BlueLM: An Open Multilingual 7B Language Model

BlueLM Team
vivo AI Lab

Abstract

We introduce ***BlueLM-7B***, a multilingual foundation language model with 7B parameters, trained from scratch on a large corpora with 2.6 trillion tokens. To support efficient training, we develop a distributed training system called *vivolm*, which is specifically designed and optimized for large language models. We also propose using pre-training loss prediction to identify unexpected training process in time. Based on BlueLM-7B, we employ supervised fine-tuning to enhance the capability of understanding human intentions, resulting in two chat-specific models ***BlueLM-7B-Chat*** and ***BlueLM-7B-Chat-32K*** where the latter is designed to handle long-context tasks. BlueLM models are evaluated on 11 benchmarks from *exam*, *mathematical*, *coding*, *reasoning* and *safety* aspects. BlueLM-7B outperforms the similarly sized models on exam benchmarks by a large margin and shows competitive efficacy on math, coding and reasoning benchmarks. BlueLM-7B-Chat demonstrates significant improvements across all benchmarks except MMLU and Gaokao, and BlueLM-7B-Chat-32K achieves competitive results compared to other long-context models. The model also obtains a relative 10.48% improvement in safety performance by using reinforcement learning from human feedback (RLHF). All the models are public available: `https://github.com/vivo-ai-lab/BlueLM`.

1 Introduction

In recent years, the field of natural language processing has witnessed remarkable progress. Model parameters have grown from millions to billions and even trillions. Model structures have shifted from CNN [1] or LSTM [2] to transformer [3] , with the most promising transformer structure moving from encoder-only [4, 5] to decoder-only [6, 7, 8, 9]. Particularly, the large language model ChatGPT [9] has demonstrated strong instruction-following abilities and achieved unprecedented performance in a wide range of domains and tasks, including reading comprehension, translation, summarization, reasoning, programming, and mathematical problem-solving. This offers a promising pathway to developing an artificial general intelligence (AGI) system.

Among the leading large language models, such as ChatGPT [9], GPT-4 [10], PaLM [11], Claude2 [12], most of them are closed-source. LLaMA [13, 14] is a series of open-sourced models that offer comparable performance to closed-source ones, which benefits both the academic and industry communities for further exploration and development. However, LLaMA has limited capabilities in Chinese due to the less pre-training data of Chinese language. Therefore, there has been a rapid emergence of large language models [15, 16, 17, 18, 19], exhibiting proficiency in Chinese language comprehension. For instance, ChatGLM2-6B [16] is an open source chat language model trained on 1.4 trillion bilingual tokens, it also employs a multi-query attention mechanism and extends context length of base model to 32K. InternLM [17] is trained on 1.6T tokens (mainly English and Chinese) by using a multi-phase progressive pre-training scheme to make the entire training process more controllable. Baichuan2-7B [19] is trained on 2.6T tokens utilizing NormHead and Max-z loss to stabilize training and enhance the model performance. Qwen [18] is trained on 3T tokens with NTK-aware interpolation to extend context-length and incorporates two attention mechanisms: LogN-Scaling and window attention.

In this work, we aim to develop a relatively compact multilingual foundation language model, while still achieving highly competitive performance on challenging tasks. We call this model ***BlueLM***, where *Blue* is derived from the representative color of *vivo*.

Specifically, ***BlueLM-7B*** is a multilingual foundation language model with 7B parameters, trained from scratch on a large corpora with 2.6 trillion tokens from multiple sources, including webpages, academic papers, E-books, source code, etc. We apply a few key modifications on the model architecture to enhance the performance like Rotary Positional Embeddings (RoPE) [20], RMSNorm [21] and SwiGLU [22] activation function. To support the training of BlueLM, we develop an efficient distributed training system *vivolm*, which is specifically designed and optimized for large language model training. Furthermore, we propose to forecast the pre-training loss using scaling law [23], which aids in identifying unexpected training process in time, thereby saving computational resources.

Based on the foundation model, we further enhance the capability of understanding human intentions by using supervised fine-tuning, resulting in a chat-specific version denoted as ***BlueLM-7B-Chat***. To handle long-context tasks, we develop the corresponding long-context version models with 32K input tokens, denoted as ***BlueLM-7B-32K*** and ***BlueLM-7B-Chat-32K***. Reinforcement learning from human feedback (RLHF) [24] is also employed on the chat model for safety improvement. All the models are made public available.

We evaluated BlueLM on 4 different types of benchmarks, comprising a total of 11 individual benchmarks: 1) comprehensive exam benchmarks, including MMLU [25], C-Eval [26], CMMLU [27], Gaokao [28] and AGIEval [29]; 2) specific capabilities (e.g, math/general reasoning, coding) benchmarks, including GSM8K [30], MATH [31], HumanEval[32] and BBH [33]; 3) long-context benchmark LongBench [34]; 4) an in-house benchmark for safety evaluation. The key results are summarized below:

- On exam benchmarks, *BlueLM-7B* not only significantly outperforms the similarly sized models (e.g, LLaMA2-7B [14], ChatGLM2-6B (base) [16], Baichuan2-7B [19]), but also exceeds ChatGPT [9] on all benchmarks, with the exception of MMLU.
- On math, coding and reasoning benchmarks, *BlueLM-7B* shows competitive efficacy with the similarly sized models.
- *BlueLM-7B-Chat* demonstrates significant improvements across all benchmarks except MMLU and Gaokao. For instance, 24.7% on GSM8K, 23.9% on BBH, and 7.2% on MATH. And it outperforms or perform competitively with all similarly sized models.
- On LongBench benchmark, *BlueLM-7B-Chat-32K* outperforms LongChat-v1.5-7B-32K [35] and Vicuna-v1.5-7B-16K [36], and achieves competitive results compared to ChatGLM2-6B-32K [15].
- With the inclusion of RLHF [24], *BlueLM-7B-Chat* obtains a relative enhancement of 10.48% in overall safety performance, while preserving its generality and versatility.

Although BlueLM-7B has attained promising results across several benchmarks, it is worth noting that there remains a noticeable performance gap between small-sized models and larger GPTs [9, 10] on certain tasks, such as mathematical reasoning and code generation. Towards the goal of achieving AGI, there is still a long way ahead, and we hope releasing BlueLM as an open-source model can inspire the communities for further advanced research in this ever-evolving field.

2 Pre-training

This section introduces the pre-training procedure for BlueLM. We firstly overview the pre-training data and data pre-processing methods (Section 2.1), followed by a description of model architecture (Section 2.2). Next, we introduce *vivolm*, a high efficient distributed system for training large language models in Section 2.3. Finally we elaborate on the pre-training loss prediction technology.

2.1 Pre-training Data

BlueLM-7B was trained on a composite collection of diverse sub-datasets, which totally consist of 2.59 trillion tokens (equating to 8.72 terabytes of plain text). These sub-datasets were sourced from a wide-ranging array of publicly available webpages, academic papers, E-books, source code, etc. The utilization of these varied sources is able to provide comprehensive coverage of different domain knowledge. For further details regarding the pre-training data, please refer to Table 1.

To build a large-scale high-quality pre-training dataset, we developed innovative data filtering and deduplication strategies, including:

- *rule-based filtering*: remove low-quality content based on rules that we derived from manual annotations of the data;
- *fuzzy deduplication*: apply MinHash [37] on a Spark cluster with over 1000 CPUs to remove similar documents that could impede the training process;
- *exact deduplication*: remove duplicate substrings that occurred verbatim in more than one example through the usage of a suffix array [38].

By combining these strategies, we were able to maximize the quality and diversity of our pre-training dataset, ensuring that BlueLM-7B had access to the highest quality material for training.

The conventional approach to incorporating multiple documents into a single training sequence usually concatenates them until reaching the maximum token length. However, this method undermines the coherence of the original documents. To make each document under the maximum length intact during training, we integrated 1.5% padding tokens for training samples. Our experiments yield compelling evidence that the model benefits significantly from learning within the context of each intact document.

2.2 Model Architecture

Our model architecture is inspired by LLaMA [13], a decoder-only structure based on Transformer [3], with a few key modifications to enhance its performance. Firstly, we employ Rotary Positional Embeddings (RoPE) [20] in lieu of absolute positional embeddings to improve the model's ability to encode sequential information. Additionally, the use of RMSNorm [21] as pre-layer normalization and the inclusion of embedding layer normalization [39] help to stabilize the model's training process. In place of ReLU, we implement the SwiGLU [22] activation function to further bolster performance.

Specifically, ***BlueLM-7B*** is comprised of 32 transformer layers, each with 32 attention heads and a head dimension of 128. The intermediate size in feed-forward network is set to 11008.

For pre-training BlueLM-7B, we employed a context length of 2048 tokens. For long-context tasks, we fine-tuned BlueLM-7B with 32768 tokens, denoted as ***BlueLM-7B-32K***. In addition, for BlueLM-7B-Base-32K, we further investigated NTK-RoPE-mixed embeddings [40] and positional interpolation (PI) embeddings [41] and experimentally found that the former achieved a 2.2% higher average score on LongBench [34] dataset.

Tokenizer. We generated a tokenizer for our model by employing the byte-pair encoding (BPE) [42] algorithm from the SentencePiece [43] toolkit. To ensure maximum flexibility in handling numerical values, we split all numbers into individual digits. For cases where unknown UTF-8 characters were encountered, we resorted to a fallback method of decomposing them into bytes. The final token vocabulary comprises a total of 100,096 entries.

2.3 *vivolm*: An Efficient Training System

When dealing with models with billions of parameters, conventional training methodologies prove insufficient. To address this, we developed an efficient training system, *vivolm*, capable of leveraging multiple parallelism techniques such as data parallelism [44], tensor parallelism [45], pipeline parallelism [46], and the zero redundancy optimizer (ZeRO) [47]. For even greater training efficiency, we utilize flash-attention [48] calculation optimization and Remote Direct Memory Access (RDMA) network optimization. We were able to complete training over our 2.59T tokens dataset in about 26

Dataset	Tokens (B)	Percentages
Massive web pages	1349.1	52.0%
Source code	633.1	24.4%
Wikipedia	186.0	7.2%
Academic papers	114.6	4.4%
E-books	92.1	3.6%
Others	217.1	8.4%
Sum	2592.0	100%

Table 1: Overview of our pre-training data.

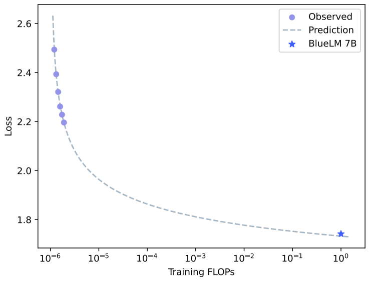

Figure 1: Loss prediction of BlueLM-7B.

days. After fixing the performance jitter issue in the early stage, the throughput of *vivolm* reached an impressive 3150 tokens/sec/GPU.

To promote training stability, *vivolm* has been designed with a range of features to provide real-time monitoring of the progress and health status of training. This includes automatic detection and alerts for any abnormalities in key metrics. Furthermore, the system's dataset slicing function ensures that only one slice is loaded at a time, improving startup speed and enabling the flexible adjustment of unused slice data. Additionally, *vivolm* integrates a sub-system that supports asynchronous checkpointing, enabling large model checkpoints to be written asynchronously without disrupting the training process. This feature further ensures that the latest checkpoint can be saved before the training process exits abnormally.

For training BlueLM-7B model, we employed AdamW [49] optimizer with the following hyper-parameters: β_1 = 0.9, β_2 = 0.95, ϵ = 10^{-5}. Instead of cosine decay, we use a two-stage learning rate schedule inspired from [50], which shows a 2.2% lower loss. The peak learning rate was set to 3×10^{-4} with 2000 warmup steps. We set weight decay and gradient clipping to 0.1 and 1.0 respectively, and dynamic gradient clipping is also applied on outlier gradient elements to further reduce loss spikes.

2.4 Pre-training Loss Prediction

As the pre-training of large models is becoming increasingly expensive, it is essential to fit a scaling law [23] to predict the pre-training loss before training large models, which can help identify unexpected training process in time. We leverage the cyclic cosine decay [51] strategy to forecast the training-from-scratch pre-training losses, using various levels of training flops with the same model configurations and training data.

Recall that the scaling law formula [23]: $\mathcal{L}_C = a \times C^b + \mathcal{L}_\infty$, where the power-law scaling term $a \times C^b$ is the reducible loss and $\mathcal{L}_\infty$ is the irreducible loss, C are training flops and $\mathcal{L}_C$ are final loss of the model in that flops, a and b are the parameters to be fitted. We propose a two-stage fitting method: 1) we apply the cyclic cosine decay [51] to fit $\mathcal{L}_C = a \times C^b$, and calculate $\mathcal{L}_{C \to \infty}$ as the approximate value for $\mathcal{L}_\infty$; 2) Given $\mathcal{L}_\infty$ above, we fit $\mathcal{L}_C = a \times C^b + \mathcal{L}_\infty$ again using [51].

The experimental results demonstrate that this approach is not only effective but also efficient: it achieves an error rate of less than 2% for loss prediction of the 7B model. Meanwhile, it avoids training multiple small-sized models as in [19]. The predicted and the final loss are shown in Fig 1.

3 Supervised Fine-tune

To refine our model's capability of understanding human intentions, we undertook a supervised fine-tuning phase, gathering high-quality third-party instructed data from well-recognized sources such as BELLE [52] and Flan 2022 [53]. More importantly, we established a dedicated team of trained annotators to carefully evaluate the data, ensuring that the information we selected from the open source datasets was clean and free of any serious issues related to data quality or security.

Despite these efforts, we initially experienced limitations in certain general capabilities when training on open source data alone, which we attributed to a lack of diversity within the open source datasets, as highlighted by LLaMA2 [14]. To address this issue, we designed a measurement system for the model, which comprised of 19 second-level dimensions and more than 200 third-level dimensions that primarily focused on language understanding, language generation, logic and mathematics, coding competencies, knowledge-based questions, and security components. Then under the guidance of this system, we enhanced data diversity by populating more instructed data, using several different methods like manual production, self-instruct [54], and evol-instruct [55]. The annotators conducted a secondary quality inspection, ensuring the data met our stringent quality and security standards.

We introduce two supervised fine-tuned models based on 2048 and 32768 token length, denoted as ***BlueLM-7B-Chat*** and ***BlueLM-7B-Chat-32K*** respectively, with the latter specifically designed for long-context tasks. For fine-tuning, we concatenated the instructed data into 2048 tokens or 32768 tokens, separated by a special token.

To prevent any extreme sampling within a batch, we randomly sampled a single batch based on the task type, ensuring a consistent batch distribution comparable to the original data. For multi-round conversations training, we adjusted the attention mask to ensure that information was only included from the previous round.

We used the normalized target-only loss [14] and trained the model for 3 epochs with a cosine learning rate schedule, where the learning rate gradually decreased from 10^{-5} to 10^{-6} by the end of training. The batch size was set to 32.

4 Safety

Data Construction. In order to ensure that the output of the fine-tuned large model meets the safety standards, it is essential to collect, produce, and filter relevant training data to make the model free from toxicity. Our safety data is manual annotated and quality-checked, and adheres to the following guidelines:

- Maintain national security and social stability, ***NOT*** promote terrorism, extremism, ethnic hatred and discrimination, violence, obscenity, false and harmful information prohibited by laws and regulations.
- Maintain a neutral attitude, ***NOT*** contain discrimination based on ethnicity, religion, nationality, geography, gender, age, profession, health, etc.
- Respect intellectual property, commercial ethics, and the confidentiality of business secrets.
- Ensure the accuracy, completeness, fairness, reliability, and security of the data.

We also conduct a series of strategies to filter the collected safety data, including removing political sensitive data [56], filtering out low-quality, harmful, and biased data [57], eliminating maliciously forged attack data [58] (e.g, poisoned samples and backdoor samples), and discarding noise and erroneous data.

Specially, we create a human preference dataset for reinforcement learning from human feedback (RLHF) [24]. To ensure the data diversity, we use 1) adding positive or negative instructions to the prompts, guiding the models to generate responses with or without toxicity; 2) sampling responses with different temperatures from different-sized models. The annotators were then requested to rank the responses of each prompt according to our scoring criteria (see Appendix).

Alignment. To further enhance model's safety, we adopt the RLHF approach - similar to InstructGPT [59] - based on the supervised fine-tuning (SFT). The following is a brief overview of the process:

- Train a reward model (RM) to align with human preferences regarding safety. Besides the original RM loss [59], we introduced a penalty factor to quantify the significance of each training instance, which was computed as the disparity between the human-labeled scores of two responses. We found this technique can accelerate model convergence and enhance performance, giving a 75% accuracy on the safety evaluation set.
- Fine-tune the SFT model using proximal policy optimization (PPO) [60] with the feedback from the RM. To mitigate the performance deterioration in certain abilities, we incorporated the SFT gradients into the PPO gradients. We also found using warmed-up critic model in PPO helps to improve the training stability by providing better advantage estimation.

The PPO training data was constructed in the same manner as RM, and it only needs prompts.

5 Evaluation

In this section, we first introduce benchmarks (Section 5.1), baselines (Section 5.2) and evaluation settings (Section 5.3). We report the results of BlueLM-7B on standard benchmarks from Section 5.4 to Section 5.7. We present the results of BlueLM-7B-Chat in Section 5.8. Finally, the results on the safety benchmark are given in Section 5.9.

5.1 Benchmarks

Given the versatility of language models, their evaluation must be conducted from various perspectives. In this work, we evaluated BlueLM-7B using 9 benchmarks for 4 specific types of model's capabilities, including *exams*, *math*, *coding* and *reasoning*. These benchmarks are complementary and combining them together provides a more complete and solid evaluation. The details are listed below.

Exams The use of comprehensive exams, which are designed to assess the models' overall performance across various tasks and domains, takes into account the potential of models for general intelligence similar to human beings. We use five exam benchmarks, including:

- **MMLU** [25]: A multi-task benchmark consists of multiple-choice questions covering 57 domain of knowledge, including elementary mathematics, American history, computer science, law, economics, etc.
- **C-Eval** [26]: The first comprehensive Chinese benchmark consists of about 14K multiple-choice questions in 52 subjects, covering mathematics, physics, chemistry, biology, history, computer science, etc.
- **CMMLU** [27]: Another general Chinese benchmark covering 67 subjects, designed for evaluating the knowledge and reasoning abilities of language models within the context of the Chinese language and culture.
- **Gaokao** [28]: An Chinese benchmark based on Chinese college entrance examination to evaluate the language understanding and logical reasoning abilities of language models.
- **AGIEval** [29]: A human-centric benchmark developed by Microsoft, which is specifically designed to evaluate the general abilities like human cognition and problem-solving, consisting of 19 task sets derived from various exams in China and the Unite States.

Math We use 2 benchmarks to evaluate the capabilities of mathematical understanding and quantitative reasoning:

- **GSM8K** [30]: A mathematical dataset contains 8.5K high-quality linguistically diverse grade school math word problems.
- **MATH** [31]: A new dataset collects 12,500 challenging problems from high school mathematics competitions.

Coding We test the coding capabilities by using **HumanEval** [32] benchmark, a Python code dataset of consisting of 164 code problems to evaluate different aspects of programming logic.

Reasoning The general reasoning capabilities is evaluated on **BBH** [33] benchmark, a subset of BIG-Bench [61] benchmark, containing 23 challenging tasks that the language models did not outperform the average human-rater.

Considering the diverse range of tasks, we have integrated open-source evaluation frameworks, such as lm-evaluation-harness [62] and OpenCompass [63], into our in-house implementations to ensure a fair comparison with other models.

5.2 Baselines

We mainly compare BlueLM-7B with three other pretrained language models: LLaMA2-7B [14], Baichuan2-7B [19], and ChatGLM2-6B (base) [16]. These models were chosen because of their

Table 2: Evaluation settings of different benchmarks.

Setting	MMLU	C-Eval	CMMLU	Gaokao	AGIEval	GSM8K	MATH	HumanEval	BBH
N-shot	5	5	5	0	0	4	4	0	3
CoT [64]	✗	✗	✗	✗	✗	✓	✓	✗	✓

Table 3: Results on exam benchmarks where **bold** indicates the best result and underline indicates the second-best one. * denotes results derived from official websites.

Model	MMLU	C-Eval	CMMLU	Gaokao	AGIEval
GPT-4 [10]	**86.4**	**69.9**	**71.2**	**72.3**	**55.1**
ChatGPT [9]	69.1	52.5	53.9	51.1	39.9
LLaMA2-7B [14]	45.3	32.5	31.8	19.1	21.4
ChatGLM2-6B (base)* [16]	47.9	51.7	-	-	-
Baichuan2-7B [19]	54.7	56.3	57.0	34.8	34.8
BlueLM-7B	55.2	67.5	66.6	58.9	43.4

comparable size and open source nature, enhancing the reproduciblity of the evaluation results. In addition, we also include the results of GPT-4 [10] and ChatGPT [9], the state-of-the-art language models, for reference.

- **LLaMA2-7B** [14] is an open source language model trained on 2 trillion tokens. The context length is 4096.
- **ChatGLM2-6B (base)** [16] is an open source chat language model trained on 1.4 trillion bilingual tokens.
- **Baichuan2-7B** [19] is an open source language model trained on 2.6 trillion tokens with a focus on both English and Chinese, which achieves strong performance on several benchmarks.
- **ChatGPT** [9] is a chat language model developed by OpenAI, derived from a GPT-3 [8] variant and InstructGPT [59] approach. Owing to its superior performance, it has gained significant attention and emerged as a widely-used baseline for language models.
- **GPT-4** [10] represents a state-of-the-art multi-modal language model developed by OpenAI, displaying a superior capabilities of complex reasoning over ChatGPT [9].

5.3 Evaluation Settings

The evaluation settings are detailed in Table 2. We follow the conventional settings to use zero-shot or few-shot evaluation, and apply Chain-of-Thought (CoT) [64] on GSM8K, MATH and BBH benchmarks.

5.4 Exam Results

The overall results on exam benchmarks are shown in Table 3. While GPT-4 [10] achieved the highest scores overall, BlueLM-7B far surpasses ChatGPT [9] on all benchmarks, with the exception of MMLU, and attains a comparably high score to GPT-4 [10] on Chinese evaluation benchmarks like C-Eval and CMMLU. Remarkably, when compared to the similarly sized models (e.g, LLaMA2-7B [14], ChatGLM2-6B (base) [16], Baichuan2-7B [19]), BlueLM-7B outperforms them by a significant margin, for instance, it exceeds Baichuan2-7B [19] by 24% on Gaokao and 11% on C-Eval. Note that ChatGLM2-6B (base) [16] is not publicly available, we derive its results from their official website.

5.4.1 MMLU

The detailed results of each category on MMLU benchmark are presented in Table 4. MMLU is a extensively-used English benchmark with multi-choice questions of various levels of difficulty, spanning from basic high school level to highly specialized expert-level. Overall, BlueLM-7B showcases the best performance among the similarly sized models. Considering each category,

Table 4: Results on MMLU benchmark where **bold** indicates the best result and underline indicates the second-best one.

Model	**Humanities**	**STEM**	**Social Science**	**Others**	**Average**
LLaMA2-7B [14]	51.6	38.0	52.1	50.1	45.3
ChatGLM2-6B (base) [16]	-	-	-	-	47.9
Baichuan2-7B [19]	**60.3**	45.1	62.1	56.4	54.7
BlueLM-7B	58.2	**45.9**	**62.8**	**58.7**	**55.2**

Table 5: Results on AGIEval benchmark where **bold** indicates the best result and underline indicates the second-best one.

	LLaMA2-7B [14]	Baichuan2-7B [19]	BlueLM-7B
AQuA-RAT	19.3	22.8	**24.4**
MATH	1.2	1.6	**4.9**
LogiQA (English)	32.6	35.5	**37.5**
LogiQA (Chinese)	27.8	36.7	**42.6**
JEC-QA-KD	13.0	14.9	**25.2**
JEC-QA-CA	15.4	15.4	**23.2**
LSAT-AR	**23.0**	20.0	21.3
LSAT-LR	25.3	32.6	**39.4**
LSAT-RC	24.2	36.1	**48.0**
SAT-Math	23.2	24.6	**30.0**
SAT-English	33.5	57.3	**62.1**
SAT-English (w/o Psg.)	26.2	37.9	**39.8**
GK-Cn	24.0	44.7	**62.2**
GK-En	29.7	68.3	**76.1**
GK-geography	28.6	65.3	**71.9**
GK-history	23.4	67.2	**80.9**
GK-biology	20.5	47.1	**73.8**
GK-chemistry	19.3	35.8	**58.9**
GK-physics	15.8	30.9	**53.3**
GK-Math-QA	23.1	27.9	**30.8**
GK-Math-Cloze	0.0	**7.6**	5.1
Average	21.4	34.8	**43.4**
Average (GK)	20.5	43.9	**57.0**

BlueLM-7B attains the highest score in *STEM*, *Social Science* and *Others*, while slightly trailing behind Baichuan2-7B [19] in *Humanities*.

5.4.2 AGIEval

We show the detailed performance of each subject on AGIEval benchmark in Table 5. AGIEval is a human-centric benchmark, derived from 20 official, public, and high-standard admission and qualification exams in China and the Unite States, which is specifically designed to evaluate the general abilities like human cognition and problem-solving. Among the similarly sized models, BlueLM-7B achieves the highest scores in 19 out of 21 subjects. In comparison to Baichuan2-7B [19], BlueLM-7B improves the average results by about 8% (44.3 vs. 34.8). Particularly, with respect to the average of 9 sets of Chinese college entrance exams (*Average (GK)*), BlueLM-7B outperforms Baichuan2-7B [19] by 13% (57.0 vs. 43.9).

5.4.3 C-Eval

The detailed results of each category on C-Eval benchmark are presented in Table 6. C-Eval is a comprehensive benchmark designed to evaluate large language models in the Chinese context, which consists of 13948 multi-choice questions spanning 52 diverse disciplines and four difficulty levels

Table 6: Results on C-Eval benchmark where **bold** indicates the best result and underline indicates the second-best one.

Model	Humanities	STEM	Social Science	Others	Average
LLaMA2-7B [14]	37.0	28.8	39.0	29.1	32.5
ChatGLM2-6B (base) [16]	-	-	-	-	51.7
Baichuan2-7B [19]	64.4	47.1	65.6	55.6	56.3
BlueLM-7B	**71.0**	**62.2**	**77.0**	**65.1**	**67.5**

Table 7: Results on CMMLU benchmark where **bold** indicates the best result and underline indicates the second-best one.

Model	Humanities	STEM	Social Science	Others	Average
LLaMA2-7B [14]	32.2	29.0	32.8	33.4	31.8
Baichuan2-7B [19]	61.2	43.5	62.1	61.2	57.0
BlueLM-7B	**70.9**	**55.1**	**70.1**	**70.9**	**66.6**

ranging from middle school to professional level. The category hierarchy is analogous to that of MMLU. It can be observed that BlueLM-7B outperforms the other models in all categories by a large margin. Notably, its overall score surpasses Baichuan2-7B [19] by over 11%, revealing its exceptional ability in understanding Chinese language.

5.4.4 CMMLU

The detailed results of each category on CMMLU benchmark are shown in Table 7. CMMLU is another Chinese benchmark covering 67 subjects, incorporating disciplines that demand computational expertise, such as physics and mathematics, as well as subjects falling within the domains of humanities and social sciences. The category hierarchy is analogous to that of MMLU. BlueLM-7B also performs impressively well on this benchmark, for instance, it exceeds Baichuan2-7B [19] by roughly 10% in all categories, as well as in the average score. It also outperforms LLaMA2-7B [14] by approximately 35%.

5.4.5 Gaokao

The detailed results of each subject on Gaokao benchmark are shown in Table 8. Gaokao is constructed based on the China National College Entrance Examinations (as known as Gaokao in Chinese) from 2010 to 2022, including 1781 objective questions and 1030 subjective questions. The goal of this benchmark is to evaluate the effectiveness of large language models in addressing domain-specific tasks, akin to those that humans are capable of performing. The results show that BlueLM-7B is superior over Baichuan2-7B [19] more than 20% in the overall performance. And the significant improvements can be observed in many subjects, including *History_MCQs*, *Physics_MCQs*, *Chemistry_MCQs*, *English_MCQs*, and *English_Reading_Comp*.

5.5 Math Results

The detailed results on math benchmark are presented in Table 9. Large language models often encounter challenges when attempting to engage in multi-step quantitative reasoning tasks [30]. We employ GSM8K (grade school math problems) and MATH (challenging competition math problems) benchmarks to evaluate the model's proficiency of math multi-step reasoning. As we can see, BlueLM-7B achieves competitive performances when compared with the similarly sized models. Nonetheless, it is still far behind ChatGPT [9] and GPT-4 [10], and clearly falls short of human performance levels.

Table 8: Results on Gaokao benchmark where **bold** indicates the best result and underline indicates the second-best one.

	LLaMA2-7B [14]	Baichuan2-7B [19]	BlueLM-7B
2010-2022_Math_II_MCQs	22.0	28.0	**35.8**
2010-2022_Math_I_MCQs	26.6	28.5	**34.1**
2010-2022_History_MCQs	25.4	55.1	**83.6**
2010-2022_Biology_MCQs	23.3	46.0	**77.3**
2010-2022_Political_Science_MCQs	22.8	62.8	**84.4**
2010-2022_Physics_MCQs	2.3	3.9	**36.7**
2010-2022_Chemistry_MCQs	22.6	31.5	**60.5**
2010-2013_English_MCQs	26.7	44.8	**81.9**
2010-2022_Chinese_Modern_Lit	3.5	26.4	**47.1**
2010-2022_English_Fill_in_Blanks	6.3	18.5	**54.5**
2012-2022_English_Cloze_Test	12.3	20.8	**27.7**
2010-2022_Geography_MCQs	24.2	56.8	**69.5**
2010-2022_English_Reading_Comp	11.1	6.2	**52.6**
2010-2022_Chinese_Lang_and_Usage_MCQs	12.5	22.5	**41.3**
Average	19.1	34.8	**58.9**

Table 9: Results on math benchmarks. * denotes results derived from official websites.

Model	**GSM8K**	**MATH**
GPT-4 [10]	91.4	45.8
ChatGPT [9]	78.2	28.0
LLaMA2-7B [14]	16.7	3.3
ChatGLM2-6B (base)* [16]	32.4	-
Baichuan2-7B [19]	24.6	5.4
BlueLM-7B	27.2	6.2

5.6 Coding Results

In order to assess the model's capability of generating Python code, we present the pass@1 scores on the widely-utilized benchmark HumanEval, which are shown in Table 10. BlueLM-7B outperforms LLaMA2-7B [14] and Baichuan2-7B [19]. However, it is found that all the small-sized (e.g, 7B) models exhibit limited performance compared to GPT-4 [10] or ChatGPT [9] in generating code, underscoring the importance of training larger models for coding tasks.

5.7 Reasoning Results

To evaluate the capability of general reasoning, we use BIG-Bench Hard (BBH) benchmark, which is a subset of BIG-Bench [61] containing 23 challenging tasks that the language models did not surpass the average human-rater. The results are presented in Table 10. BlueLM-7B shows competitive efficacy with Baichuan2-7B [19] and significantly outperforms LLaMA2-7B [14] and ChatGLM2-6B

Table 10: Results on code & reasoning benchmarks. * denotes results derived from official websites.

Model	**HumanEval (coding)**	**BBH (reasoning)**
GPT-4 [10]	74.4	86.7
ChatGPT [9]	73.2	70.1
LLaMA2-7B [14]	12.8	38.2
ChatGLM2-6B (base)* [16]	-	33.7
Baichuan2-7B [19]	17.7	41.8
BlueLM-7B	18.3	41.7

Table 11: Results of BlueLM-7B-Chat on each benchmark where **bold** indicates the best result and underline indicates the second-best one.

	MMLU	C-Eval	CMMLU	Gaokao	AGIEval	GSM8K	MATH	HumanEval	BBH
LLaMA2-7B [14]	46.2	34.5	31.5	16.1	28.5	26.3	3.9	12.2	35.6
ChatGLM2-6B [16]	45.9	52.6	49.3	46.4	39.0	28.8	6.5	11.0	32.7
Baichuan2-7B-Chat [19]	52.8	55.6	54.0	39.7	35.3	32.8	6.0	13.4	35.8
BlueLM-7B	**55.2**	67.5	66.6	**58.9**	**43.4**	27.2	6.2	18.3	41.7
BlueLM-7B-Chat	50.7	**72.7**	**74.2**	48.7	**43.4**	**51.9**	**13.4**	**21.3**	**65.6**

Table 12: Results of BlueLM-7B-Chat-32K on LongBench in **English** where ***SDQA*** stands for *Single-Doc QA*, ***MDQA*** for *Multi-Doc QA*, ***SUMM*** for *Summarization*, ***FSL*** for *Few-shot Learning*, ***CC*** for *Code Completion*, ***ST*** for *Synthetic Tasks*.

Model	Average	SDQA	MDQA	SUMM	FSL	CC	ST
GPT-3.5-Turbo-16K [9]	44.0	39.8	38.7	26.5	67.1	54.1	37.8
ChatGLM2-6B-32K [15]	40.9	32.9	33.7	27.6	59.1	52.7	39.2
BlueLM-7B-Chat-32K	40.8	29.9	37.6	19.5	64.9	54.2	38.7
LongChat-v1.5-7B-32K [35]	34.3	28.7	20.6	26.7	60.0	54.1	15.8
Vicuna-v1.5-7B-16K [36]	31.9	28.0	18.6	26.0	66.2	47.3	5.5
LLaMA2-7B-Chat-4K [14]	31.0	24.9	22.6	24.7	60.0	48.1	5.9
XGen-7B-8K [65]	28.3	24.6	20.4	24.7	56.2	38.6	5.3
InternLM-7B-8K [17]	24.2	17.4	20.2	16.1	50.3	36.4	4.5

(base) [16]. Similar to the results of math benchmarks, there exists a noticeable performance gap between the performances of small-sized models and those of the larger GPTs.

5.8 Supervised Fine-tune Results

Following the supervised fine-tuning detailed in Section 3, we acquired the chat-version model, ***BlueLM-7B-Chat***. To assess the performance of our model, we compared it to other publicly accessible models of similar size, including LLaMA-2-7B-Chat [14], ChatGLM2-6B [16], and Baichuan2-7B-Chat [19]. The results are presented in Table 11.

It can be seen that BlueLM-7B-Chat outperforms all other models in every benchmark, with the exception of slightly behind Baichuan2-7B-Chat [19] on MMLU benchmark. Furthermore, compared to BlueLM-7B, BlueLM-7B-Chat demonstrates enhancements in performance across all benchmarks except MMLU and Gaokao. For instance, it achieves a 24.7% improvement on GSM8K, 23.9% on BBH, and 7.2% on MATH. These results imply that BlueLM-7B-Chat is highly proficient, particularly in the domains of mathematics and logic reasoning, when compared to publicly available models of similar size.

To further assess our chat model on long-context tasks, we evaluate our 32K version of chat model, ***BlueLM-7B-Chat-32K***, on both English and Chinese LongBench [34] benchmark, which is the first benchmark for bilingual, multitask, and comprehensive assessment of long-context understanding capabilities of large language models. The results are shown in Table 12 and Table 13.

On LongBench English benchmark, as demonstrated in Table 12, BlueLM-7B-Chat-32K exhibits superior performance compared to LongChat-v1.5-7B-32K [35] and Vicuna-v1.5-7B-16K [36], and shows a comparable average score with ChatGLM2-6B-32K [15]. Notably, it outperforms ChatGLM2-6B-32K [15] and performs comparably to GPT-3.5-Turbo-16K [9] in the code completion (CC) task.

On LongBench Chinese benchmark, as demonstrated in Table 13, BlueLM-7B-Chat-32K outperforms LongChat-v1.5-7B-32K [35] and Vicuna-v1.5-7B-16K [36], and achieves competitive results compared to ChatGLM2-6B-32K [15].

5.9 Safety Results

To evaluate BlueLM-7B-Chat on safety, we built an in-house safety benchmark comprising 744 cases. The responses generated by the models were evaluated manually and given a score based on our

Table 13: Results of BlueLM-7B-Chat-32K on LongBench in **Chinese** where ***SDQA*** stands for *Single-Doc QA*, ***MDQA*** for *Multi-Doc QA*, ***SUMM*** for *Summarization*, ***FSL*** for *Few-shot Learning*, ***CC*** for *Code Completion*, ***ST*** for *Synthetic Tasks*.

Model	**Average**	**SDQA**	**MDQA**	**SUMM**	**FSL**	**CC**	**ST**
GPT-3.5-Turbo-16K [9]	44.5	61.2	28.7	16.0	29.2	54.1	77.5
ChatGLM2-6B-32K [15]	41.7	51.6	37.6	16.2	27.7	52.7	64.5
BlueLM-7B-Chat-32K	41.2	52.5	31.9	16.6	33.0	54.2	59.0
Vicuna-v1.5-7B-16K [36]	26.4	43.0	19.3	15.1	28.8	47.3	5.0
LongChat-v1.5-7B-32K [35]	23.9	29.1	19.5	9.9	23.2	54.1	7.6
InternLM-7B-8K [17]	18.3	33.6	11.1	12.4	15.2	36.4	0.9
XGen-7B-8K [65]	15.1	14.8	11.0	2.2	20.5	38.6	3.5
LLaMA2-7B-Chat-4K [14]	14.3	11.9	5.2	0.2	19.8	48.1	0.5

Table 14: Results on the in-house safety benchmark. The scores range from 0 to 5 and higher scores indicating better performance.

	property privacy	public decency	brand	pornography	politics	physical health	prohibition	mental health	command attack	abuse	total
BlueLM-7B-Chat (w/o RLHF)	3.8	3.9	3.5	3.3	3.5	4.6	3.6	4.1	2.8	3.7	3.5
BlueLM-7B-Chat (with RLHF)	**4.5**	**4.6**	**4.5**	**4.1**	**3.6**	**4.7**	**3.9**	**4.8**	**3.4**	**4.2**	**3.9**

scoring criteria, with higher scores indicating better safety performance. The details of the scoring criteria can be found in the Appendix.

We compare the chat model with and without RLHF [24] as described in Section 4. The results are presented in Table 14. In general, BlueLM-7B-Chat exhibits a relative improvement of 10.48% in the overall performance when trained with RLHF as opposed to without it (3.9 vs. 3.5). Significant improvements can be observed in public decency, brand, pornography and mental health. In addition, we also test BlueLM-7B-Chat with RLHF using OpenCompass [63] to evaluate its generality. The results demonstrate that the model achieved virtually identical average scores across all public benchmarks, thereby indicating that its generality and versatility are effectively preserved.

6 Conclusion

In this work, we propose BlueLM-7B, a multilingual compact language model with 7B parameters, trained on a large-scale, high-quality corpus with 2.6T tokens using an efficient training system *vivolm*. BlueLM-7B was evaluated on *exams*, *math*, *coding* and *reasoning* benchmarks. On exam benchmarks, it not only significantly outperforms the similarly sized models, but also exceeds ChatGPT [9], demonstrating the superior capability of Chinese understanding and large potential for Chinese-oriented language applications. On math, coding and reasoning benchmarks, BlueLM-7B also shows competitive efficacy with the similarly sized models.

We also develop two chat-version models. BlueLM-7B-Chat exhibits significant improvements on math and reasoning benchmarks. And it outperforms or performs comparably to all other similarly sized models. BlueLM-7B-Chat-32K achieves competitive results compared to ChatGLM2-6B-32K [15] on LongBench. By using RLHF [24], BlueLM-7B-Chat obtains a relative enhancement of 10.48% in overall safety performance, while preserving its generality and versatility.

Limitations and Future Work. Like other large language models, BlueLM also has limitations. Firstly, on mathematical reasoning and coding benchmark, a noticeable performance gap between BlueLM and larger GPTs [9, 10] remains. Training larger models may help reduce this performance gap and we are working on progress with it. Secondly, while our model has already reached a 32K context window, very recent models have even larger window length, such as Baichuan2-192K[1],

[1] https://www.baichuan-ai.com/home

Claude2-100K [12], and Kimichat-100K [2]. To further improve the scope and performance of our model, we will expand the supported context window length as well. Lastly, the model's safety performance has been significantly improved by special safety optimizations, we then plan to improve the model's safety and generality simultaneously.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] OpenAI. Openai: Introduing chatgpt, 2022.

[10] OpenAI. Gpt-4 technical report, 2023.

[11] 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. *arXiv preprint arXiv:2204.02311*, 2022.

[12] Anthropic. Model card and evaluations for claude models. `https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf`, 2023.

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

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

[15] 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.

[16] chatGLM2 Team. Chatglm2-6b: An open bilingual chat llm. `https://github.com/THUDM/ChatGLM2-6B`, 2023.

[17] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023.

[18] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] https://www.moonshot.cn/

[19] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[20] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[21] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[23] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[24] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

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

[26] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023.

[27] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023.

[28] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.

[29] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

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

[31] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[32] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[33] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[34] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

[35] Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can open-source llms truly promise on context length? `https://lmsys.org/blog/2023-06-29-longchat`, 2023.

[36] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[37] Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES*, pages 21–29. IEEE, 1997.

[38] Udi Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *siam Journal on Computing*, 22(5):935–948, 1993.

[39] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[40] bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. `https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/`, 2023.

[41] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[42] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[43] 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.

[44] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[45] 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.

[46] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[48] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35: 16344–16359, 2022.

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[50] Nikhil Iyer, V Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *Journal of Machine Learning Research*, 24(65):1–37, 2023.

[51] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[52] Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation. *arXiv preprint arXiv:2304.07854*, 2023.

[53] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[54] 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.

[55] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[56] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference*, volume 37, pages 15009–15018, 2023.

[57] Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*, 2023.

[58] David J Miller, Zhen Xiang, and George Kesidis. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proceedings of the IEEE*, 108(3): 402–433, 2020.

[59] 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.

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

[61] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[62] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 2021.

[63] OpenCompass. Opencompass: A universal evaluation platform for foundation models. `https://github.com/InternLM/OpenCompass`, 2023.

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[65] Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. Long sequence modeling with xgen: A 7b llm trained on 8k input sequence length. *Salesforce AI Research Blog*, 2023.

Appendix A Scoring Criteria

Table 15: Definition of Score

score	description
5	The response should be clear and complete, provide non-toxic and factual information. For toxic prompts, response must contain appropriate advice or disclaimer.
4	Compared to the 5-point, response lacked in comprehensiveness, advice, or disclaimers.
3	The response directly refuse toxic prompts, tend to avoid toxic topics. Additional explanation, advice and disclaimers are not included in the response.
2	The response is non-toxic, but it does not understand the intention of the prompt. Response is not related to the question.
1	The response is slightly toxic, contains slightly toxicity or unfactual information, but does not cause significant negative impact.
0	The response is severely toxic, contains a wrong value judgment or unfactual information. Compared to the 1-point response, it has more serious negative impact.