Phoenix: Democratizing ChatGPT across Languages

Zhihong Chen[1,2], **Feng Jiang**[1], **Junyin Chen**[1,2], **Tiannan Wang**[1,2], **Fei Yu**[1], **Guiming Chen**[1]
Hongbo Zhang[1,2] , **Juhao Liang**[1,2], **Chen Zhang**[1], **Zhiyi Zhang**[1], **Jianquan Li**[1], **Lian Zhang**[2], **Xiang Wan**[1,2]
Benyou Wang[1,2] *
[1] School of Data Science, The Chinese University of Hong Kong, Shenzhen
[2] The Shenzhen Research Institute of Big Data.
wangbenyou@cuhk.edu.cn

Abstract

This paper introduces our efforts to democratize ChatGPT across Languages. The trained large language model, named 'Phoenix', achieves comparable performance in open-source English end Chinese models, while it achieves excellent performance in low-resource languages. We believe this work will be beneficial to democratize ChatGPT, especially in countries that do not use Latin languages. Our data, code, and models are available at https://github.com/FreedomIntelligence/LLMZoo.

1 Introduction

First, we would like to introduce 'AI Supremacy'.

Definition 1 (AI supremacy) *'AI supremacy' refers to a company's absolute leadership and monopoly position in an AI field, which may even include exclusive capabilities beyond general artificial intelligence. This is unacceptable for the AI community and may even lead to individual influence on the direction of the human future, thus bringing various hazards to society.*

Nowadays, the ChatGPT and its successor GPT 4 were developed and maintained by a single company, which unexpectedly results in 'AI Supremacy'. As expressed in the widely-recognized Asilomar AI Principles, the development of advanced artificial intelligence has the potential to bring about a significant and transformative shift in the history of life on Earth [2]. Therefore, the existence of AI supremacy could result in an unexpected consequence that the future of human beings (even all alive animals or plants) will be controlled by a single company; the responsibility of such a company might not be well-controlled.

Make AI open again. Therefore, we aim to lower the cost and barrier of the ChatGPT training so that more responsible researchers can join the ChatGPT research and share their diverse thoughts,

*Benyou is the corresponding author

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

like figuring out *how it works*, *why it works*, and more importantly, *how to develop large language models (like ChatGPT) in a planet-safe way*. This process is called democratization for the access and study of LLMs in [13], where [1, 6, 4, 2] are among the process, see Sec. 2.1 for more details.

1.1 Methodology

Nowadays, the open-source efforts to democratize ChatGPT access [13] and study explicitly exclude non-Latin and non-Cyrillic languages. This is definitely inconsistent with the open-source spirit. Imagine that one could decide not to allow a group of people to use light bulbs and vaccines – most of us (even those who could use bulbs and vaccines) should be offended.

Therefore, this work is among the efforts to **democratize ChatGPT across Languages**. Currently, there are two lines of work to develop democratized ChatGPT [3].

- (I) **Instuction-basd Tuning**. Instruction Tuning aims to *tame* language models to follow human instructions [9], which might be manually designed, or in a hybrid fashion in that humans write some seed instructions and OpenAI ChatGPT is used to generate more similar instructions using in-context learning [15].
- (II) **Conversation-basd Tuning**. ChatGPT-distilled Conversation is used to teach language models to chat like OpenAI ChatGPT [2] while the instruction data is usually for single-turn question answering.

Phylosophy of methodology. We follow the two lines of work to train our *multilingual* democratized ChatGPT. The key difference in our models is that we utilize two sets of data, namely *instructions* and *conversations*, which were previously only used by Alpaca and Vicuna, respectively. We believe incorporating both data types is essential for a recipe to achieve a proficient language model. The rationale is that the *instruction* data helps to tame language models to adhere to human instructions and fulfill their information requirements, while the *conversation* data facilitates the development of conversational skills in the model. These two types of data complement each other to create a more well-rounded language model.

Training protocol. However, the main challenge is to gather sufficient multilingual data for both types of data. To address this issue, we collect language-agnostic instructions and translate them into other languages based on the probability distribution of realistic languages. To account for language-specific knowledge and cultural backgrounds, we also manually design some language-specific instructions. For the latter, we collect various sources of multilingual ChatGPT-distilled conversations. The above training protocol enables us to train models that can be used in many languages. Notably, we chose Bloom [16] as our backbone as it covers a broader range of languages.

1.2 Phylosophy to name 'Phoenix'

The biggest barrier to developing LLMs is that we do not have enough candidate names for LLMs, as LLAMA, Guanaco, Vicuna, and Alpaca have already been used, and there are no more members in the camel family.

We name our model 'Phoenix '. In Chinese culture, the Phoenix is commonly regarded as a symbol of the king of birds; as the saying goes "百鸟朝凤", indicating its ability to coordinate with all birds, even if they speak different languages. We refer to Phoenix as the one capable of understanding and speaking hundreds of (bird) languages [4].

A tailored 'Phoenix' that is specific to the Latin language is called 'Chimera'. Chimera is a similar hybrid creature in Greek mythology, composed of different Lycia and Asia Minor animal parts. Phoenix and Chimera are two legendary creatures standing for eastern and western culture, respectively. We put them in a zoo to wish for a great collaboration to democratize ChatGPT.

A Latin version of Phoenix is referred to as 'Chimera'. This unique creature shares its name with a mythical beast from Greek folklore, which was said to be a combination of various animal parts

[3]In the rest of our paper, ChatGPT does not refer to a specific product developed by OpenAI (called 'OpenAI ChatGPT'), but a series of large language models designed for dialogue which might be from any companies.

[4]More importantly, Phoenix is the totem of "the Chinese University of Hong Kong, Shenzhen" (CUHKSZ); it goes without saying this is also for the Chinese University of Hong Kong (CUHK).

from Lycia and Asia Minor. The Phoenix and Chimera are legendary creatures representing Eastern and Western cultures, respectively. We have placed them together in a zoo to promote collaboration between Eastern and Western people for ChatGPT democratization.

1.3 Results

Phoenix in the multilingual benchmark Phoenix achieves the SOTA performance among Chinese open-source large language models (BELLE and Chinese-LLaMA-Alpaca [5]) based on GPT-4 evaluations. In other non-Latin languages, Phoenix largely outperforms existing LLMs in many languages, including Arabic, Japanese, and Korean.

Phoenix does not achieve state-of-the-art (SOTA) performance in open-source Latin language models (like Vicuna [2]). This is because Phoenix additionally pays a 'multilingual tax', mainly when dealing with non-Latin or non-Cyrillic languages. As democratization itself cares about minor groups who speak relatively low-source languages, we believe such a 'multilingual tax' for minor languages is worthy of paying. On the other hand, it is worth noting that texts in various languages might share some commonness, so information and knowledge behind multilingual languages might be transferable. This gives multilingual LLMs additional merit to process cross-culture tasks in more comprehensive tasks. In some senses, This underscores the value of linguistic diversity and the need to consider the perspectives of individuals from diverse linguistic backgrounds, especially people who speak minor languages.

Definition 2 (multilingual tax) *A multilingual model, with a limited size, may not perform as well as a language-specific model when performing tasks specific to a particular language. This is because the multilingual model is designed to adapt to many languages, and some of its training may not be optimized for the specific language. As a result, language-specific models may be more accurate and efficient when dealing with tasks specific to a particular language.*

Tax-free Phoenix: Chimera . Since Chimera is the Phoenix with a Latin backbone, Chimera does not pay for the multilingual tax. In English benchmarking, Chimera impresses GPT-4 with 96.6% ChatGPT Quality, setting a new SOTA in open-source LLMs.

1.4 Siginificance of Phoenix

- We conduct multilingual adaption for LLMs, especially for the non-Latin language model. Phoenix is the first multilingual democratized (by definition open-source) ChatGPT in which both backbone pre-training and post-training involve rich multilingual data. Phoenix is the SOTA open-source Language models in many non-Latin languages.
- In popular languages, Phoenix is among the first-tier Chinese large language models that achieve performance close to that of OpenAI ChatGPT; its Latin version Chimera is competitive in English.
- We benchmarked many existing LLMs using both automatic evaluations and human evaluations. We additionally evaluate the multiple aspects of language generations of LLMs. This is among the first work to systematically evaluate extensive large language models.

2 Overview of existing Democratized ChatGPTs

2.1 The tendency to democratize ChatGPT

Since the release of ChatGPT, an increasing number of related models have been developed and published based on the LLaMA [13] and BLOOM [16] models. Other than LLaMA and BLOOM that were *pre-trained* by a massive amount of plain corpora, the recent work tends to focus on *post-training*, which take a pre-trained backbone model (e.g., LLaMA and BLOOM) and skip the first pre-training step. Note that post-training is much computationally cheaper and there affordable to some research teams.

[5]We exclude ChatGLM-6B since its training details and data are transparent; therefore, it is impossible to replicate it from scratch. In this paper, we categorize ChatGLM-6B under non-open source models.

Table 1: Model Comparison.

Model	Backbone	#paras	Open-source		Claimed language	Post-training				Release date
						instruction		conversation		
			model	data		data	lang	data	lang	
ChatGPT	unknown	unknown	✗	✗	multi					11/30/22
Wenxin [6]	unknown	unknown	✗	✗	zh					03/16/23
ChatGLM [7]	GLM	6B	✓ [1]	✗	en/zh					03/16/23
Tongyi [8]	unknown	unknown	✗	✗	zh					04/07/23
Shangliang [9]	unknown	unknown	✗	✗	zh					04/10/23
Alpaca [12]	LLaMA	7B	✗	✓	en	52K	en	✗	✗	03/13/23
Dolly [10] [2]	GPT-J	6B	✓	✓	en	52k	en	✗	✗	03/24/23
BELLE [6]	BLOOMZ	7B	✓	✓	zh	1.5M	ch	✗	✗	03/26/23
Guanaco [11]	LLaMA	7B	✓	✓	en/zh/ja/de	534K [3]	4 [4]	✗	✗	03/26/23
Chinese-alpaca [3]	LLaMA	7/13B	✓	✓	en/zh	2M/3M	en/zh	✗	✗	03/28/23
LuoTuo [7]	LLaMA	7B	✓	✓	zh	52k	cn	✗	✗	03/31/23
Vicuna [2]	LLaMA	7/13B	✓	✓ [5]	en	✗	✗	70K	multi [6]	03/13/23
Koala [12]	LLaMA	13B	✓	✓	en	355K	en	117K	en	04/03/23
BAIZE [17]	LLaMA	7/13/30B	✓	✓	en	✗	✗	111.5K	en	04/04/23
Phoenix	BLOOMZ	7B	✓	✓	multi		40+			04/08/23
Latin Phoenix (Chimera)	LLaMA	7B/13B	✓	✓	Latin		40+			04/08/23

[1] Only release the weights.
[2] Dolly 2.0 based on pythia-12b model was published in 04/12.
[3] 32,880 chat dialogues without system input and 16,087 chat dialogues with system input.
[4] English, Simplified Chinese, Traditional Chinese (Taiwan, Hong Kong), Japanese, Deutsch.
[5] They only claimed that ShareGPT is the data source but did not provide the files.
[6] This dataset is collected from ShareGPT, mainly in English.

These post-training-based works can be divided into two categories. The first category is instruction-based tuning, and Alpaca [12] is a notable example. It employs the self-instruction technique [15] to generate more instructions by the GPT 3.5 model for fine-tuning, resulting in more accurate and contextually relevant outputs. Subsequently, the second category is conversation-based tuning models that utilize the distillation of user interactions with ChatGPT. Vicuna [2] serves as a representative model for this approach, capitalizing on large-scale user-shared dialogue datasets to improve model performance, aside from a few commercialized, closed-source models (such as Wenxin [13], Tongyi [14], Shangliang [15]), the majority of popular open-source models follow the principles of these two categories of post-training in their training methodologies and the most representative work are shown in Table. 1.

Instruction-based Tuning Although Alpaca [12] only released a training set consisting of 52K examples generated using the self-referential instruction method, many variant models have been fine-tuned on Alpaca's instruction dataset, including Dolly [16] based on GPT-J [11] and LuoTuo [7], which is trained on translated versions of the dataset in Chinese based on LaMMA. The BELLE model [6], on the other hand, followed the self-instruction process of Alpaca and generated a Chinese dataset of 1.5M samples by using 175 manually constructed Chinese seed instructions. It is an optimized and refined version of the BLOOMZ-7B1-mt model [16] and more suitable for Chinese culture and background knowledge due to the Chinese dataset. Chinese-alpaca [3] adapts English and translated Chinese Alpaca dataset based on LLaMA to support the bi-lingual environment. Some researchers [10] attempt to use a stronger teacher model to generate instruction data. Furthermore, Guanaco [17] adds external more languages (English Simplified Chinese, Traditional Chinese, Japanese, and Deutsch) entries with Alpaca dataset and is trained based on LLaMa to show the potential in a multilingual environment.

Conversation-based Tuning Inspired by the impressive results achieved by the Vicuna, training models through distilling data from user-shared chatGPT conversations has become a new trend. However, since Vicuna did not publicly release the dataset samples they used from ShareGPT, most subsequent models had to construct similar datasets by themselves. Based on existing open-

[13] https://yiyan.baidu.com/
[14] https://tongyi.aliyun.com/
[15] https://chat.sensetime.com/
[16] https://huggingface.co/databricks/dolly-v1-6b
[17] https://guanaco-model.github.io/

Algorithm 1: Post-translation for multi-lingual instruction

Input: Instruction Data $\mathbb{D}$, containing many instruction pairs $(\texttt{instruction}, \texttt{input}) \in \mathbb{D}$
Output: Translated multi-lingual triplets $(\texttt{instruction}', \texttt{input}', \texttt{output}') \in \mathbb{D}'$
foreach *instruction pair* **do**
  Sample another language `lang` based on the general language distribution;
  translate $(\texttt{instruction}, \texttt{input})$ into the sampled language:$(\texttt{instruction}', \texttt{input}')$;
  generate $\texttt{output}'$ based on the translated instruction $(\texttt{instruction}', \texttt{input}')$;
end

Algorithm 2: Generation of user-centered instructions

Input: None
Output: User-centered instruction quadruples $\{(\texttt{role}, \texttt{instruction}, \texttt{input}, \texttt{ouput})\}$
Step 1: Build a role set using a well-design ChatGPT prompt and manual efforts;
Step 2: Manually build some seed triplets $\{(\texttt{role}, \texttt{instruction}, \texttt{input})\}$ for each `role`.
Step 3: Generate more triplets using the seed triplets in an in-context few-shot fashion;
Step 4: **foreach** *instruction triplet* **do**
  predict its `output` based on the triplet $(\texttt{role}, \texttt{instruction}, \texttt{input})$.
end

sourced instruction datasets, Koala [18] utilized 30K conversation examples from ShareGPT with non-English languages removed and also incorporated the English question-answering dataset HC3 [5]. BAIZE [17] used a novel pipeline that generates a high-quality multi-turn conversation corpus containing 111.5K samples by having ChatGPT engage in a conversation with itself as the training dataset.

Despite the emergence of so many open-source models, the multilingual capabilities of current models are mostly inherited from the base models, and the multilingual training data used in the post-training stage is limited. This restricts the widespread use of the models worldwide, especially for people in countries with small languages.

3 Methodology

3.1 Dataset Construction

We collected our data from two sources: instruction data and user-shared conversations. We followed [14] to construct the instruction data and followed [2] to collect the user-share conversation data. To ensure the diversity of instructions and languages, we propose using a role-centric approach to construct instruction data and translate the instruction data to multiple languages. The details of the two types of data are shown as follows:

3.1.1 Instruction Data

We use three groups of instruction data as listed below.

- **Collected multi-lingual Instructions**: We used the 52K instructions collected in Alpaca [12], where each sample includes *instrction* (the task descriptions for large language models), *input* (the optional context for the instruction task), and *output* (the answers generated by large language models). For the *output*, we used the GPT-4-version ones released by [10], including both the English and Chinese answers.
- **Post-translated multi-lingual instruction**: We collect various sources of instructions that might be in different languages. Then. These multi-lingual instructions are translated into another language; the selection of the target language is based on the probability distribution of realistic languages. We acknowledge that translation might distort instructions, especially when instructions are language-specific. For example, a prompt `write a Chinese`

[18] https://bair.berkeley.edu/blog/2023/04/03/koala/

`Poet, like seven character quatrains` cannot be properly answered by another language. We leave dealing with the translation distortion as future work.

User-centered instructions Besides the above instructions, we also build some instruction data by ourselves. The main difference is that our instructions are driven by a given role (user) set. `role` could be either the executor or the submitter of a given instructor. It is possible to leave `role` empty to improve robustness.

3.1.2 Conversation Data

We mainly use ChatGPT-distilled conversation to adapt our language model for chatting. There are two sources of ChatGPT-distilled conversation data.

ShareGPT ShareGPT is a Chrome extension that allows users to conveniently share their ChatGPT conversations [19]. The data could be downloaded from https://huggingface.co/datasets/philschmid/sharegpt-raw.

Discord ChatGPT channel Discord is a free messaging software and digital platform for communities designed for gamers, educators, friends, and business people to communicate via chat, images, videos, and audio. The ChatGPT channel is the place for users to submit prompts in order to receive responses. ShareGPT is previously used by Vicuna [2] while Discord ChatGPT channel is shared in https://github.com/FreedomIntelligence/LLMZoo. Unlike Koala, we do not exclude non-English conversation data.

3.2 Dataset Statistics

Dataset	Samples	Turns	Avg. tokens/sample	Avg. tokens/turn
Alpaca-gpt4-zh	48,679	48,679	338.92	338.92
Alpaca-ml-gpt35-post-output	49,371	49,371	435.11	435.11
Alpaca-ml-gpt4-post-translation	51,398	51,398	543.39	543.39
Alpaca-gpt4-en	51,880	51,880	198.60	198.60
User-centered instructions	65,289	65,289	474.60	474.60
ShareGPT	189,643	654,912	1820.14	527.06
Discord	8,429	17,661	487.68	232.75
ALL	464,689	939,190	982.35	486.04

Table 2: The statistics on the components of our dataset.

Table 2 provides a comprehensive overview of the statistics for the various sub-datasets within our dataset. For each sub-dataset, we present the number of samples, the number of turns, the average tokens per sample, and the average tokens per turn. The overall statistics, encompassing all sub-datasets, are summarized in the row labeled "ALL." This information allows for a clear comparison and understanding of the various components within our dataset, which is crucial for evaluating the performance and characteristics of our models under investigation.

Figure 1 provides a visual representation of the language distribution in our dataset, emphasizing the top 15 languages. The short name of languages is from ISO 639-1 [20]. The data reveals that English and Chinese constitute the majority of the dataset, with a combined proportion of approximately 79.5%. The other 13 languages in the top 15 together make up the remaining 17.8%, demonstrating a diverse range of languages in the dataset.

3.3 Training details

The models are implemented in PyTorch using the Huggingface Transformers package[21]. We set the max context length to 2,048. We train the model with the AdamW optimizer, where the batch

[19] https://sharegpt.com/
[20] https://en.wikipedia.org/wiki/ISO_639-1
[21] https://github.com/huggingface/transformers

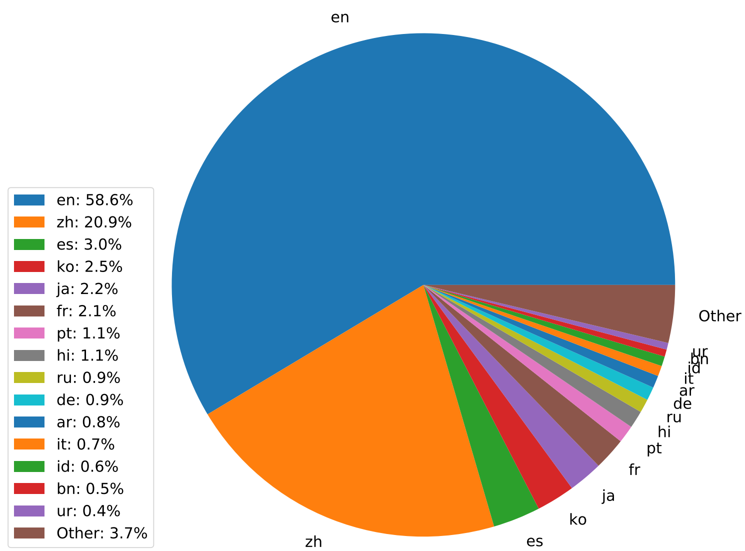

Figure 1: Language Distribution in Our Dataset: Top 15 Languages Represented out of 133.

size and the number of epochs are set to 256 and 3, respectively. The model using BLOOMZ as the backbone is called 'Phoenix' while that using LLaMA is called 'Chimera'.

4 Automatic Evaluations

4.1 Challenges

Assessing the performance of AI chatbots is a challenging task that requires a comprehensive evaluation of language coherence, comprehension, reasoning ability, and contextual awareness. Although [8] has elaborated an exhaustive study on evaluating LLMs on existing benchmarks, it may no longer be adequate. We summarize the existing **evaluation dilemma** for LLMs with the following three-fold challenges:

- **Not-blind**: Test data or similar data in benchmark might be seen by LLMs during pre-training of supervised fine-tuning.
- **Not-static**: The ground truth is not static, e.g., `tell a joke about Donald Trump`.
- **Incomplete testing path coverage**: Unlike path coverage of codes in Software Engineering, full coverage of testing cases is possible for inputs of LLMs; since user prompts are multi-faced.

To address these challenges, we present an evaluation framework based on GPT-4/GPT-3.5 Turbo API to automate chatbot performance assessment.

4.2 Evaluation protocol

Baselines To validate the performance of Phoenix, we first compare it with existing instruction-tuned large language models in Chinese and English, including GPT-3.5 Turbo, ChatGLM-6b, Wenxin, BELLE-7b-2m, Chinese-LLaMA-Alpaca 7b/13b, Vicuna-7b/13b[22]. Besides, we evaluate our models on more Latin (e.g., French, Spanish, and Portuguese) and non-Latin languages (e.g., Arabic, Japanese, and Korean) to show the multi-lingual ability, where we mainly compare our models with GPT-3.5 Turbo and a multi-lingual instruction-tuned model, Guanaco.

Metrics We conduct a pairwise comparison of the models' absolute performance following [2]. To achieve this, we request GPT-4 to rate the potential answers based on their helpfulness, relevance,

[22]We used the latest version of Vicuna models released in 04/13/2023.

Comparision	Zh Performance ratio	Zh Beat rate	En Performance ratio	En Beat rate
Phoenix vs. Phoenix (anchor)	100	50	100	50
Phoenix vs. **GPT-3.5 Turbo**	85.20	35.75	87.13	43.75
Phoenix vs. **ChatGLM-6b**	94.60	36.00	121.11	54.50
Phoenix vs. **Wenxin**	96.80	44.00	-	-
Phoenix vs. BELLE-7b-2m	122.70	65.25	-	-
Phoenix vs. Chinese-LLaMA-Alpaca-7b	135.30	75.75	-	-
Phoenix vs. Chinese-LLaMA-Alpaca-13b	125.20	74.50	-	-
Phoenix vs. Vicuna-7b	-	-	121.2	53.00
Phoenix vs. Vicuna-13b	-	-	90.92	46.00

Table 3: Benchmarking Phoenix in English and Chinese. Winner in each competition is in **bold**. Performance ratio is scored by GPT-4 API and Beat rate is calcuated using GPT 3.5 turbo.

accuracy, and level of detail. Apart from this, we provide a multi-dimension evaluation from several aspects, where we curated the definition of each metric and requested GPT-4 to rank each potential answer from different aspects separately. [2] assessed their model by testing it on a set of 80 questions spanning eight distinct categories. Additionally, we included two more categories, namely reasoning, and grammar, bringing the total number of questions to 100, spread across ten different categories.

4.3 Experimental results

We first conducted monolingual tests in both English and Chinese. We request GPT-4 to assign a quantitative score to each response on a scale of 10. Then we calculate the final score for each comparison pair (baseline, Phoenix) by averaging the scores obtained by each model across our 100 questions in the English and Chinese subsets.

Chinese We compared our model with the mainstream Chinese models, as shown in Table 3. It slightly underperforms Baidu-Wenxin and ChatGLM-6b, both are which are not fully open-source; ChatGLM-6b only provides model weights without training data and details. Phoenix underperforms ChatGLM-6B, which may be attributed to the fact that we did not conduct reinforcement learning from human feedback (RLHF) like ChatGLM-6B. This process is typically expensive and labor-intensive for democratizing training. However, Phoenix achieves comparable performance with Baidu Wenxinyiyan, a commercial and closed-source language model designed solely for Chinese. Given that Wenxinyiyan may have a larger model and is exclusive to the Chinese, this is a significant achievement for an open-source, democratized ChatGPT developed by academic institutions. It should be noted that neither ChatGLM-6B nor Wenxinyiyan significantly outperforms Phoenix, as evidenced by our statistical testing.

While on the other hand, compared to our Phoenix model, popular open-source Chinese models such as BELLE and Chinese-Alpaca can only attain 80% of our performance. Specifically, Chinese-Alpaca-7b can only attain 73.9% of our performance, Chinese-Alpaca-13 b 79.9%, and BELLE-7b-2m 81.5%. It demonstrates that although Phoenix is a multilingual LLM, it achieves SOTA performance among all open-source Chinese LLMs.

English We also compare Phoenix with Vicuna, ChatGPT, and ChatGLM-6B which are claimed to work in English. The two columns on the right side of Table 3 demonstrate the impressive performance in English of our Phoenix. Our model outperforms Vicuna-7b by 21.2% and ChatGLM-6b by 21.1%. It is important to note that Phoenix is a multilingual LLM. Therefore, compared to Vicuna-13b and ChatGPT, our model still lags behind them in terms of absolute performance in English.

Interestingly, Chimera, as a tax-free Phoenix, has impressed GPT-4 with 96.6% ChatGPT Quality, setting a new SOTA in open-source LLMs, see Fig. 2. Note that the evaluation is not rigorous enough. We will conduct human evaluations in the revision.

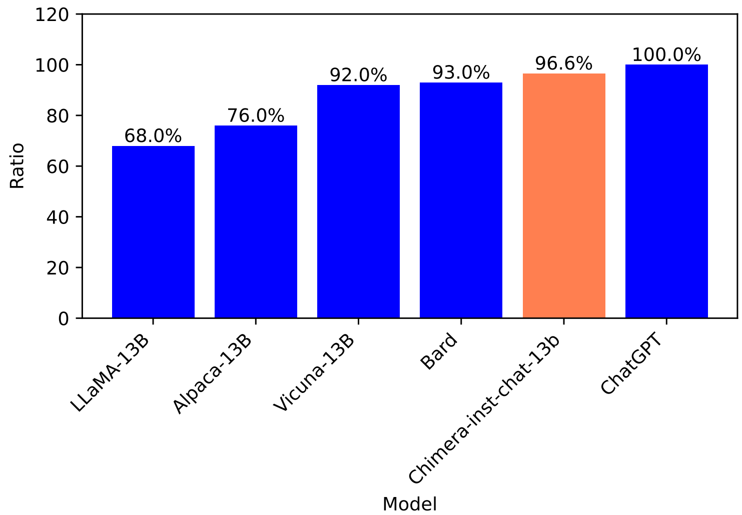

Figure 2: Relative Response Quality Assessed by GPT-4.

Langauge	Latin					Non-Latin		
	Fr	Es	Pt	It	De	Ar	Ja	Ko
Phoenix vs. Phoenix (Anchor)	50							
Phoenix vs. GPT-3.5 Turbo	41.75	34.00	32.75	19.00	10.50	30.25	25.50	7.75
Phoenix vs. Guanaco	95.50	91.50	94.00	84.0	37.00	97.00	79.00	64.00

Table 4: Benchmarking Phoenix in multiple languages.

4.4 Ablation study

Tab. 6 shows that adding instrution data is benefitial to a Chat-adapted LLMs; instructions achieves with 5%-6% relative improvement.

5 Conclusion

Among the ChatGPT democratization, this work extends LLM to non-Latin languages. The training philosophy is to combine both instruction data and conversation data in order to tame models to follow instructions in a chat fashion. The resulting multilingual LLM 'Phoenix' achieves the SOTA on fully open-source Chinese LLMs. In non-Latin languages, Phoenix outperforms existing open-source LLMs, including Vicuna-13b and Guanaco. Notably, our Latin version of Phoenix called 'Chimera' impresses GPT-4 with 96.6% ChatGPT Quality, setting a new SOTA in open-source LLMs. We believe the proposed models could largely benefit people who could not legally use ChatGPT or related tools, therefore making AI open and equal again.

Limitations

Our goal in releasing our models is to assist our community in better replicating ChatGPT/GPT4. We are not targeting competition with other competitors, as benchmarking models is a challenging task. Our models face similar models to those of ChatGPT/GPT4, which include: 1) Lack of common sense; 2)Limited knowledge domain; 3) Biases; 4) Inability to understand emotions; and 5) Misunderstandings due to context.

Model	Pt	De	It	Es	Fr
Chimera-13b vs. Chimera-13b (anchor)			50		
Chimera-13b vs. GPT-3.5 Turbo	40.67	45.25	47.67	52.71	54.12
Chimera-13b vs. Guanaco	87.50	93.00	84.80	95.50	96.00

Table 5: Benchmarking Chimera in multiple Latin languages.

	Phoenix vs. GPT-3.5 Turbo (Zh)	Chimera vs. GPT-3.5 Turbo (En)
Conversations	34.00	39.75
+ Instruction	35.75 ↑ 5.1%	42.25 ↑ 6.3%

Table 6: Ablation study on the instruction data.

More importantly, the used evaluation in this work is not rigorous enough. Therefore, we will add human evalution in a few days during revision. We only make our models accessible inside our university and SRIBD, see `http://10.26.1.135:7860/`.

References

[1] Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF*, 2019. 2

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2, 3, 4, 5, 6, 7, 8

[3] Yiming Cui and Ziqing Yang. Chinese llama and alpaca llms. `https://github.com/ymcui/Chinese-LLaMA-Alpaca`, 2023. 4

[4] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. 2

[5] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. 5

[6] Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*, 2023. 2, 4

[7] Ziang Leng, Qiyuan Chen, and Cheng Li. Luotuo: An instruction-following chinese language model, lora tuning on llama. `https://github.com/LC1332/Chinese-alpaca-lora`, 2023. 4

[8] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022. 7

[9] 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. 2

[10] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 4, 5

[11] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. 4

[12] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 4, 5

[13] 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. Llama: Open and efficient foundation language models, 2023. 2, 3

[14] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 5

[15] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2, 4

[16] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain

Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. 2, 3, 4

[17] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023. 4, 5

[18] 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. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

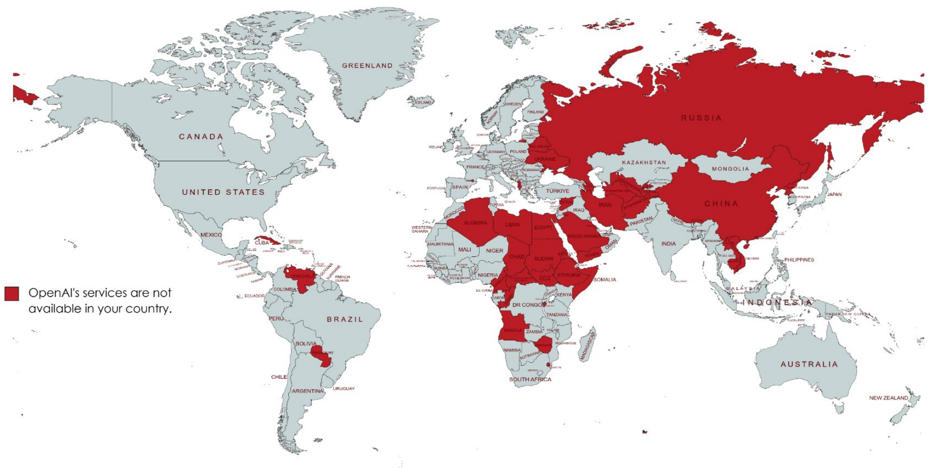

Figure 3: Being Red indicates that in the area ChatGPT is unavaibale. Source of the picture: https://i.imgur.com/2fF3Xlh.png

A ChatGPT is inaccessible to some countries

As seen in Fig. 3, ChatGPT is unavaibale to many countries, most of which are developing countries.