A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery

Anonymous ACL submission

Abstract

In many scientific fields, large language models (LLMs) have revolutionized the way with which text and other modalities of data (*e.g.,* molecules and proteins) are dealt, achieving superior performance in various applications and augmenting the scientific discovery process. Nevertheless, previous surveys on scientific LLMs often concentrate on one to two fields or a single modality. In this paper, we aim to provide a more holistic view of the research landscape by unveiling cross-field and cross-modal connections between scientific LLMs regarding their architectures and pre-training techniques. To this end, we comprehensively survey over 250 scientific LLMs, discuss their commonalities and differences, as well as summarize pre-training datasets and evaluation tasks for each field and modality. Moreover, we investigate how LLMs have been deployed to benefit scientific discovery. Resources related to this survey are available at `https://anonymous.4open.science/r/SciLLM-72F8`.

1 Introduction

The emergence of large language models (LLMs) (Zhao et al., 2023c) brings a new paradigm to natural language processing (NLP) by replacing specialized models designed for each task with unified models that are reasonably effective for a wide spectrum of problems. In the scientific domain, such a paradigm not only reshapes people's strategies to handle tasks related to natural language (*e.g.,* scientific papers, medical records, and climate reports) but also inspires analogous ideas to deal with other types of data (*e.g.,* molecules, proteins, tables, and metadata). In addition to understanding existing scientific data, LLMs have shown their potential to accelerate scientific discovery (Wang et al., 2023c; Zhang et al., 2023f; Wang et al., 2024b) through generation, planning, *etc.*

Given the broad and profound impact of LLMs in various scientific fields across diverse modalities, it becomes necessary to comprehensively review related work in this direction. However, existing scientific LLM surveys typically focus on either one to two fields (*e.g.,* biomedicine (Wang et al., 2023a; He et al., 2024; Pei et al., 2024) and chemistry (Xia et al., 2023; Zhang et al., 2024c)) or one modality (*e.g.,* text (Ho et al., 2024)) only. In fact, if we take a holistic view of the research landscape, we can observe similar and interrelated techniques used to develop LLMs for different fields and modalities.

Figure 1 depicts three major types of scientific LLM pre-training strategies (*i.e.,* COLUMNS 1 to 3), for each of which we give 4 examples (*i.e.,* TYPES A to D). In COLUMN 1, following BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), existing studies use masked language modeling (MLM) to pre-train encoder language models. Here, the input can be naturally sequential (*e.g.,* papers in each field, protein/DNA/RNA sequences in the FASTA format (Lipman and Pearson, 1985)) or artificially linearized (*e.g.,* molecules in the SMILES format (Weininger, 1988), sequences of venue/author/reference nodes in citation graphs). In COLUMN 2, inspired by GPT (Brown et al., 2020) and LLaMA (Touvron et al., 2023a), previous studies adopt next token prediction to pre-train (encoder-)decoder language models, some of which further adopt instruction tuning and preference optimization (Ouyang et al., 2022). Other than plain text input (*e.g.,* question-answer pairs from knowledge bases or exams), we see more ways to sequentialize complex scientific data, such as flattening table cells and using particle coordinates to describe crystals. Even for images, there are studies in both mathematics (Gao et al., 2023) and biomedicine (Li et al., 2023a) that exploit a vision encoder to project an image onto several visual tokens and prepend them to text tokens as linearized LLM input. In COLUMN 3, following DPR (Karpukhin et al., 2020) and CLIP (Radford et al., 2021), two encoders are pre-trained to map relevant data pairs closer in the latent space via contrastive learning. When both modalities are sequential (*e.g.,* text-text or text-protein), the model is built upon two LLM encoders. When we prefer

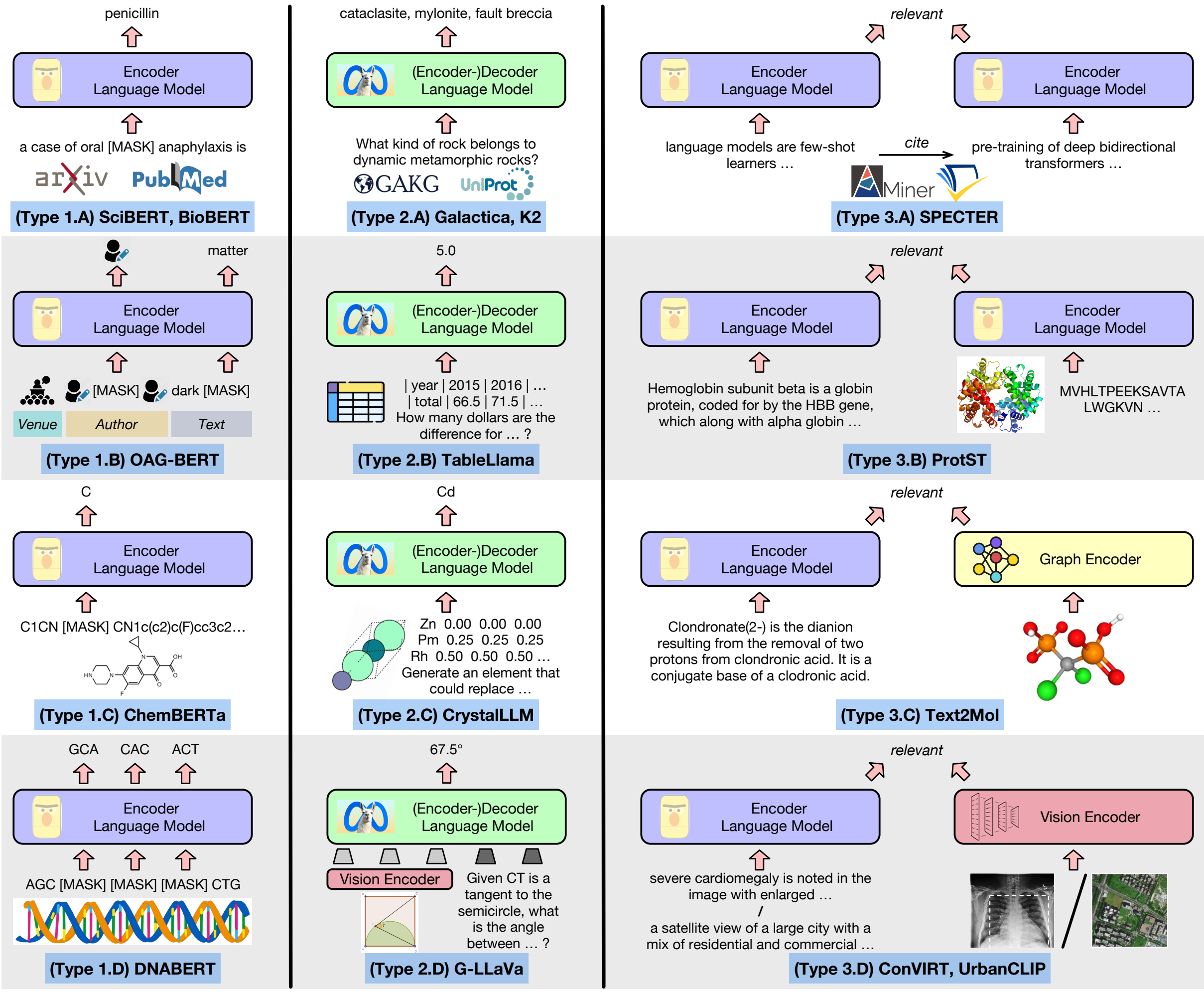

Figure 1: Three major types of scientific LLM pre-training techniques. (COLUMN 1): Pre-training encoder LLMs with sequentialized scientific data (*e.g.,* text, academic graphs, molecules, biological sequences) via masked language modeling. (COLUMN 2): Pre-training (encoder-)decoder LLMs with sequentialized scientific data (*e.g.,* text, tables, crystals, images) via next token prediction (possibly with instruction tuning). (COLUMN 3): Mapping text and relevant sequences/graphs/images closer in the latent space via contrastive learning.

to keep the non-sequential nature of one modality (*e.g.,* molecular graphs (Edwards et al., 2021), chest X-rays (Zhang et al., 2022), and aerial views (Yan et al., 2024)), the corresponding graph or image encoder can be employed. To summarize, a cross-field cross-modal survey will more accurately draw the connections between different scientific LLMs, demonstrate their commonalities, and potentially guide their future designs.

Contributions. In this paper, motivated by the discussions above, we systematically survey over 250 scientific LLMs encompassing various fields (*e.g.,* general science, mathematics, physics, chemistry, materials science, biology, medicine, and geoscience), modalities (*e.g.,* language, graph, vision, table, molecule, protein, genome, and climate time series), and sizes (from ∼100M to ∼100B parameters). For each field/modality, we investigate commonly adopted pre-training datasets, model architectures, and evaluation tasks of scientific LLMs. Following our motivation, when we talk about model architectures in detail, we link them back to Figure 1 to build cross-field cross-modal connections. Moreover, we provide a structured summary of these scientific LLMs in Table A1-Table A6 (Appendix A). Furthermore, for different fields, we introduce how LLMs have been deployed to benefit science by augmenting different aspects and stages of the scientific discovery process, such as hypothesis generation, theorem proving, experiment design, drug discovery, and weather forecasting.

2 LLMs in General Science (Table A1)

2.1 Language

The most commonly used pre-training corpora for scientific LLMs are research papers from bibliographic databases, such as AMiner (Tang et al., 2008), Microsoft Academic Graph (MAG) (Sinha et al., 2015), and Semantic Scholar (Ammar et al., 2018). Some of these sources (*e.g.,* S2ORC (Lo et al., 2020)) contain paper full-text information, while the others have titles and abstracts only.

The evolution of scientific LLMs bears similarity to that of general-domain LLMs. Specifically, pioneering models utilize paper text in a self-supervised way during pre-training, aiming to acquire scientific knowledge from large-scale unlabeled corpora. For example, masked language modeling (MLM) is the default pre-training task for scientific LLMs with a BERT backbone (TYPE 1.A in Figure 1, *e.g.,* SciBERT (Beltagy et al., 2019)); next token prediction is widely used for GPT-based scientific LLMs (TYPE 2.A in Figure 1, *e.g.,* SciGPT (Luu et al., 2021)). More recently, inspired by the fact that LLMs can be trained to follow natural language instructions (Wei et al., 2022a; Ouyang et al., 2022), researchers have put more effort into tuning LLMs with instructions to solve complex scientific problems (TYPE 2.A, *e.g.,* Galactica (Taylor et al., 2022) and SciGLM (Zhang et al., 2024a)). The instruction tuning data are often derived from datasets for downstream tasks, such as exam question answering (Welbl et al., 2017), and further filtered/augmented by humans or existing LLMs (*e.g.,* GPT-4 (Achiam et al., 2023)).

General scientific LLMs are usually evaluated on common NLP tasks, such as named entity recognition (NER), relation extraction (RE) (Luan et al., 2018), question answering (QA) (Wang et al., 2023g), and classification (Cohan et al., 2019).

2.2 Language + Graph

Beyond plain text, scientific papers are associated with rich metadata including venues, authors, and references (Zhang et al., 2023h). Such metadata connect papers into a graph that complements text signals for characterizing paper semantics. To exploit metadata, some studies (TYPE 1.B, *e.g.,* OAG-BERT (Liu et al., 2022b)) concatenate paper text with venues/authors as input and perform MLM on both text and metadata; others (TYPE 3.A, *e.g.,* SPECTER (Cohan et al., 2020)) take citation links as supervision and train LLMs to encode linked papers closer in the embedding space. Recent approaches further modify the Transformer architecture in LLMs with Adapters (Singh et al., 2023), GNN-nested Transformers (Jin et al., 2023b), and Mixture-of-Experts Transformers (Zhang et al., 2023g) to better capture graph signals.

Graph-aware scientific LLMs are often evaluated on tasks regarding the relation between two text units (*e.g.,* paper-paper or query-paper), including link prediction, retrieval, recommendation, and author name disambiguation. SciDocs (Cohan et al., 2020) and SciRepEval (Singh et al., 2023) are widely adopted benchmark datasets.

2.3 Applications in Scientific Discovery

Performant scientific LLMs can work alongside researchers throughout the entire scientific discovery process. Leaving field-specific applications for later sections, here we underscore LLMs' general usefulness in brainstorming and evaluation: Lahav et al. (2022) integrate LLMs into a search engine for the discovery of scientific challenges and directions; Wang et al. (2023f) and Baek et al. (2024) leverage LLMs to generate novel scientific ideas grounded in prior literature; Zhang et al. (2023i) rely on LLMs to find expert reviewers for each submission; Liu and Shah (2023), Liang et al. (2023a), and D'Arcy et al. (2024) explore the capacity of GPT-4 to provide useful feedback on research papers to facilitate automatic review generation.

3 LLMs in Mathematics (Table A2)

3.1 Language

The pre-training text corpora for math LLMs can be categorized into two classes: (1) multiple-choice QA, the representative datasets of which include MathQA (Amini et al., 2019), Ape210K (Zhao et al., 2020), and Math23K (Wang et al., 2017); as well as (2) generative QA, the representative datasets of which include GSM8K (Cobbe et al., 2021a), MATH (Hendrycks et al., 2021), and MetaMathQA (Yu et al., 2023b).

Similarly to general science LLMs, the backbone model of pioneering math LLMs is BERT (TYPE 1.A, *e.g.,* GenBERT (Geva et al., 2020) and MathBERT (Shen et al., 2021)). These models are mostly trained via MLM, with the only exception being BERT-TD (Li et al., 2022c), where a contrastive loss is adopted. For GPT-based math LLMs (TYPE 2.A, *e.g.,* GSM8K-GPT (Cobbe et al., 2021b) and NaturalProver (Welleck et al., 2022)), we find a diversity of pre-training tasks: supervised fine-tuning, next token prediction, and instruction tuning. The most recent math LLMs (TYPE 2.A, *e.g.,* Rho-Math (Lin et al., 2024a) and MAmmoTH2 (Yue et al., 2024)) are based on LLaMA and are trained to follow natural language instructions. However, when an enormous pre-training dataset (*e.g.,* 55 billion tokens) is available, next token prediction is still favored as the mere pre-training task (Azerbayev et al., 2023; Lin et al., 2024a) or the companion task (Shao et al., 2024; Ying et al., 2024) to build base models.

QA and math world problems (MWP) have been the most common evaluation tasks. In addition, quantitative reasoning contains more difficult problems, as the model has to provide a complete and self-contained solution without relying on external tools (Shao et al., 2024; Lin et al., 2024a). We see a dominance of use from GSM8K and MATH for QA, and from MathQA and Math23K for MWP. For quantitative reasoning, MMLU-STEM (Hendrycks et al., 2020) and Big-Bench Hard (Suzgun et al., 2023) are the most widely adopted.

3.2 Language + Vision

Geometry is one of the most important branches of mathematics, and it expresses the settings jointly in text and diagrams. As such, it is mandatory to involve the vision modality for geometric LLMs. The most commonly used pre-training datasets for geometric LLMs include Geometry3K (Lu et al., 2021) and GeoQA (Chen et al., 2021), both of which contain multiple-choice geometry problems.

The key to incorporating the vision modality to LLMs is to encode the images and obtain linearized visual representations. Specifically, Inter-GPS (Lu et al., 2021) (TYPE 2.D) uses RetinaNet (Lin et al., 2017) to transform images into a set of relationships and then applies BART (Lewis et al., 2020a) to produce the solution; G-LLaVA (Gao et al., 2023) (TYPE 2.D) encodes visual input via a pre-trained vision Transformer (ViT), concatenates visual embeddings with textual embeddings, and then feeds the concatenation into LLaMA-2 (Touvron et al., 2023b). These models are by default pre-trained via sequence-to-sequence tasks, where the problem is the input, and the ground-truth answer with optional rationale is the output. Auxiliary loss such as masked image modeling, image construction, or text-image matching, is optionally added for better visual modeling.

Geometric LLMs are evaluated through geometry problem solving, where the model is asked to select the correct answer given the diagram and its caption, the question, and answer options. Renowned evaluation datasets include Geometry3K (Lu et al., 2021), GEOS (Seo et al., 2015), and MathVista (Lu et al., 2023b).

3.3 Table

A large proportion of math knowledge is stored in the form of tabular data. For the "Table" modality, notable resources for pre-training include WikiTableQuestions (Pasupat and Liang, 2015), WikiSQL (Zhong et al., 2017), and WDC Web Table (Lehmberg et al., 2016).

The challenge in tables is similar to that in diagrams, namely to obtain linearized table representations. In most cases, tables are squeezed into linear text sequences as part of the context and are prepended with the question text as the model input. As one of the first works in this line of research, TAPAS (Herzig et al., 2020) (TYPE 1.A) adopts the MLM object to predict the masked token in textual and tabular context. The most recent developments (Li et al., 2023c; Zhang et al., 2024d) resemble the design of TableLlama (Zhang et al., 2023d) (TYPE 2.B), with LLaMA-2 as the backbone and instruction tuning as the pre-training task.

Table LLMs are validated through table QA, where the model is asked to produce the correct answer given the table structure, data values, and a question text. Most existing studies have been evaluated on the WikiTableQuestions and WikiSQL datasets. TableInstruct (Zhang et al., 2023d) is the most recently developed comprehensive benchmark integrating 14 datasets across 11 tasks.

3.4 Applications in Scientific Discovery

Math LLMs have great potential to assist humans in offering potential solutions. For instance, AlphaGeometry (Trinh et al., 2024) combines an LLM with a symbolic deduction engine, where the LLM generates useful constructs and the symbolic engine applies formal logic to find solutions. AlphaGeometry solves 25 out of 30 classical geometry problems adapted from the International Mathematical Olympiad. Sinha et al. (2024) extend AlphaGeometry by adding Wu's method (Chou, 1988), further solving 27 out of 30, surpassing human gold medalists. FunSearch (Romera-Paredes et al., 2024) integrates LLM with program search. One notable achievement of FunSearch is its ability to find a new solution to the cap set problem in combinatorial optimization. The solutions generated can be faster and more efficient than those devised by human experts. In Li et al. (2024a), LLMs iteratively propose and critique statistical models by leveraging in-context learning and chain-of-thought reasoning (Wei et al., 2022b).

4 LLMs in Physics (Table A3)

Existing physics LLMs largely focus on astronomy and the "Language" modality. As a derivative of BERT, astroBERT (Grezes et al., 2021) (TYPE 1.A) is further pre-trained using astronomy-related papers via MLM and next sentence prediction. It is evaluated on the NER task. Likewise, AstroLLaMA (Nguyen et al., 2023b) (TYPE 2.A) fine-tunes LLaMA-2 using over 300,000 astronomy abstracts from arXiv. It is evaluated on paper generation and paper recommendation tasks. AstroLLaMA-chat (Perkowski et al., 2024) (TYPE 2.A) is the chat version of AstroLLaMA. It is continually trained on a GPT-4 generated domain-specific dialogue dataset.

5 LLMs in Chemistry and Materials Science (Table A4)

5.1 Language

LLM pre-training corpora in chemistry and materials science typically come from research papers and databases (*e.g.,* Materials Project (Jain et al., 2013)). Besides, recent works adopt domain-specific instruction tuning datasets (e.g., Mol-Instructions (Fang et al., 2023a) and SMolInstruct (Yu et al., 2024)) derived from PubChem (Kim et al., 2019), MoleculeNet (Wu et al., 2018), *etc.*

Early studies on chemistry LLMs mostly adopt a moderate-sized encoder-only architecture pre-trained with MLM (TYPE 1.A, *e.g.,* ChemBERT (Guo et al., 2022), MatSciBERT (Gupta et al., 2022), and BatteryBERT (Huang and Cole, 2022)). These models are usually evaluated on downstream tasks including reaction role labeling (Guo et al., 2022) and abstract classification (Gupta et al., 2022). Recently, researchers have focused more on large-scale decoder-only LLMs trained with next token prediction and instruction tuning (TYPE 2.A). Examples include ChemDFM (Zhao et al., 2024), ChemLLM (Zhang et al., 2024b), and LlaSMol (Yu et al., 2024). Given the desired generalization capability of such models, they are evaluated on a diverse set of tasks such as name conversion (Kim et al., 2019), reaction prediction (Jin et al., 2017), retrosynthesis (Schneider et al., 2016), text-based molecule design (Edwards et al., 2022), and crystal generation (Antunes et al., 2023; Flam-Shepherd and Aspuru-Guzik, 2023; Gruver et al., 2024).

5.2 Language + Graph

Graphs are appropriate data structures for characterizing molecules (Jin et al., 2023a). Popular datasets containing molecular graphs include ChEBI-20 (Edwards et al., 2021, 2022), ZINC (Sterling and Irwin, 2015), and PCDes (Zeng et al., 2022).

In some scenarios, molecular graphs appear simultaneously with text information, thus existing works have explored how to encode both effectively. The first type of such models adopt a GNN as the graph encoder and an LLM as the text encoder. The two modalities are connected through contrastive learning (Liu et al., 2023d) (TYPE 3.C). For example, Text2Mol (Edwards et al., 2021) uses GCN (Kipf and Welling, 2016) and SciBERT to encode a molecule and its corresponding natural language description, respectively, for text-to-molecule retrieval. The second type of models utilize an LLM to encode text and graphs simultaneously (Zeng et al., 2022). Graphs can be either linearized to SMILES strings (Edwards et al., 2022) (TYPE 2.C) or projected onto virtual tokens with graph encoders (Zhao et al., 2023a; Liu et al., 2023f) (TYPE 2.D). For instance, 3D-MoLM (Li et al., 2024b) uses a 3D molecular encoder to represent molecules as tokens, and feed them together with instructions into LLaMA-2 for molecule-to-text retrieval and molecule captioning.

5.3 Language + Vision

Complementing text and graph modalities, molecular images form the vision modality in chemistry. Existing works adopt a similar philosophy to BLIP-2 (Li et al., 2023b), which represents each image as tokens and feed them into an LLM (TYPE 2.D). For example, GIT-Mol (Liu et al., 2024) projects all modalities, including graphs and images, into the latent text space and conducts encoding and decoding with T5 (Raffel et al., 2020).

5.4 Molecule

Different from subsection 5.2, this subsection introduces models dealing with molecules without associated text information. That being said, comparable approaches inspired by LLMs are utilized to develop molecular language models (Flam-Shepherd et al., 2022). To be specific, most studies adopt SMILES or SELFIES (Krenn et al., 2020) strings as the sequential representation of molecules. Similar to the trend in the “Language” modality, pioneering molecular LLMs focus on representation learning with bidirectional Transformer encoders (TYPE 1.C, *e.g.,* SMILES-BERT (Wang et al., 2019) and MoLFormer (Ross et al., 2022)). For instance, ChemBERTa (Chithrananda et al., 2020) adopts the architecture and pre-training strategy similar with those of RoBERTa (Liu et al., 2019). These models exhibit extraordinary abilities in molecular understanding tasks such as molecular property prediction (*e.g.,* toxicity classification (Wu et al., 2018) and atomization energy regression (Ramakrishnan et al., 2014)) as well as virtual screening (Riniker and Landrum, 2013). Later works explore representing molecules in an autoregressive fashion (TYPE 2.C, *e.g.,* BARTSmiles (Chilingaryan et al., 2022) and ChemGPT (Frey et al., 2023)). For instance, T5Chem (Lu and Zhang, 2022) adopts the T5 backbone and a sequence-to-sequence pre-training objective. These models are evaluated in generative tasks that include molecule generation (Gaulton et al., 2017), reaction prediction, and retrosynthesis. Besides linearizing molecules, there are studies modifying the Transformer architecture to admit molecular graphs, such as MAT (Maziarka et al., 2020) and R-MAT (Maziarka et al., 2024).

5.5 Applications in Scientific Discovery

Previous studies have shown that LLMs facilitate autonomous chemical research. For example, Bran et al. (2024) present a chemistry LLM agent, ChemCrow, that can integrate expert-designed tools for organic synthesis, drug discovery, and materials design; Boiko et al. (2023) develop an LLM-empowered intelligence system, Coscientist, that can design, plan, and perform chemical research. LLMs also help with drug and catalyst design. For instance, ChatDrug (Liu et al., 2023e) explores drug editing using LLMs with a prompt module, a domain feedback module, and a conversation module; DrugAssist (Ye et al., 2023a) is proposed as an LLM-based interactive model for molecule optimization through human-machine dialogue; Sprueill et al. (2023, 2024) use LLMs as agents to search for effective catalysts through Monte Carlo Tree Search and the feedback from an atomistic neural network model.

6 LLMs in Biology and Medicine (Table A5)

6.1 Language

Besides research articles (*e.g.,* titles/abstracts from PubMed (Lu, 2011) and full text from PMC (Beck and Sequeira, 2003)), pre-training corpora for biomedical LLMs include electronic health records (*e.g.,* MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023)), knowledge bases (e.g., UMLS (Bodenreider, 2004)), and health-related social media posts (*e.g.,* COVID-19 tweets (Müller et al., 2023)). Recent studies further collect supervised fine-tuning and preference optimization datasets from medical exam questions, knowledge graphs, and doctor-patient dialogues. Examples include ChiMed (Ye et al., 2023b), MedInstruct-52k (Zhang et al., 2023e), and BiMed1.3M (Acikgoz et al., 2024), many of which have non-English components (*e.g.,* Chinese and Arabic).

The watershed moment of biomedical LLM evolution is still the emergence of billion-parameter architectures and instruction tuning. Before that, a wide variety of moderate-sized backbones are explored, including both encoder-based (TYPE 1.A, *e.g.,* BioBERT (Lee et al., 2020), Bio-ELECTRA (Ozyurt, 2020), BioRoBERTa (Lewis et al., 2020b), BioALBERT (Naseem et al., 2022), and Clinical-Longformer (Li et al., 2022a)) and (encoder-)decoder-based ones (TYPE 2.A, *e.g.,* SciFive (Phan et al., 2021), BioBART (Yuan et al., 2022a), and BioGPT (Luo et al., 2022)). Evaluation tasks for these models range from biomedical NER, RE, sentence similarity estimation, document classification, and QA (*i.e.,* the BLURB bechmark (Gu et al., 2021)) to natural language inference (NLI) (Romanov and Shivade, 2018) and entity linking (Doğan et al., 2014). After the watershed, the trend becomes instruction-tuning billion-parameter LLMs (TYPE 2.A, *e.g.,* Med-PaLM (Singhal et al., 2023a), MedAlpaca (Han et al., 2023), and BioMistral (Labrak et al., 2024)). Accordingly, evaluation tasks become single-round QA (Jin et al., 2021; Pal et al., 2022) and multi-round dialogue (Wang et al., 2023h). Meanwhile, there are studies proposing a Bi-Encoder architecture (TYPE 3.A, *e.g.,* Jin et al. (2023c) and Xu et al. (2024)) that specifically targets biomedical retrieval tasks, the benchmarks of which are NFCorpus (Boteva et al., 2016), TREC-COVID (Voorhees et al., 2021), *etc.*

6.2 Language + Graph

Biomedical ontologies capture rich types of relations between entities. Analogously, citation links characterize connections between biomedical papers. Intuitively, jointly leveraging text and such graph information paves the way for multi-hop reasoning in QA. For instance, Yasunaga et al. (2022a) propose to use an LLM and a GNN to encode text and ontology signals, respectively, and deeply fuse them (TYPE 3.C); Yasunaga et al. (2022b) concatenate text segments from two linked papers together and feed the sequence into an LLM for pre-training, which is essentially appending a metadata neighbor (*i.e.,* reference) as context for MLM (TYPE 1.B). Both approaches demonstrate significant improvement on QA tasks that require complex reasoning.

6.3 Language + Vision

Biomedical text-image pairs typically come from two sources: (1) medical reports, such as chest X-rays (*e.g.,* MIMIC-CXR (Johnson et al., 2019)) and pathology reports (Huang et al., 2023b); as well as (2) figure-caption pairs extracted from biomedical papers (*e.g.,* ROCO (Pelka et al., 2018) and MedICaT (Subramanian et al., 2020)).

Most biomedical vision-language models exploit the CLIP architecture (Radford et al., 2021), where a text encoder and an image encoder is jointly trained to map the paired text and image closer via contrastive learning (TYPE 3.D). The choice of the text encoder evolves from BERT (Zhang et al., 2022) and GPT-2 (Huang et al., 2023b) to LLaMA (Wu et al., 2023) and LLaMA-2 (Liu et al., 2023b), while the image encoder evolves from ResNet (Huang et al., 2021) to ViT (Zhang et al., 2023c) and Swin Transformer (Thawkar et al., 2023). MLM, masked image modeling, and text-text/image-image contrastive learning (*i.e.,* by creating augmented views within the language/vision modality) are sometimes adopted as auxiliary pre-training tasks. Besides CLIP, other general-domain vision-language architectures, such as LLaVA (Li et al., 2023a), PaLM-E (Tu et al., 2024), and Gemini (Saab et al., 2024), have been explored. For instance, LLaVA-Med (TYPE 2.D) encodes images onto several visual tokens and prepends them to text tokens as the LLM input. Evaluation tasks of these models encompass image classification, segmentation, object detection, vision QA, text-to-image/image-to-text retrieval, and report generation, the benchmarks of which include CheXpert (Irvin et al., 2019), PadChest (Bustos et al., 2020), SLAKE (Liu et al., 2021a), *etc.*

6.4 Protein, DNA, RNA, and Multiomics

The FASTA format (Lipman and Pearson, 1985) naturally represents proteins as amino acid sequences and DNAs/RNAs as nucleotide sequences, enabling models to treat them as "languages". Representative resources of such sequences include UniRef (Suzek et al., 2015) and Swiss-Prot (Bairoch and Apweiler, 2000) for proteins, GRCh38 (Harrow et al., 2012) and the 1000 Genomes Project (Consortium, 2015) for DNAs, as well as RNAcentral (Consortium, 2019) for RNAs.

Encoder-only protein, DNA, and RNA LLMs (TYPE 1.D), such as ESM-2 (Lin et al., 2023b), DNABERT (Ji et al., 2021), and RNABERT (Akiyama and Sakakibara, 2022), adopt BERT-like architectures and MLM as the pre-training task (*i.e.*, predicting masked amino acids, nucleotides, k-mers, or codons); decoder-only models, such as ProGen (Madani et al., 2023) and DNAGPT (Zhang et al., 2023a), exploit GPT-like architectures and next token prediction as the pre-training task. There are also studies jointly considering text and protein modalities. For instance, ProtST (Xu et al., 2023b) matches protein sequences with their text descriptions (*i.e.,* names and functions) via contrastive learning (TYPE 3.B); BioMedGPT (Luo et al., 2023c) first projects proteins onto tokens and then inputs these tokens together with text into LLaMA-2 for instruction tuning, bearing similarity with TYPE 2.D.

Existing multiomics LLMs mainly focus on single-cell transcriptomics (*e.g.,* scRNA-seq) data, such as the expression levels of genes within a single cell (Franzén et al., 2019). Besides BERT-based (*e.g.,* Geneformer (Theodoris et al., 2023)) and GPT-based (*e.g.,* scGPT (Cui et al., 2024)) architectures, Performer (Yang et al., 2022a; Hao et al., 2024) is widely used due to its linear attention complexity in handling long scRNA-seq data.

6.5 Applications in Scientific Discovery

Similarly to chemistry, LLMs can automate experiments in biology and medicine research. For example, CRISPR-GPT (Huang et al., 2024) augments an LLM agent with domain knowledge to enhance the design process of CRISPR-based gene-editing experiments. Moreover, LLMs can encode biological sequences to capture structural properties, guide protein design, and evaluate the evolutionary fitness of viral variants. For instance, ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2023b) enable accurate structure prediction of proteins without expensive and time-consuming experiments; Ferruz and Höcker (2022) fine-tune LLMs on protein families, which can generate highly divergent but still potentially functional novel sequences; Hie et al. (2021) develop LLMs that can predict viral escape mutations.

7 LLMs in Geography, Geology, and Environmental Science (Table A6)

7.1 Language

Geoscience research papers, climate-related news articles, Wikipedia pages, corporate sustainability reports, knowledge bases (*e.g.,* GAKG (Deng et al., 2021)), and point-of-interest (POI) data (*e.g.,* OpenStreetMap (Haklay and Weber, 2008)) constitute the pre-training corpora of geoscience LLMs.

Preliminary research on geoscience LLMs focuses on pre-training bidirectional LLMs with the Transformer encoder backbone (TYPE 1.A, *e.g.,* ClimateBERT (Webersinke et al., 2021), SpaBERT (Li et al., 2022b), and MGeo (Ding et al., 2023)). For instance, SpaBERT and MGeo perform MLM on a sequence of geolocations for geographic entity linking and query-POI matching, respectively. More recently, related studies concentrate on scaling up decoding-style autoregressive LLMs in geoscience (TYPE 2.A, *e.g.,* K2 (Deng et al., 2024), OceanGPT (Bi et al., 2023b), and GeoGalactica (Lin et al., 2024b)). For instance, K2 and OceanGPT adapt LLaMA to geoscience and ocean science, respectively, via supervised fine-tuning with domain-specific instructions curated by human experts and/or augmented by general-domain LLMs. Evaluations of such models are conducted on geoscience benchmarks, such as GeoBench (Deng et al., 2024) and OceanBench (Bi et al., 2023b), which contain a broad range of tasks including QA, classification, knowledge probing, reasoning, summarization, and generation.

7.2 Language + Graph

Some geoscience applications involve graph signals, such as heterogeneous POI networks and knowledge graphs. To handle such signals and text jointly, ERNIE-GeoL (Huang et al., 2022) introduces a transformer-based aggregation layer to deeply fuse text and POI information within the BERT-based architecture; PK-Chat (Deng et al., 2023) combines an LLM with a pointer generation network on a knowledge graph to build a knowledge-driven dialogue system.

7.3 Language + Vision

Aerial views, together with location descriptions, profile urban regions. To deal with language and vision modalities jointly, UrbanCLIP (Yan et al., 2024) considers the CLIP architecture (TYPE 3.D), which is also widely adopted by biomedical vision-language models as mentioned in subsection 6.3, to perform text-image contrastive learning for urban indicator prediction.

7.4 Climate Time Series

The intuitions and methodologies used in LLMs also facilitate the construction of climate foundation models. Based on the ERA5 (Hersbach et al., 2020) and CMIP6 (Eyring et al., 2016) datasets of climate time series, previous studies exploit the ViT and Swin Transformer architectures to pre-train foundation models for weather forecasting. Representative models include FourCastNet (Pathak et al., 2022), Pangu-Weather (Bi et al., 2023a), *etc.*

7.5 Applications in Scientific Discovery

In geography, Wang et al. (2023b) and Zhou et al. (2024) highlight the potential of LLMs in urban planning from the sustainability, living, economic, disaster, and environmental perspectives. In geology, besides climate and weather forecasting, foundation models have been applied to simultaneous earthquake detection and phase picking (Mousavi et al., 2020). In environmental science, ChatClimate (Vaghefi et al., 2023) enhances GPT-4 by providing access to external, scientifically accurate knowledge on climate change to build a climate science conversational AI.

8 Challenges and Future Directions

In this survey, we compile literature that elucidates the data, architectures, and tasks used for scientific LLM pre-training, as well as how scientific LLMs have been applied to downstream applications in scientific discovery. In particular, we underscore analogous architectures, tasks, and trends observed during the evolution of scientific LLMs across different fields and modalities. Beyond reviewing prior research, we present several challenges to inspire further exploration of this topic.

Diving into Fine-Grained Themes. Most existing scientific LLMs target a coarse-grained field (*e.g.,* chemistry), while some tasks rely on highly specialized knowledge of a fine-grained theme (*e.g.,* Suzuki coupling). When LLMs are pre-trained on more general corpora, frequently appeared signals may dominate the model parameter space, and domain-specific tail knowledge may swipe out. We believe automatically curating in-depth, theme-focused knowledge graphs (Hope et al., 2021) and using them to guide the generation process will be a promising direction to tackle this issue.

Generalizing to Out-of-Distribution Scientific Data. In the scientific domain, it is common that the testing distribution shifts from the training distribution (Zhang et al., 2023f): novel scientific concepts keep emerging in newly published papers; unseen molecules with different scaffolds and unseen proteins with different numbers of peptide chains may appear during testing. Handling such out-of-distribution data remains a challenge for pre-trained scientific LLMs. To our knowledge, invariant learning (Arjovsky et al., 2019) can serve as the theoretical foundation for out-of-distribution analyses, and how to integrate it into LLM pre-training is worth exploring.

Facilitating Trustworthy Predictions. LLMs can generate plausible-sounding but factually incorrect output, commonly known as hallucination (Ji et al., 2023), which is particularly dangerous in high-stake scientific domains such as chemistry and biomedicine. To mitigate this issue, retrieval-augmented generation (RAG) provides LLMs with relevant, up-to-date, and trustworthy information. However, previous RAG studies in the scientific domain mainly focus on retrieving text (Xiong et al., 2024) and knowledge (Jin et al., 2024), while scientific data are heterogeneous and multi-modal. We envision that cross-modal RAG (*e.g.,* guiding text generation with relevant chemicals and proteins) will present additional opportunities to further enhance the trustworthiness of scientific LLMs.

Limitations

This survey mainly covers LLMs in mathematics and natural sciences. We are aware that LLMs can also significantly impact social sciences by achieving remarkable performance in representative tasks (Ziems et al., 2024) and serving as agents for social simulation experiments (Horton, 2023), but we leave the survey of these efforts as future work due to space limitations. In addition, this paper focuses on LLMs pre-trained on scientific data or augmented with domain-specific knowledge to benefit scientific discovery. There are studies (Wang et al., 2023g; Guo et al., 2023) proposing new benchmark datasets of scientific problems but evaluating the performance of general-purpose LLMs only, and we do not include these works in our survey. Furthermore, some LLMs may belong to more than one field or modality category given our classification criteria in the paper. For instance, BioMedGPT (Luo et al., 2023c) is pre-trained on biology and chemistry data jointly; GIT-Mol (Liu et al., 2024) considers the language, graph, and vision modalities simultaneously. For succinctness, we introduce each of them in only one subsection.

References

Hisham Abdel-Aty and Ian R Gould. 2022. Large-scale distributed training of transformers for chemical fingerprinting. *Journal of Chemical Information and Modeling*, 62(20):4852–4862.

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *AAAI'24*, pages 10757–10765.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emre Can Acikgoz, Osman Batur İnce, Rayene Bench, Arda Anıl Boz, İlker Kesen, Aykut Erdem, and Erkut Erdem. 2024. Hippocrates: An open-source framework for advancing large language models in healthcare. *arXiv preprint arXiv:2404.16621*.

Manato Akiyama and Yasubumi Sakakibara. 2022. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4(1):lqac012.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL'19*, pages 2357–2367.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. In *NAACL'18*, pages 84–91.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. 2023. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. 2022. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Amos Bairoch and Rolf Apweiler. 2000. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Research*, 28(1):45–48.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR'23*, pages 15016–15027.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. In *EMNLP'20*, pages 3122–3137.

Jeff Beck and Ed Sequeira. 2003. Pubmed central (pmc): An archive for literature from life sciences journals. *The NCBI Handbook*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP'19*, pages 3615–3620.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023a. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023b. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV'22*, pages 1–21.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *ECIR'16*, pages 716–722.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski, Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. 2020. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36(21):5255–5261.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS'20*, pages 1877–1901.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of EMNLP'20*, pages 4766–4777.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *COLING'20*, pages 669–679.

Jinho Chang and Jong Chul Ye. 2024. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1):2323.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022a. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *EMNLP'22*, pages 3313–3323.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of ACL'21*, pages 513–523.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. 2022b. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. 2023b. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. 2024. Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction. *Briefings in Bioinformatics*, 25(3):bbae163.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. 2023c. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023d. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023e. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022c. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *MICCAI'22*, pages 679–689.

Zhihong Chen, Guanbin Li, and Xiang Wan. 2022d. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *ACM MM'22*, pages 5152–5161.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. 2023. Prior: Prototype representation joint learning from medical images and reports. In *CVPR'23*, pages 21361–21371.

Zhoujun Cheng, Haoyu Dong, Ran Jia, Pengfei Wu, Shi Han, Fan Cheng, and Dongmei Zhang. 2022. Fortap: Using formulas for numerical-reasoning-aware table pretraining. In *ACL'22*, pages 1150–1166.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Lusine Khondkaryan, Karen Hambardzumyan, Zaven Navoyan, Hrant Khachatrian, and Armen Aghajanyan. 2022. Bartsmiles: Generative masked language models for molecular representations. *arXiv preprint arXiv:2211.16349*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Shang-Ching Chou. 1988. An introduction to wu's method for mechanical theorem proving in geometry. *Journal of Automated Reasoning*, 4(3):237–267.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M Church, et al. 2022. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *ICML'23*, pages 6140–6157.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. 2024. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, pages 1–12.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL'19*, pages 3586–3596.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *ACL'20*, pages 2270–2282.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, 526(7571):68–74.

The RNAcentral Consortium. 2019. Rnacentral: a hub of information for non-coding rna sequences. *Nucleic Acids Research*, 47(D1):D221–D229.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Cheng Deng, Yuting Jia, Hui Xu, Chong Zhang, Jingyao Tang, Luoyi Fu, Weinan Zhang, Haisong Zhang, Xinbing Wang, and Chenghu Zhou. 2021. Gakg: A multimodal geoscience academic knowledge graph. In *CIKM'21*, pages 4445–4454.

Cheng Deng, Bo Tong, Luoyi Fu, Jiaxin Ding, Dexing Cao, Xinbing Wang, and Chenghu Zhou. 2023. Pk-chat: Pointer network guided knowledge driven generative dialogue model. *arXiv preprint arXiv:2304.00592*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *WSDM'24*, pages 161–170.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL'19*, pages 4171–4186.

Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: Multimodal geographic language model pre-training. In *SIGIR'23*, pages 185–194.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *EMNLP'22*, pages 375–413.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *EMNLP'21*, pages 595–607.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. 2023. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE TPAMI*, 44(10):7112–7127.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. 2016. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023a. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. 2023b. Domain-agnostic molecular generation with self-feedback. *arXiv preprint arXiv:2301.11259*.

Noelia Ferruz and Birte Höcker. 2022. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348.

Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. 2023. Gena-lm: A family of open-source foundational dna language models for long sequences. *bioRxiv*, pages 2023–06.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. 2023. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. 2022. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293.

Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. 2019. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046.

Nathan C Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor W Coley, and Vijay Gadepally. 2023. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. 2017. The chembl database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *ACL'20*, pages 946–958.

Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avirup Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. 2021. Capturing row and column semantics in transformer based question answering over tables. In *NAACL'21*, pages 1212–1224.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, et al. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. 2024. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. 2022. Automated chemical reaction extraction from scientific literature. *Journal of Chemical Information and Modeling*, 62(9):2035–2045.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *NeurIPS'23*.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pages 1–11.

Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. 2012. Gencode: the reference human genome annotation for the encode project. *Genome Research*, 22(9):1760–1774.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. 2024. Foundation model for advancing healthcare: Challenges, opportunities, and future directions. *arXiv preprint arXiv:2404.03264*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *NeurIPS'21*.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *ACL'20*, pages 4320–4333.

Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. 2021. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288.

Xanh Ho, Anh Khoa Duong Nguyen, An Tuan Dao, Junfeng Jiang, Yuki Chida, Kaito Sugimoto, Huy Quoc To, Florian Boudin, and Akiko Aizawa. 2024. A survey of pre-trained language models for processing scientific text. *arXiv preprint arXiv:2401.17824*.

Zhi Hong, Aswathy Ajith, James Pauloski, Eamon Duede, Kyle Chard, and Ian Foster. 2023. The diminishing returns of masked language models to science. In *Findings of ACL'23*, pages 1270–1283.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel S Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021.

Extracting a knowledge base of mechanisms from covid-19 papers. In *NAACL'21*, pages 4489–4503.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *ICML'22*, pages 8946–8970.

Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *KDD'22*, pages 3029–3039.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV'21*, pages 3942–3951.

Shu Huang and Jacqueline M Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, 62(24):6365–6377.

Weijian Huang, Hongyu Zhou, Cheng Li, Hao Yang, Jiarun Liu, and Shanshan Wang. 2023a. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *arXiv preprint arXiv:2309.05904*.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023b. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. Tabbie: Pretrained representations of tabular data. In *NAACL'21*, pages 3446–3456.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. In *NeurIPS'23*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI'19*, pages 590–597.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. Omnitab: Pretraining with natural and synthetic data for few-shot table-based question answering. In *NAACL'22*, pages 932–942.

Zhanming Jie, Jierui Li, and Wei Lu. 2022. Learning to reason deductively: Math word problem solving as complex relation extraction. In *ACL'22*, pages 5944–5955.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023a. Large language models on graphs: A comprehensive survey. *arXiv preprint arXiv:2312.02783*.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*.

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023b. Patton: Language model pretraining on text-rich networks. In *ACL'23*, pages 7005–7020.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023c. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Wengong Jin, Connor W Coley, Regina Barzilay, and Tommi Jaakkola. 2017. Predicting organic reaction outcomes with weisfeiler-lehman network. In *NIPS'17*, pages 2604–2613.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP'20*, pages 6769–6781.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *ISBI'21*, pages 1033–1036.

Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. 2024. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, pages 1–12.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Christopher Kuenneth and Rampi Ramprasad. 2023. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature Communications*, 14(1):4099.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. 2022. A search engine for discovery of scientific challenges and directions. In *AAAI'22*, pages 11982–11990.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata. In *WWW'16*, pages 75–76.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL'20*, pages 7871–7880.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. In *NeurIPS'22*, pages 3843–3857.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS'23*.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR Medical Informatics*, 7(3):e14830.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML'23*, pages 19730–19742.

Michael Y Li, Emily B Fox, and Noah D Goodman. 2024a. Automated statistical model discovery with language models. *arXiv preprint arXiv:2402.17879*.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023c. Table-gpt: Table-tuned gpt for diverse table tasks. *arXiv preprint arXiv:2310.09263*.

Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. 2023d. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *MICCAI'23*, pages 374–383.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024b. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al. 2023e. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, pages 2023–09.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific Reports*, 10(1):7155.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022a. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023f. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022b. Spabert: A pretrained language model from geographic data for geo-entity representation. In *Findings of EMNLP'22*, pages 2757–2769.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023g. Geolm: Empowering language models for geospatially grounded language understanding. In *EMNLP'23*, pages 5227–5240.

Zhongli Li, Wenxuan Zhang, Chao Yan, Qingyu Zhou, Chao Li, Hongzhi Liu, and Yunbo Cao. 2022c. Seeking patterns, not just memorizing procedures: Contrastive learning for solving math word problems. In *Findings of ACL'22*, pages 2486–2496.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023a. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023b. Unimath: A foundational and multimodal mathematical reasoner. In *EMNLP'23*, pages 7126–7133.

Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. Mwp-bert: Numeracy-augmented pre-training for math word problem solving. In *Findings of NAACL'22*, pages 997–1009.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV'17*, pages 2980–2988.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *MICCAI'23*, pages 525–536.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023b. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024a. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, et al. 2024b. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

David J Lipman and William R Pearson. 1985. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021a. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ISBI'21*, pages 1650–1654.

Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023a. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *MICCAI'23*, pages 637–647.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021b. Self-alignment pretraining for biomedical entity representations. In *NAACL'21*, pages 4228–4238.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023b. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022a. Tapex: Table pre-training via learning a neural sql executor. In *ICLR'22*.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. 2023c. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023d. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2023e. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*.

Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022b. Oag-bert: Towards a unified backbone language model for academic knowledge services. In *KDD'22*, pages 3418–3428.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023f. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP'23*, pages 15623–15638.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *ACL'20*, pages 4969–4983.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. 2023a. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *CVPR'23*, pages 19764–19775.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL'21*, pages 6774–6786.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP'18*, pages 3219–3232.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *JAMIA*, page ocae037.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Yizhen Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023b. Molfm: A multi-modal molecular foundation model. *arXiv preprint arXiv:2307.09484*.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023c. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A Smith. 2021. Explaining relationships between scientific documents. In *ACL'21*, pages 2130–2144.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106.

Xin Man, Chenghong Zhang, Jin Feng, Changyu Li, and Jie Shao. 2023. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. 2020. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*.

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. 2024. Relative molecule self-attention transformer. *Journal of Cheminformatics*, 16(1):3.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. In *NeurIPS'21*, pages 29287–29303.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. 2021a. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129.

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021b. Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. In *EMNLP'21*, pages 4672–4681.

Giacomo Miolo, Giulio Mantoan, and Carlotta Orsenigo. 2021. Electramed: a new pre-trained language representation model for biomedical nlp. *arXiv preprint arXiv:2104.09585*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022. Lila: A unified benchmark for mathematical reasoning. In *EMNLP'22*, pages 5807–5832.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE JBHI*, 26(12):6070–6080.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *ML4H'23*, pages 353–367.

S Mostafa Mousavi, William L Ellsworth, Weiqiang Zhu, Lindsay Y Chuang, and Gregory C Beroza. 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1):3952.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022. Joint learning of localized representations from medical images and reports. In *ECCV'22*, pages 685–701.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *NAACL'22*, pages 4453–4470.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. 2022. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC Bioinformatics*, 23(1):144.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton M Rabideau, Yoshua Bengio, et al. 2023a. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In *NeurIPS'23*.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciuca, Charles O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Jason Jingsh Li, et al. 2023b. Astrollama: Towards specialized foundation models in astronomy. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 49–55.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. 2023c. Climax: A foundation model for weather and climate. In *ICML'23*, pages 25904–25938.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. 2023. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.

Maizhen Ning, Qiu-Feng Wang, Kaizhu Huang, and Xiaowei Huang. 2023. A symbolic characters aware model for solving geometry problems. In *ACM MM'23*, pages 7767–7775.

Janghoon Ock, Chakradhar Guntuboina, and Amir Barati Farimani. 2023. Catalyst energy prediction with catberta: Unveiling feature exploration strategies through large language models. *ACS Catalysis*, 13(24):16032–16044.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *EMNLP'22*, pages 11670–11688.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS'22*, pages 27730–27744.

Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 104–112.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *CHIL'22*, pages 248–260.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *ACL'15*, pages 1470–1480.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. 2024. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *EMNLP'23*, pages 1102–1123.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology objects in context (roco): a multimodal image dataset. In *7th Joint International Workshop, CVII-STENT and 3rd International Workshop, LABELS, Held in Conjunction with MICCAI'18*, pages 180–189.

Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jiraskova, Rickmer Braren, and Nassir Navab. 2023. Xplainer: From x-ray observations to explainable zero-shot diagnosis. In *MICCAI'23*, pages 420–429. Springer.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Ernest Perkowski, Rui Pan, Tuan Dung Nguyen, Yuan-Sen Ting, Sandor Kruk, Tong Zhang, Charlie O'Neill, Maja Jablonska, Zechang Sun, Michael J Smith, et al. 2024. Astrollama-chat: Scaling astrollama with conversational and diverse datasets. *Research Notes of the AAS*, 8(1):7.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. *arXiv preprint arXiv:2402.13253*.

Haoke Qiu, Lunyang Liu, Xuepeng Qiu, Xuemin Dai, Xiangling Ji, and Zhao-Yan Sun. 2024. Polync: a natural and chemical language model for the prediction of unified polymer properties. *Chemical Science*, 15(2):534–544.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML'21*, pages 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. Msa transformer. In *ICML'21*, pages 8844–8856.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86.

Sereina Riniker and Gregory A Landrum. 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15):e2016239118.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *ACL'18*, pages 1586–1596.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. 2023. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *EMNLP'23*, pages 15745–15756.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56(12):2336–2346.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. 2021a. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. 2021b. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP'15*, pages 1466–1476.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *IJCAI'19*, pages 5953–5959.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340*.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: larger biomedical domain language model. In *EMNLP'20*, pages 4700–4706.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *EMNLP'23*, pages 5548–5566.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW'15*, pages 243–246.

Shiven Sinha, Ameya Prabhu, Ponnurangam Kumaraguru, Siddharth Bhat, and Matthias Bethge. 2024. Wu's method can boost symbolic ai to rival silver medalists and alphageometry to outperform gold medalists at imo geometry. *arXiv preprint arXiv:2404.06405*.

Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2021. Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In *CIKM'21*, pages 1692–1702.

Henry Sprueill, Carl Edwards, Mariefel Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. 2023. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. In *Findings of EMNLP'23*, pages 8348–8365.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*.

Teague Sterling and John J Irwin. 2015. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. 2023. Bioclip: A vision foundation model for the tree of life. *arXiv preprint arXiv:2311.18803*.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2023. Saprot: protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. In *Findings of EMNLP'20*, pages 2112–2120.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. 2015. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of ACL'23*, pages 13003–13051.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *KDD'08*, pages 990–998.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *CVPR'23*, pages 7433–7442.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4).

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *SIGIR Forum*, volume 54, pages 1–12.

Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202*.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. In *NeurIPS'23*.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Dongjie Wang, Chang-Tien Lu, and Yanjie Fu. 2023b. Towards automated urban planning: When generative and chatgpt-like ai meets urban planning. *arXiv preprint arXiv:2304.03892*.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NeurIPS'22*, pages 33536–33549.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023c. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2024a. Drg-llama: tuning llama model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, 7(1):16.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023d. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023e. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023f. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*.

Qingyun Wang, Carl Edwards, Heng Ji, and Tom Hope. 2024b. Towards a human-computer collaborative scientific paper lifecycle: A pilot study and hands-on tutorial. In *COLING'24*, pages 56–67.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *ACM BCB'19*, pages 429–436.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023g. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023h. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *EMNLP'17*, pages 845–854.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. Tuta: Tree-based transformers for generally structured table pre-training. In *KDD'21*, pages 1780–1790.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP'22*.

Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. 2021. Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations. *Bioinformatics*, 37(3):404–412.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *ICLR'22*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS'22*, pages 24824–24837.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. In *NeurIPS'22*, pages 4913–4927.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. 2023. Cellplm: Pre-training of cell language model beyond single cells. *bioRxiv*, pages 2023–10.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *JAMIA*, page ocae045.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530.

Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z Li. 2023. A systematic survey of chemical pre-trained models. In *IJCAI'23*, pages 6787–6795.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. 2023. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Changwen Xu, Yuyang Wang, and Amir Barati Farimani. 2023a. Transpolymer: a transformer-based language model for polymer property predictions. *npj Computational Materials*, 9(1):64.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023b. Protst: Multi-modality learning of protein sequences and biomedical texts. In *ICML'23*, pages 38749–38767.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D Wang, Joyce C Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*.

Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. 2022. A japanese masked language model for academic domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 152–157.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *WWW'24*, pages 4006–4017.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022a. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. 2024a. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *AAAI'24*, pages 19368–19376.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *JAMIA*, 27(12):1935–1942.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022b. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Xianjun Yang, Junfeng Gao, Wenxin Xue, and Erik Alexandersson. 2024c. Pllama: An open-source large language model for plant science. *arXiv preprint arXiv:2401.01600*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. In *NeurIPS'22*, pages 37309–37323.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. Linkbert: Pretraining language models with document links. In *ACL'22*, pages 8003–8016.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2023a. Drugassist: A large language model for molecule optimization. *arXiv preprint arXiv:2401.10334*.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023b. Qilinmed: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Junqi Yin, Sajal Dash, Feiyi Wang, and Mallikarjun Shankar. 2023. Forge: pre-training open foundation models for science. In *SC'23*, pages 1–13.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *ACL'20*, pages 8413–8426.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *MICCAI'23*, pages 101–111.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Fei Yu, Anningzhe Gao, and Benyou Wang. 2023a. Ovm, outcome-supervised value models for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023b. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2021. Grappa: Grammar-augmented pre-training for table semantic parsing. In *ICLR'21*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. 2023. Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4(2):025035.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13(1):862.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv preprint arXiv:2401.07950*.

Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. 2023a. Dnagpt: A generalized pretrained tool for multiple dna sequence analysis tasks. *bioRxiv*, pages 2023–07.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024b. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023b. Huatuogpt, towards taming language model to be a doctor. In *Findings of EMNLP'23*, pages 10859–10885.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024c. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023c. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023d. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, et al. 2024d. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023e. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. 2023f. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. 2024e. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3.

Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023g. Pre-training multi-task contrastive learning models for scientific literature understanding. In *Findings of EMNLP'23*, pages 12259–12275.

Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, and Jiawei Han. 2023h. The effect of metadata on scientific literature tagging: A cross-field cross-model study. In *WWW'23*, pages 1626–1637.

Yu Zhang, Yanzhen Shen, Xiusi Chen, Bowen Jin, and Jiawei Han. 2023i. "Why should i review this paper?" unifying semantic, topic, and citation factors for paper-reviewer matching. *arXiv preprint arXiv:2310.14483*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *MLHC'22*, pages 2–25.

Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. 2023j. Text-guided foundation model adaptation for pathological image classification. In *MICCAI'23*, pages 272–282.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023a. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. In *NeurIPS'23*.

Suyuan Zhao, Jiahuan Zhang, and Zaiqing Nie. 2023b. Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv preprint arXiv:2306.04371*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023c. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022. Reastap: Injecting table reasoning skills during pre-training via synthetic reasoning examples. In *EMNLP'22*, pages 9006–9018.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. 2024. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. 2023. Structure-informed language models are protein designers. In *ICML'23*, pages 42317–42338.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

Zhilun Zhou, Yuming Lin, Depeng Jin, and Yong Li. 2024. Large language model for participatory urban planning. *arXiv preprint arXiv:2402.17161*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, et al. 2023. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705.

A Summary Tables of Scientific LLMs

Table A1-Table A6 summarize the modality, number of parameters, model architecture, pre-training data, pre-training task(s), and evaluation task(s) of scientific LLMs in each field. Within each field, we categorize models according to their modality; within each modality, we sort models chronologically. To be specific, if a paper has a preprint (*e.g.,* arXiv or bioRxiv) version, its publication date is according to the preprint service. Otherwise, its publication date is according to the conference proceeding or journal.

Table A1: Summary of LLMs in general science. “L”: Language; “L+G”: Language + Graph; “∼”: generally adopting the architecture but with modifications; “MLM”: masked language modeling; “NSP”: next sentence prediction; “NER”: named entity recognition; “RE”: relation extraction; “QA”: question answering.

Model	Modality	Size	Architecture	Pre-training Data	Pre-training Task(s)	Evaluation Task(s)
SciBERT (Beltagy et al., 2019)	L	110M	BERT	Semantic Scholar	MLM, NSP	NER, RE, classification, parsing
SciGPT2 (Luu et al., 2021)	L	117M	GPT-2	S2ORC	next token prediction	paper relationship explanation
CATTS (Cachola et al., 2020)	L	406M	BART	SciTLDR	sequence to sequence	paper summarization
SciNewsBERT (Smeros et al., 2021)	L	110M	BERT	news headlines	MLM, NSP	scientific claim extraction
ScholarBERT (Hong et al., 2023)	L	340M, 770M	BERT	Public.Resource.Org, Wikipedia, BookCorpus	MLM	NER, RE, classification
AcademicRoBERTa (Yamauchi et al., 2022)	L	125M	RoBERTa	CiNii	MLM	classification, author identification
Galactica (Taylor et al., 2022)	L	125M, 1.3B, 6.7B, 30B, 120B	Galactica	papers, code, reference materials, knowledge bases, web crawl data, instructions	next token prediction, instruction tuning	QA, link prediction, knowledge probing, quantitative reasoning, chemical name conversion, molecule classification, protein function prediction
DARWIN (Xie et al., 2023)	L	7B	LLaMA	papers, QA pairs, instructions	instruction tuning	QA, classification, regression
FORGE (Yin et al., 2023)	L	1.4B, 13B, 22B	GPT-NeoX	CORE, AMiner, MAG, SCOPUS, arXiv	next token prediction	QA, classification, regression
SciGLM (Zhang et al., 2024a)	L	6B, 32B	ChatGLM	SciInstruct	instruction tuning	QA, quantitative reasoning
SPECTER (Cohan et al., 2020)	L+G	110M	BERT	Semantic Scholar	link prediction	classification, link prediction, recommendation
OAG-BERT (Liu et al., 2022b)	L+G	110M	∼BERT	AMiner, PubMed, OAG	MLM	classification, link prediction, recommendation, retrieval, author name disambiguation
ASPIRE (Mysore et al., 2022)	L+G	110M	BERT	S2ORC	link prediction	paper similarity estimation
SciNCL (Ostendorff et al., 2022)	L+G	110M	BERT	Semantic Scholar	link prediction	classification, link prediction, recommendation
SPECTER 2.0 (Singh et al., 2023)	L+G	113M	Adapters	SciRepEval	classification, regression, link prediction, retrieval	classification, regression, link prediction, retrieval, author name disambiguation, paper-reviewer matching
SciPatton (Jin et al., 2023b)	L+G	–	GraphFormers	MAG	MLM, link prediction	classification, link prediction
SciMult (Zhang et al., 2023g)	L+G	138M	MoE	MAG, Semantic Scholar, SciRepEval	classification, link prediction, retrieval	classification, link prediction, recommendation, retrieval, patient-article/patient matching

Table A2: Summary of LLMs in mathematics. “L+V”: Language + Vision; “MWP”: math word problems. Other notations have the same meaning as in previous tables.

Model	Modality	Size	Architecture	Pre-training Data	Pre-training Task(s)	Evaluation Task(s)
GenBERT (Geva et al., 2020)	L	110M	BERT	Wikipedia	MLM, sequence to sequence	QA, MWP
MathBERT (Shen et al., 2021)	L	110M	BERT	arXiv, math curricula, syllabi, textbooks	MLM	classification, auto-grading
MWP-BERT (Liang et al., 2022)	L	110M	BERT	Ape210K	MLM, regression, classification	QA, MWP
BERT-TD (Li et al., 2022c)	L	110M	BERT	Math23K, MathQA	sequence to sequence, contrastive learning	QA, MWP
GSM8K-GPT (Cobbe et al., 2021b)	L	6B, 175B	GPT-3	GSM8K	supervised fine-tuning	QA, MWP
DeductReasoner (Jie et al., 2022)	L	125M	RoBERTa	MAWPS, Math23K, MathQA, SVAMP	sequence to sequence	QA, MWP
NaturalProver (Welleck et al., 2022)	L	175B	GPT-3	NaturalProofs	supervised fine-tuning	mathematical proof generation
Minerva (Lewkowycz et al., 2022)	L	8B, 62B, 540B	PaLM	arXiv, math web pages	next token prediction	QA, MWP, quantitative reasoning
Bhāskara (Mishra et al., 2022)	L	2.7B	GPT-Neo	Līla	instruction tuning	QA, MWP, knowledge probing
WizardMath (Luo et al., 2023a)	L	7B, 13B, 70B	LLaMA-2	GSM8K, MATH	instruction tuning	QA, MWP
MAmmoTH (Yue et al., 2023)	L	7B, 13B, 34B, 70B 7B	LLaMA-2 Mistral	MathInstruct	instruction tuning	QA, MWP
MetaMath (Yu et al., 2023b)	L	7B, 13B, 70B 7B	LLaMA-2 Mistral	MetaMathQA	instruction tuning	QA, MWP
ToRA (Gou et al., 2023)	L	7B, 13B, 34B, 70B	LLaMA-2	ToRA-Corpus	instruction tuning	QA, MWP
MathCoder (Wang et al., 2023e)	L	7B, 13B, 34B, 70B	LLaMA-2	MathCodeInstruct	instruction tuning	QA, MWP
Llemma (Azerbayev et al., 2023)	L	7B, 34B	LLaMA-2	Proof-Pile-2	next token prediction	QA, MWP, quantitative reasoning
OVM (Yu et al., 2023a)	L	7B 7B	LLaMA-2 Mistral	GSM8K	supervised fine-tuning	QA, MWP, quantitative reasoning
DeepSeekMath (Shao et al., 2024)	L	7B	DeepSeek	math web pages, instructions	next token prediction, instruction tuning	QA, MWP, quantitative reasoning, formal translation
InternLM-Math (Ying et al., 2024)	L	7B, 20B	InternLM2	Knowledge Pile, Proof-Pile-2, instructions	next token prediction, instruction tuning	QA, MWP, quantitative reasoning, formal translation
OpenMath (Toshniwal et al., 2024)	L	7B, 13B, 34B, 70B 7B	LLaMA-2 Mistral	OpenMathInstruct-1	instruction tuning	QA, MWP
Rho-Math (Lin et al., 2024a)	L	1B 7B	∼LLaMA-2 Mistral	OpenWebMath, SlimPajama, StarCoderData	next token prediction	QA, MWP, quantitative reasoning
MAmmoTH2 (Yue et al., 2024)	L	8B 7B 8×7B	LLaMA-3 Mistral Mixtral	WebInstruct	instruction tuning	QA, MWP, quantitative reasoning
Inter-GPS (Lu et al., 2021)	L+V	–	∼BART + RetinaNet	Geometry3K, GEOS	sequence to sequence	geometry problem solving
Geoformer (Chen et al., 2022a)	L+V	–	VL-T5 + ResNet	UniGeo	sequence to sequence	geometry problem solving

(Mathematics, Table Continued)						
Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
SCA-GPS (Ning et al., 2023)	L+V	–	RoBERTa + ViT	GeoQA, Geometry3K	masked image modeling, sequence to sequence	geometry problem solving
UniMath-Flan-T5 (Liang et al., 2023b)	L+V	–	Flan-T5 + VQ-VAE	SVAMP, GeoQA, TabMWP	image reconstruction, sequence to sequence	MWP, geometry problem solving
G-LLaVA (Gao et al., 2023)	L+V	7B, 13B	LLaVA	GeoQA+, Geometry3K	text-image matching, instruction tuning	geometry problem solving
TAPAS (Herzig et al., 2020)	Table	110M, 340M	BERT	Wikipedia	MLM	table QA
TaBERT (Yin et al., 2020)	Table	110M, 340M	BERT	Wikipedia, WDC Web Table	MLM, cell value recovery	table QA
GraPPa (Yu et al., 2021)	Table	355M	RoBERTa	Wikipedia	MLM, SQL semantic prediction	table QA
TUTA (Wang et al., 2021)	Table	110M	BERT	Wikipedia, WDC Web Table, spreadsheets	MLM, cell-level cloze, table context retrieval	cell type classification, table type classification
RCI (Glass et al., 2021)	Table	12M	ALBERT	WikiSQL, TabMCQ, WikiTableQuestions	classification	table QA
TABBIE (Iida et al., 2021)	Table	110M	ELECTRA	Wikipedia, VizNet	MLM, replaced token detection	column/row population, column type classification
TAPEX (Liu et al., 2022a)	Table	140M, 406M	BART	WikiTableQuestions	sequence to sequence	table QA
FORTAP (Cheng et al., 2022)	Table	110M	BERT	spreadsheets	MLM, numerical reference prediction, numerical calculation prediction	table QA, formula prediction, cell type classification
OmniTab (Jiang et al., 2022)	Table	406M	BART	Wikipedia	sequence to sequence	table QA
ReasTAP (Zhao et al., 2022)	Table	406M	BART	Wikipedia	sequence to sequence	table QA, table fact verification, table-to-text generation
Table-GPT (Li et al., 2023c)	Table	175B –	GPT-3.5 ChatGPT	instructions	instruction tuning	table QA, column-finding, missing-value identification, column type classification, data transformation, table matching, data cleaning
TableLlama (Zhang et al., 2023d)	Table	7B	LLaMA-2	TableInstruct	instruction tuning	table QA, RE, entity linking, column type classification, column/row population, table fact verification, cell description
TableLLM (Zhang et al., 2024d)	Table	7B, 13B	LLaMA-2	WikiTQ, FeTaQA, TAT-QA, WikiSQL, Spider	instruction tuning	table QA, table updating, table merging, table charting

Table A3: Summary of LLMs in physics. Notations have the same meaning as in previous tables.

Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
astroBERT (Grezes et al., 2021)	L	110M	BERT	NASA Astrophysics Data System	MLM, NSP	NER
AstroLLaMA (Nguyen et al., 2023b)	L	7B	LLaMA-2	arXiv	next token prediction	paper generation, paper similarity estimation
AstroLLaMA-Chat (Perkowski et al., 2024)	L	7B	LLaMA-2	QA pairs, LIMA, OpenOrca, UltraChat	instruction tuning	QA

Table A4: Summary of LLMs in chemistry and materials science. "L+G+V": Language + Graph + Vision; "KG": knowledge graph; "SMILES": simplified molecular-input line-entry system. Other notations have the same meaning as in previous tables.

Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
ChemBERT (Guo et al., 2022)	L	110M	BERT	chemistry journals	MLM	NER
MatSciBERT (Gupta et al., 2022)	L	110M	BERT	ScienceDirect	MLM	NER, RE, classification
MatBERT (Trewartha et al., 2022)	L	110M	BERT	materials science journals	MLM	NER
BatteryBERT (Huang and Cole, 2022)	L	110M	BERT	Elsevier, Springer, RSC	MLM	QA, classification
MaterialsBERT (Shetty et al., 2023)	L	110M	BERT	materials science journals	MLM, NSP	NER
CatBERTa (Ock et al., 2023)	L	125M	RoBERTa	OC20	regression	regression
LLM-Prop (Rubungo et al., 2023)	L	37M	T5 (encoder)	Materials Project	classification, regression	classification, regression
ChemDFM (Zhao et al., 2024)	L	13B	LLaMA	chemistry papers, textbooks, instructions	next token prediction, instruction tuning	QA, classification, name conversion, molecule captioning, text-based molecule design, reaction prediction, retrosynthesis
CrystalLLM (Gruver et al., 2024)	L	7B, 13B, 70B	LLaMA-2	Materials Project	instruction tuning	crystal generation
ChemLLM (Zhang et al., 2024b)	L	7B	InternLM2	QA pairs, ChemData	instruction tuning	QA, classification, name conversion, molecule captioning, text-based molecule design, reaction prediction, retrosynthesis
LlaSMol (Yu et al., 2024)	L	6.7B 7B 7B	Galactica LLaMA-2 Mistral	SMolInstruct	instruction tuning	QA, classification, regression, name conversion, molecule captioning, text-based molecule design, reaction prediction, retrosynthesis
Text2Mol (Edwards et al., 2021)	L+G	–	BERT + GCN	PubChem, ChEBI-20	text-graph matching	text-to-molecule retrieval
KV-PLM (Zeng et al., 2022)	L+G	110M	BERT	S2ORC, PubChem	text-graph matching	NER, RE, classification, text-to-molecule retrieval, molecule-to-text retrieval
MolT5 (Edwards et al., 2022)	L+G	60M, 220M, 770M	T5	C4, ZINC, ChEBI-20	sequence to sequence	molecule captioning, text-based molecule design
MoMu (Su et al., 2022)	L+G	–	BERT + GIN	S2ORC, PubChem	text-graph matching	classification, text-to-molecule retrieval, molecule-to-text retrieval, molecule captioning, text-based molecule design

(Chemistry and Materials Science, Table Continued)						
Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
MoleculeSTM (Liu et al., 2023d)	L+G	–	BERT + GIN	PubChem	text-graph matching	classification, text-to-molecule retrieval, molecule-to-text retrieval, text-based molecule design
Text+Chem T5 (Christofidellis et al., 2023)	L+G	60M, 220M	T5	Pistachio, ChEBI-20, experimental procedures	sequence to sequence	molecule captioning, text-based molecule design, reaction prediction, retrosynthesis, paragraph-to-action generation
GIMLET (Zhao et al., 2023a)	L+G	60M	∼T5	ChEMBL	instruction tuning	classification, regression
MolFM (Luo et al., 2023b)	L+G	–	∼BERT + GIN	S2ORC, PubChem	MLM, KG embedding, text-graph matching	classification, text-to-molecule retrieval, molecule-to-text retrieval, molecule captioning, text-based molecule design
MolCA (Liu et al., 2023f)	L+G	–	Galactica + GIN	PubChem	text-graph matching, graph-to-text generation	classification, name conversion, molecule-to-text retrieval, molecule captioning, functional group counting
InstructMol (Cao et al., 2023)	L+G	–	LLaMA + GIN	PubChem, MoleculeNet, ChEBI-20, USPTO	text-graph matching, instruction tuning	classification, regression, molecule captioning, reaction prediction, retrosynthesis, reagent selection
3D-MoLM (Li et al., 2024b)	L+G	–	LLaMA-2 + Uni-Mol	PubChem, 3D-MoIT	text-graph matching, graph-to-text generation, instruction tuning	QA, regression, molecule-to-text retrieval, molecule captioning
GIT-Mol (Liu et al., 2024)	L+G+V	–	BERT + GIN + Swin	PubChem, ChEBI-20	text-graph/image/text matching, supervised fine-tuning	classification, molecule captioning, text-based molecule design, molecule image recognition
SMILES-BERT (Wang et al., 2019)	Molecule	–	∼BERT	ZINC	MLM	classification
MAT (Maziarka et al., 2020)	Molecule	–	∼BERT	ZINC	masked node prediction	classification, regression
ChemBERTa (Chithrananda et al., 2020)	Molecule	125M	RoBERTa	PubChem	MLM	classification
MolBERT (Fabian et al., 2020)	Molecule	110M	BERT	ChEMBL	MLM, regression, SMILES equivalence	classification, regression, virtual screening
rxnfp (Schwaller et al., 2021b)	Molecule	110M	BERT	Pistachio, USPTO	classification	classification, reaction representation learning
RXNMapper (Schwaller et al., 2021a)	Molecule	770K	∼ALBERT	USPTO	MLM	atom-mapping
MoLFormer (Ross et al., 2022)	Molecule	47M	linear attention	PubChem, ZINC	MLM	classification, regression
Chemformer (Irwin et al., 2022)	Molecule	45M, 230M	∼BART	USPTO, ChEMBL, MoleculeNet	sequence to sequence, regression	regression, reaction prediction, retrosynthesis, molecule generation
R-MAT (Maziarka et al., 2024)	Molecule	–	∼BERT	ZINC, ChEMBL	masked node prediction, regression	classification, regression
MolGPT (Bagal et al., 2022)	Molecule	6M	∼GPT-1	ZINC, ChEMBL	next token prediction	molecule generation
T5Chem (Lu and Zhang, 2022)	Molecule	–	∼T5	PubChem	sequence to sequence	classification, regression, reaction prediction, retrosynthesis
ChemGPT (Frey et al., 2023)	Molecule	4.7M, 19M, 1.2B	∼GPT-Neo	PubChem	next token prediction	–
TransPolymer (Xu et al., 2023a)	Molecule	–	∼RoBERTa	PI1M	MLM	regression
polyBERT (Kuenneth and Ramprasad, 2023)	Molecule	86M	DeBERTa	density functional theory, experiments	MLM, regression	regression
MFBERT (Abdel-Aty and Gould, 2022)	Molecule	–	∼RoBERTa	GDB-13, ZINC, PubChem, ChEMBL, USPTO	MLM	classification, regression, virtual screening
SPMM (Chang and Ye, 2024)	Molecule	–	∼BERT	PubChem	next token prediction, SMILES-property matching	classification, regression, reaction prediction, retrosynthesis, SMILES-to-property generation, property-to-SMILES generation
BARTSmiles (Chilingaryan et al., 2022)	Molecule	406M	BART	ZINC	sequence to sequence	classification, regression, reaction prediction, retrosynthesis
MolGen (Fang et al., 2023b)	Molecule	406M 7B	BART LLaMA	ZINC, NPASS	sequence to sequence, prefix tuning	molecule generation
SELFormer (Yüksel et al., 2023)	Molecule	58M, 87M	∼RoBERTa	ChEMBL	MLM	classification, regression
PolyNC (Qiu et al., 2024)	Molecule	220M	T5	density functional theory, experiments	sequence to sequence	classification, regression

Table A5: Summary of LLMs in biology and medicine. “Multi”: Multiomics (*e.g.,* single-cell); “NLI”: natural language inference; “VQA”: visual question answering; “EHR”: electronic health record; “EMR”: electronic medical record; “PPI”: protein-protein interaction. Other notations have the same meaning as in previous tables.

Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
BioBERT (Lee et al., 2020)	L	110M, 340M	BERT	PubMed, PMC	MLM, NSP	NER, RE, QA
BioELMo (Jin et al., 2019)	L	93M	ELMo	PubMed	next token prediction, previous token prediction	NER, NLI
ClinicalBERT (Alsentzer et al., 2019)	L	110M	BERT	MIMIC-III	MLM, NSP	NER, NLI
ClinicalBERT (Huang et al., 2019)	L	110M	BERT	MIMIC-III	next token prediction, previous token prediction	word similarity estimation, hospital readmission prediction
BlueBERT (Peng et al., 2019)	L	110M, 340M	BERT	PubMed, MIMIC-III	MLM, NSP	NER, RE, NLI, classification, sentence similarity estimation
BEHRT (Li et al., 2020)	L	–	∼BERT	Clinical Practice Research Datalink	MLM	disease prediction
EhrBERT (Li et al., 2019)	L	–	∼BERT	MADE 1.0	entity linking	entity linking
Clinical XLNet (Huang et al., 2020)	L	110M	XLNet	MIMIC-III	permutation language modeling	mortality prediction
ouBioBERT (Wada et al., 2020)	L	110M	BERT	PubMed	MLM, NSP	NER, RE, NLI, classification, sentence similarity estimation
COVID-Twitter-BERT (Müller et al., 2023)	L	340M	BERT	COVID-19 tweets	MLM, NSP	classification, sentiment analysis, stance prediction
Med-BERT (Rasmy et al., 2021)	L	–	∼BERT	Cerner Health Facts	MLM, classification	disease prediction
Bio-ELECTRA (Ozyurt, 2020)	L	110M	ELECTRA	PubMed	MLM, replaced token detection	NER, QA

(Biology and Medicine, Table Continued)						
Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
BiomedBERT (Gu et al., 2021)	L	110M, 340M	BERT	PubMed, PMC	MLM, NSP	NER, RE, QA, classification, sentence similarity estimation
MCBERT (Zhang et al., 2020)	L	110M	BERT	Chinese media, encyclopedia, EHRs	MLM, NSP	NER, QA, classification, retrieval, paraphrase identification
BRLTM (Meng et al., 2021a)	L	–	∼BERT	EHRs	MLM	disease prediction
BioRedditBERT (Basaldella et al., 2020)	L	110M	BERT	Reddit	entity linking	entity linking
BioMegatron (Shin et al., 2020)	L	345M	BERT	PubMed, PMC	MLM, NSP	NER, RE, QA
SapBERT (Liu et al., 2021b)	L	110M	BERT	UMLS	synonym alignment	entity linking
ClinicalTransformer (Yang et al., 2020)	L	110M 125M 12M 110M 110M 149M 86M	BERT RoBERTa ALBERT ELECTRA XLNet Longformer DeBERTa	MIMIC-III	MLM, NSP, sentence order prediction, replaced token detection, permutation language modeling	NER
BioRoBERTa (Lewis et al., 2020b)	L	125M, 355M	RoBERTa	PubMed, PMC, MIMIC-III	MLM	NER, RE, NLI, classification
RAD-BERT (Bressem et al., 2020)	L	110M	BERT	radiology reports	MLM, NSP	classification
BioMedBERT (Chakraborty et al., 2020)	L	340M	BERT	BREATHE	MLM, NSP	NER, RE, QA, retrieval
LBERT (Warikoo et al., 2021)	L	–	∼BERT	PubMed	RE	RE
ELECTRAMed (Miolo et al., 2021)	L	110M	ELECTRA	PubMed	MLM, replaced token detection	NER, RE, QA
KeBioLM (Yuan et al., 2021)	L	110M	BERT	PubMed, UMLS	MLM, NER, entity linking	NER, RE, knowledge probing
SciFive (Phan et al., 2021)	L	220M, 770M	T5	PubMed, PMC	sequence to sequence	NER, RE, QA, NLI, classification
BioALBERT (Naseem et al., 2022)	L	12M, 18M	ALBERT	PubMed, PMC, MIMIC-III	MLM, sentence order prediction	NER, RE, QA, NLI, classification, sentence similarity estimation
Clinical-Longformer (Li et al., 2022a)	L	149M 110M	Longformer BigBird	MIMIC-III	MLM	NER, QA, NLI, classification
BioBART (Yuan et al., 2022a)	L	140M, 406M	BART	PubMed	sequence to sequence	NER, entity linking, summarization, dialogue
BioGPT (Luo et al., 2022)	L	355M, 1.5B	GPT-2	PubMed	next token prediction	RE, QA, classification, generation
Med-PaLM (Singhal et al., 2023a)	L	8B, 62B, 540B	PaLM	instructions	instruction tuning	QA
GatorTron (Yang et al., 2022b)	L	345M, 3.9B, 8.9B	BERT	Wikipedia, PubMed, PMC, MIMIC-III, clinical narratives	MLM	NER, RE, QA, NLI, sentence similarity estimation
ChatDoctor (Li et al., 2023f)	L	7B	LLaMA	HealthCareMagic	instruction tuning	dialogue
DoctorGLM (Xiong et al., 2023)	L	6B	ChatGLM	medical dialogues	instruction tuning	dialogue
BenTsao (Wang et al., 2023d)	L	7B	LLaMA	instructions	instruction tuning	QA, dialogue
MedAlpaca (Han et al., 2023)	L	7B, 13B	LLaMA	medical flash cards, Stack Exchange, WikiDoc	instruction tuning	QA
PMC-LLaMA (Wu et al., 2024)	L	7B, 13B	LLaMA	biomedical papers, books, instructions	next token prediction, instruction tuning	QA
Med-PaLM 2 (Singhal et al., 2023b)	L	8B, 62B, 540B	PaLM 2	instructions	instruction tuning	QA
HuatuoGPT (Zhang et al., 2023b)	L	7B, 13B	BLOOM	instructions	instruction tuning	QA, dialogue
MedCPT (Jin et al., 2023c)	L	110M	BERT	PubMed search logs	retrieval	classification, link prediction, recommendation, retrieval, sentence similarity estimation
Zhongjing (Yang et al., 2024b)	L	13B	Ziya-LLaMA	textbooks, QA pairs, knowledge bases, EHRs, EMRs, clinical reports, instructions	next token prediction, instruction tuning	QA
DISC-MedLLM (Bao et al., 2023)	L	13B	Baichuan	instructions	instruction tuning	QA, dialogue
DRG-LLaMA (Wang et al., 2024a)	L	7B, 13B	LLaMA	MIMIC-IV	classification	diagnosis-related group prediction
Qilin-Med (Ye et al., 2023b)	L	7B	Baichuan	ChiMed-CPT, ChiMed-SFT, ChiMed-DPO	next token prediction, instruction tuning	QA, dialogue
AlpaCare (Zhang et al., 2023e)	L	7B, 13B 7B, 13B	LLaMA LLaMA-2	MedInstruct-52k	instruction tuning	QA, summarization
BianQue (Chen et al., 2023d)	L	6B	ChatGLM	BianQueCorpus	instruction tuning	dialogue
HuatuoGPT-II (Chen et al., 2023a)	L	7B, 13B, 34B	Baichuan 2	instructions	instruction tuning	QA, dialogue
Taiyi (Luo et al., 2024)	L	7B	Qwen	instructions	instruction tuning	NER, RE, QA, classification
MEDITRON (Chen et al., 2023e)	L	7B, 70B	LLaMA-2	GAP-Replay	next token prediction, instruction tuning	QA
PLLaMa (Yang et al., 2024c)	L	7B, 13B	LLaMA-2	plant science journals, instructions	next token prediction, instruction tuning	QA
BioMistral (Labrak et al., 2024)	L	7B	Mistral	PMC	next token prediction	QA
Me LLaMA (Xie et al., 2024)	L	13B, 70B	LLaMA-2	PubMed, PMC, MIMIC-III, MIMIC-IV, MIMIC-CXR, RedPajama, instructions	next token prediction, instruction tuning	NER, RE, QA, NLI, classification, summarization
BiMediX (Pieri et al., 2024)	L	8×7B	Mixtral	BiMed1.3M	instruction tuning	QA
BioMedLM (Bolton et al., 2024)	L	2.7B	∼GPT-2	PubMed, PMC	next token prediction	QA
Hippocrates (Acikgoz et al., 2024)	L	7B 7B	LLaMA-2 Mistral	PubMed, PMC, medical guidelines, instructions	next token prediction, instruction tuning	QA
BMRetriever (Xu et al., 2024)	L	410M, 1B 2B 7B	Pythia Gemma Mistral	biomedical papers, textbooks, QA pairs, instructions	contrastive learning, instruction tuning	QA, recommendation, retrieval, entity linking, sentence similarity estimation
G-BERT (Shang et al., 2019)	L+G	–	BERT + GAT	MIMIC-III, ICD-9, ATC	MLM, diagnosis prediction, medication prediction	medication recommendation
CODER (Yuan et al., 2022b)	L+G	110M	BERT	UMLS	link prediction	entity linking, link prediction, entity similarity estimation
MoP (Meng et al., 2021b)	L+G	–	Adapters	UMLS	link prediction	QA, NLI, classification
BioLinkBERT (Yasunaga et al., 2022b)	L+G	110M, 340M	BERT	PubMed	MLM, link prediction	NER, RE, QA, classification, sentence similarity estimation
DRAGON (Yasunaga et al., 2022a)	L+G	360M	∼BERT + ∼GAT	PubMed, UMLS	MLM, link prediction	QA
ConVIRT (Zhang et al., 2022)	L+V	–	BERT + ResNet	MIMIC-CXR, musculoskeletal text-image pairs	text-image matching	classification, text-to-image retrieval, image-to-image retrieval
MMBERT (Khare et al., 2021)	L+V	–	BERT + ResNet	ROCO	MLM	VQA

(Biology and Medicine, Table Continued)

Model	Modality	Size	Architecture	Pre-training Data	Pre-training Task(s)	Evaluation Task(s)
MedViLL (Moon et al., 2022)	L+V	–	BERT + ResNet	MIMIC-CXR	MLM, text-image matching	VQA, classification, text-to-image retrieval, image-to-text retrieval, report generation
GLoRIA (Huang et al., 2021)	L+V	–	BERT + ResNet	CheXpert	text-image matching	classification, segmentation, image-to-text retrieval
LoVT (Müller et al., 2022)	L+V	–	BERT + ResNet	MIMIC-CXR	text-image matching	segmentation, detection
BioViL (Boecking et al., 2022)	L+V	–	BERT + ResNet	MIMIC-CXR	MLM, text-image matching	NLI, classification, segmentation, phrase grounding
M^{3}AE (Chen et al., 2022c)	L+V	–	RoBERTa + ViT	ROCO, MedICaT	MLM, masked image modeling, text-image matching	VQA, classification, text-to-image retrieval, image-to-text retrieval
ARL (Chen et al., 2022d)	L+V	–	BERT + ViT	ROCO, MedICaT, MIMIC-CXR	MLM, masked image modeling, text-image matching	VQA, classification, text-to-image retrieval, image-to-text retrieval
CheXzero (Tiu et al., 2022)	L+V	–	Transformer + ViT	MIMIC-CXR	text-image matching	classification
MGCA (Wang et al., 2022a)	L+V	–	BERT + ResNet / ViT	MIMIC-CXR	text-image matching	classification, segmentation, detection
MedCLIP (Wang et al., 2022b)	L+V	–	BERT + Swin	MIMIC-CXR, CheXpert	text-image matching	classification, image-to-text retrieval
BioViL-T (Bannur et al., 2023)	L+V	–	BERT + ResNet	MIMIC-CXR	MLM, text-image matching	classification, report generation, sentence similarity estimation
BiomedCLIP (Zhang et al., 2023c)	L+V	–	BERT + ViT	PMC figure-caption pairs, fine-grained text-image pairs	text-image matching	VQA, classification, text-to-image retrieval, image-to-text retrieval
PMC-CLIP (Lin et al., 2023a)	L+V	–	BERT + ResNet	PMC figure-caption pairs, subfigure-subcaption pairs	MLM, text-image matching	VQA, classification, text-to-image retrieval, image-to-text retrieval
Xplainer (Pellegrini et al., 2023)	L+V	–	BERT + ResNet	MIMIC-CXR	text-image matching	classification
RGRG (Tanida et al., 2023)	L+V	–	GPT-2 + ResNet	MIMIC-CXR	detection, classification, next token prediction	report generation
Med-UniC (Wan et al., 2023)	L+V	–	BERT + ResNet / ViT	MIMIC-CXR, PadChest	text-image matching, contrastive learning	classification, segmentation, detection
LLaVA-Med (Li et al., 2023a)	L+V	7B	LLaVA	PMC figure-caption pairs, instructions	text-image matching, instruction tuning	VQA
MI-Zero (Lu et al., 2023a)	L+V	–	BERT + CTransPath	histopathology figure-caption pairs	text-image matching	classification
XrayGPT (Thawkar et al., 2023)	L+V	–	LLaMA + Swin	MIMIC-CXR, Open-i	text-image matching	VQA
MONET (Kim et al., 2024)	L+V	–	BERT + ViT	PMC and textbook figure-caption pairs	text-image matching	classification, data auditing, model auditing
QuiltNet (Ikezogwo et al., 2023)	L+V	–	BERT + ViT	Quilt-1M	text-image matching	classification, text-to-image retrieval, image-to-text retrieval
MUMC (Li et al., 2023d)	L+V	–	BERT + ViT	ROCO, MedICaT, ImageCLEFmedical Caption	MLM, text-image matching	VQA
M-FLAG (Liu et al., 2023a)	L+V	–	BERT + ResNet	MIMIC-CXR	text-image matching	classification, segmentation, detection
PRIOR (Cheng et al., 2023)	L+V	–	BERT + ResNet	MIMIC-CXR	text-image matching, image reconstruction, sentence prototype generation	classification, segmentation, detection, image-to-text retrieval
Med-PaLM M (Tu et al., 2024)	L+V	12B, 84B, 562B	PaLM-E	MultiMedBench	instruction tuning	QA, VQA, classification, report generation, report summarization
CITE (Zhang et al., 2023j)	L+V	–	BERT + ViT	PatchGastric	text-image matching, prompt tuning	classification
Med-Flamingo (Moor et al., 2023)	L+V	–	Flamingo	PMC figure-caption pairs, textbooks	next token prediction	VQA
RadFM (Wu et al., 2023)	L+V	14B	LLaMA + ViT	MedMD, RadMD	next token prediction, instruction tuning	VQA, classification, report generation
PLIP (Huang et al., 2023b)	L+V	–	GPT-2 + ViT	Twitter text-image pairs, PathLAION	text-image matching	classification, text-to-image retrieval, image-to-image retrieval
MaCo (Huang et al., 2023a)	L+V	–	BERT + ViT	MIMIC-CXR	masked image modeling, text-image matching	classification, segmentation, phrase grounding
CXR-CLIP (You et al., 2023)	L+V	–	BERT + ResNet / Swin	MIMIC-CXR, CheXpert, ChestX-ray14	text-image matching, contrastive learning	classification, image-to-text retrieval
Qilin-Med-VL (Liu et al., 2023b)	L+V	–	LLaMA-2 + ViT	ChiMed-VL-Alignment, ChiMed-VL-Instruction	text-image matching, instruction tuning	VQA
BioCLIP (Stevens et al., 2023)	L+V	–	GPT-2 + ViT	TreeOfLife-10M	text-image matching	classification
M3D (Bai et al., 2024)	L+V	–	LLaMA-2 + ViT	M3D-Cap, M3D-VQA, M3D-RefSeg, M3D-Seg	text-image matching, instruction tuning	VQA, segmentation, text-to-image retrieval, image-to-text retrieval, report generation, 3D positioning
Med-Gemini (Saab et al., 2024)	L+V	–	Gemini	MedQA, LiveQA, HealthSearchQA, MedicationQA, MIMIC-III, SLAKE, PathVQA, ROCO, PAD-UFES-20, MIMIC-CXR, ECG-QA	instruction tuning	QA, VQA, signal QA, video QA, classification, long-form text generation, long EHR understanding
Med-Gemini-2D/3D/Polygenic (Yang et al., 2024a)	L+V	–	Gemini	SLAKE, MIMIC-CXR, Digital Knee X-ray, CXR-US2, NLST, CT-US1, PathVQA, Histopathology, PAD-UFES-20, EyePACS, PMC-OA, VQA-Med, UK Biobank	VQA, captioning, instruction tuning	VQA, classification, report generation, disease risk prediction
ProtTrans (Elnaggar et al., 2021)	Protein	420M 224M 409M 420M 3B, 11B	∼BERT ∼ALBERT ∼XLNet ∼ELECTRA T5	UniRef50, UniRef100, BFD	MLM, permutation language modeling, replaced token detection, sequence to sequence	secondary structure prediction, function prediction

(Biology and Medicine, Table Continued)						
Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
ESM-1b (Rives et al., 2021)	Protein	650M	~BERT	UniRef50, UniRef100	MLM	secondary structure prediction, contact prediction, remote homology detection
MSA Transformer (Rao et al., 2021)	Protein	100M	~BERT	UniRef50	MLM	secondary structure prediction, contact prediction
ESM-1v (Meier et al., 2021)	Protein	650M	~BERT	UniRef90	MLM	mutation effect prediction
AminoBERT (Chowdhury et al., 2022)	Protein	–	~BERT	UniParc	MLM, chunk permutation prediction	secondary structure prediction, contact prediction
ProteinBERT (Brandes et al., 2022)	Protein	16M	~BERT	UniRef90, Gene Ontology	MLM	secondary structure prediction, remote homology detection, fitness prediction
ProtGPT2 (Ferruz et al., 2022)	Protein	738M	GPT-2	UniRef50	next token prediction	secondary structure prediction, disorder prediction, protein sequence generation
ESM-IF1 (Hsu et al., 2022)	Protein	142M	Transformer + GVP-GNN	UniRef50	next token prediction	fixed backbone protein design, mutation effect prediction
ProGen (Madani et al., 2023)	Protein	1.6B	CTRL	UniParc, UniprotKB, Pfam, NCBI Taxonomy	next token prediction	protein sequence generation
ProGen2 (Nijkamp et al., 2023)	Protein	151M, 764M, 2.7B, 6.4B	~GPT-3	UniRef90, BFD	next token prediction	protein sequence generation, fitness prediction
ESM-2 (Lin et al., 2023b)	Protein	8M, 35M, 150M, 650M, 3B, 15B	~BERT	UniRef50, UniRef90	MLM	secondary structure prediction, contact prediction, 3D structure prediction
Ankh (Elnaggar et al., 2023)	Protein	450M, 1.1B	~T5	UniRef50	sequence to sequence	secondary structure prediction, contact prediction, embedding-based annotation transfer, remote homology detection, fitness prediction, localization prediction
ProtST (Xu et al., 2023b)	Protein	–	~BERT	Swiss-Prot	MLM, text-protein matching	fitness prediction, localization prediction, function annotation
LM-Design (Zheng et al., 2023)	Protein	659M	~BERT + ProtMPNN	CATH, UniRef50	MLM	fixed backbone protein design
ProteinDT (Liu et al., 2023c)	Protein	–	~BERT	Swiss-Prot	text-protein matching	text-to-protein generation, text-guided protein editing, secondary structure prediction, contact prediction, remote homology detection, fitness prediction
Prot2Text (Abdine et al., 2024)	Protein	256M, 283M, 398M, 898M	~BERT + R-GCN + ~GPT-2	Swiss-Prot	sequence to sequence	protein-to-text generation
BioMedGPT (Luo et al., 2023c)	Protein	10B	LLaMA-2 + GraphMVP + ESM-2	S2ORC, PubChemQA, UniProtQA	next token prediction, instruction tuning	QA
SaProt (Su et al., 2023)	Protein	35M, 650M	~BERT	UniRef50	MLM	mutation effect prediction, fitness prediction, localization prediction, function annotation, PPI prediction
BioT5 (Pei et al., 2023)	Protein	220M	T5	C4, ZINC, UniRef50, PubMed, PubChem, Swiss-Prot	sequence to sequence	molecule property prediction, protein property prediction, drug-target interaction prediction, PPI prediction, molecule captioning, text-based molecule design
ProLLaMA (Lv et al., 2024)	Protein	7B	LLaMA-2	UniRef50, instructions	next token prediction, instruction tuning	protein sequence generation, protein property prediction
DNABERT (Ji et al., 2021)	DNA	110M	BERT	GRCh38	MLM	chromatin profile prediction, promoter prediction, splice site prediction, functional genetic variant identification
GenSLMs (Zvyagin et al., 2023)	DNA	25M, 250M, 2.5B, 25B	~GPT-2	prokaryotic gene sequences	next token prediction	SARS-CoV-2 genome evolution prediction
Nucleotide Transformer (Dalla-Torre et al., 2023)	DNA	50M, 100M, 250M, 500M	~BERT	GRCh38, 1000 Genomes, multispecies genomes	MLM	chromatin profile prediction, enhancer prediction, promoter prediction, epigenetic marks prediction, splice site prediction
GENA-LM (Fishman et al., 2023)	DNA	110M, 340M 110M	BERT BigBird	T2T-CHM13, 1000 Genomes, multispecies genomes	MLM	enhancer prediction, promoter prediction, epigenetic marks prediction, splice site prediction, species classification
DNABERT-2 (Zhou et al., 2023)	DNA	110M	BERT	GRCh38, multispecies genomes	MLM	chromatin profile prediction, promoter prediction, epigenetic marks prediction, splice site prediction, species classification, SARS-CoV-2 variant prediction, enhancer-promoter interaction
HyenaDNA (Nguyen et al., 2023a)	DNA	0.4M, 3.3M, 6.6M	Hyena	GRCh38	next token prediction	chromatin profile prediction, enhancer prediction, promoter prediction, epigenetic marks prediction, splice site prediction, species classification
DNAGPT (Zhang et al., 2023a)	DNA	0.1B, 3B 6.6M	~GPT-3	Ensembl	next token prediction, sequence order prediction, regression	genome generation, chromatin profile prediction, promoter prediction, genomic signals and regions recognition
RNABERT (Akiyama and Sakakibara, 2022)	RNA	–	~BERT	RNAcentral	MLM	RNA structural alignment, RNA clustering

(Biology and Medicine, Table Continued)						
Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
RNA-FM (Chen et al., 2022b)	RNA	–	∼BERT	RNAcentral	MLM	secondary structure prediction, 3D structure prediction, protein-RNA interaction, mean ribosome load prediction
SpliceBERT (Chen et al., 2024)	RNA	19.4M	∼BERT	UCSC genome browser	MLM	human branchpoint prediction, splice site prediction
RNA-MSM (Zhang et al., 2024e)	RNA	–	∼BERT	Rfam	MLM	secondary structure prediction, solvent accessibility prediction
CodonBERT (Li et al., 2023e)	RNA	–	∼BERT	mRNA sequences	MLM, homologous sequences prediction	mRNA property prediction
UTR-LM (Chu et al., 2024)	RNA	–	∼BERT	5' UTR sequences	MLM, classification, regression	mean ribosome load prediction, mRNA property prediction, internal ribosome entry site prediction
scBERT (Yang et al., 2022a)	Multi	–	Performer	PanglaoDB	MLM	cell type annotation, novel cell type discovery
scGPT (Cui et al., 2024)	Multi	–	∼GPT-3	CELLxGENE	MLM	cell type annotation, perturbation response prediction, multi-batch integration, multi-omic integration, gene network inference
scFoundation (Hao et al., 2024)	Multi	100M	Transformer + Performer	scRNA-seq data	MLM	cell clustering, drug response prediction, perturbation response prediction, cell type annotation, gene network inference
Geneformer (Theodoris et al., 2023)	Multi	10M, 40M	∼BERT	Genecorpus-30M	MLM	gene dosage sensitivity prediction, chromatin dynamics prediction, network dynamics prediction
CellLM (Zhao et al., 2023b)	Multi	–	Performer	PanglaoDB, CancerSCEM	MLM, classification, contrastive learning	cell type annotation, drug sensitivity prediction
CellPLM (Wen et al., 2023)	Multi	82M	Transformer	scRNA-seq data, spatially-resolved transcriptomic data	MLM	cell clustering, scRNA-seq denoising, spatial transcriptomic imputation, cell type annotation

Table A6: Summary of LLMs in geography, geology, and environmental science. "Climate": Climate Time Series; "POI": point of interest. Other notations have the same meaning as in previous tables.

Model	**Modality**	**Size**	**Architecture**	**Pre-training Data**	**Pre-training Task(s)**	**Evaluation Task(s)**
ClimateBERT (Webersinke et al., 2021)	L	82M	DistilRoBERTa	climate-related news, papers, corporate climate reports	MLM	classification, fact-checking
SpaBERT (Li et al., 2022b)	L	110M, 340M	BERT	OpenStreetMap	MLM, masked entity prediction	entity typing, entity linking
MGeo (Ding et al., 2023)	L	213M	∼BERT	text-geolocation pairs	MLM, masked geographic modeling, contrastive learning	query-POI matching
K2 (Deng et al., 2024)	L	7B	LLaMA	geoscience papers, Wikipedia, instructions	next token prediction, instruction tuning	QA
OceanGPT (Bi et al., 2023b)	L	7B	LLaMA-2	ocean science papers, instructions	next token prediction, instruction tuning	QA, classification, extraction, knowledge probing, commonsense reasoning, summarization, generation
ClimateBERT-NetZero (Schimanski et al., 2023)	L	82M	DistilRoBERTa	Net Zero Tracker	classification	classification
GeoLM (Li et al., 2023g)	L	110M, 340M	BERT	OpenStreetMap, Wikipedia	MLM, contrastive learning	NER, RE, entity typing, entity linking
GeoGalactica (Lin et al., 2024b)	L	30B	Galactica	geoscience papers, code, Wikipedia, instructions	next token prediction, instruction tuning	QA, knowledge probing, quantitative reasoning, summarization, generation
ERNIE-GeoL (Huang et al., 2022)	L+G	–	Transformer + graph aggregation	Baidu Maps (POI database, search logs)	MLM, geocoding	classification, query-POI matching, address parsing, geocoding, next POI recommendation
PK-Chat (Deng et al., 2023)	L+G	132M	∼UniLM	Geoscience Academic Knowledge Graph	next token prediction, bag-of-words prediction, classification	task-oriented dialogue
UrbanCLIP (Yan et al., 2024)	L+V	–	Transformer + ViT	satellite images, location descriptions,	next token prediction, text-image matching	urban indicator prediction
FourCastNet (Pathak et al., 2022)	Climate	–	∼ViT	ERA5	regression	weather forecasting
Pangu-Weather (Bi et al., 2023a)	Climate	–	∼Swin	ERA5	regression	weather forecasting
ClimaX (Nguyen et al., 2023c)	Climate	–	∼ViT	CMIP6	regression	weather forecasting, climate projection, climate model downscaling
FengWu (Chen et al., 2023b)	Climate	–	Transformer	ERA5	regression	weather forecasting
W-MAE (Man et al., 2023)	Climate	–	ViT	ERA5	masked image modeling	weather forecasting
FuXi (Chen et al., 2023c)	Climate	–	∼Swin V2	ERA5	regression	weather forecasting