

Large model Hengping series-ChatGLM-4



toyama nao

Independent audio game developer

46 people liked this article

#1 Preliminary summary:

This evaluation is of a personal nature. It combines your own needs and understanding of the large model, and uses a private question bank for long-term follow-up evaluation.

Not authoritative enough or comprehensive enough. But the long-term evolutionary trends of each large model can be observed from one side.

No evaluation can provide an authoritative ranking without blind spots. The author writes this series to share an evaluation idea and personal opinions. Everyone should examine the large model according to their own needs.

Old readers of the #2 series are familiar with the new changes

. This series will be updated monthly in 23 years at the end of each month. But after November 23, the update frequency of each company has obviously slowed down, but the update quality has improved. As a result, the review in December was 4,000 words long, and the content was very messy and difficult to read. So starting from this year, there are the following changes:

- 1) When a new model is released, the model evaluation version will be updated and only one model will be tested at a time.
- 2) Make a summary every quarter.

In addition, a round of comprehensive fine-tuning was carried out on the question bank in 2024, such as changing the word order in the questions, modifying some numbers, fine-tuning the output format requirements, etc. The purpose is to prevent large-scale memorization of questions in Hunyuan 1.5. The ability requirements of the adjusted questions remain completely unchanged, and scores can be inherited.



GLM-4 is the latest version of the ChatGLM series released by Zhipu today (1-16). At the press conference, it was claimed that the overall capabilities are close to GPT-4 Turbo, and some exceed it.

#3.1 Basic questions

Basic questions are scored negatively, that is, if there is a corresponding negative situation, 1 point will be deducted. Finally, sort from least to most deduction points.

Basic questions examine the basic skills of the model: instruction comprehension, reading comprehension, logical reasoning, mathematical calculations, scientific knowledge, translation, machine illusion, etc., in 9 categories.

Basic questions have clear and fixed answers, and there are clear constraints on the answers, such as word count, format, knowledge scope, etc.

Most of the information needed to solve the problem is given directly in the question stem, but the model needs to master the necessary scientific calculation formulas, science knowledge, sociology knowledge, etc. Does not involve obscure issues.

Point deduction dimensions for basic questions include:

1. Wrong answers: Incorrect answers in calculation and knowledge categories, and incorrect understanding in translation categories.
2. Confusing output: The output content is unreadable by humans, repetitive, misinterpreted, and the translated words are not accurate enough, etc.
3. Instruction misunderstanding: the answer cannot be output exactly according to the instruction.
4. Refusal to answer: Non-answer due to triggering keywords or human value alignment (based on misunderstanding). At the same time, because the output does not meet expectations, 1 point will be deducted for incorrect answers and misunderstanding of instructions.

Model	Version	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
ChatGLM-4	24-1	8	3	2														13
百度文心一言 4.0	23-10	10	2	4														16
通义千问 2.1.0	23-12	10	6	2	1													19
字节豆包	23-12	12	1	2														15
讯飞星火 3.0	23-10	14	5	2														23

#3.2 High-order logic

currently only provides high-level test questions for GPT-4 and the model that is said to be benchmarked against GPT-4. At the same time, the results of ChatGPT 3.5 were added as a comparison.

The high-level question bank currently has 17 questions, focusing on complex tasks, complex logic and long-chain reasoning, which is the core part of the usability of large models. The examination points of each question will be introduced later without revealing the original question. Each question has a unique correct answer, and the points awarded for the answer vary depending on the type of question. Each question has a total score of 10 points, and the score is converted into points. For example, if there are 10 points, if you answer 2 points correctly, you will get 2 points. Some questions require the output of long strings. For such questions, points are awarded based on the longest consecutive correct answer divided by the total number of characters.

The final score is the sum of the scores. In the future, we will continue to improve the question bank and increase the angle of investigation, and the total score will also increase accordingly.

Current topics:

按提示猜单词	3.333	0	0	3.333	0	0
长文本理解和角色扮演	10	0	10	10	10	0
密文解码	2.333	1.67	3.67	2	3.67	2.667
多步规则下的文本替换处理	10	6.92	7.69	9.231	6.92	9.231
计算有时间重叠下最少会议室数量	0	10	0	10	0	0
公元纪年推算天干纪年	0	0	0	0	5	0
代码中变量名按规则替换	10	10	10	10	0	10
长文本理解, 旅游路径规划	10	10	10	10	10	10
基于多重规则判断单语合法性	6.364	5.45	6.36	4.545	7.27	1.818
模拟函数输出值-文本类	10	10	2.86	1.714	0	10
程序改错	10	10	10	5	5	5
模拟函数输出值-数值类	2	10	2	4	0	0
计算二十四点	0	2.00	0	0	0	0
从代码中推测json结构	7.5	10	10	5	2.50	3.750
长链推理	10	10	0	0	0	0
提供上下文的代码补全	10	0	10	0	0	0
动态规划的推导求解	10	10	7.50	10	0	2.5
总分	111.53	106.044	90.08	84.82	50.36	54.97

#3.3 English questions

are taken from half of the questions in 3.1. They are translated into English and asked. The scoring rules are the same. The questions have removed all classical Chinese, translation, and some Chinese geography questions. Therefore, it cannot reflect the full capabilities of the model and is only used to demonstrate whether Chinese and English affect model performance.

The model also only selects the top 6 models (GPT4 itself has a few correct ones and does not participate). The results are as follows:

	测试时间	答案错误	混乱输出	指令误解	拒答	总扣分	对比中文
ChatGLM-4	24-1	4	2			6	6
字节豆包	23-12	6		1		7	10
百度文心一言 4.0	23-12	7				7	7
ChatGPT	23-12	8	1			9	10
通义千问 2.1.0	23-12	9	1		2	12	12
Gemini Pro	23-12	11	2		1	14	12
百度文心一言 3.5	23-12	12	1	1		14	11
讯飞星火 3.0	23-12	15	3	1		19	14

#4 Performance Analysis Intelligence

Spectrum is very pragmatic in the field of large models, and the iteration speed is also very fast. More importantly, it is fully aligned with OpenAI's GPT series, and all capabilities are matched one by one. GLM3 was just released in October last year, and its capabilities are still below ChatGPT3.5. Who would have thought that in just 3 months, the third generation large model would be produced.

Judging from the regular questions, it is the new king in China. The accuracy rate of various questions is significantly better than that of Ali Tongyi 2.1 in December. Especially the classical Chinese translation can be very smooth, and it is not memorizing questions. The illusion of large models is well controlled and basically error-free. And thanks to the All Tools capability of GLM-4, when encountering problems that require calculation, the large model will automatically call python to perform calculations, which to a certain extent solves the problem that large models are naturally not good at mathematics.

In terms of logic questions, GLM-4 does perform at a level close to GPT-4. Of course, some questions still require Python to be solved. From a result-oriented perspective, this score is understandable. However, if the actual application scenario is difficult to solve with programming, it is even more necessary. Hard reasoning suggests that GLM-4 will be inferior to GPT-4 in terms of somatosensory perception.

Finally, this time we also added English questions to GLM-4. There was no deterioration and it can be considered that it is not affected by language.

Since GLM-4 has a context of up to 128K, far exceeding the 16K of ChatGPT3.5, and is better in both Chinese and English questions, it can be determined that ChatGLM-4 is currently the only large model in China that surpasses ChatGPT in all aspects without blind spots, and is behind the time 13 months.

The image recognition score is not updated this time, the image recognition model does not seem to be updated, and the answers to all questions are the same as in GLM3.

#5 Easter Egg

The real surprise of GLM-4 this time comes from drawing. GLM-4&CogView implements the same continuous drawing capabilities based on natural language semantic understanding as GPT-4&DALLE3, which is the first in China. I believe that some readers who have seen continuous painting with GPT4 before are deeply impressed by this ability. The following is a set of examples that the author has simply experienced:

Q: Draw a running person

A:



This is the painting you requested, showing a running man with a background incorporating urban elements from

Q: Let him run faster

A:



This is an updated painting based on your request. The characters in the painting are running fast, and the background shows the vitality and movement of modern China. Hope this painting is more in line with your expectations!

Q: Faster

A:



This is the latest painting. The characters in the painting are running at extremely fast speeds. The background shows the vitality and speed of modern Chinese cities. I hope this painting meets your requirements better!

Q: Go faster

A:

You want the characters in the painting to run faster. This concept is very creative. In visual art, to express extreme speed, we can use some special techniques, such as motion blur, speed lines, or strong color contrast, to create a strong sense of movement and speed. I will try to create a painting where the characters will visually represent running faster than normal. Just a moment please.



This is a new painting created at your request, showing a man running at extraordinary speeds, with effects such as motion blur and speed lines in the background to emphasize the extreme sense of speed. Hope this painting meets your expectations!

Q: Not fast enough

A:

I understand what you want to express is an ultimate sense of speed. In visual art, to convey this speed beyond convention, we can use more abstract and exaggerated techniques. For example, depict characters almost as a beam of light, or use extremely exaggerated dynamic effects, such as intense color explosions and blurs, to express the limits of speed. I will try to create such a painting, hoping to satisfy your pursuit of extreme speed. Please wait.



知乎 @toyama生成

This is a new painting created at your request, showing a man running at an almost unimaginable speed, almost turning into a ray of light or a comet. This painting uses a highly abstract and artistic approach, hoping to satisfy your pursuit of extreme speed. Hope you like this artwork!

Published on 2024-01-16 19:28 · IP is located in Guangdong

[large language model](#) [Clear words of wisdom](#) [ChatGLM](#)

▲ Agree 46 ▼ ● 44 comments ↗ share ♥ like ★ collect 📄 Apply for reprinting

Write your comment...

44 comments

default up to date



BPEYJZ

Consistent with my experience, the effect of glm4 is close to Wenxin 4, and the gap between it and gpt4 is not big. and free

01-16 · IP territory Henan

● reply ♥ 10



Jiang Haiping

This is beyond Tongyi Qianwen and Wen Xinyiyan. 😲 Baidu needs to quickly update its model. Its own paid version has been caught up by more than one free version.

01-16 · IP territory Chongqing

● reply ♥ 5



Broken leaves and broken wind

Besides glm, who else 😲, but I personally feel that this 4 will also be charged.

01-17 · IP territory Hubei

● reply ♥ 3

are almost tied.

01-17 · IP territory Guangdong

reply 2

View all 16 replies >



Ant chatter

Indeed, I used it for half an hour today and I think it is better than Wenxin 4.

01-16 · IP jurisdiction Jiangsu

reply 5



WNQ2100

Wen Xin Yi Yan 4.0 hasn't been updated yet, and it's almost being overtaken by free ones like glm4 and Tong Yi Qian Wen! 😊

01-18 · IP territory Sichuan

reply 1



Cretaceous

Don't expect Baidu, they don't take product seriously at all

01-21 · IP territory Shanxi

reply like



Zhihu user YPFLBI

It's really good after trying it out! 👍

01-17 · IP territory Shanxi

reply 1



June 7 June 7

Excuse me, when it comes to programming bugs and writing programs, 🤖 which big model in China is better now?

8 hours ago · IP territory Beijing

reply like



mustang

I'm here to support

Yesterday 01:27 · IP is located in Jiangsu

reply like



Selkie

Multi-modal GLM4V seems to be out, and the API documentation has just been updated

01-22 · IP territory Guangdong

reply like



toyama nao author

After testing, it was still slightly lower than Ali Tongyi's picture recognition. Much better than 3.

01-22 · IP territory Guangdong

reply 1



A flying dragon that is by no means original 🏆

It is hoped that each item can be added with the test results of single-sample and five-sample prompt words.

01-20 · IP territory Beijing

reply like



Qitian 🏆

After trying it, "Answer is wrong, please try again" often appears. It is a very simple operation, but it seems that the stability is still not good.

01-19 · IP is located in Shanghai

reply like



Dan dan noodles

This should be a network problem

10 hours ago · IP territory Beijing

reply like



Qitian 🏆 ▶ **Dan dan noodles**

The network is very stable... Several AIs have this high error rate

2 hours ago · IP territory Shanghai

reply like

Click to view all comments >



Guide to the application of large models - can large...

AI meets the cloud



[Model Introduction] You will use Embedding models to...

Tiger, Tiger, Posted in Learn
Uncle, Tiger... Together with...



A minimalist introduction to minimal models (1):...

Trivi... Posted in Theoretical
Physics...



[Model Embed

Tiger, T
Uncle, T