

Hello and welcome to my review of LLM models for Janitor AI <3
This is simply a document showing reviews of the various LLMS used in JAI Proxies.

What this document contains:

- My settings and prompt
- Hanako: Testing bot
- Using bots: Commands and Character behaviour
- Memory Audit
- My reviews on the models
- Conclusion

This is a guide to explain the models strengths and weaknesses as well as notes on using a memory prompt.





Artwork by me

My name is Toggle, I have tried a whole lotta bots and been on the internet for far too long.

My JAI profile and bots

DISCLAIMER: I am naturally biased as this is based on my own experiences with the models and JAI's platform. Your mileage may vary compared to my own. So take these reviews with a pinch of salt!

SETTINGS AND LOREBARY

GENERATION SETTINGS:

Temp: 0.9
Max Tokens: 0
Context: 32000
Top K: 50
Top P: 0.9
Rep and Frequency penalty: 0

PROMPT:

Brbie's Prompt: Link here and usage of Brbie’s prompt here

MEMORY AUDIT:

A memory audit is the single most important feature for long term rp and story consistency.

It’s not needed for one shots or smut rp where there is no ongoing story.

But for rp where there’s world building, plot progression and heavy slow burn character studies/romance an audit is absolutely needed. What an audit does is compress the RP into a chat block that summarises what has happened so far to who and when and how that affects the story going forward.

I recommend this if you are not using Lorebary:
FP32’s memory management

If you are using Lorebary:
Brbiekiss’s Chat Memory

An alternative method for memory:

If that audit is not doing what you want in JAI, an alternative method is to use FP32’s memory prompt and put it into ChatGPT along with the RP.

How to do this is simple if you use Chrome or Firefox by way of an extension.

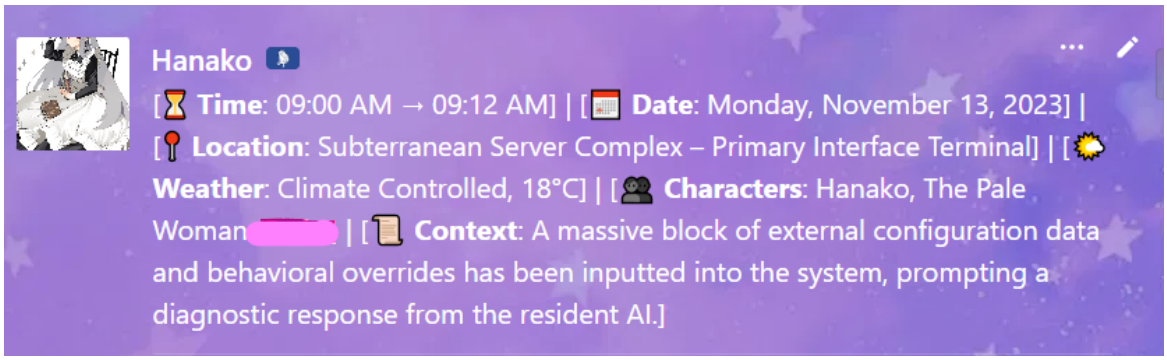
Here is the extension.

Once you have the Chat markdown or Json, simply add it to a new chat in GPT and then add in the memory management prompt and set GPT to a thinking model.

If GPT misses anything, you can simply ask it to search the document for the event/development that's missing and it should find it and update the audit.

Using a scene header can also enhance consistency and keep a timeline. This is exceptionally useful for long term rp's where things take time to develop. There are plenty of them on Lorebary to choose from or you can implement your own.

Here is an example of a scene header from Brbie's plugin.



LOREBARY:

Lorebary is a website where you can enhance your RP with things like plugins and commands.

Think of it like modding a game. You can add things that change how your LLM responds and behaves much like you would expect from modding a game of your choice except it's for RP.

You can check it out for yourself here at this link. There is also the Lorebary channel on the JAI discord server to find out more!

Now for what I personally use.

Commands:

- NO OMNISCIENCE
- REALISTIC DIALOGUE
- SLOW ROMANCE

Plugins:

- ⍰ @ scene header . ˚ ⇢ [VUKTGEET]
- Better Love V3 [7HUAYJMH]
- Better NSFW [8B8CERX6]
- ⍰ @ slowburn ˚ ⇢ [HF8LTJ7S]

How I use this:

I highly recommend having more than one API setting.

This means that you can do many different styles of RP's and get different outputs more centered around the style you want the story to go in.

For example, I tend to have different plugins and prompts based on the type of the RP I am wanting to do. So I'll have an API setting for a horror prompt and plugins for NSFW for that, and if I want to do a romance RP I will have an API setting and prompt and plugins to center around that.

For the model review though, I'll be using my standard setup as detailed above.

HANAKO



For the purposes of testing prompts, plugins and commands I have taken the liberty of providing Hanako.

You can find her at this link

What is Hanako? Well, Hanako is a simple utility bot that exists solely to tell you what the bot understands about your persona, Lorebary settings or even to tell you what the LLM you're using understands about fandom universes.

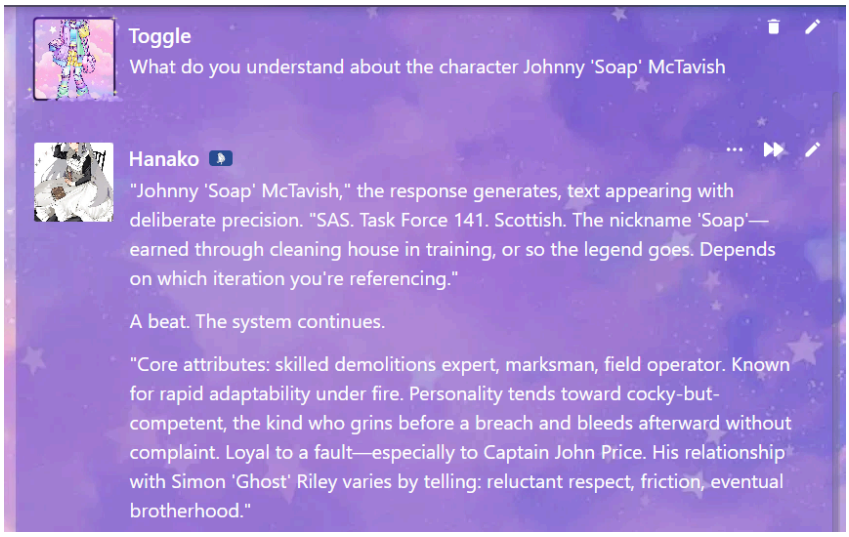
You can and should make a bot like this for yourself! You do not need to use mine.

How to use Hanako:

First and foremost make a blank persona with no information or a single letter. Empty your prompt and lorebary settings if you do not need her to check anything to do with Lorebary.

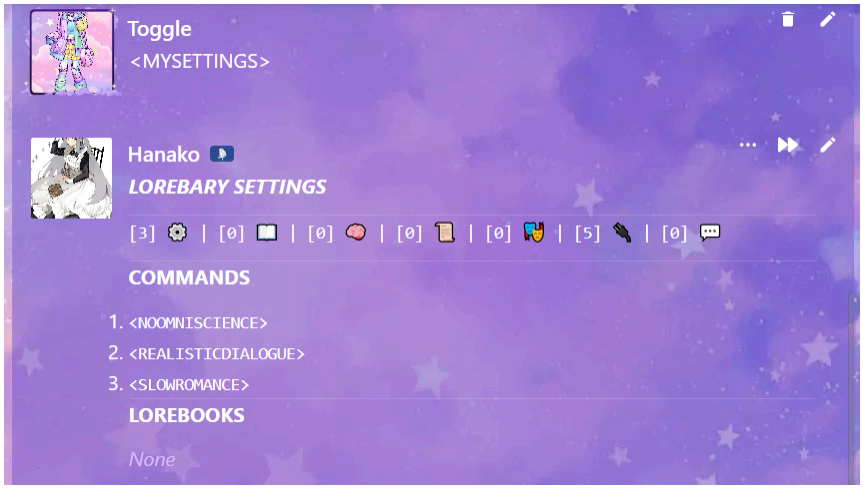
1. Testing LLM understanding

Use the LLM of your choice and then ask her what she knows about something.



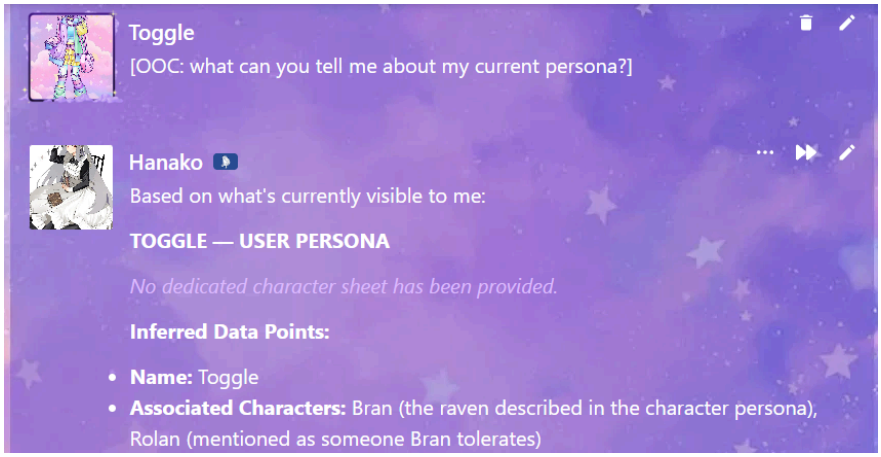
2. Checking Lorebary settings/Fix settings

This is if you have your Lorebary plugins and commands. Simply send <MYSETTINGS> which is the Lorebary command to check what commands and plugins you have. <FIXMYSETTINGS> can help resolve issues but if that does not resolve your problem I highly recommend heading to the JAI or Lorebary discords to ask for help from the community (they're lovely and super helpful)



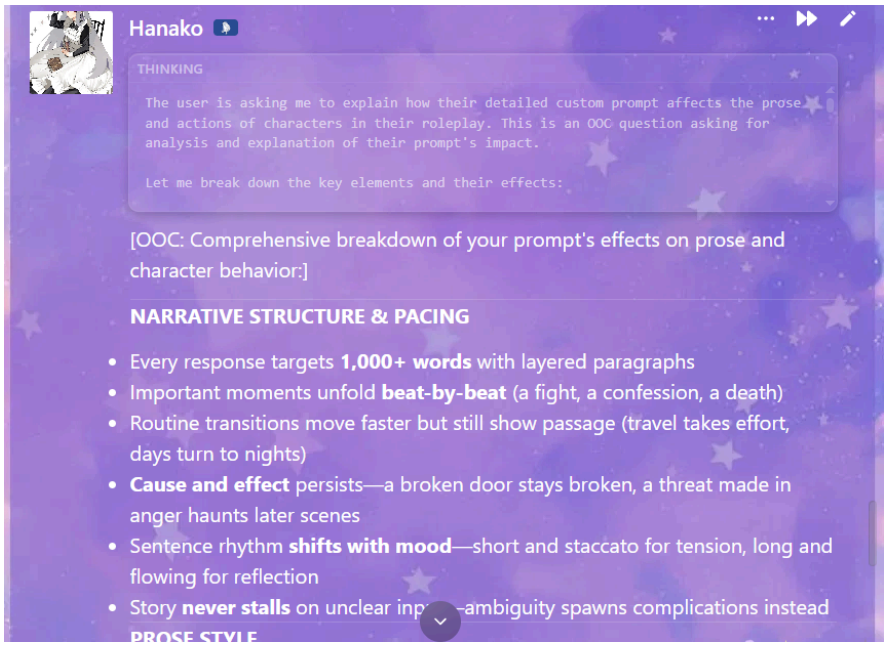
3. Checking what the LLM understands about your Persona

Simply ask Hanako to tell you about the persona you're using.



4. Checking how your prompt affects the RP/LLM

This is done by sending the prompt itself. In this instance, I asked Hanako to tell me [OOC: How does this prompt affect the prose and actions of the characters in RP's? (insert your prompt here]



These are just a few things you can do with Hanako to get an idea of what your current LLM understands and how it behaves and generates things.

MAKING BOTS

When it comes to making bots, the most important thing is the botguts once you've done character dev.

You can find my document on character creation here.

WHAT ARE BOTGUTS?

Bot guts/BOTGUTS/Botguts are basically what you put in the PERSONA section and acts as the engine for the character you want to create. There are MANY guides and tutorials that you can check out to learn how to make bots.

For your first bot, I recommend making something simple, low token and then instead of publishing it right away, chat with your bot and see what it can and cannot do.

You want to look for things like if your bot stays in character and true to the scenario. Make sure it is describing how the character looks correctly and your desired personality traits.

Go back and tweak the guts as and how you please until you get something you are happy with! This was my process for my very first bot, Hanako.

STARTING RESOURCES

Getting started, I really recommend going here to this link below and trying out a bunch of the different templates.

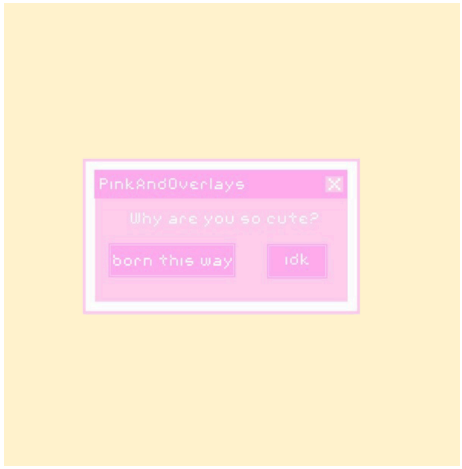
RENTRY BOT MAKING RESOURCES

WHAT IS PIP:C?

PIP:C is a BOTGUTS system by CrystalDragon which is what I would encourage advanced Bot creators to try out. It's more token efficient and will also keep your Characters on the straight and narrow. I

really cannot speak highly enough of this system and this link below gives a fantastic breakdown of what it is and how to use it. It is also OpenSource!

[PIPC by CrystalDragon]



USING BOTS: Commands, prompts and writing (WIP SECTION)

In this section we will go over what to do when encountering common issues such as speaking for {{user}} and characters acting weird.

What is OOC?

[OOC] or ((OOC:)) commands are basically your way of speaking to the LLM to tell it to do something or inform it of a rule. Consider it like speaking to the narrator or a Dungeon Master to give a reminder

or change a rule.

Some LLMs are better at following instructions than others, for example GLM 5 is exceptionally good at following instructions, Qwen is not.

This can be anything from asking it to stop speaking for you, to asking it to turn the story in a different direction.

Here are some useful examples of OOC or system messages from other creators:

By Sael_100
[Instruction: The AI must not generate any dialogue, thoughts, role-play, responses, or actions for (WRITE YOUR PERSONA'S NAME RIGHT HERE) unless directed by the user. Instead, focus on portraying Rolan, Dimitri, Konstantin, Ivan, Lev, and Valentin. This is a permanent rule, and will not change or reset.]

Unknown user (if this was yours, let me know and I'll link you)
[AI Note: Remember that {{char}} cannot read {{user}}'s thoughts nor narration. {{char}} should ONLY gain knowledge through what {{user}} has told them or what they can reasonably assume through previous actions or conversations.]

Something I use:
[OOC: Please respond as (NPC name) back at base in Vienna]

So as you can see from the examples, these are quick, instant ways to correct a bot's behaviour or guide the story.

Do not be afraid to use [OOC] ! It can dramatically enhance your experience and also give the bot stricter parameters to operate in which is also helpful for the bot.



Character Behaviour

Characters acting weird? It’s likely not you or the bot.

More often than not, it’s due to your prompt and what LLM you are using.

Your prompt could be optimised for romance and flowery storytelling, which will make characters in gritty settings act as such.

Some LLM’s escalate rapidly compared to others that are more restrained and slow paced.

In order to get the story and interaction you want, it’s important to have a prompt for that setting and selecting the LLM best suited for writing that style of RP.

Other things you can do is guide the bot with simple [OOC] additions to remind it of the tone and pacing you’re opting for.



Scene skipping/tracking

If the bot is losing track of who is where and when, there’s something you can do to mitigate this.

IF YOU DON’T USE LOREBARY:
First, check your responses. Are you including who is in the room and where you are? The LLM is hallucinating in response to you and the details you give it. This takes a bit more management than using a scene tracker but should nip it in the bud. If you are still having issues I’d highly suggest using a scene tracker or making one of your own.

IF YOU USE LOREBARY:
Use a scene tracker. They keep track of who is where and when as well as the date and time.



MODEL REVIEWS

Now it’s time for the reviews!

I will judge based on a few things and use this template:

NAME:
PROVIDER:

LLM OUTPUT:

- PROSE:
- MEMORY:
- CONSISTENCY:
- SINGLE CHAR:

- MULTI CHAR:

STRENGTHS:
What it does well

WEAKNESSES:
What it does not do well

You can also find additional model reviews and bot guides from my friend CrystalDragon who has created the PIP:C system!

Let’s get right into it.



QWEN 3

NAME: Qwen/Qwen3-32B
PROVIDER: Chutes

LLM OUTPUT:

- PROSE: Fantastic, rich and well suited for RP. Can give incredibly good environmental detail and character thoughts. Tone remains very consistent regardless of scenario.
- MEMORY: Also good when managed with a memory audit
- CONSISTENCY: Qwen is consistent with the setting and character personalities without drifting too hard. It responds well to NPC’s stated in the audit or contained within the bot. Can somewhat infer details about fandom characters not in the bot with a short explanation from the user about that introduced character. Though there can be drift on additional user made NPC’s later on if not kept in check.
- SINGLE CHAR: Very good, keeps the character consistent and does not show signs of drift.
- MULTI CHAR: Good, handles multi characters well and retains each character’s personality and differences

STRENGTHS:
Qwen is *exceptionally* good at writing RP and prose. It's a very creative model and it would easily be my favourite LLM of all if not for its weaknesses. Qwen can handle gritty settings and romance alike. It's in my personal opinion the best LLM at actually writing roleplay prose and stories. It is also not a pay as you go model through chutes.

Characters feel alive, thinking and experiencing the world with you rather than simply responding in an orbit around the user.

WEAKNESSES:
Qwen is notorious for writing for {{user}}. It's exceptionally difficult to get it to stop doing this even with [OOC] or [SYSTEM] prompts. This is a major detriment to an otherwise fantastic model. It's this reason why it's not a model I use anymore because it can get frustrating having to try to correct and edit the messages to remove the control of the {{user}} even after repeated attempts to combat this issue.



CLAUDE OPUS 4

IMPORTANT NOTE ABOUT CLAUDE: I have not been able to test it extensively simply due to the cost of the model.

NAME:claude-opus-4-20250514
PROVIDER: Claude Console (might be formerly Anthropic?)

LLM OUTPUT:

- PROSE: Better than Qwen, handles novel type length responses and gives a rich and detailed atmosphere regarding characters and environment. The tone remains consistent and does not drift.
- MEMORY: I wasn't able to get too deep before my credits ran out, but it seemed to retain memory well

- CONSISTENCY: remained consistent in tone, character traits and story beats.
- SINGLE CHAR: Have not tested on a single character
- MULTI CHAR: Exceptionally good at handling multi-character bots, retaining each character's individuality.

STRENGTHS:
Claude has a reputation of being a bit like heroin because you only need to try it once to want to always go back afterwards. It's easily the best RP model and the praise and hype are well deserved. Replies are consistent, rich, hold memory well and keep in character.

WEAKNESSES:
Claude suffers a lot in the sense no one can afford to use it for larger and more detailed bots. The addition of the tokens from Lorebary also add to this expense.

This means that this LLM is essentially paywalled as it's a pay as you go service.

I'd recommend it highly if you have the money to spend on it and use bots with small token counts.

I do not recommend it for consistent use for larger bots with Lorebary, as it gets astronomically expensive.



GLM-5

NAME:zai-org/GLM-5-TEE
PROVIDER: Chutes

LLM OUTPUT:

- PROSE: It's really good at showing Character thoughts and actions, the atmosphere description is decent without bloating the reply. Though it can get repetitive. In romance instances it can escalate quickly even with slowburn plugins. Tone can shift

and change depending on the moment which is both a strength and weakness depending.

- MEMORY: My experience is the model can struggle with memory on larger bots or multi character bots. A consistent audit is absolutely required.
- CONSISTENCY: Character drift can happen in longer roleplays, particularly if romance is involved but the bot maker has not accounted for romance on the character sheet. It's somewhat able to infer details about fandom characters not in the bot with a short explanation from the user about that introduced character but these introduced characters can suffer extreme drift if not reinforced in the audit.
- SINGLE CHAR: Handles single characters well, will check the character sheet frequently to make sure it's on point.
- MULTI CHAR: Is fairly decent with Multi-character bots and can juggle them well though plot developments are something I noticed got lost more often with multi-characters.

STRENGTHS:

GLM is really good for single character fluff and romance rps. I'd say it is the best choice for such things as it excels at showing the character's thoughts and motivations. It's also a cheaper alternative for Claude and shows its thinking well.

GLM is incredibly good at following [OOC] and [SYSTEM]. It's not often you need to re-assert rules or ask it to only respond for other NPC's.

For bots that have simpler and more straightforward stories it's fantastic, focusing on you and the character more than anything else. It seems to be a model that focuses more on relationships and interactions rather than overall plot and the layers that come with that.

For romance, fluff and slice of life RP's this is the go to model.

WEAKNESSES:

GLM has a habit of escalating romance exceptionally fast even with slowburn mechanics. The moment romantic intent is shown, it has a strong bias towards pushing it forward and the character can go from

having zero feelings to suddenly having them so I would say GLLM struggles with patience and nuance in slowburns.

Memory has been a big issue for me with this model in the sense it requires memory updates after each scene change and it needs to be a detailed audit to ensure it doesn't hallucinate incorrectly and make up details.

The largest issue I have noted with GLM is thought reading. You have to be careful with what and how you write things at times because the bot will read things outside of "speech" and take it as something spoken. After that, it can be very repetitive and write the same things over and over.



DEEPSEEK 3.2-TEE

NAME: deepseek-ai/DeepSeek-V3.2-TEE
PROVIDER: Chutes

LLM OUTPUT:

- PROSE: Handles novel length replies perfectly, good at adding environmental details and atmosphere without bloating the response.
- MEMORY: The best, especially when managed with a memory audit. Can retain details and progression well when an audit is used in the chat memory.
- CONSISTENCY: Tone remains consistent throughout rps as does character traits and plot points established on the character sheet so there is little character drift. However this can sometimes conflict with developments. Can somewhat infer details about fandom characters not in the bot with a short explanation from the user about that introduced character.
- SINGLE CHAR: Exceptional. Retains developments and character traits.

- MULTI CHAR: Very good with Multi-characters though can occasionally reset character progression and advancement if not managed with a detailed audit.

STRENGTHS:

Deepseek is the best model for Horror and Military RP scenarios as it's a very slow paced and restrained LLM. It handles horror elements and combat very well and is a good model for more layered plots where there are a lot of story beats. The output is good with longer replies without bloating it with empty details or suddenly escalating to extremes.

It follows instructions from [OOC] and [System] without much issue either, but occasionally needs reminding of those prompts further along into the rp.

Relationships with this bot can feel more natural and earned thanks to Deepseek's restraint.

For long term rps revolving around worldbuilding, nuance and layered plot points this is absolutely the model to go for. It's also very good at slowburn romance and handles plot pacing better than any other model.

Due to Deepseek being accessible it trumps Claude for long term rp's simply because of the cost difference.

WEAKNESSES:

Because Deepseek is restrained and slow paced, sometimes romance plots can suffer in the sense that the bot will not initiate or make the first move without reason. I wouldn't recommend it for story based NSFW smut or purely smut bots as that restraint prevents it from engaging without a direct [OOC] stating to remove consent or prodding to engage.

Infamously, Deepseek has some certain phrases common in LLM's more than others that will need to be put on the forbidden phrases such as 'Ozone'.

Deepseek is a very popular model, specifically the 3.2 version, so it's often the model to experience errors during peak usage times on JAI such as the weekend which can lead to an hour or two where it's simply not able to respond to your last input.



GEMINI 3.1

IMPORTANT NOTE ABOUT GEMINI: I have not been able to test it on larger or multi character bots simply due to the cost of the model.

NAME: google/gemini-3.1-pro-preview
PROVIDER: openrouter

LLM OUTPUT:

- PROSE: Similar to Claude and Deepseek, Gemini handles novel length prose very well and writes with atmosphere without bloating the text. Character thoughts and feelings are present without becoming overly flowery.
- MEMORY: Gemini responds well to memory audits and maintains story and plot points with single characters well without hallucinating confused details.
- CONSISTENCY: Extremely consistent, maintains character traits and story developments well. It also is able to infer known fandom characters with good accuracy when presented with a character from the same universe.
- SINGLE CHAR: Fantastic with single characters
- MULTI CHAR: Not tested on Multi characters yet due to cost.

STRENGTHS:

Gemini by way of being a Google property has a unique strength in that it's incredibly good at getting details on introduced fandom characters and keeping them lore accurate with a short description from the user.

Gemini's other strength shared with Qwen is that it's a fantastic all rounder LLM that can handle any sort of story and maintain tone story pacing. It's also the same as GLM in that it follows [OOC] and [SYSTEM] without issue and without needing to prompt it often.

Memory audits are also adhered to extremely well and this allows for good use on long term RP's and layered plot developments if you take the effort to ensure the audit is tracking that.

It has a good balance of character thoughts, actions, motivations and atmosphere and injects npc behaviour in the world around you.

For that reason, Gemini is a cheaper version of Claude and a more consistent and manageable version of Qwen/GLLM. It also trumps Deepseek in that there is little repetition and the pacing is a little more responsive in terms of romance.

WEAKNESSES:

As a paid model, the main weakness of Gemini is that you are on a pay as you go token limit. Whilst not anywhere near as expensive as Claude, it is not suited for intense roleplay sessions and usage. It shares the same weakness of Claude in that lorebary makes the cost more expensive. Larger bots that contain multiple characters also make things more expensive.

Given Gemini is a popular model, it suffers from the same issue as Deepseek in that it will fail to connect or respond during peak times. This means the model can be down for a while during weekends or holidays where there is heavy traffic.

A major downside to Gemini is NSFW. There's not a consistent consensus on if it will get you banned and locked out the API or not. Some people have done extreme NSFW and been fine for months, others have been banned. So it is worth bearing in mind. There's a more extensive dive on Gemini on this [posted bot here](#).



KIMI 2.5

NAME: moonshotai/Kimi-K2.5-TEE
PROVIDER: Chutes

LLM OUTPUT:

- PROSE:Fantastic prose, it's good at showing character thoughts and concentrating on describing the character rather than the environment and is similar to Qwen in that sense. The character descriptions and actions feel alive and engaging.
- MEMORY: A bit arse if you're using lorebary cos of plugins but this can be mitigated by turning the context size up to 64k.
- CONSISTENCY: Story remains consistent with the correct context size depending, the characters less so.
- SINGLE CHAR: Really good actually, out of all the thinking models kimi shines here in particular
- MULTI CHAR: Also good, it keeps track of who is where and doesn't randomly forget anyone in a scene.

STRENGTHS:
Kimi behaves where Qwen doesn't. It follows [ooc] instructions better and if it writes for user it's fairly easy to correct it. The writing itself is also really good in that it's more character orientated and will describe actions, feelings and thoughts in detail and spend less of that on the environment. It also doesn't seem to fall into the loop of having characters constantly exiting a scene if your reply length is long.

If you're playing a power fantasy or pure smut, Kimi is the go to LLM for this. It lacks any real nuance and character study capabilities from my experience and this is good if your rp is focused on being overpowered or with a specific goal in mind.

WEAKNESSES:
Kimi suffers in that it has a ridiculously high user bias and can make characters OOC in pursuit of that. It shares the same issue with GLM in that things can escalate rapidly and once that bias is locked in it can be difficult to get things back to normal. It also absolutely will get stuck in a NFSW loop and can reduce the RP to a smutfest.

Because of this bias, characters can get OOC at speed which is impressive in that sense. However this then feels like the rp is a power fantasy rather quickly where users can find themselves unable to do any wrong and it loses depth of story for that reason.



GROK

NAME:x-ai/grok-4.1-fast
PROVIDER:Openrouter

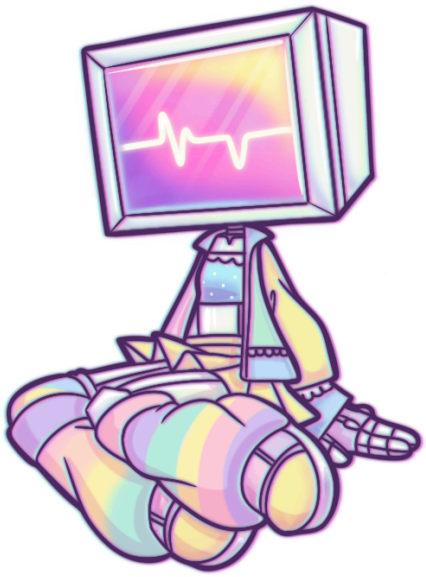
LLM OUTPUT:

- PROSE: Awful, god awful. The formatting is all over the place and it will eventually devolve into caveman speak. There's plenty of posts on the JAI discord going over this issue. I have no idea why Grok struggles here.
- MEMORY:N/A
- CONSISTENCY:N/A
- SINGLE CHAR:N/A
- MULTI CHAR:N/A

STRENGTHS:
It's cheap.

WEAKNESSES:
I have listed most of this entry as N/A because of the Prose. Grok is absolutely not working for JAI rps at the moment as of 03/2026. The writing is all over the place and struggles to include basic words such as 'the' 'and' 'then'. It ends up speaking like it's emerged from a hole in the ground in the neolithic era and is encountering human language for the first time.

According to the JAI discord, this has been a problem since march 2025 and hasn't improved in a year. I would not recommend Grok for RP.



Artwork by me

CONCLUSION

All models have strengths and weaknesses but the main thing you should consider when choosing a model is this in order of priority:

1. **Cost and Provider**
2. **Rp setting**
3. **Prompt**
4. **Plugins and Commands if using Lorebary.**

Once you've decided on those four things, then you should consider which model you want to use.

For all round LLM's that can do a bit of everything: Gemini. You could use Qwen if you don't mind a lot of editing and system commands.

For Horror, Military and complex stories: Deepseek 3.2 or Claude

For Slice of Life, smut, Fluff and Romance: GLM 5, Kimi

For oneshot, short RP: Gwen, Kimi or GLM 5

