

Transformer-Based Language Models for Contextual Information Extraction From Scientific Documents

Dr. Aino Virtanen

Associate Professor, Helsinki Institute of Artificial Intelligence Research, Finland

Abstract

Scientific literature contains substantial quantities of specialized information concerning methods, materials, datasets, processes, findings, and relationships among research concepts. Automated extraction of this information is challenging because scientific terminology is highly domain-specific and the interpretation of individual expressions frequently depends on surrounding textual context. Conventional information-extraction approaches often rely on manually engineered features or fixed word representations, which may inadequately represent context-dependent meanings. Transformer-based language models provide an alternative by producing contextualized representations in which the interpretation of words and phrases depends on their surrounding linguistic environment. This hypothetical methodological study examines transformer-based language models for contextual information extraction from scientific documents. The proposed framework evaluates named scientific entities and semantic relations using contextual representations. A fictional corpus containing 900 scientific documents is divided into hypothetical training, validation, and testing partitions. Three conceptual extraction configurations are compared: a BiLSTM-CRF baseline using conventional embeddings, a general BERT-based extractor, and a scientific-domain transformer model.

The illustrative analysis evaluates entity extraction and relation extraction separately through precision, recall, and F1 score. Under the assigned hypothetical values, the scientific-domain transformer records the strongest entity and relation extraction performance. The improvement is conceptually associated with domain-specific contextual representation rather than with transformer architecture alone. The numerical results are illustrative and are not outputs obtained from implemented models. The study concludes that transformer-based language models provide a useful methodological foundation for scientific information extraction because they can represent context-dependent terminology and relationships more effectively than fixed word representations. However, practical performance depends on annotation quality, disciplinary vocabulary, document length, relation complexity, and model generalization. Reproducible experiments using real annotated scientific corpora are therefore necessary before claims of extraction superiority can be established.

Keywords: Transformer Language Models, Scientific Information Extraction, BERT, SciBERT, Named Entity Recognition, Relation Extraction, Scientific Documents, Contextual Embeddings, Natural Language Processing, Knowledge Extraction.

1. Introduction

Scientific publishing produces a continuously expanding body of documents containing information about research methods, experimental procedures, datasets, theoretical concepts, materials, and findings. Manually identifying relevant information within large collections can become difficult when researchers must examine thousands of articles. Automated information extraction provides computational methods for transforming unstructured scientific text into structured entities and relations. Such systems can support literature analysis, scientific knowledge graphs, document retrieval, and evidence organization when their outputs accurately represent the information expressed within source documents.

Earlier natural language processing approaches commonly represented words through sparse features, manually designed linguistic indicators, or fixed distributed vectors. Word-embedding methods improved semantic representation by mapping related words into continuous vector spaces, but a single representation was generally assigned to a word regardless of context. Scientific terminology creates difficulties for such approaches because the meaning or function of an expression may vary across disciplinary and sentence contexts. Context-sensitive representation therefore became increasingly important for extracting scientific concepts and their relationships.

Transformer architectures changed sequence modeling by relying primarily on attention mechanisms rather than recurrent processing. Vaswani et al. introduced the Transformer architecture, while Devlin et al. subsequently developed BERT for deep bidirectional contextual representation. These developments enabled each token representation to incorporate information from surrounding words. Such contextualization is particularly relevant to scientific information extraction, where distinguishing a method, material, task, metric, or dataset may require interpretation of the complete sentence rather than the target expression alone.

Domain adaptation represents another important development because general-language models may not fully represent specialized scientific vocabulary. Beltagy et al. introduced SciBERT, which was pretrained on a large collection of scientific publications. Luan et al. developed SciERC and a multi-task framework for identifying entities, relations, and coreference in scientific documents. These studies demonstrated the value of domain-specific representations and task-oriented scientific annotations. They also established scientific information extraction as a distinct challenge involving terminology and relational structures not always represented adequately in general text corpora.

The present hypothetical study examines transformer-based contextual information extraction through a structured methodological framework. Three conceptual systems are compared: a conventional sequence-modeling baseline, a general transformer model, and a scientific-domain transformer. Entity and relation extraction are evaluated separately because improvements in entity recognition do not automatically imply equivalent improvements in relationship identification. The study uses fictional data and illustrative calculations to demonstrate evaluation procedures rather than presenting unverified experimental findings.

2. Objectives of the Study

The primary objective of this study is to examine the methodological potential of transformer-based language models for extracting context-dependent information from scientific documents. The research focuses on entity identification, semantic relation extraction, contextual representation, and challenges associated with specialized scientific language.

2.1 Examination of Contextual Scientific Language Representation

The first objective is to examine how contextualized transformer representations may support identification of scientific concepts whose meanings depend on surrounding text. Scientific expressions may describe methods, materials, tasks, datasets, metrics, or other research entities according to their sentence context. Transformer-based models can conceptually represent these differences through attention across neighboring tokens. The study therefore evaluates contextual modeling as an alternative to fixed word representations commonly used by earlier sequence-labeling approaches.

The objective also examines the potential value of domain-specific pretraining. General-language transformers learn broad linguistic relationships but may encounter specialized terminology rarely represented in ordinary text. Scientific-domain models are pretrained on research literature and may consequently develop representations more closely aligned with scholarly vocabulary and syntax. The study compares general and scientific-domain transformer configurations conceptually to demonstrate how domain adaptation can be evaluated rather than assuming that specialized pretraining always produces superior extraction.

2.2 Evaluation of Entity and Relation Extraction

The second objective is to evaluate scientific entity recognition independently from semantic relation extraction. Entity extraction determines the boundaries and categories of scientific concepts, whereas relation extraction identifies structured connections between recognized concepts. The tasks are related but involve different sources of difficulty. A model can identify two entities correctly while assigning an incorrect relationship between them. Separate evaluation therefore provides a clearer representation of system behavior than one aggregated information-extraction score.

The study compares hypothetical BiLSTM-CRF, general BERT, and scientific-domain transformer configurations using precision, recall, and F1 score. Entity evaluation uses a conceptual set of one thousand reference entity mentions, while relation evaluation uses six hundred reference relations. The purpose is to demonstrate internally consistent performance calculations under shared testing conditions. The values do not represent actual model outputs, and no conclusion regarding algorithmic superiority should be drawn from the illustrative differences.

2.3 Identification of Scientific Information-Extraction Challenges

The third objective is to identify difficulties that influence contextual extraction from scientific literature. These include specialized terminology, abbreviations, overlapping concepts, long-range dependencies, disciplinary differences, limited annotation, and complex relations extending across sentence boundaries. Scientific documents also vary substantially in structure and writing style. These characteristics create challenges for both conventional and transformer-based systems and can influence whether a model trained within one scientific domain transfers effectively to another.

The objective further considers reproducibility and evaluation reliability. Information-extraction experiments should use clearly defined entity and relation categories, independently separated training and testing data, consistent tokenization, and transparent matching criteria. Domain-specific transformer models should be compared with meaningful baseline systems rather than weak alternatives. Error analysis should accompany numerical metrics to determine whether performance differences arise from boundary detection, label confusion, terminology, or relation reasoning rather than from overall classification behavior alone.

3. Review of Literature

3.1 From Static Embeddings to Contextual Representation

Mikolov et al. demonstrated that neural representation learning could encode semantic relationships among words within continuous vector spaces. Pennington et al. subsequently introduced GloVe, combining global corpus statistics with distributed word representations. These methods significantly improved many natural language processing tasks. However, a word generally retained one principal vector regardless of its sentence context, creating difficulties for expressions with multiple meanings or context-dependent functional roles. Scientific language frequently contains such terminology, increasing the importance of contextual representation.

Peters et al. developed deep contextualized word representations in which embeddings were derived from the complete linguistic context. Transformer-based representation subsequently expanded this principle through self-attention. Vaswani et al. introduced the Transformer architecture, while Devlin et al. developed bidirectional pretraining through BERT. These advances shifted language modeling away from exclusively static representations and provided a foundation for task-specific fine-tuning. Contextual modeling became particularly valuable for sequence labeling and relation extraction because surrounding text could influence representation directly.

3.2 Scientific Information Extraction

Luan et al. introduced SciERC, an annotated scientific information-extraction dataset containing entities, relations, and coreference information. Their multi-task scientific information extractor demonstrated how shared span representations could support construction of scientific knowledge graphs. The work established a useful benchmark for evaluating extraction across research literature and highlighted the interconnected nature of entities and relations. Scientific information extraction consequently developed beyond isolated named entity recognition toward more integrated representations of research concepts and their connections.

Wadden et al. developed DyGIE++, a multi-task framework using contextualized span representations for entity, relation, and event extraction. Their approach incorporated local and cross-sentence contextual information and demonstrated the relevance of contextual embeddings for information-extraction tasks. Such research shows that identifying scientific knowledge requires more than recognizing isolated terminology. Relationship interpretation may depend on surrounding entities, sentence organization, coreference, and broader context, particularly when scientific claims extend across multiple textual units.

3.3 Domain-Specific Transformer Models

Beltagy et al. introduced SciBERT, a BERT-based language model pretrained on scientific publications from multiple domains. The model was evaluated across sequence tagging, sentence classification, and dependency parsing tasks and demonstrated the value of scientific-domain pretraining. Domain-specific models may provide improved representations of terminology, abbreviations, and scholarly syntax compared with models trained predominantly on general-language corpora. However, the magnitude of improvement depends on the downstream task and similarity between pretraining and target documents.

BioBERT extended transformer pretraining toward biomedical literature, while later scientific representation models explored document-level semantics and scholarly similarity. These developments demonstrate that domain adaptation can operate at several levels, including vocabulary, sentence

representation, entity extraction, and document retrieval. For contextual information extraction, the main methodological question is whether domain-specific representations improve recognition of scientific entities and relations beyond gains already provided by general contextual language models.

4. Research Methodology

4.1 Research Design and Model Framework

The present study adopts a hypothetical methodological research design centered on contextual information extraction from scientific text. Three conceptual model configurations are compared. The first uses a bidirectional recurrent sequence model with a conditional random field output layer. The second uses a general BERT-based contextual encoder with task-specific classification layers. The third uses a scientific-domain transformer conceptually based on SciBERT-style pretraining. None of these models was actually trained for the present manuscript.

Entity extraction is treated as a sequence- or span-labeling problem in which scientific concepts are assigned category labels. Relation extraction is treated as classification over candidate entity pairs. The transformer configurations use contextual token or span representations to predict labels, while the baseline relies on conventional word representations and recurrent sequence encoding. All models conceptually use the same train, validation, and test partitions to ensure that differences are not produced by unequal evaluation data.

4.2 Hypothetical Scientific Document Corpus

The fictional corpus contains 900 scientific documents representing computer science, engineering, biomedical informatics, and environmental research. Six hundred thirty documents are assigned to training, one hundred thirty-five to validation, and one hundred thirty-five to testing. The test partition conceptually contains one thousand annotated scientific entity mentions and six hundred annotated semantic relations. These figures are entirely hypothetical and do not represent SciERC, PubMed, or any other existing scientific corpus.

Entity categories conceptually include Method, Task, Dataset, Material, Metric, and Scientific Concept. Relation categories include Used-for, Evaluated-on, Part-of, Compared-with, and Associated-with. Annotation boundaries are assumed to have been created consistently for methodological illustration. A genuine study would require detailed annotation guidelines, inter-annotator agreement, disciplinary sampling information, and independent adjudication procedures. The present analysis does not claim that these activities were actually conducted.

4.3 Evaluation Procedure and Performance Measures

Entity and relation extraction are evaluated through precision, recall, and F1 score. Precision measures the proportion of predicted entities or relations that match reference annotations, while recall measures the proportion of reference annotations successfully identified. The F1 score combines both dimensions through their harmonic mean. An exact-match procedure is conceptually assumed for entity boundaries, meaning that partially overlapping spans are not counted as correct unless their complete boundaries and labels correspond to the reference.

The principal equations are $\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$, $\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$, and $\text{F1} = 2\text{PR}/(\text{P}+\text{R})$. Entity extraction uses one thousand conceptual reference mentions, while relation extraction uses six hundred conceptual reference relations. The numerical counts assigned below are designed only

to illustrate performance calculation. In a genuine investigation, repeated fine-tuning runs, confidence intervals, model-selection controls, and statistical comparison procedures would also be necessary.

5. Analysis

5.1 Hypothetical Scientific Entity Extraction Performance

870. False-positive predictions decrease across the configurations from 160 to 125 and 100.

The assigned pattern illustrates how contextual and domain-specific representations might be evaluated, but it does not represent performance obtained from an implemented scientific information-extraction experiment.

The scientific-domain transformer records hypothetical precision of 89.69%, recall of 87.00%, and F1 score of 88.32%. The general BERT configuration produces an F1 score of approximately 84.62%, while the conventional baseline records approximately 79.17%. These illustrative differences demonstrate how separate false-positive and false-negative counts influence the overall F1 score. A real experiment would require repeated model training and statistical evaluation before concluding that scientific-domain pretraining improves entity recognition.

Table 1: Hypothetical Scientific Entity Extraction Performance

Model Configuration	TP	FP	FN	Precision (%)	Recall (%)	F1 Score (%)
BiLSTM-CRF Baseline	760	160	240	82.61	76.00	79.17
General BERT Model	825	125	175	86.84	82.50	84.62
Scientific-Domain Transformer	870	100	130	89.69	87.00	88.32

Note: All counts and metrics are illustrative. No real model or scientific corpus produced these results.

5.2 Hypothetical Scientific Relation Extraction Performance

The second analysis examines six hundred hypothetical reference semantic relations. The recurrent baseline is assigned 390 correctly identified relations, compared with 440 for general BERT and 480 for the scientific-domain transformer. Relation extraction produces lower illustrative recall than entity extraction across all three configurations, reflecting the greater conceptual difficulty assigned to identifying relationships between scientific concepts. This pattern is hypothetical and should not be interpreted as a general quantitative property of relation extraction.

The scientific-domain transformer records hypothetical precision of 87.27%, recall of 80.00%, and F1 score of 83.48%. General BERT produces approximately 77.88% F1, while the baseline produces approximately 70.91%. These values illustrate the methodological advantage of reporting relation extraction separately from entity recognition. A model with strong entity identification can still make substantial errors when classifying semantic connections. Actual relation performance would depend heavily on annotation definitions and candidate-pair generation.

Table 2: Hypothetical Scientific Relation Extraction Performance

Model Configuration	TP	FP	FN	Precision (%)	Recall (%)	F1 Score (%)
BiLSTM-Based Relation Model	390	110	210	78.00	65.00	70.91
General BERT Relation Model	440	90	160	83.02	73.33	77.88
Scientific-Domain Transformer	480	70	120	87.27	80.00	83.48

Note: The relation counts and performance metrics are fictional and do not constitute empirical model results.

6. Discussion

6.1 Contextual Representation and Scientific Entity Recognition

The hypothetical entity results illustrate the methodological rationale for contextualized language representation. A fixed embedding represents a term similarly across sentences, whereas transformer representations can change according to surrounding words. This distinction can be important in scientific writing because the same expression may function as a method, material, task, or general concept depending on context. BERT demonstrated the value of deeply bidirectional contextual representation, while SciBERT extended similar modeling principles to scientific corpora.

The hypothetical advantage assigned to the scientific-domain model should nevertheless not be interpreted as evidence that domain pretraining is always preferable. General BERT can provide strong contextual representations, especially when downstream fine-tuning data are sufficiently informative. Scientific pretraining may be most valuable when the target text contains specialized terminology and syntactic patterns resembling the pretraining corpus. Controlled experiments are therefore needed to separate gains produced by architecture, vocabulary, pretraining data, and downstream annotation size.

6.2 Relation Extraction and Contextual Dependencies

Relation extraction requires models to determine whether two scientific concepts are connected and which semantic relationship is expressed. Correct entity detection does not automatically resolve this task because several candidate relationships may be plausible within a sentence. Contextualized span representations can provide information about intervening language and surrounding concepts. Luan et al. and Wadden et al. demonstrated the importance of joint or contextual approaches for scientific relation identification and broader information-extraction tasks.

Long-range relations introduce additional difficulty because connected concepts may appear in separate clauses or sentences. Standard sentence-level classification can miss relationships whose interpretation depends on coreference or broader document context. DyGIE++ addressed some of these issues through contextual span representations and graph propagation. Future transformer-based scientific extraction should therefore examine both local attention and document-level context rather than assuming that sentence-limited contextualization captures every scientifically meaningful relationship.

6.3 Domain Adaptation and Scientific Knowledge Extraction

Scientific-domain pretraining may improve representation of vocabulary that occurs rarely in general language. Specialized abbreviations, methodological terminology, chemical or biomedical expressions, and technical noun phrases can create tokenization and semantic challenges. SciBERT was specifically trained on scientific publications and demonstrated improved performance on several scientific NLP tasks. BioBERT similarly illustrated the value of continuing contextual pretraining on biomedical literature for domain-oriented text-mining applications.

However, a model pretrained on broad scientific literature may still encounter disciplinary differences. Terminology in computer science, biomedical research, materials engineering, and environmental science can follow different conventions. Domain adaptation should therefore be considered a continuum rather than a single solution. Further pretraining on a narrower corpus, task-specific supervision, or vocabulary adaptation may be required when the downstream domain differs substantially from the model's original scientific training distribution.

7. Challenges

7.1 Specialized Terminology and Annotation Ambiguity

Scientific documents contain specialized terminology, abbreviations, symbols, and multiword expressions that can make entity boundaries difficult to determine. The same phrase may also receive different labels depending on disciplinary context. Annotators may disagree regarding whether a term represents a method, task, dataset, or broader concept. Such ambiguity directly influences model evaluation because the reference annotations themselves determine which predictions are classified as correct or incorrect.

Annotation quality consequently represents a fundamental limitation of scientific information extraction. High-performing models cannot compensate for inconsistent category definitions or incomplete relation labeling. Genuine empirical research should therefore provide detailed annotation guidelines, measure inter-annotator agreement, and document adjudication procedures. Error analysis should distinguish model failures from annotation uncertainty. Without such transparency, small numerical differences between transformer models may appear meaningful even when they fall within the variability introduced by human annotation decisions.

7.2 Document Length and Long-Range Context

Scientific articles are substantially longer than the short text sequences for which many language models are commonly fine-tuned. Important relationships may connect information introduced in different sections, while standard transformer input limits can require long documents to be divided into smaller segments. Such segmentation may remove context needed to interpret abbreviations, experimental settings, or relationships between methods and findings. Document-level information extraction therefore remains more difficult than sentence-level sequence classification.

Processing longer context also increases computational requirements because standard self-attention has substantial memory and processing costs as input length grows. Researchers may use sliding windows, hierarchical representations, selective retrieval, or specialized long-context architectures to address this limitation. Each strategy can alter what information is available during extraction. Evaluation should therefore document the context window and document-segmentation procedure rather than attributing differences entirely to the underlying language model.

7.3 Generalization and Computational Requirements

Transformer models can require substantial computational resources for pretraining and repeated task-specific fine-tuning. Scientific institutions with limited hardware may find it difficult to train large domain models from the beginning. Reusing pretrained models reduces this requirement but still introduces memory and inference costs. A more complex model should therefore demonstrate meaningful extraction benefits relative to computationally cheaper baselines before additional resource requirements are considered justified.

Generalization represents another important challenge. A scientific-domain transformer trained and evaluated on similar disciplines may perform less effectively when applied to unfamiliar research fields. Vocabulary differences, writing conventions, relation distributions, and annotation categories can change substantially across domains. Reliable evaluation should include cross-domain or temporally separated datasets wherever possible. This provides stronger evidence of transferability than random splitting of documents drawn from one relatively homogeneous corpus.

8. Conclusion

Transformer-based language models provide a methodologically important approach to contextual information extraction from scientific documents. Their principal advantage lies in constructing token and span representations that incorporate surrounding linguistic information rather than assigning fixed meanings independently of context. The present hypothetical study examined this capability through entity and relation extraction and compared a conventional recurrent baseline with general and scientific-domain transformer configurations.

The illustrative analysis assigned the highest precision, recall, and F1 scores to the scientific-domain transformer for both entity and relation extraction. These numerical values are entirely hypothetical and do not establish empirical superiority. Their purpose is to demonstrate a transparent evaluation structure in which scientific entities and semantic relations are assessed separately. Genuine performance comparisons would require implemented systems, real annotated corpora, repeated fine-tuning experiments, confidence estimates, and detailed error analysis.

The study concludes that transformer models are particularly relevant to scientific information extraction when contextual terminology and specialized language must be interpreted accurately. Nevertheless, domain-specific pretraining does not eliminate challenges involving ambiguous annotations, long scientific documents, cross-domain variation, or computational cost. Future empirical research should evaluate domain transfer, long-context processing, relation reasoning, and annotation quality to determine the conditions under which contextual transformer models provide reliable improvements in scientific knowledge extraction.

References

1. Beltagy, I., Lo, K., & 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 (pp. 3615–3620). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1371>
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

- Technologies (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. <https://arxiv.org/abs/1706.03762>
 4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. <https://arxiv.org/abs/1907.11692>
 5. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
 6. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), Article 2. <https://doi.org/10.1145/3458754>
 7. Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3219–3232). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1360>
 8. Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7506–7516). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.670>
 9. Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 5784–5789). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1585>
 10. Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., & Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics. <https://aclanthology.org/2020.nlp-covid19-acl.1/>
 11. Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4969–4983). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.447>
 12. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.207>
 13. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv. <https://arxiv.org/abs/2004.05150>

14. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17283–17297). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html>
15. Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 8003–8016). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.551>
16. Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 8086–8098). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.556>
17. Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv. <https://arxiv.org/abs/1904.05342>
18. Gu, Y., Wang, X., Yu, J., et al. (2023). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 24(2), Article bbad077. <https://doi.org/10.1093/bib/bbad077>
19. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv. <https://arxiv.org/abs/2108.07258>
20. Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). AMMU: A survey of transformer-based biomedical pretrained language models. Journal of Biomedical Informatics, 122, Article 103867. <https://doi.org/10.1016/j.jbi.2021.103867>