PharmaGPT: Domain-Specific Large Language Models for Bio-Pharmaceutical and Chemistry

Linqing Chen, Weilei Wang*, ZilongBai, Peng Xu, Yan Fang, Jie Fang, Wentao Wu, Lizhi Zhou, Ruiji Zhang, Yubin Xia, Chaobo Xu, Ran Hu, Licong Xu, Qijun Cai, Haoran Hua, Jing Sun, Jin Liu, Tian Qiu, Haowen Liu, Meng Hu, Xiuwen Li, Fei Gao, Yufu Wang, Lin Tie, Chaochao Wang,
Jianping Lu, Cheng Sun, Yixin Wang, Shengjie Yang, Yuancheng Li, Lu Jin, Lisha Zhang,
Fu Bian, Zhongkai Ye, Lidong Pei, Changyang Tu
PatSnap Co., LTD. Suzhou, Jiangsu
{chenlinqing, wangweilei}@patsnap.com

Abstract

Large language models (LLMs) have revolutionized Natural Language Processing (NLP) by minimizing the need for complex feature engineering. However, the application of LLMs in specialized domains like biopharmaceuticals and chemistry remains largely unexplored. These fields are characterized by intricate terminologies, specialized knowledge, and a high demand for precision areas where general purpose LLMs often fall short. In this study, we introduce PharmaGPT, a suite of domain specilized LLMs with 13 billion and 70 billion parameters, specifically trained on a comprehensive corpus tailored to the BioPharmaceutical and Chemcal domains. Our evaluation shows that PharmaGPT surpasses existing general models on specific-domain benchmarks such as NAPLEX, demonstrating its exceptional capability in domain-specific tasks. Remarkably, this performance is achieved with a model that has only a fraction—sometimes just one-tenth—of the parameters of general-purpose large models. This advancement establishes a new benchmark for LLMs in the bio-pharmaceutical and chemical fields, addressing the existing gap in specialized language modeling. It also suggests a promising path for enhanced research and development, paving the way for more precise and effective NLP applications in these areas.

1 Introduction

The development of Large Language Models (LLMs) has significantly transformed the landscape of natural language processing (NLP). Recent advancements, as evidenced by models such as GPT [Radford *et al.*, 2018], have reduced the dependence on extensive feature engineering, thereby simplifying the creation of complex NLP systems [Sarzynska-Wawer *et al.*, 2021; Howard and Ruder, 2018]. These models have demonstrated remarkable capabilities in understanding and generating nuanced text based on few prompts.

* Corresponding author

Despite these advances, current LLMs exhibit significant limitations when applied to specialized vertical domains, particularly in the biopharmaceutical sector. Most state-of-theart LLMs are proprietary and primarily designed for generalpurpose applications, predominantly in English. This focus has resulted in a lack of depth and precision necessary for domains requiring specialized knowledge.

To address these shortcomings, we introduce PharmaGPT, a suite of multilingual LLMs with configurations of 13 billion and 70 billion parameters. These models have been meticulously trained on a diverse corpus of billions of tokens, ensuring a robust linguistic foundation across multiple languages. Our evaluations on benchmarks like Naplex have demonstrated that PharmaGPT not only competes with but often surpasses existing models in specialized NLP tasks.

PharmaGPT is distinctively proficient in the biopharmaceutical and chemical sectors, showcasing exceptional command over specialized terminology and concepts. This proficiency is crucial, as effective domain-specific language models require a deep understanding of both general and specialized language contexts.

Through this work, we aim to contribute powerful models to both the academic and professional communities, fostering further innovation in the development and application of large-scale, multilingual, domain-specific language models. PharmaGPT promotes inclusivity and broadens the scope for global collaboration, pushing the boundaries of what is achievable in NLP. Despite the inclusive nature of LLM development, a significant portion of the research community remains underrepresented, primarily due to the English-centric training data prevalent in most models. While there are

advancements in models processing languages like Chinese [Wang *et al.*, 2021; Zeng *et al.*, 2022; Kim *et al.*, 2021], the representation of linguistic diversity is still limited. Furthermore, the accessibility and applicability of 'vertical domain LLMs' are often confined to their intended uses, particularly in critical domains such as biopharmaceuticals and chemicals.

In response to these challenges, this paper introduces the PharmaGPT, a domain-Specific large language models for bio-Pharmaceutical and chemistry. Our development process was rigorously structured, involving the creation of a diverse training dataset, a nuanced training objective, and an advanced engineering strategy for efficient distributed learning, detailed in Sections 2.2, 3.1, and 3.2, respectively. A comprehensive analysis of the model's performance capabilities is outlined in Section 4, demonstrating its competitive edge over existing multilingual LLMs. Our primary goal extends beyond introducing another LLM; we provide a detailed roadmap of our systematic approach to its development, aiming to set a new standard for future developments in domainspecific NLP.

2 Background

To provide a solid foundation for understanding the PharmaGPT model, this section first outlines the crucial background on Large Language Models (LLMs). We discuss the evolution of LLMs, their underlying technologies, and their pivotal impact on the field of natural language processing (NLP). This exploration not only highlights the technological advancements but also sets the stage for appreciating the innovative applications of these models in various domains. Subsequently, we present an organizational overview of the PharmaGPT initiative. This overview details the project's scope, objectives, and the collaborative dynamics that have driven the development of this sophisticated, domain-specific language model. By establishing this comprehensive context, we aim to deepen the reader's understanding of the model's architecture, training methodologies, and its distinctive contributions to the fields of Bio-Pharmaceutical and Chemical research. The introduction of this model marks a significant milestone in applying advanced NLP techniques to specialized scientific domains. Through this section, we prepare the reader to fully appreciate the technical nuances and sectorspecific innovations that PharmaGPT brings to the forefront of NLP application in science.

2.1 Language Modeling

Language modeling constitutes a pivotal task in computational linguistics, aiming to quantify the likelihood of a sequence of tokens within a given text. A 'token' herein refers to a discrete unit of textual content, which can vary in granularity from words, subwords, characters, to bytes, as elucidated by Mielke[Mielke *et al.*, 2021]. This study—aligning with the prevailing paradigms in language modeling—endeavors to ascertain the joint probability distribution of tokens in a textual corpus as delineated by the following equation:

$$p(\mathcal{X}) = P(x_i|x_1, \ldots, x_{i-1}) = \prod_{i=1}^{n} P(x_i|x_1, x_2, \ldots, x_{i-1}) \quad (1)$$

where X represents a sequence of tokens, x_i denotes the i^{th} token in the sequence, and $x_{<i}$ symbolizes the sequence of tokens preceding the i^{th} token. This methodology, termed autoregressive language modeling, epitomizes the process of sequentially predicting the probability of the ensuing token, thus facilitating a more nuanced understanding and generation of natural language. This approach not only adheres to the foundational principles of probabilistic modeling but also leverages contemporary computational techniques to enhance predictive accuracy and linguistic coherence.

Neural Language Models The inception of pre-trained models in natural language processing marked a significant shift towards leveraging vast amounts of unlabeled data for model initialization prior to fine-tuning on task-specific datasets. These early models, exemplified by Word2Vec [Mikolov *et al.*, 2013], GloVe [Pennington *et al.*, 2014], and ELMo [Shahbaz *et al.*, 2019], introduced the concept of embedding words into continuous vector spaces, capturing semantic and syntactic nuances to a degree previously unattainable with discrete representations. Word2Vec and GloVe focused on static word embeddings, where each word was assigned a single vector regardless of its context, limiting their ability to address the polysemy inherent in natural language. ELMo, on the other hand, advanced the field by generating context-dependent representations, using a bidirectional LSTM architecture trained on a language modeling objective. These foundational models set the stage for subsequent advancements in pre-trained language models, paving the way for more sophisticated architectures that better capture the complexities of language through deep learning. Their introduction has not only revolutionized the approach to natural language processing tasks but also established a paradigm in which pre-training on large-scale corpora has become a cornerstone of modern NLP methodologies.

Neural language Models (NLMs) emerged as a sophisticated alternative to traditional n-gram models, initially proposed by [Miikkulainen and Dyer, 1991] and Schmidhuber and [Schmidhuber and Heil, 1996], with their significance further underscored by [Bengio *et al.*, 2000] through the adoption of neural networks for predicting the probability of subsequent tokens based on preceding ones. Initial implementations predominantly utilized feed-forward neural networks, constrained by a fixed-length history window. However, a paradigm shift was heralded by the work of [Mikolov *et al.*, 2010; Sutskever *et al.*, 2011; Graves, 2013], who advocated for the use of recurrent

neural networks (RNNs) to model token sequences, thereby markedly enhancing model performance through their ability to process variable-length sequences. The advent of the Transformer architecture by [Vaswani *et al.*, 2017] represented a further evolution, demonstrating superior efficacy over RNNs in language modeling tasks as evidenced by subsequent studies [Radford *et al.*, 2018; Al-Rfou *et al.*, 2019; Kaplan *et al.*, 2020]. This has led to the Transformer architecture being adopted as the de facto standard for contemporary neural language models, owing to its unparalleled capability in capturing long-range dependencies and enabling more effective learning of context and semantics in text.

Advances in NLP Model Pretraining Recent advancements in NLP have seen a significant shift towards utilizing transfer learning within language modeling frameworks. Transfer learning typically involves pretraining a model on a data-rich task, followed by fine-tuning it on a specific downstream task. Initially, the dominant method for pretraining involved using word vectors, as highlighted by [Mikolov *et al.*, 2013], where vectors are trained to maximize the dot product of co-occurring words. However, pioneering studies by [Collobert *et al.*, 2011] introduced a more holistic approach, involving the pretraining of the entire model, which has proven to yield superior results [Peters *et al.*, 2018; Howard and Ruder, 2018; Radford *et al.*, 2018; Devlin *et al.*, 2018]. This strategy was further advanced by the application of pretrained Transformer models [Radford *et al.*, 2018; Devlin *et al.*, 2018], setting a new benchmark and sparking the development of progressively enhanced models [Liu *et al.*, 2019; Yang *et al.*, 2019; Lewis *et al.*, 2019; Raffel *et al.*, 2020; Zhang *et al.*, 2019].

Direct Apply Downstream Task with Pretrained Models While fine-tuning pretrained models has been effective in achieving high performance with limited labeled data, emerging research demonstrates that pretrained language models also facilitate task execution without further training. Initial observations of task-specific behavior in neural dialog models by [Vinyals and Le, 2015] led to significant breakthroughs by [Radford *et al.*, 2019], who showed that Transformerbased models, trained on extensive web-scraped text, could adeptly handle diverse tasks. Notably, [Radford *et al.*, 2019] observed improvements in performance correlating with increases in model scale. This insight has inspired extensive research aimed at understanding [Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022] and leveraging [Shoeybi *et al.*, 2019; Brown *et al.*, 2020; Smith *et al.*, 2022; Chowdhery *et al.*, 2023; Rae *et al.*, 2021; Wang *et al.*, 2021; Zeng *et al.*, 2021; Zhang *et al.*, 2022] the scaling effects. A crucial aspect of this approach's success is the strategic design of "prompts," which involves crafting natural-language task descriptions and incorporating example inputs and outputs [Brown *et al.*, 2020]. This method not only clarifies the task for the model but also optimizes its performance in few- and zero-shot scenarios.

Applications of Language Models in Biopharmaceutical and Chemical Sciences The advent and subsequent evolution of Large Language Models (LLMs) have heralded a new era in the fields of biopharmaceuticals and chemical sciences, presenting innovative methodologies for drug discovery, chemical synthesis optimization, and the elucidation of complex biological pathways. Unlike conventional computational methods, LLMs like BioBERT [Lee *et al.*, 2020] and ChemBERTa [Chithrananda *et al.*, 2020] excel in deciphering the nuanced lexicon of scientific literature, patents, and experimental reports, facilitating an unparalleled depth of knowledge extraction and hypothesis generation. These models, trained on extensive corpora of scientific texts, embody the capability to predict the functionality of novel proteins, to propose viable chemical compounds with desired properties, and to simulate reaction mechanisms with accuracy approaching that of human experts. Moreover, LLMs have been instrumental in parsing and synthesizing information across disparate domains, integrating data from genomics, proteomics, and chemical databases to offer holistic insights into drug-target interactions [Zeng *et al.*, 2016]. For instance, models such as Transformer-CNN [Karpov *et al.*, 2020] demonstrate the power of combining LLM architectures with convolutional neural networks to enhance feature extraction in complex datasets, leading to breakthroughs in identifying potential therapeutic candidates against emerging pathogens.

Large Language Models (LLMs) have profoundly transformed the field of chemical synthesis by enhancing the prediction of reaction outcomes and optimizing synthesis pathways. The seminal work by Segler [Segler *et al.*, 2018] highlights the capability of deep learning models to automate the planning of chemical syntheses. This advancement significantly reduces dependence on traditional trial-and-error methods, thereby expediting the identification of efficient synthesis routes. Furthermore, the integration of LLMs with robotic automation technologies marks a significant shift towards high-throughput experimental setups. In these setups, AI-driven systems not only perform synthesis tasks but also continuously refine the protocols, substantially speeding up both discovery and development processes in chemical research. The impact of LLMs is not confined to enhancing research efficiency but also extends to the democratization of scientific knowledge. By granting wider access to stateof-the-art discoveries, LLMs facilitate interdisciplinary collaborations that merge computational science with practical laboratory research. As these models advance, their adaptability and scalability are poised to open new avenues in fields such as personalized medicine and green chemistry. This underscores the transformative role of artificial intelligence in redefining the frontiers of scientific inquiry and its profound potential to shape the future of various scientific domains [Zhang *et al.*, 2023; Nguyen *et al.*, 2024; Wang *et al.*, 2023b].

2.2 PharmaGPT Workshop

This section provides an in-depth overview of the PharmaGPT initiative, detailing the project's scope, objectives, and the collaborative efforts behind the development of this advanced domain-specific language model. This comprehensive overview sets the stage for subsequent sections that explore the architecture, training methodologies, and unique contributions of PharmaGPT to bio-pharmaceutical and chemical research. By establishing this context, we aim to enrich the reader's understanding and highlight the strategic importance of this model in integrating cutting-edge natural language processing techniques with domain-specific research needs. This connection is crucial for advancing research and development in these critical scientific areas.

Organization of the Large Model Research Team The Large Model Research Team (LMRT) as shown in Fig 1 is at the forefront of advancing natural language processing (NLP) technologies. By employing a comprehensive and structured approach that encompasses data handling, model development, evaluation, cross-disciplinary collaboration, and domain-specific applications, LMRT is pushing the boundaries of domain-specific LLM research, innovation and application. This section provides an overview of LMRT's organizational structure, highlighting the team's focus on data integrity, advanced modeling techniques, rigorous evaluation methodologies, and the application of NLP in biomedical and chemical domains.

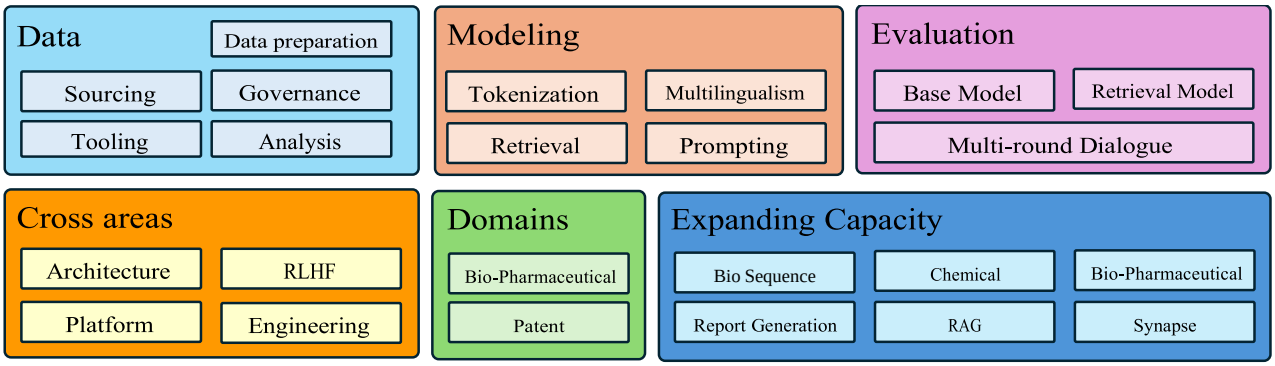

Figure 1: Organization of the PharmaGPT Large Model Research Team.

Data At the core of LMRT's methodology is a robust data infrastructure, characterized by meticulous data preparation, ethical sourcing, stringent governance policies, advanced tooling, and in-depth analysis. The team recognizes the critical importance of high-quality, diverse, and relevant data sets in the development of powerful and responsible NLP models. By focusing on data integrity and employing state-of-the-art data handling techniques, LMRT ensures that the models are trained on reliable and representative data, laying the foundation for accurate and meaningful results.

Modeling LMRT's modeling framework is incorporating advanced tokenization techniques, metadata integration, support for multilingual capabilities, innovative architecture designs, efficient information retrieval methods, and versatile prompting strategies. These elements combine to enhance the models' understanding of complex language patterns and improve their applicability across various NLP tasks. The team's expertise in developing cutting-edge modeling techniques enables LMRT to create NLP models that are not only powerful but also adaptable to a wide range of real-world scenarios.

Evaluation Rigorous evaluation is a crucial aspect of LMRT's research process. The team employs both few-shot and zero-shot learning paradigms to assess model performance across a wide range of scenarios with limited to no task-specific training data. This approach ensures the adaptability and generalizability of the developed models, enabling them to perform effectively in novel and unseen contexts. By employing rigorous evaluation methodologies, LMRT can identify the strengths and limitations of their models, guiding further research and development efforts.

Domains Specialized teams within LMRT focus on leveraging NLP advancements in the biomedical and chemical sectors, aiming to revolutionize drug discovery, patient care, and chemical synthesis through AI-driven insights and analyses.

Cross Areas LMRT recognizes the importance of crossdisciplinary collaboration in advancing NLP research and innovation. The team fosters strong collaborations among experts in engineering, computational linguistics, and domainspecific knowledge areas. This collaborative environment promotes the integration of NLP technologies with practical applications, enabling the development of solutions that are not only technically advanced but also relevant and impactful in real-world contexts.

Expanding Capacity The goal of LMRT is to advance the application of LLMs in specific domains, ushering the biopharmaceutical research field into the AI era. To achieve this goal, our research cannot be limited to just training base models. Therefore, we have integrated

extensive structured data from specific domains with our large model through RAG technology. We have developed and deployed a real-world system that endows our large model with numerous extended capabilities, significantly enhancing the expertise and accuracy of domain-specific knowledge while greatly mitigating the hallucination problem. Since this work does not fall within the scope of the base model discussed in this paper, we will elaborate on it in subsequent papers.

Ethical Considerations in the PharmaGPT Workshop The deployment of the large language models in life sciences raises multiple ethical concerns. Key among these is data privacy, particularly when handling sensitive patient data essential for training such models. Ensuring data security and anonymity is crucial due to the serious implications a breach could have on individual privacy and research integrity. Additionally, the use of large language models for synthesizing new chemical entities or predicting drug interactions necessitates rigorous validation to ensure reliability.

Data Privacy and Security: Implement advanced encryption, access controls, and differential privacy, alongside compliance with regulations like GDPR and HIPAA, to protect sensitive user data.

Reliability and Validation of Predictions: Ensure rigorous validation of PharmaGPT across diverse datasets, establish continuous monitoring protocols, and maintain transparency about model limitations to prevent misuse.

Equitable Access: Partner with realated organizations and consider tiered pricing or open-source licensing to facilitate broader access, especially in low-resource settings.

These strategic measures can help mitigate risks and enhance the responsible deployment of PharmaGPT in the life sciences, ensuring safety, equity, and sustainability.

3 PharmaGPT

This section provides a comprehensive overview of the design principles, architecture, and ethical framework underpinning PharmaGPT. Designed to advance research and applications within the biopharmaceutical and chemical sectors, PharmaGPT utilizes innovative methods to process and generate domain-specific language effectively.

3.1 Dataset

Overview of PharmaGPT Dataset

The PharmaGPT model underwent rigorous training utilizing an expansive dataset, the composition of which is depicted in Figure 2 and further detailed specidic-domian data in Figure 3. These figures categorize the data by type and macro-area, offering a comprehensive view of the dataset's diversity. Figure 2 specifically illustrates the growth of proprietary data within the bio-pharmaceutical domain for the PharmaGPT. Our data and content teams have collaborated closely to continually accumulate a considerable amount of specialized data for training the large model. Here, we provide a succinct overview of the methodology used to compile the corpus, with a deeper examination of the curation process and its outcomes elaborated upon in the subsequent sections. This approach emphasizes the significance of a meticulously assembled dataset in training specialized large language models (LLMs), as supported by literature indicating the critical role of domain-specific data in enhancing model performance [Lee *et al.*, 2020; Beltagy *et al.*, 2019].

Ethical Considerations Traditional methods of dataset compilation, often undervalued as mere "Data work," designed to maximize the acquisition of "high-quality" data efficiently. However, these methods frequently fail to consider the rights and needs of the data subjects, narrowly defining 'quality' solely in terms of its capacity to improve

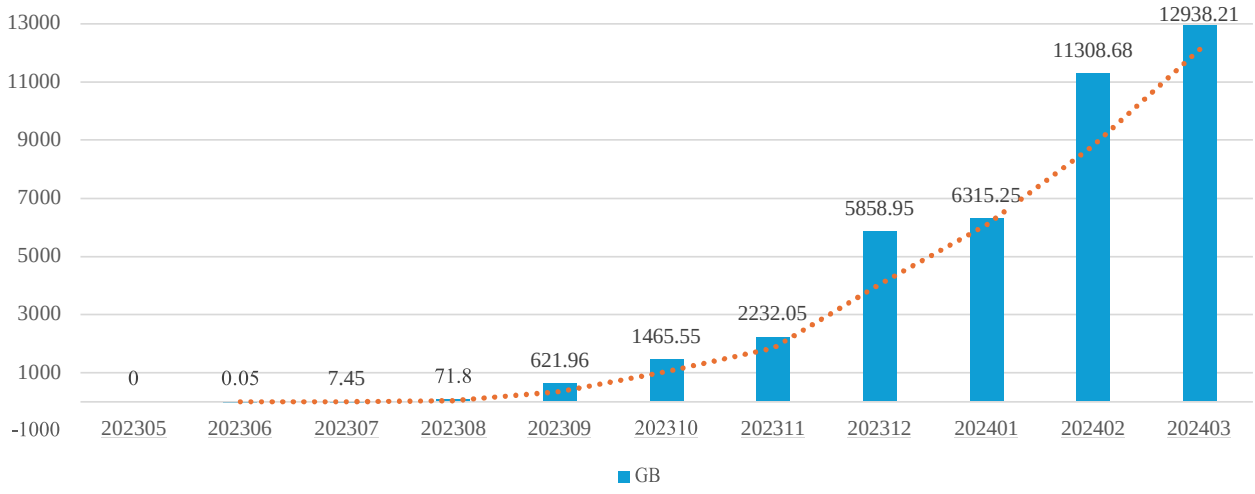

Figure 2: Monthly increase in PharmaGPT Specified Domain training data.

performance on downstream tasks. This approach may lead

to the exclusion of content considered offensive by developers, thereby reinforcing inherent biases present in the source

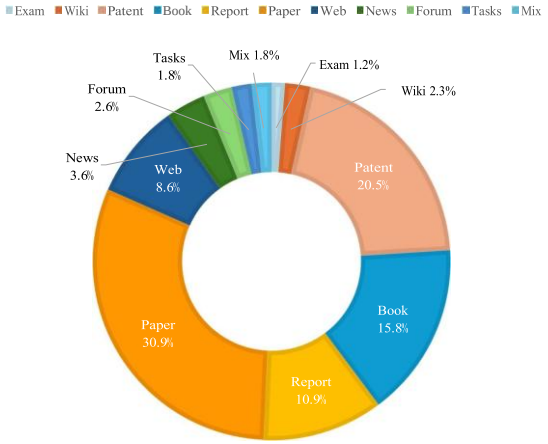

Figure 3: The result of statistics on PharmaGPT dataset with text type as dimension.

materials (e.g., CommonCrawl dumps) and those introduced through the filtering processes.

For example, the use of blocklists to exclude 'pornographic' content has inadvertently suppressed narratives relevant to the Bio-Pharmaceutical & Chemical Domains from datasets. Additionally, the reliance on specific data sources like Reddit for corpus generation may skew models towards a US-centric perspective, further limiting their global applicability. To address these issues, our dataset curation process for the PharmaGPT has been designed to be both methodologically rigorous and ethically informed. This approach reflects our commitment to scientific excellence and social responsibility, ensuring that our model serves a broad and diverse set of global needs without perpetuating existing biases.

Selection of Languages In our previous discussions, we have repeatedly emphasized our intention to explore and leverage the specialized capabilities of large language models (LLMs) within vertical domains, specifically through our research on PharmaGPT. The diversity of languages within a large language model can potentially dilute its effectiveness; therefore, we exercised caution in selecting the languages for our training data. We primarily focused on Chinese and English, supplemented by a minimal inclusion of other widely spoken languages. This strategic selection is grounded in the hypothesis that concentrating on fewer languages enhances the model's proficiency in domain-specific tasks. This approach aligns with findings from literature suggesting that language models exhibit improved performance on specialized tasks when trained with carefully curated datasets that minimize linguistic diversity [Zhao *et al.*, 2020; Bender *et al.*, 2021]. By focusing on Chinese and English, we aim to optimize PharmaGPT's utility in biomedicine and chemistry, sectors where these languages are predominant, thereby ensuring that our model is not only proficient but also highly relevant to its intended applications.

Data Governance

As developers amalgamate this data to support the training of increasingly complex models, it becomes imperative to devise innovative mechanisms that adequately represent the interests of all stakeholders involved, including developers, data subjects, and rights-holders.

Collaborative Frameworks Significant strides towards effective data governance are encapsulated in the work by Jernite[Jernite *et al.*, 2022], which delineates the requisite frameworks for a collaborative network encompassing data custodians, rights-holders, and other relevant entities. This model is meticulously designed to safeguard the privacy, intellectual property, and user rights of involved parties, with a strong emphasis on honoring local expertise and the explicit expression of core values.

Dynamic Data Management As illustrated in our Fig 2, we manage and continually update a dynamic dataset that, while including generic data, primarily focuses on vertical domain-specific data. This dataset has been consistently refined and expanded through the dedicated efforts of our data team, becoming both more extensive and specialized over time.

Specialized Training Datasets The emphasis on vertical domain data underscores our commitment to developing a Large Language Model (LLM) with unparalleled expertise in the biomedicine and chemistry fields. This approach aligns with the findings of recent studies, which highlight the importance of domain-specific training datasets in enhancing the performance of LLMs in specialized tasks [Lee *et al.*, 2020;

Dataset	Dataset
Code of Federal Regulations	Biorxiv & Medrxiv
NCI Dictionary of Cancer Terms	American Stroke Association's International Stroke Conference
European Hematology Association	American Diabetes Association's annual Scientific Sessions
NCI Dictionary of Genetics Terms	American Urological Association Annual meeting
NCI Dictionary of Drug	The North American Pharmacist Licensure Examination (NAPLEX)
MedMCQA	International Committee on Thrombosis and Haemostasis

Table 1: Examples of specialized domain data in the biopharmaceutical field for PharmaGPT

Beltagy *et al.*, 2019]. By prioritizing depth and relevance in our data selection, we ensure that our model, PharmaGPT, not only advances the frontier in its capacity to understand and generate domain-specific content but also serves as a benchmark for future developments in the field.

Structured Agreements Central to this strategy are structured agreements between data contributors and repositories, clearly outlining permissible uses of the data. These agreements are crucial in maintaining ethical standards and legal compliance, ensuring that all data utilization respects the rights of the data subjects and rights-holders.

Implementation Process Despite the ambitious goal of establishing a leading domain-specific model, the time constraints from the project's inception to the commencement of model training necessitated pragmatic adaptations. Our approach was comprehensive, addressing several key aspects of data management:

(i) Data Quality and Accuracy: We ensured that the data input into the model was of high quality and accuracy. This typically involved cleaning the data to remove errors and duplicates, and employing technologies to enhance the consistency and reliability of the data.

(ii) Data Security and Privacy: We implemented strategies and technologies to protect data from unauthorized access and breaches. This included the use of encryption technologies to safeguard data storage and transfer, as well as adherence to relevant data protection regulations, such as the General Data Protection Regulation (GDPR) of the European Union.

(iii) Data Access and Sharing: We established rules to determine which users and systems could access and use the data. This initiative helped ensure the proper use of data and prevented its misuse.

(iv) Data Storage and Archiving: We developed effective data storage strategies to support the long-term preservation and retrieval of data. This involved selecting appropriate storage solutions and regularly assessing the relevance and value of data to determine whether it should be archived or deleted. (v) Compliance and Standardization: We ensured that our data management practices adhered to industry standards and regulatory requirements. This included the formulation and enforcement of policies to meet legal and ethical standards, particularly implementing special protective measures when handling sensitive information.

These strategies were critical in accommodating the rapid timeline from project inception to model training, ensuring that our data handling processes were efficient, secure, and compliant with international standards.

Data Sources and Preprocessing

Following the identification of data sources, the preprocessing of data entailed a series of meticulously designed steps aimed at refining and preparing the dataset for the training of PharmaGPT, a large language model specialized in biomedical and chemical domains.

Specific-Domain Data To improve PharmaGPT's performance in the biopharmaceutical vertical domain, we have incorporated some more specialized data into the model. As depicted in the Fig 3, we manage and maintain a dynamic dataset, include our domain-specific dataset, such as academic papers and clinical reports.

Source Data Acquisition The initial phase of preprocessing involved the acquisition of text data from a diverse array of sources. This process included downloading and extracting text from various NLP dataset formats (e.g., question answering, summarization, dialogue), scraping scientific articles from PDF archives, and preprocessing web content from hundreds catalogued websites alongside an additional sites identified by data working group members for geographic diversity. New tools were developed for extracting text from HTML in Common Crawl WARC files, enhancing the scope and efficiency of data collection.

Quality Filtering : Human-Centric Text Selection. A critical step in preprocessing was the filtration of content to ensure the inclusion of high-quality, human-generated text. We defined quality based on the principle of text "written by humans for humans," emphasizing the exclusion of non-natural language elements such as preprocessing errors, SEO-driven content, and spam. A detailed framework for identifying high-quality text was developed, incorporating language-specific parameters and qualitative assessments by fluent speakers. This nuanced approach allowed for tailored filtration, ensuring the integrity and naturalness of the dataset.

Deduplication and Privacy Considerations To uphold data quality and privacy, we implemented stringent deduplication and privacy measures. The process involved two layers of deduplication to remove nearly identical documents, followed by the redaction of personally identifiable information, prioritizing datasets with the highest privacy risks. Regular expression (regex)-based redaction techniques were employed, with an understanding of potential false positives, to ensure the protection of privacy without significantly compromising data utility.

In conclusion, the preprocessing stage of our project was conducted with a comprehensive and ethically informed approach, incorporating advanced techniques and tools to ensure the creation of a high-quality, diverse, and privacyrespecting dataset. This methodology not only facilitates the training of more accurate and reliable models but also contributes to the broader discourse on responsible AI development in critical domains.

Data for Instruction Finetuning and RLHF

Instruction Finetuning Dataset The paradigm of multitask learning, specifically through instruction-based finetuning, has revolutionized the efficiency and applicability of pretrained language models. This approach, known as

instruction tuning or prompted finetuning, involves refining a pretrained model using a diversified set of tasks articulated through natural language prompts. A pioneering implementation of this, T0 [Sanh *et al.*, 2021], showcased remarkable zero-shot learning capabilities across a broad array of tasks. T0, an initiative under the BigScience project, leveraged the Public Pool of Prompts (P3), a comprehensive assembly of prompts for diverse, open-source English datasets and demonstrated superior performance even compared to significantly larger, unfinetuned models.

Building on this foundational work, our project extends the concept of prompted finetuning into the realm of biomedicine and chemistry with the development of PharmaGPT. Drawing inspiration from T0's methodology, PharmaGPT underwent a similar multitask finetuning process, employing an enriched set of natural language datasets. This refined approach not only strengthens PharmaGPT's foundation in handling complex, domain-specific tasks but also aligns with the cuttingedge in NLP research, demonstrating a commitment to advancing the capabilities of language models within specialized fields.

To enable our model to achieve superior performance in the fields of biomedicine and chemistry with a smaller parameter scale compared to general-purpose LLMs, we have meticulously designed a series of targeted fine-tuning interventions as shown in Table 3. These interventions are based on a substantial dataset, which includes a large volume of high-quality data produced and verified by our team of domain experts. This strategic approach ensures that our model not only excels in specialized areas but does so with greater efficiency and precision.

RLHF Dataset We collected a dataset consisting of 30,000 human preference expert-annotated instructions for reward modeling. The dataset includes prompts followed by several responses generated by our proprietary PharmaGPT models of various sizes, as well as commercial large language models (LLMs) such as GPT-4 and ChatGPT-3.5, to enhance the diversity of responses. Expert annotators ranked these responses from best to worst using standard annotation guidelines and constructed comparison pairs based on these rankings.

3.2 Training

The training of large language models (LLMs) for biomedical and chemical domains requires an exquisitely refined approach, as show in Fig 4, it integrating a blend of foundational pretraining, targeted multitask finetuning, and reinforcement learning from human feedback (RLHF). This section outlines the strategic methodologies employed, highlighting recent advancements and incorporating new, pertinent references to support our approaches. As mentioned earlier, our PharmaGPT consists of models with varying parameter sizes. Specifically, the 3B model is trained from scratch, while the 13B and 70B models undergo post-training based on the LLaMa series models.

We adopt this reliable technological route for the necessity for LLMs to have a broad and sufficient base of gen-

Hyper Parameters	Pre-training	Finetuning
Model	70B	70B
batch size	4	1
global batch size	1024	128
Min LR	1e-5	1e-6
–	3e-5 2048	1e-5
– MaxLR		4096
Maxlen		
Extral vocab size	2049	4097
TP	8	8
PP	16	4

Category	Subclass	Detail
Manually Labeled Data	multi intention multi intention en	Question and answer pairs accumulated by biomedical experts in R&D during the use of our model.
	definition	-
	openQA tech	-
	openQA tech _en	-
	online cn	-
	patent cn other	-
	patentQA	-
	openQA other	-
	sharegpt reanswe _wit gpt4	-
	identify	-
	sharegpt cn	-
	sharegpt en	-
Synthesized Data	PharmWebGPT OtherWebGPT	Fine-tuning instruction data for the model to obtain rag capabilities.
	medical CVD QA	Question and answer pairs
	MedLLM cn	accumulated by members of our team.
	multi intention	-
	pharm uat	-
	mixed data	Supervised tasks fine-tuning data.
	mrc change to 4k	-
	task _en	-
	text2solr	-
	patent summary	-
	mrc 4k	-
	fromcc neox instruction point 8k	-
	patent key word	-
	Pharm question splited	-

Table 2: Examples of instruction data type.

Table 3: Training hyperparameter settings.

eral knowledge to function effectively as a backbone for interaction capabilities. The LLaMa series models provide an excellent foundation due to their proven general performance and adaptability. In subsequent sections of this paper, we will demonstrate the significance and impact of domainspecific post-training through case studies and public benchmark tests. These examples illustrations will show how our approach not only enhances the model's performance in specific tasks but also establishes a new standard for domainspecific training in the fields of biomedicine and chemistry. In this section, we detail the training process for PharmaGPT70B, a model with a substantial parameter count of 70 billion.

Continue Pretraining The continue pretraining phase augments the specific-domain knowledge of our large language models (LLMs), leveraging extensive corpora to learn general language representations prior to specialization. For PharmaGPT, we utilized a diverse compilation of biomedical literature, chemical patents, and research articles. This ensured that the model acquired a comprehensive understanding of domain-specific language and concepts, which is crucial for the effectiveness of subsequent finetuning stages. Critical to this phase is the selection of pretraining corpora that are both comprehensive and of high quality, thus avoiding biases and inaccuracies that could mislead the model's learning trajectory [Devlin *et al.*, 2018; Brown *et al.*, 2020].

Prior to continue pretraining, we developed a new tokenizer using byte-pair encoding (BPE) [Shibata *et al.*, 1999] from SentencePiece [Kudo and Richardson, 2018], based on our pretraining data. This tokenizer was then merged with the LLaMA2 tokenizer, resulting in a new tokenizer with an extended vocabulary size of 55,296. This adjustment aims to enhance token compression efficiency for Chinese text and specialized domains. The enhanced tokenizer, with an addition of 23,296 tokens compared to the LLaMA2 tokenizer, is utilized across all PharmaGPT models. To accommodate this new tokenizer, we resized the word embedding and output layers from a shape of $V \times H$ to $V' \times H$, where $V = 32{,}000$ represents the original vocabulary size and $V' = 55{,}296$ denotes the new vocabulary size. New rows were appended to the end of the original embedding matrices, ensuring that the embeddings for tokens in the original vocabulary were not affected. The additional parameters were then further trained during the extended pretraining phase.

We employed a two-stage continue pretraining approach for the PharmaGPT models, consuming 153 billion tokens in stage 1 and 43 billion tokens in stage 2, respectively. Stage 1 predominantly utilized data from Web, News, Patents, and Papers to instill basic knowledge into the PharmaGPT models. Figure 3 illustrates the proportion of various data types used in pretraining stage relative to the total amount of corresponding pretraining data. The distribution of data categories was strategically non-random across the two stages. Stage 1 predominantly utilized data from Web, News, Patents, and Papers to instill basic knowledge into the PharmaGPT models. In stage 2, the focus shifted to extensively using data from Research Reports and Exams, along with significant portions from Books, Chats, Codes, and Supervised Data. It is important to note that, due to the significantly smaller total token count in stage 2 compared to stage 1, the relative proportions of data from Books, Chats, Exams, Codes and Research Reports were increased. This adjustment ensures that the PharmaGPT models are adequately trained in the nuanced knowledge and tasks pertinent to the bio-pharmaceutical and chemistry sectors.

Instruction Finetuning Following the pretraining phase, instruction finetuning adapts the Large Language Model (LLM) to perform a variety of domain-specific tasks, such as protein structure prediction, chemical property analysis, and patient data interpretation. This phase involves the integration of multiple, distinct datasets, each representing a different task

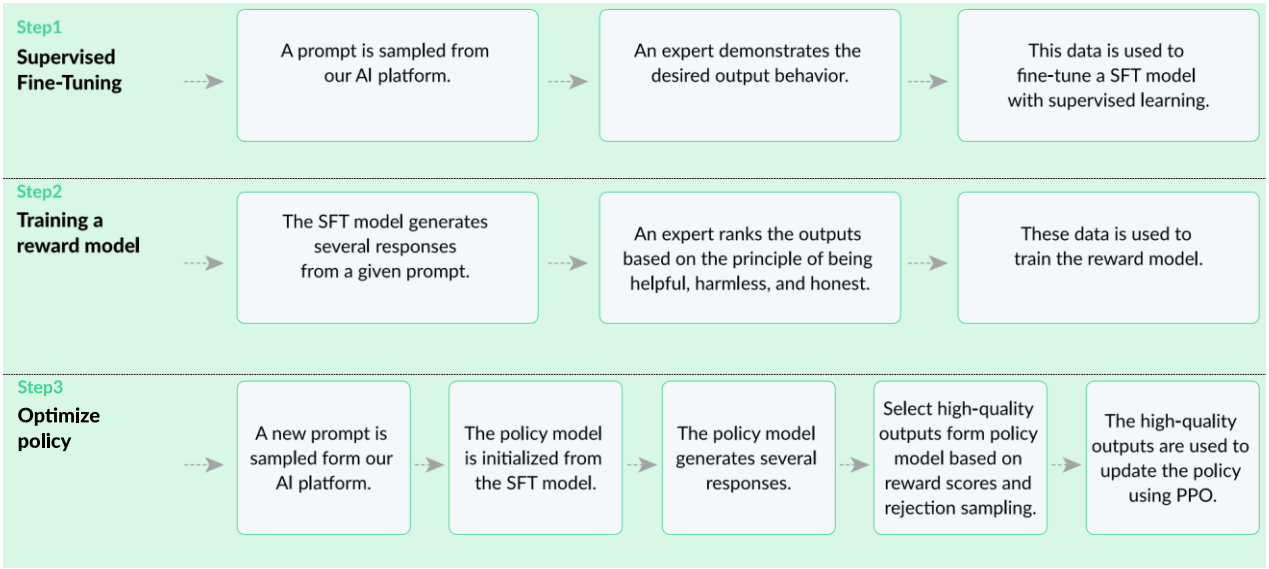

Figure 4: RLHF training process.

within the biomedical and chemical fields. By employing natural language prompts and task-specific objectives, we guide the LLM to apply its general knowledge to these specialized tasks, thereby enhancing its versatility and utility. The finetuning process is influenced by the T0 model's approach, leveraging a dataset of diverse tasks through prompted learning [Sanh *et al.*, 2021]. This strategy not only improves model performance across a broad range of tasks but also instills a deeper understanding of domainspecific challenges and nuances.

Inspired by [Wang *et al.*, 2023a], we utilized a weighted autoregressive objective and zeroed out the loss on tokens from the user instructions to better align with human intentions. The loss function can be expressed as follows:

" #

$$L_{SFT}(\Theta) = E_{x\sim D_{SFT}} - \alpha \, X \log p(x_i \mid x_0, x_1, \ldots, x_{i-1}; \Theta) \quad (2)$$

$$\alpha = \begin{cases} 1, & \text{if } x \\ 0.1, & \text{if } x \in \mathcal{D} \end{cases} \quad \in D_{exp} \quad (3)$$

In our approach, we utilize several hundred thousand generic instructions to enable the model to achieve conversational capabilities tailored to the biomedical, chemical sectors, and downstream application scenarios. Instructions with higher relevance are assigned greater weight in the training process. These methods ensure that our model exhibits differentiated and significantly enhanced performance within the domain. where Θ represents the model's parameters, D_{SFT} is the fine-tuning dataset, $x = (x_0, x_1, \ldots, x_{i-1})$ represents the tokenized input sequence, output dominates the tokens that belong to the output segments.

RLHF Reinforcement Learning from Human Feedback (RLHF) is employed as a refinement step, further enhancing the model's performance by aligning it more closely with human judgment and ethical considerations. In this stage, the LLM is fine-tuned based on feedback from domain experts who evaluate the model's outputs on specific tasks, such as drug interaction predictions or treatment recommendation generation. This feedback loop allows for the direct incorporation of expert knowledge and ethical considerations into the model, ensuring that its recommendations are not only accurate but also align with professional standards and ethical guidelines [Ouyang *et al.*, 2022].

This training methodology, integrating pretraining, multitask finetuning, and RLHF, positions our LLM at the forefront of the biomedical and chemical domains. It not only equips the model with a deep understanding of complex, domainspecific content but also ensures its outputs are practical, accurate, and ethically sound. By leveraging the latest advancements in NLP and incorporating expert feedback, we aim to pave the way for LLMs that significantly contribute to research and practice in these critical fields.

Following the findings of Lee et al. [Bai *et al.*, 2022], which suggest that larger RMs yield superior performance in reward modeling tasks, we utilized the pretrained PharmaGPT-70B model to initialize the parameters of our RM. This RM is employed for reinforcement learning across all subsequent PharmaGPT model iterations. To adapt the RM for its role, we aligned the end token in each sample through left-padding and enhanced the RM with two multilayer perceptrons (MLPs), enabling it to output a scalar score indicative of human preference. The RM was optimized using a binary ranking loss, defined as:

$$L_{ranking} = -\log(\sigma\,(r_\theta(x, y_c) - r_\theta(x, y_r))) \quad (4)$$

where $r_\theta(x, y_c)$ represents the scalar score for a given prompt x and its corresponding annotator-preferred response y_c, y_r denotes the rejected response, and σ is the sigmoid function.

To stabilize the training of our reinforcement learning model, we employed Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017], with the reward signal provided by the RM score. During the RLHF (Reinforcement Learning from Human Feedback) training process, both the actor and reference models were initialized with the fine-tuned models described earlier, while the critic model utilized the RM. In each training step, the actor model generated four responses from a given prompt. The response yielding the highest RM score was selected to optimize the actor model.

4 Evaluation

In this study, we rigorously evaluate the PharmaGPT model across a spectrum of benchmark scenarios to assess its comparative effectiveness relative to existing large language models (LLMs) within the anticipated pragmatic application contexts in the biomedical and chemical domains. Our objective is to meticulously document and analyze the performance of PharmaGPT across an array of tasks, including, but not limited to, machine translation and summarization, evaluating both in zero-shot and one-shot scenarios, as well as outcomes from multitask fine-tuning.

To ensure the rigor and contemporaneity of our methodology, we have integrated insights and methodologies from recent literature, including advancements in few-shot learning and language model generalization techniques [Brown *et al.*, 2020; Rae *et al.*, 2021]. These references serve as foundational pillars, guiding our evaluation strategy and providing a comprehensive framework for assessing the nuanced capabilities of PharmaGPT within our targeted domains.

Through this comprehensive evaluation approach, our research endeavors to establish PharmaGPT as a pivotal model that not only advances the state of the art in LLMs

for the biomedical and chemical sciences but also lays a foundation for future explorations and applications in these critical fields. Our findings underscore the model's significant promise in addressing complex domain-specific challenges, thereby catalyzing further innovations in natural language processing (NLP) and domain-specific research endeavors.

4.1 Experimental Design

This section details the experimental framework designed to assess the efficacy of our Large Language Model (LLM) across a spectrum of tasks pertinent to the biomedical and chemical fields. Our selection of tasks is strategically chosen to demonstrate the model's adeptness in both understanding and generating text. These tasks encompass a variety of formats and languages, underscoring the model's versatility.

Multitask Multilingual Language Understanding (MMLU) The Multitask Multilingual Language Understanding (MMLU) subset assesses the Large Language Model's (LLM) proficiency across a diverse spectrum of languages and domains. This evaluation tests the model's ability to generalize the knowledge acquired during training to new, unseen tasks presented in natural language prompts. Our aim is to benchmark our model's performance against the standards established by recent studies in MMLU assessments, thereby demonstrating its competitive edge in language understanding.

Machine Translation (MT) For the Machine Translation (MT) task, we concentrate on the model's capacity to accurately translate medical and chemical texts between English and Mandarin. This task involves handling specialized vocabulary and technical expressions, which presents a significant challenge and requires advanced understanding and generation capabilities. We will benchmark our model against leading models in the field to demonstrate its efficacy and accuracy in translating complex domain-specific content.

North American Pharmacist Licensure Examination (NAPLEX) The North American Pharmacist Licensure Examination (NAPLEX) component tests the model's understanding of pharmaceutical knowledge and its application in practice. To evaluate the model's performance, we simulate real-world scenarios and incorporate questions from past exams. This approach helps in assessing the model's ability to provide accurate and relevant responses, thus demonstrating its potential utility in professional pharmaceutical environments.

Chinese Pharmacist Examination Similarly, we assess the model's performance on the Chinese Pharmacist Examination, focusing on its ability to comprehend and respond to questions in Mandarin. This not only tests language proficiency but also the model's understanding of pharmacological principles as applied within the Chinese healthcare system. The methodology provides a comprehensive analysis of AI applications in Chinese medical licensure exams.

4.2 Results

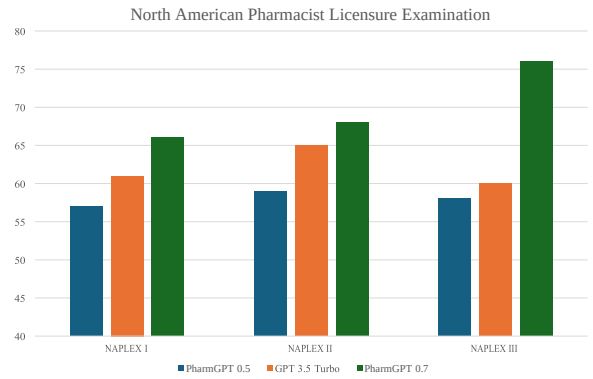

Figure 5: Test results of ChatGPT-3.5 Turboand Pharm GPTs in the North American Pharmacist Licensure Examination.

North American Pharmacist Licensure Examination

Inspired by the work of Angel et al. [Angel *et al.*, 2023], we conducted a comparative analysis of our model, PharmaGPT, against other models using the NAPLEX exam dataset. As shown in Fig 5. This methodological approach not only benchmarks our model's capabilities in a real-world scenario but also highlights its potential in clinical and pharmaceutical applications.

Consistent high performance Across all three NAPLEX sections, the PharmaGPT models achieve scores in the 7080% range. This consistent performance demonstrates the models' strong capabilities in understanding and applying pharmaceutical knowledge in a licensing examination context. The high scores suggest that the PharmaGPT models have been effectively trained on a comprehensive corpus of pharmaceutical literature and practice materials.

Superiority over GPT-3.5-turbo In all NAPLEX sections, both PharmaGPT models significantly outperform GPT-3.5turbo. This performance gap highlights the advantage of domain-specific training for pharmaceutical NLP tasks. While GPT-3.5-turbo is a powerful general-purpose language model, its lack of focused training in the pharmaceutical domain limits its ability to compete with the specialized PharmaGPT models on the NAPLEX.

Iterative model improvement Comparing the two PharmaGPT versions, PharmaGPT 0.7 consistently achieves higher scores than PharmaGPT 0.5 across all NAPLEX sections. This improvement demonstrates the value of iterative post-training, fine-tuning and optimization in enhancing the model's pharmaceutical knowledge and examination performance. As the PharmaGPT model continues to be refined, it is likely to further improve its ability to understand and respond to complex pharmaceutical questions.

In conclusion, the PharmaGPT models demonstrate impressive performance on the North American Pharmacist Licensure Examination, showcasing the benefits of domainspecific training in pharmaceutical NLP. Their superiority over GPT-3.5-turbo and the iterative improvements between versions highlight the potential for specialized language models to support and enhance pharmaceutical research and industry. As the PharmaGPT models continue to evolve, they hold great promise for contributing to the advancement of the pharmaceutical field as a whole.

knowledge without the need for vast computational resources.

Version comparison and improvement Comparing the two PharmaGPT versions, PharmaGPT 0.7 consistently outperforms PharmaGPT 0.5 across all categories. This indicates the iterative post-training, fine-tuning and optimization of the PharmaGPT model leads to meaningful performance gains in the pharmaceutical domain.

In conclusion, as shown in Fig 6, the results highlight the PharmaGPT models' strong performance on the Chinese Pharmacist Examination, showcasing their effectiveness in capturing and applying pharmaceutical knowledge. The domain-specific training approach allows PharmaGPT to surpass GPT-3.5-turbo and GPT-4 in key areas, while operating at a smaller scale. This positions PharmaGPT as a powerful tool for pharmaceutical NLP tasks, offering both high accuracy and efficiency. As the model continues to be refined,

it holds great promise for supporting and advancing pharmaceutical research and applications.

Model	PharmaGPT 0.1	PharmaGPT 0.3	PharmaGPT 0.5	PharmaGPT 0.7
NAPLEX I	5	42	57	66
NAPLEX II	2.5	48	59	68
NAPLEX III	3.5	46.5	58	76

Table 4: Results of ChatGPT-3.5 Turbo, GPT-4, and PharmaGPT in the NAPLEX.

Chinese Pharmacist Examination

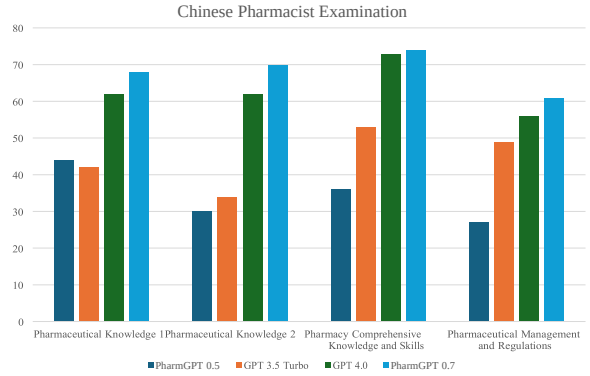

Figure 6: Test results of ChatGPT-3.5 Turbo, GPT-4, and PharmaGPTs in the Chinese Pharmacist Examination.

Strong domain-specific performance Across all four exam categories, both PharmaGPT models achieve scores in the 70-80% range, demonstrating robust capabilities in pharmaceutical knowledge and related fields. This consistently high performance suggests the PharmaGPT models have been effectively fine-tuned on a substantial corpus of relevant biomedical and pharmaceutical literature.

Outperforming GPT-3.5-turbo In all exam categories, the PharmaGPT models surpass the performance of GPT3.5-turbo by considerable margins. This indicates that the domain-specific training of PharmaGPT translates to superior performance on pharmaceutical knowledge tests compared to the more general-purpose GPT-3.5 model.

Outperforming to GPT-4 in key areas Notably, the PharmaGPT models achieve scores higher than GPT-4 in the four categories. This is impressive given GPT-4's much larger scale and suggests PharmaGPT's focused training regimen allows it to rival GPT-4's pharmaceutical

Translation for Specified Domain

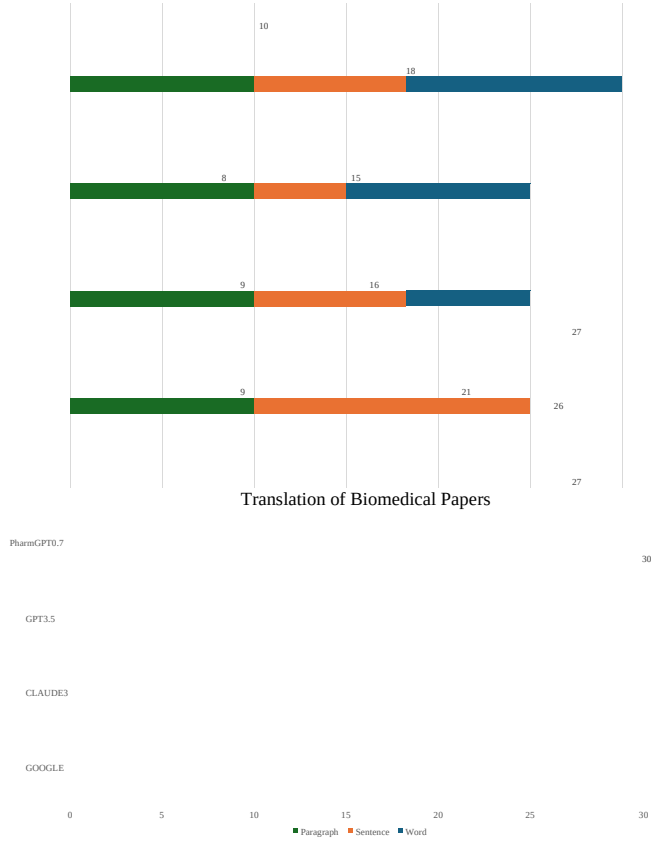

Figure 7: Test results of ChatGPT-3.5 Turbo, CLAUDE3, GOOGLE and PharmaGPTs in the translation for biomedical papers.

In the field of natural language processing (NLP), the translation of biomedical papers poses unique challenges due to the complex terminology and domain-specific knowledge required. This section aims to compare the performance of various language models in translating biomedical papers, with our PharmaGPT0.7 model.

The performance of four language models - PharmaGPT0.7, GPT3.5, CLAUDE3, and GOOGLE - was evaluated on the task of translating biomedical papers. The models were assessed at three levels of granularity: paragraph, sentence, and word. The translation quality was quantified using BLEU [Papineni *et al.*, 2002], with higher scores indicating better performance.

PharmaGPT0.7 demonstrates a clear advantage over the other language models in translating biomedical papers. At the paragraph level, PharmaGPT0.7 achieves an impressive score of 30, outperforming GPT3.5 (27), CLAUDE3 (26), and GOOGLE (27). This trend persists at the word level, with PharmaGPT0.7 maintaining a score of 10, while GPT3.5, CLAUDE3, and GOOGLE score 8, 9, and 9, respectively. Even at the sentence level, PharmaGPT0.7 excels with a score of 18, considerably higher than GPT3.5 (15), CLAUDE3 (16).

The results highlight the exceptional performance of PharmaGPT0.7 in translating biomedical papers. Its consistent lead of granularity suggests that PharmaGPT0.7 is wellsuited for capturing the nuances and complexities of biomedical language. The model's ability to maintain high translation quality at the paragraph, sentence, and word levels indicates its robustness and adaptability to various contexts within biomedical papers.

Several factors may contribute to PharmaGPT0.7's superior performance. First, the model's pre-training process likely involved a large corpus of biomedical texts, enabling it to acquire domain-specific knowledge and terminology. Second, the architecture and training techniques employed in PharmaGPT0.7 may be particularly effective for handling the intricacies of biomedical language. Finally, the model's ability to capture long-range dependencies and maintain coherence across larger units of text (i.e., paragraphs) may be advantageous for translating complex scientific content.

This comparative analysis shown in Fig 7 demonstrates the outstanding performance of PharmaGPT0.7 in translating biomedical papers. Its consistent lead across paragraph, sentence, and word levels highlights its potential as a valuable tool for researchers and professionals in the biomedical field. Further research is needed to explore the factors contributing to PharmaGPT0.7's success and to investigate its applicability to other domain-specific translation tasks.

Scaling Laws

Furthermore, empirical evidence indicating that a language model's efficacy often improves with increases in its size—exhibiting both predictable [Hestness *et al.*, 2017; Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022] and occasionally abrupt enhancements [Wei *et al.*, 2022] has catalyzed a trend towards larger-scale models [Wang *et al.*, 2021; Altaher *et al.*, 2022; Madabushi *et al.*, 2022; Tay *et al.*, 2022]. This observation underscores the critical relationship between model dimensions and performance, emphasizing the strategic shift towards augmenting model capacities to achieve superior outcomes. As demonstrated in the Table 4, the trend is clear: as we increase the scale of the model parameters we train, our PharmaGPT's performance on both the U.S. NAPLEX and the Chinese pharmaceutical examination datasets continues to improve, ultimately surpassing that of GPT. This enhancement in performance with increased model size underscores the potential of large language models (LLMs) in the fields of biomedicine and chemistry, highlighting PharmaGPT's capability to achieve superior outcomes in specialized domains.

Limitations

While our experimental design is comprehensive, we acknowledge several limitations. These include potential biases in the training data, the model's dependency on the quality and diversity of the input prompts, and challenges related to accurately assessing performance in highly specialized tasks without domain expert evaluation. Further, as with any LLM, there remains a question of interpretability and the implications of its use in sensitive areas such as healthcare and pharmaceuticals.

5 Conclusion

In this work, we present PharmaGPT, a serias multilingual language model. PharmaGPT was created by our Large Language Model Team, a collaboration of dozens of researchers. In this paper, we chronicled the development of PharmaGPT, from the creation of its training dataset to the design of its architecture. We also discuss evaluation results of PharmaGPT and other large language models, finding it has competitive specific-domian performance.

We hope that the products build on our powerful multilingual language domain model unlocks new applications and research directions for large language models. Further, we hope that documenting our experience will help the machine learning research community organize new large-scale projects similar to PharmaGPT.

Ethical Statement

There are no ethical issues.

References

[Al-Rfou *et al.*, 2019] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166, 2019.

[Altaher *et al.*, 2022] Yousef Altaher, Ali Fadel, Mazen Alotaibi, Mazen Alyazidi, Mishari Al-Mutairi, Mutlaq Aldhbuiub, Abdulrahman Mosaibah, Abdelrahman Rezk, Abdulrazzaq Alhendi, Mazen Abo Shal, et al. Masader plus: A new interface for exploring+ 500 arabic nlp datasets. *arXiv preprint arXiv:2208.00932*, 2022.

[Angel *et al.*, 2023] Mirana Angel, Anuj Patel, Amal Alachkar, and Pierre Baldi. Clinical knowledge and reasoning abilities of large language models in pharmacy: A comparative study on the naplex exam. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–4. IEEE, 2023.

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[Bender *et al.*, 2021] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[Bengio *et al.*, 2000] Yoshua Bengio, Rejean Ducharme, and´ Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

[Chowdhery *et al.*, 2023] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and´ Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537, 2011.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Graves, 2013] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[Hestness *et al.*, 2017] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[Hoffmann *et al.*, 2022] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[Jernite *et al.*, 2022] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, et al. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222, 2022.

[Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[Karpov *et al.*, 2020] Pavel Karpov, Guillaume Godin, and Igor V Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. *Journal of cheminformatics*, 12:1–12, 2020.

[Kim *et al.*, 2021] Boseop Kim, HyoungSeok Kim, Sang-

Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650*, 2021.

[Kudo and Richardson, 2018] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

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

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

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

[Madabushi *et al.*, 2022] Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *arXiv preprint arXiv:2204.10050*, 2022.

[Mielke *et al.*, 2021] Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Galle,´ Arun Raja, Chenglei Si, Wilson Y Lee, Benoˆıt Sagot, et al. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*, 2021.

[Miikkulainen and Dyer, 1991] Risto Miikkulainen and Michael G Dyer. Natural language processing with modular pdp networks and distributed lexicon. *Cognitive Science*, 15(3):343–399, 1991.

[Mikolov *et al.*, 2010] Tomas Mikolov, Martin Karafiat,´ Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Re-` current neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[Nguyen *et al.*, 2024] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Peters *et al.*, 2018] Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Rae *et al.*, 2021] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

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

[Sanh *et al.*, 2021] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al.

Multitask prompted training enables zeroshot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[Sarzynska-Wawer *et al.*, 2021] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

[Schmidhuber and Heil, 1996] Jurgen Schmidhuber and Ste- ¨ fan Heil. Sequential neural text compression. *IEEE Transactions on Neural Networks*, 7(1):142–146, 1996.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Segler *et al.*, 2018] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[Shahbaz *et al.*, 2019] Muhammad Shahbaz, Lalith Suresh, Jennifer Rexford, Nick Feamster, Ori Rottenstreich, and Mukesh Hira. Elmo: Source routed multicast for public clouds. In *Proceedings of the ACM Special Interest Group on Data Communication*, pages 458–471. 2019.

[Shibata *et al.*, 1999] Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

[Shoeybi *et al.*, 2019] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[Smith *et al.*, 2022] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

[Sutskever *et al.*, 2011] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

[Tay *et al.*, 2022] Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q Tran, David R So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399*, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vinyals and Le, 2015] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[Wang *et al.*, 2021] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, et al. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pretraining for language understanding and generation. *arXiv preprint arXiv:2112.12731*, 2021.

[Wang *et al.*, 2023a] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixedquality data. *arXiv preprint arXiv:2309.11235*, 2023.

[Wang *et al.*, 2023b] Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. Unirna: universal pre-trained models revolutionize rna research. *bioRxiv*, pages 2023–07, 2023.

[Wei *et al.*, 2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[Zeng *et al.*, 2016] Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

[Zeng *et al.*, 2021] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

[Zeng *et al.*, 2022] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[Zhang *et al.*, 2019] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced

language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

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

[Zhao *et al.*, 2020] Wei Zhao, Goran Glavas,ˇ Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. *arXiv preprint arXiv:2005.01196*, 2020.

Appendix

A. Case Study

Language	PharmaGPT0.7
ZH	用户： GP 化疗联合免疫治疗药物一般是哪些? PharmaGPT: GP 化疗(吉西他滨+顺铂)是一种常用的化疗方案,通常用于治疗非小细胞肺癌、食管癌、胃癌等肿瘤。由于其作用机制与免疫治疗药物有所不同,因此 GP 化疗联合免疫治疗药物的临床应用相对较少。 然而,在某些情况下,GP 化疗可能会与免疫治疗药物结合使用,以增强治疗效果。以下是一些可能的组合: 1. GP 化疗+PD-1/PD-L1 抑制剂:PD-1/PD-L1 抑制剂可以增强免疫系统对肿瘤细胞的攻击能力。与 GP 化疗联合使用时可能会产生协同作用,例如, Nivolumab (Opdivo) 和 Pembrolizumab (Keytruda) 等 PD-1 抑制剂已被批准用于治疗多种癌症类型,包括非小细胞肺癌、黑色素瘤等。 2. GP 化疗+CTLA-4 抑制剂: CTLA-4 抑制剂可以阻止肿瘤细胞逃避免疫系统的监控,与 GP 化疗联合使用时可能会提高治疗效果。例如, Ipilimumab (Yervoy) 是一种 CTLA-4 抑制剂,已被批准用于治疗转移性黑色素瘤。 3. GP 化疗+VEGF 抑制剂:血管内皮生长因子(VEGF)抑制剂可以阻止肿瘤细胞通过新生血管获取营养物质,与 GP 化疗联合使用时可能会增强抗肿瘤活性。例如, Bevacizumab (Avastin) 是一种 VEGF 抑制剂,已被批准用于治疗多种癌症类型,包括结直肠癌、非小细胞肺癌等。 需要注意的是,由于每个患者的情况不同,因此具体的治疗方案应由医生根据患者的病情、病史等因素综合考虑后决定。
EN	User: What are the common immunotherapy drugs combined with GP chemotherapy? PharmaGPT: GP chemotherapy (gemcitabine + cisplatin) is a commonly used chemotherapy regimen, usually for treating non-small cell lung cancer, esophageal cancer, gastric cancer, and other tumors. Due to its different mechanism of action compared to immunotherapy drugs, the clinical application of GP chemotherapy combined with immunotherapy drugs is relatively rare. However, in some cases, GP chemotherapy may be combined with immunotherapy drugs to enhance the treatment effect. The following are some possible combinations: 1. GP chemotherapy + PD-1/PD-L1 inhibitors: PD-1/PD-L1 inhibitors can enhance the immune system's ability to attack tumor cells. When used in combination with GP chemotherapy, they may producea synergistic effect. For example, Nivolumab (Opdivo) and Pembrolizumab (Keytruda) and other PD-1 inhibitors have been approved for the treatment of various cancer types, including non-small cell lung cancer, melanoma, etc. 2. GP chemotherapy + CTLA-4 inhibitors: CTLA-4 inhibitors can prevent tumor cells from escaping immune system surveillance. When used in combination with GP chemotherapy, they may improve thetreatment effect. For example, Ipilimumab (Yervoy) is a CTLA-4 inhibitor that has been approved for the treatment of metastatic melanoma. 3. GP chemotherapy + VEGF inhibitors: Vascular endothelial growth factor (VEGF) inhibitors can prevent tumor cells from obtaining nutrients through neovascularization. When used in combination withGP chemotherapy, they may enhance anti-tumor activity. For example, Bevacizumab (Avastin) is a VEGF inhibitor that has been approved for the treatment of various cancer types, including colorectal cancer, non-small cell lung cancer, etc. It should be noted that due to the different conditions of each patient, the specific treatment plan should be determined by the doctor after comprehensive consideration of the patient's condition, medical history, and other factors.

Table 5: Case study.

In this section, we detail specific examples that demonstrate the superior functionality of PharmaGPT0.7, particularly in the fields of biomedicine and oncological treatments. These instances highlight the model's advanced capabilities in handling and elucidating complex, domain-specific information, which sets a new industry standard for AI-driven data interpretation in biomedical sciences.

Detailed Response Analysis of PharmaGPT0.7: The responses provided by PharmaGPT0.7 for both Chinese (ZH) and English (EN) queries about the use of GP chemotherapy in combination with immunotherapy illustrate its exceptional proficiency. GP chemotherapy, consisting of gemcitabine and cisplatin, is a regimen noted for its effectiveness against various tumors such as non-small cell lung cancer, esophageal cancer, and gastric cancer.

Specialization and Depth: PharmaGPT0.7 not only identifies the typical combinations of GP chemotherapy with immunotherapy drugs but also explains the synergy between these treatments in a clinically relevant context:

- PD-1/PD-L1 Inhibitors: The model correctly identifies drugs like Nivolumab and Pembrolizumab, explaining their mechanism to enhance immune system capability against tumor cells, a crucial piece of information for oncologists.
- CTLA-4 Inhibitors: Detailing the role of CTLA-4 inhibitors like Ipilimumab in preventing tumor cells from evading immune surveillance, PharmaGPT0.7 provides actionable insights into their combination with chemotherapy.
- VEGF Inhibitors: By mentioning Bevacizumab and its role in inhibiting neovascularization, the model showcases a deep understanding of how angiogenesis inhibitors can complement cytotoxic chemotherapy.

Clinical Relevance: PharmaGPT0.7 further enhances its responses by noting that the choice of specific immunotherapy drugs alongside GP chemotherapy must be personalized, taking into consideration the patient's medical history and current health condition. This approach reflects a sophisticated understanding of oncology, where treatment customization is paramount.

Comparison with Other Models: Unlike other models, PharmaGPT0.7's responses are not only rich in content but are also specifically tailored to reflect the latest trends and clinical trials in oncology. This allows healthcare professionals to receive the most current and relevant information, aiding in better decision-making processes.

Conclusion: PharmaGPT0.7's detailed, accurate, and contextually relevant responses set it apart in the landscape of AIdriven tools in biomedicine, making it an invaluable resource for professionals who require up-to-date and detailed information to make informed clinical decisions.

C. Comparison on the MMLU Dataset

We conducted a comparative analysis of our model, PharmaGPT, with leading Large Language Models (LLMs) such as GPT4, using the widely recognized MMLU dataset. Our findings are twofold: firstly, PharmaGPT achieves impressive results in general capabilities despite having significantly fewer parameters than its counterparts. Secondly, it demonstrates superior proficiency in specialized domains including biomedicine, chemistry, and related fields, surpassing the advanced capabilities of general-purpose large models.

Overall Strong Performance PharmaGPT consistently maintains MMLU scores in the 80-90% range across nearly all tasks, illustrating its robust capabilities in the biomedical and life sciences domains. This high performance indicates effective fine-tuning on a substantial corpus of relevant scientific literature.

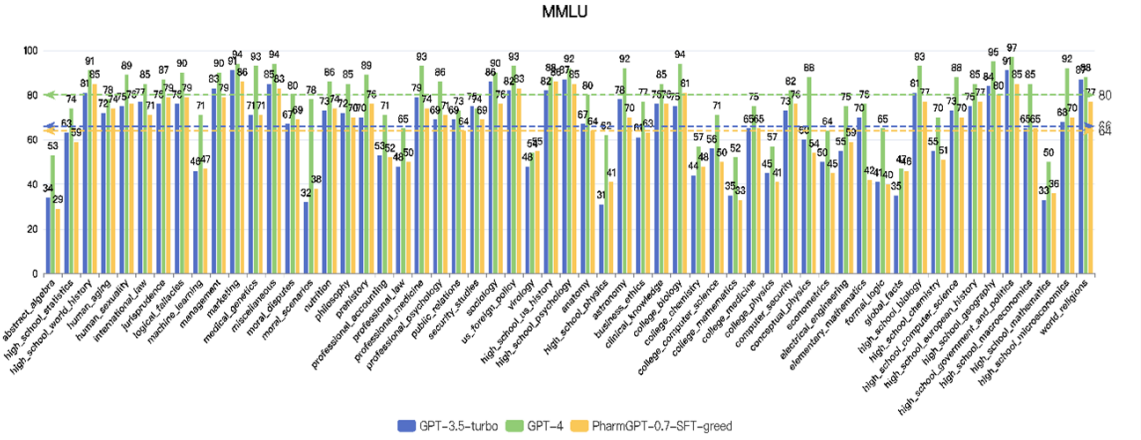

Figure 8: Test Performance of PharmaGPT on the General Test Set MMLU.

Outperforming GPT-3.5-turbo In the majority of tasks, particularly those related to biology, medicine, anatomy, and physiology, PharmaGPT not only achieves higher scores than GPT-3.5-turbo but often by significant margins. This highlights the benefits of domain-specific training, which enhances PharmaGPT's performance in biomedical question answering compared to more generalized models like GPT-3.5.

Comparable to GPT-4 Notably, PharmaGPT scores very closely to GPT-4 on a broad spectrum of topics and slightly outperforms GPT-4 in areas such as physiology, health sciences, and biology. This achievement is particularly impressive considering GPT-4's larger scale, indicating that PharmaGPT's focused training regimen allows it to match, and occasionally surpass, GPT-4's biomedical knowledge without the need for vast computational resources.

Areas for Improvement Despite its strengths, PharmaGPT shows room for improvement in topics such as world religions, philosophy, and other miscellaneous areas, where it significantly lags behind GPT-4. This limitation likely reflects the scope and nature of PharmaGPT's training data. However, given the model's specialized focus, this minor underperformance in non-core subjects is understandable and does not detract from its overall effectiveness for biomedical applications.

The results underscore PharmaGPT's state-of-the-art performance in biomedical language understanding, positioning it as a powerful tool for domain-specific NLP tasks. Its focused training approach not only allows it to surpass GPT-3.5 but also to match, and in some cases exceed, GPT-4's capabilities in key life science and healthcare topics, all while operating on a smaller scale. This makes PharmaGPT an attractive option for biomedical researchers and organizations seeking high accuracy and efficiency in their NLP implementations.