

GeMTeX: Ein reales klinisches Textkorpus für datengetriebene medizinische Sprachverarbeitung

Worum geht's?

Developing the German Medical Text Corpus (GeMTeX): Legal Compliance and Semantic Enrichment

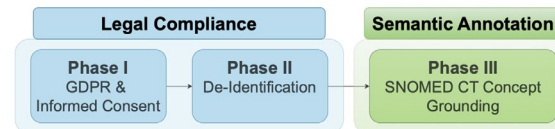
Justin Hofenbitzer¹, Christina Lohr¹, Andrea Riedel^{1,3}, Rebekka Kiser¹,
Allaksandra Shutsko^{1,4}, Abanoub Abdelmalak^{1,5}, Peter Klügl¹, Jutta Romberg^{2,6},
Sarah Riepenhausen¹, Miriam Schechner¹, Jakob Falter^{1,3}, Frank Meineke¹,
Luise Modersohn¹, Markus Löffler¹, Juliane Fluck^{1,5}, Udo Hahn¹,
Stefan Schulz^{1,5}, Martin Boeker¹

¹Technical University of Munich, Germany, ²Leipzig University, Germany,
³FAU Erlangen-Nürnberg, Germany, ⁴Uniklinikum Erlangen, Germany, ⁵Zg MED, Germany,
⁶Averbis GmbH, Germany, ⁷University of Bonn, Germany,
⁸Charité-Universitätsmedizin Berlin, Germany, ⁹Berlin Institute of Health @ Charité, Germany,
¹⁰University of Münster, Germany, ¹¹LMU University Hospital Munich, Germany,
¹²MedUni Graz, Austria

justin.hofenbitzer@tum.de, martin.boeker@tum.de



- Einführung des größten deutschsprachigen klinischen Textdatensatz: **German Medical Text Corpus (GeMTeX)**
- **Rechtliche** Rahmenbedingungen & **De-Identifizierung**
- **Semantische Annotation** mit SNOMED CT
- Aktuelle **Korpuszahlen** & **Nutzungsmöglichkeiten**



Towards conversational diagnostic artificial intelligence

[Tao Tu](#) , [Mike Schaekermann](#) , [Anil Palepu](#), [Khaled Saab](#), [Jan Freyberg](#), [Ryutaro Tanno](#), [Amy Wang](#), [Brenna Li](#), [Mohamed Amin](#), [Yong Cheng](#), [Elahe Vedadi](#), [Nenad Tomasev](#), [Shekoofeh Azizi](#), [Karan Singhal](#), [Le Hou](#), [Albert Webson](#), [Kavita Kulkarni](#), [S. Sara Mahdavi](#), [Christopher Semturs](#), [Juraj Gottweis](#), [Joelle Barral](#), [Katherine Chou](#), [Greg S. Corrado](#), [Yossi Matias](#), ... [Vivek Natarajan](#)  [+ Show authors](#)

[Nature](#) **642**, 442–450 (2025) | [Cite this article](#)

ORIGINAL ARTICLE

Chatbots, Agents, LLMs: **Sprachverarbeitung** wird in der Medizin immer relevanter

Präzise Systeme brauchen **echte klinische** Daten

or LLM-Based Medical Consultation
t Collaboration

awen Duan^{1*}, Jianxin Wang¹,

CURRENT ISSUE ▾ RECENTLY PUBLISHED university

.csu.edu.cn

f X in e w

Generative AI Chatbot

udeau, B.A. , Sukanya Bhattacharya,
gail J. Salzhauer, B.A. , Tess Z.
s
VOL. 2 NO. 4 | Copyright © 2025

📖 © 📄 🔍

ic Medical

Dialogue with Grounded Reasoning

An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial

[Xinge Tao](#), [Shuya Zhou](#), [Kai Ding](#), [Sairan Li](#), [Yanzeng Li](#), [Boyou Wu](#), [Qirui Huang](#), [Wangyue Chen](#), [Muzi Shen](#), [En Meng](#), [Xiaowang Chen](#), [Hong Hu](#), [Jinchao Zhang](#), [Jie Zhou](#), [Lei Zou](#), [Libing Ma](#)  & [Shasha Han](#) 

[Nature Medicine](#) **32**, 934–942 (2026) | [Cite this article](#)

Yuyuan Liu
University of Oxford

le Wu
of Oxford

Klinische Textkorpora:

- Die meisten Realkorpora sind **Englisch (US)**:
 - MIMIC (Johnson et al., 2023)
 - i2b2 & n2c2 corpora (Kumar et al., 2015, Stubbs et al., 2019)
 - SemEval corpora (e.g., Pradhan et al., 2014)

- **Deutsche Realkorpora:**

Referenz	# Doks	# Tokens
Richter-Pechanski et al., 2023	500	993k
Lenz et al., 2025	150	31k
Dewald et al., 2023	1.9k	200k
Kittner et al., 2021	150	75k

Semantisch Annotierte Korpora:

- **Englisch:**
 - MCN Corpus (Luo et al., 2019)
 - SNOMED CT Linking Challenge (Davidson et al., 2025)
- **Schwedisch:** Skeppstedt et al. (2012)
- **Spanisch:** González-Agirres et al. (2019)
- **Mehrsprachig:** Miranda-Escalada et al. (2022)
- **Deutsch:** Borchert et al. (2022)

German Medical Text Corpus (GeMTeX):

- Ein Projekt, um das **größte deutschsprachige klinische Korpus** zu erstellen
- **Nationales** Konsortium innerhalb der **MII**

- **Annotationspartner**

- *TUM Klinikum*
- *Universitätsklinikum Leipzig*
- *Uniklinikum Erlangen*
- *Charité Berlin*
- *Universitätsklinikum Essen*
- *Universitätsklinikum Carl Gustav Carus Dresden*

- **Methodische Partner**

- *ID GmbH & Co. KGaA*
- *Technische Universität Darmstadt*
- *Averbis GmbH*
- *MHH Hannover*
- *Universitätsklinikum Heidelberg*
- *ZB Med*
- *LMU Klinikum München*
- *Universität Münster*
- *Hasso Plattner Institut*
- *Universitätsklinikum Tübingen*
- *MedUni Graz*



Rechtlicher Rahmen & De-Identifizierung

Rechtlicher Rahmen

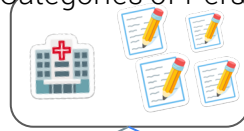
MII Broad Consent

Bild et al. (2020); Zenker et al.
(2024a,b)

Pseudonymisierung

Anonymisierung

Personally Identifiable Information
(PII)
Special Categories of Personal Data



Gesetzes-
grundlage

Informierte
Einwilligung

Safeguards

De-
Identifizierungs-
prozess

Lohr et al. (2024,
2025)

Art. 9(1) DSGVO

Art. 9(2) DSGVO

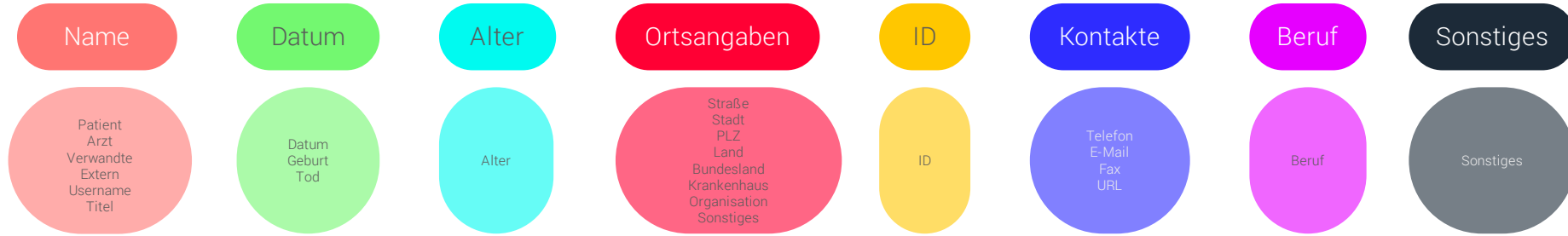
Art. 89 DSGVO

+ Art. 10 EU AI Act

De-Identifizierung

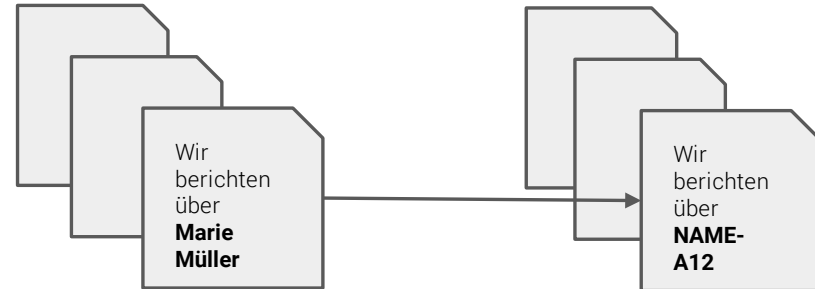
Annotation - De-Identifizierung:

- Erstes, umfassendes deutschsprachiges **De-Identifizierungsschema** (Lohr et al., 2024, 2025)
- Basierend auf **HIPAA**



Ersetzung:

- Relevante Spannen werden **pseudonymisiert** & **semantisch ersetzt**
- Eigens entwickeltes Tool **"Surrogator"**



De- Identifizierung

- Was bringt die De-Identifizierung?
 - GeMTeX als klinische **Referenztextressource**
 - **Breite** Nutzbarkeit
 - **Domänenadaption** möglich
- Aber: Welche Informationen stecken in den klinischen Texten?
 - **Semantische Annotation** mit **SNOMED CT**
 - KI-Modelle sollen in der Lage sein, **medizinisch-relevante Entitäten** in den Texten zu erkennen

Semantische Annotation mit SNOMED CT

- **Ontologie-gestützte** Annotation mit SNOMED CT
 - ~ 370k Konzepte
 - **Sprachunabhängige** Nutzbarkeit der Annotate
- Semantische **Annotationsleitlinien**:
 - Einteilung der Konzepte in **drei Konzeptgruppen**
→ *Core, Modifier & Qualifier Konzepte*
 - Sechs **Annotationsmaximen**
→ z.B. Concept before Structure oder In Doubt? Be Broad?



Annotationsrichtlinie:



Core Konzepte

Osteoarthritis of right knee joint (disorder)

Wir stellen Arthrose am re. Knie fest.

Epidermal burn of head (disorder)

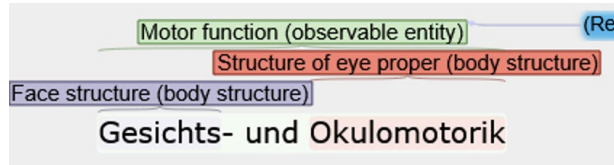
Der Kopf ist erstgradig verbrannt.

Visual field defect (finding)

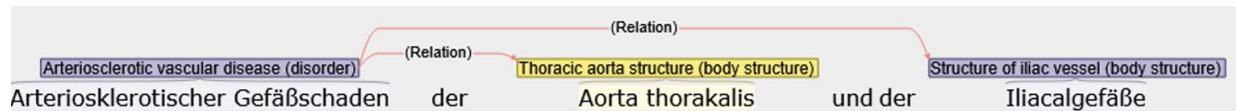
Gesichtsfelddefekte,

- **“Vollwertiges”** klinisches Konzept
→ = *klinischer Fakt*
- **Keine Kombination** mit anderen Konzepten
notwendig
- Beispiele:
 - Prozeduren
 - Konditionen (Befunde, Erkrankungen, ...)
 - Beobachtungen
 - Events
 - Medikation

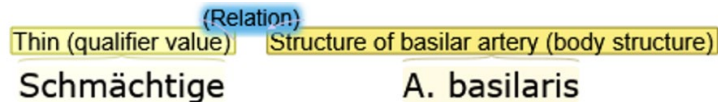
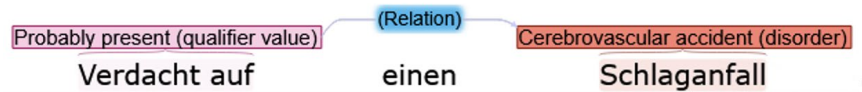
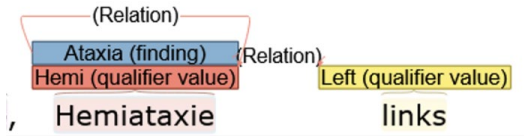
Modifizier Konzepte



- **Nicht** die zentrale klinische Information
→ *Oft Kombination notwendig*
- Spezifische Information zu **strukturellen**, **physiologischen** & **biologischen** Eigenschaften
- Beispiele:
 - Körperteile
 - Physische Objekte
 - Organismen



Qualifier Konzepte



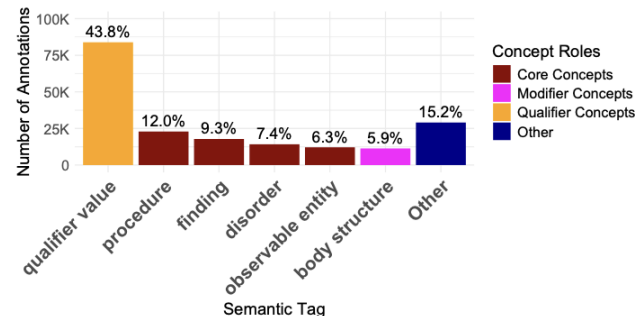
- **Kontextualisierung** von klinischen Fakten
- Oft **natürlichsprachliche Entsprechung** zu gängigen Adjektiven
- Beispiele:
 - Einheiten
 - Temporalinformation
 - Faktualitäten
 - Negation

Annotationsprozess:

- **INCEpTION** Annotationsplattform (Klie et al., 2018; Eckart de Castilho et al., 2024)
- Industrielle **Vorannotation** (Averbis Health Discovery) & **Recommender** (ID Berlin)
- Ein **geschulter Annotator** pro Dokument
- **1%** aller Dokumente werden von **mehreren** Annotatoren annotiert
- GraSCCo Korpus (Modersohn et al., 2022) wird **über Standorte hinweg** annotiert

Inter-Annotator Agreement & Gold Standard

- **Synthetisches Goldstandarddokument** *Amanda Alzheimer* aus GraSCCo
- 3.7k Annotationen (8 Annotatoren)
- Konservatives **Spannen-basiertes Agreement** mit Bootstrapping:
 - Konzeptebene: $\alpha = 0.65$ [0.63, 0.68]
 - Semantic Tag Ebene: $\alpha = 0.71$ [0.69, 0.74]



Zum Download des
Goldstandarddokuments:



Semantische Annotation

- **Umfassende** Semantische Annotation
→ *Ca. 250-300 Konzepte pro 1000 Tokens*
- Downstream Tasks:
 - Automatisches klinisches Coding
 - Knowledge-Retrieval
 - Negations- & Faktualitätsanalysen
 - Frage-Antwort Systeme
 - ...

Aktuelle Zahlen: Wie groß ist GeMTeX?

De-Identifizierung

>18.6k
Dokumente

>26M Tokens

>680k Annotationen

Semantische Annotation

>1.3k
Dokumente

>2.3M
Tokens

>556k Annotationen

Schon jetzt: Das größte deutschsprachige klinische Textkorpus!

Accessibility: Wie kann GeMTeX für Forschung genutzt werden?

- GeMTeX ist ein **verteiltes Korpus**
- **Strukturierte Daten** sind zusätzlich zu den Texten ausleitbar über etablierte MII Prozesse
 - **Schritt 1:** Studienprotokoll
 - **Schritt 2:** Ethikvotum
 - **Schritt 3:** Datennutzungsantrag via *Forschungsdatenportal Gesundheit* (FDPG)



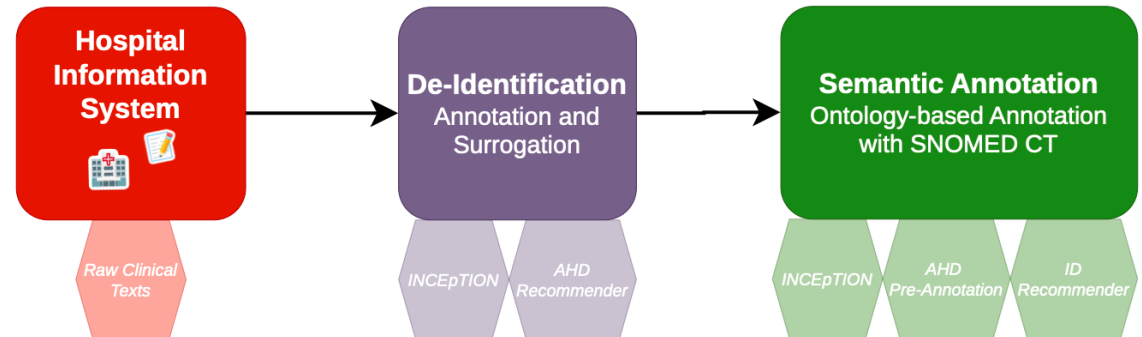
Wir unterstützen bei Nutzungs- & freuen uns über Kooperationsanfragen!



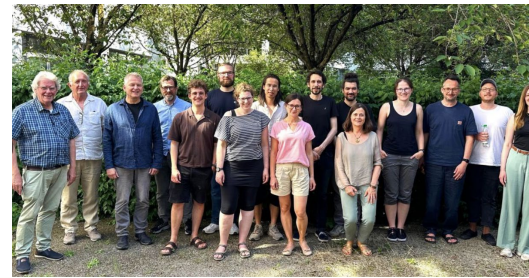
gemtex.mi@mh.tum.de

GeMTeX: Ein reales klinisches Textkorpus für datengetriebene medizinische Sprachverarbeitung

- **Größtes** deutschsprachiges klinisches Textkorpus
- **De-Identifizierung** mit semantischer Ersetzung
- Ontologie-gestützte **Semantische Annotation** mit SNOMED CT
- Zugang über **etablierte Prozesse** der MII (FDPG)



**Vielen Dank für Ihre
Aufmerksamkeit!**



GeMTeX Projekt

– Modul 3 Projekt –

PIs:

Martin Boeker (TUM), Markus Loeffler (UKL)

Arbeitspaketleitung:

Justin Hofenbitzer (TUM), Christina Lohr (UKL),

Luise Modersohn (TUM), Frank Meineke (UKL)



Beteiligte Standorte:

Technische Universität München, Universität Leipzig, Uniklinikum Erlangen, Charité Berlin, Universitätsklinikum Essen, Universitätsklinikum Carl Gustav Carus Dresden, ID GmbH & Co. KGaA, Technische Universität Darmstadt, Averbis GmbH, MHH Hannover, Universitätsklinikum Heidelberg, ZB Med, LMU Klinikum München, Universität Münster, Hasso Plattner Institut, Universitätsklinikum Tübingen, MedUni Graz

With funding from the:



**Federal Ministry
of Research, Technology
and Space**