Computational Molecular Biology 2025, Vol.15, No.4, 160-170
http://bioscipublisher.com/index.php/cmb

Review Article Open Access

Large Language Models for Biological Knowledge Extraction

Hongpeng Wang, Minghua Li
Biotechnology Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, China
Corresponding author: minghua.li@cuixi.org

Computational Molecular Biology, 2025, Vol.15, No.4 doi: 10.5376/cmb.2025.15.0016
Received: 01 May, 2025
Accepted: 10 Jun., 2025
Published: 01 Jul., 2025

Preferred citation for this article:
Wang H.P., and Li M.H., 2025, Large language models for biological knowledge extraction, Computational Molecular Biology, 15(4): 160-170 (doi: 10.5376/cmb.2025.15.0016)

Abstract The surge in biomedical literature has led to severe information overload for researchers, necessitating automated knowledge extraction tools. Large Language Models (LLMs), which have emerged in recent years, demonstrate superior performance in text understanding and generation, providing a new approach for biological knowledge extraction. This study reviews the applications of LLMs in tasks such as named entity recognition, relation extraction, and event extraction, and discusses their latest advancements in subfields such as genomics, proteomics, and pharmacology. The advantages of LLMs over traditional methods in contextual understanding and semantic representation are analyzed, along with the optimization effects of domain adaptation, fine-tuning, and cue engineering on model performance. A case study of extracting gene-disease associations using the BioGPT model demonstrates the application process and effectiveness of LLMs, while also analyzing challenges related to data quality, model illusion, and privacy protection. The future directions of LLM integration with knowledge graphs, multimodal data integration, and knowledge verification are discussed, along with related ethical considerations. These advancements are expected to provide new paradigms for future biomedical research.

Keywords Large language model; Biomedical text mining; Knowledge extraction; Knowledge graph; Cue engineering; engineering

1 Introduction

The growth rate of knowledge in the biomedical field is almost suffocating. In 2016 alone, more than 860,000 new papers were added to PubMed, with an average of one new paper emerging every minute. Researchers are often submerged in this flood of information, and it becomes increasingly difficult to sort out and absorb key information in a timely manner. This is precisely the so-called "information overload" problem (Brown et al., 2020). As a result, people began to pay more attention to information extraction technology, hoping to use it to automatically identify useful biological entities, relationships and events from vast amounts of unstructured text, and transform disordered words into structured knowledge that machines can understand. The problem is that, in the face of such a huge scale of data, traditional text mining methods seem somewhat struggling. Early practices mainly relied on experts to formulate rules, depend on ontology libraries, or combine artificial features to train machine learning models. However, in the context of biomedical literature where terminology is complex and sentence structures are lengthy, these methods often have limited effects. It was not until the emergence of deep learning, especially the rise of large language models, that hope was reignited in this field (Topol, 2019).

The so-called large language model is actually a giant neural network trained on massive corpora, which "remembers" a large amount of language knowledge with a huge number of parameters (Wiggins and Tejani, 2021). Since Vaswani et al. (2017) proposed the Transformer architecture of "attention mechanism is everything", the capabilities of language models have almost multiplied with scale. BERT is one of the milestones, which achieves deep understanding of context through bidirectional Transformer. Then, models such as GPT-3 have even broken through the task boundary and can complete the extraction class task with only a few examples. The knowledge and language rules learned in the general corpus make LLM a "basic model" that can be easily adapted to various biological information extraction scenarios through fine-tuning.

Compared with those models in the past that relied on artificial rules or feature engineering, the "smartness" of large language models lies in their ability to understand the context more naturally, generate coherent text, and

handle long sentences and complex dependencies commonly seen in the biomedical field (Ouyang et al., 2022). The self-attention mechanism of Transformer enables the model to directly "focus" on the relevant parts in the sentence when decoding an entity or relation, solving the problem that previous models had difficulty capturing long-distance semantics. It is precisely for this reason that LLM is regarded as a powerful tool to break through the current bottleneck of biological knowledge extraction. This article will focus on its application progress in this field, discussing its advantages, limitations and future development directions.

2 The Research Background of Biological Knowledge Extraction

2.1 Definition and key tasks of knowledge extraction

In essence, the extraction of biological knowledge is about automatically identifying those "knowledge points" from disorganized information such as literature and databases that machines can understand and connect. This knowledge is often hidden in unstructured text, such as gene names, protein names, disease or compound names, etc.. To sort out this information clearly, several steps are usually involved: first, identify the key nouns (named entity recognition), then determine the relationships between them (relation extraction), and finally capture more complex biological events (event extraction). It sounds natural and logical, but it is not easy to do. Take NER as an example. The same protein may have different names, and some may even have symbols or numbers. The rule system can easily mistake or miss them. The task of relation extraction goes further. For example, to identify the causal relationship between genes and diseases from sentences such as "*BRCA1* gene mutation increases the risk of breast cancer". Such ability is particularly important for building biomedical knowledge graphs because it can connect scattered entities into a network. However, event extraction is the most complex - it not only needs to know "who is related to whom", but also understand "what happened". For example, in the sentence "TP53 mutation leads to cell cycle arrest", "mutation" is the trigger word, while "TP53" and "cell cycle arrest" are the thesis elements. Such tasks often require the model to have a deeper understanding of the context, where traditional methods seem inadequate.

2.2 Traditional biological text mining methods and their limitations

Before the popularity of large language models, biological text mining mostly relied on rules and machine learning. The early approach was more like "writing programs to teach machines to recognize words": first list the dictionary, then write the matching rules, and manually design a whole set of templates to recognize terms (Beltagy et al., 2019). This method is acceptable in fields where the terminology is relatively stable, but it is prone to collapse once encountering new concepts and new names (Wang et al., 2018). The problem of synonyms in biomedicine is particularly troublesome. A drug may have multiple alternative names, and it is difficult to capture all of them by string matching. Later, with the rise of statistical machine learning, support vector machines, hidden Markov models, etc. began to be used, and the degree of automation improved somewhat. However, various features still needed to be designed manually, such as context Windows, part-of-speech tags, syntactic paths, etc. Once these features change domains, the model performance often drops. What's more troublesome is that these models cannot effectively utilize unlabeled large corpora and are not flexible enough in capturing complex sentences and long-distance relationships.

2.3 Technological evolution from statistical learning to deep learning

The turning point of technology emerged with the rise of deep learning. Around 2010, statistical models were still dominant, but neural networks soon took over. RNN and CNN have been successively introduced into relation extraction and entity recognition tasks. The models have begun to be able to automatically learn features and no longer rely entirely on human intervention. CNN performs well in drug-protein relationship extraction, and the architecture of bidirectional LSTM combined with CRF also outperforms the traditional CRF model in gene and chemical entity recognition. The year 2018 can be said to be a crucial one. ULMFiT proposed the idea of pre-training and fine-tuning, proving that it is more effective to learn languages from large corpora first and then perform specific tasks (Howard and Ruder, 2018). Immediately afterwards, the emergence of BERT completely changed the rules of the game. It implemented deep modeling of bidirectional contexts using the Transformer architecture. This enables the model to "understand" the context rather than merely memorize the key words.

When the era of large language models truly arrives, the speed of technological evolution can almost be described as a "leap". The model parameters range from tens of millions to hundreds of billions, and the capacity has skyrocketed. Moreover, researchers have found that continuing to pre-train on biomedical corpora can enable the model to better understand the language of professional domains. This is known as domain adaptive pre-training (Gururangan et al., 2020). PubMedBERT is a typical example. It was trained from scratch entirely with PubMed texts and outperformed BioBERT and General BERT in multiple tasks, indicating that the value of domain corpora is much greater than imagined.

3 The Current Application Status of the Three Major Language Models in the Field of Biology

3.1 Mainstream large language models and their performance in biological texts

In the current field of biomedical information extraction, the stage is almost dominated by various large language models. Some researchers directly use general models, such as the GPT series. Some have simply developed versions specifically for biomedicine, such as BioBERT, PubMedBERT, and BioGPT. As soon as these models appeared, the traditional methods immediately seemed a bit "clumsy". The GPT series (such as GPT-2 and GPT-3) have been trained very thoroughly on general text, and their language understanding and generation capabilities are quite strong. Some people have attempted to apply them in biomedical question answering or literature abstract tasks, and can produce quite reasonable results even without specific task training. However, problems arise as well - their understanding of professional terms is often inadequate and their explanations are not precise enough. It is precisely for this reason that the academic community has begun to shift towards "domain-specialized" large models. For instance, BioBERT is based on BERT and further trained with a large amount of biological literature from PubMed and PMC, enabling the model to truly "understand" professional language. Experiments show that BioBERT performs approximately 1 to 3 percentage points higher than ordinary BERT in tasks such as biomedical NER and relation extraction (Lee et al., 2020). PubMedBERT goes a step further by training directly from scratch on the domain corpus, verifying the importance of the proprietary corpus. Microsoft's BioGPT, on the other hand, follows a generative approach and is trained on biomedical texts using the GPT architecture. It can not only generate knowledge descriptions, but also automatically continue the content of molecular interactions. In most of the six biological NLP tasks, it outperforms previous discriminative models. SciBERT is even more "cross-disciplinary". It has been trained on a large number of scientific publications, including multi-disciplinary texts such as those in biomedicine. The results prove that training with corpora and vocabularies in the field of science can indeed improve the performance of the model in academic texts, especially in long-length and terminologically intensive tasks.

3.2 Applications of the model in different fields of biology

If we turn our attention to the different branches of biology, LLMS are almost everywhere. In genomics, it is used to extract the associations between genes and diseases, as well as genes and phenotypes, from a vast number of papers, helping researchers screen candidate genes. The research of Huang et al. (2024) also pointed out that on datasets such as ChemDisGene, the performance of Transformer-like models has approached the optimal level. In the field of proteomics, models are used in a different way - they can more accurately identify the interaction relationships between proteins. Models like SciBERT have surpassed the old rule system on BioCreative's PPI data. More complex tasks, such as identifying protein modifications or binding events, have also begun to rely on the long-range dependency modeling capabilities of LLMS to handle cross-sentence semantic associations. As for pharmacology and clinical medicine, the applications are simply too numerous to count. Adverse drug reactions, interactions, etc. previously required manual literature screening. Now, the model can automatically extract and update the drug knowledge base. BioGPT even achieved an F1 score of 83% in drug-drug interaction extraction, which is almost comparable to the expert-labeled results. The general model has not been idle either - ChatGPT achieved a near-pass score in the USMLE, while GPT-4 has a higher accuracy rate. Google's Med-PaLM was also used for medical Q&A, and it was said that the performance was not much different from that of clinical experts (Singhal et al., 2023). These results suggest that LLMS are becoming increasingly "doctor-like" in terms of medical knowledge expression and reasoning (Shah et al., 2023). However, different fields have different requirements for models. Genomics places more emphasis on the description of biological mechanisms, while

pharmacological texts focus on details such as dosage and drug names. Therefore, specialized models like CancerBERT emerged, focusing on oncology tasks, and their performance is indeed better than that of the general BioBERT. In the future, more models targeting specific subdomains may emerge, and the trend of segmentation has become a foretold conclusion.

3.3 LLM-driven knowledge graph construction and biological knowledge base update

If the extraction of a single piece of literature is the "point", then the construction of a knowledge graph is the "surface". The capabilities of LLMS are automating the generation and update of large-scale knowledge graphs. The biomedical knowledge graph is essentially a huge network, with nodes being entities such as genes, diseases, and drugs, and edges representing the relationships among them. In the past, literature had to be organized manually, but now, most of the work can be accomplished by batch extraction of models. For instance, the gene-disease relationship network constructed by Percha and Altman (2018) integrates tens of thousands of pieces of association information. The PubMed Knowledge Graph of Xu et al. (2020) is even larger. It extracts entities and relationships from nearly 30 million abstracts and obtains a graph with millions of nodes and hundreds of millions of edges. All of these can be regarded as early achievements of domain models. Nowadays, researchers are further exploring the integration of LLMS and knowledge graphs, enabling "graphs" to assist "text" and also allowing "text" to complement "graphs". RAG (Retrieval Enhanced Generation) is a representative of this idea. It retrieves the knowledge base simultaneously when generating answers to improve the accuracy of the content. In biomedical scenarios, this mechanism can help models update knowledge in real time. For instance, when new literature reports that a certain mutation is related to a disease, the system can automatically extract and complete the knowledge graph. In addition to expanding knowledge, LLM can also assist in semantic normalization, merge and simplify repetitive or conflicting relationships, and even generate explanatory text to assist in verifying the relationships in the knowledge graph and enhance the readability and credibility of the knowledge base.

4 Knowledge Extraction Mechanisms and Optimization Strategies of the Four Major Language Models

4.1 Advantages of context understanding and semantic representation

A major reason why large language models perform well in knowledge extraction is that they "see further". Traditional models are often limited by Windows and can only understand local information, while LLM can capture context relationships on a larger scale by relying on the self-attention mechanism of Transformer. This enables it to break away from the local binding of words when analyzing sentences. For instance, in the statement "Inhibiting the expression of the *EGFR* gene can reduce the survival rate of lung cancer cells", it is not misled by the surface word order but can match "reducing the survival rate" with "inhibiting the expression of EGFR" to draw the correct causal relationship. Compared with the past models that only relied on matching neighboring words, this global modeling ability is obviously more reliable. What's more interesting is that LLMS that have undergone pre-training have actually "internalized" a considerable amount of biological semantics. In models like BioBERT, the activation of certain neurons can even correspond to protein-protein interaction patterns. This enables them to make reasonable inferences through the semantic space when encountering unfamiliar terms in the extraction task. Sentences in biomedical literature are often lengthy and complex in hierarchy, with attributives and interjections interwoven. Traditional methods often get "lost" in this kind of sentence structure, while LLMS can bypass interfering words with multi-head attention and directly grasp the core relationship. For instance, in the sentence "We observed that the incidence of breast tumors significantly increased in mouse models carrying *BRCA1* mutations", the model can automatically connect "*BRCA1* mutations" with "tumor incidence", and the correct causal clues can be found without analyzing word by word. This ability to capture complex semantics is the true strength of LLMS.

4.2 Domain adaptation and fine-tuning techniques

However, no matter how powerful a general model is, when placed in a professional field, it is impossible to become immediately proficient. To make it "speak in a professional way", domain adaptation and fine-tuning are necessary. Domain adaptation means continuing to feed the model with biomedical corpora to familiarize it with professional vocabulary and language habits. Many general LLMS 'vocabularies do not contain complex genetic

symbols or compound names, which can easily lead to word splitting errors. Training the segmenter directly with biological corpora like PubMedBERT can significantly reduce this problem (Gu et al., 2021). Meanwhile, continuing to conduct pre-training on biological texts can also make the model more "close to" the domain expression. Fine-tuning, on the other hand, enables the model to "practice fine skills" for specific tasks. For instance, by adjusting parameters on only a few thousand labeled samples, an F1 score of nearly 80% can be achieved in the task of extracting chemical protein relationships. For traditional models to achieve the same effect, they often require multiple times the amount of data. There are still some lightweight fine-tuning strategies now, such as freezing the underlying Transformer layer, training only the upper-level classifier, or using methods like LoRA to reduce computing power consumption. It can be said that domain adaptation makes the model "understand the trade", while fine-tuning enables it to "do the job". The combination of the two is the key to truly unleashing the potential of LLMS. They work well together. The model can not only extract known relations but also flexibly generalize to new types of text.

4.3 Application of prompt engineering and instruction fine-tuning in knowledge extraction

Looking further ahead, the plasticity of large language models is also reflected in "prompts". Just tell it what to do and it will do it. This is the magic of prompt engineering. For example, if we directly ask ChatGPT to "extract the relationship between genes and diseases from the following text", it can output a structured result without even fine-tuning (Kung et al., 2023). However, how the prompts are written has a significant impact. Even a slight difference in wording can lead to vastly different effects. Providing examples and adjusting the format often enable the model to understand the task intent more accurately. For instance, in relation extraction, a question pattern can be used to prompt: "Entity 1 [GENE], Entity 2 [DISEASE]: Does Entity 1 cause Entity 2?" The model can answer with "Yes/No" or specific relationship types, and this format is often more natural than traditional classification. In addition to manually designed prompts, fine-tuning of instructions further enhances the model's "obedience" ability. Models such as InstructGPT are trained with large-scale instruction data to better understand users' intentions. In the knowledge extraction task, this means that as long as the input is "list the diseases mentioned in the text and their related genes", the model can directly generate results that meet the requirements. However, LLMS are not immune to mistakes either. Sometimes it "makes up confidently", which is what is called an illusion. To this end, researchers introduced verification mechanisms, such as cross-checking the relationships extracted by the model with the database, or having it provide the original text basis simultaneously. Although these practices complicate the process a bit, they can make the extraction results more reliable and also give humans more confidence when trusting the model.

5 Case Study: Gene-Disease Relationship Extraction Based on BioGPT

5.1 Case background and data sources (PubMed abstract and gene database)

In biomedical research, the relationship between genes and diseases is almost everywhere. The pathogenesis of many diseases can be traced back to mutations or abnormal expressions of certain genes, which is also one of the most core knowledge in genetics and precision medicine. To more intuitively demonstrate the role of large language models in knowledge extraction, we have selected a typical task - "gene-disease relationship extraction" - as an example for illustration. Here we use BioGPT. This model is trained based on large-scale biomedical literature and has strong professional semantic understanding and generation capabilities. We want it to automatically identify triples like "Gene A is associated with disease B" from unlabeled literature abstracts, and then compare these results with the existing associations in authoritative databases to see if it truly "understands biology".

The selection of the data section actually also requires careful consideration. We use the abstracts of biomedical literature from PubMed, which often contain rich information such as gene functions and disease mechanisms, and are very suitable for input in relation extraction. As a reference standard, we selected the DisGeNET database (Pinero et al., 2020). This database integrates human gene-disease association data from multiple sources and contains tens of thousands of entries. It is currently recognized as an authoritative knowledge base (Pinero et al., 2020). We selected a batch of highly reliable associations from them, mapped them to the corresponding PubMed

literature, and after manual review, sorted out a labeled training set. What BioGPT aims to do is to be able to identify similar relationships in new literature after learning these examples.

However, this task is not merely about "matching words". The expression styles of biomedical papers are diverse, and the same meaning often has different ways of being written. For instance, "*BRCA1* mutations can lead to an increase in breast cancer susceptibility" and "*BRCA1* is a susceptibility gene for breast cancer" refer to the same thing, but the sentence structures are completely different. To enable the model to recognize such semantically equivalent expressions, we deliberately retained a variety of sentence patterns when constructing the training set. This design can help BioGPT better understand the context, thereby improving the recall rate, and also make it less likely to be confused by the surface wording when dealing with literature of different writing styles.

5.2 Model design and training process

For this part, we adopted a relatively common approach - "pre-trained model + downstream fine-tuning". In simple terms, it involves first using a model that has already learned to "understand biological language" as a foundation, and then training it for specific tasks. BioGPT is just right. It has been pre-trained on a large number of biomedical literatures and has accumulated sufficient language patterns and domain knowledge (Luo et al., 2022), and has a certain understanding of the concepts of genes and diseases. Next, what we need to do is to make it more "focused", and through supervised fine-tuning, enable the model to learn to extract gene-disease relationships from the literature.

However, we did not let it perform the standard task of classification, but transformed the problem into a generation task. The specific operation is as follows: Feed the literature abstract to the model and add a prompt at the beginning. Please extract the genes mentioned in the text and their related diseases:" In this way, the model knows that it needs to "list" the results instead of continuing to write the article. This design enables BioGPT to directly generate the output format we need and is more in line with its characteristics as a generative language model.

The training data comes from the DisGeNET annotation set mentioned earlier. Each sample includes a summary and the corresponding gene-disease pairing. We let the model learn to "repeat" these paired contents and complete the training by minimizing the difference between its output results and the standard answers. Since BioGPT itself is a generative model, we have retained its original generation mechanism, enabling it to output multiple relationships in a single generation without the need to split them into multiple independent samples.

We made some conservative adjustments to the training parameters. The parameter scale of BioGPT is very large. A learning rate that is too high can easily cause it to forget the original knowledge. Therefore, we used a relatively small learning rate, approximately 2e-5, and controlled the training rounds within three rounds (Devlin et al., 2019). To save computational effort, we also froze the underlying parameters of the model and only fine-tuned the high-level Transformer layer and the output part. Doing so can maintain the stability of the model language generation and also avoid overfitting.

Finally, we use the untrained samples from DisGeNET to evaluate the model's performance. The evaluation criteria remain Precision, Recall and F1 value. To determine whether the extraction is correct, we require that the gene names and disease names generated by the model must be consistent with the standard answers, and the relationship types should also match semantically. Of course, we allow for some differences in expression. For instance, synonyms like "breast cancer" and "breast cancer" are regarded as equivalent. After all, in the biomedical context, such differences do not affect the accuracy of knowledge points.

5.3 Experimental results, performance comparison and analysis of result interpretability

The trained BioGPT performed well in the extraction of gene-disease relationships. The validation set results show that the Precision is approximately 0.80, the Recall is approximately 0.78, and the F1 value is close to 0.79. This level is already similar to the current best supervised model. If the unfine-tuned BioGPT is directly used for extraction, the Precision is only around 0.60. It is evident that fine-tuning for the task is still indispensable. We

also compared it with the traditional scheme of BioBERT plus classifier, whose F1 value is approximately 0.75, which is about 4 percentage points lower than that of our generative model. It seems that the end-to-end generation method of BioGPT can indeed understand the text semantics more comprehensively, especially performing better when dealing with implicit relations.

Of course, the model is not perfect either. After a careful analysis of the errors, it was found that most of the problems lay in the recall rate - it missed some relationships that should have been extracted. These missed detections can be roughly divided into two situations: One is that the relationship is written too implicitly and can only be identified by combining background knowledge; Another category is because extremely obscure terms or rare gene names are used in the text, and the model is not very familiar with them. For instance, an abstract mentioned a rare disease and an uncommon gene, but the model simply did not extract them. In fact, even human readers have to look up literature to confirm such sentences. So from this perspective, the model's performance can be considered stable.

In terms of interpretability, we conducted a small experiment, asking the model to simultaneously indicate the "reason" or the basis sentence when outputting each relationship. The approach is to adjust the prompt, requiring it to list the gene-disease pairs while marking the original sentence from the summary. BioGPT can identify key sentences in most cases and output them along with the extraction results. For instance, when it extracts "BRCA1-breast cancer", it will attach "*BRCA1* gene mutations are significantly enriched in patients, suggesting a correlation with breast cancer susceptibility" as evidence, which is consistent with manual judgment. This ability makes the results more trustworthy for researchers because they can see the evidence directly.

However, there are exceptions. Sometimes, the model will quote entire sentences or even entire paragraphs. As long as it contains a little relevant information, it will all be uploaded, which seems a bit redundant. This indicates that while pursuing interpretability, some post-processing steps are also needed to make the long sentences provided by the model more compact. We also noticed another type of error: the model occasionally misjudges "co-occurrence" relationships as causal ones, such as only mentioning genes and diseases simultaneously, but not stating that the former leads to the latter. This type of error requires particular caution in knowledge graphs. In the future, it can be considered to add causal reasoning or discrimination modules to help the model be more "cautious" during extraction.

6 Challenges and Limitations

6.1 Data quality and standardization of biological terminology issues

Although large language models have demonstrated considerable potential in extracting biological knowledge, problems follow one after another when they are actually implemented. The first problem often encountered is not the "intelligence quotient" of the model itself, but rather the confusion of data and terminology. Biomedical texts are filled with various abbreviations, aliases and non-uniform names. Different researchers and even different databases have different names for the same gene and the same disease. For instance, A gene is represented by its full name in Paper A, abbreviated in Paper B, and then assigned a different code in the database. This is especially true for diseases, where sometimes scientific names and common names are used interchangeably. The result is that when the model is training or reasoning, it will recognize the same entity as several different objects, and the extraction results will be divided into multiple parts, or even contradictory to each other.

What's more troublesome is that the quality of data annotation is often not satisfactory either. Manual annotation requires biomedical experts, and such labor is both scarce and expensive (Esteva et al., 2019). However, due to insufficient labeling and the model's inability to acquire high-quality knowledge, it can only rely on a small number of samples for training, which easily leads to overfitting to specific expressions. Once encountering rare relationships or novel descriptions, the generalization ability of the model appears insufficient. Moreover, if the training data itself is biased or labeled incorrectly, the model will accept all these errors and even magnify them during generation.

For the issue of inconsistent terminology, researchers usually make some "fixes" before and after the model. One approach is to perform vocabulary mapping during the input stage, such as uniformly replacing common aliases with standard names. Another approach is to perform comparative correction using the database at the output stage (Pinero et al., 2020). Some people have also attempted to directly integrate medical vocabularies (such as UMLS) into the model reasoning process, using them to constrain the naming of the generated entities. However, at present, LLMS mainly generate text based on statistical correlations and do not automatically follow medical naming conventions. This means that even if the model generation results are smooth, additional post-processing steps are still needed to "clean" them; otherwise, the constructed knowledge base may still be chaotic.

As for the issue of data quality, several compromise approaches are currently being explored. Relying solely on manual annotation is too slow. Therefore, some studies adopt "silver standard" data, that is, allowing the model to generate annotations by itself or automatically using rules (Habibi et al., 2017). Although it is not as precise as manual labor, it can make up for the shortage in terms of quantity. Then fine-tuning with a small number of "gold standard" samples can compensate for the deviation to a certain extent. There is also the idea of active learning, which enables the model to help pick out samples with large amounts of information for priority labeling, achieving higher results with less labor costs. In other words, the problem of data cannot be solved in the short term, but by using data smartly, perhaps the model can learn more intelligently from "dirty data".

6.2 Model illusion and challenge of result verifiability

The "illusion" problem of large language models is almost an insurmountable hurdle. It sometimes solemnly generates non-existent content that looks decent but is actually groundless. This situation is particularly evident in knowledge extraction tasks - the model may "fabricate" a pair of associations between genes and diseases based on impression, even if it is not mentioned at all in the original text. The reason for this is that LLMS study the statistical laws of language rather than the facts themselves. It will guess based on semantic "inertia", and even in the absence of evidence, it may force out a seemingly reasonable answer.

This is no small problem in the field of biomedicine. Both scientific research and clinical practice emphasize "evidence". Once a model outputs incorrect relationships, the consequences could be very serious. For instance, if it fabricates the interaction between a certain drug and a target, researchers might waste experimental resources in vain. Or, if a gene is wrongly associated with a disease and the doctor interprets it this way, it may lead to misjudgment (Shah et al., 2023). In other words, the "confidence error" of a model is more dangerous in scientific research than silence.

To deal with this situation, a common approach is to add an extra layer of "verification". After the model extracts the relationships, it does not rush to store them in the database. Instead, it checks the original text through the search or discrimination module to see if there are any relevant sentences to support it. If not found, it is marked as low credibility and not included in the final result (Nori et al., 2023). The RAG model proposed by Lewis et al. (2020) follows this idea. It retrives relevant literature simultaneously during generation, making the output more "well-grounded". In our experiment, we also tried to have BioGPT attach the original sentence as a basis when generating relationships. Although not always accurate, it does enhance the verifiability of the results.

Another approach is to have the model learn to "admit not knowing" during the training phase. For example, some samples are added during instruction fine-tuning to enable the model to answer "not mentioned" or "unable to judge" in the absence of evidence. This approach is somewhat like adding a "refusal mechanism" to the model, enabling it to learn to shut up when it is uncertain. Although this cannot completely eliminate the illusion, it can reduce its tendency to fabricate facts at will.

In addition, researchers are also attempting to evaluate the reliability of the model output in a more systematic way. For instance, use the existing knowledge graph to conduct consistency checks on the generated results to see if there are any conflicts with the known facts. If the model claims that "mitochondrial DNA causes certain skin diseases", and biological common sense explicitly denies this causality, then this output can be identified as an illusion and eliminated. Some people even let the model "reflect on itself" and evaluate the credibility of their

answers in the form of dialogue (Ji et al., 2023). Although these methods are not yet perfect, at least they indicate that AI is no longer merely regarded as an "output machine", but is gradually being taught how to "verify itself".

6.3 Privacy protection and intellectual property issues

The application of large language models in the biomedical field is not merely a technological breakthrough; it has also brought about new concerns regarding privacy and copyright. Models often have to deal with a large amount of medical text during training, among which there are many sensitive contents, such as patient cases or electronic health records (EHR). Although most of our current research is based on publicly available scientific literature, the situation will be completely different once the model is applied to the clinical field. LLMS sometimes "remember" training data. If they inadvertently disclose information such as patients' names and case numbers during generation, it would be a serious privacy violation. Such risks blur the ethical boundaries of models and force people to re-examine the bottom line of data usage.

To prevent privacy leakage, the most direct approach is to clean up data from the source. Desensitization processing must be carried out before training to remove or replace the information that can identify the individual's identity (Pinero et al., 2020). Some studies are still attempting to enhance protection through technical means, such as incorporating differential privacy mechanisms during model training, enabling the model to learn overall patterns rather than specific individuals. Even if this results in a slight loss of accuracy, it is worth it. For knowledge extraction tasks, focusing on statistical information or public knowledge at the group level is a more reliable approach (Mesko and Topol, 2023). For instance, only extract collective patterns such as gene-disease relationships and avoid involving individual case descriptions. If the model is really used in sensitive data scenarios such as clinical notes in the future, these protective measures are likely not optional but "mandatory".

Apart from privacy, copyright issues are equally thorny. The training data of large language models often contains protected texts such as full papers and patent descriptions, which turns "whether the model output constitutes infringement" into a gray area. If the sentences output by the model during the generation process are highly similar to the training text, even if they merely state facts, they may be questioned as improper quotations. In our experiment, BioGPT was required to output the basis sentences in the literature. Strictly speaking, this approach might be on the verge of copyright infringement, but in academic research, it is usually considered fair use.

To reduce disputes, models can be taught to "paraphray" rather than "copy". That is to say, let it summarize the facts in its own words instead of copying the original text word for word (Moor et al., 2023). Meanwhile, the source should be clearly indicated in the results, which is both a respect for the original author and convenient for others to verify. This approach not only complies with legal requirements but also conforms to the principle of transparency in scientific research.

However, fundamentally speaking, the cleanest solution still lies at the data level. Nowadays, more and more open-licensed medical datasets and knowledge graphs have emerged, and researchers can fully train models based on these open-source resources. The LLM obtained in this way has no worries about copyright and is both legal and safe. If the training of future domain models can all be based on these public data, the predicament of privacy and copyright might be alleviated from the source.

7 Future Outlook

The potential of future large language models in extracting biological knowledge is far from being fully unleashed. The real challenge is not merely to "make the model smarter", but to "make the model more trustworthy". The biggest problem at present still lies in interpretability and verifiability - the model can output results, but it is difficult to explain "why". So the focus of the next research is likely to be on making the model "clearly state what it is doing". Only in this way can researchers truly trust it.

In fact, large language models and knowledge graphs are essentially a combination of two ways of thinking. One is good at capturing patterns from jumbled text, while the other excels at organizing knowledge with logic and structure. If the two can be truly integrated, an effect of "1+1>2" may occur. At that time, we might witness a new

framework for knowledge representation - one that can both understand the ambiguity of semantics and maintain the rigor of structured reasoning.

Think more boldly. In the future, models may no longer merely "answer questions", but be able to directly communicate with knowledge bases. When researchers input a natural language question, the model can not only retrieve the answer from the huge database, but also automatically supplement it to the knowledge graph after obtaining new discoveries. Moreover, this update process is self-checking. The model will determine by itself whether the new information is consistent with the old knowledge. This human-machine collaborative way of knowledge update can keep scientific databases "fresh in real time", especially in a rapidly changing field like biology, which is of great significance.

In the long run, the changes brought about by large language models may not only be at the technical level, but also a transformation in the way research is conducted. The speed of knowledge accumulation will accelerate. Researchers will no longer spend a lot of time screening literature but rather think more about how to utilize existing knowledge to innovate and verify. AI assistants might become a standard feature in laboratories. They won't replace scientists, but they will change the pace of their work.

However, the more powerful a tool is, the more it needs to be used correctly. Knowledge is a kind of power, and if power lacks restraint, it may be abused. In the future, when building an AI knowledge base, the scientific research community needs to always be vigilant about ethical risks and maintain openness and transparency. We are standing at the threshold of an era brimming with imagination - from laboratory notes to computer screens, the ways knowledge is acquired and disseminated are being rewritten. As long as the scientific spirit and ethical bottom line remain firm, this scientific research revolution jointly written by humans and large language models will eventually bring about a new chapter in the field of biology.

Acknowledgments

The authors extend sincere thanks to two anonymous peer reviewers for their invaluable feedback on the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

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 (EMNLP-IJCNLP), pp.3615-3620.
https://doi.org/10.18653/v1/D19-1371

Wiggins W., and Tejani A., 2021, On the opportunities and risks of foundation models, Radiology: Artificial Intelligence, 2022, 4(4): e220119.
https://doi.org/10.1148/ryai.220119

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., and Amodei D., 2020, Language models are few-shot learners, Advances in Neural Information Processing Systems, 33: 1877-1901.

Devlin J., Chang M.W., Lee K., and 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, volume 1, pp.4171-4186.
https://doi.org/10.18653/v1/N19-1423

Esteva A., Robicquet A., Ramsundar B., Kuleshov V., DePristo M., Chou K., Cui C., Corrado G., Thrun S., and Dean J., 2019, A guide to deep learning in healthcare, Nature Medicine, 25(1): 24-29.
https://doi.org/10.1038/s41591-018-0316-z

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, 3(1): 1-23.
https://doi.org/10.1145/3458754

Gururangan S., Marasović A., Swayamdipta S., Lo K., Beltagy I., Dauphin Y.N., and Smith N.A., 2020, Don’t stop pretraining: adapt language models to domains and tasks, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.8342-8360.
https://doi.org/10.18653/v1/2020.acl-main.740

Habibi M., Weber L., Neves M., Wiegandt D.L., and Leser U., 2017, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics, 33(14): i37-i48.
https://doi.org/10.1093/bioinformatics/btx228

Howard J., and Ruder S., 2018, Universal language model fine-tuning for text classification, In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.328-339.
https://doi.org/10.18653/v1/P18-1031

Huang M.S., Han J.C., Lin P.Y., You Y.T., Tsai R., and Hsu W.L., 2024, Surveying biomedical relation extraction: a critical examination of current datasets and a new resource, Briefings in Bioinformatics, 25(3): bbad132.
https://doi.org/10.1186/s12859-024-05749-y

Ji Z., Lee N., Fries J.A., Yu T., and Finn C., 2023, Hallucination in natural language generation, ACM Computing Surveys, 55(12): 248.
https://doi.org/10.1145/3571730

Kung T.H., Cheatham M., Medenilla A., Sillos C., De Leon L., Elepaño C., Madriaga M., Aggabao R., Diaz-Candido G., Maningo J., and Tseng, V., 2023, Performance of ChatGPT on USMLE: potential for AI-assisted medical education, PLOS Digital Health, 2(2): e0000198.
https://doi.org/10.1371/journal.pdig.0000198

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.
https://doi.org/10.1093/bioinformatics/btz682

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., and Kiela D., 2020, Retrieval-augmented generation for knowledge-intensive NLP, Advances in Neural Information Processing Systems, 33: 9459-9474.

Luo R., Sun L., Xia Y., Qin T., Zhang S., Poon H., and Liu T.Y., 2022, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics, 23(6): bbac409.
https://doi.org/10.1093/bib/bbac409

Meskó B., and Topol E.J., 2023, The imperative for regulatory oversight of large language models in healthcare, NPJ Digital Medicine, 6(1): 120.
https://doi.org/10.1038/s41746-023-00873-0

Moor M., Banerjee O., Abad Z.S.H., Krumholz H.M., Leskovec J., Topol E.J., and Rajpurkar P., 2023, Foundation models for generalist medical artificial intelligence, Nature, 616(7956): 259-265.
https://doi.org/10.1038/s41586-023-05881-4

Nori H., King N., McKinney S.M., Carignan D., and Horvitz E., 2023, Capabilities of GPT-4 on medical challenge problems, arXiv Preprint, 2303: 13375.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., and Lowe R., 2022, Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems, 35: 27730-27744.

Percha B., and Altman R.B., 2018, A global network of biomedical relationships derived from text, Bioinformatics, 34(15): 2614-2624.
https://doi.org/10.1093/bioinformatics/bty114

Shah N.H., Entwistle D., and Pfeffer M.A., 2023, Creation and adoption of large language models in medicine, JAMA, 330(9): 866-867.
https://doi.org/10.1001/jama.2023.14217

Singhal K., Azizi S., Tu T., Mahdavi S.S., Wei J., Chung H.W., Scales N.,Tanwani A., Cole-Lewis H., Pfohl S., Payne P., Seneviratne M., Gamble P., Kelly C., Babiker A., Schärli N., Chowdhery A., Mansfield P., Demner-Fushman D., Arcas B., Webster D., Corrado G., Matias Y., Chou K., Gottweis J., Tomasev N., Liu Y., Rajkomar A., Barral J., Semturs C., Karthikesalingam A., and Natarajan V., 2023, Large language models encode clinical knowledge, Nature, 620(7972): 172-180.
https://doi.org/10.1038/s41586-023-06291-2

Topol E.J., 2019, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine, 25(1): 44-56.
https://doi.org/10.1038/s41591-018-0300-7

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., and Polosukhin I., 2017, Attention is all you need, Advances in Neural Information Processing Systems, 30: 5998-6008.

Wang Y., Wang L., Rastegar-Mojarad M., Moon S., Shen F., Afzal N., Liu S., Zeng Y., Mehrabi S., Sohn S., and Liu H., 2018, Clinical information extraction applications: a literature review, Journal of Biomedical Informatics, 77: 34-49.
https://doi.org/10.1016/j.jbi.2017.11.011

Xu J., Kim S., Song M., Jeong M., Kim D., Kang J., Rousseau J., Li X., Xu W., Torvik V., Bu Y., Chen C., Akef Ebeid I., Li D., and Ding Y., 2020, Building a PubMed knowledge graph, Scientific Data, 7(1): 205.
https://doi.org/10.1038/s41597-020-0543-2