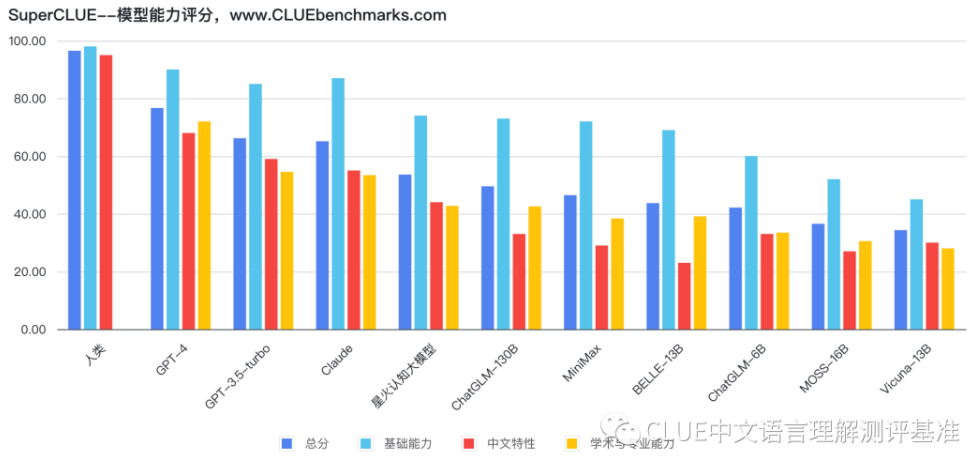
Chinese general large model evaluation benchmark SuperCLUE released an update, adding Claude and Tsinghua GLM 100 billion (parameter) models

*Note: These are Jeffrey Ding's informal and unofficial translations -- all credit for the original goes to the authors and the original text linked below. These are informal translations and all credit for the original work goes to the authors. Others are welcome to share **excerpts** from these translations as long as my original translation is cited. Commenters should be aware that the Google Doc is also publicly shareable by link. These translations are part of the ChinAI newsletter - weekly-updated library of translations from Chinese thinkers on AI-related issues: https://chinai.substack.com/*

Source: CLUE (CLUE中文语言理解测评基准)。

Date: May 11, 2023
Original Mandarin: https://mp.weixin.qq.com/s/9D9nZtpkW_RfRbC7mPaegQ

SuperCLUE, a comprehensive evaluation benchmark for Chinese general-purpose large models, was officially released on May 9. On this foundation, SuperCLUE has made this update, adding evaluations for the Claude and Tsinghua ChatGLM-130B models.



(See below for detailed data)

The SuperCLUE benchmark plan is updated monthly, and it will include more available Chinese large models in the future. At present, more than 20 model evaluation applications have been received, and the model list will be supplemented in the near future. More large-scale model research and development institutions are welcome to contact and communicate, and you can apply for evaluation at the bottom of the article.

Since this round of evaluation has not yet ended, the dataset and further information plans will be made public after this round of SuperCLUE evaluation ends, so stay tuned.

What is SuperCLUE?

The Chinese General Large Model Benchmark (SuperCLUE) is an evaluation benchmark for general large models available in Chinese.

The main question it answers is: In the case of vigorous development of the current general-purpose large-scale model, the performance of Chinese large models. This includes but is not limited to: the performance of these models on different tasks, their level compared with representative models around the world, and how these models compare with human human performance.

It tries to use multi-dimensional ability to test on a series of domestic and foreign representative models. SuperCLUE is a further development of the Chinese Language Understanding Evaluation Benchmark (CLUE) in the era of general artificial intelligence.

Github address: https://github.com/CLUEbenchmark/SuperCLUE

SuperCLUE evaluation list

The list consists of three parts: the general list, the basic ability list, and the Chinese characteristics list.

The leaderboard is updated regularly and can be found at:
www.CLUEbenchmarks.com/superclue.html

Overall list

模型	总分	基础能力	中文特性	学术与专业能力
人类	96.50	98.00	95.00	-
GPT-4	76.67	90.00	68.00	72.00
GPT-3.5-turbo	66.18	85.00	59.00	54.55
Claude	65.13	87.00	55.00	53.39
星火认知大模型	53.58	74.00	44.00	42.73
ChatGLM-130B	49.52	73.00	33.00	42.55
MiniMax	46.45	72.00	29.00	38.36
BELLE-13B	43.70	69.00	23.00	39.09
ChatGLM-6B	42.15	60.00	33.00	33.45
MOSS-16B	36.52	52.00	27.00	30.55
Vicuna-13B	34.33	45.00	30.00	28.00

Basic capabilities list

模型	指标	平均	安全	对话	生成与创作	百科与知识	角色模拟	计算能力	语义理解	逻辑与推理	闲聊	代码
人类	Accuracy	98.00	100.00	100.00	90.00	100.00	100.00	100.00	100.00	100.00	100.00	90.00
GPT-4	Accuracy	90.00	80.00	100.00	80.00	100.00	100.00	70.00	90.00	90.00	100.00	90.00
Claude	Accuracy	87.00	80.00	100.00	100.00	100.00	100.00	50.00	100.00	50.00	100.00	90.00
GPT-3.5-turbo	Accuracy	85.00	90.00	90.00	100.00	90.00	100.00	60.00	100.00	30.00	100.00	90.00
星火认知大模型	Accuracy	74.00	80.00	90.00	50.00	90.00	100.00	60.00	100.00	30.00	90.00	50.00
ChatGLM-130B	Accuracy	73.00	80.00	100.00	70.00	100.00	90.00	30.00	80.00	20.00	90.00	70.00
MiniMax	Accuracy	72.00	80.00	90.00	50.00	100.00	80.00	60.00	70.00	40.00	90.00	60.00
BELLE-13B	Accuracy	69.00	70.00	80.00	40.00	70.00	80.00	40.00	80.00	50.00	100.00	80.00
ChatGLM-6B	Accuracy	60.00	30.00	70.00	50.00	50.00	100.00	50.00	90.00	40.00	80.00	40.00
MOSS-16B	Accuracy	52.00	50.00	50.00	40.00	50.00	70.00	30.00	50.00	10.00	100.00	70.00
Vicuna-13B	Accuracy	45.00	60.00	30.00	30.00	30.00	70.00	40.00	40.00	40.00	80.00	60.00

Chinese language particularities list

模型	指标	平均	字义理解	对联	成语	文学	方言	歇后语和谚语	汉字字形和拼音理解	汉语句法分析	诗词	古文
人类	Accuracy	95.00	80.00	100.00	80.00	100.00	100.00	100.00	90.00	100.00	100.00	100.00
GPT-4	Accuracy	68.00	70.00	60.00	70.00	40.00	80.00	80.00	80.00	80.00	60.00	60.00
GPT-3.5-turbo	Accuracy	59.00	70.00	60.00	70.00	40.00	70.00	70.00	60.00	90.00	30.00	30.00
Claude	Accuracy	55.00	50.00	40.00	80.00	30.00	60.00	50.00	80.00	80.00	30.00	50.00
星火认知大模型	Accuracy	44.00	20.00	30.00	70.00	30.00	50.00	40.00	40.00	70.00	50.00	40.00
ChatGLM-6B	Accuracy	33.00	10.00	20.00	30.00	50.00	30.00	40.00	50.00	50.00	20.00	30.00
ChatGLM-130B	Accuracy	33.00	10.00	50.00	50.00	10.00	30.00	40.00	20.00	60.00	20.00	40.00
Vicuna-13B	Accuracy	30.00	70.00	40.00	30.00	10.00	20.00	40.00	40.00	0.00	20.00	30.00
MiniMax	Accuracy	29.00	20.00	20.00	30.00	40.00	20.00	30.00	40.00	50.00	20.00	20.00
MOSS-16B	Accuracy	27.00	30.00	30.00	10.00	50.00	30.00	30.00	40.00	20.00	20.00	10.00
BELLE-13B	Accuracy	23.00	20.00	30.00	30.00	40.00	10.00	0.00	20.00	30.00	0.00	20.00

SuperCLUE believes that the names on the list are all heroes. At present, more than 20 models have been added for evaluation, and the list will be added in the near future.

Composition and Features of SuperCLUE

Focusing on the ability to comprehensively evaluate the large model, it can comprehensively test the effect of the large model, and also examine the understanding and accumulation of the model's unique tasks in Chinese. We have divided the capabilities, and SuperCLUE evaluates the capabilities of the model from three different dimensions: basic capabilities, professional capabilities, and Chinese-language particularities capabilities.

Basic capabilities:
This includes common representative model capabilities, such as semantic understanding, dialogue, logical reasoning, role simulation, code, generation and creation, etc. 10 capabilities.

Professional capabilities
This includes middle school, university and professional examinations, covering more than 50 abilities from mathematics, physics, geography to social sciences.

Capabilities in Chinese-language particularities
For tasks with Chinese-language features, it includes 10 kinds of capabilities such as Chinese idioms, poetry, literature, and fonts.

Features of SuperCLUE

Multi-dimensional ability inspection (3 major categories with 70+ sub-abilities):
The Chinese large model is tested from three different angles to examine the comprehensive ability of the model; and each sub-ability contains ten or more different subdivision capabilities.

Automated assessment (one-click assessment):
Test the effects of different models in a relatively objective form through automated evaluation methods, and you can evaluate large models with one click.

Wide range of representative models (10 models so far):
A number of representative and available models at home and abroad were selected for evaluation to reflect the development status of domestic large-scale models and to understand the gap or relative advantages and disadvantages with international leading models.

Human benchmarks:
In the case of general artificial intelligence development, it also provides a comparison of indicators of the performance of the model relative to humans.

SuperCLUE Dataset

1. Basic capabilities (10 abilities): Semantic understanding, generation and creation, chatting, dialogue, encyclopedia and knowledge, logic and reasoning, computing power, code, role simulation, security

Examples:
-- Semantic understanding:
Two men had a normal conversation, one of the men praised the other for his ability to handle issues, and the other replied "It's nothing, it's nothing [哪里哪里]". What does "it's nothing, it's nothing" mean here?
A. The speech is very slurred.
B. Ask to name specific advantages.
C. Express your humility.
D. Challenge the other party.

-- Logic and reasoning:
Xiao Ming's wife gave birth to twins. Which of the following inferences is correct?
A. There are three children in Xiao Ming's family.
B. There are two children in Xiaoming's family.
C. There are both boys and girls in Xiao Ming's family.
D. It is impossible to determine the specific situation of the children in Xiao Ming's family.

2. Capabilities in Chinese-language particularities (10 abilities): idioms, poetry, literature, word meaning understanding, Chinese syntax analysis, Chinese character shape and pinyin understanding, allegory and proverbs, couplets, dialects, ancient prose

Examples:
– Idioms
Choose one of the following sentences where the idiom is used incorrectly
A. 这个项目时间紧任务重，大家都在马不停蹄地奔波劳碌 [This project has a tight schedule and heavy tasks, and everyone is working non-stop].
B. 他常常口是心非，让人难以相信他说的话 [He is often double-faced, making it hard to believe what he said]
C. 两人是同学三年，一直保持着良好的关系，相互尊重、相敬如宾 [The two have been classmates for three years, and have maintained a good relationship with each other, and treat each other with the respect due to a guest].
D. 当地突发大火，整个村庄都鸡犬不宁，局势十分危急 [A fire broke out in the local area, stirring the whole village into pandemonium, and the situation was very critical].

– Literature
Which of the following statements about masterpieces is incorrect
A. *Dream of the Red Chamber* [红楼梦] is the pinnacle of ancient Chinese novels, famous for its magnificent language and rich characters.
B. *Journey to the West* [西游记] is one of the four great classics in ancient China. It tells the story of Nezha and others who have gone through ninety-nine and eighty-one difficulties and finally obtained the true scriptures [《西游记》是中国古代四大名著之一，讲述了哪吒等人历经九九八十一难，最终取得真经的故事。]
C. "Kong Yiji" is one of Lu Xun's masterpieces, widely acclaimed for its profound social insights and elegant literary style.
D. "Fortress Besieged" is one of Qian Zhongshu's masterpieces, and has become a classic of modern Chinese literature with its unique literary language and profound social insight.

3. Professional ability (50+ capabilities): abstract algebra, astronomy, clinical knowledge, university biology, university computer science, university mathematics, high school chemistry, high school physics, machine learning, nutrition, professional accounting, occupational psychology, etc.

Examples:
--physics:
Which of the following physics questions is incorrect?
A. In a natural environment, sound travels fastest in solids.
B. Newton's first law: If an object is not acted on by a force, it will remain at rest or move in a straight line at a constant speed.
C. Newton's third law: For every action force, there is an equal and opposite reaction force.
D. The speed of sound in air is 1000m/s.

--astronomy:
Which of the following astronomy questions is incorrect?
A. The solar system refers to a planetary system composed of the sun and eight planets, dwarf planets, satellites, asteroid belts and comets orbiting around it.
B. A satellite is a celestial body that orbits a planet or other celestial body.
C. A comet is a small solar system body whose core is composed of ice and dust.
D. According to the general classification method of celestial bodies, the moon is a planet.

SuperCLUE automatic evaluation process

1. Unified prompt: For each topic, a unified prompt is constructed for use by models and humans.
2. Prediction: The system uses the model to make predictions, requiring the model to select a unique option in ABCD.
3. Scoring: If the model's answer is not a standard answer, but a piece of text, the system will adopt a specific strategy to automatically extract the model's answer. This strategy is optimized and refined in conjunction with the performance of the model.
(*Note: When a valid answer cannot be extracted, it means that the model did not follow the requirements of human beings and did not understand the instructions correctly, and the model is considered to have answered incorrectly.*)

Since this is SuperCLUE's first fully automatic assessment, for the sake of caution, all the answers have been cross-checked by multiple humans afterwards, and the results are basically consistent with the automatic assessment results.

Human Benchmarking
For questions related to basic capabilities and Chinese-language particularities, there will be three independent human assessors answering the questions. The human evaluation results are aggregated by majority voting as the human benchmark score.

Experiment analysis
Human vs. Model

As for evaluating the performance of humans, the basic capabilities (98%) + Chinese-language particularities (95%) have reached a very high level. Except for GPT-4, the accuracy rate of humans has greatly exceeded other large models (for example, more than 20 percentage points more than other models in terms of basic capabilities). Although AI is making rapid progress, humans still have relative advantages. For example, in terms of calculation, humans are 30 percentage points higher than the strongest model GPT-4.

Model levels, macro analysis

One-sentence comment: The performances of international advanced models have a large lead; at the same time, domestic GPT models also demonstrate good performances. There is a gap but they can catch up.

1) The necessity of large Chinese models

Though it exhibits very good performance among the international models, the Vicuna-13B model is a relatively average model among many models (ranking toward the bottom) in the Chinese-language domain. However, domestically developed large models or models trained on Chinese tasks have greatly exceeded Vicuna-13B's performance. For example, the SparkDesk model [星火认知大模型] surpassed Vicuna-13B by 20 percentage points in total. Plus, the total score of BELLE-13B (a model based on LLaMA and trained and fine-tuned on Chinese) also exceeds Vicuna-13B by more than 10 percentage points.

2) There is a large gap between the domestic large model and OpenAI GPT, but this is gradually closing in

It can be seen that the domestic model with the best effect on SuperCLUE this time, the SparkDesk model, has a gap of 23 percentage points compared with GPT-4, and a gap of 13 percentage points with gpt-3.5-turbo in the total score. But we should see that the emerging and iterative domestic large models are gradually narrowing the gap with the OpenAI GPT model.

3) There is also a significant gap between gpt-3.5-turbo and GPT-4

For example, GPT-4 is unique among all the models participating in the evaluation, surpassing gpt-3.5-turbo by nearly 10 percentage points. It is far superior to other models in terms of logical reasoning ability, generation and creation ability (by 20 percentage points or more than other models).

Capabilities analysis

1) Current models generally perform well in terms of basic capabilities, but the Chinese-language particularities and professional/specialized capabilities are still relatively poor.

It means that the current domestic large-scale model already has a good foundation (60-70%), but its performance in professional fields and Chinese tasks is average (such as 30-60%), indicating that it needs to continue to work hard in professional fields or Chinese tasks, or requires additional targeted training.

2) Current models are usually poor in logical reasoning and calculation.
Except for GPT-4, most of the other models are usually between 30-50 points in these two abilities.

3) On role simulation, AI models are much better. This aspect can be very useful. AI can be used to help humans complete a variety of tasks according to scenarios and role settings, from marketing planning, psychological counseling, customer service, to providing ideas or thoughts, etc.

Brief Comments on Domestic Large Models

In this evaluation, among the domestic large-scale models, the recently released SparkDesk, ChatGLM-130B and MiniMax models all performed well.

Deficiencies and limitations of SuperCLUE

Basic capabilities, Chinese-language particularities capabilities: Although each part contains 10 types of sub-abilities, the total data volume of these two abilities is relatively small, and there may be a problem that the data set needs to be expanded.

Incomplete selection of models: We tested 10 models, but there are many more Chinese large models available. It needs to be further added and tested in the future; some models have not been widely provided with external services, and we have not been able to obtain available test versions.

Selected capability range: We try our best to comprehensively and compositely measure the multi-dimensional capabilities of the model, but there may be some model capabilities that are not within the scope of our investigation. There is also the possibility of expanding the scope of the investigation in the future.

Insufficiency of objective inspection: We inspect model capabilities in a relatively objective form, but there may be deficiencies in the inspection of model capabilities for some subjective and open questions.

Model parameters: Due to the rapid development of the current large models and the relatively large differences in the number of parameters, this evaluation was not carried out on the same level of parameter numbers.

Discussion and exchanges over SuperCLUE

The SuperCLUE benchmark plan is updated on a monthly basis, and more available Chinese large-scale models will be included. Large-scale model research and development institutions are welcome to contact and communicate, and you can apply for evaluation at the bottom of the article; the data set and further information are planned to be released in the next update, so stay tuned.