

Natural language processing and language models for Dutch clinical text: a systematic review

Artuur M. Leeuwenberg and Ruurd Kuiper

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht,
Utrecht University, Utrecht, the Netherlands

Preprint of preliminary results

Abstract

Background. Increasingly natural language processing (NLP) tools and applications - including those using large language models (LLMs) - are developed and used in electronic health records (EHRs). As generalization to specific languages and EHR settings or departments is not guaranteed, having a comprehensive overview about tool availability and accuracy across EHR settings is increasingly important for their effective application and reuse of existing tools. No such overview was available for Dutch, focused on language technology for EHRs that covers the past decade.

Objective. To identify and describe existing NLP tools, including those using LLMs, that have been developed or evaluated in real-world Dutch EHRs.

Methods. A literature search was conducted in Scopus and Pubmed up to November 11, 2025. Information about the NLP task, patient group, healthcare setting, code and model availability were extracted.

Results. A total of 44 studies were included, describing 794 models and 792 evaluations. Most studies focused on information extraction (73%), followed by de-identification (14%), generative applications (11%), and language modeling tasks (9%). Rule-based methods were most frequently used at the study level (50%), while transformer-based approaches accounted for the majority of models and evaluations (55%). Prompting LLMs was used in 16% of studies and accounted for 32% of models and evaluations. Code was shared in 43% of studies, covering 91% of models, whereas only 4% of models were publicly available.

Conclusions. A diverse set of NLP models has been developed for Dutch clinical text, with an increasing use of transformer-based and LLM-based approaches. However, model availability remains limited. This review provides a structured overview of available tools and evaluations to support their application and reuse in Dutch clinical settings.

Introduction

Natural language processing, including large language models (LLMs), have large potential to at scale utilise free text recorded in electronic health records (EHRs) for healthcare research and practice. Most current models are developed and evaluated predominantly in English health data.¹⁻³ As their generalizability to non-English and especially clinical settings is not guaranteed⁴⁻⁶, before applying NLP tools to practical clinical or research settings, knowing what NLP tools and models are available and have been evaluated in relevant data to each local setting is critical.⁷ The - to our knowledge - last review on NLP tools for Dutch clinical care text was conducted more than a decade ago⁸, after which there have been significant advances in the field. Therefore, the current study systematically reviews what NLP models, including LLMs, have been developed for or evaluated in Dutch EHR texts.

Methods

Search

Scopus and PubMed were searched for studies for studies including up to November 11, 2025. The search string (**Supplement A**) was a conjunction of search terms related to (1) NLP tools or tasks, (2) the use of patient care data, and (3) the Dutch setting. Arxiv and medRxiv were searched for preprints that appeared in 2025. Additional articles and preprints known by the authors were added.

Selection and eligibility

Eligibility was screened after deduplication by one reviewer (AL). Studies were eligible for inclusion if they described the development, evaluation or application of an NLP model in original Dutch patient care data (i.e., text created in the process of the care process). This includes models for named entity recognition, relation extraction, text normalisation (abbreviation resolution, spelling correction), parsing, text classification, concept extraction, automatic coding, and summarization. Also text representation or preprocessing models were included (e.g., language or word embedding models, and de-identification tools). Prognostic models using free text input were excluded. Both peer reviewed scientific articles and preprints were included. Studies were explicitly excluded if they were conducted published before 2010 or only used non-original Dutch care data (e.g., translated non-Dutch EHR data or vignettes).

Data extraction and analysis

For each included study, data were extracted about each model presented in the study by two reviewers (AL, RK) on (1) the NLP task (2) model type, (3) code and model availability, and (4) the healthcare setting and patient group and size of the data used for development, evaluation or application.

Table 1. Information collected from included articles.

Data extracted	Description
NLP task	Categorization of NLP usages: <ul style="list-style-type: none"> • Information extraction: the filling of values for predefined structured variables (e.g., smoking status) based on their presence in the free text. • De-identification: the removal (or replacement with a placeholder) of identifying information occurrences in the text (e.g., names). • Generative applications: NLP used to generate text as output (e.g., EHR summarization). • Language normalisation, representation and modelling: NLP models to preprocess or represent free text (e.g., spelling correction, word vector or language models).
Model architecture	The type of model architecture used, categorized into: <ul style="list-style-type: none"> • Rule-based systems (e.g., regular expressions) • Traditional machine learning (e.g., random forests) • Non-transformer-based deep learning (e.g., long-short term memory networks) • Fine-tuned or pre-trained transformers (e.g., BERT models) • Prompting LLMs (e.g., chatGPT) • Ensembles (e.g., combinations of architectures) • Others
Code availability	Code to develop, evaluate or apply the models is shared online.
Model availability	The model itself is shared online (e.g., parameters, rules).
Model development, evaluation or application data	Details about the data used: <ul style="list-style-type: none"> • Patient inclusion criteria • Sample size • Healthcare setting (e.g., primary care, hospital care) • Geographical region (e.g., Groningen, Rotterdam, Antwerp) • Text types (e.g., discharge summaries, radiology reports)

For each of the data categories descriptive statistics about the different extracted data categories are calculated, on the level of individual models as well as articles. And interactive overviews are constructed, to facilitate model identification.

Results

In total, 1,044 references were screened, of which 44 studies were included in the review, describing 794 models and tools and 792 evaluations. The study selection process is presented in Figure 1.

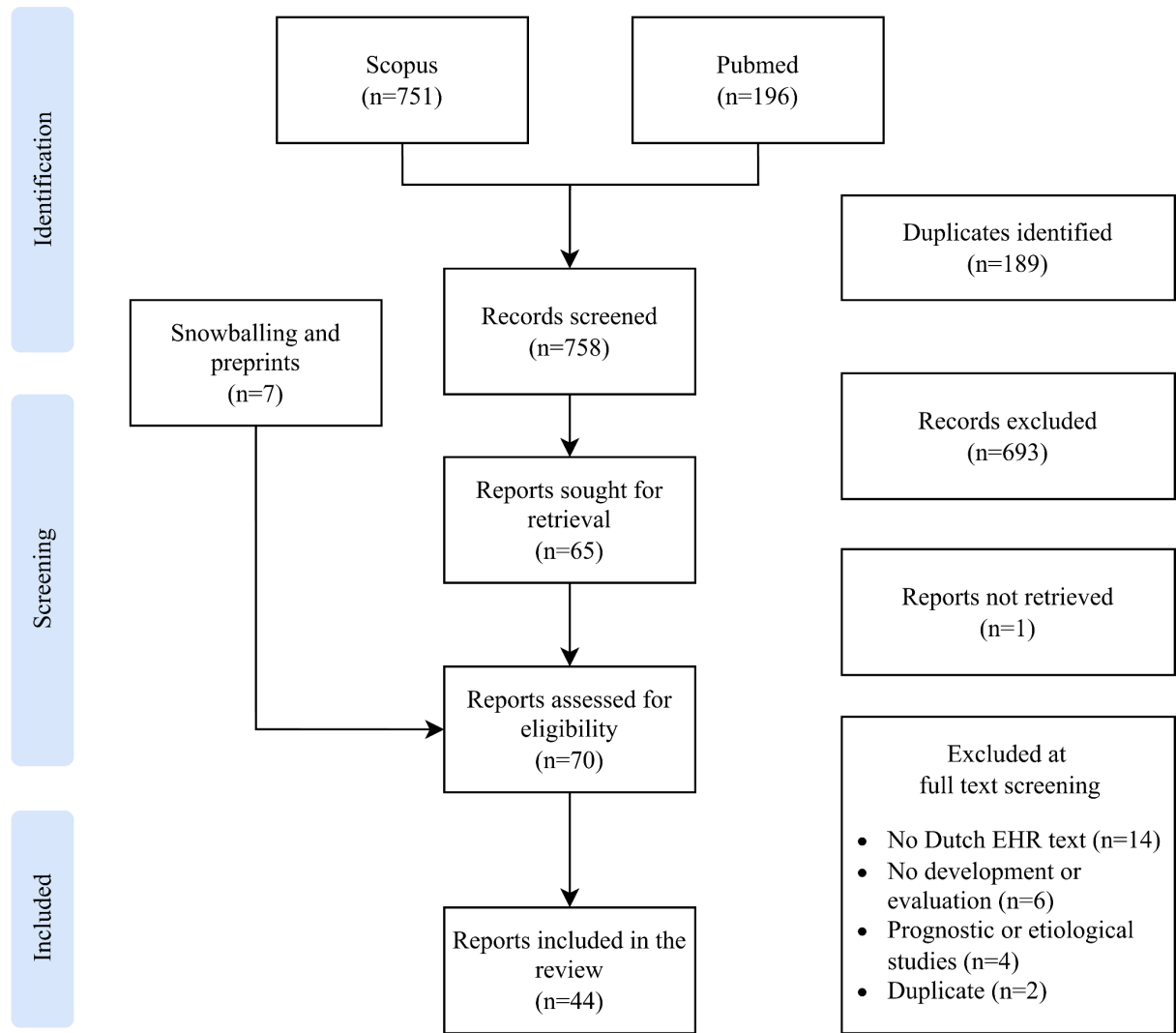


Figure 1. Flow diagram of selected articles

Most studies addressed information extraction (32/44, 73%), covering 744 models and 741 evaluations. De-identification was addressed in 6 studies (14%), covering 31 models and 33 evaluations. Generative applications were described in 5 studies (11%), covering 4 models and 3 evaluations. Language normalization, representation, and modeling were addressed in 4 studies (9%), covering 16 models and 15 evaluations.

A range of modeling architectures was applied across studies. Rule-based approaches were used in 50% of studies (22/44), followed by fine-tuned or pre-trained transformer models in 25% (11/44). Traditional machine learning methods (23%, 10/44) and non-transformer deep learning approaches (21%, 9/44) were also applied. Prompting-based large language models were used in 16% of studies (7/44). At the model level, transformer-based approaches accounted for 55% of models and evaluations, while prompting-based large language models accounted for 32%. Ensemble and other approaches were used in a small number of studies ($\leq 7\%$).

A variety of EHR text types were used. Twenty studies (46%) used general or broad descriptions of the text data (e.g., “all notes” or “EHR data”). Fourteen studies (32%) used radiology reports, 6 studies (14%) primary care notes, 5 studies (11%) discharge documentation, 3 studies (7%) echocardiography reports, and 2 studies (5%) pathology reports.

Patient populations were heterogeneous. Most studies focused on disease-specific cohorts (29/44, 66%), although these were fragmented across conditions. In addition, 22 studies (50%) did not clearly specify the patient population. Among specified groups, cardiovascular (7/44, 16%) and oncology populations (5/44, 11%) were most frequently represented.

All studies used Dutch-language clinical data. Most studies were conducted in the Netherlands (41/44, 93%), with a small number including data from Belgium (3/44, 7%). Data were obtained from multiple university medical centers, regional hospitals, and primary care networks.

Code was shared in 43% of studies (19/44), covering 91% of models. However, only 4% of the actual models were publicly available.

A model overview dashboard is provided at: <http://tuur.github.io/files/dutchnlpllmoverview.html>. A list of the shared models, including their repositories or location is provided in Supplement B.

Discussion

This review provides a comprehensive overview of NLP models and tools developed for Dutch clinical text over the past decade. Most studies focused on information extraction tasks, with relatively limited work on generative applications and language modeling. This reflects the continued emphasis on structured data extraction from EHRs.

A wide range of modeling approaches was observed. Rule-based methods were most frequently used at the study level, while transformer-based approaches accounted for the majority of models and evaluations. Prompting-based large language models were applied in a smaller number of studies but contributed substantially to the total number of models evaluated. This indicates a shift toward transformer-based and LLM-driven methods, while earlier approaches remain in use. Generative applications were relatively scarce and typically focused on summarization or text generation tasks. These studies were often exploratory in nature and relied on subjective evaluation metrics, such as Likert-scale assessments. Standardized and task-specific evaluation frameworks for generated clinical text remain limited.

Studies were conducted across a variety of healthcare settings and patient populations. However, reporting of these characteristics was often incomplete. Many studies used broadly defined or unspecified datasets, making it difficult to assess generalizability. While data were predominantly sourced from the Netherlands, studies covered multiple institutions and regions, suggesting broad national representation.

Code sharing was relatively common, but actual model sharing was rare. Although code was available for a substantial proportion of models, only a small fraction of models themselves were publicly accessible. This limits reproducibility and reuse of existing tools.

Finally, the included studies demonstrate substantial heterogeneity in tasks, datasets, and evaluation approaches. This fragmentation may hinder comparison across studies and reuse of models.

Several limitations should be considered. Study selection was performed by a single reviewer, which may have introduced selection bias. Some studies may not have been identified if NLP methods were not explicitly described in titles or abstracts. In addition, not all NLP tools developed in practice are reported in the scientific literature, particularly those developed in industry or local clinical settings.

Conclusions

A broad range of NLP models and tools has been developed for Dutch clinical text, predominantly focused on information extraction tasks. Transformer-based and LLM-based approaches account for the majority of recent models and evaluations, although rule-based and traditional methods remain in use. Generative applications are emerging but remain limited. Despite increasing methodological diversity, reporting of data characteristics and availability of models remains incomplete. While code sharing is relatively common, model sharing is rare, limiting reproducibility and reuse. This overview provides a structured summary of available models and evaluations, which may support the selection and application of NLP tools in Dutch EHR settings and guide future development.

References

1. Névél, A., Dalianis, H., Velupillai, S., Savova, G. & Zweigenbaum, P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *J. Biomed. Semantics* **9**, 12 (2018).
2. Shaitarova, A., Zaghir, J., Lavelli, A., Krauthammer, M. & Rinaldi, F. Exploring the latest highlights in medical Natural Language Processing across multiple languages: A survey. *Yearb. Med. Inform.* **32**, 230–243 (2023).
3. Klug, K. *et al.* From admission to discharge: a systematic review of clinical natural language processing along the patient journey. *BMC Med. Inform. Decis. Mak.* **24**, 238 (2024).
4. Jin, Y. *et al.* Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. in *Proceedings of the ACM Web Conference 2024* vol. 35 2627–2638 (ACM, New York, NY, USA, 2024).
5. Elvas, L. B., Almeida, A. & Ferreira, J. C. Natural language processing in medical text

- processing: A scoping literature review. *Int. J. Med. Inform.* **204**, 106049 (2025).
6. Spaanderman, D. J. *et al.* Evaluating open-weight large language models for structured data extraction from narrative medical reports across multiple use cases and languages. *arXiv [cs.CL]* (2025) doi:[10.48550/arXiv.2511.10658](https://doi.org/10.48550/arXiv.2511.10658).
 7. Gong, E. J., Bang, C. S., Lee, J. J. & Baik, G. H. Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. *J. Med. Internet Res.* **27**, e84120 (2025).
 8. Cornet, R., Van Eldik, A. & De Keizer, N. Inventory of tools for Dutch clinical language processing. *Studies in Health Technology and Informatics* vol. 180 245–249 Preprint at <https://doi.org/10.3233/978-1-61499-101-4-245> (2012).
 9. Cara, I. *et al.* Automating performance status annotation in oncology using Llama-3. *Stud. Health Technol. Inform.* **327**, 868–869 (2025).
 10. Trienes, J. *et al.* Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records. *CEUR Workshop Proceedings* vol. 2551 3–11 Preprint at <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081652100&partnerID=40&md5=f542e34a749fdd0b57e85a1041192de5> (2020).
 11. Klappe, E. S., van Putten, F. J. P., de Keizer, N. F. & Cornet, R. Contextual property detection in Dutch diagnosis descriptions for uncertainty, laterality and temporality. *BMC Med. Inform. Decis. Mak.* **21**, (2021).
 12. Menger, V., Scheepers, F., van Wijk, L. M. & Spruit, M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telemat. Inform.* **35**, 727–736 (2018).
 13. Verschueren, M. V. *et al.* Development and Portability of a Text Mining Algorithm for Capturing Disease Progression in Electronic Health Records of Patients with Stage IV Non-Small Cell Lung Cancer. *JCO Clin. Cancer Inform.* **8**, (2024).
 14. van de Burgt, B. W. M. *et al.* Development of a text mining algorithm for identifying adverse

- drug reactions in electronic health records. *JAMIA Open* **7**, ooae070 (2024).
15. Arends, B. *et al.* Diagnosis extraction from unstructured Dutch echocardiogram reports using span- and document-level characteristic classification. *arXiv [cs.CL]* (2024).
 16. Muizelaar, H., Haas, M., van Dortmont, K., van der Putten, P. & Spruit, M. Extracting patient lifestyle characteristics from Dutch clinical text with BERT models. *BMC Med. Inform. Decis. Mak.* **24**, (2024).
 17. Krastman, P. *et al.* Incidence of hand and wrist disorders in primary care: a retrospective cohort study. *BJGP Open* (2024) doi:[10.3399/BJGPO.2023.0240](https://doi.org/10.3399/BJGPO.2023.0240).
 18. Mosteiro, P., Wang, R., Scheepers, F. & Spruit, M. Investigating DE-identification methodologies in Dutch medical texts: A replication study of Deduce and Deidentify. *Electronics (Basel)* **14**, 1636 (2025).
 19. Olthof, A. W. *et al.* Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Comput. Methods Programs Biomed.* **208**, (2021).
 20. Verkijk, S. & Vossen, P. Creating, anonymizing and evaluating the first medical language model pre-trained on Dutch Electronic Health Records: MedRoBERTa.nl. *Artif. Intell. Med.* **167**, 103148 (2025).
 21. Kim, J. *et al.* Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. *2022 Language Resources and Evaluation Conference, LREC 2022* 4577–4585 Preprint at <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144431747&partnerID=40&md5=fbeealdf41612e172627bac061430653> (2022).
 22. van Es, B. *et al.* Negation detection in Dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC Bioinformatics* **24**, (2023).
 23. Nobel, J. M. *et al.* T-staging pulmonary oncology from radiological reports using natural language processing: translating into a multi-language setting. *Insights Imaging* **12**, (2021).
 24. Bosma, J. S. *et al.* The DRAGON benchmark for clinical NLP. *NPJ Digit. Med.* **8**, 289 (2025).

25. Hendrix, W. *et al.* Trends in the incidence of pulmonary nodules in chest computed tomography: 10-year results from two Dutch hospitals. *European Radiology* **33**, 8279–8288 (2023).

Supplement A Search query

Scopus

TITLE-ABS-KEY(clinical OR medical) AND (TITLE-ABS-KEY(NLP OR "language processing" OR "text mining" OR "language model") OR (TITLE-ABS-KEY(extract* OR recogni* OR detect* OR tagg* OR stemm* OR tokeniz* OR pars* OR normali* OR "de-identif*" OR "spelling correction") AND TITLE-ABS-KEY(notes OR letters OR reports OR text))) AND TITLE-ABS-KEY(Dutch) AND PUBYEAR > 2009

PubMed

((("clinical"[Title/Abstract] OR "medical"[Title/Abstract]) AND ("NLP"[Title/Abstract] OR "language processing"[Title/Abstract] OR "text mining"[Title/Abstract] OR "language model"[Title/Abstract] OR ("extract*" [Title/Abstract] OR "recogni*" [Title/Abstract] OR "detect*" [Title/Abstract] OR "tagg*" [Title/Abstract] OR "stemm*" [Title/Abstract] OR "tokeniz*" [Title/Abstract] OR "pars*" [Title/Abstract] OR "normali*" [Title/Abstract] OR "de identif*" [Title/Abstract] OR "spelling correction"[Title/Abstract]) AND ("notes"[Title/Abstract] OR "letters"[Title/Abstract] OR "reports"[Title/Abstract] OR "text"[Title/Abstract]))) AND "Dutch"[Title/Abstract]) AND (2010:2025[pdat])

Supplement B Overview of shared models

This supplement provides an overview of the studies with openly shared NLP models identified in the review, including the corresponding study, task, short description, and model location. Models are grouped by study and deduplicated at the model level.

Ref.	Model	NLP task	Description	Model location
⁹	Cara-Llama-3-classify	Information extraction	WHO performance status classification	https://github.com/IreneZelda/NLP-Performance-Status-project
⁹	Cara-Llama-3-score	Information extraction	WHO performance score regression	https://github.com/IreneZelda/NLP-Performance-Status-project
¹⁰	Trienes-BiLS TM-CRF	Deidentification	Removal of Protected Health Information	https://github.com/nedap/deidentify
¹⁰	Trienes-CRF	Deidentification	Removal of Protected Health Information	https://github.com/nedap/deidentify
¹⁰	Trienes-DED UCE	Deidentification	Removal of Protected Health Information	https://github.com/nedap/deidentify
¹¹	Klappe-RB	Information extraction	Contextual property detection: uncertainty, laterality, temporality extraction	https://github.com/evaklappe/UnLaTem
¹²	DEDUCE	Deidentification	Removal of Protected Health Information	https://github.com/vmenger/deduce
¹³	Verschueren-CTCue-rule-based	Information extraction	Capturing Disease Progression in Electronic Health Records of Patients With Stage IV Non-Small Cell lung cancer	Table S1a+b+c
¹⁴	Burgt-TM	Information extraction	Extraction of the occurrence and severity of eleven commonly described cardiac characteristics	Supplement
¹⁵	Arends-ALLRule	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler
¹⁵	Arends-BOW	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler

15	Arends-CNN	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
15	Arends-MedCAT	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
15	Arends-MedRoBERTa	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
15	Arends-SetFit-RobBERT	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
15	Arends-SpanCategorizer	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
15	Arends-biGRU	Information extraction	Diagnosis extraction	https://github.com/umcu/EchoLabeler .
16	Muizelaar-StrMatch	Information extraction	Lifestyle characteristics classification, specifically on smoking, alcohol and drug usage	https://github.com/hielkemui/zelaar/clinical-dutch-lifestyle-extraction/
17	Krastman-HW	Information extraction	To identify patients with diagnoses of a hand or wrist disorder	Supplement
18	DEDUCE	Deidentification	Removal of Protected Health Information	https://github.com/vmenger/deduce
18	Trienes-BiLSTM-CRF	Deidentification	Removal of Protected Health Information	https://github.com/nedap/deidentify
19	Olthof-RB-raw	Information extraction	Classify radiology reports in orthopaedic trauma for the presence of injuries	Supplement
20	MedRoBERTa.nl	Language normalization, representation, and modeling	Encoder style language model development	https://medroberta.nl/
21	Kim-MedRoberta-FT	Information extraction	Classification to ICF (Classification of Functioning, Disability and Health)	https://github.com/cltl/a-proof-zonmw https://huggingface.co/CLTL/models

22	Es-RoBERTa	Information extraction	Negation detection	https://github.com/umcu/negation-detection/tree/master
23	Nobel-RB	Information extraction	Quantify T-stage of pulmonary tumors according to the tumor node metastasis classification	https://github.com/putssander/medstruct-config/tree/final-results
24	joeranbosma/dragon-bert-base-domain-specific	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2166
24	joeranbosma/dragon-bert-base-mixed-domain	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2166
24	joeranbosma/dragon-longformer-base-domain-specific	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2173
24	joeranbosma/dragon-longformer-base-mixed-domain	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2172
24	joeranbosma/dragon-longformer-large-domain-specific	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2175
24	joeranbosma/dragon-longformer-large-mixed-domain	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2174
24	joeranbosma/dragon-roberta-base-domain-specific	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2169

²⁴	joeranbosma/dragon-roberta-base-mixed-domain	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2168
²⁴	joeranbosma/dragon-roberta-large-domain-specific	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2171
²⁴	joeranbosma/dragon-roberta-large-mixed-domain	Language normalization, representation, and modeling	Encoder style language model development	https://doi.org/10.57967/HF/2170
²⁵	Hendrix-regex	Information extraction	To identify pulmonary nodules described in radiology reports	Table E1-E4