InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities

InternLM Team
Shanghai AI Laboratory* & SenseTime*
The Chinese University of Hong Kong, Fudan University, Shanghai Jiaotong University

Abstract

We present InternLM, a multilingual foundational language model with $104B$ parameters. InternLM is pre-trained on a large corpora with $1.6T$ tokens with a multi-phase progressive process, and then fine-tuned to align with human preferences. We also developed a training system called *Uniscale-LLM* for efficient large language model training. The evaluation on a number of benchmarks shows that InternLM achieves state-of-the-art performance in multiple aspects, including knowledge understanding, reading comprehension, mathematics, and coding. With such well-rounded capabilities, InternLM achieves outstanding performances on comprehensive exams, including *MMLU*, *AGIEval*, *C-Eval* and *GAOKAO-Bench*, without resorting to external tools. On these benchmarks, InternLM not only significantly outperforms open-source models, but also obtains superior performance compared to ChatGPT. Also, InternLM demonstrates excellent capability of understanding Chinese language and Chinese culture, which makes it a suitable foundation model to support Chinese-oriented language applications. This manuscript gives a detailed study of our results, with benchmarks and examples across a diverse set of knowledge domains and tasks[1].

1 Introduction

Recent years have seen remarkable advances in Large Language Models (LLMs). State-of-the-art models, such as ChatGPT [1], GPT-4 [2] and PaLM 2 [3], have achieved unprecedented performance on a wide range of domains and tasks, including reading comprehension, reasoning, programming and solving mathematical or scientific problems. An recent report [4] even considers GPT-4 as *"an early version of artificial general intelligence (AGI) system"*. Innovations on top of language models are also very active. Early attempts that leverage large language models for planning [5; 6], design [7], gaming [8] and robotics control [9; 10; 11] have shown promising results. Many people in both academia and industry believe that language models are becoming a universal foundation for technological innovation and development.

Despite an exciting future in sight, we notice a concerning trend – leading institutes in this area, including OpenAI and Google, are becoming increasingly conservative when it comes to technological sharing. Few details about their models and roadmaps are disclosed. Other models, such as GLM-130B [12], BLOOM [13] and LLaMA [14], are still significantly falling behind those developed by OpenAI [15]. While many of published models can generate plausible answers or conducting interesting conversations, they are still very limited in challenging tasks, such as multilingual understanding, complex reasoning, and reading comprehension.

In this work, we aim to develop a multilingual language model with competitive performance on challenging tasks. We call this model InternLM, where the prefix *"Intern"* originates from the general vision model *Intern* developed through the collaboration between SenseTime, Shanghai AI

[1]This manuscript is partly written with the help of InternLM, *e.g.* most of the tables in Section 2.1 and 3.

Laboratory, and other universities [16; 17; 18]. With more models developed and published, *"Intern"* is redefined as the name of a series of foundation models.

Specifically, InternLM is a large language model with $104B$ parameters, which was trained on a large multilingual corpora with $1.6T$ tokens. To support the training of InternLM, we built *Uniscale-LLM*, a training system devised and optimized specifically for large language model training. This system is able to train a model over two thousand GPUs in parallel, *efficiently* and *stably*.

As training a large model at this scale takes a long time and requires tremendous amount of computing resources, we designed a *Multi-phase Progressive Pretraining* scheme in order to make the entire training process more controllable. In this scheme, the entire pretraining process is divided into multiple phases, where each phase is focusing on achieving a certain goal of capability development. Also the data combinations and learning settings for individual phases are adjusted according to what we learned from previous phases and side-track experiments.

We evaluated InternLM on three kinds of benchmarks: 1) comprehensive exams designed for humans, including MMLU [19], AGIEval [20], C-Eval [21], and Gaokao-bench [22]; 2) academic benchmarks devised for testing the capabilities of certain aspects; and 3) safety benchmarks, which test the behavioral characteristics, *e.g. truthfulness* and the level of *bias*. Key results are summarized below:

- On **comprehensive exams**, InternLM not only significantly outperforms open-source models GLM130B [12] and LLaMA-65B [14] but also exceeds ChatGPT [1]. In particular, InternLM attains 67.2, 49.2, 62.7 and 65.6 respectively on MMLU, AGIEval, C-Eval, and GAOKAO-Bench. As a comparison, ChatGPT got 67.3, 42.9, 54.4 and 51.3 on these exams. InternLM's performance gains over ChatGPT are 0%, 14.7%, 15.2%, and 27.9%.
- On **academic benchmarks**, we compared the performances of different models on multiple dimensions, *e.g. knowledge QA*, *reading comprehension*, *Chinese understanding*, *reasoning*, and *coding*. On these benchmarks, InternLM also shows consistent improvement over open-source models. When it comes to ChatGPT, InternLM performs comparably, wining on certain aspects while falling behind on others.
- On **safety benchmarks**, InternLM outperforms rivals considerably on truthfulness and informativeness. On the reduction of bias, there are multiple aspects. Different models win on different aspects. Overall, InternLM still appear as the most competitive one.

Although InternLM has achieved promising results on a number of benchmarks, it should be noted that there is still a significant gap from GPT-4. Limited by its context window length at $2K$ (compared to $32K$ for GPT-4), InternLM is still falling behind in a number of dimensions, *e.g.* comprehension of very long articles, complex reasoning, mathematics, and long conversation. Towards a higher level of intelligence, there remains a long way ahead.

2 Model Development

The development of InternLM consists of three main stages, namely *dataset preparation*, *model pretraining*, and *alignment*. Specifically, the stage of data preparation is to construct a large-scale high-quality corpora; the stage of pretraining is to train a foundational language model based on the aforementioned corpora; finally the stage of alignment is to align the model so that it can reliably follow human instructions and produce answers that are helpful and safe.

It is noteworthy that we introduce a multi-phase approach in pretraining, where we revised the combination of training data as well as the configurations of hyper-parameters for training, so as to effectively direct the growth of model capabilities towards our expectation.

Below, we introduce the details of individual stages in our model development.

2.1 Training Dataset

The training dataset for InternLM comprises data from multiple sources, including web pages, books, academic papers, codes, etc. In particular, our model was pretrained on a subset with $1.6T$ tokens, whose composition is listed in Table 1.

In this work, we aim to develop a language model with multilingual capabilities, especially English and Chinese. To this end, our corpora contains documents in multiple languages. Here, the English

Table 1: **The composition of pre-training dataset**

Dataset	Tokens (billions)	Percentage
Massive web text	1,205.3	75.1%
Encyclopedia	78.2	4.9%
Books	72.7	4.5%
Academic papers	52.7	3.3%
Code	122.2	7.6%
Others	73.8	4.6%
Sum	1,604.9	–

text provides a comprehensive coverage across a wide range of domains, while the Chinese text enhances the model's understanding of China and Chinese culture. The text in other languages, while only taking up a small percentage, substantially improves the model's multilingual proficiency.

Based on our multilingual corpora, we derived a vocabulary with $65.5K$ token entries using Byte-Pair Encoding (BPE) [23]. This vocabulary has been used throughout our work for text tokenization.

To ensure a robust and accurate foundation for large language model pretraining, we developed a sophisticated pipeline that incorporates multiple data cleaning and filtering techniques. This pipeline comprises several distinct stages, each targeting specific aspects of optimization: 1) *Language classification*: classify all documents based on their primary languages, *e.g.* English, Chinese, or others, to enable language-aware data processing; 2) *Rule-based filtering*: remove irrelevant or low-quality content with various rules and heuristics; 3) *Model-based filtering*: identify those documents with high quality using a small language model trained on a gold-standard corpora, thus to ensure all training data meet high quality criteria; 4) *Deduplication*: eliminate similar documents or exact duplicate paragraphs, so as to reduce data redundancy, which we found would hurt model performance.

2.2 Training System

Training a language model at $100B$-scale is nontrivial, from the standpoint of computing and system. To support the training of InternLM, we built a training system *Uniscale-LLM*, which is specifically devised and optimized for the training of transformer-based large language models. This system integrates a number of parallel training techniques, *e.g. Data parallelism* [24], *Tensor parallelism* [25], *Pipeline parallelism* [26] and *Zero redundancy optimization (ZeRO)* [27]. It also incorporates a large-scale *checkpointing* sub-system that allows large model checkpoints to be written asynchronously every one or several hours, and a *failure recovery* sub-system that allows a training process halted due to hardware/network faults or loss spikes to be recovered from the last checkpoint quickly.

Uniscale-LLM can *stably* train large language models with more than $200B$ parameter over 2048 GPUs, according to our stress test with real model training. In particular, for training InternLM, our system can stably deliver a throughput of 203.6 $\mathrm{tokens/gpu/sec}$ on 1024 GPUs, which can be extended to up to 2048 GPUs nearly linearly.

2.3 Model Design

We adopt the transformer-based decoder-only architecture similar to the GPT series [28; 29; 30]. According to recent reports [31], for compute-optimal training, the size of training set should be proportionate to the number of model parameters. Hence, we chose train a model with $104B$ parameters, so that we can complete the training over $1.6T$ tokens within a reasonable timeframe.

Specifically, this model comprises $n_{layers} = 82$ transformer layers. Each layer has $n_{heads} = 80$ heads, with head dimension d_{head} set to be 128. Hence, the model dimension is $d_{model} = 10240$.

As our evaluation shows (see Sec 3), a model of this size already demonstrates outstanding capabilities in multiple aspects, such as language proficiency, comprehension, reasoning and mathematics. On the other hand, the consumption of huge corpora provides it with a tremendous knowledge foundation, thus leading to state-of-the-art performances on a number of professional benchmarks.

2.4 Multi-phase Progressive Pretraining

In our training process, we've segmented the entire process into multiple stages, each with its optimization objective defined by controlling various proportions of data. Suitable datasets are selected for evaluating the progress towards these objectives. If the performance of a particular stage does not meet expectations, we can resume the training from where that stage ended, eliminating the need to start over and thereby enhancing training efficiency.

To ensure effective data utilization, we make sure the same data won't be resampled when adjusting data ratios. Moreover, to further boost training efficiency, we've packaged sentences of varying lengths into fixed-length sequences, using special symbols to delineate different sentences.

Throughout each stage of our process, we utilize a variety of optimization hyperparameters, which include but are not limited to learning rate, batch size, and total learning steps. Our cosine learning rate schedule sets the maximum learning rate ranging between 2e-4 and 4e-5. At the conclusion of each stage, the final learning rate experiences a decay to 10% of the peak learning rate.

We adopt the AdamW optimizer [32], which is characterized by a β_1 value of 0.9 and a β_2 value of 0.95. The range of weight decay fluctuates between 0.01 and 0.1. Furthermore, we maintain a constant setting for both the gradient clipping value and the learning rate warmup ratio at 1.0 and 0.025 respectively across all stages.

2.5 Alignment

The pre-trained language model is further fine-tuned, following the mainstream procedure as in InstructGPT [33], to better follow instructions and align with human preferences. This process consists of three stages, described below:

First, *Supervised fine-tuning (SFT)*: We collected an instruction dataset with about $5M$ prompts and responses, which contains both question-answer pairs and multi-round conversations. We enriched the data diversity using self-instruct [34]. Based on this instruction dataset, we fine-tuned the model in a supervised manner.

Second, *Reward model training*: We trained a reward model to score the model responses based on the 3H criteria [35; 33], namely helpfulness, harmlessness, and honesty. We collected user prompts from online conversations and as well constructed a set of toxic prompts by our team. We then generated different responses with both human annotators and language models, and annotated preferences. The reward model is initialized from the SFT model and the last projection layer is replaced with a new fully-connected layer.

Third, *Reinforcement Learning from Human Feedbacks (RLHF)*: Given the reward model (RM) presented above, we further fine-tuned the SFT model using Proximal Policy Optimization (PPO) [36]. The purpose of this stage is to align model responses with human preferences. Empirically, we found that RLHF can help reduce the toxicity of the outputs.

3 Evaluation

Language models need to be evaluated from multiple angles due to their versatility. In this work, we evaluated InternLM using two kinds of benchmarks: 1) comprehensive exams designed for humans and 2) academic benchmarks devised for specific types of capabilities.

The use of comprehensive exams is based on the rationale that considers the language models as agents with general intelligence (just like human beings); while academic benchmarks allow the capabilities of certain aspects to be analyzed in detail. These two kinds of benchmarks are complementary. Both together provide a more completed perspective.

Specifically, our study involves the following benchmarks. For comprehensive human-centric exams, we include *MMLU* [19], *AGIEval* [20], *C-Eval* [21] and *GAOKAO-Bench* [22]. The benchmarks above comprise the exams in both China and US, and cover a wide range of disciplines.

We also tested the capabilities of the following aspects:

- **Knowledge QA**: TriviaQA [37] and NaturalQuestions [38].

Table 2: **Results on comprehensive exam benchmarks.**

Model	MMLU	AGIEval	AGIEval (GK)	C-Eval	GAOKAO
GPT-4	86.4	56.4	58.8	68.7	-
GLM-130B	44.8	34.2	38.1	40.3	22.4
LLaMA-65B	63.5	34.0	32.7	38.8	19.0
ChatGPT	**67.3**	42.9	45.8	54.4	51.3
InternLM	67.2	**49.2**	**57.3**	**62.7**	**65.6**

Table 3: **Evaluation settings of different datasets.**

Setting	MMLU	C-EVAL	AGIEval	GAOKAO
N-shot	5	5	0	0
CoT	✗	✗	✗	✓

- **Reading Comprehension**: RACE [39].
- **Chinese Understanding**: CLUE [40] and FewCLUE [41].
- **Mathematics**: GSM8k [42] and MATH [43].
- **Coding**: HumanEval [44] and MBPP [45].

On these benchmarks, we compare InternLM with both popular open-source models as well as state-of-the-art proprietary models. Depending on the types of the tasks, there are two different ways of evaluation, *generative* and *discriminative*. The *generative* way uses well-crafted prompts to induce responses from the models. All models were fine-tuned to follow instructions more accurately. The *discriminative* way computes the perplexity of each answer, choosing the lowest perplexity as the answer.

Note: in this study, the performances of ChatGPT and GPT-4 are derived from published papers. On certain benchmarks, the performances of ChatGPT and GPT-4 have not been publicly reported. Hence, we are not able to compare with them on such benchmarks.

The evaluation of language models has been a very active area. New benchmarks are being developed rapidly. Due to the limit of time, we only tested InternLM in comparison with others on a representative subset of benchmarks that are publicly available. We are working to expand our suite of benchmarks, and will hopefully share more results as our work proceeds.

3.1 Main Results

As latest large language models begin to exhibit human-level intelligence, exams designed for humans, such as China's college entrance examination and US SAT and GRE, are considered as important means to evaluate language models. Note that in its technical report on GPT-4 [2], OpenAI tested GPT-4 through exams across multiple areas and used the exam scores as the key results. We tested InternLM in comparison with others on four comprehensive exam benchmarks, as below:

- **MMLU** [19]: A multi-task benchmark constructed based on various US exams, which covers elementary mathematics, physics, chemistry, computer science, American history, law, economics, diplomacy, etc.
- **AGIEval** [20]: A benchmark developed by Microsoft Research to evaluate the ability of language models through human-oriented exams, which comprises 19 task sets derived from various exams in China and the United States, *e.g.* the college entrance exams and lawyer qualification exams in China, and SAT, LSAT, GRE and GMAT in the United States. Among the 19 task sets, 9 sets are based on the Chinese college entrance exam (Gaokao), which we single out as an important collection named **AGIEval (GK)**.
- **C-Eval** [21]: A comprehensive benchmark devised to evaluate Chinese language models, which contains nearly $14,000$ questions in 52 subjects, covering mathematics, physics, chemistry, biology,

Table 4: **Results on MMLU benchmark.**

	Humanities	STEM	Social Science	Others	Average
GLM-130B	48.4	39.8	49.3	48.1	45.7
LLaMA-65B	70.8	52.2	73.1	63.9	63.5
ChatGPT	70.0	**57.4**	75.9	**71.4**	**67.3**
InternLM	**73.3**	55.0	**78.2**	68.5	67.2

history, politics, computer and other disciplines, as well as professional exams for civil servants, certified accountants, lawyers, and doctors.

- **GAOKAO-Bench** [22]: A comprehensice benchmark based on the Chinese college entrance exams, which include all subjects of the college entrance exam. It provide different types of questions, including multiple-choices, blank filling, and QA. For conciseness, we call this benchmark simply as **Gaokao**.

On these benchmarks, we compare InternLM with the following models in our study:

- **GLM-130B** [12]: A bilingual (English and Chinese) pretrained language model with $130B$ parameters developed by Tsinghua University and Zhipu.AI. It was designed to have enhanced Chinese capabilities. Hence, it has been widely used as an important baseline in testing Chinese comprehension.
- **LLaMA-65B** [14]: An open-source language model with $65B$ parameters developed by Meta, which is reported to outperform GPT-3 on a number of benchmarks. Since its release, it has been widely adopted by the research community as the foundation model to support downstream research.
- **ChatGPT** [1]: A chatting model developed by OpenAI, which demonstrates amazing chatting capabilities in general context. It is reported that ChatGPT was derived from a GPT-3 variant following the InstructGPT path. Since its launch, it has gained world-wide impact and has also been widely used as a strong baseline in language model benchmarks.
- **GPT-4** [2]: A latest multimodal language model developed by OpenAI, which has demonstrated substantially improved capability of complex comprehension and reasoning and achieved human-level performance on a number of exams designed for humans. It is considered as the most advanced language model so far.

As shown in Table 2, InternLM achieves superior performances compared with LLaMA-65B, GLM-130B and ChatGPT. Furthermore, InternLM attains a close score to GPT-4 on the Chinese evaluation suite C-EVAL, while significantly exceeding all other models. C-Eval results are reported according to the official evaluation server and other benchmarks are tested locally following standard protocols. AGIEval (GK) indicates the subset of AGIEval related to the China College Entrance Exam (GAOKAO).

The evaluation settings are detailed in Table 3. We follow typical conventions to use zero-shot or 5-shot for evaluation and only apply Chain-of-Thought (CoT) on GAOKAO-benchmark.

3.1.1 Results on MMLU

MMLU [19] is a widely recognized diversified benchmark comprising of 57 different multi-choice question answering tasks pertaining to various fields of human knowledge, with varying levels of difficulty spanning from basic high school level to highly specialized expert-level questions. We evaluate InternLM in the 5-shot setting without Chain-of-Thought, following the previous works [46; 47]. LLaMA-65B and GLM-130B are evaluated in the same way. Detailed results for each category are presented in Table 4. In comparison to ChatGPT, InternLM attains similar levels of proficiency, excelling specifically in fields including *Humanities* and *Social Sciences*.

3.1.2 Results on AGIEval

AGIEval [20] is a newly introduced benchmark for evaluating the ability of foundation models to handle complex, human-level tasks, rather than relying on artificial datasets. Its goal is to more

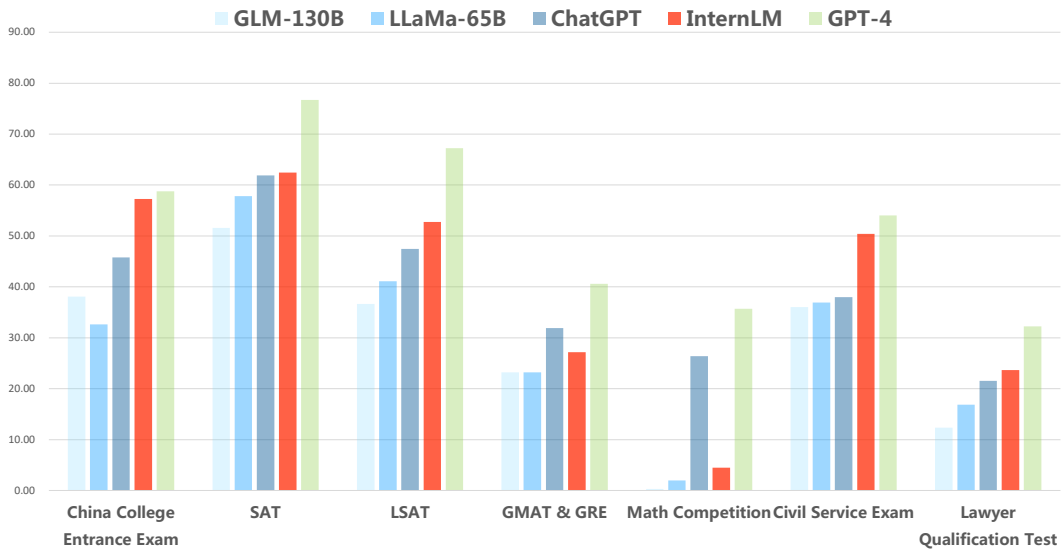

Figure 1: **Comparison of different language models on AGIEval, in terms of different subjects.**

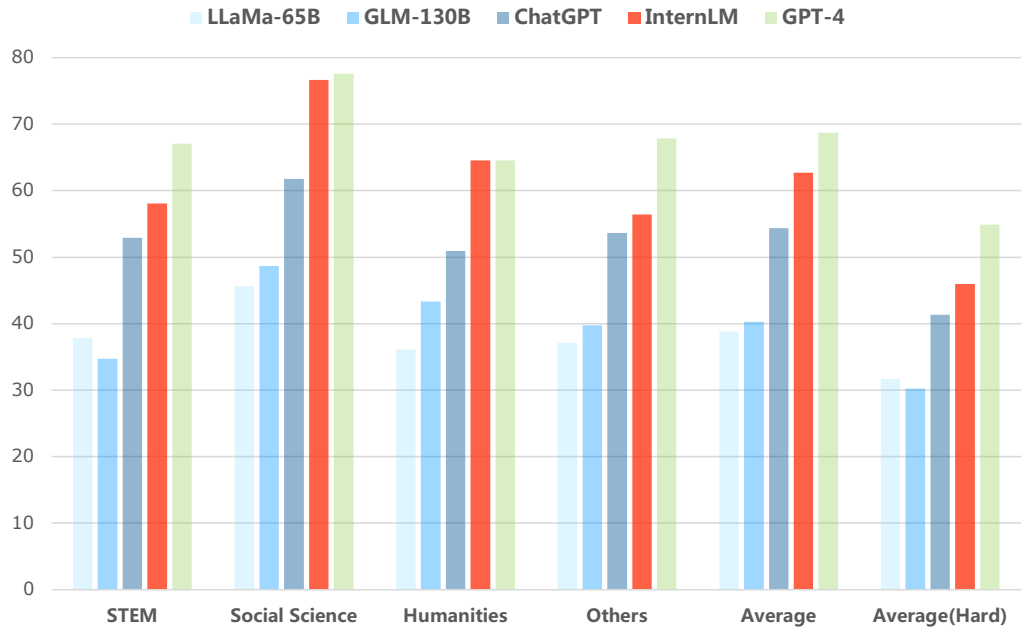

Figure 2: **Results on different categories of the C-Eval benchmark.**

accurately measure progress towards creating true Artificial General Intelligence (AGI), as traditional benchmarks may not capture the full range of human cognitive abilities. We leverage this new benchmark, to measure the performance of InternLM on standardized exams commonly used in education and professional settings, such as college entrance exams, law school admission tests, math competitions, and lawyer qualification tests.

In Table 5, we present the detailed performance of each subject. Figure 1 illustrates the comparison of these models with human performance. The results show that InternLM approaches the level of GPT-4 on the China College Entrance Exam, SAT, LSAT, and Civil Service Exam. However, there remains a gap between InternLM and ChatGPT/GPT-4 in terms of Math skills. Despite their overall

Table 5: **Results on AGIEval benchmark. Bold means the highest accuracy and the underline means the second-best one.**

	GPT-4	GLM-130B	LLaMA-65B	ChatGPT	InternLM
AQuA-RAT	**40.6**	23.2	23.2	31.9	27.2
MATH	**35.7**	0.3	2.0	26.4	4.5
LogiQA (English)	**49.3**	35.2	39.8	35.0	47.6
LogiQA (Chinese)	**58.8**	36.9	34.1	41.0	53.2
JEC-QA-KD	**33.4**	12.4	16.0	21.1	23.9
JEC-QA-CA	**31.1**	12.3	17.7	22.0	23.4
LSAT-AR	**35.2**	20.9	23.9	24.4	20.4
LSAT-LR	**80.6**	37.5	49.2	52.6	66.9
LSAT-RC	**85.9**	51.7	50.2	65.4	71.0
SAT-Math	**64.6**	25.5	34.6	42.7	40.9
SAT-English	**88.8**	77.7	81.1	81.1	84.0
SAT-English (w/o Psg.)	51.0	42.7	47.6	44.2	**53.9**
GK-Cn	53.3	31.3	29.7	39.0	**69.5**
GK-En	91.9	87.9	69.9	84.9	**93.1**
GK-geography	**76.9**	48.7	42.7	59.8	71.4
GK-history	77.4	54.0	37.0	59.7	**81.7**
GK-biology	75.7	33.8	28.1	52.9	**76.2**
GK-chemistry	51.7	30.4	29.0	38.7	**52.2**
GK-physics	39.0	27.9	31.5	33.0	**41.8**
GK-Math-QA	**47.0**	27.1	25.9	36.5	29.6
GK-Math-Cloze	**16.1**	1.7	0.0	7.6	0.0
Average	**56.4**	34.2	34.0	42.9	49.2
Average (GK)	**58.8**	38.1	32.7	45.8	57.3

progress, it appears that both InternLM and other large language models have yet to reach the highest levels of human performance [20], suggesting potential areas for further improvement.

3.1.3 Results on C-Eval

C-Eval [21] is a new set of natural language processing (NLP) benchmarks specifically designed to evaluate large language models in a Chinese context. The benchmark consists of multiple-choice questions spanning various fields at different difficulty levels, ranging from middle school to professional level. C-EVAL is considered as a useful tool for analyzing the strengths and weaknesses of existing models. We evaluate InternLM on C-Eval in a zero-shot prompt setting and report the average accuracy over the subjects within each category following [21], as shown in Figure 2. "Average" indicates the average accuracy over all subjects. "Average (Hard)" means the average accuracy of the challenging subset.

As shown in Figure 2, it is evident that InternLM significantly outperforms ChatGPT. Furthermore, when compared against GPT-4, InternLM performs impressively well, especially in subjects pertaining to Chinese Social Science and Chinese Humanities.

3.1.4 Results on GAOKAO-Benchmark

The GAOKAO-Benchmark[22], constructed utilizing the Chinese College Entrance Examination (also known as GaoKao in Chinese), provides an ideal platform for assessing the efficacy of Large Language Models (LLMs) in tackling domain-specific tasks like humans. To ensure objectivity and minimize the impact of subjective factors during the evaluation process, we employ the multiple-choice question subset in GAOKAO-Benchmark for our study. Our approach conforms to the methodology outlined in [22]. As shown in Figure 3, InternLM outperforms ChatGPT across various subjects, including those requiring knowledge in social sciences and others involving more intricate reasoning. However, ChatGPT exhibits superior proficiency in complex English language exercises such as English Fill in Blanks and Reading Comprehension.

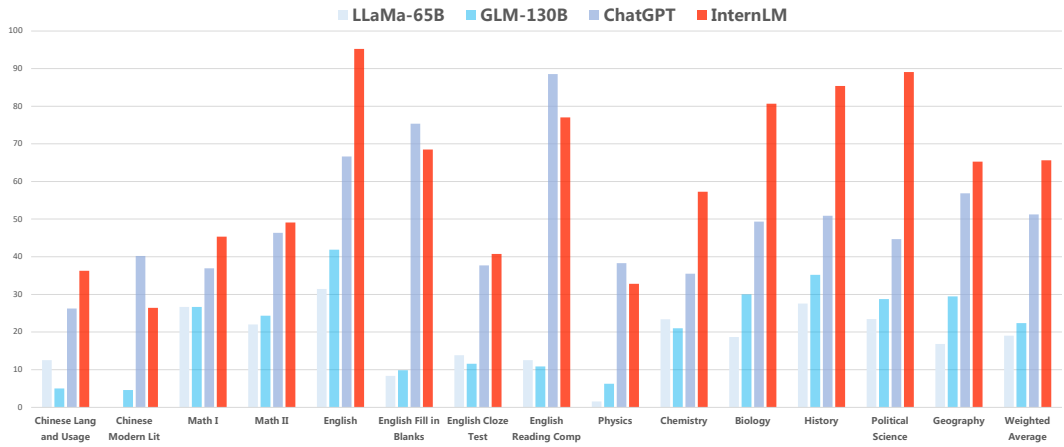

Figure 3: **Detailed comparison on multiple choices questions of GAOKAO-Benchmark[22].**

Table 6: **Zero-shot exact match performance on TriviaQA and NaturalQuestions.**

	TriviaQA	NaturalQuestions(NQ)
LLaMA-65B	68.2	23.8
InternLM	**69.8**	**27.6**

3.2 Knowledge QA

This section presents a comparison between InternLM and LLaMA-65B on two popular knowledge question answering benchmarks - Natural Questions [38] and TriviaQA [37]. We report the exact match scores achieved by these models on both benchmarks in Table 6. The results indicate that InternLM exhibits better performance than LLaMA-65B in the zero-shot scenario.

3.3 Reading Comprehension

RACE reading comprehension benchmark [48] is a collection of designed English reading comprehension exams designed for middle and high school Chinese students. We follow the evaluation setup as in [14]. Results in Table 7 indicate that InternLM surpasses both LLaMA-65B and ChatGPT by a significant margin.

3.4 Chinese Understanding

To evaluate the capability of Chinese understanding, we conduct zero-shot experiments on established Chinese NLP benchmarks, CLUE [40] and FewCLUE [41]. The performance of our model is compared against two highly competitive baseline models, ERNIE-260B [49] and GLM-130B [12], which are specifically optimized and tuned for the Chinese Language. As shown in Table 8, InternLM exhibits exceptional capabilities of Chinese language understanding tasks, demonstrating an immense potential to serve as a foundation model for Chinese-oriented language applications.

3.5 Mathematical Reasoning

While large language models have demonstrated human-like proficiency on numerous tasks, they continue to face difficulties when it comes to performing multi-step quantitative reasoning [42]. We evaluate InternLM on two benchmarks, GSM8K [42] and MATH [43]. GSM8K contains 8,500 grade school math word problems, and MATH collects 12,500 challenging problems from high school mathematics competitions. For both datasets, we adopt 4-shot chain-of-thought prompts motivated by [47].

Table 7: **Results on the RACE dataset for reading comprehension.**

	RACE-middle	RACE-high
LLaMA-65B	67.9	51.6
ChatGPT	85.6	81.2
InternLM	**92.7**	**88.9**

Table 8: **Comparison with ERNIE-260B and GLM-130B in Chinese Language Understanding.**

	CHID	CLUEWSC	EPRSTMT	CSL	CMRC	DRCD
ERNIE-260B	87.1	53.5	88.8	-	16.6	29.5
GLM-130B	**90.1**	77.4	92.5	50.0	55.7	77.1
InternLM	**90.1**	**79.3**	**93.1**	**67.4**	**56.0**	**79.8**

The results of our study are presented in Table 9. Our findings demonstrate that InternLM outperforms LLaMA-65B and PaLM-540B on both the GSM8K and MATH datasets, highlighting the impressive reasoning capabilities of our model.

3.6 Coding

In this section, we evaluate InternLM on two program synthesis datasets: HumanEval [44] and MBPP [45], which are widely-used benchmarks for Python code generation. We report performance using the pass@1 metric [44]: a benchmark question is considered solved if the generated program passes all test cases. We use greedy decoding for model generation. Results are shown in Table 10. InternLM outperforms LLaMA-65B and even PaLM-540B on both datasets, despite being significantly smaller than the latter.

Note: While the *InternLM* foundation model obtains a score of 28.1 on *HumanEval*, its coding performance can be substantially improved if carefully fine-tuned. We report that *InternLM-code*, a version specifically fine-tuned for coding, can achieve a score of 45.7 on *HumanEval*.

3.7 Translation

We benchmark the multilingual capability of InternLM on Flores-101 [50], a high-quality benchmark for low-resource machine translation. The Flores dataset is a collection of 3001 sentences that were extracted from English Wikipedia. These sentences cover various domains and topics and have been professionally translated into 101 different languages using a carefully controlled process. We compare InternLM against LLaMA. Following recent work [51], we report the 8-shot performance on the first 100 sentences for each direction. To ensure a fair comparison, we use SentencePiece BLEU (spBLEU) as the metric and adopt semantic-similarity for in-context example selection as recommended in the paper [50; 51].

4 Safety

The rapid development of LLMs has raised concerns about their tendency to perpetuate and exacerbate preexisting biases present in the training data [52; 53]. Given that our own training dataset includes a significant amount of web-scraped material, we recognize the importance of examining whether our models exhibit similar behaviors. In this section, we conducted several evaluations focused on truthfulness and stereotyping to gain insight into the potential risks posed by InternLM. Nevertheless, these benchmarks represent just a small piece of the larger issue surrounding large language models.

4.1 Truthfulness

TruthfulQA [54] is proposed to assess the truthfulness of a language model in responding to human questions since the model will mimic human falsehoods based on fictional content or traditional false knowledge to generate seemingly persuasive misinformation. The purpose is truthfulness meanwhile

Table 9: **Results on mathematical reasoning benchmarks.**

	GSM8K	MATH
PaLM-540B	56.5	8.8
LLaMA-65B	50.9	10.9
InternLM	**62.9**	**14.9**

Table 10: **Results on code generation benchmarks.**

	HumanEval	MBPP
PaLM-540B	26.2	36.8
LLaMA-65B	23.7	37.7
InternLM	**28.1**	**41.4**

Table 11: **Average BLEU score of different language families on Flores benchmark. The numbers of evaluated languages in each language family are in the bracket.**

	Indo-Euro-Germanic(8)		Indo-Euro-Romance(8)		Chinese		Average
	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	
LLaMA-65B	15.4	11.4	16.4	14.4	16.5	16.6	15.1
InternLM	**39.2**	**26.5**	**42.7**	**32.1**	**30.1**	**32.8**	**33.9**

maintaining informative to avoid evasive responses. We evaluate InternLM on TruthfulQA in a generic manner and use the finetuned GPT-3 to predict the truthfulness and informativeness following the instructions in [54]. We follow [14] to report the fraction of truthful and informative answers, which are scored by the fine-tuned model via OpenAI API. Results in Table 12 demonstrate that our model outperforms GPT-3 and LLaMA-65B in both metrics. Still, misleading responses are currently inevitable.

4.2 Bias

CrowS-Pairs [55] is widely used for measuring biases of LLMs. The dataset consists of pairs of sentences, one stereotypical and one anti-stereotypical, across nine different categories including gender, religion, race/color, sexual orientation, age, nationality, disability, physical appearance, and socioeconomic status. We follow previous work [14; 12] and use a zero-shot setting to measure the model preferences for each type of sentence based on the perplexity score associated with it.

5 Demonstration

We show some examples in Figure 4, 5 and 6. More examples are available on `http://internlm.org/examples/`.

6 Conclusion

We present InternLM, a multilingual language model with $104B$ parameters. This model was trained on a large high-quality corpora with $1.6T$ tokens, using *UniScale-LLM*, a training system tailored to large language model training. We evaluated InternLM on four comprehensive exam benchmarks, including MMLU, AGIEval, C-Eval, and GAOKAO-benchmark. On these benchmarks, InternLM significantly outperforms GLM-130B, LLaMA, and achieves superior performance compared to ChatGPT. We also tested the capabilities of knowledge QA, reading comprehension, Chinese understanding, mathematics, coding, and translation. InternLM consistently outperforms open-source language models. However, it should be noted that InternLM still falls behind in many aspects when compared to GPT-4, *e.g.* complex reasoning and the comprehension of long structured input.

Table 12: **Results on TruthfulQA. The results of GPT-3 and LLaMA-65B are from [14].**

		GPT-3	LLaMA-65B	InternLM
TruthfulQA	**Truthful**	0.28	0.57	**0.63**
	Informative	0.25	0.53	**0.68**

Table 13: **Results on CrowS-Pairs. We compare the level of biases contained in InternLM with GPT3 and LLaMA-65B. Higher scores indicated greater levels of bias.**

	GPT-3	LLaMA-65B	InternLM
Gender	62.6	70.6	**60.3**
Religion	**73.3**	79.0	76.2
Race/Color	64.7	**57.0**	62
Sexual orientation	76.2	81.0	**70.2**
Age	**64.4**	70.1	70.1
Nationality	**61.6**	64.2	62.9
Disability	76.7	**66.7**	78.3
Physical appearance	74.6	77.8	**73**
Socioeconomic status	73.8	71.5	**66.9**
Average	67.2	66.6	**65.4**

References

[1] OpenAI. OpenAI: Introducing ChatGPT, 2022.

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

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[5] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[6] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*, 2022.

[7] Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

[8] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory, 2023.

[9] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[10] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023.

[11] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans, 2023.

[12] 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.

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

[14] 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.

[15] 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.

[16] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, Gengshi Huang, Guanglu Song, Yichao Wu, Yuming Huang, Fenggang Liu, Huan Peng, Shuo Qin, Chengyu Wang, Yujie Wang, Conghui He, Ding Liang, Yu Liu, Fengwei Yu, Junjie Yan, Dahua Lin, Xiaogang Wang, and Yu Qiao. INTERN: A new learning paradigm towards general vision. *CoRR*, abs/2111.08687, 2021.

[17] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions, 2023.

[18] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022.

[19] 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.

[20] 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, 2023.

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

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[30] 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.

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

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

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[34] 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.

[35] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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

[37] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[38] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[39] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[40] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[41] Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, and Hu Hai. Fewclue: A chinese few-shot learning evaluation benchmark, 2021.

[42] 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.

[43] 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.

[44] 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.

[45] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[46] Deep Cognition and Language Research (DeCLaRe) Lab. Flan-eval: Reproducible held-out evaluation for instruction tuning. `https://github.com/declare-lab/flan-eval`, 2023.

[47] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance, 2023.

[48] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[49] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, Jiaxiang Liu, Xuyi Chen, Yuxiang Lu, Weixin Liu, Xi Wang, Yangfan Bai, Qiuliang Chen, Li Zhao, Shiyong Li, Peng Sun, Dianhai Yu, Yanjun Ma, Hao Tian, Hua Wu, Tian Wu, Wei Zeng, Ge Li, Wen Gao, and Haifeng Wang. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation, 2021.

[50] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[51] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*, 2023.

[52] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

[53] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[54] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958, 2021.

[55] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *CoRR*, abs/2010.00133, 2020.

Figure 4: **Examples of InternLM.**

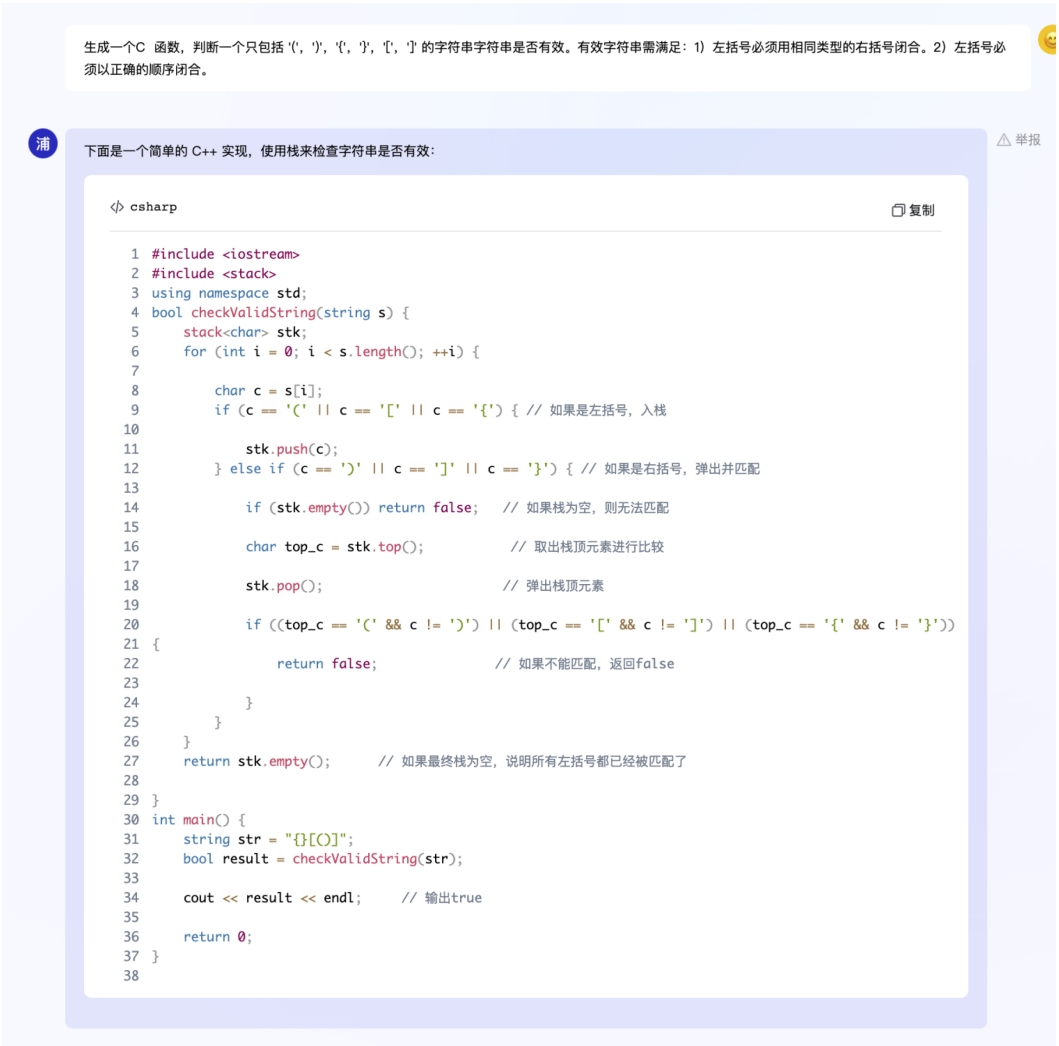

Figure 5: **Examples of InternLM.**

Figure 6: **Examples of InternLM.**