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

Simona Doneva[1], Sijing Qin[2], Beate Sick[3], Tilia Ellendorff[4], Gerold Schneider[5], and Benjamin Victor Ineichen[6]

[1,6]**Center for Reproducible Science, University of Zurich, Zurich, Switzerland**
[2,4,5]**University of Zurich, Zurich, Switzerland**
[3]**ZHAW School of Engineering, Winterthur, Switzerland**

ABSTRACT

This survey seeks to offer a comprehensive summary of the present landscape in biomedical and clinical Natural Language Processing (NLP) research and application. In particular, we focus on the emerging role of Large Language Models (LLMs). By providing a clear and structured overview of current practices and trends, we aim at helping researchers and practitioners make informed decisions about which models and techniques are best suited for specific NLP challenges. Furthermore, this study contributes to the broader understanding of the capabilities and limitations of LLMs in a highly specialized and impactful domain like biomedicine, ultimately aiding in the development of more sophisticated and effective tools for biomedical text analysis. Finally, we seek to gain insights into the openness and accessibility of LLM research, which is particularly crucial in the context of biomedical applications where reliability and ethical considerations are paramount.

Keywords: Natural Language Processing, Bioinformatics, Biomedicine, Large Language Models

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 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's a continuous demand for up-to-date literature reviews and overviews.

In our current literature review, we are focused on thoroughly exploring the role of LLMs in processing and analyzing biomedical texts. Within the field of Biomedical Natural Language Processing (BioNLP), LLMs hold the potential to revolutionize various applications. These include interpreting medical literature, enhancing literature reviews, facilitating drug discovery by analyzing biological data, assisting in clinical decision-making, and making medical information more accessible for patient education. Our review will cover three primary areas.

Firstly, we seek to identify and categorize the specific tasks within BioNLP that are currently being addressed using LLMs. These tasks encompass a range of activities, such as Named Entity Recognition, Information Extraction, and Text Classification. The study aims to provide a clear overview of how LLMs are changing the way automated processing and understanding of complex biomedical texts is performed, as well as the most affected areas like drug discovery, clinical decision support, and personalized medicine.

Secondly, we explore 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.

Additionally, this study incorporates an evaluation of 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

TODO Discuss related work such as: (Yang et al., 2023b), (Thirunavukarasu et al., 2023), (Wang et al., 2023), (Sallam, 2023).

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).

METHODS

Study registration

We registered the study protocol on the ... platform. (TODO: add link)

Search

Our literature search was conducted on MEDLINE via PubMed and Embase, employing a strategy developed with a librarian's assistance. The details of this search strategy are available in our protocol. Additionally, for a more extensive search, we adapted this strategy to Google Scholar to include articles published in notable Machine Learning Conferences (ACL and EMNLP). We defined specific criteria to guide the selection of articles for our analysis.

Inclusion and exclusion criteria

The inclusion criteria encompassed original research articles that employed LLMs to analyze extensive collections of biomedical texts, including scientific publications, (pre)clinical trial registries, patents, and grey literature. Additionally, systematic reviews and meta-analyses that utilized LLMs for automating tasks like abstract screening were also considered. Our focus was particularly on works published from 2017 onwards, a period marking the inception of LLMs with the publication of the Transformer architecture (Vaswani et al., 2017).

Our exclusion criteria ruled out conference abstracts or proceedings. While reviews were not included in our main analysis, they were retained as a supplementary source for additional references. Non-English publications, as well as text data derived from clinical questionnaires or surveys, were also excluded from our analysis.

Study selection and data extraction

We conducted a two-phase data screening and extraction process, each phase executed by two independent reviewers. In the initial phase, we reviewed papers published up to April 2023. Subsequently we expanded to papers released from May to December 2023. The details of the review process are described below.

Selection of studies

We screened all retrieved publications using the ASReview software (Van De Schoot et al., 2021). This tool utilizes machine learning algorithms to prioritize relevant studies, making the review process more efficient. Based on initial relevance classifications of the abstracts by the reviwer, the machine learning model prioritizes the remaining abstracts, bringing potentially relevant studies to the top of the list. The model learns from the user's ongoing input, continuously improving its prioritization accuracy. Our predefined stopping criteria was based on the number of consecutively irrelevant abstracts. After seeing thirty irrelevant abstracts, the review process was interrupted and the remaining abstracts were excluded from further analysis. In cases when it was not clear from the abstract if the study is relevant, we included it in ASReview, and the final decision to make it part of the final analysis was based on the full text.

Data synthesis and analysis

We followed a structured approach to data synthesis and analysis consisting of the following steps:

1. **Pre-specification of key outcomes:** We began by pre-specifying the main outcomes of our review. These outcomes were carefully defined and included details on how each outcome would be extracted from the paper. For instance, one of the key outcomes focused on providing an overview of the core Large Language Models (LLMs) used, such as BERT.

2. **Pilot extraction:** To validate our outcome definitions, a pilot extraction was performed on a subset of 20 papers. Two reviewers independently read the full texts of these papers and manually extracted the relevant information into a structured Excel sheet. This pilot phase allowed us to improve the definitions of our outcomes and reduce any possible ambiguity.

3. **Full text review and data extraction:** Finally, the full text of each paper was read and all relevant information pertaining to the specified outcomes was extracted.

Accessibility of data

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

Results of the search

In total, 18,065 (TODO) original publications were retrieved from our comprehensive database search. After abstract and title screening, 655 (TODO) publications were eligible for full-text search. After screening the full text of these studies, 122 (TODO) articles (4% of deduplicated references) were included for information extraction.

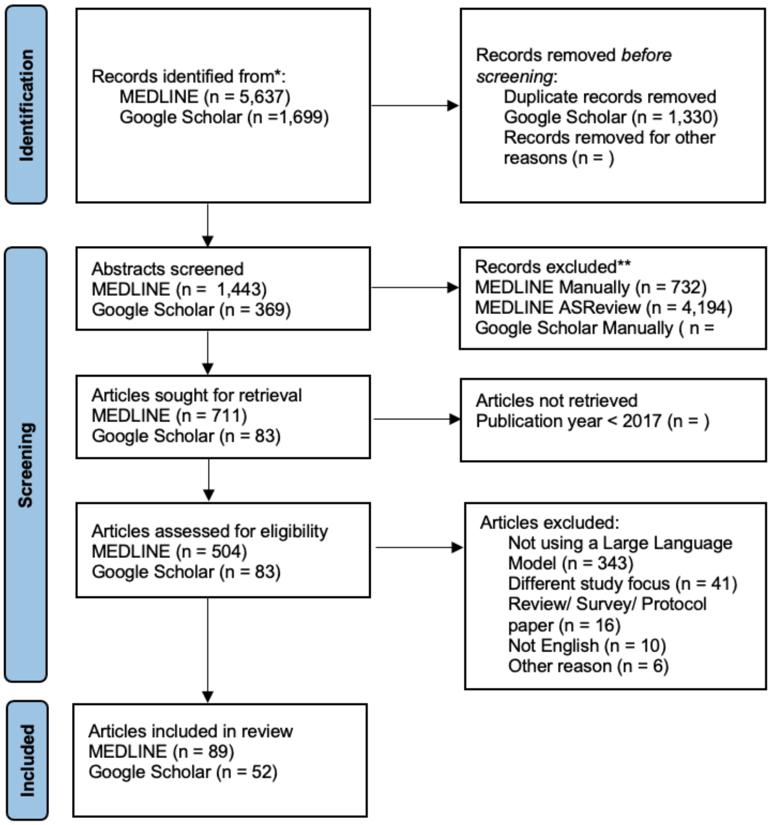
Identification
Records identified from*:
MEDLINE (n = 5,637)
Google Scholar (n =1,699)
Records removed before screening:
Duplicate records removed
Google Scholar (n = 1,330)
Records removed for other reasons (n =)
Screening
Abstracts screened
MEDLINE (n = 1,443)
Google Scholar (n = 369)
Records excluded**
MEDLINE Manually (n = 732)
MEDLINE ASReview (n = 4,194)
Google Scholar Manually (n =
Articles sought for retrieval
MEDLINE (n = 711)
Google Scholar (n = 83)
Articles not retrieved
Publication year < 2017 (n =)
Articles assessed for eligibility
MEDLINE (n = 504)
Google Scholar (n = 83)
Articles excluded:
Not using a Large Language Model (n = 343)
Different study focus (n = 41)
Review/ Survey/ Protocol paper (n = 16)
Not English (n = 10)
Other reason (n = 6)
Included
Articles included in review
MEDLINE (n = 89)
Google Scholar (n = 52)

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

Overview of included papers

A total of 199 articles were included in our final review. From those 138 resulted from the biomedical database search, and 61 were included from the NLP conference outputs. We can see the distribution of articles over time and source in Figure 2.

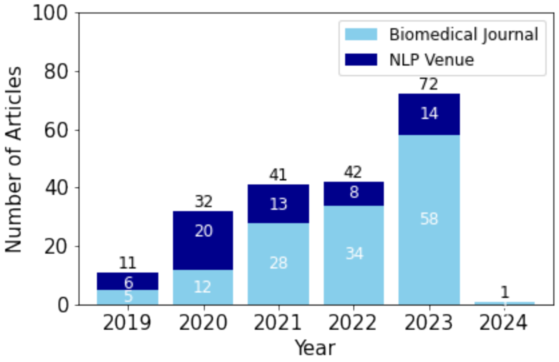

Figure 2. Number Articles Over Publication Year and Publication Type.

Figure 3 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 biology, medicine, and information technology, with a focus on new methodologies and techniques that address real-world biomedical or clinical problems.

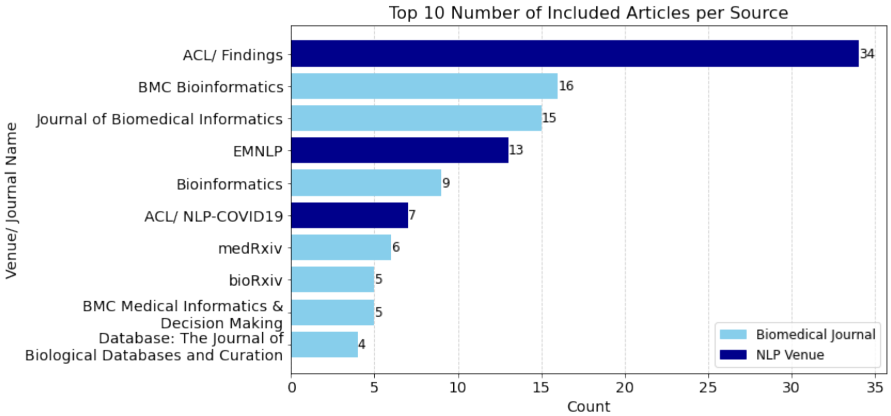

Figure 3. Number Articles Over Venue/Journal.

[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

Applications

Data Sources and Data Types

The robustness and efficacy of LLMs are fundamentally contingent upon the quality and variety of the datasets employed for their training and fine-tuning. These datasets are integral not just for the foundational training of the algorithms, but also in shaping the breadth and potential use-cases of the resulting models within their intended application areas. In this chapter, we provide a high-level overview of the most common data sources and types. In chapter LLMs we further show the concrete data sets that were used.

Figure 4 introduces the top 10 data sources used to inform the training and refinement phases of the LLMs development. Here we summarize the most prevalent ones:

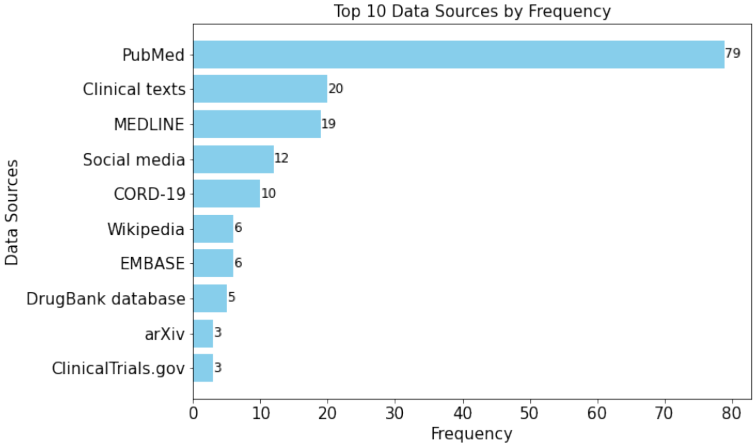

Figure 4. Frequency of specific data sources used for model development and testing.

1. **PubMed:** A free search engine accessing primarily the MEDLINE database, life science journals, and online books. It is widely used for its comprehensive coverage of research articles.[5]

2. **Clinical Texts:** These refer to a wide array of textual data generated in clinical settings, such as electronic health records (EHRs) and clinical notes. They are rich in patient-specific information, crucial for personalized medicine and clinical decision support.

3. **MEDLINE:** A premier bibliographic database of the U.S. National Library of Medicine. It contains millions of references and abstracts from life sciences and biomedical literature. MEDLINE includes literature published in more than 5600 journals worldwide.[6]

4. **Social Media:** This includes platforms like Twitter, Reddit, and health-focused forums where individuals share health-related experiences. Social media data can provide insights into patient experiences, disease surveillance, and public health trends.

5. **CORD-19:** The COVID-19 Open Research Dataset (CORD-19) is a repository of scientific papers on COVID-19 and related coronaviruses, launched by the Allen Institute for AI and partners. It includes over 140,000 papers, facilitating text mining and information retrieval for COVID-19 research. (Wang et al., 2020).

Further we focus on the various types of raw input data, which were extracted from these data sources to be processed by the LLMs. An overview of this data is provided in Figure 5.

The high prevalence of abstracts could be due to several practical considerations. They are more accessible, as full texts can be restricted or behind a paywall, and they require significantly less computational resources for processing, making them a more efficient choice. The conciseness of abstracts

[5] https://pubmed.ncbi.nlm.nih.gov/

[6] https://www.nlm.nih.gov/medline/

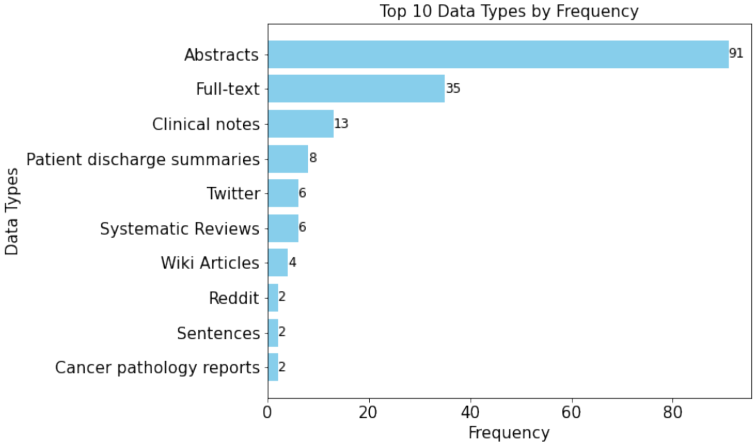

Figure 5. Frequency of specific data types used for model development and testing.

might aid in efficient data processing, allowing models to quickly grasp key findings and methodologies. Furthermore, data annotators can work more quickly when dealing with the structured and focused content of abstracts, leading to the faster development of annotated datasets, a critical resource for fine-tuning and evaluating LLMs. While full texts offer a comprehensive view of research, encompassing detailed methodologies and nuanced discussions, the conciseness of abstracts and the availability of corresponding annotated resources seem to make them the more pragmatic and favored choice for many NLP tasks in the biomedical field.

From the clinical texts data source, we see the usage of clinical notes, patient discharge summaries and cancer pathology reports. This is indicative of the need for LLMs to interpret and analyze detailed patient records for specific applications such as explainable prediction of sepsis and mortality from free-text notes (Feng et al., 2020). However, the relative rarity might be due to several challenges. For example, systems storing electronic health records are not typically designed for research, making data access and navigation difficult. This data is often irregular and noisy due to its primary focus on patient care rather than data collection quality (Johnson et al., 2023). Another important point is privacy concerns. Patient data is highly sensitive, and its use is strictly regulated by law. Automated approaches for de-identification have the potential to increase the volume of clinical data available for NLP research. However, their limited deployment due to uncertainties about their performance on unseen datasets restricts the expansion of research in this field (Hartman et al., 2020).

The use of social media data, while not predominant, indicates the potential for health-related applications such as pharmacovigilance (Hussain et al., 2021), responding to mental health issues (Sawhney et al., 2020), and normalizing medical concepts in social media text (Kalyan and Sangeetha, 2021).

Systematic reviews as a data source refers to the output from a comprehensive, methodical compilation and synthesis of research studies on a specific question or topic area (Clarke et al., 2007). 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).

Biomedical Application Domains

Based on the paper contents, we manually devised 8 main categories of biomedical applications for LLMs (see Table 1). Figure 6 shows the distribution of the number of articles across those various domains of application. Further details about concrete applications within the top 4 most frequent domains are provided in Figure 7. These sub-categories were derived from the reported objectives stated in each reviewed paper. Note that the applications behind ”General biomedical text mining” are explored in the next section.

Knowledge Management is the most frequently represented domains. This points to the importance of solutions for curating, extracting, and synthesizing useful information from the vast amount of biomedical literature. The high prominence of General Biomedical Text Mining might indicate an emphasis on

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.
Pharma	Drug discovery, development, and optimization.
Media for Health Care	Analyzing media content to extract health-related information, trends, or public sentiments for healthcare applications.
Clinical	Enhanceing clinical decision-making, patient care, and medical record management.
Evidence-Based Medicine	Automated evidence extraction and synthesis from biomedical texts.
Data Privacy	Ensuring the confidentiality, integrity, and secure handling of sensitive healthcare information.
Biological Functions	Understanding and categorizing biological processes, mechanisms, and interactions.

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

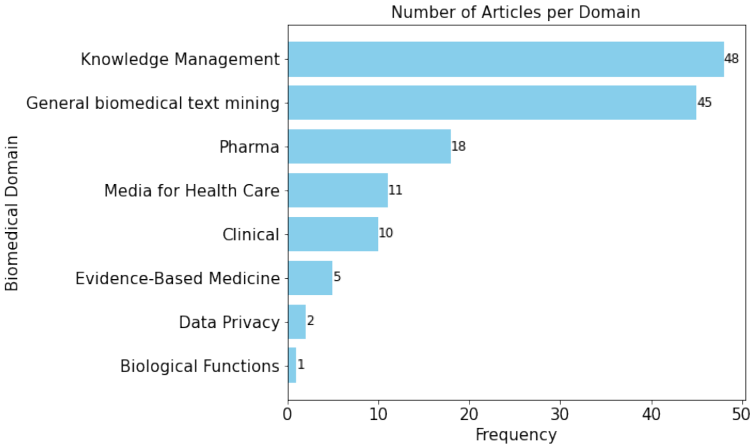

Figure 6. Number of articles assigned to each domain.

developing and improving methodologies for various NLP tasks, such as named entity recognition and dependency parsing. The Pharma domain reflects the interest in applying LLMs to drug discovery, development, and optimization, underscoring the potential of these models in accelerating and enhancing pharmaceutical research. Media for Health Care and Clinical domains might point towards a growing trend in utilizing LLMs for analyzing health-related information from media content and enhancing clinical decision-making and patient care. The relatively lower representation in other domains may indicate emerging areas of research in LLM applications or areas that currently face more challenges in integration and implementation.

NLP tasks

Figure 8 shows the distribution of various NLP tasks addressed by LLMs in the biomedical field. When classifying each paper to one of those tasks, we considered the main objective reported in the paper.

The sankey diagram in Fig. 9 allows to analyse which domains are most active in specific NLP applications. These relationships demonstrate the diversity of NLP tasks applied and the frequency analysis reveals that certain tasks are more prevalent within specific domains.

We can see that the Pharma domain significantly 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)

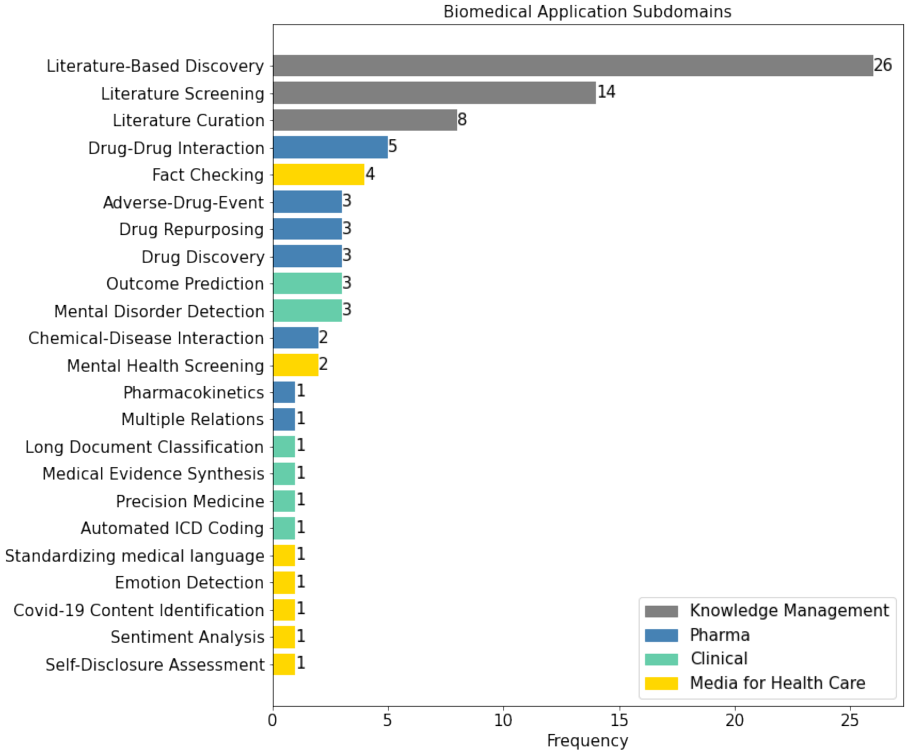

Figure 7. Number of articles assigned to each subdomain.

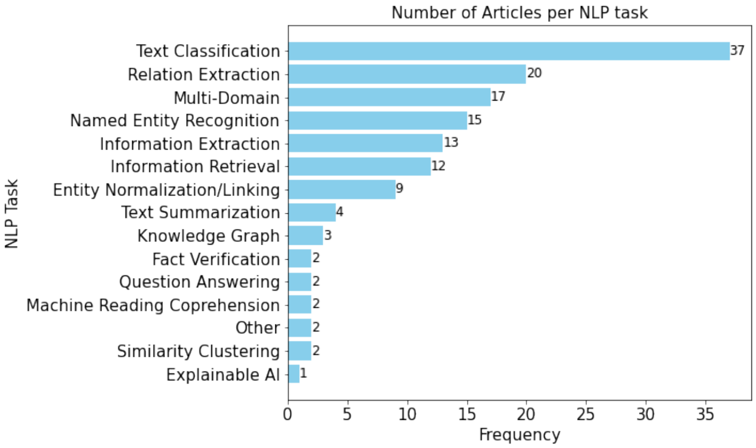

Figure 8. Number of articles assigned to each domain.

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.

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).

In Knowledge Management, LLMs could significantly improve literature analysis and 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 also plays a crucial role in Knowledge Management, aiming to facilitate efficient navigation through vast literature databases. 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).

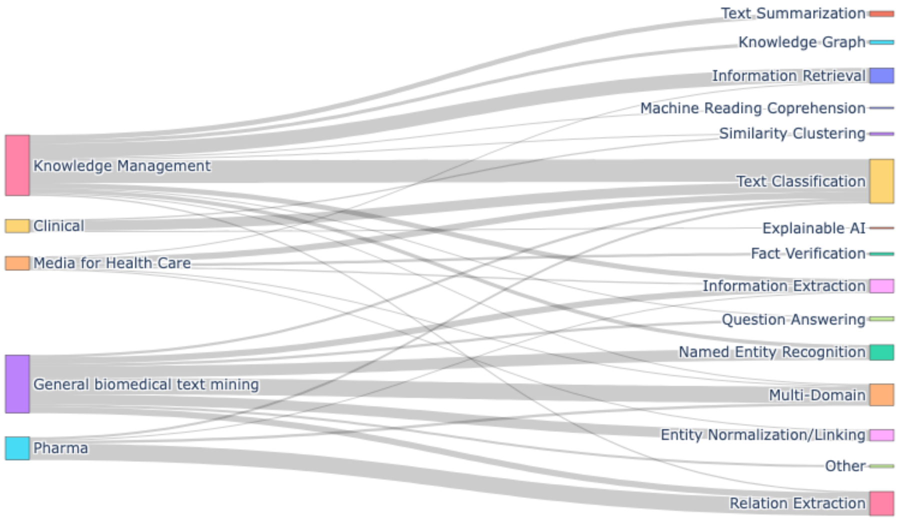

Figure 9. Sankey Diagram representing the relationships between the biomedical domains and the utilized NLP applications. Each domain is represented as a source node, while the associated NLP applications are shown as the target nodes. The thickness of the flows between the two is proportional to the number of articles that exhibit this connection.

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 also 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).

Large Language Models

Models Overview

Figure 10 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.

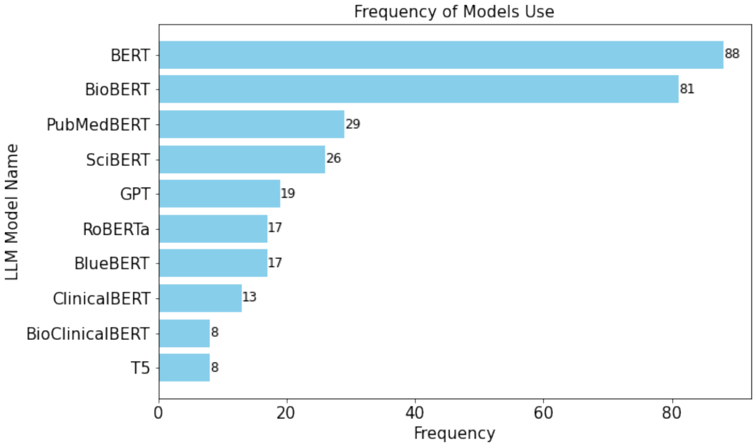

Figure 10. Most frequently used models.

The GPT-like and BERT-like models represent two distinct LLM approaches, differentiated by their architecture, training methods, and use cases (Yang et al., 2023a).

BERT-style Language Models

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. To improve the performance in biomedical NLP tasks, the models can be trained on biomedical text corpus in two ways (Gu et al., 2021):

1. Mixed-Domain Pretraining: The weights are first initialized with the general-domain BERT model and training is continued using biomedical texts. In the case of BioBERT (Lee et al., 2020) this includes PubMed abstracts and PubMed Central (PMC) full-text articles. BlueBERT (Peng et al., 2019) uses both PubMed abstracts and de-identified clinical notes from MIMIC-III (Johnson et al., 2016).

2. Domain-Specific Pretraining: In this setup the language model is trained using purely in-domain data. PubMedBERT (Gu et al., 2021) follows this approach and is pretrained from scratch on abstracts and full-texts from PubMed only.

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-style Language Models

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

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

Table 2. Training Corpora for Different Models

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.

Trends Over Time

Figure 11 shows the top 5 language models based on their reported usage for each year in the analysed literature. The data reveals evolving trends in the usage of language models from 2019 to 2023, highlighting a shift from general-purpose models like BERT, which saw peak usage in 2020, towards more specialized models such as BioBERT and SciBERT. This shift indicates a preference for domain-specific performance, as seen in the increasing use of BioBERT, especially in 2021. In 2023 GPT models gained substantial recognition with 18 mentions. Overall, the data underscores a dynamic landscape in language model utilization, with a clear trend towards specialization and new model adoption.

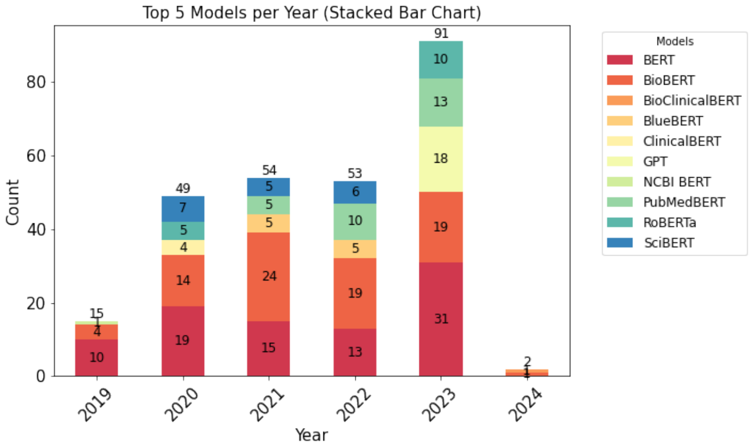

Figure 11. Top 5 most frequently utilized models for each year.

Figure 12 shows a line graph with confidence intervals that visualizes the average number of NLP models used by papers published each year. The upward trend in the average could indicate an increasing reliance on multiple models, possibly reflecting the growing complexity and diversity of NLP tasks and the need to combine different models to achieve better results in various experiments and research studies. Furthermore, the growing number of available models could present an opportunity for improved benchmarking, as researchers have a wider array of models to choose from when conducting experiments and evaluating NLP performance.

Technical setup

Figure 13 presents a summary of the hardware types reported in the development of large language models. The data reveals a significant reliance on GPUs in large language model development. GPUs, with their robust parallel processing capabilities, are evidently preferred for their efficiency in handling complex matrix operations typical in deep learning tasks.

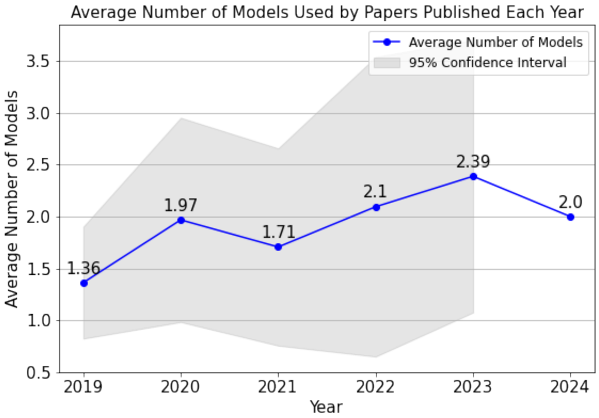

Figure 12. Average number of different models used per paper each year.

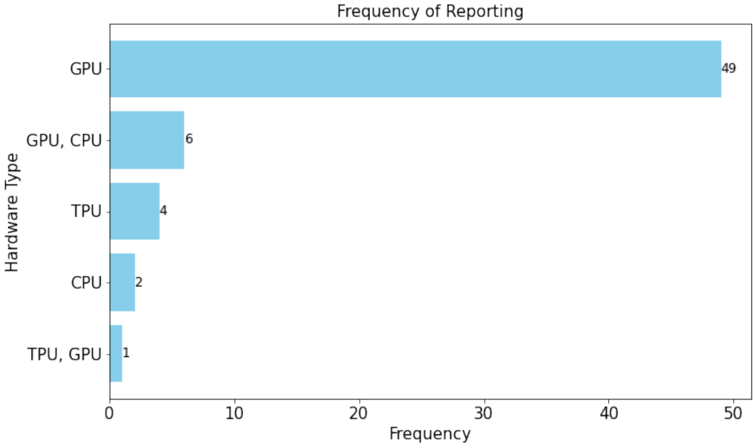

Figure 13. Reported hardware used.

In a high number of instances (79) the hardware used was not reported. This lack of reporting could point to a gap in the documentation practices within the field and raises questions about the reproducibility and comparability of these models.

Figure 14 provides a breakdown of the programming languages reported in the development of large language models. The overwhelming preference for Python (105 instances) underscores its status as the de facto standard in this field. Python's extensive libraries, community support, and readability make it highly conducive for rapid prototyping and complex machine learning tasks.

The occasional use of Python in combination with other languages like Java, C, Perl, and AWK (totaling 9 instances) reflects the diverse requirements of language model development. For instance, Java's use alongside Python could be attributed to its robustness in handling large-scale systems, while C might be chosen for performance-critical components. The inclusion of Perl and AWK, albeit rare, indicates specific use cases, probably related to their strengths in text processing.

Notably, there was a relatively high number of instances where the programming language was not reported (27). This could suggest that the choice of programming language is considered an obvious or trivial detail, not worth reporting. Alternatively, it could reflect a lack of standardization in reporting practices within the field.

Figure 15 summarizes the reported usage of different computational libraries in the development of large language models. The data highlights a diverse range of computational libraries used in large

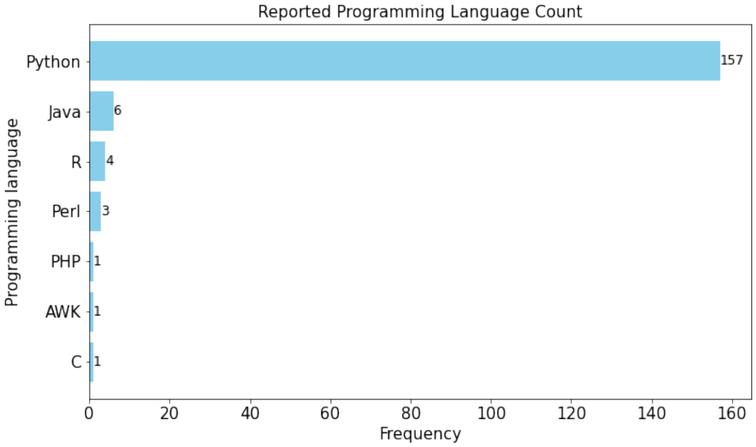

Figure 14. Reported programming language used.

language model development, with HuggingFace and PyTorch leading in popularity. 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.

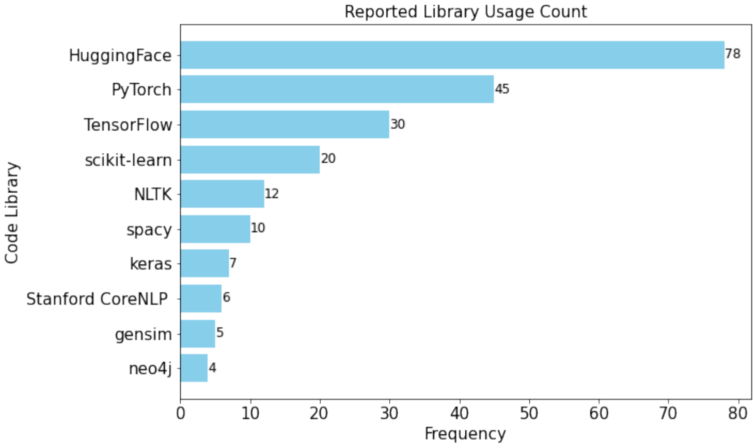

Figure 15. Reported programming language used.

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.

The usage of specialized libraries like Stanford CoreNLP and spaCy indicates the importance of sophisticated natural language processing capabilities in language model development. Libraries like Keras, NLTK, and Torch, though less prevalent compared to HuggingFace and PyTorch, highlight the diversity of tools researchers utilize to address different aspects of language modeling.

Notably, a significant number of studies (45 instances) did not report the computational library used. This non-reporting could suggest an oversight in detailing the development environment.

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 3 provides a comprehensive compilation of the most commonly used standardized benchmarks, on which LLM performance was reported.

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 3. Common Fine-Tuning tasks and related Datasets

Competitions and Shared Tasks

While Table 3 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.

Custom-developed Datasets

Finally, we observed 36% of papers from biomedical journals and 29% of papers from NLP conferences that reported the development of a new dataset. The bar chart in Figure 16 visualizes the distribution of these custom datasets across different NLP tasks, showing the areas where there is a focus on dataset creation. The emphasis on ’Text Classification’, ’Information Retrieval’, and ’Named Entity Recognition’ indicates significant research interest and potential advancements in these areas.

The prominence of ’Text Classification’ in the figure could be partially attributed to the relative ease of annotating data for this task. Text classification involves assigning predefined labels to text, which is a more straightforward annotation process compared to more complex tasks such as ’Named Entity Recognition’ or ’Information Extraction’, which require detailed, context-specific annotations.

Transparency of methods

The increasing application of Large Language Models (LLMs) in biomedical text research underscores the need for methodological transparency. This transparency is crucial for reproducibility, ethical considerations, and the advancement of the field. We focus on three transparency dimensions:

- Source Code Availability - the degree to which the algorithms and computational methods are accessible for review and reuse.

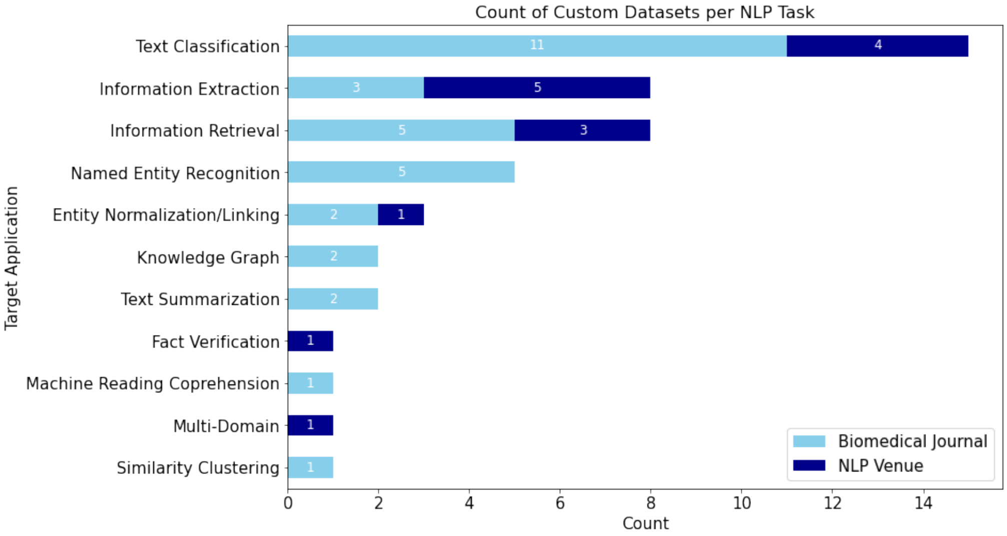

Figure 16. Target applications for which custom data has been developed.

- Data Used Availability - the extent to which the datasets used are available for independent verification.
- Hosted Application for End-Users - the availability of user-friendly interfaces or applications that demonstrate the practical application of these methods.

Figure 17 provides an overview of the percentage of publications that have explicitly reported on the availability those parameters. The data suggests a moderate level of code transparency among traditional biomedical journals, with over half of the papers making their code available for review and potential reuse. In contrast, only 40.38% of publications in natural language processing (NLP) venues shared their source code.

A substantial 75.28% of biomedical journal publications shared their data, indicating a commendable commitment to data transparency. This high percentage suggests that authors recognize the importance of making their datasets available for replication and further research. Publications in NLP venues excelled in this aspect, with 86.54% providing access to the data they used. This higher availability of data used might reflect the focus of development of new methods that are tested on established and publicly available benchmark datasets. Furthermore, privacy concerns and data access restrictions in the biomedical field may lead to more restricted data sharing practices.

Finally, while biomedical research shows a somewhat higher propensity to develop end-user applications, possibly driven by the direct applicability of their research in clinical settings, NLP venues appear more invested in foundational research and less in user-oriented products. This generally low interface availability suggests that while code and data may be accessible, opportunities for non-technical users to interact with the models are limited.

DISCUSSION

Our data shows that, to date, most text processing systems make use of information collected from abstracts only. However, there is evidence that the reported information in research abstracts is commonly not 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

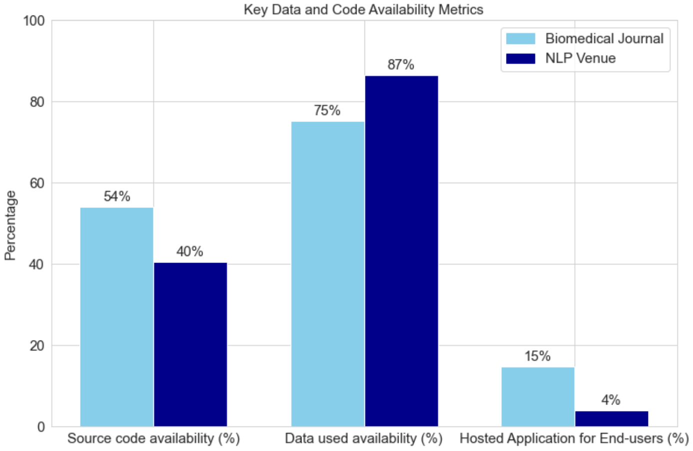

Figure 17. Transparency of LLM Methods: Source Code, Data, and Hosted Application Availability..

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.

We observed a diverse range of applications for LLMs in biomedicine, showcasing the broad applicability of NLP techniques in extracting meaningful information, facilitating knowledge discovery, and supporting decision-making processes in the biomedical field.

It is important to note that while LLMs demonstrate powerful capabilities, these models are not without limitations, and implementing them in critical scenarios introduces various potential hazards. The limitations, challenges and risks associated with LLMs the context of biomedical and health domains have been discussed in previous research (Tian et al., 2024; Thirunavukarasu et al., 2023). Among the important considerations for LLMs in this setting are maintaining accuracy and relevance due to fast-evolving medical knowledge, difficulty in interpreting complex medical terms and nuances, risk of perpetuating data biases, privacy concerns with sensitive patient data, and ethical issues around AI's role in clinical decisions, impacting accountability and patient care quality.

LIMITATIONS

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).

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

Our systematic literature review focused on publications since the inception of the Transformer architecture and explored the diverse applications of LLMs within the realms of biomedicine and health. We analysed the utilization of various input data sources and types, application domains, and NLP tasks. We examined different models, including BERT variants and highlighted the emergence of GPT-like models in the field. We gave a comprehensive overview of the prominent datasets used to fine-tune and evaluate those models.

We investigated the reported transparency regarding technical setup, methods, data availability, and source code.

The findings from our 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, our review also acknowledges the inherent risks and challenges associated with the deployment of LLMs like ChatGPT in these sensitive areas. 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.

Clarke, M., Hopewell, S., and Chalmers, L. (2007). Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. *Journal of the Royal Society of Medicine*, 100(4):187–190.

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.

Feng, J., Shaib, C., and Rudzicz, F. (2020). Explainable clinical decision support from text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1478–1489.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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*.

Hussain, S., Afzal, H., Saeed, R., Iltaf, N., and Umair, M. Y. (2021). Pharmacovigilance with transformers: A framework to detect adverse drug reactions using bert fine-tuned with farm. *Computational and Mathematical Methods in Medicine*, 2021.

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.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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.

Kalyan, K. S. and Sangeetha, S. (2021). Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. *Artificial Intelligence in Medicine*, 112:102008.

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.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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).

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

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.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., et al. (2020). Cord-19: The covid-19 open research dataset. *ArXiv*.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., and Hu, X. (2023a). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., and Liu, N. (2023b). 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*.

APPENDIX

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.

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.