

GLM-5.2 vs Claude Sonnet 5: Which AI Model Should You Actually Use in 2026?

TechRounder PDF Edition

Live article:

<https://www.techrounder.com/ai/glm-5-2-vs-claude-sonnet-5-which-ai-model-should-you-actually-use-in-2026/>

By Vipin PG | Published July 1, 2026 | Updated October 7, 2026 | Format: Comparison | 4 min read

Bottom line

If you need an AI that can see images, read screenshots, or reliably handle multi-step business tasks, go with Claude Sonnet 5. If you mainly need heavy-duty coding, math, or research work done at the lowest possible cost - and don't mind it being text-only - GLM-5.2 is the smarter pick, often cutting costs by 70-80%.

Two major AI models launched within weeks of each other in mid-2026, Zhipu AI's GLM-5.2 and Anthropic's Claude Sonnet 5. Both are being used heavily for coding, research, and automation, but they're built differently and priced differently too. If you're trying to figure out which one actually makes sense for your work, here's a breakdown in plain terms, without the technical fluff.

What Are GLM-5.2 and Claude Sonnet 5?

GLM-5.2 comes from Zhipu AI, a Chinese AI company, and was released in mid-June 2026. What makes it stand out is that it's open-weight, meaning anyone can download it and run it on their own hardware, completely free of licensing restrictions. Claude Sonnet 5 comes from Anthropic and launched at the end of June 2026. Unlike GLM-5.2, it's a closed, proprietary model, you can only access it through Anthropic's API or apps that use it, like Claude.ai, Claude Code, or cloud platforms such as AWS and Google Cloud. Both models support large context windows (around 1 million tokens, meaning they can "read" huge amounts of text in one go), but beyond that, they take different approaches.

Pricing: Which One Actually Costs Less?

This is where the two models diverge the most. GLM-5.2 is dramatically cheaper on paper.

Model: GLM-5.2 | Input Cost (per 1M tokens): \$1.40 | Output Cost (per 1M tokens): \$4.40

Model: Claude Sonnet 5 (promo, until Aug 31, 2026) | Input Cost (per 1M tokens): \$2.00 | Output Cost (per 1M tokens): \$10.00

Model: Claude Sonnet 5 (standard pricing) | Input Cost (per 1M tokens): \$3.00 | Output Cost (per 1M tokens): \$15.00

At standard rates, Sonnet 5's output is more than three times pricier than GLM-5.2's. For a business running large volumes of AI tasks daily, that difference adds up fast, potentially tens of dollars a day, which becomes thousands over a year. But there's a catch on both sides. Claude Sonnet 5 uses a new way of counting text (called a tokenizer) that counts roughly 30% more tokens for the same amount of writing compared to its previous version, quietly pushing up real costs. GLM-5.2, on the other hand, tends to "overthink", generating far more output text than necessary even for simple questions, which eats into its price advantage. So neither sticker price tells the whole story; the actual cost depends on how each model behaves in practice.

Coding and Problem-Solving

Both models are genuinely strong at coding , this is one area where the gap has nearly closed. On real-world coding tests that measure how well an AI can fix actual GitHub issues, Claude Sonnet 5 scores slightly higher than GLM-5.2. When it comes to working directly in a terminal (writing scripts, managing files, running commands), the two are essentially tied, both performing at an elite level. Where Sonnet 5 pulls ahead is in producing code that's genuinely ready to merge into a real project without extra cleanup , useful for teams shipping production software. GLM-5.2 is still the strongest coding model available in the open-source space, easily beating other free alternatives.

Math and Scientific Reasoning

This is where GLM-5.2 clearly outperforms Claude Sonnet 5. On tough, PhD-level science questions and competition-level math problems, GLM-5.2 scores significantly higher , by a wide margin in some tests. If your work involves heavy research, scientific analysis, or advanced math, GLM-5.2 has a real edge here. That said, Claude Sonnet 5 tends to do better when a task requires using outside tools step by step , like searching, checking results, and adjusting its answer along the way , rather than solving a problem purely through internal reasoning.

Can They "See" Images? Vision Capabilities

This is one of the biggest practical differences between the two. Claude Sonnet 5 is fully multimodal , it can read images, look at screenshots, understand documents, and even control a computer screen by seeing what's on it and clicking through interfaces on its own. GLM-5.2 is text-only. It has no ability to process images at all. If you try to show it a screenshot or a photo, it simply can't work with it. This makes GLM-5.2 unsuitable for tasks like visual debugging, reading charts, or automating anything that involves looking at a screen. If your work involves images, PDFs with visuals, or automating tasks inside apps and browsers, Claude Sonnet 5 is the only real option between the two.

Safety, Security, and Restrictions

Anthropic has built Claude Sonnet 5 with heavy safety restrictions, especially around cybersecurity. It's intentionally limited when it comes to finding or creating security exploits, making it safe to use inside a company but not useful for security research work. GLM-5.2, being fully open-weight, has no such built-in restrictions. This makes it genuinely useful for legitimate security research and vulnerability testing, but it also means there are fewer guardrails in place, so it needs to be used responsibly.

So, Which One Should You Choose?

Go with Claude Sonnet 5 if:

- You need the AI to read images, screenshots, or visual documents
- You're automating tasks inside apps, browsers, or desktop software
- You need strong built-in safety and security compliance for business use
- You want easy integration with AWS, Google Cloud, or Microsoft's cloud tools

Go with GLM-5.2 if:

- You're running high-volume tasks and want to keep costs as low as possible
- Your work is mostly text-based , coding, research, data processing
- You want the flexibility to run the model on your own servers
- You need strong performance on math and scientific reasoning

Final Thoughts

Neither model is a clear winner across the board , they're built for different priorities. Claude Sonnet 5 is the safer, more polished choice for businesses that need reliability and visual understanding. GLM-5.2 is the better choice if cost efficiency and raw reasoning power matter more than convenience. Many teams are actually starting to use both , Claude Sonnet 5 for tasks involving visuals or client-facing work, and GLM-5.2 for the heavy, high-volume text processing in the background. That combination gets you the strengths of both without being stuck with the weaknesses of either.

References

1. z.ai - blog / glm-5.2 - <https://z.ai/blog/glm-5.2>