

Forum:[Hack Forums](#)

Thread: [GLM 5.2 Might Be The Best Model For Vibe Coding](#)

JoseMar - June 18, 2026, 4:21 pm

GLM 5.2 Might Be The Best Model
For Vibe Coding

? loading encrypted image...
Hello again HackForums ?

Today I made a little more time for another thread, because I saw an update that honestly deserves an article and I hope you will enjoy it.

So, the Chinese company Z.ai made an update from GLM 5.1 to GLM 5.2, it is an AI model that, according to independent evaluations, is positioned at the top. It is a powerful model that you can run locally, without depending on a single external provider, but YES, you need a computer that can help you do that.

I will explain what GLM 5.2 actually is, and why it deserves attention:

GLM 5.2 is an LLM developed by Z.ai and released under the MIT license, which also allows commercial use. This means that you can download it and run it locally.

The company Z.ai announced the model on June 13, 2026, which was initially offered to subscribers and a few days later, it became publicly available.

So, GLM 5.2 is a model that thinks step by step before answering us. It has ~753 billion parameters in total, but activates only about 40 billion for each processed token.

Specifications confirmed for GLM 5.2:Parameters: ~753 billion total, ~40 billion active/token
1 million tokens
Maximum output: up to 131,072 tokens per response
This is text --> text only (meaning, it does not create images)
License: MIT (commercial use allowed)
Public, available on Hugging Face

What is the most spectacular thing? 1 million tokens,? so in practice, better said, it is approximately five times more than 5.1

The performance of GLM 5.2 is, according to Artificial Analysis, the smartest public model, meaning you can run it locally and do not have to pay.

Also according to Artificial Analysis, GLM 5.2 achieves:75% on Terminal-Bench
89% on GPQA Diamond
71% on the long-context reasoning test.

? loading encrypted image...

This means that it is a model built for programming and multi-step workflows, where the AI writes, runs and reviews code throughout an entire project.

Let me tell you the speed: this update delivers approximately 112 tokens per second, above the

average of its class, so it is fast in practical use.

What does the MIT license mean?

The MIT license means, in practical terms, freedom and control. You can use the model:
commercially
you can modify it
you can adapt it
you can run it locally

For a business, this difference is more than a legal nuance. When you build something important on a proprietary model accessible only through an API, you depend on the decisions of an external provider: price
availability
access conditions.

A model like this one you can install once and run it for days, months or years if that is what you want

Local AI Without Limits ?

A model such as GLM 5.2 can be hosted on company servers or locally if only you use it. So documents, conversations and sensitive information are not sent to an external service under another jurisdiction. For regulated sectors, public administration, healthcare, legal services, finance... etc

How much does it cost and how can you access GLM 5.2?

GLM 5.2 can be accessed in several ways:
downloaded for free and run locally, as I already told you,
used through the Z.ai API or accessed through a monthly subscription to the company's coding plan.

Being under the MIT license, the "basic" cost of running it yourself is only the cost of PC resources.

For access through the API hosted by Z.ai, the prices reported by Artificial Analysis are approximately \$1.40 per million input tokens and \$4.40 per million output tokens, with a significant discount when reusing context (cache). The same source notes that, for its class of open models, GLM 5.2 is relatively expensive through the API, however, it remains well below the cost of top proprietary models.

One detail that I believe is worth keeping in mind: the model is relatively "talkative", meaning it uses more tokens to answer than average. At a cost per token, this translates directly into the bill (meaning you pay for something you should not, it is like a woman when she is angry so to speak... she talks a lot... :))))), so for large-scale use, optimizing the way you prompt it matters.

With a model like GLM 5.2 you can build ? : Practical AI tools integrated directly into your workflows.

Coding assistants and development automation, the model is built explicitly for programming and multi-step tasks.

Smart search in your own documents, the large context window allows analysis of serious volumes of text in a single query.

Virtual assistants for websites and customer relations, let's say a chatbot that answers based on real information.

Internal automations, message classification, summarization, generation of standard responses.
and many more things, you only need imagination...

? loading encrypted image...
Is GLM 5.2 Worth It ?

WAIT, we also need to be careful with GLM 5.2, because it is a really powerful model, but it is not suitable for everything or EVERYONE. It processes only text, it does not "see" images and it does not generate images or video.

My sincere and practical recommendation: do not choose a model "because it is trendy", but because it fits your use case. For tasks such as automation and working with your own documents, under data control conditions, GLM 5.2 is a serious and actually good option plus vibe-coding.

Now you should also know that GLM 5.2 compared to top models and according to Artificial Analysis, is the smartest text to text model that can be run locally or on a server, which stands out a lot at vibe-coding. On the GENERAL intelligence index it is surpassed by a few top proprietary models. but for those you need some \$\$.

Links:<https://docs.z.ai/guides/llm/glm-5.2>
<https://huggingface.co/blog/zai-org/glm-52-blog>

Entering The Brotherhood Tutorial Contest.?
[7 JoshAllen](#) [DemonAsylum](#) [Dystortion](#)

MadScientist - June 18, 2026, 5:21 pm

Going to try it out, just setup ComfyUI and was installing a few different LLMs to try out. Namely Claude

JoseMar - June 18, 2026, 5:33 pm

(Jun 18, 2026 05:21 PM)MadScientist Wrote: [\[CITE\]](#) Going to try it out, just setup ComfyUI and was installing a few different LLMs to try out. Namely Claude

Let me know what you think once you've tried it. ? Honestly, if you use it for vibe coding, I think you'll be surprised.?

Freemasons - June 19, 2026, 6:04 am

(Jun 18, 2026 05:33 PM)JoseMar Wrote: [\[CITE\]](#) Let me know what you think once you've tried it. ? Honestly, if you use it for vibe coding, I think you'll be surprised.?
Is the a uncensored LLM or still filtered?

SIM-2470866 - June 19, 2026, 7:14 am

Just picked up a MAX Package on them, gonna give it a shot. I was normally using Gemini & Claude.

Gonna take it for a spin the entire week now. Lets see how good it is

JoseMar - June 19, 2026, 7:54 am

(Jun 19, 2026 06:04 AM)Freemasons Wrote: [\[CITE\]](#) Is the a uncensored LLM or still filtered?
Depends on how you run it. Via the API, you're gonna hit some censorship wall. But run it locally on your own PC, and it's a whole different game.

(Jun 19, 2026 07:14 AM)SIM-2470866 Wrote: [\[CITE\]](#) Just picked up a MAX Package on them, gonna give it a shot. I was normally using Gemini & Claude.

Gonna take it for a spin the entire week now. Lets see how good it is
You will definitely notice a big difference, especially with MAX. Don't forget to let us know what you think about it.?

DramaKing - June 20, 2026, 4:35 am

The model really is impressive. It beats GPT 5.5 and trades blows with Opus 4.8. Design Arena even has it pulling ahead of Fable 5. IndexShare also looks good.

However, you are entirely wrong about being able to run a 744B MoE model on a computer. The download size at INT4 quantization is still 400-500 GB. A new Vera Rubin GPU from NVIDIA won't be able to run it. Even AMD's upcoming Radeon Instinct MI455X will need the model reduced to fit in the gigantic framebuffer.

(Jun 19, 2026 06:04 AM)Freemasons Wrote: [\[CITE\]](#) Is the a uncensored LLM or still filtered? All LLMs have guardrails, and GLM is built by a Chinese company, so while safety is not a priority, CCP censorship is. The model on HuggingFace from Z.ai is no different from the cloud model.

JoseMar - June 20, 2026, 5:15 am

(Jun 18, 2026 04:21 PM)JoseMar Wrote: [\[CITE\]](#) It is a powerful model that you can run locally, without depending on a single external provider, but YES, you need a computer that can help you do that.

(Jun 20, 2026 04:35 AM)DramaKing Wrote: [\[CITE\]](#) The model really is impressive. It beats GPT 5.5 and trades blows with Opus 4.8. Design Arena even has it pulling ahead of Fable 5. IndexShare also looks good.

However, you are entirely wrong about being able to run a 744B MoE model on a computer. The download size at INT4 quantization is still 400-500 GB. A new Vera Rubin GPU from NVIDIA won't be able to run it. Even AMD's upcoming Radeon Instinct MI455X will need the model reduced to fit in the gigantic framebuffer.

Thanks for the correction, I appreciate it. That's a fair clarification.??

SIM-2502833 - June 20, 2026, 8:42 am

IDK I just hopped over to it since i hit my claude token limit and I was oing a pretty complex backend rework for a game system, and it keeps looping on itself trying to figure out what's going on. It's used 2% of my 5 hour usage and still hasn't even done anything yet on the \$72/mo plan.

DramaKing - June 20, 2026, 8:56 am

(Jun 20, 2026 08:42 AM)SIM-2502833 Wrote: [\[CITE\]](#) IDK I just hopped over to it since i hit my claude token limit and I was oing a pretty complex backend rework for a game system, and it keeps looping on itself trying to figure out what's going on. It's used 2% of my 5 hour usage and still hasn't even done anything yet on the \$72/mo plan.

You have a Git repo for the game?

SIM-2502833 - June 20, 2026, 9:03 am

(Jun 20, 2026 08:56 AM)DramaKing Wrote: [\[CITE\]](#) You have a Git repo for the game?

<https://github.com/Eric-Gurt/StarDefenders2D>

I was in the middle of a rework for flying bases, current code moves every item individually, claude was about halfway through rewriting so that it would create a giant static object and move it as a whole except edge cases.

Edit: Ended up giving up on 5.2 and telling it to revert everything how it was. I had been using GPT 5 for auditing and writing tasks and PRDs, then having claude implement them. I decided to give 5.2 a try because all the code was written for a dry run and it just needed to implement it, but it couldn't figure it out.

JoseMar - June 20, 2026, 10:08 am

(Jun 20, 2026 09:03 AM)SIM-2502833 Wrote: [\[CITE\]](#) <https://github.com/Eric-Gurt/StarDefenders2D>

I was in the middle of a rework for flying bases, current code moves every item individually, claude was about halfway through rewriting so that it would create a giant static object and move it as a whole except edge cases.

Edit: Ended up giving up on 5.2 and telling it to revert everything how it was. I had been using GPT 5 for auditing and writing tasks and PRDs, then having claude implement them. I decided to give 5.2 a try because all the code was written for a dry run and it just needed to implement it, but it couldn't figure it out.

Personally, I recommend MAX. Plus, from what I've seen, pretty much everyone online is saying that MAX is the best option for vibe coding.

But, it will still make mistakes. That said, like many others, I find it to be quite good overall.?

SIM-2502833 - June 20, 2026, 8:31 pm

(Jun 20, 2026 10:08 AM)JoseMar Wrote: [\[CITE\]](#) Personally, I recommend MAX. Plus, from what I've seen, pretty much everyone online is saying that MAX is the best option for vibe coding.

But, it will still make mistakes. That said, like many others, I find it to be quite good overall.?Unless you're referring to claude max

The only difference between the two is 80 dollars and the amount of tokens you can spend. I am fully aware of how tokens work, I don't need more tokens for an API that spends 3m+ tokens just looping over the clavix PRD without making a single code change. During my test run, (which I only decided to do because everyone said it was trading blows with Opus) it spend 3M tokens just analyzing code, saying 'Oh, I understand now!' then saying 'Oh wait!' Then looping back to where it was. It made 3 file changes over 3 mil tokens, and then resumed it's loop. I ended up just telling it to revert everything and let Claude handle it. I'll let my subscription run out at the end of the month and milk it out on projects that interns could complete.

JoseMar - June 21, 2026, 1:20 am

(Jun 20, 2026 08:31 PM)SIM-2502833 Wrote: [\[CITE\]](#) Unless you're referring to claude max

The only difference between the two is 80 dollars and the amount of tokens you can spend. I am fully aware of how tokens work, I don't need more tokens for an API that spends 3m+ tokens just looping over the clavix PRD without making a single code change. During my test run, (which I only decided to do because everyone said it was trading blows with Opus) it spend 3M tokens just analyzing code, saying 'Oh, I understand now!' then saying 'Oh wait!' Then looping back to where it was. It made 3 file changes over 3 mil tokens, and then resumed it's loop. I ended up just telling it to revert everything and let Claude handle it. I'll let my subscription run out at the end of the month and milk it out on projects that interns could complete.

GLM-5.2 Max?

[Video: <https://www.youtube.com/watch?v=VDZIAA9pWNE>]

securer - June 28, 2026, 7:58 am

This is very true. I canceled my claude and chatgpt for this one. I'm also selling credits in GLM 5.2 for 50% off. Willing to give a vouch to uber and up. I just need your userid. :)

(Jun 18, 2026 04:21 PM)JoseMar Wrote: [\[CITE\]](#) GLM 5.2 Might Be The Best Model For Vibe Coding

? loading encrypted image...

Hello again HackForums ?

Today I made a little more time for another thread, because I saw an update that honestly deserves

an article and I hope you will enjoy it.

So, the Chinese company Z.ai made an update from GLM 5.1 to GLM 5.2, it is an AI model that, according to independent evaluations, is positioned at the top. It is a powerful model that you can run locally, without depending on a single external provider, but YES, you need a computer that can help you do that.

I will explain what GLM 5.2 actually is, and why it deserves attention:

GLM 5.2 is an LLM developed by Z.ai and released under the MIT license, which also allows commercial use. This means that you can download it and run it locally.

The company Z.ai announced the model on June 13, 2026, which was initially offered to subscribers and a few days later, it became publicly available.

So, GLM 5.2 is a model that thinks step by step before answering us. It has ~753 billion parameters in total, but activates only about 40 billion for each processed token.

Specifications confirmed for GLM 5.2: Parameters: ~753 billion total, ~40 billion active/token
1 million tokens

Maximum output: up to 131,072 tokens per response

This is text --> text only (meaning, it does not create images)

License: MIT (commercial use allowed)

Public, available on Hugging Face

What is the most spectacular thing? 1 million tokens,? so in practice, better said, it is approximately five times more than 5.1

The performance of GLM 5.2 is, according to Artificial Analysis, the smartest public model, meaning you can run it locally and do not have to pay.

Also according to Artificial Analysis, GLM 5.2 achieves: 75% on Terminal-Bench

89% on GPQA Diamond

71% on the long-context reasoning test.

? loading encrypted image...

This means that it is a model built for programming and multi-step workflows, where the AI writes, runs and reviews code throughout an entire project.

Let me tell you the speed: this update delivers approximately 112 tokens per second, above the average of its class, so it is fast in practical use.

What does the MIT license mean?

The MIT license means, in practical terms, freedom and control. You can use the model: commercially

you can modify it

you can adapt it

you can run it locally

For a business, this difference is more than a legal nuance. When you build something important on a proprietary model accessible only through an API, you depend on the decisions of an external provider: price

availability

access conditions.

A model like this one you can install once and run it for days, months or years if that is what you

want

Local AI Without Limits ?

A model such as GLM 5.2 can be hosted on company servers or locally if only you use it. So documents, conversations and sensitive information are not sent to an external service under another jurisdiction. For regulated sectors, public administration, healthcare, legal services, finance... etc

How much does it cost and how can you access GLM 5.2?

GLM 5.2 can be accessed in several ways:

downloaded for free and run locally, as I already told you,

used through the Z.ai API or accessed through a monthly subscription to the company's coding plan.

Being under the MIT license, the "basic" cost of running it yourself is only the cost of PC resources.

For access through the API hosted by Z.ai, the prices reported by Artificial Analysis are approximately \$1.40 per million input tokens and \$4.40 per million output tokens, with a significant discount when reusing context (cache). The same source notes that, for its class of open models, GLM 5.2 is relatively expensive through the API, however, it remains well below the cost of top proprietary models.

One detail that I believe is worth keeping in mind: the model is relatively "talkative", meaning it uses more tokens to answer than average. At a cost per token, this translates directly into the bill (meaning you pay for something you should not, it is like a woman when she is angry so to speak... she talks a lot... :))), so for large-scale use, optimizing the way you prompt it matters.

With a model like GLM 5.2 you can build ? Practical AI tools integrated directly into your workflows.

Coding assistants and development automation, the model is built explicitly for programming and multi-step tasks.

Smart search in your own documents, the large context window allows analysis of serious volumes of text in a single query.

Virtual assistants for websites and customer relations, let's say a chatbot that answers based on real information.

Internal automations, message classification, summarization, generation of standard responses. and many more things, you only need imagination...

? loading encrypted image...

Is GLM 5.2 Worth It ?

WAIT, we also need to be careful with GLM 5.2, because it is a really powerful model, but it is not suitable for everything or EVERYONE. It processes only text, it does not "see" images and it does not generate images or video.

My sincere and practical recommendation: do not choose a model "because it is trendy", but because it fits your use case. For tasks such as automation and working with your own documents, under data control conditions, GLM 5.2 is a serious and actually good option plus vibe-coding.

Now you should also know that GLM 5.2 compared to top models and according to Artificial Analysis, is the smartest text to text model that can be run locally or on a server, which stands out a lot at vibe-coding. On the GENERAL intelligence index it is surpassed by a few top proprietary

models. but for those you need some \$\$.

Links: <https://docs.z.ai/guides/llm/glm-5.2>
<https://huggingface.co/blog/zai-org/glm-52-blog>

Entering The Brotherhood Tutorial Contest.?

[7 JoshAllen](#) [DemonAsylum](#) [Dystortion](#)

CL1B - June 28, 2026, 10:52 am

What a great project [SIM-2502833](#)

I just inspected the repo and played the game for a bit, you should push this out onto a game platform!