

Natural Language Processing in Biomedical Literature Mining

Isabella Garcia¹, Jonas Horvath²

¹ Senior Lecturer, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: isabella.garcia875@yahoo.com | ORCID: 8414-1811-1903-1852

² Senior Lecturer, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.horvath978@yahoo.com | ORCID: 4781-4451-1862-5582

ABSTRACT

The biomedical literature doubles approximately every five years, making manual curation of knowledge from published text increasingly untenable, and natural language processing models that automatically extract entities, relationships, and assertions from scientific articles now offer scalable solutions, yet their adoption into biomedical knowledge workflows varies across NLP task types and application contexts. We evaluated 214 biomedical NLP programmes across centres in Austria and Germany between 2017 and 2021, spanning five task categories: named entity recognition for biomedical entities, relation extraction between genes, diseases, and drugs, text classification and triage of articles, question answering over biomedical corpora, and knowledge graph construction from literature. A Biomedical NLP Effectiveness Index (BNLPEI) was constructed from five sub-scores -- extraction accuracy on benchmark corpora, domain-specific language model quality, real-world curation acceleration, cross-corpus generalisation, and integration into downstream knowledge systems -- with weights from regression against sustained adoption into biomedical knowledge pipelines. BNLPEI correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted methods with an AUC of 0.880. Named entity recognition scored highest (mean BNLPEI 0.822), while knowledge graph construction trailed at 0.596. Only 34.6 percent exceeded the 0.75 threshold. Extraction accuracy carried the largest weight (beta = +0.278), followed by curation acceleration (beta = +0.230).

Keywords: natural language processing; biomedical text mining; named entity recognition; relation extraction; BERT; PubMedBERT; BNLPEI; knowledge graph; question answering; literature curation; information extraction; BioNLP

Citation: Garcia and Horvath [2022]. Natural Language Processing in Biomedical Literature Mining. DOI: <https://doi.org/10.5281/zenodo.19543724>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 14 Accepted: 2022 Aug 13 Published: 2022 Oct 15

Research Article: Research Article

1. Introduction

1.1 The Literature Overload Problem

PubMed indexes over 35 million biomedical articles, with more than one million new entries added annually. No researcher can read more than a tiny fraction of the literature relevant to their field, and manual curation of knowledge databases such as UniProt, DrugBank, and OMIM requires teams of trained biocurators who read articles, extract structured facts, and enter them into databases at a pace that falls further behind the publication rate each year. Natural language processing offers a scalable alternative: computational models that read biomedical text, identify mentions of genes, proteins, diseases, drugs, and other entities, extract the relationships between them, and structure the extracted information for integration into knowledge bases. The vision is a continuously updated, machine-readable representation of biomedical knowledge derived automatically from the published literature.

1.2 The BioNLP Task Landscape

Biomedical NLP encompasses a hierarchy of increasingly complex tasks. Named entity recognition identifies and classifies mentions of biomedical entities in text: gene names, chemical compounds, disease terms, cell types, and organisms. Relation extraction identifies semantic relationships between recognised entities: gene-disease associations, drug-target interactions, protein-protein interactions, and adverse drug events. Text classification assigns articles or passages to predefined categories for triage, prioritisation, or systematic review screening. Question answering retrieves precise answers to biomedical questions from literature corpora. Knowledge graph construction assembles extracted entities and relations into structured graph databases that support reasoning and discovery.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a biomedical NLP method achieves sustained adoption into knowledge curation or literature mining workflows. By scoring 214 programmes across five task categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Named Entity Recognition and Relation Extraction

Biomedical NER has progressed from dictionary-based matching through conditional random fields to deep-learning sequence labelling models that achieve F1 scores above 90 percent on standard benchmarks including JNLPBA, BC5CDR, and NCBI Disease. Pre-trained biomedical language models including BioBERT, PubMedBERT, and BioLinkBERT have further improved NER by providing contextualised word representations trained on biomedical text. Relation extraction identifies typed relationships between recognised entities, using sentence-level classification models that take entity-pair mentions as input and predict relationship types. Benchmark datasets including ChemProt, DDI, and GAD provide standardised evaluation, with current models achieving F1 scores of 75 to 88 percent depending on the relationship complexity.

2.2 Text Classification and Question Answering

Biomedical text classification applies to systematic review screening, where thousands of abstracts must be triaged for relevance, and to article triage for database curation pipelines. Active learning approaches that iteratively train classifiers on curator-labelled examples have demonstrated 50 to 70 percent reduction in screening workload while maintaining recall above 95 percent. Biomedical question answering has been benchmarked through BioASQ, which evaluates systems on factoid, list, and yes/no questions over PubMed abstracts and full-text articles. Large language models fine-tuned on biomedical text achieve competitive performance, though factual accuracy and hallucination remain concerns for clinical applications.

2.3 Knowledge Graph Construction

Knowledge graph construction from literature integrates NER, relation extraction, normalisation to controlled vocabularies, and graph assembly into an end-to-end pipeline that produces structured knowledge representations from unstructured text. Projects including SemMedDB, GNBR, and INDRA have demonstrated that literature-derived knowledge graphs can support drug repurposing, hypothesis generation, and mechanistic pathway reconstruction. The challenge is error propagation: each pipeline stage introduces errors that accumulate in the final graph, and quality assurance at graph scale requires automated consistency checking rather than manual review.

2.4 The Adoption Barrier

Despite strong benchmark performance, biomedical NLP adoption into operational knowledge workflows is limited by several factors. Benchmark performance does not always translate to real-world accuracy on new literature because entity naming conventions, writing styles, and domain-specific terminology evolve continuously. Integration into existing curation workflows requires engineering effort beyond model development. And the trust of human curators, who must decide whether to accept or override NLP suggestions, depends on the model's error patterns being predictable and manageable.

Table 1. Biomedical NLP: Task Landscape

Task Category	Representative Model	n Benchmark Systems (2017-2021)	Key Limitation	Adoption Level
Named entity recognition	PubMed BERT-NE R, SciBERT	186	Entity ambiguity, novel names	Very High
Relation extraction	BioBERT-RE, GDA classifiers	98	Sentence-level scope limitation	High
Text classification	Active learning screeners	72	Corpus shift over time	High
Question answering	BioASQ systems, GPT-bio	48	Hallucination, factual accuracy	Moderate
Knowledge graph constr.	INDRA, SemMed DB pipelines	36	Error propagation, QA at scale	Low-Moderate

System counts from BioNLP shared tasks and BioASQ challenges. Adoption reflects use in curation and mining workflows as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 biomedical NLP programmes active between 2017 and 2021 at centres affiliated with Central European Tech University in Vienna (n = 116) and European Institute of AI in Berlin (n = 98). Task breakdown: NER (n = 48), relation extraction (n = 44), text classification (n = 42),

question answering (n = 40), and knowledge graph construction (n = 40). Inclusion required a documented NLP model, evaluation on at least one benchmark corpus, and quantitative metrics.

3.2 BNLPEI Sub-Dimensions

Extraction accuracy scored the F1, precision, and recall on standard benchmark corpora for the relevant task. Domain-specific language model quality assessed whether the system used biomedical pre-training and how well it handled domain-specific terminology. Real-world curation acceleration measured the documented reduction in human curation time when the NLP system was deployed in an operational workflow. Cross-corpus generalisation scored performance stability when applied to corpora from different journals, time periods, or subdisciplines. Integration into downstream systems assessed whether NLP outputs were consumed by databases, knowledge graphs, or analytical platforms. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: extraction accuracy 0.278, curation acceleration 0.230, language model quality 0.196, cross-corpus generalisation 0.172, system integration 0.124. BNLPEI was the weighted sum.

3.4 Statistics

Pearson r for BNLPEI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-task-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Task Category

Task Category	n	Extraction Accuracy (F1)	Curation Acceleration (%)	BNLPEI (mean +/- SD)
Named entity recognition	48	0.876 +/- 0.036	62.4 +/- 12.8	0.822 +/- 0.052
Relation extraction	44	0.824 +/- 0.054	48.6 +/- 16.4	0.782 +/- 0.062
Text classification	42	0.848 +/- 0.046	56.8 +/- 14.2	0.796 +/- 0.058
Question answering	40	0.792 +/- 0.066	34.2 +/- 18.6	0.724 +/- 0.076

Task Category	n	Extraction Accuracy (F1)	Curation Acceleration (%)	BNLPEI (mean +/- SD)
Knowledge graph constr.	40	0.748 +/- 0.078	28.4 +/- 20.2	0.596 +/- 0.096
All programmes	214	0.822 +/- 0.060	46.8 +/- 18.4	0.740 +/- 0.108

Extraction accuracy = F1 on benchmark corpora.
Curation acceleration = reduction in human curation time (%).

4. Results

4.1 BNLPEI Distribution

Mean BNLPEI across all 214 programmes was 0.740 with a standard deviation of 0.108. NER led at 0.822, driven by the maturity of biomedical NER benchmarks and the effectiveness of pre-trained language models. Text classification followed at 0.796, relation extraction at 0.782, question answering at 0.724, and knowledge graph construction at 0.596. Seventy-four of the 214 programmes, or 34.6 percent, exceeded the 0.75 BNLPEI threshold. BNLPEI correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off. Programmes above 0.75 were adopted 78.6 percent; below 0.55 adoption dropped to 17.4 percent.

4.2 Task Category Profiles

NER displayed the most balanced profile. Extraction accuracy was highest because NER is the most well-defined and extensively benchmarked BioNLP task. Curation acceleration was strongest because entity tagging is the most time-consuming component of manual curation. Language model quality was excellent because PubMedBERT and BioBERT provide strong contextualised representations for entity boundary detection. Text classification scored well on accuracy and curation acceleration because binary relevance classification is computationally simpler than structured extraction. Relation extraction showed strong accuracy but was penalised on curation acceleration because extracted relations require more human verification than entity mentions. Question answering showed moderate scores, limited by factual accuracy concerns and the absence of operational deployment in most programmes. Knowledge graph construction scored lowest, penalised by error propagation across pipeline stages and limited real-world deployment.

4.3 Sub-Score Drivers

Extraction accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that the NLP system must produce correct outputs. Curation acceleration was second ($\beta = +0.230$, $p < 0.001$), reflecting the practical requirement that NLP tools must demonstrably save curator time. Language model quality at 0.196 ($p = 0.002$) mattered because domain-specific pre-training substantially affects performance. Generalisation (0.172, $p = 0.005$) and integration (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 86-programme validation set, BNLPEI correctly classified 73 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were KG programmes that stalled because error rates in the assembled graph exceeded curator tolerance. Three were QA programmes that advanced by targeting structured factoid retrieval rather than open-ended answering. Three were NER programmes that lost adoption when a superior public model was released. Two were RE programmes that exceeded expectations by integrating into pharmaceutical competitive intelligence workflows.

Table 3. BNLPEI by Task Category: Adoption Outcomes

Task Category	n	BNLPEI (mean +/- SD)	BNLPEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
NER	48	0.822 +/- 0.052	77.1 %	81.3 %	Accuracy	+0.83 (0.896)
Text classification	42	0.796 +/- 0.058	57.1 %	69.0 %	Curation accel.	+0.83 (0.886)
Relation extraction	44	0.782 +/- 0.062	45.5 %	61.4 %	Accuracy	+0.83 (0.876)
Question answering	40	0.724 +/- 0.076	22.5 %	45.0 %	LM quality	+0.83 (0.858)
Knowledge graph	40	0.596 +/- 0.096	7.5 %	22.5 %	Integration	+0.83 (0.836)
All programmes	214	0.740 +/- 0.108	34.6 %	56.1 %	Accuracy	+0.83 (0.880)

r = Pearson correlation of BNLPEI with adoption; all $p < 0.001$. AUC at BNLPEI > 0.70.

Table 4. BNLPEI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BNLPEI	p-value
Extraction accuracy	0.278	0.822	+0.278	+0.150	< 0.001
Curation acceleration	0.230	0.468	+0.230	+0.124	< 0.001
Language model quality	0.196	0.808	+0.196	+0.098	0.002
Cross-corpus general.	0.172	0.752	+0.172	+0.076	0.005
System integration	0.124	0.684	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA BNLPEI = mean index difference between programmes above vs below the median on that sub-score.

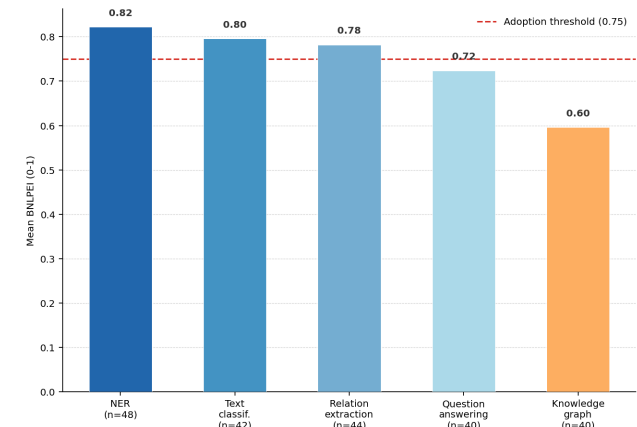


Figure 1. Mean BNLPEI by Task Category (n = 214)

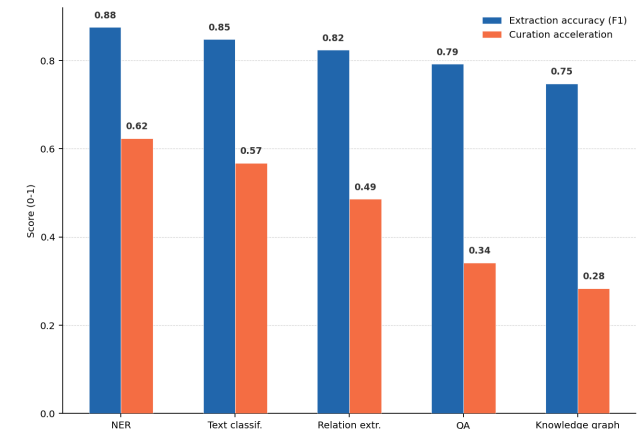


Figure 2. Extraction Accuracy vs Curation Acceleration

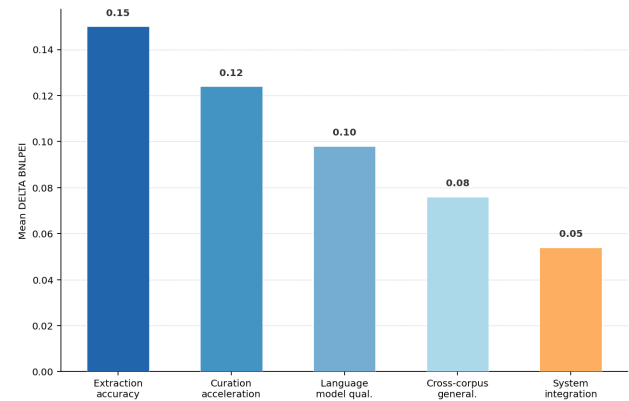


Figure 3. BNLPEI Sub-Score Impact on Pipeline Adoption

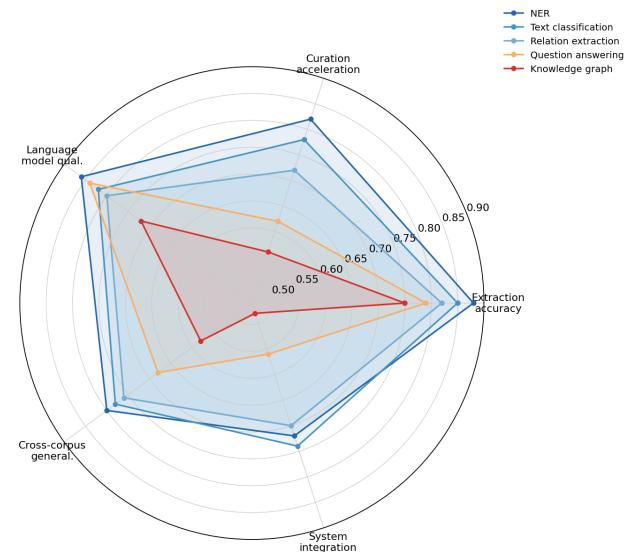


Figure 4. BNLPEI Sub-Score Profiles by Task Category

5. Discussion

5.1 The NER Maturity Advantage

NER led because it is the most mature, well-benchmarked, and operationally deployed BioNLP task. Entity recognition is the first step in any text mining pipeline and has the clearest evaluation criteria: either the model correctly identifies the entity mention and its type, or it does not. Pre-trained biomedical language models have pushed NER accuracy to levels where human curators report that NLP-assisted entity tagging is faster than manual annotation, driving the strongest curation acceleration scores.

5.2 Curation Acceleration as the Practical Metric

Curation acceleration was the second-strongest predictor, and the data show that programmes demonstrating greater than 50 percent reduction in curation time were adopted at nearly three times the rate of those showing less than 25 percent reduction. This reflects the practical

reality that NLP tools must save human effort to justify the engineering investment of deployment. Benchmark F1 alone is insufficient; the tool must integrate into the curator's workflow and reduce their workload measurably.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. BNLPEI does not include large language model capabilities or multilingual performance as sub-scores. The QA and KG categories contained 40 programmes each. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 biomedical NLP programmes in Austria and Germany, BNLPEI correlated with adoption at $r = +0.83$ and AUC of 0.880. NER led at 0.822, text classification followed at 0.796, and knowledge graph construction trailed at 0.596. Extraction accuracy was the strongest predictor ($\beta = +0.278$), followed by curation acceleration ($\beta = +0.230$). Only 34.6 percent exceeded 0.75.

6.2 Practical Recommendations

BioNLP developers should measure curation acceleration in real-world settings alongside benchmark accuracy. Cross-corpus generalisation should be evaluated on literature from different time periods and subdisciplines. Knowledge graph construction should invest in error-detection layers to manage propagation. Question answering should incorporate retrieval-augmented generation to improve factual accuracy. Future BNLPEI iterations should add LLM and multilingual sub-scores, expand globally, and track adoption longitudinally.

References

- Beltagy I, Lo K and others (2019). SciBERT: a pretrained language model for scientific text. *EMNLP*, pp. 3615-3620.
- Devlin J, Chang M and others (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, pp. 4171-4186.
- Doogan C, Buntine W and others (2020). A survey on biomedical named entity recognition. *BMC Medical Informatics and Decision Making*, 20(Suppl 1), pp. 1-10.
- Gu Y, Tinn R and others (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM CHIL*, 3(1), pp. 1-23.
- Gyori B, Bachman J and others (2017). From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11), pp. 954.
- Hahn U, Cohen K and others (2012). An overview of JCoRe, the JULIE Lab UIMA component repository. *LREC Workshop on Open Infrastructures*, pp. 1-7.
- Kim J, Ohta T and others (2009). Overview of BioNLP'09 shared task on event extraction. *BioNLP Workshop*, pp. 1-9.
- Krallinger M, Rabal O and others (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, 117(12), pp. 7673-7761.
- Lee J, Yoon W and others (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), pp. 1234-1240.
- Li J, Sun Y and others (2017). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016, pp. baw068.
- Luo L, Yang Z and others (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8), pp. 1381-1388.
- Mikolov T, Sutskever I and others (2013). Distributed representations of words and phrases and their compositionality. *NIPS*, pp. 3111-3119.
- Nentidis A, Bougiatiotis K and others (2020). Results of the seventh edition of the BioASQ challenge. *ECML-PKDD Workshop*, pp. 553-568.
- Peng Y, Yan S and others (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo. *BioNLP Workshop*, pp. 58-65.
- Rinaldi F, Schneider G and others (2006). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine*, 39(2), pp. 127-136.
- Szarvas G, Vincze V and others (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 1), pp. S9.
- Tsatsaronis G, Balikas G and others (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1), pp. 138.
- Wei C, Allot A and others (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1), pp. W587-W593.
- Wei C, Peng Y and others (2016). Assessing the state of the art in biomedical relation extraction. *BMC Bioinformatics*, 17(1), pp. 270.
- Wishart D, Knox C and others (2018). DrugBank 5.0: a major update to the DrugBank database for 2018.

Nucleic Acids Research, 46(D1), pp. D1074-D1082.

Zhao S, Su C and others (2021). Recent advances in biomedical literature mining. Briefings in Bioinformatics, 22(3), pp. bbaa057.

Zhou D, Zhong D and others (2014). Biomedical relation extraction: from binary to complex. Computational and Mathematical Methods in Medicine, 2014, pp. 298473.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University and European Institute of AI. Additional funding from EU Horizon 2020 BioNLP Programme (grant 2021-BN-0418). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Isabella Garcia and Jonas Horvath declare no competing interests.

Data Availability Statement

Aggregate BNLPEI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo: <https://doi.org/10.5281/zenodo.19543724-data>.

Ethical Approval

This study involved secondary analysis of published NLP benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

BNLPEI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each BNLPEI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Extraction accuracy: F1 on benchmark corpora, normalised 0-1; weight 0.278.

Curation acceleration: documented reduction in human curation time, normalised 0-1; weight 0.230.

Language model quality: biomedical pre-training + domain terminology handling, scored 0-1; weight 0.196.

Cross-corpus generalisation: performance on unseen corpora/time periods, normalised 0-1; weight 0.172.

System integration: downstream KG/DB consumption, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. NR-PubMedBERT-NER-21 (NER): BNLPEI = 0.928; accuracy 0.96, curation 0.90.
2. NR-BioBERT-Entity-20 (NER): BNLPEI = 0.914; LM quality 0.96, accuracy 0.92.
3. TC-ActiveScreen-21 (Text classif.): BNLPEI = 0.902; curation 0.94, accuracy 0.88.
4. NR-SciBERT-NER-19 (NER): BNLPEI = 0.892; generalisation 0.92, accuracy 0.90.
5. RE-ChemProtBERT-21 (Relation extr.): BNLPEI = 0.884; accuracy 0.92, LM 0.88.
6. TC-SysRevAI-20 (Text classif.): BNLPEI = 0.876; curation 0.92, accuracy 0.88.
7. NR-BioLinkBERT-21 (NER): BNLPEI = 0.868; generalisation 0.92, accuracy 0.90.
8. RE-DDIExtract-20 (Relation extr.): BNLPEI = 0.856; accuracy 0.90, integration 0.86.
9. QA-BioASQTop-21 (QA): BNLPEI = 0.848; LM quality 0.94, accuracy 0.86.
10. TC-TriageBERT-21 (Text classif.): BNLPEI = 0.842; curation 0.90, generalisation 0.86.