Synthetische Datensätze in der CLS

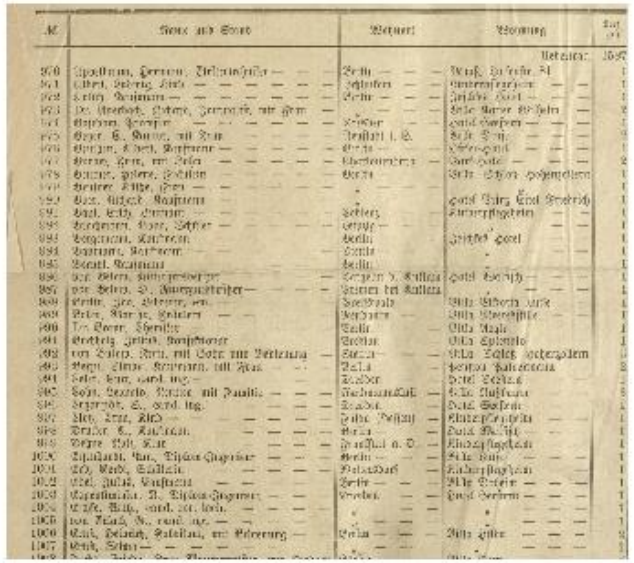
  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann

- Text+ LLM Service

  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder

SwineBad

Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger




```
{
    "Nummer": "970",
    "Vorname": "Hermann",
    "Nachname" : "Appelbaum",
    "Titel": null,
    "Beruf": "Elektrotechniker",
    "Sozialer Stand": null,
    "Begleitung" : null,
    "Wohnort": "Berlin",
    "Wohnung": "Maaß, Lotsenstr. 81",
    "Personenanzahl": "1"
},
```


- Tabellen ankommender Badegäste 1910–1932
- Software Pipeline zur Extraktion und Strukturierung

1. Segmentierung der Tabellen
2. OCR der Tabellendaten
3. LLM-basierte Korrektur und Zusammenfassung
4. LLM-basierte Strukturierung

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- ~~Reaching for the stars: Integration von LLMs in komplexe automatisierte Workflows~~

  ~~Daniela Schneider~~

- SwineBad: Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger

  Steffen Steiner, Frank Krüger

- SwissGB4Science - ein Volltext Korpus für die Forschung

  Eric Dubey, Matteo Lorenzini, Martin Reisacher, Tim Rüdiger

- Synthetische Datensätze in der CLS

  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann

- Text+ LLM Service

  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder

- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik

  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider

Swiss Google Books for Research

90 Millionen historische Seiten für die Forschung?

Universität Basel

ZB Zentralbibliothek

UNIVERSITÄT BERN

zhb Zentral- und Hochschulbibliothek Luzern

Forschung Anforderungen

Infrastruktur Realitätscheck

Szenarien

State of the Art Möglichkeiten

Daten Grenzen

NVIDIA

KEYCLOAK

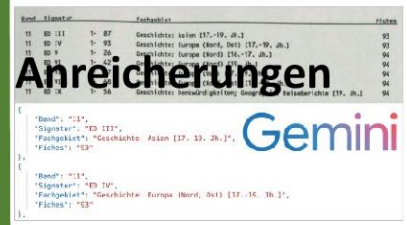
iiif

elastic

ARK Alliance

LLMs + RAG + Benchmarking

Anreicherungen



Zusammenfassungen

Unterschiedliche Modelle

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- SwineBad: Tabellenextraktion und Informationsstrukturierung aus dem Swinemünder Badeanzeiger

  Steffen Steiner, Frank Krüger

- SwissGB4Science - ein Volltext Korpus für die Forschung

  Eric Dubey, Matteo Lorenzini, Martin Reisacher, Tim Rüdiger

- Synthetische Datensätze in der CLS

  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann

- Text+ LLM Service

  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder

- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik

  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider

- “Computer, was bedeutet ‘Tiki-Taka’?” Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS

  Alexander Geyken, Gregor Middell

UNIVERSITÄT BIELEFELD

Fakultät für Linguistik und Literaturwissenschaft

Synthetische Datensätze in den Computational Literary Studies

Eines der typischen Probleme in den CLS:

- Annotieren ist schwer und dauert
- Textstellen zum Annotieren finden ist schwer und dauert

Was tun!?

Lösungen müssen her:

- Sind LLMs eine Lösung?
- Falls nein: können wir zumindest mit LLMs weitere Trainingsdaten für ein eigenes Modell erzeugen?

Falls ihr auf die letzte Frage Antworten (und ein angeregtes Gespräch) haben wollt, kommt bei unserem Poster vorbei!

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- SwissGB4Science - ein Volltext Korpus für die Forschung
  Eric Dubey, Matteo Lorenzini, Martin Reisacher, Tim Rüdiger
- Synthetische Datensätze in der CLS
  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann
- Text+ LLM Service
  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder
- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik
  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider
- "Computer, was bedeutet 'Tiki-Taka'?" Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS
  Alexander Geyken, Gregor Middell
- "fRAG Deine Daten doch selbst" – Potenziale des Einsatzes von Retrieval Augmented Generation für Forschungsdaten und Forschungsdateninfrastrukturen
  Timm Lehmberg

Task Area Infrastructure/Operations

LLM Service: Text+ AI Assistant

in progress

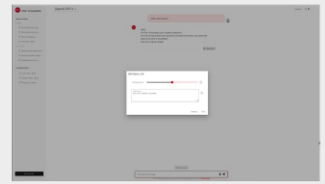
AI as a Service

Text- and language-based humanities offer extensive use-cases for Large Language Models (LLMs). Through GWDG, an additional web service will be made available on the Text+ website providing access to open-source and custom fine-tuneable LLMs, including:

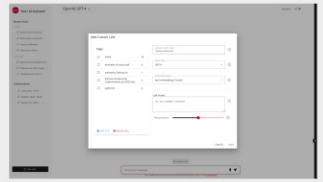
- (Meta) LLaMA
- Mixtral
- Qwen
- Codestral
- Chat-GPT

Interface

Base LLM selection, add, remove



Edit base LLM



Add custom LLM

Sources section in RAG

Features

- Free use of various open-source models
- Users can create, edit and delete custom LLMs
- Collaborations section allows users to invite collaborators to chat with the custom LLMs
- Retrieval-augmented generation (RAG) on personal documents (currently .pdf, .txt, .csv)
- Sources section on generated answers for users to check and enable citations
- Compliance with legislative requirements and user privacy interests
- No data leakage
- With Open AI's Chat-GPT, no single user related data is externally transmitted, as the current implementation makes all users appear as one
- Currently, the assistant is available for project participants who log in via Academic Cloud

The AI Assistant is based on GWDG's LLM Service whose architecture and features are described in detail in these articles:

- Decker, Hosseini, Meisel: "The New GWDG LLM Service", GWDG Nachrichten 03/2024.
- M. Azampour: "LLM Services and the AI Act of the EU", GWDG Nachrichten 04-05/2024.

A plethora of useful information an AI adoption can be sourced in this white paper:

- https://docs.kisski.de/white_paper/Barriers_to_AI_adoption_Report.pdf

Use Cases

- Retrieval Augmented Generation (RAG)
- Entity Linking
- Enrichment of GermaNet
- Query generation for search strings
- Data preprocessing for NER model
- Historical normalizations
- Runtime environment

In this poster session, the respective task force should exhibit a poster specifically on this topic:

Feedback

We believe in agile development. As such, the Text+ AI Assistant is a first installment of a service to our user base. Currently, it does not meet all of our requirements, but shows the way to the (future) use cases.

We depend on you to deliver feedback.

Try it out, and let us know about your experiences and suggestions directly or via the Text+ helpdesk !

text-plus.org/ai-assistant

DFG Deutsche Forschungsgemeinschaft

Text+ is funded by the German Research Foundation (DFG) – 460033370.

Text+ is part of the nfdi Nationale Forschungsdaten Infrastruktur

LLM services provided by GWDG

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Synthetische Datensätze in der CLS

  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann

- Text+ LLM Service

  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder

- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik

  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider

- “Computer, was bedeutet ‘Tiki-Taka’?” Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS

  Alexander Geyken, Gregor Middell

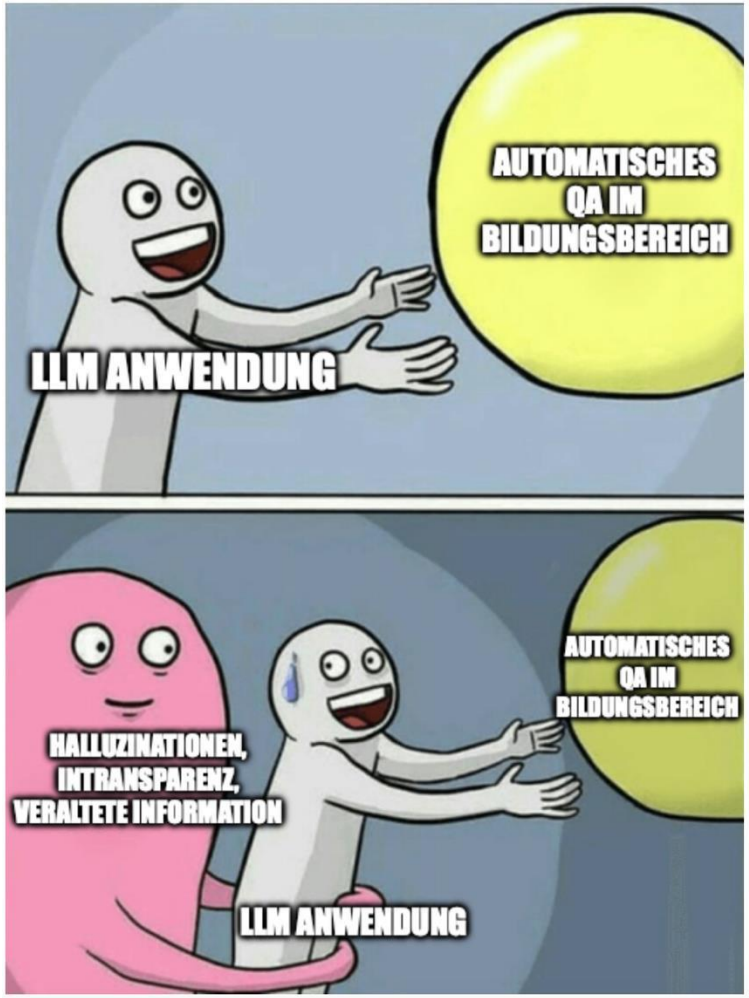
- “fRAG Deine Daten doch selbst” – Potenziale des Einsatzes von Retrieval Augmented Generation für Forschungsdaten und Forschungsdateninfrastrukturen

  Timm Lehmberg

- “Nun sag', wie hast du's mit den LLMs?” – Antworten der Text+ Community auf die Gretchenfrage

  Stine Ziegler, Philippe Genêt

WISSEN, WEN MAN FRAGT…



- **Ausgangslage:** Auch in der Domäne Grammatik mildert RAG Probleme wie Halluzinationen ab und verbessert die Antwortqualität. Dies gilt jedoch **nicht für alle Fragetypen** (vgl. Lang et al. 2024).

- **Strategie:** Dynamische **Steuerung** der Antwortprozesse in Abhängigkeit des Fragetyps

- **Umsetzung:** Prototyp einer Agentic RAG-Chain: Generierendes LLM fungiert u. a. als **Router** zur passgenauen Anbindung externer Ressourcen für unterschiedliche Fragetypen.

https://imgflip.com/memegenerator; Zugriff 01.10.2024

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Synthetische Datensätze in der CLS

  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann

- Text+ LLM Service

  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder

- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik

  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider

- “Computer, was bedeutet ‘Tiki-Taka’?” Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS

  Alexander Geyken, Gregor Middell

- “fRAG Deine Daten doch selbst” – Potenziale des Einsatzes von Retrieval Augmented Generation für Forschungsdaten und Forschungsdateninfrastrukturen

  Timm Lehmberg

- “Nun sag', wie hast du's mit den LLMs?” – Antworten der Text+ Community auf die Gretchenfrage

  Stine Ziegler, Philippe Genêt

LLMs für Wörterbuchdefinitionen

Alexander Geyken und Gregor Middell (BBAW)

Aufgabe: definiere *Mausohr*

1. eine Pflanze mit herzförmigen Blättern, die an die Form von Mäuseohren erinnern. (ChatGPT)

Damit zusammenhängende Fragen:

- Halluzination oder nicht?
- Erkennen das die Nutzerinnen und Nutzer?
- Gibt es leichtere oder schwerere Wörter für LLMs (Kriterien)?
- Prompting Strategien und weitere Optimierungen für LLMs?

↪ Poster: “Computer, was bedeutet Tiki-Taka?”

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Synthetische Datensätze in der CLS
  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann
- Text+ LLM Service
  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder
- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik
  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider
- "Computer, was bedeutet 'Tiki-Taka'?" Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS
  Alexander Geyken, Gregor Middell
- "fRAG Deine Daten doch selbst" – Potenziale des Einsatzes von Retrieval Augmented Generation für Forschungsdaten und Forschungsdateninfrastrukturen
  Timm Lehmberg
- "Nun sag', wie hast du's mit den LLMs?" – Antworten der Text+ Community auf die Gretchenfrage
  Stine Ziegler, Philippe Genêt

OpenAI. (2024). *Image generated using DALL-E, October 2024.* Retrieved from OpenAI platform.

Agenda

Collections
Lexical Resources
Editions
Infrastructure/ Operations

- Synthetische Datensätze in der CLS
  Daniel Kababgi, Emilie Sitter, Robin Martin Aust, Marie-Christine Boucher, Berenike Herrmann
- Text+ LLM Service
  Alexander Steckel, Umut Basaran, Stefan Buddenbohm, Maik Wegener, Philipp Wieder
- Wissen, wen man fragt – Agentic RAG für Automatisches Question Answering in der Domäne deutscher Grammatik
  Christian Lang, Ngoc Duyen Tanja Tu, Roman Schneider
- "Computer, was bedeutet 'Tiki-Taka'?" Eine Studie zur Generierung von Definitionsparaphrasen für Bedeutungswörterbücher am Beispiel des DWDS
  Alexander Geyken, Gregor Middell
- "fRAG Deine Daten doch selbst" – Potenziale des Einsatzes von Retrieval Augmented Generation für Forschungsdaten und Forschungsdateninfrastrukturen
  Timm Lehmberg
- "Nun sag', wie hast du's mit den LLMs?" – Antworten der Text+ Community auf die Gretchenfrage
  Stine Ziegler, Philippe Genêt

„Nun sag‘, wie hast du‘s mit denn LLMs?“

Antworten der Text+ Community auf die Gretchenfrage

Text+

Vielen Dank an alle Vortragenden und Zuhörenden!

Die Postersession findet ab sofort bis 18:00 Uhr in O 138 (Fuchs-Petrolub-Saal) statt.

Collections
Lexical Resources
Editions
Infrastructure/ Operations

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Förderungsnummer 460033370

Teil der
nfdi Nationale Forschungsdaten Infrastruktur

Die vorliegende Präsentation wurde im Rahmen des Konsortiums Text+ im Kontext der Arbeit des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. verfasst. NFDI wird von der Bundesrepublik Deutschland und den 16 Bundesländern finanziert, und das Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370. Die Autor:innen bedanken sich für die Förderung sowie Unterstützung. Ein Dank geht außerdem an alle Einrichtungen und Akteur:innen, die sich für den Verein und dessen Ziele engagieren.