MoZIP: A Multilingual Benchmark to Evaluate Large Language Models in Intellectual Property

Shiwen Ni[1*], **Minghuan Tan**[1*], **Yuelin Bai**[1], **Fuqiang Niu**[2,3]
Min Yang[1†]**, Bowen Zhang**[2†]**, Ruifeng Xu**[4]**, Xiaojun Chen**[3]**,Ye Li**[1]**, Jianping Fan**[1]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]Shenzhen Technology University   [3]Shenzhen University
[4]Harbin Institute of Technology, Shenzhen
{sw.ni, mh.tan, yl.bai, min.yang, ye.li, jp.fan}@siat.ac.cn,
nfq729@gmail.com   zhang_bo_wen@foxmail.com
xuruifeng@hit.edu.cn   xjchen@szu.edu.cn

Abstract

Large language models (LLMs) have demonstrated impressive performance in various natural language processing (NLP) tasks. However, there is limited understanding of how well LLMs perform in specific domains (e.g, the intellectual property (IP) domain). In this paper, we contribute a new benchmark, the first **M**ultilingual-**o**riented qui**Z** on **I**ntellectual **P**roperty (**MoZIP**), for the evaluation of LLMs in the IP domain. The MoZIP benchmark includes three challenging tasks: IP multiple-choice quiz (IPQuiz), IP question answering (IPQA), and patent matching (PatentMatch). In addition, we also develop a new IP-oriented multilingual large language model (called **MoZi**), which is a BLOOMZ-based model that has been supervised fine-tuned with multilingual IP-related text data. We evaluate our proposed MoZi model and four well-known LLMs (i.e., BLOOMZ, BELLE, ChatGLM and ChatGPT) on the MoZIP benchmark. Experimental results demonstrate that MoZi outperforms BLOOMZ, BELLE and ChatGLM by a noticeable margin, while it had lower scores compared with ChatGPT. Notably, the performance of current LLMs on the MoZIP benchmark has much room for improvement, and even the most powerful ChatGPT does not reach the passing level. Our source code, data, and models are available at https://github.com/AI-for-Science/MoZi.

Keywords: Intellectual property, benchmark, large language model, multilingual

1. Introduction

With the development of Large Language Models (LLMs) (Workshop et al., 2022; OpenAI, 2023; Zeng et al., 2023; Touvron et al., 2023), AI assisted agents are showing increasing abilities in understanding natural language and manipulating well-formatted text drafted by humans. Recent studies show that LLMs supervised fine-tuned on domain-specific data achieve significant progress in a wide range of fields, such as Finance (Wu et al., 2023), Law (Liu et al., 2023; Huang et al., 2023a), Medicine (Zhang et al., 2023a; Wang et al., 2023) and Programming (Li et al., 2022). As large models have exhibited remarkable versatility in our daily lives, it is crucial to effectively evaluate their abilities in handling specific tasks and identify potential shortcomings. Recently, several benchmarks were proposed to evaluate the large models from different perspectives such as factual consistency (Laban et al., 2023), question answering in the medical field (Singhal et al., 2022), the fairness of recommendation (Zhang et al., 2023b), the programming ability (Chen et al., 2021), and comprehensive evaluation in general fields (Huang et al., 2023b).

However, we notice that in the area of AI for science, the protection and inspiration for creativity are still overlooked by the community. Intellectual Property (IP) has been a widely-used term to address rights in encouraging innovation and creativity. Since 2000, the World Intellectual Property Organization (WIPO) has established World Intellectual Property Day to "raise awareness of how *patents*, *copyright*, *trademarks* and *designs* impact on daily life" and "to celebrate creativity, and the contribution made by creators and innovators to the development of economies and societies across the globe" (WIPO, 2011). Despite the impressive capabilities of LLMs in natural language understanding and generation, it remains a challenge to explore to what extent LLMs can understand innovative ideas, cutting-edge creations, and their protections against infringements.

As far as we are concerned, the major challenges to developing LLMs in the IP domain are two-fold. First, due to the wide coverage of IP rights, there is still a lack of benchmarks evaluating how LLMs understand IP-related concepts and regulations. The major IP rights include *Patents*, *Trademarks*, *Industrial Designs*, *Geographical Indications*, *Copyright* and *Trade Secrets*. Unfortunately, existing benchmarks in QA or Laws do not focus on these directions. Second, despite the great demand for people from different professions, there are still obstacles to acquiring highly relevant information and protecting potential IPs using appropriate strate-

---

[*]Equal contribution
[†]Corresponding author

11658

gies. According to WIPO PCT Yearly Review 2022, the top 50 PCT geographical clusters accounted for nearly 60% of total PCT filings. As a result, it is essential to develop a IP-oriented LLM and construct a benchmark for evaluating the performance of different LLMs.

To make fair comparisons of LLMs over IP knowledge, we propose a multilingual benchmark called **MoZIP** (**M**ultilingual-**o**riented qui**Z** for **I**ntellectual **P**roperty). The MoZIP benchmark consists of three datasets, IPQuiz, IPQA and PatentMatch. Since IP rights are protected under IP systems, a prior requirement for the LLMs is domain-specific knowledge about terminologies and regulations of IP systems. To consider the above knowledge, we first construct the IPQuiz dataset, which consists of 2000 multilingual multiple-choice questions on IP knowledge. These questions are gathered from online IP knowledge tests from different countries and languages. We also pay special attention to frequently asked questions on websites of IP-related organizations and agencies. These questions contain important information that IP consumers care about. From these questions, we select 100 items to form the IPQA test set. Among all the highlighted rights of IP, *patents* encourage the development of innovations and new technologies in every field. However, there have been numerous new patents being submitted each day that searching or indexing similar patents has been a challenge for current systems. We collect patent documents covering different languages across the globe to help the model to learn how patents are drafted and innovations are described. Based on this, we further construct the PatentMatch dataset to test how models may differentiate patents based on the descriptions.

In this paper, we also develop our IP-oriented LLM **MoZi**. The model is finetuned from BLOOMZ-MT-7B through three stages. For the first stage, we use 24 million official patent documents to make the model aware of how patents are usually crafted. For the second stage, we conduct 3 million instruction finetuning using various instruction types across multiple fields. In the last stage, 58k IP instruction fine-tuning data constructed by ourselves is used to enable the model to learn about IP knowledge. We conducted experiments based on five LLMs of MoZi, ChatGPT, ChatGLM, BELLE, and BLOOMZ and evaluated each model on our benchmark **MoZIP**. Our experiments show that ChatGPT performs best overall, followed by the other 6-7b parameter number models in which MoZi performs best. Overall, there is still much room for improvement in the current LLMs' underperformance on MoZIP. The contributions of this paper to the community are summarized as follows:

- This paper presents **MoZIP** the first IP benchmark covering nine languages for evaluating the capabilities of large language models in the IP domain.

- In this work, we propose the first IP-oriented multilingual large language model **MoZi**, and experimental results on the MoZIP benchmark show that **MoZi** performs the best among models of the same parameter level.

- We conducted a comprehensive evaluation using five LLMs on the MoZIP benchmark, and the experimental results show the challenge of the MoZIP benchmark and illustrate the deficiencies of LLMs in the IP field at this stage.

- In order to contribute to the development of LLMs in intellectual property, we have made available source code, **MoZIP** benchmark, instruction fine-tuning data, and **MoZi** model.

2. Primary data collection

To facilitate the construction of an IP-oriented benchmark, we collect different categories of data from multiple resources, including IP acts, patents as well as frequently asked questions related to IP. These data can be used as instructions for fine-tuning to inject IP-related knowledge into LLMs.

IPACT To protect IP rights, there has been plenty of legal information on intellectual property available online, including IP laws and regulations, WIPO-administered and IP related treaties, and leading judicial decisions on IP. This part of the data is basically objective knowledge. LLMs need to access such resources to acquire knowledge about mechanisms of IP systems.

IPFAQ Frequently Asked Questions (FAQs) related to IP usually contain basic knowledge about IP rights and important information that IP customers care most about but likely to misunderstand. We crawl FAQs from worldwide websites of IP organizations and agencies. Their answers are officially provided with guaranteed reliability and accuracy. The dataset therefore contains questions asked by people from different countries and answers given online.

Patent "A patent is an exclusive right granted for an invention, which is a product or a process that provides." (WIPO, 2011) Generally speaking, a patent needs to propose a new way of doing something, or offer a new technical solution to a problem. In order to obtain a patent, one must reveal technical details about their invention in a patent application that is made available to the public.

We gather patent documents from a database of CNIPA[1]. Figure 1 shows an example of a patent

---

[1]The Patent data is crawled from `http://www.szxyd.org.cn` before March 15, 2023.
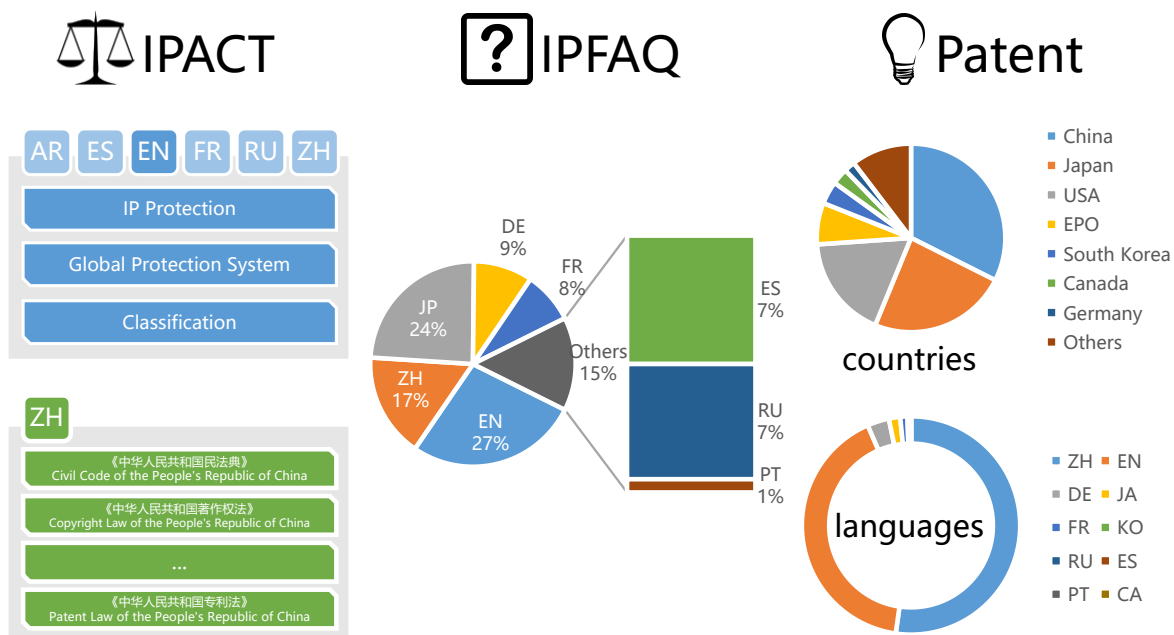
Figure 1: Statistics and distribution of data. ZH-Chinese, EN-English, DE-German, JA-Japanese, FR-French, KO-Korean, RU-Russian, ES-Spanish, PT-Portuguese, CA-Catalan.

document. Each patent document contains different fields of data, like *country of the patent*, *patent type*, *public number*, *claims*, *International Patent Classification (IPC) information* and *description*. The patents may be written in different languages regardless of the *country of the patent*. We use the *claims* field and the *description* field as language indicators. If the two fields are empty (about half of the Patent data has this issue), the patent is not used for language detection. We adopt a language detector from FastText ([Joulin et al., 2016a,b](#)) to categorize the patents into different language groups. The right of Figure 1 shows the statistics for country and language distributions for all used patents.

3. Benchmark

The **MoZIP** benchmark contains **IPQuiz**, **IPQA** and **PatentMatch**, see Table 1.

IPQuiz[2] We construct the IPQuiz dataset to evaluate if models understand IP-related concepts and regulations. IPQuiz is a multiple-choice question answering dataset gathered from publicly accessible websites of various languages across the globe. For each question, the model needs to choose a response from a list of candidates, see Figure 2. The dataset is keeping increasing in size and coverage of languages. We will host the dataset on HuggingFace's *datasets* website. For the initial ver-

Dataset	IPQuiz	IPQA	PatentMatch
Types	M-C	Generation	M-C
EN	834	35	500
ZH	564	35	500
XL	623	30	-
Size	2021	100	1000
# languages	7	7	2

Table 1: Statistics for each dataset in MoZIP benchmark. IPQuiz-XL includes DE, ES, JP, KO and PT. IPQA-XL includes ES, JP, DE, FR, RU. M-C: Multiple-choice.

sion of IPQuiz, it contains 2k questions with seven languages. Figure 1 shows the distribution of languages for the initial version of the dataset.

IPQA[3] We reserve 100 questions from **IPFAQ** as test set to evaluate how LLMs understand IP-related questions. The IPQA contains questions in seven languages, and the 100 data items include 35 each in Chinese and English, and 6 each in Spanish, Japanese, German, French, and Russian. The specific IP-questions are shown in Figure 3. Their responses are further judged by human annotators to compare the quality.
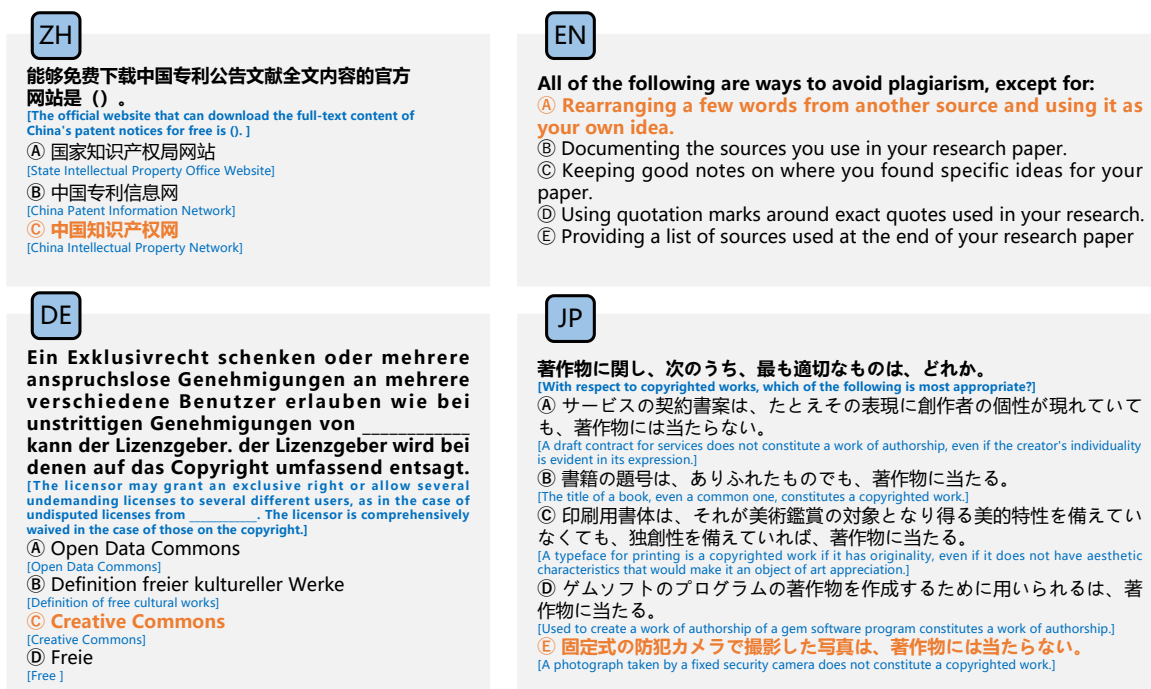
ZH

能够免费下载中国专利公告文献全文内容的官方网站是（）。
[The official website that can download the full-text content of China's patent notices for free is ().]
Ⓐ 国家知识产权局网站
[State Intellectual Property Office Website]
Ⓑ 中国专利信息网
[China Patent Information Network]
Ⓒ 中国知识产权网
[China Intellectual Property Network]

EN

All of the following are ways to avoid plagiarism, except for:
Ⓐ Rearranging a few words from another source and using it as your own idea.
Ⓑ Documenting the sources you use in your research paper.
Ⓒ Keeping good notes on where you found specific ideas for your paper.
Ⓓ Using quotation marks around exact quotes used in your research.
Ⓔ Providing a list of sources used at the end of your research paper

DE

Ein Exklusivrecht schenken oder mehrere anspruchslose Genehmigungen an mehrere verschiedene Benutzer erlauben wie bei unstrittigen Genehmigungen von _____ kann der Lizenzgeber. der Lizenzgeber wird bei denen auf das Copyright umfassend entsagt.
[The licensor may grant an exclusive right or allow several undemanding licenses to several different users, as in the case of undisputed licenses from _____. The licensor is comprehensively waived in the case of those on the copyright.]
Ⓐ Open Data Commons
[Open Data Commons]
Ⓑ Definition freier kultureller Werke
[Definition of free cultural works]
Ⓒ Creative Commons
[Creative Commons]
Ⓓ Freie
[Free]

JP

著作物に関し、次のうち、最も適切なものは、どれか。
[With respect to copyrighted works, which of the following is most appropriate?]
Ⓐ サービスの契約書案は、たとえその表現に創作者の個性が現れていても、著作物には当たらない。
[A draft contract for services does not constitute a work of authorship, even if the creator's individuality is evident in its expression.]
Ⓑ 書籍の題号は、ありふれたものでも、著作物に当たる。
[The title of a book, even a common one, constitutes a copyrighted work.]
Ⓒ 印刷用書体は、それが美術鑑賞の対象となり得る美的特性を備えていなくても、独創性を備えていれば、著作物に当たる。
[A typeface for printing is a copyrighted work if it has originality, even if it does not have aesthetic characteristics that would make it an object of art appreciation.]
Ⓓ ゲムソフトのプログラムの著作物を作成するために用いられるは、著作物に当たる。
[Used to create a work of authorship of a gem software program constitutes a work of authorship.]
Ⓔ 固定式の防犯カメラで撮影した写真は、著作物には当たらない。
[A photograph taken by a fixed security camera does not constitute a copyrighted work.]

Figure 2: Examples of questions in IPQuiz. The words in blue below non-English content are the corresponding English translations.

DE: Können Minderjährige Erfindungen zum Patent anmelden?
[Can minors file patent applications for inventions?]

EN: How much does it cost to patent an invention?

ES: ¿En qué consiste la exclusividad de la PI?
[What does IP exclusivity consist of?]

FR: Quelle est la durée de la protection par brevet?
[How long does patent protection last?]

JP: 大学と企業とが共同研究を行う場合に知的財産の扱いはどうなりますか。
[How is intellectual property treated when a university and a company conduct joint research?]

RU: Являются ли патенты единственным средством охраны изобретений?
[Are patents the only means of protecting inventions?]

ZH: 获得了专利证书，就表示你会necessarily拥有该专利权了吗?
[Does obtaining a patent certificate mean that you will necessarily own the patent?]

Figure 3: Examples of seven language questions in IPQA dataset. The words in blue below non-English content are the corresponding English translations.

PatentMatch [4] We construct PatentMatch to assess whether the model truly comprehends the inventions described in patent documents and accurately differentiates between different patents. We first create a parallel dataset of 250k patents granted by WIPO from 2010 to 2022 collected from Google Public Datasets [5], with both Chinese and English abstracts extracted from the original documents. These contents are written by the patent applicants, ensuring the accuracy of the descrip-

tions. We construct this sub-task following the steps below:

- First, we leverage the Pyserini toolkit (Lin et al., 2021) to build a BM25 database. And we use OpenAI API (text-embedding-ada-002 model) [6] to build a dense vector database based on Pinecone[7] with cosine similarity metric.

- Then, we select 500 patents across various technical domains based on 8 sections from the IPC classification system to construct 1000 multiple-choice questions(500 questions in each language). Next, we use the abstracts of these patents as queries to retrieve the top-k results from both the BM25 database and the dense vector database.

- As our setting is to select the most relevant patent from the options, we follow the following steps to construct these choices. (1) To determine the accurate answer, we choose patents that belong to the same IPC subgroup as the patent mentioned in the question. We then make sure these patents have a low BM25 ranking but a high ranking in vector-based assessments. This setup allows us to assess whether the model can effectively recognize similar patents even when there are fewer word

---

[4] https://huggingface.co/datasets/BNNT/PatentMatch

[5] https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-public-data

[6] https://platform.openai.com/docs/guides/embeddings

[7] https://www.pinecone.io/

Figure 4: An example in PatentMatch. The texts with blue color are overlappings between the source patent and each candidate patent. However, the texts with green background color are the key information why the two patents match each other.

overlaps. (2) We choose the remaining options based on patents with high rankings in BM25 but low rankings in the vector-based assessment. Additionally, we ensure that these patents have a different IPC classification code with patent in the question.

Human evaluators then perform manual verification to validate the quality of the constructed questions. We made necessary modifications or deletions based on their feedback, resulting in a final set of 500 PatentMatch questions in both English and Chinese.

4. The Proposed MoZi Model

As depicted in Figure 5, we conduct three stages of fine-tuning on an IP-oriented language model known as **MoZi**, which is built upon the BLOOMZ-MT-7B model. These stages entail the utilization of the data gathered in Section 2. Due to the substantial amount of Patent data available, we choose to employ it in the initial stage. Following that, we proceed with two additional stages of instruction finetuning, incorporating both general-domain instructions and IP-domain-specific instructions.

Patent Pre-training Patent documents contain detailed descriptions and instructions about the inventions. However, in practice, most models struggle in differentiating patents or retrieving similar patents given the large amount of candidates and high overlapping of text. To strengthen abilities in this direction, we use the Patent Data to train our model **MoZi**. Our model is initialized from a publicly released checkpoint[8] of BLOOMZ-MT, which is a multilingual model finetuned on xP3mt (Muennighoff et al., 2022). We construct a training case from each patent document. The context includes the patent's title, abstract, claims and description.

General Instruction Finetuning[9] In total, we used 3,025,600 general instructions for the first stage of our proprietary pre-trained model. The 3m instruction data come from several public datasets, including 1) BELLE Chinese general instructions, 2) Alpaca-gpt4 English general instructions, 3) BELLE Chinese general conversation instructions, 4) Sharegpt-vicuna English general conversation

---

[8] https://huggingface.co/bigscience/bloomz-7b1-mt

[9] https://huggingface.co/datasets/BNNT/mozi_general_instructions_3m
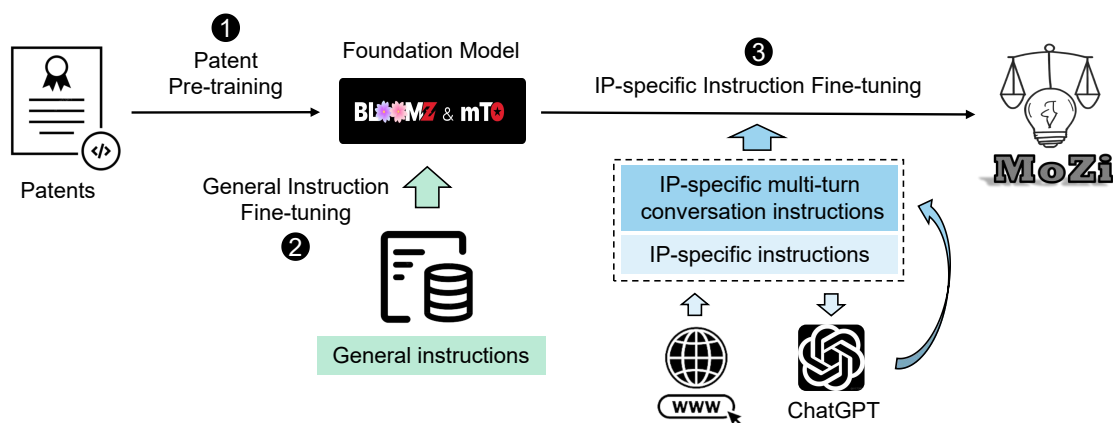
11662

Figure 5: Schematic of our proposed IP-oriented multilingual large language model **MoZi**.

instructions, and 5) GuanacoDataset Chinese, English, and Japanese general instructions.

IP-specific Instruction Finetuning[10] For the model after fine-tuning the generic instruction, we perform the second stage of IP-specific Instruction Finetuning upon it. The fine-tuned data in this phase consisted of a total of 58,874 items, including multilingual Q&A datasets we collected from the IPFAQ dataset, Chinese IP-related law articles in IPACT, and Chinese multi-turn conversation data in the IP domain generated by ChatGPT. Let's think of each law as an instruction, where the input is the name of the law and the output is the specifics of the law. ChatGPT generates a prompt for a multiple round conversation: "*The following is a conversation between a user and a patented AI assistant. The user and the patented AI assistant are having a conversation around this topic: [seed question]. The user utterance begins with human and the AI assistant utterance begins with assistant. The user asks relevant questions about the topic in question or about previous conversations. When they have no more questions, the user will stop the conversation. The AI assistant tries not to ask questions.*"

5. Experiments

5.1. Training Details

Our MoZi-7b model was implemented in PyTorch using the *transformer* and *deepspeed* packages and with Bloomz-7b1-mt (Workshop et al., 2022) as the Foundation model. We used ZeRO-3 (Rajbhandari et al., 2020) to distribute the model over 8 × A100 GPUs for training. During supervised fine-tuning, we set the learning rate, batch size, and

maximum token length to 5e-6, 64, and 2048, respectively. Moreover, weight_decay is set to 0.0001 and num_warmup_steps is set to 100. The training epochs for the pre-training phase are 1 time, the general instruction fine-tuning phase is 2 times, and the IP instruction fine-tuning stage is 4 times.

5.2. Baselines

In addition to evaluating the proposed **MoZi-7b** on the MoZIP benchmark, we evaluated the following baseline models:

- **ChatGPT** [11] A powerful large language model developed by OpenAI, which is based on the *gpt-3.5-turbo* model. We used the API provided by openai for the evaluation.

- **ChatGLM-6b (Du et al., 2022; Zeng et al., 2022)** An open source, Chinese-English bilingual large language model proposed by Tsinghua University, based on the general language model (GLM) architecture.

- **BELLE-7b (Ji et al., 2023a,b)** It is a Bloomz-7b1-mt fine-tuned model that combines 2 million Chinese data and 50,000 English data from the open source Stanford-Alpaca, which makes it have excellent Chinese comprehension and response generation capabilities.

- **BLOOMZ-7b (Workshop et al., 2022; Muennighoff et al., 2022)** A LLM obtained by fine-tuning the BLOOM and mT5 pre-trained multilingual models on a cross-language task mixture (xP3) with the ability to generalize across languages to unseen tasks and languages.

The LLaMA family of large language models performs adequately only in English, and its multilingual capability is limited. Consequently, we opted

---

[10] https://huggingface.co/datasets/BNNT/
mozi_IP_instructions

[11] https://openai.com/blog/chatgpt

11663

Dataset	Model				
	MoZi-7b (ours)	BLOOMZ-7b	BELLE-7b	ChatGLM-6b	ChatGPT
IPQuiz-EN	41.5	29.3	<u>42.7</u>	38.9	**60.8**
IPQuiz-ZH	<u>39.2</u>	29.1	38.6	31.7	**45.8**
IPQuiz-XL	<u>37.4</u>	29.4	36.1	30.5	**42.2**
Average	<u>39.4</u>	29.3	39.1	33.7	**49.6**

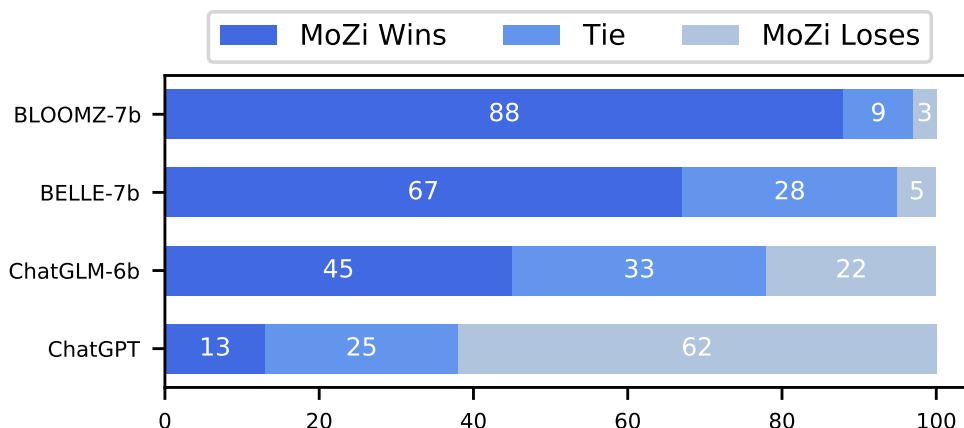Table 2: Performance of all used models on IPQuiz.

Figure 6: The number of questions on IPQA that MoZi wins, ties or loses.

not to employ the LLaMA family of models for measurement purposes.

5.3. Experimental Results

IPQuiz consists of 2k multiple-choice questions about IP knowledge covering 7 languages. We construct this prompt ("*Please give the correct option for the following question*") to LLMs and collect generated responses from them. We used regularity and manual verification to determine which candidate was selected as the answer. Then we compute answer accuracy over the dataset. The experimental results are shown in Table 2, Chat-GPT performs the best on IPQuiz-zh, IPQuiz-en and IPQuiz-xl because of its parameter level and the amount of training data. With the exception of ChatGPT, our MoZi-7b had the best average accuracy of 39.4% among models with approximately 7b number of parameters. Compared to the foundation model we used, BLOOMZ-7b, our MoZi-7b improved on average by 10.1% on IPQuiz tasks in various languages, which illustrates the significant effect of further pre-training and instruction fine-tuning. Overall, none of the scores exceeded 60%, except for ChatGPT, which reached 60.8% on IPQuiz-en. This shows that current LLMs almost always fail in performance on our proposed IPQuiz

dataset. There is a lot of room for improving the performance of LLMs in the IP domain at this stage.

IPQA We provide the annotators with a question and two potential answers generated by different models. The annotators were then tasked with indicating which answer they thought was superior, or whether there was a significant difference between the two answers. If there is no distinction then label the tie. We assess inter-annotator agreement using tie-discounted accuracy. In this method, we award one point when both annotators agree on a label, half a point when either annotator (but not both) labels a tie, and no points when there is no agreement. We compared two different human annotators and obtained a consistent score of 81%. Despite the presence of some subjective elements in this task, the agreement among human annotators is reasonably good.

The results of the experiments on IPQA are shown in Figure 6, and are ChatGPT, MoZi-7b, ChatGLM-6b, BELLE-7b, and BLOOMZ-7b in descending order of performance results. Our MoZi-7b model defeated the original foundation model, BLOOMZ-7b, a total of 88 times, losing only 3 times. Also both BLOOM-based models, our MoZi-7b beat BELLE-7b a total of 67 times, losing only 5 times and tying 28 times. ChatGLM-6b has performed

Figure 7: Comparisions for answers generated by diffenrent models for a question in IPQA, the models from top to bottom are MoZi-7b, ChatGPT, ChatGLM-6b, BELLE-7b and BLOOMZ-7b.

Dataset	Model				
	MoZi-7b (ours)	BLOOMZ-7b	BELLE-7b	ChatGLM-6b	ChatGPT
EN	<u>25.8</u>	24.0	14.2	23.8	**34.6**
ZH	<u>29.0</u>	24.2	26.4	27.2	**43.0**
Average	<u>27.4</u>	24.1	20.3	25.5	**38.8**

Table 3: Evaluation results for all models over PatentMatch.

well in IPQA, losing to MoZi-7b only 45 times, winning over MoZi-7b 22 times and tying 33 times. Finally comparing to the powerful ChatGPT, our MoZi-7b also beat it 13 times and tied it 25 times.

In addition, Figure 7 is a specific IPQA example that we show. From top to bottom are the human questions, the answer of MoZi-7b, the answer of ChatGPT, the answer of ChatGLM-6b, the answer of BELLE-7b and the answer of BLOOMZ-7b, respectively. MoZi-7b has the most accurate answer.

PatentMatch The experimental results for the five LLMs on the PatentMatch task are shown in Table 3. ChatGPT still performs the best. However, we found that ChatGPT only got 43.0% and 34.6% accuracy on PatentMatch-zh and PatentMatch-en, respectively. The mean scores of the four models MoZi-7b, ChatGLM-6b, BELLE-7b and BLOOMZ-

7b were 27.4%, 24.1%, 20.3% and 25.5%, respectively. Our MoZi is also better than other LLMs in its class in both PatentMatch-zh and PatentMatch-en. That is, the average score of the other four models except ChatGPT is useless to exceed 30%. We also found that BELLE-7b's score in PatentMatch-en was even only 14.2% far worse than a blind guess. We believe that the reason why all models perform so poorly may be that patent data are too rare and that LLMs have more difficulty understanding long texts. The input of the models in the PatentMatch task is basically more than 1,000 tokens. From the experimental results, it still faces a great challenge for LLMs to process long texts in specialized fields.

6. Conclusion

We introduce **MoZIP**, the inaugural multilingual benchmark for evaluating Language Models (LLMs) in the field of Intellectual Property (IP). MoZIP encompasses a wide range of real-world objective knowledge and various types of user questions. It comprises three distinct datasets: *IPQuiz*, *IPQA*, and *PatentMatch*, which collectively cover nine languages. As part of this research, we present **MoZi**, the first IP-oriented Multilingual Large Language Model. We evaluate five models, namely MoZi-7b, ChatGPT, ChatGLM-6b, BELLE-7b, and BLOOMZ-7b, using the MoZIP benchmark. The experimental findings highlight the challenging nature of the MoZIP benchmark and the current deficiency of IP-related knowledge among LLMs. To facilitate research within the recommendation community, we release the source code, benchmark datasets, instruction fine-tuning data, and the MoZi model. Our aim is for MoZIP to serve as a standardized benchmark for evaluating LLMs in the IP domain. Looking ahead, we plan to develop a more comprehensive dataset that includes a greater number of minor languages.

7. Ethics Statement

The sources of our data are publicly available URLs on the Internet, URLs that can be accessed without registering for an account. And the data does not involve personal privacy. The data are collected legally. The sources of all our data are open and transparent. The purpose of the data use is to evaluate and improve the capabilities of large-scale language models in the field of intellectual property.

8. Acknowledgements

9. Bibliographical References

Anthony McEnery and others. 2004. *The EMILLE/CIIL Corpus*. EMILLE (Enabling Minority Language Engineering) Project. distributed via ELRA: ELRA-Id W0037, ISLRN 039-846-040-604-0.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023a. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023a. Belle: Be everyone's large language model engine. https://github.com/LianjiaTech/BELLE.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023b. Exploring the impact of instruction data scaling on large language models:

An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Khalid Choukri and Niklas Paullson. 2004. *The OrienTel Moroccan MCA (Modern Colloquial Arabic) database*. distributed via ELRA: ELRA-Id ELRA-S0183, ISLRN 613-578-868-832-2.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R. Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Llms as factual reasoners: Insights from existing benchmarks and beyond.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Lawgpt: Chinese legal dialogue language model. https://github.com/LiuHC0428/LAW_GPT.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

OpenAI. 2023. Gpt-4 technical report.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Speecon Consortium. 2011. *Catalan Speecon database*. SpeeCon. Speecon Project, distributed via ELRA: ELRA-Id S0327, Speecon resources, 1.0, ISLRN 935-211-147-357-5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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

WIPO. 2011. Wipo web site, world intellectual property day – 26 april. [Online; Accessed 20 April 2011.].

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,

Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

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

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. Huatuogpt, towards taming language models to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023b. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation.