Large Language Models to process, analyze, and synthesize biomedical texts – a scoping review

Simona Emilova Doneva[1], Sijing Qin[1], Beate Sick[2], Tilia Ellendorff[3], Gerold Schneider[3], and Benjamin Victor Ineichen[1,4]

[1]**Center for Reproducible Science, University of Zurich, Zurich, Switzerland.**
[2]**ZHAW School of Engineering, Winterthur, Switzerland.**
[3]**Department of Computational Linguistics, University of Zurich, Zurich, Switzerland.**
[4]**Clinical Neuroscience Center, University of Zurich, Zurich, Switzerland.**

ABSTRACT

The advent of large language models (LLMs) such as BERT and, more recently, GPT, is transforming our approach of analyzing and understanding biomedical texts. To stay informed about the latest advancements in this area, there is a need for up-to-date summaries on the role of LLM in Natural Language Processing (NLP) of biomedical texts. Thus, this scoping review provides a detailed overview of the current state of biomedical NLP research and its applications, with a special focus on the evolving role of LLMs. We conducted a systematic search of PubMed, EMBASE, and Google Scholar for studies and conference proceedings published from 2017 to December 19, 2023, that develop or utilize LLMs for NLP tasks in biomedicine. We evaluated the risk of bias in these studies using a 3-item checklist. From 13,823 references, we selected 199 publications and conference proceedings for our review. LLMs are being applied to a wide array of tasks in the biomedical field, including knowledge management, text mining, drug discovery, and evidence synthesis. Prominent among these tasks are text classification, relation extraction, and named entity recognition. Although BERT-based models remain prevalent, the use of GPT-based models has substantially increased since 2023. We conclude that, despite offering opportunities to manage the growing volume of biomedical data, LLMs also present challenges, particularly in clinical medicine and evidence synthesis, such as issues with transparency and privacy concerns.

Keywords: Natural Language Processing, Bioinformatics, Biomedicine, Large Language Models, BERT, evidence synthesis

INTRODUCTION

Natural Language Processing (NLP) has become an essential tool for processing data, widely adopted in biomedical research and clinical applications (Zhou et al., 2022). The advent of Large Language Models (LLMs) has further expanded NLP's capabilities, revolutionizing how we analyze and interpret complex biomedical texts. Those models are characterized by their extensive scale, both in terms of the data they are trained on and their architectural complexity. They utilize deep learning techniques, especially Transformer models, to process and generate human language with a high degree of proficiency. These models stand out due to their ability to understand context over lengthy passages and their versatility across a wide range of language tasks, including text generation, translation, and question answering. Examples of LLMs include OpenAI's GPT series and Google's BERT (Brown et al., 2020; Devlin et al., 2018). Given the fast-paced nature of NLP, marked by regular introductions of new tools and solutions, there is a critical need for up-to-date literature overviews.

Thus, in our scoping review, we aim to summarize the role of LLMs in processing and analyzing biomedical texts, i.e., in the field of Biomedical Natural Language Processing (BioNLP). Concretely, the goal here is threefold: Firstly, we aim to identify and categorize the specific tasks within BioNLP that are currently being addressed using LLMs, e.g., in evidence synthesis and using tasks such as named entity recognition and information extraction.

Secondly, we aim to map the various LLM architectures employed in these BioNLP tasks. Understanding the architecture is crucial, as it directly impacts the efficacy, accuracy, and applicability of these models in handling the unique challenges presented by biomedical literature.

FInally, this study sets out to assess the transparency of methods used in the development of these LLMs. This includes assessing the availability of source code and data, as well as reporting on the hardware and software utilized. Such transparency is essential for replicability, trustworthiness, and further advancement in the field.

Related Work

Existing review has argued that in the age of artificial intelligence, LLMs can greatly assist in literature search in biomedical fields, especially in terms of changing the way users interact with biomedical literature (Jin et al., 2024).

Additionally, in a systematic survey about pre-trained language models in biomedical domain, Wang et al. (2023) summarized available generative pre-trained language models, protein/DNA language models, and pre-trained vision-language models. Furthermore, they specifically presented an overview of pre-training models fine-tuned for diverse biomedical downstream tasks. Similarly, Tian et al. (2024) also sorted specialized biomedical LLMs targeting different downstream tasks such as question answering and information extraction, evaluated and compared their performance based on represented datasets.

Moreover, Thirunavukarasu et al. (2023) chose ChatGPT as an illustrative example to explore state-of-the-art LLM applications in medicine, especially the clinical, educational, and research aspects. The authors described the development and iteration of different GPT versions, illustrated the evolution from LLM to generative AI chatbot and discussed how the LLMs can be utilized in the clinical setting.

In another review highlighting the ChatGPT utility, Sallam (2023) retrieved several English records and categorized them primarily by benefits and applications of or risks and concerns toward ChatGPT to examine ChatGPT in healthcare education, research and practice.

A recent survey has been conducted with the objective to comprehensively explore biomedical and clinical NLP research in languages other than English. This survey specifically focuses on data resources, language models, and prevalent NLP tasks in these languages. They also report on the trend towards the use of transformer-based language models for various NLP tasks in medical fields (Lavelli et al., 2023).

Latest analysis also focused on the applications of LLMs in healthcare. The review suggested that LLMs have the potential to improve the patient experience. Patient's history of complaints, along with additional information such as comorbidities, risk factors, and medication lists, can be used by LLMs to predict possible disease categories during the pre-consultation phase. Patients can accordingly obtain the recommended medical subspecialties and make appointments. (Yang et al., 2023).

METHODS

Study registration

We registered the study protocol on the Open Science Framework (OSF) platform (`https://osf.io/bjv24/`).

Literature search

We conducted a search for studies in PubMed and EMBASE, spanning from 2017 (i.e., the inception of transformer-based LLMs (Vaswani et al., 2017)) to December 19, 2023. Additionally, we searched Google Scholar for EMNLP and ACL machine learning conference proceedings through the "Publish or Perish" interface, employing a search string translated from the one used in our PubMed query. This search string was created by an information specialist at the University of Zurich's library. The complete search string can be found in the supplementary data.

Inclusion and exclusion criteria

We included original research articles that employed LLMs to analyze extensive collections of biomedical texts, e.g., scientific publications, (pre)clinical trial registries, patents, and grey literature. Additionally, systematic reviews and meta-analyses that leveraged LLMs for automating certain tasks like abstract screening were also considered.

We excluded grey literature (e.g., non peer-reviewed conference abstracts), Non-English publications as well as text data derived from clinical questionnaires or surveys. Reviews were also excluded but were retained as a supplementary source for additional references.

Study selection

All retrieved publications were screened using ASReview which incorporates a machine learning algorithms to prioritize relevant studies (Van De Schoot et al., 2021). We a priori defined to stop abstract screening after thirty irrelevant abstracts have been presented in a row, and the remaining abstracts were excluded from further analysis.

Data extraction

The following data were extracted from available full texts of eligible articles: Target application (grouped into pre-specified classes), domain of automated approach, target database, data type (e.g., abstracts), data availability, hosted Application for end-users, LLM (e.g., BioBERT), programming language (e.g., Python), library/framework (e.g., HuggingFace), reported performance metrics (e.g., F1-score), source code availabilityy, pretraining corpus origin (e.g., PubMed abstracts), Fine-tuning corpus data/task (e.g., bc5cdr), fine-tuning corpus size, number of tasks/datasets for performance evaluation, hardware type (e.g., GPU). Interrater agreement for data extraction was assessed by independent duplicate extraction of 10 studies by two authors (SED and SQ). These criteria were developed and iteratively improved over a pilot extraction round conducted by two authors (SED and BVI).

Risk of bias assessment

We assessed the quality of each study against three predefined criteria: 1) Is the process of splitting training from validation data described? 2) Are there methods described for avoiding overfitting or underfitting? And 3) Does the dataset or assessment measure provide a possibility to compare to other tools in the same domain?

Data synthesis and analysis

Findings were summarized in a narrative fashion bolstered by descriptive statistics of extracted parameters.

Data and code availability

The complete data extraction sheets and code for the data analysis is made available in a GitHub repository .

Changes from protocol

While in our initial protocol, we were planning to evaluate all ML-based approaches for BioNLP, our focus changed to specifically evaluating the emerging field of LLMs.

RESULTS

Eligible publications

In total, 13,823 publications were retrieved from our database search. 224 publications were eligible for full-text screening (Figure 1). A total of 196 articles were included in our final review, with 137 coming from the biomedical journals and 59 from the NLP conference outputs (Figure 2).

Overview of included papers

Figure 2 illustrates the distribution of article counts across various journals. In the realm of NLP venues, the journal "ACL/ Findings," associated with the Association for Computational Linguistics (ACL) conference[1], and the Conference on Empirical Methods in Natural Language Processing (EMNLP)[2], are notable for their substantial contributions. These conferences are renowned for their focus on computational linguistics and empirical methods in NLP. On the biomedical front, "BMC Bioinformatics" and "Journal of Biomedical Informatics" emerge as prominent sources. BMC Bioinformatics[3] is acclaimed for its focus on computational algorithms, software, and systems biology, with the overall goal of advancing bioinformatics. The Journal of Biomedical Informatics[4] specializes in the intersection of

[1] https://dl.acm.org/conference/acl
[2] https://dl.acm.org/conference/emnlp
[3] https://bmcbioinformatics.biomedcentral.com/
[4] https://www.sciencedirect.com/journal/journal-of-biomedical-informatics

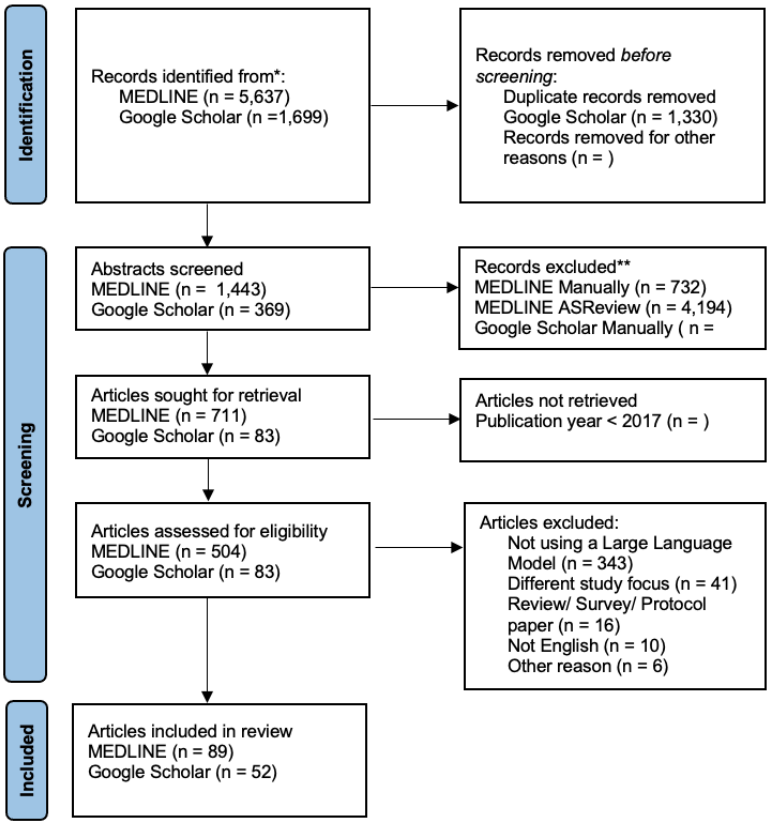

Figure 1. PRISMA Flow Diagram (Page et al., 2021).

biology, medicine, and information technology, with a focus on new methodologies and techniques that address real-world biomedical or clinical problems.

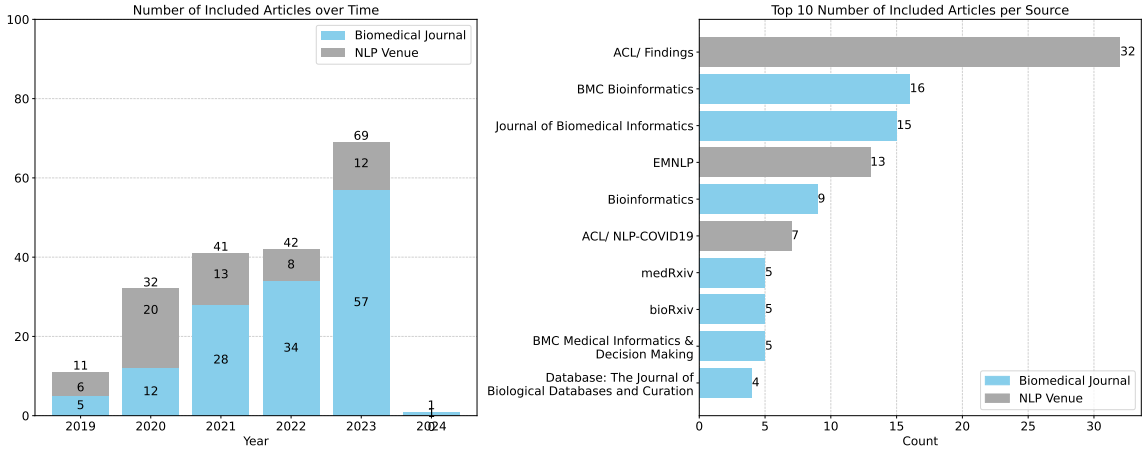

Figure 2. Number Articles Over Venue/Journal.

Risk of bias of eligible articles

The overall risk of bias TODO: based on ML reporting best-practices; we don't compare any outcomes? was moderate with XX% of studies reporting how/whether training and validation data were splitted, XX% of studies reported measures to mitigate over-/underfitting of models, and XX% of studies report on

benchmarking their approaches against other methods.

Applications

Data Sources and Data Types

Figure 3 introduces the top 10 data sources used to inform the training and refinement phases of the LLMs development. The most frequently reported ones were PubMed/Medline (98 publications, 49%) , clinical texts (i.e., textual data generated in clinical settings like electronic health records; 20, 10%), social media (12, 6%), and CORD (The COVID-19 Open Research Dataset; 10, 5%).

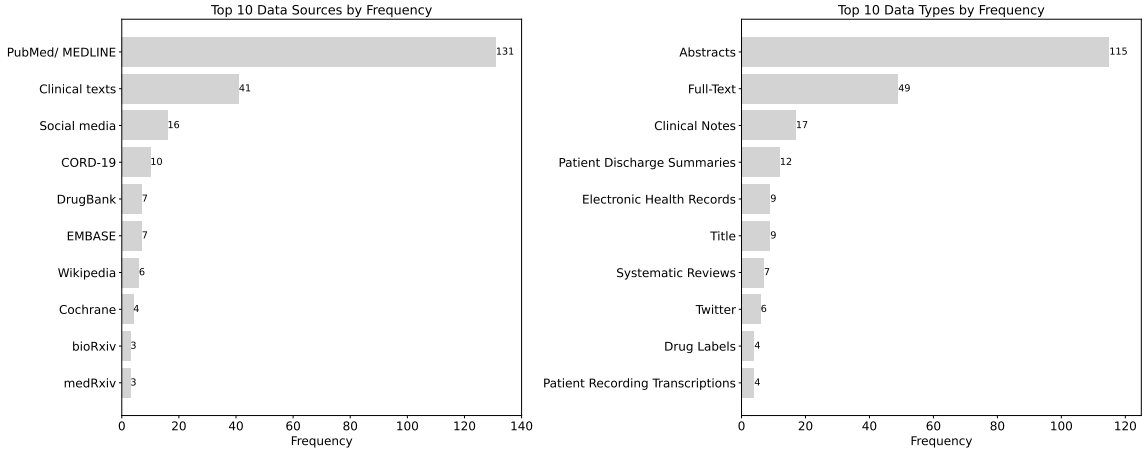

Figure 3. Frequency of specific data sources and data types used for model development and testing.

The most commonly used data types were abstracts (91 publications, 46%) and full texts of scientific articles (36, 18%) as well as clinical notes (13, 7%) (Figure 3).

Biomedical Application Domains

Eligible publications were grouped into 8 main categories of biomedical applications for LLMs (see Table 1). The top three categories were knowledge management (48 publications, 24%), general biomedical text mining (45, 23%), and Pharmacology (18, 9%) (Figure 4). Further details about concrete applications within the top 4 most frequent domains are provided in the supplementary materials in Figure 10.

Domain Category	Definition
General biomedical text mining	Developing a new/improved methodology for an NLP task (e.g., NER, dependency parsing).
Knowledge Management	Knowledge management in NLP involves tasks such as literature-based discovery, curation, and screening, focusing on recommendation, summarization, annotation, and categorization of scientific literature.
Pharmacology	Development, testing, and understanding of therapeutic agents.
Media for Health Care	Analyzing media content to extract health-related information, trends, or public sentiments for healthcare applications.
Clinical	Enhancing clinical decision-making, patient care, and medical record management.
Evidence Synthesis	Automated evidence extraction and synthesis from biomedical texts.
Biology	Understanding and categorizing biological processes, mechanisms, and interactions.

Table 1. Main Domains of Application of Large Language Models in Biomedicine

NLP tasks

A variety of NLP tasks has been employed by the eligible articles, most commonly text classification (37 publications, 19%), relation extraction (20, 10%), and multi-domain applications (17, 9%) (Figure 4).

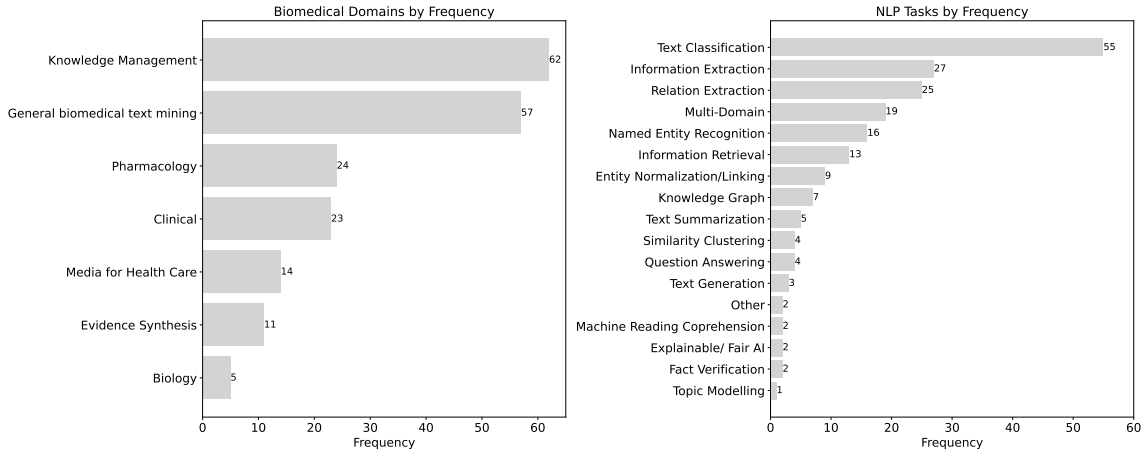

Figure 4. Biomedical domains and NLP tasks by article frequency.

Matching the main domains of application with the NLP tasks shows notable links (Figure 5):

In knowledge management, LLMs have been used for two main applications: First text classification and second, information retrieval. Text Classification technologies, in particular, streamline the organization of large literature volumes by assigning predefined categories or labels to documents. This approach has been effectively applied to enhance reference prioritization in systematic reviews (Mantas et al., 2021; Aum and Choe, 2021; Ambalavanan and Devarakonda, 2020; Habets et al., 2022; Jimeno Yepes and Verspoor, 2023). Text Classification has been also used to better understand the structure of papers, for example by automatically predicting sections and headers in Electronic Health Records (Rosenthal et al., 2019). Information Retrieval on the other hand plays a key role in knowledge management, aiming to facilitate navigation through vast volumes of literature. Various applications have been developed to improve the discovery of novel insights by more effectively aggregating data from existing knowledge databases, as well as improving article recommendation service(Martenot et al., 2022; Kart et al., 2022). For instance, pubmedKB serves as a web server designed to extract and visualize semantic relationships between genes, diseases, chemicals, and variants within PubMed abstracts (Li et al., 2022).

In the clinical domain, text classification is frequently utilized for predicting patient outcomes. For instance, in-hospital mortality predictions are made by combining time series data from various medical devices with clinical notes found in electronic health records (Zhang et al., 2022; Deznabi et al., 2021) Additionally, there are applications in mental health, such as the automated detection of mental conditions using transcribed patient recordings (Duan et al., 2023; Aich et al., 2022).

Social media platforms have been utilized for mental health screening, employing text classification to identify suicidal risk and predict mental health disorders from user-generated content, such as Reddit and Twitter posts (Zanwar et al., 2023; Sawhney et al., 2022, 2020). Additionally, this technology has been applied for detecting nuanced emotional states within online health communities (Sosea and Caragea, 2020). A recent development in this area has been the use of fact verification techniques to authenticate statements related to COVID-19 (Hossain et al., 2020; Liu et al., 2020).

Biomedical text mining spans various tasks, with Multi-Domain analysis standing out significantly. This area focuses on designing methods or frameworks capable of addressing multiple NLP tasks through a unified strategy. An illustration of this is BioGPT, a generative Transformer language model, which, after being pre-trained on an extensive biomedical literature corpus, demonstrates its versatility across six different biomedical NLP challenges (Luo et al., 2022). Named Entity Recognition and Entity Normalization are also commonly tackled challenges. For example (Tong et al., 2021) develop a multi-task model that simultaneously learns sentence-level and token-level labels for NER, utilizing BioBERT for text encoding and sharing hidden states across tasks. Furthermore, for specific applications like COVID-19, techniques have been developed, such as a neural BERT-based model for concept wikification, efficiently performing end-to-end entity linking by processing sentences through the BERT framework and assigning a unique concept name to each token Lymperopoulos et al. (2020)

The Pharma domain utilizes relation extraction (13). This task involves identifying and extracting relationships between entities such as drugs, genes, or proteins from text, which is crucial for pharmacological research and drug development. For example, Guan and Devarakonda (2019) utilize BERT and Edge sampling, a technique for selecting negative training samples, to enhance adverse drug events extraction from clinical notes and patient discharge summaries. KafiKang and Hendawi (2023) introduce an approach for identifying and classifying drug-drug interactions by combining Relation BioBERT and Bidirectional Long Short-Term Memory.

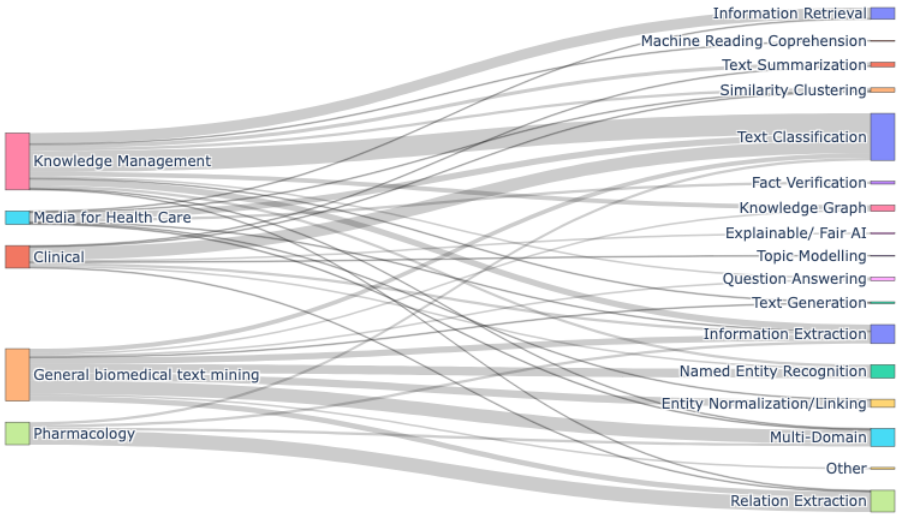

Figure 5. Sankey Diagram showing the relationships between the biomedical domains and the utilized NLP applications. Each domain is represented as a source node, while the associated NLP applications are shown as the target nodes. The thickness of the flows between the two is proportional to the number of articles that exhibit this connection.

Large Language Models

Models Overview and Trends Over Time

Figure 6 provides an overview for the most frequently used LLM models across all evaluated papers. The results show a focus on encoder-based BERT architectures. However there is also a notable presence of GPT models.

The most prominently used LLMs were encoder-based BERT models (279 models, 91%), followed by GPT models (19, 5%) (Figure 6).

Figure 6 presents the annual usage trends of the top five LLMs. While BERT models are prominently used in earlier years, GPT models see increasing use in 2023. The data also shows that general-purpose models like BERT are replaced with more specialized models such as BioBERT and SciBERT.

Technical setup: hardware and programming languages

The most commonly employed hardware by LLM papers with 90% of reported hardware were GPUs (56/62) (Figure 7). Of note, there was a lack of reporting on hardware use in many instances (79 publications).

The vast majority of studies used the Python programming language for their technical setup (157 publications, 91%) (Figure 7). JAVA, R, Perl, PHP, AWK, and C were rarely used. Again, many studies did not report the used programming language (27 publications).

HuggingFace (78 publications, 39%), PyTorch (45, 23%), and TensorFlow (30, 15%) were the most commonly reported computational libraries for using LLMs (Figure 7) . Some studies used more specialized libraries like Stanford CoreNLP, spaCy, Keras, NLTK, and Torch. Notably, many studies (45) did not report the computational library used.

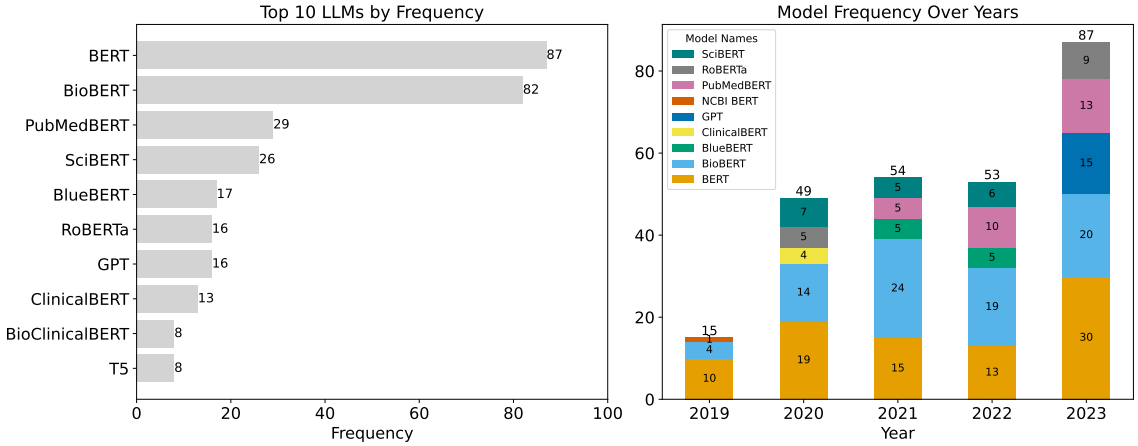

Figure 6. Most frequently used models and trends over time.

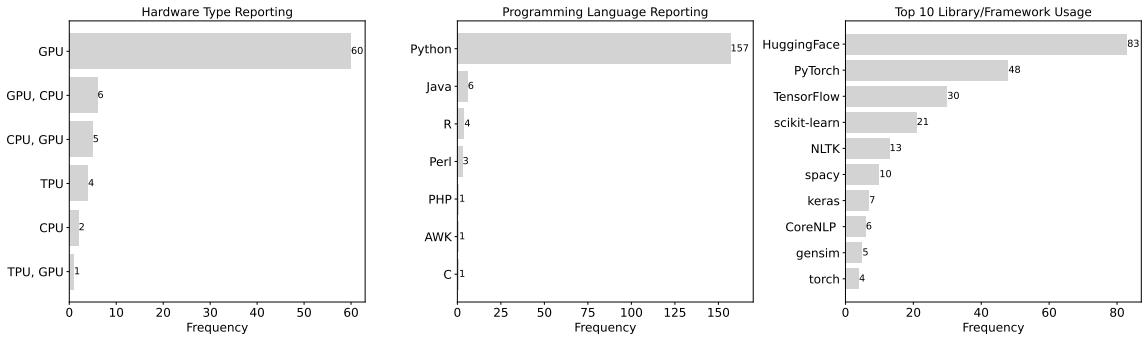

Figure 7. Reported technical setup.

Fine-Tuning Tasks and Datasets

Well-established Benchmark Datasets

The field of BioNLP relies heavily on the availability of standardized datasets and benchmarks to assess and compare the performance of various language models and algorithms. In this context, Table 2 provides a comprehensive compilation of the most commonly used standardized benchmarks, on which LLM performance was reported.

Competitions and Shared Tasks

While Table 2 compiles the most frequently used standard benchmarks for model evaluation, several research papers have also made use of datasets from prominent competitions and shared tasks in the field of NLP. These collaborative efforts provide standardized datasets and evaluation metrics, facilitating fair comparisons among different research teams. Key competitions and shared tasks include:

- **Informatics for Integrating Biology and the Bedside (i2b2) Challenges:** Focused on clinical data, these challenges have addressed topics such as temporal relations in clinical narratives (2012) and de-identification (2014).
- **BioCreative Workshops:** These workshops host challenges related to the extraction and annotation of biological entities (e.g., genes, proteins) and relationships from scientific literature. Datasets like BioCreative-LitCovid and DrugProt have been utilized for various applications.
- **Text REtrieval Conference (TREC):** TREC is an ongoing series of workshops that cover a wide range of information retrieval topics. Datasets associated with TREC, such as TREC-COVID and Health Misinformation, have been employed in numerous NLP research papers.

Task	Dataset	Frequency	Type	Train	Dev	Test	Evaluation Metrics
NER	NCBI-disease	18	Disease	5134	787	960	F1 entity-level
	BC5CDR-disease	13	Disease	4182	4244	4424	F1 entity-level
	BC5CDR-chem	10	Chemical/ Drug	5203	5347	5385	F1 entity-level
	BC2GM	9	Protein/ Gene	15197	3061	6325	F1 entity-level
	JNLPBA	7	Protein/ Gene	46750	4551	8662	F1 entity-level
	BC4CHEMD	5	Chemical/ Drug	3500	3500	3000	F1 entity-level
	Species-800	5	Species from NCBI Taxonomy	800 abstracts	-	-	F1 entity-level
	LINNAEUS	5	Species from NCBI Taxonomy	100 full-text	-	-	F1 entity-level
Relation Extraction	ChemProt	9	Chemical-protein interactions	18035	11268	15745	Micro F1
	DDIExtractions 2013	9	Drug-Drug interactions	25296	2496	5716	Micro F1
Text Classification	MIMIC-III	5	Clinical data	112,000 clinical reports	-	-	
Multi-Task	2010 i2b2/VA	5	Clinical data	394 clinical reports	-	477 clinical reports	

Table 2. Common Fine-Tuning tasks and related Datasets

Custom-developed Datasets

We observed 36% of papers from biomedical journals and 29% of papers from NLP conferences that reported the development of new datasets. These datasets were mostly in the area of 'Text Classification', 'Information Retrieval/Extraction', and 'Named Entity Recognition' (Figure 8).

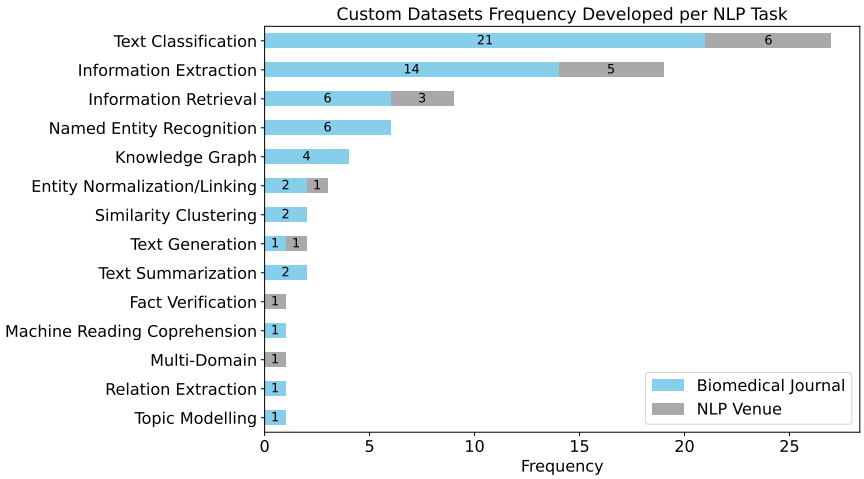

Figure 8. Target applications for which custom data has been annotated or collected.

Transparency of methods

There was a moderate level of code (e.g., algorithms and computational methods) transparency with approximately half of the studies making their code available alongside the publication (Figure 9). There was a high level of data transparency with over three quarters of studies making their datasets fully available. Finally, only a minority of studies hosted applications for end-users with a higher propensity of biomedical journals compared to NLP venues.

DISCUSSION

Main Findings

The objective of our scoping review was to explore and organize the specific tasks in BioNLP that are being tackled with LLMs. In the biomedical domain, LLMs have found applications across a diverse set of tasks, including knowledge management, mining biomedical texts, drug discovery, clinical applications, and synthesizing evidence. Within these areas, tasks such as text classification, relation extraction, and named entity recognition/information extraction stand out for their prominence. The primary sources of data for these studies are often PubMed/Medline or clinical documents, such as electronic health records.

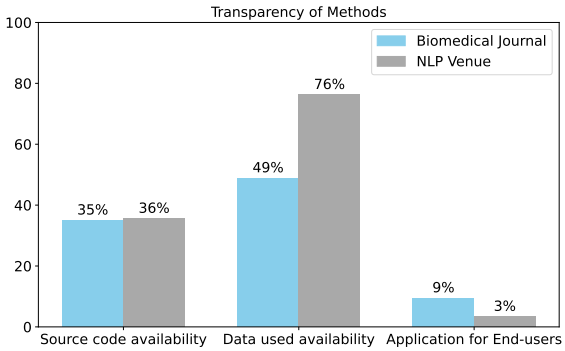

Figure 9. Transparency of LLM Methods: Source Code, Data, and Hosted Application Availability..

While BERT-based models continue to be widely used, there has been a noticeable surge in the adoption of GPT-based models since 2023. Moreover, the technical frameworks for these studies typically involve the Python programming language, with HuggingFace and PyTorch being the most popular computational libraries.

Findings in the Context of Existing Evidence

1. Tasks

LLMs have been used for a variety of NLP tasks. mostly for knowledge management and biomedical text mining, i.e., for curating, extracting, and synthesizing useful information from the vast volume of biomedical literature. With this, they seem to provide an ample opportunity to overcome the increasing volumes of biomedical data being published. One notable application of LLMs was in the realm of evidence synthesis, i.e., the methodological approach used to systematically review, critically appraise, and combine results from multiple studies to draw more robust and comprehensive conclusions about a specific research question or area of interest. Although LLMs could be of immense help to streamline the often time-consuming task of manually screening papers for relevance or extracting data, these approaches do not come withouth their limitation: Despite their capabilities, these models have limitations: First, their 'black box' operation means these models often lack transparency, which is a key requirement in evidence synthesis. Second, their memorized knowledge does not comprise primary biomedical data and they are thus not specifically tailored to scientific literature. Third, these models tend to hallucinate, i.e. generate factually incorrect information, which poses a risk for automated evidence synthesis. Together, these limitations mean they fall short of the rigorous standards essential for systematic reviews aimed at minimizing bias. Therefore, although LLMs present an opportunity to address some of the challenges in MS research, they cannot yet substitute for the detailed, methodical approach of systematic reviews and meta-analyses in health care. Dedicated, more specialized methods are necessary to meet the stringent requirements of evidence synthesis in the medical field.

2. Sources

As information source, most LLM approaches make use of information collected from scientific abstracts only, likely in part representing the easy and abundant availability as well as the condensed format of this text type. However,information from abstracts might not be well aligned with the corresponding full reports. For example, Li et al. (2017) analyse the state of reporting of primary biomedical research and find inconsistencies with respect to the reported sample sizes, outcome measures, result presentation and interpretation, and conclusions or recommendations. This suggests the need to be cautious when developing applications that rely only on the information reported in abstracts, especially for evidence-synthesis. Article abstracts and article bodies can differ not only in content, but also in their structural, linguistic, and semantic composition (Cohen et al., 2010). LLMs are generally well equipped to handle a wide variety of linguistic styles and structures due to their vast training on diverse text corpora. However, those models can still struggle with the subtlety and specificity required in academic texts, especially if their training data does not sufficiently cover the breadth and depth of academic language. Commonly, the article collections from existing systematic reviews were used as a fine-tuning dataset to train a model

to identify relevant papers to the topic in an automated way (Aum and Choe, 2021). In contrast, clinical notes were less commonly used, likely due to data privacy concerns (Hartman et al., 2020).

3. Employed models and technical setup

While BERT-based models have been the predominantly used LLMs in recent years, there was a notable surge in GPT-based models in 2023. BERT employs a bidirectional framework, analyzing text in both directions simultaneously, which is enabled by its use of the Transformer encoder architecture. This bidirectionality allows BERT to understand the context of a word based on its entire surrounding text, making it good at tasks that require a deep understanding of language context, such as sentiment analysis, question answering, and named entity recognition. BERT's pre-training involves masked language modeling and next sentence prediction, tasks that help it learn a comprehensive understanding of language structure and flow. While most reported models follow a BERT-based architecture, they differ in the pretraining corpus used. The standard BERT (Devlin et al., 2018) model is pretrained on texts from Wikipedia and BookCorpus, which is considered general-domain. SciBERT (Beltagy et al., 2019) is also trained from scratch on a purely scientific corpus from Semanitic Scholar. However its pretraining corpus is a mixture of biomedical and computer science texts. GPT-like models operate on a unidirectional framework, processing text from left to right and utilizing the Transformer decoder architecture, making it well fit for generative tasks like text completion and content creation. Its training is focused on predicting the next word in a sequence, optimizing the model for generating coherent and contextually relevant text. While BERT-like architectures typically require task-specific fine-tuning to achieve optimal performance, models like GPT-3 demonstrate few-shot and zero-shot learning capabilities. Few-shot learning refers to the model's ability to perform tasks with a very limited amount of training data, while zero-shot learning refers to its ability to perform tasks without any task-specific training data. This is achieved through advanced prompting techniques and in-context learning, where the model generates responses based on the context provided in the prompt (Zhang and Li, 2021). TODO: add more details on which gpt version has been used.

Training Corpus	BERT	BioBERT	PubMedBERT	SciBERT	BlueBERT
General	✓	✓	✗	✗	✓
PMC	✗	✓	✓	✗	✗
PubMed	✗	✓	✓	✗	✓
Semantic Scholar	✗	✗	✗	✓	✗
Clinical Notes	✗	✗	✗	✗	✓

Table 3. Training Corpora for Different Models

Regarding used programming languages and computational libraries, HuggingFace's prominence can be attributed to its comprehensive collection of pre-trained models and easy-to-use interfaces, making it highly appealing for both research and application purposes. PyTorch, known for its flexibility and dynamic computation graph, appeals to researchers for its ease of experimentation and prototyping. Similarly, TensorFlow's significant usage reflects its robustness and scalability, particularly in production environments. The presence of scikit-learn underscores its role in data preprocessing, feature extraction, and traditional machine learning tasks, which remain relevant even in the context of advanced language models.

4. Data transparency

Finally, although only a minority of studies provide end-user applications, most of the eligible studies made their datasets and code freely available which is in line with the current open science movement.

LIMITATIONS

First, the current literature review is primarily centered on text-based applications of LLMs in biomedicine. However, biomedical data are inherently multi-modal, encompassing text, imaging, tabular, time-series, and structured sequence data such as proteins and DNA. Notably, the integration of text with imaging data represents a significant area of research exploration, leading to the development of large biomedical vision-and-language models. Furthermore language models have been used for genomic and proteomic analysis. These models treat biological sequences, e.g, amino acids in proteins, similar to textual data,

enabling the prediction of sequence functions, structures, and variations. An overview of these two research areas has been given in other work (Wang et al., 2023).

second, we omitted articles published in languages other than English, which might restrict the scope of this review, as we might miss out on related papers written in other languages.

CONCLUSION

This scoping review provides an overview how LLMs are used for NLP tasks in biomedicine and health sciences. The findings from our scoping review underline the rapid progress of LLMs, emphasizing their potential in accelerating discovery and enhancing health outcomes. These advancements signal a promising avenue for leveraging LLMs in biomedicine and health, given their capacity to process and analyze complex biomedical texts with high proficiency. However, we also acknowledges the inherent risks and challenges associated with the deployment of LLMs like ChatGPT in these sensitive areas, including evidence synthesis. Issues such as the generation of fabricated information and concerns over legal and privacy implications pose significant hurdles to the safe and ethical application of these technologies. These challenges highlight the need for careful consideration and management to mitigate risks associated with the use of LLMs in biomedicine and health.

ACKNOWLEDGMENTS

REFERENCES

Aich, A., Quynh, A., Badal, V., Pinkham, A., Harvey, P., Depp, C., and Parde, N. (2022). Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887.

Ambalavanan, A. K. and Devarakonda, M. V. (2020). Using the contextual language model bert for multi-criteria classification of scientific articles. *Journal of biomedical informatics*, 112:103578.

Aum, S. and Choe, S. (2021). srbert: automatic article classification model for systematic review using bert. *Systematic reviews*, 10(1):1–8.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11:1–10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Deznabi, I., Iyyer, M., and Fiterau, M. (2021). Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4026–4031.

Duan, J., Wei, F., Liu, J., Li, H., Liu, T., and Wang, J. (2023). CDA: A contrastive data augmentation method for alzheimer's disease detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1819–1826.

Guan, H. and Devarakonda, M. (2019). Leveraging contextual information in extracting long distance relations from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2019, page 1051. American Medical Informatics Association.

Habets, P. C., van IJzendoorn, D. G., Vinkers, C. H., Härmark, L., de Vries, L. C., and Otte, W. M. (2022). Development and validation of a machine-learning algorithm to predict the relevance of scientific articles within the field of teratology. *Reproductive Toxicology*, 113:150–154.

Hartman, T., Howell, M. D., Dean, J., Hoory, S., Slyper, R., Laish, I., Gilon, O., Vainstein, D., Corrado, G., Chou, K., et al. (2020). Customization scenarios for de-identification of clinical notes. *BMC medical informatics and decision making*, 20(1):1–9.

Hossain, T., Iv, R. L. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Jimeno Yepes, A. J. and Verspoor, K. (2023). Classifying literature mentions of biological pathogens as experimentally studied using natural language processing. *Journal of Biomedical Semantics*, 14(1):1.

Jin, Q., Leaman, R., and Lu, Z. (2024). Pubmed and beyond: biomedical literature search in the age of artificial intelligence. *Ebiomedicine*, 100.

KafiKang, M. and Hendawi, A. (2023). Drug-drug interaction extraction from biomedical text using relation biobert with blstm. *Machine Learning and Knowledge Extraction*, 5(2):669–683.

Kart, Ö., Mestiashvili, A., Lachmann, K., Kwasnicki, R., and Schroeder, M. (2022). Emati: a recommender system for biomedical literature based on supervised learning. *Database*, 2022:baac104.

Lavelli, A., Krauthammer, M., and Rinaldi, F. (2023). Exploring the latest highlights in medical natural language processing across multiple languages: A survey. *Yearbook of Medical Informatics*, 32(01):230–243.

Li, G., Abbade, L. P., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., Wang, M., Bhatt, M., Zielinski, L., Sanger, N., et al. (2017). A scoping review of comparisons between abstracts and full reports in primary biomedical research. *BMC medical research methodology*, 17:1–12.

Li, P.-H., Chen, T.-F., Yu, J.-Y., Shih, S.-H., Su, C.-H., Lin, Y.-H., Tsai, H.-K., Juan, H.-F., Chen, C.-Y., and Huang, J.-H. (2022). pubmedkb: an interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Research*, 50(W1):W616–W622.

Liu, Z., Xiong, C., Dai, Z., Sun, S., Sun, M., and Liu, Z. (2020). Adapting open domain fact extraction and verification to covid-fact through in-domain language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2395–2400.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Lymperopoulos, P., Qiu, H., and Min, B. (2020). Concept wikification for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Mantas, J. et al. (2021). The classification of short scientific texts using pretrained bert model. *Public Health and Informatics: Proceedings of MIE 2021*, 281:83.

Martenot, V., Masdeu, V., Cupe, J., Gehin, F., Blanchon, M., Dauriat, J., Horst, A., Renaudin, M., Girard, P., and Zucker, J.-D. (2022). Lisa: an assisted literature search pipeline for detecting serious adverse drug events with deep learning. *BMC medical informatics and decision making*, 22(1):1–16.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, 372(n71).

Rosenthal, S., Barker, K., and Liang, Z. (2019). Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4864–4873.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Sawhney, R., Joshi, H., Gandhi, S., and Shah, R. (2020). A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Sawhney, R., Neerkaje, A. T., and Gaur, M. (2022). A risk-averse mechanism for suicidality assessment on social media. *Association for Computational Linguistics 2022 (ACL 2022)*.

Sosea, T. and Caragea, C. (2020). Canceremo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Tian, S., Jin, Q., Yeganova, L., Lai, P.-T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D. C., et al. (2024). Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Tong, Y., Chen, Y., and Shi, X. (2021). A multi-task approach for improving biomedical named

entity recognition by incorporating multi-granularity information. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4804–4813.

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., and Fu, J. (2023). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., and Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Zanwar, S., Li, X., Wiechmann, D., Qiao, Y., and Kerz, E. (2023). What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8926–8940.

Zhang, M. and Li, J. (2021). A commentary of gpt-3 in mit technology review 2021. *Fundamental Research*, 1(6):831–833.

Zhang, Y., Zhou, B., Song, K., Sui, X., Zhao, G., Jiang, N., and Yuan, X. (2022). PM2F2N: Patient multi-view multi-modal feature fusion networks for clinical outcome prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1985–1994.

Zhou, B., Yang, G., Shi, Z., and Ma, S. (2022). Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*.

SUPPLEMENTARY DATA

Search string

Supplementary search string (’data mining’/exp OR ((data NEXT/1 mining):ti,ab,kw)) AND (literature:ti,ab OR abstract:ti,ab OR abstracts:ti,ab OR text:ti,ab OR articles:ti,ab) OR (((analyz* OR analys* OR extract* OR screen* OR evaluat* OR classif* OR ’natural language processing’) NEAR/3 (literature OR abstract OR abstracts OR text OR articles)):ti,ab,kw) OR ((text NEXT/1 mining):ti,ab,kw) AND ’natural language processing’/exp OR ’artificial neural network’/exp OR ’support vector machine’/exp OR ’machine learning’/de OR ’automated pattern recognition’/exp OR ’artificial intelligence’/de OR ’semi automation’:ti,ab,kw OR automation:ti,ab,kw OR ’artificial intelligence’:ti,ab,kw OR ai:ti,ab,kw OR ’natural language processing’:ti,ab,kw OR ((machine NEXT/1 (intelligence OR learning)):ti,ab,kw) OR ((’text mining’ OR ’data-mining’) NEXT/3 (tool* OR technique* OR system)):ti,ab,kw) OR (((deep OR comput* OR model* OR convolutional OR artificial OR algorithmic OR connectionist OR mathematical) NEXT/1 ’neural network*’):ti,ab,kw) OR ((ann NEXT/3 (analysis OR approach OR method* OR model* OR output OR technique* OR training*)):ti,ab,kw) OR ((connectionist NEXT/1 (model OR network)):ti,ab,kw) OR ((’support vector’ NEXT/1 (machine* OR classif* OR network OR regression)):ti,ab,kw) OR ((automated NEXT/3 (tool* OR technique* OR system OR ’pattern recognition’ OR analyz* OR analys* OR extract* OR screen* OR evaluat* OR classif*)):ti,ab,kw)

Target Biomedical Applications

Number of tested models per paper over time

The mean number of LLMs used by individual papers ranges between 1.36 and 2.39 , with a slight upward trend over time (Figure 11).

Overview of custom-annotated datasets

Target Data	Type	Train	Dev	Test
PubMed	if publication reports on pharmacokinetic parameters obtained in vivo - relevant, else not relevant	3992	-	800
	annotated sentences with one of the casual relation types: 1) correlational, 2) conditional causal, 3) direct causal, 4) no relationship	3061	-	-
	abstracts assigned one of the following study labels: 1) a randomized controlled trial, 2) a human study, 3) a systematic review without meta-analysis, 4) a systematic review with meta analysis, 5) a study protocol, 6) a rodent study or 7) any other abstract type	50000	-	5000
	if publication reports on artificial intelligence (AI) applications in neurosurgery - relevant, else not			
	Classified articles for Article type: 1) original study, 2) systematic review, or 3) evidence-based guideline; Purpose categories: 1) treatment, 2) primary prevention, 3) diagnosis, 4) harm from clinical interventions, 5) economics, 6) overall prognosis, 7) clinical prediction guide, or 8) quality improvement. Methodological quality criteria - yes/no.			
Cochrane Reviews	annotated for quality of evidence: 1) RoB, 2) imprecision, 3) inconsistency, 4) indirectness, 5) publication bias			
Elsevier	if publication reports on COVID-19 - relevant, else not			
MEDLINE	if publication reports on a RCT - relevant, else not			
	one line per citation in which the PMID, and pairs (pathogen term, ncbi-id) indicate the active pathogens manually annotated for that citation			
EMBASE	if publication reports on adverse events related to pharmaceutical products of Bayer - relevant, else not			
Several	risk of bias domains annotated in full-text pre-clinical studies: 1) Random Allocation, 2) Blinded Assessment of Outcome 3) Compliance with Animal Welfare Regulations 4) Conflict of Interests 5) Animal Exclusions			
	full-text annotations for the resource role types and the resource function types of citations in scientific literature. 3 general Role types: Material, Method, Supplement. 9 fine-grained Role types: Data, Tool, Code, Algorithm, Document, Website, Paper, License, Media. 6 Function types: Use, Produce, Introduce, Extend, Compare, Other			

Table 4. Custom-annotated datasets for Text Classification.

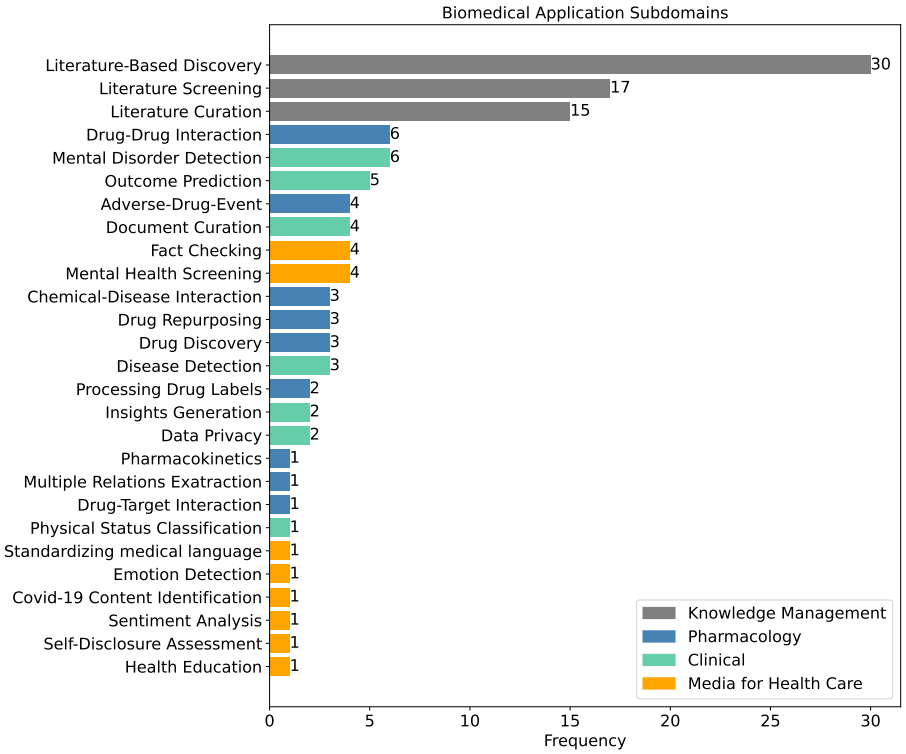

Figure 10. Number of articles assigned to each subdomain.

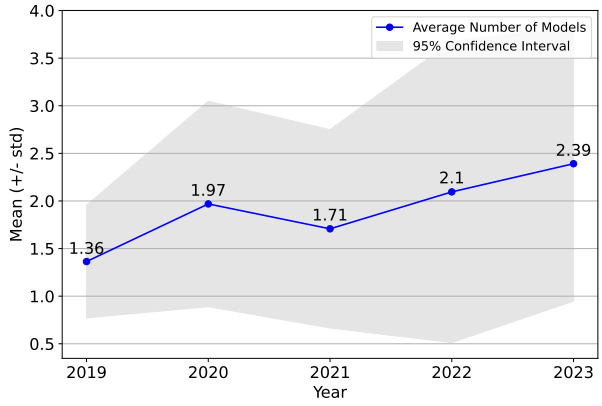

Figure 11. Average number of different models used per paper each year.

Target Data	Type	Train	Dev	Test
PubMed	if publication reports on Vitamin B's impact on health - relevant, else not relevant	3992	-	800
	annotated sentences with on of event seriousness types: 1) serious 2) important medical event 3) none	6859	-	917
	if publication reports on Covid-19- relevant, else not relevant abstracts assigned a topic: 1) general information, 2) mechanism, 3) transmission, 4) diagnosis, 5) treatment, 6) prevention, 7) case report, or 8) epidemic forecasting geolocation annotations: country, city, state and nationality, mapped into countries		-	
	if publication reports on long Covid-19 - relevant, else not relevant a nnotated three entity types: Covid-19 virus strains, vaccines, and vaccine funders			
	sentences mapped to three-dimensional locations in the human atlas			
	headline from a news article linked to the abstracts of the research publications cited by that article			
ClinVar	sentence-level triplet (genetic variant, <association>, disease) annotated. <association>can be Cause-associated, Appositive, and In-patient			
DisGeNET	sentence-level odds ratio statistics annotated			

Table 5. Custom-annotated datasets for Information Retrieval.