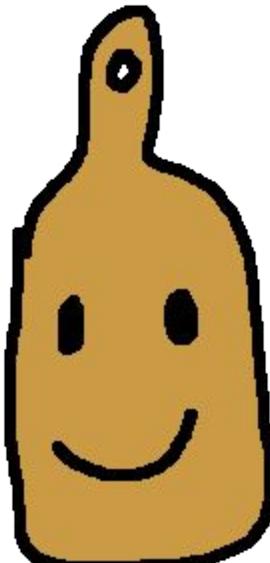


FROM INFRASTRUCTURE
TO PRODUCTION



PADDLER



Two Paths AHEAD of OSS LLMs (Both AWESOME)

1. ...
PURE HOBBY

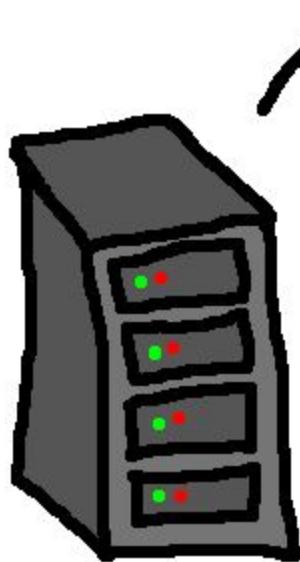
 ... 2.
EXPERIMENTS

WHERE DOES

PODDER

FIT IN?

Where can you
deploy paddle?



EC2 Image Builder

[Image pipelines](#)

Images

▼ Saved resources

Components

Image recipes

Container recipes

Infrastructure configurations

Distribution settings Updated

Image workflows

Lifecycle policies

▼ Security overview

Security findings

Security scanning settings

▼ AWS Marketplace

Subscriptions

Discover products Updated

Documentation ↗

paddler-image AMI

Summary

Description

-

IAM role

-

Enhanced metadata collection

Enabled

Security scanning status

-

YOU CAN PUT
PADDLER
IN THE CLOUD

[Output images](#)[Image recipe](#)[Image workflows](#)[Infrastructure configuration](#)[Distribution settings](#)[Security findings](#)[EventBridge rules](#)[Tags](#)

Output images (9)

Output images produced by the pipeline

[Retry Image](#)[Cancel](#) Search output image

<input type="checkbox"/>	Version ↴	Type	Creation time	▼	Image status	ARN	Log
<input type="checkbox"/>	1.0.0/8	AMI	January 07, 2026, 15:29 (UTC+01:00)	✓ Available	<input type="checkbox"/>		
<input type="checkbox"/>	1.0.0/7	AMI	January 01, 2026, 14:09 (UTC+01:00)	✓ Available	<input type="checkbox"/>		
<input type="checkbox"/>	1.0.0/6	AMI	December 31, 2025, 20:17 (UTC+01:00)	✓ Available	<input type="checkbox"/>		
<input type="checkbox"/>	1.0.0/5	AMI	December 23, 2025, 15:25 (UTC+01:00)	✓ Available	<input type="checkbox"/>		
<input type="checkbox"/>	1.0.0/5	AMI	December 22, 2025, 11:57 (UTC+01:00)	✗ Failed	<input type="checkbox"/>		

LLAMA-SERVER REWRITE



llama

1.  SEMANTIC
VERSIONING
2. WE CAN ADD MORE
CUSTOM FEATURES
3. DECOUPLING

GET /slots : Returns the current slots processing state

Warning

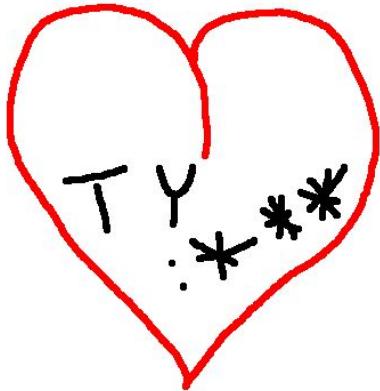
This endpoint is intended for debugging and may be modified in future versions. For security reasons, we strongly advise against enabling it in production environments.

<https://github.com/ggml-org/llama.cpp/tree/ddef99522d1ba74193b7394e803fab8db5c78bae/tools/server>

So is there a quick way to achieve my goal? Before to get hand dirty to write a simple Rust HTTP REST API server wrapping llama.cpp inference logic, just like I did in [dots.ocr.runner](#), I searched “rust llama-server” again in google. Then a new project has come into my view:

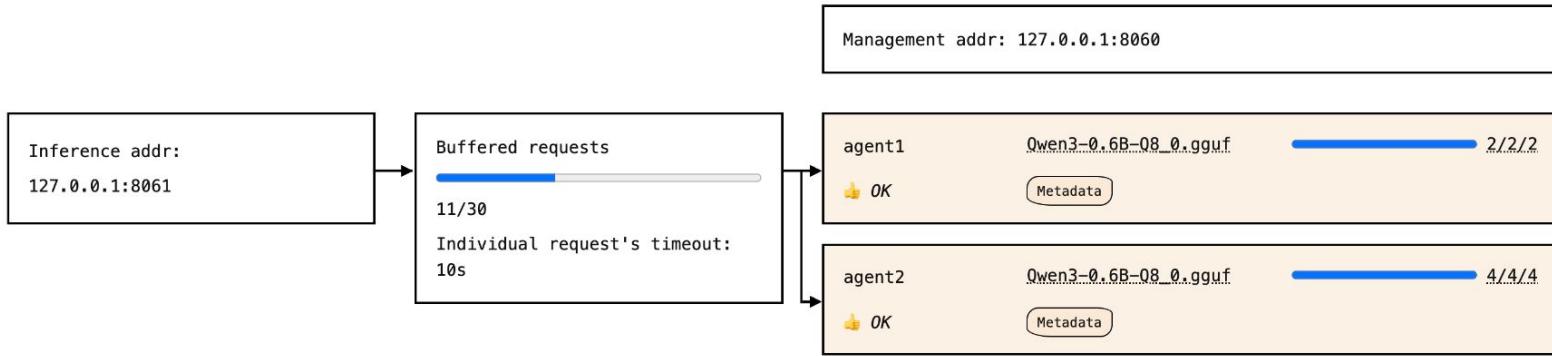
“I just rewrote llama.cpp server in Rust (most of it at least), and made it scalable”

Cool! that's what I want!



Today I tried to use llama-server to serve as immersivetranslate extension backend with the new HY-MT1.5-1.8B-GGUF model. But it seems there is memory leak in llama-server. So I searched "rust llama-server" and found paddler. Great job for paddler! Looking forward new features like multi-modal support!





(?) [How to add an agent?](#)

Model URI

```
https://huggingface.co/Qwen/Qwen3-32B-GGUF/blob/main/Qwen3-32B-Q
```

Chat Template

How should Paddler obtain the chat template?

Use the chat template provided by the model (recommended)

- Most models support this, but not all. If the model does not have a chat template, it will fail to load.

Use my chat template

- Only use this option if you know the model does not come with a chat template, or you want to modify it to suit your needs.

[Edit chat template](#)

Inference Parameters

What are these parameters?

batch_n_tokens	512
context_size	4096
min_p	0,05
penalty_frequency	0
penalty_last_n	-1
penalty_presence	1,5
penalty_repeat	1
temperature	0,6
top_k	40
top_p	0,8
enable_embeddings	<input type="checkbox"/>
pooling_type	- Last ← enable embeddings to turn this on

[Apply changes](#)

Paddler is based on `llama.cpp`, and it supports models in the `GGUF` format.

Supported sources:

Hugging Face 😊

Each agent will individually download the model, and cache it locally.

For example, you can use the following URL to download the Qwen-3 0.6B model:

```
https://huggingface.co/Qwen/Qwen3-0.6B-GGUF/blob/main/Qwen3-0.6B-Q8_0.gguf
```

Local File

File path is relative to the agent's working directory.

If you want all the agents to use the same model, you need to ensure that the file is present in the same path on all agents.

```
agent:///path/to/your/model.gguf
```

huggingface.co/Qwen/Qwen3-0.6B-GGUF

Best Practices

To achieve optimal performance, we recommend the following settings:

1. Sampling Parameters:

- For thinking mode (`enable_thinking=True`), use `Temperature=0.6, TopP=0.95, TopK=20, MinP=0, and PresencePenalty=1.5`. **DO NOT use greedy decoding**, as it can lead to performance degradation and endless repetitions.
- For non-thinking mode (`enable_thinking=False`), we suggest using `Temperature=0.7, TopP=0.8, TopK=20, MinP=0, and PresencePenalty=1.5`.
- We recommend setting presence_penalty to 1.5 for quantized models to suppress repetitive outputs.** You can adjust the `presence_penalty` parameter between 0 and 2. A higher value may occasionally lead to language mixing and a slight reduction in model performance.

← [Back to all parameters](#)

```

1 v  {%- if tools %}
2 v    {{- '<|im_start|>system\n' }}
3 v    {%- if messages[0].role == 'system' %}
4 v      {{- messages[0].content + '\n\n' }}
5 v    {%- endif %}
6 v    {{- "# Tools\n\nYou may call one or more functions to assist with the user query.\n\nYou are provided\n" }}
7 v    {%- for tool in tools %}
8 v      {{- "\n" }}
9 v      {{- tool | toJSON }} 
10 v    {%- endfor %}
11 v    {{- "\n</tools>\n\nFor each function call, return a json object with function name and arguments within\n" }}
12 v    {%- else %}
13 v      {%- if messages[0].role == 'system' %}
14 v        {{- '<|im_start|>system\n' + messages[0].content + '<|im_end|>\n' }}
15 v      {%- endif %}
16 v    {%- endif %}
17 v    {%- set ns = namespace(multi_step_tool=true, last_query_index=messages|length - 1) %}
18 v    {%- for index in range(ns.last_query_index, -1, -1) %}
19 v      {%- set message = messages[index] %}
20 v      {%- if ns.multi_step_tool and message.role == "user" and not('<tool_response>' in message.content and
21 v          {%- set ns.multi_step_tool = false %}
22 v          {%- set ns.last_query_index = index %}}
23 v      {%- endif %}
24 v    {%- endfor %}
25 v    {%- for message in messages %}
26 v      {%- if (message.role == "user") or (message.role == "system" and not loop.first) %}
27 v        {{- '<|im_start|>' + message.role + '\n' + message.content + '<|im_end|>' + '\n' }}
28 v      {%- elif message.role == "assistant" %}
29 v        {%- set content = message.content %}
30 v        {%- set reasoning_content = '' %}
31 v        {%- if message.reasoning_content is defined and message.reasoning_content is not none %}
32 v          {%- set reasoning_content = message.reasoning_content %}
```

Model URI

<https://huggingface.co/Qwen/Qwen3-0.6B-GGUF/blob/main/Qwen3-0.6E>

Chat Template

How should Paddler obtain the chat template?

Use the chat template provided by the model (recommended)

- Most models support this, but not all. If the model does not have a chat template, it will fail to load.

Use my chat template

Only use this option if you know the model does not come with a chat template, or you want to modify it to suit your needs.

[Edit chat template](#)

You:

How can I make my cat happy?

AI:

Okay, the user is asking how to make their cat happy. Let me start by recalling what cats generally enjoy. They need good food, clean water, and a safe environment. But the user probably wants more specific tips beyond the basics. Also, consider the user's possible situation. They might be a new cat owner or someone looking for more ways to bond. Tailoring the advice to be practical and actionable. Maybe add a tip about respecting their cat's personality, as some are more independent. Ending with a positive note to reinforce the bond between them and their cat.

To make your cat purr-fectly happy 🐱, try these tips:

1. **Playtime!** 🎪 Use feather toys, laser pointers, or puzzle feeders to spark their inner hunter.
2. **Cozy nooks** 🏠 Provide a soft bed, sunny windowsill, or a cat tree for climbing and napping.
3. **Groom & bond** 💆 Brush their fur regularly (and enjoy the cuddles!) and keep their litter box clean.
4. **Treats & meals** 🍴 Offer high-quality food and occasional cat-safe snacks (avoid chocolate/onions!).
5. **Enrichment** 🌱 Rotate toys, add catnip, or create a "catio" for safe outdoor adventures.
6. **Quality time** 😊 Talk to them, sing, or just sit nearby—they love your presence!

Every cat is unique, so observe their quirks and respect their moods. ❤️
You're already doing great! 🐱✨

Type your prompt here...





Keep AI on your own servers

What is it?

Paddler is an open-source LLM load balancer and serving platform that lets teams run inference and deploy LLMs on their own infrastructure.

Who made this?

This is an open-source project originally started by [Intentee](#).

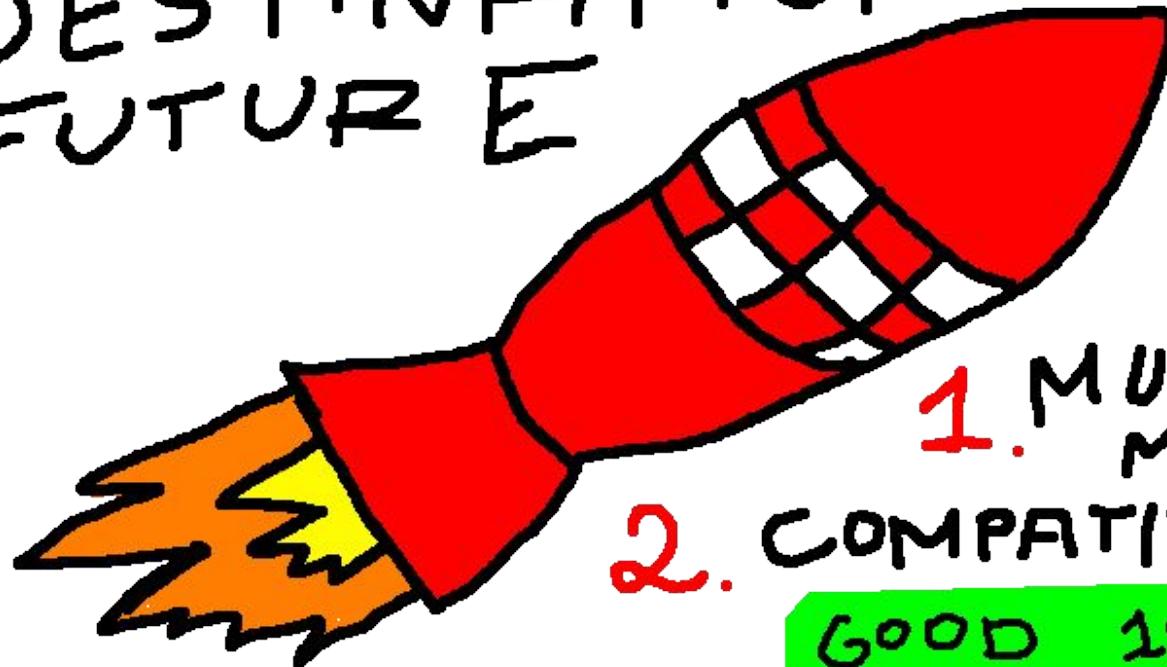
Why?

- Cost control
- Data privacy
- Reliability
- Extensive model support (uses llama.cpp for inference)

[Get Started](#)

[Install](#)

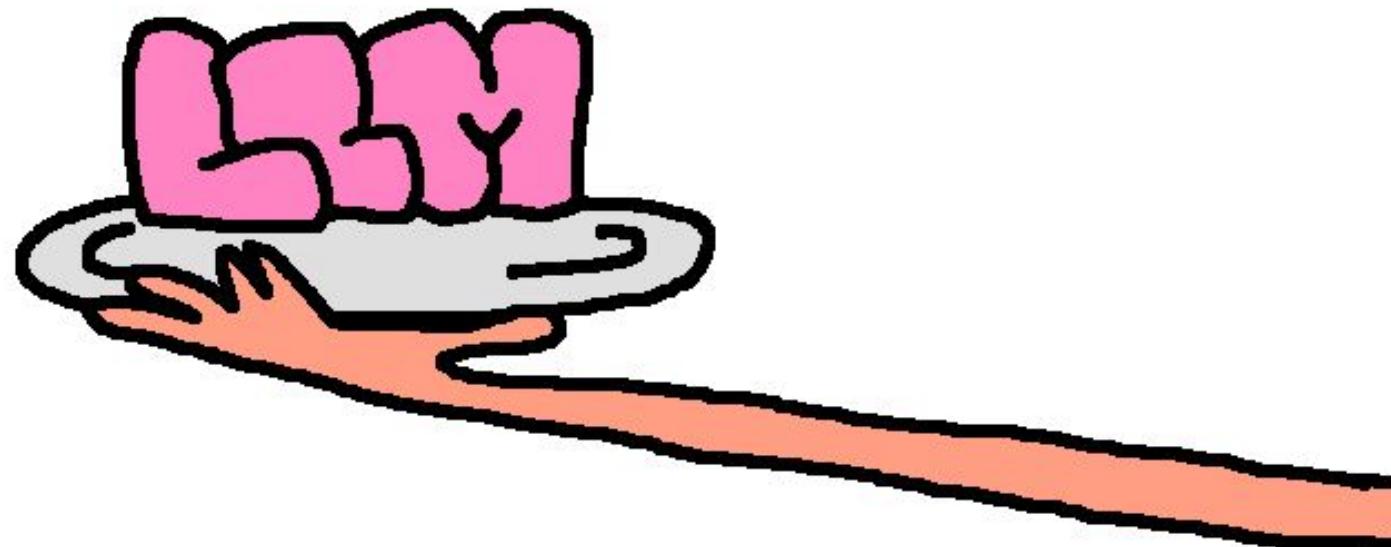
DESTINATION FUTURE



1. MULTIMODAL
2. COMPATIBILITY
3. BROWSER AGENTS

GOOD 1st ISSUE

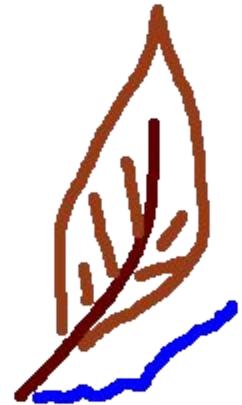
WHAT SELF-HOSTING UNIQUELY ENABLES?



WHAT ARE WE BUILDING NEXT

1. expanding the
ecosystem

2. first step: POET



What is Unique

about

~~being a~~

P O E T

?

Exhibit 1

WHY SELF-HOSTED
MODELS
IN
POET



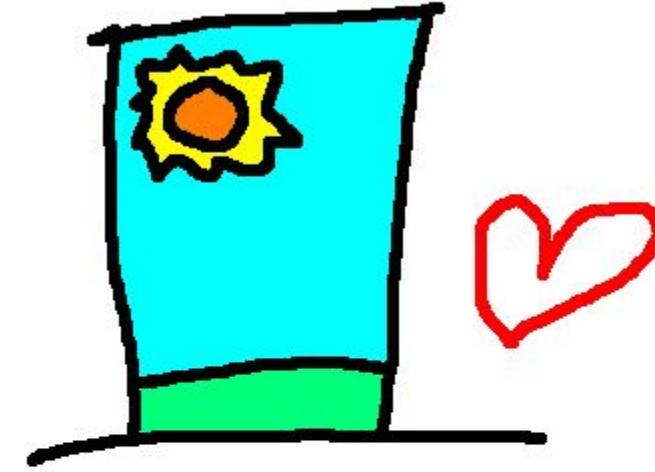
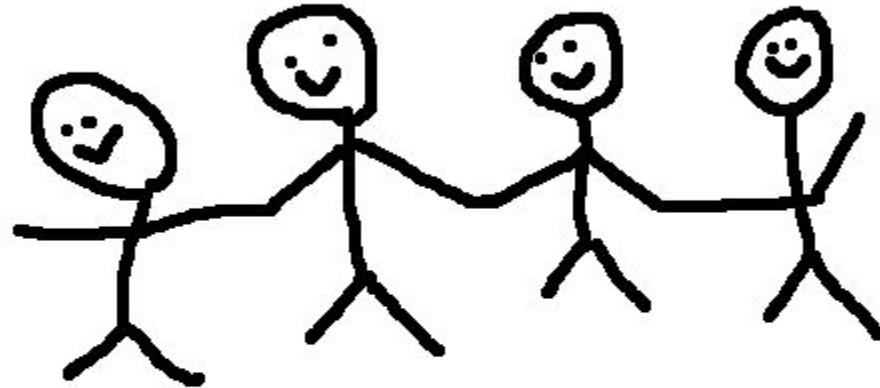


EXHIBIT 2

FIRE SPARQL



Mama



When you leave
this room . . .



github.com/intentee/paddler

WE
GIVE
AWAY
PINS

