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

Why was the study done? Brief background information

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

INTRODUCTION

Related systematic reviews and overviews

Here I will add a brief background and similar systematic reviews. NB: need to investigate for references!

Aim

The aim of this study is to comprehensively map the landscape of Large Language Models (LLMs) in the realm of processing and analyzing biomedical texts, a rapidly growing area of Biomedical Natural Language Processing (BioNLP). This exploration addresses two fundamental aspects:

- 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.

The importance of this study lies in its potential to guide future research and applications in BioNLP. By providing a clear and structured overview of current practices and trends, it can help researchers and practitioners make informed decisions about which models and techniques are best suited for specific BioNLP 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.

METHODS

Study registration

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

Search

We searched ... electronic databases, using the search methods previously described in our protocol. We searched MEDLINE, PubMed and Embase, using a search strategy developed with the help of an

information specialist. For a follow-up search of Machine Learning Conference (ACL, EMNLP) articles we used adaptations of this search strategy, which were made by the review authors.

Originally, we planned to include all ML methods, but we decided to focus on large language models.

Inclusion and exclusion criteria

Inclusion:

- Original research articles for methods using LLMs applied to large collections of Biomedical texts. This includes scientific publications, (pre)clinial trial registries, patents, grey literature, and potentially other sources.
- Systematic reviews / meta-analysis that have used a LLM to automate any of the associated systematic reviews tasks like abstract screening.
- Works published from 2017 onwards (representing the beginning of LLMs).

Exclusion: Conference abstracts or proceedings will be excluded. Reviews will be excluded but retained as a source for additional references. Non-English will be excluded. Medical reports or text data from (clinical) questionnaires/surveys will be excluded.

Study selection and data extraction

Selection of studies

We screened all retrieved publications using the ASReview software. 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.

Data synthesis and analysis

Full-text was read and information extracted into a Google Excel sheet. The full spreadsheet of items extracted from each included reference is available in ....

Accessibility of data

All data and code are free to access. A detailed list of sources is given in the 'Data availability' and 'Software availability' sections.

Changes from protocol

Describe shift of focus from general ML to 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 edxtraction.

Overview of included papers

A total of 141 articles were included for review. From those 89 resulted from the biomedical database search, and 52 were included from the NLP conference outputs. We can see the distribution of articles over time and source in Figure 2. The publications distribution seems to indicate an increase of focus in application-oriented work in the biomedical domain and a potential decline in foundational research related to biomedical NLP. (TBD, not sure we can say that; also not sure if we used a good search query? e.g. the ACL BioNLP workshop had 59 publications in 2022?)

Figure 3 shows the distribution of article counts across different journals and publication types. Among the top entries, the journal "ACL/ Findings" and "EMNLP" predominantly contribute to the category of

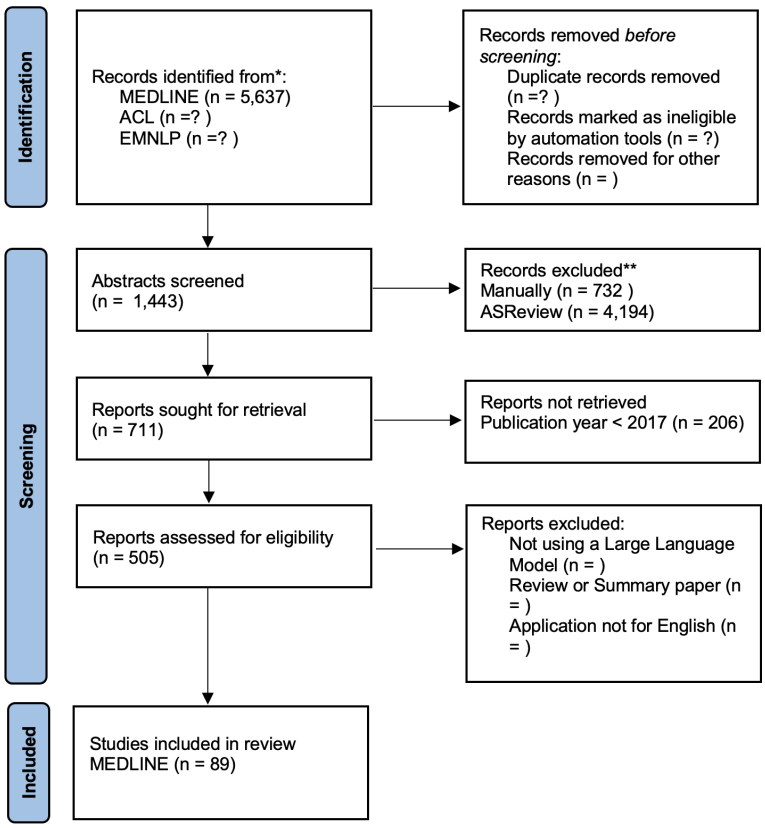

Figure 1. PRISMA Flow Diagram (Page et al., 2021).

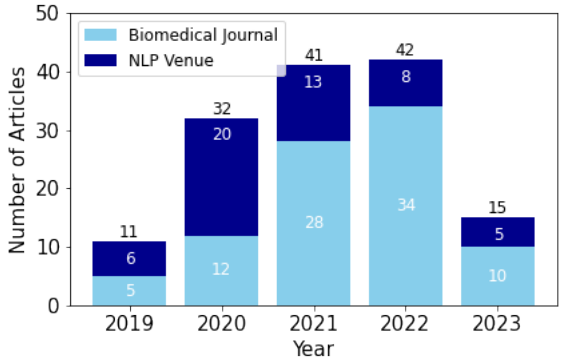

Figure 2. Number Articles Over Publication Year and Publication Type.

natural language processing (NLP) venues, with counts of 25 and 13, respectively. On the biomedical side, ”BMC Bioinformatics” and ”Journal of Biomedical Informatics” stand out, each featuring 12 and 11 articles, respectively.

Applications

Data Sources and Data Types

The robustness and efficacy of Large Language Models (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. Figure 4 introduces the top 10

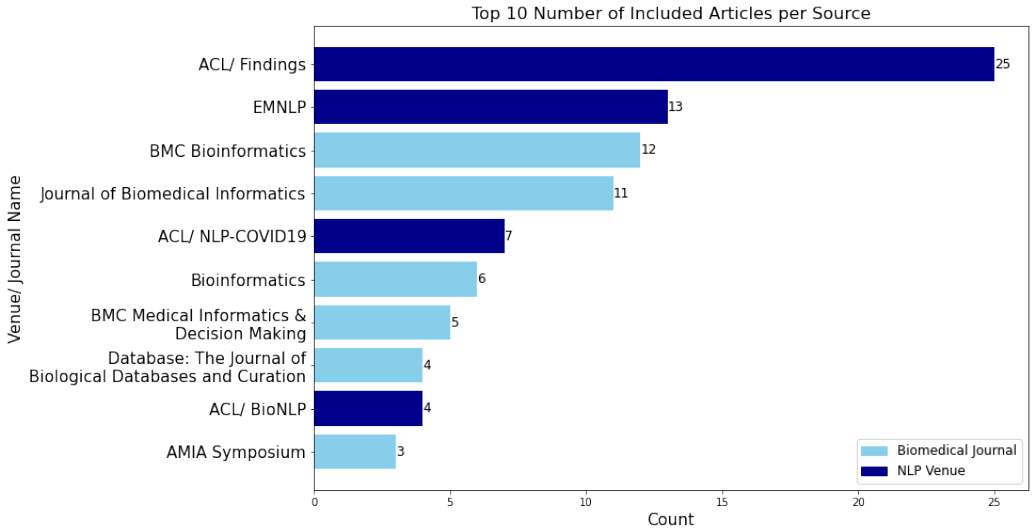

Figure 3. Number Articles Over Venue/Journal.

data sources used to inform the training and refinement phases of the LLMs development.

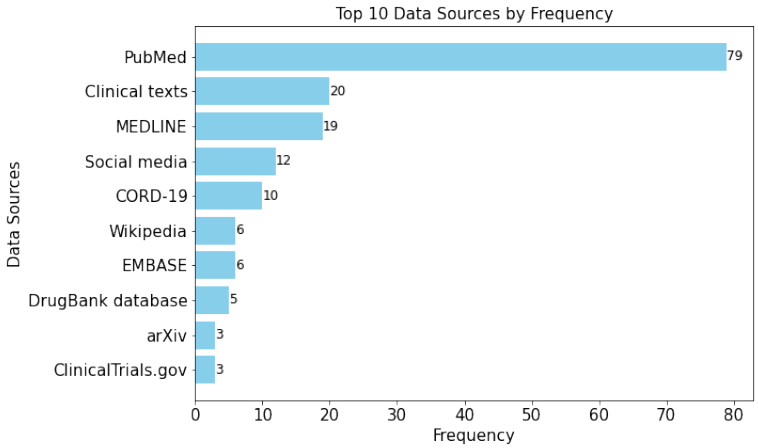

Figure 4. Frequency of specific data sources used for model development and testing.

Here we summarize the most prevalent ones:

1. **PubMed (79):** A free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. It is widely used for its comprehensive coverage of research articles, including journals and online books.

2. **Clinical Texts (20):** 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 (19):** 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.

4. **Social Media (12):** 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 (10):** The COVID-19 Open Research Dataset is a resource of over 280,000 scholarly articles, including over 100,000 with full text, about COVID-19 and the coronavirus family of viruses.

Further, we can explore the the specific types of data used from these data sources. By understanding the nature and format of the information that LLMs are trained upon, we can further elucidate their potential capabilities and limitations. Figure 5 provides an overview of the top 10 data types to which LLMs have been applied.

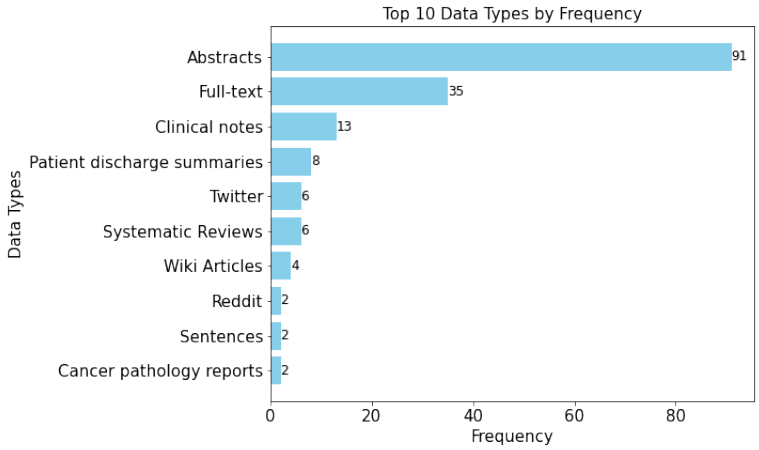

Figure 5. Frequency of specific data types used for model development and testing.

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 aids in efficient data processing and helps prevent information overload, allowing models to quickly grasp key findings and methodologies. However, full texts include comprehensive information such as experimental methodologies, detailed results, in-depth discussions, and conclusions. The necessity of such level of detail was highlighted in papers developing applications for advanced analytical tasks.

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 like post-treatment care and cancer diagnosis. However, the relative rarity might be due to privacy concerns and the need for de-identification, which makes acquiring large datasets challenging.

The use of social media data, while not predominant, indicates an interest in public health trends, patient experiences, and community discussions. The lower frequency could be due to the unstructured and often noisy nature of social media data, which can pose challenges for processing and analysis.

Biomedical Application Domains

Based on the paper contents, we manually devised 11 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.

The domain of General Biomedical Text Mining leads with 45 articles, indicating a significant emphasis on developing and improving methodologies for various NLP tasks, such as Named Entity Recognition and dependency parsing, in biomedicine. This is followed by Literature-Based Discovery with 25 articles, highlighting the importance of recommendation and retrieval systems in managing and synthesizing the vast amounts of scientific literature.

The Pharma domain, with 18 articles, reflects the growing interest in applying LLMs to drug discovery, development, and optimization, underscoring the potential of these models in accelerating and enhancing

Domain Category	Definition
General biomedical text mining	Developing a new/improved methodology for an NLP task (e.g., NER, dependency parsing).
Literature-Based Discovery	Recommendation, summarization, retrieval systems for relevant scientific literature or articles based on user preferences, research topics, or queries.
Pharma	Drug discovery, development, and optimization.
Literature Screening	Reviewing and categorizing scientific literature to identify relevant information for specific research or applications.
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.
Biomedical Literature Curation	Approaches to enrich text to improve a downstream NLP task (e.g., MeSH terms annotation, quality assessment).
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.
Precision Medicine	Tailoring medical treatments and interventions based on individual patient characteristics.

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

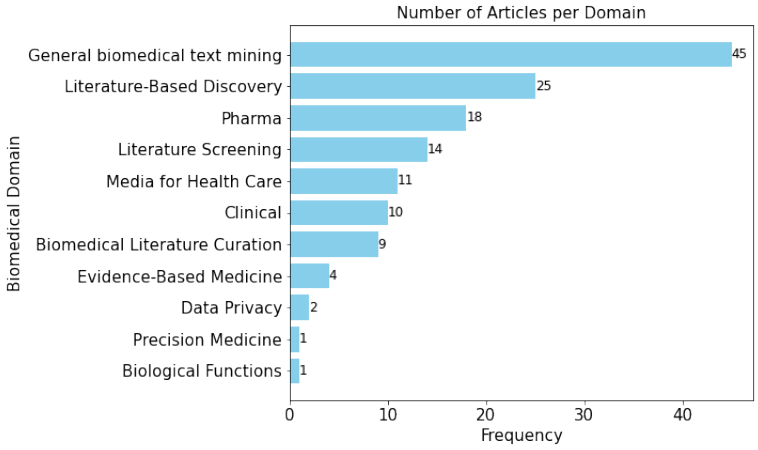

Figure 6. Number of articles assigned to each domain.

pharmaceutical research. Literature Screening, having 14 articles, shows the relevance of LLMs in reviewing and categorizing scientific literature, an essential process in research and evidence gathering.

Media for Health Care and Clinical domains, with 11 and 10 articles respectively, point towards a growing trend in utilizing LLMs for analyzing health-related information from media content and enhancing clinical decision-making and patient care. Biomedical Literature Curation, with 9 articles, signifies the role of LLMs in enriching biomedical texts to improve downstream NLP tasks.

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. Fields like Evidence-Based Medicine and Precision Medicine might be encountering unique challenges in integrating LLMs, such as the need for highly specialized datasets, ethical considerations, and the complexity of interdisciplinary data integration. The usage of LLMs for understanding low-level biological processes

also has limitations as there might be the need of integrating non-textual data to perform the analysis. Furthermore, there is still the prevalence of using rule-based approaches based on text patterns in some fields (Elangovan et al., 2022).

Overall, these results underscore a diverse range of applications for LLMs in biomedicine, with a clear focus on text mining, literature handling, and pharmaceutical applications, and emerging interests in more specialized areas.

NLP tasks

Figure 7 shows the distribution of various NLP tasks addressed by LLMs in the biomedical field.

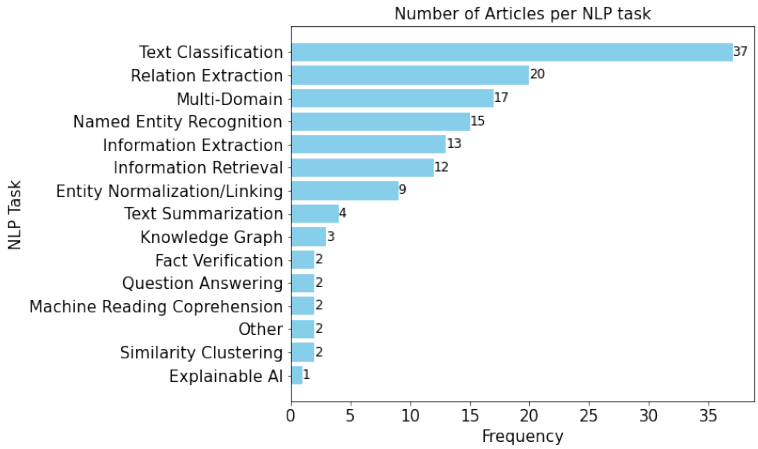

Figure 7. Number of articles assigned to each domain.

The sankey diagram in Fig. 8 allows to analyse which domains are most active in specific NLP applications. These relationships demonstrate the diversity of NLP tasks applied across various biomedical domains. Furthermore, 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.

General biomedical text mining encompasses a wide range of tasks, with the Multi-Domain category being the most prominent. This refers to methods or architectures developed to tackle multiple NLP tasks using a single unified approach. This concept is of high importance in the field of NLP, as it signifies a move towards more versatile and efficient processing models.

Literature Screening benefits the most from Text Classification, emphasizing its requirement for efficiently categorizing literature. This task involves assigning predefined categories or labels to text documents, enabling efficient organization and retrieval of the literature.

Literature-Based Discovery showcases a varied application of NLP, with Information Retrieval being prominent for sifting through extensive literature, alongside applications in Text Summarization, Knowledge Graph creation, and Machine Reading Comprehension. This suggests a strategic use of NLP in enhancing literature analysis and information discovery processes.

Lastly, Media for Health Care, though less dominant, shows an application of NLP in Text Classification, Fact Verification, and Information Retrieval, potentially indicating an emerging interest in applying NLP for media analysis in healthcare contexts.

The distribution of NLP tasks across different biomedical domains showcases the broad applicability of NLP techniques in extracting meaningful information, facilitating knowledge discovery, and supporting decision-making processes in the biomedical field.

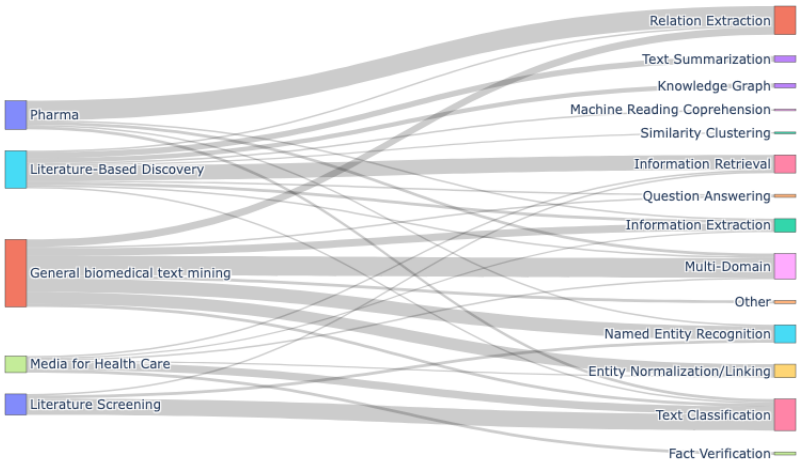

Figure 8. Sankey Diagram representing the relationships between the biomedical domains and the utilized NLP applications. Each domain is represented as a source node, while the associated NLP applications are shown as the target nodes. The thickness of the flows between the two is proportional to the number of articles that exhibit this connection.

Large Language Models

Models Overview

Figure 9 provides an overview for the most frequently used LLM models across all evaluated papers. The results show a focus on BERT (Bidirectional Encoder Representations from Transformers) architectures, which is a widely adopted language representation model in the field of NLP. It has achieved state-of-the-art results on a range of NLP tasks by leveraging large-scale unsupervised pretraining followed by task-specific fine-tuning.

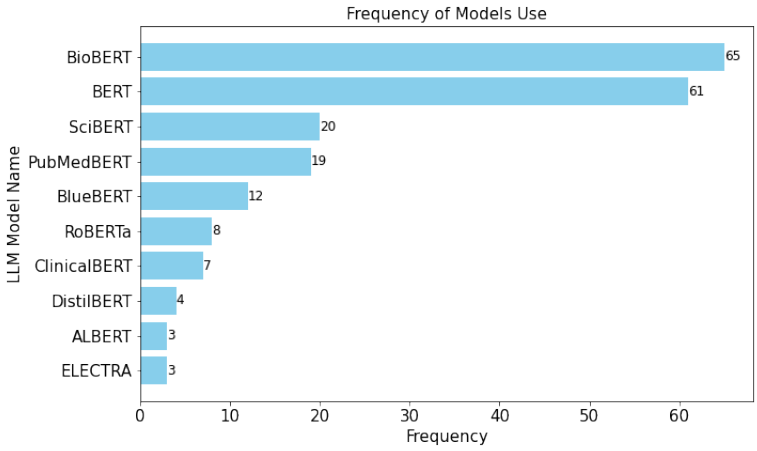

Figure 9. Most frequently used models.

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.

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

Table 2. Training Corpora for Different Models

Trends Over Time

Figure 10 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. Newer models like PubMedBERT and ASPIRE also emerge in later years, reflecting continuous innovation in the field. However, some models like ClinicalBERT and NCBI BERT show more sporadic usage, suggesting niche applications. Overall, the data underscores a dynamic landscape in language model utilization, with a clear trend towards specialization and new model adoption.

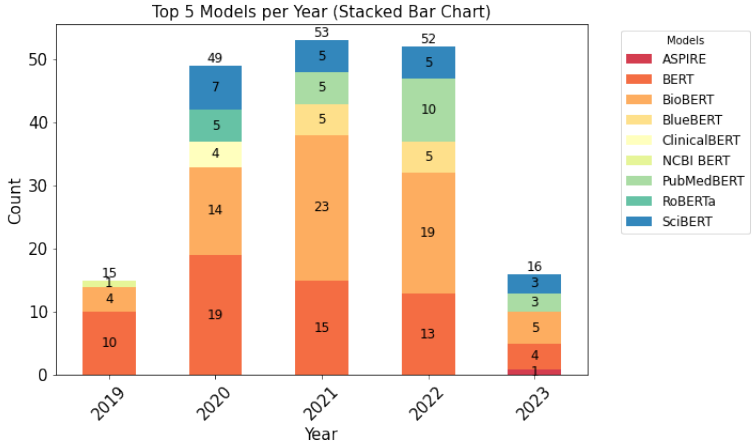

Figure 10. Top 5 most frequently utilized models for each year.

Figure 11 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.

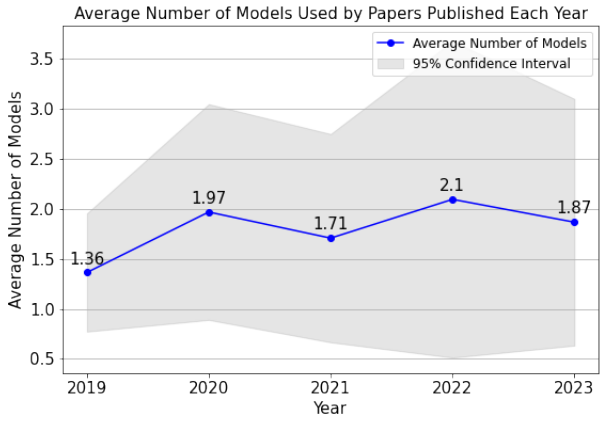

Figure 11. Average number of different models used per paper each year.

Technical setup

Figure 12 presents a summary of the hardware types reported in the development of large language models. The data reveals a significant reliance on GPUs (49 instances) 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. The relatively lower but notable usage of TPUs (4 instances), known for their even higher specialization in deep learning, suggests a trend towards hardware that is explicitly designed for machine learning tasks.

Interestingly, the combination of GPUs and CPUs (6 instances) indicates an approach to leverage the general-purpose nature of CPUs along with the specialized capabilities of GPUs, possibly to balance cost and efficiency. The minimal use of CPUs alone (2 instances) underscores their limited efficiency for large-scale deep learning tasks compared to GPUs and TPUs.

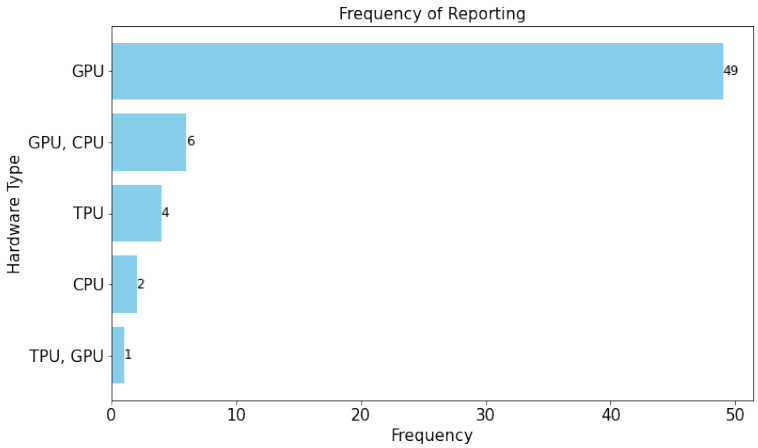

Figure 12. Reported hardware used.

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 13 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. This dominant position

of Python suggests a field that values accessibility and efficiency, leveraging Python's comprehensive ecosystem for AI and machine learning.

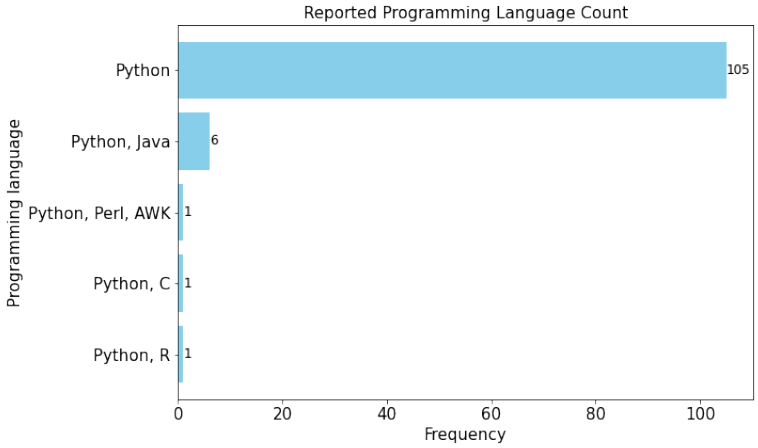

Figure 13. Reported programming language used.

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 14 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 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.

TensorFlow's significant usage reflects its robustness and scalability, particularly in production environments. Its comprehensive ecosystem and support from Google further enhance its appeal. 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.

Preprocessing

Fine-Tuning Tasks and Datasets

Well-established Benchmark Datasets

Open Shared Tasks

While Table 3 aggregates the most common standard benchmarks used for models evaluation, several papers worked with datasets coming from ongoing challenges and shared tasks. Open shared tasks are collaborative efforts where researchers and developers from around the world focus on solving a specific

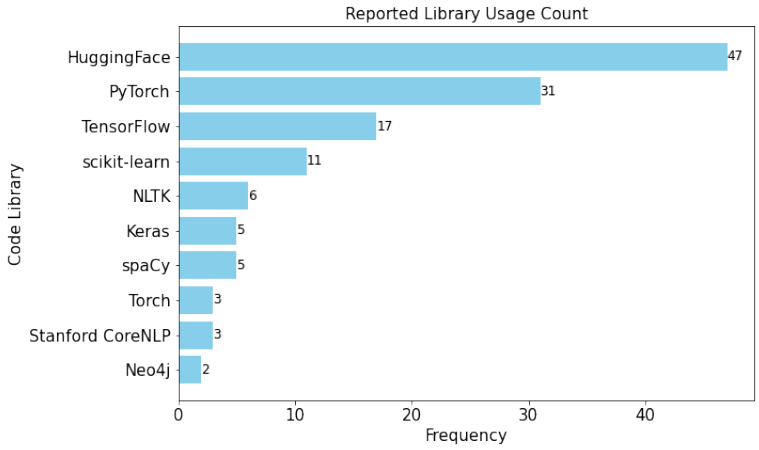

Figure 14. Reported programming language used.

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

challenge within the field of NLP. These tasks provide standardized datasets and evaluation metrics, allowing different teams to compare their approaches on a level playing field.

One example are the Informatics for Integrating Biology and the Bedside (i2b2) challenges that focus on clinical data, including aspects like temporal relations in clinical narratives (2012) and de-identification (2014). BioCreative is another workshop that hosts challenges related to the extraction and annotation of biological entities (like genes, proteins) and relationships from scientific literature. The BioCreative-LitCovid and DrugProt has been used for a few applications. Finally, the Text REtrieval Conference (TREC) is an ongoing series of workshops focusing on a range of information retrieval topics. The TREC-COVID and Health Misinformation datasets have been used in several papers.

Custom-developed Datasets

TODO: report and analyse the development of custom datasets! see also appendix data

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.
- 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 15 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.

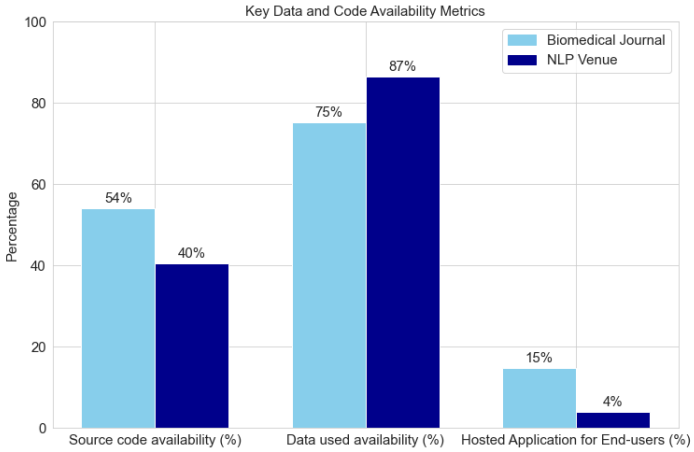

Figure 15. Transparency of LLM Methods: Source Code, Data, and Hosted Application Availability..

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

CONCLUSION

ACKNOWLEDGMENTS

Additional information can be given in the template, such as to not include funder information in the acknowledgments section.

APPENDIX

REFERENCES

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*

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.

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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*.

Elangovan, A., Li, Y., Pires, D. E., Davis, M. J., and Verspoor, K. (2022). Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated biobert. *BMC bioinformatics*, 23:1–23.

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.

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.

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.

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.