

PLGrid Forge

AI services for science
ACK Cyfronet AGH

Łukasz Flis, Szymon Mazurek



Inference in HPC environment

expert way





Inference - how to start: HPC way

STEP 1. environment preparation

- apply for HPC resources allocation
- learn how to use batch system
- install inference environment:
 - run interactive job
 - load modules environment
 - create venv
 - install vllm/sglang and dependencies
 - download desired model from huggingface or other repository /use your own



Inference - how to start: HPC way cont.

STEP 2. execution

- determine resources required for chosen model size
 - number of gpus
 - number of nodes
- prepare job script
 - vllm/sglang:
 - (optional) vllm: boot ray cluster for multi node setup / use torchrun for sglang
 - find optimal set of command line parameters and tuning settings
 - establish optimal context size for your workload
- test
 - verify performance (latency / throughput)
- submit inference job
- setup tunneling for external api access



Inference - how to start: HPC way cont.

STEP 3. additional steps & scaling

- use multiple jobs / servers to achieve desired throughput
- deploy load balancer for load distribution
- employ mechanisms for health checking and excluding malfunctioning instances (loops, crashes)



Expert way Pros and Cons

PROS:

- Flexibility of model choice: private and non-standard models, experimentation possibilities
- Private servers on dedicated resources
- Allow for non standard feature usage: structured output

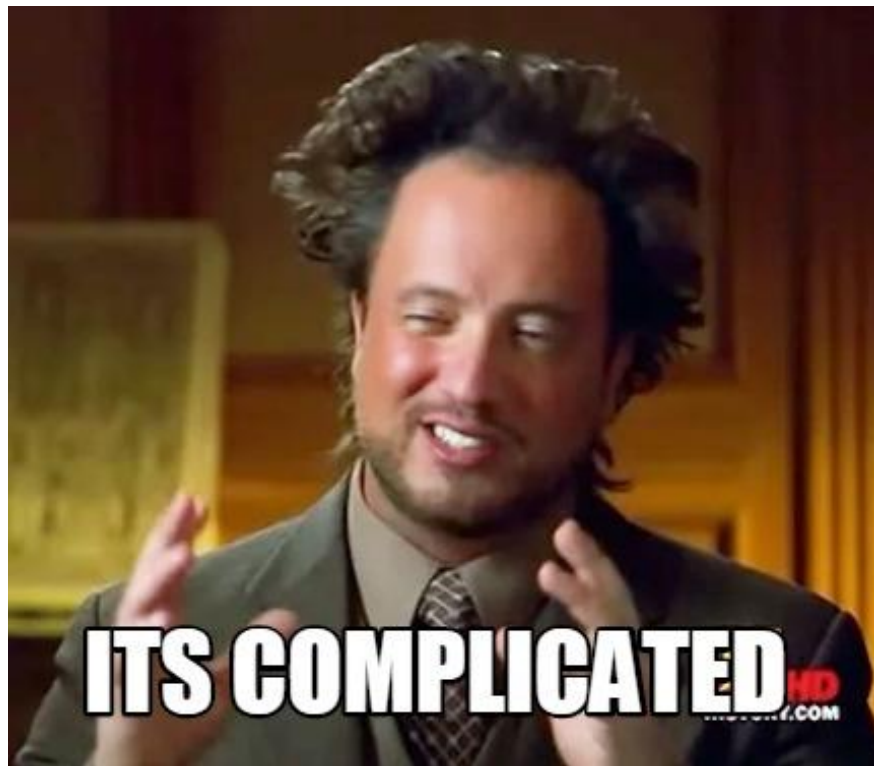
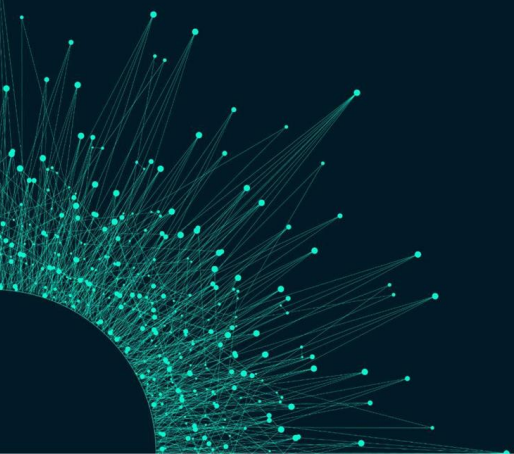
CONS

- Cluster environment experience required for setup and operation
- External access requires tunneling
- More throughput requires load balancing and more sophisticated setup



Expert way

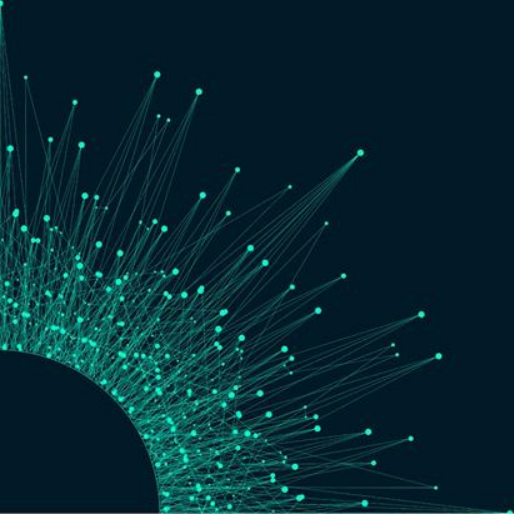
...





Inference in HPC environment

simple way





PLGrid Forge - family of AI services

Goals

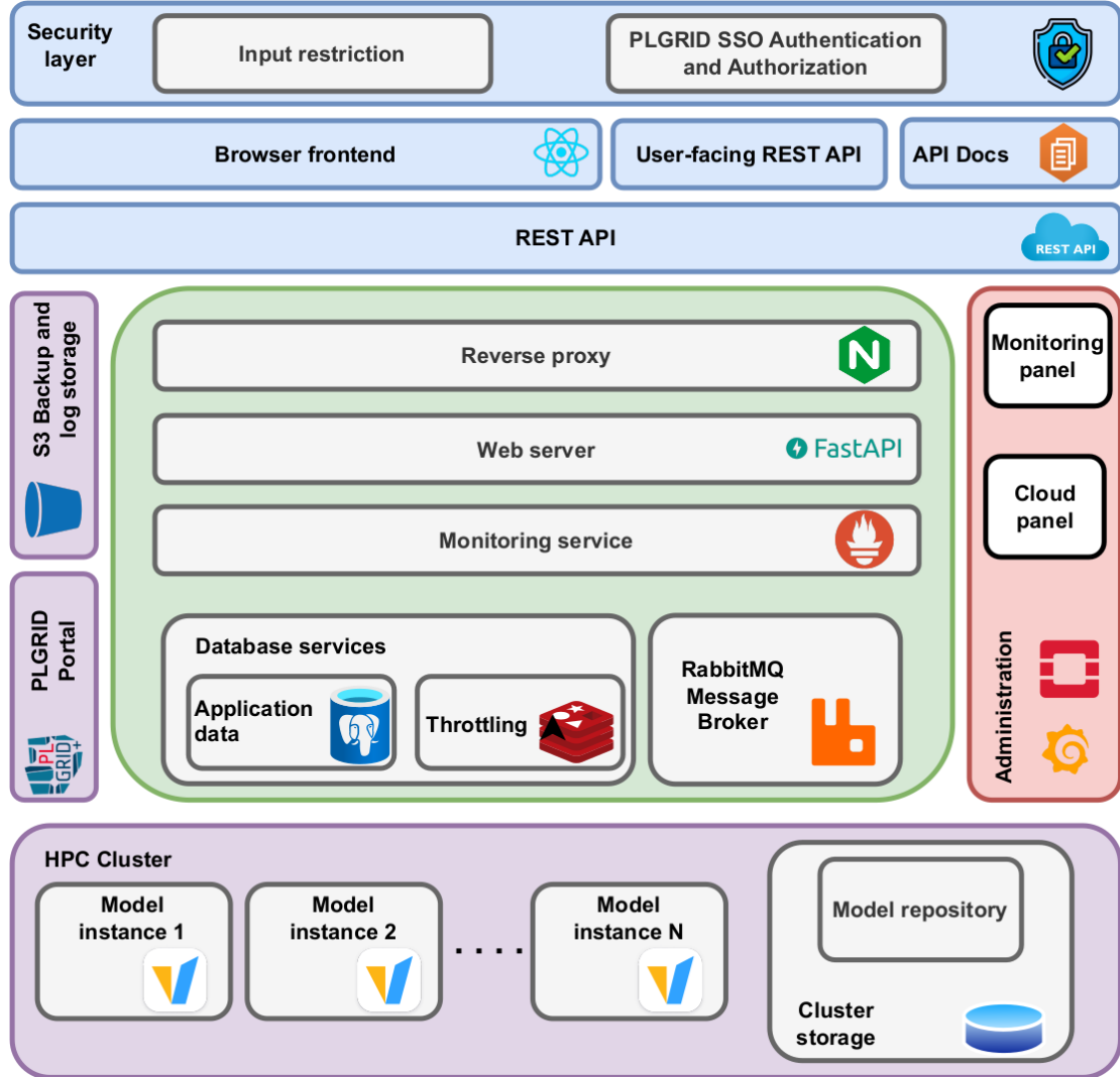
- data privacy oriented service
- provide access to popular LLM models for academic users
- service accessible for less technical and experienced users
- easy integration with desktop applications, notebooks and external services through API access
- platform for LLM experimentation by supporting vLM and speech-to-text service
- high availability, fault tolerance



PLGrid Forge - family of AI services

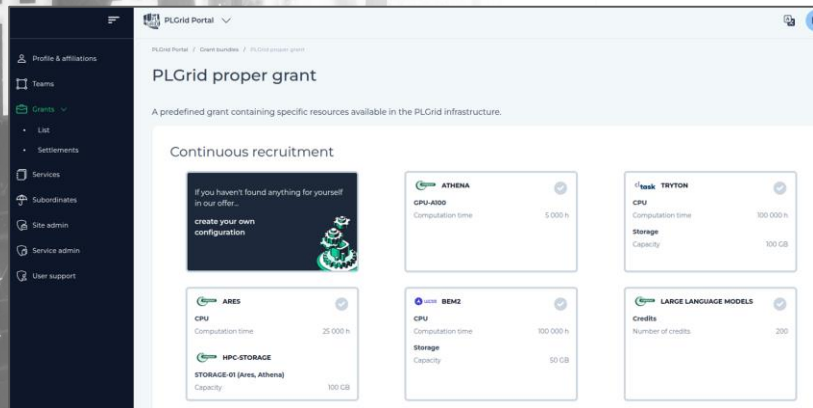
PLGrid Forge: the LLM component

- OpenAI API compatibility
- Data privacy
- Custom model support
- Per user/group: model access restrictions
- Accounting within PLGRID system based on credits
- Public API endpoint
- High availability mechanisms



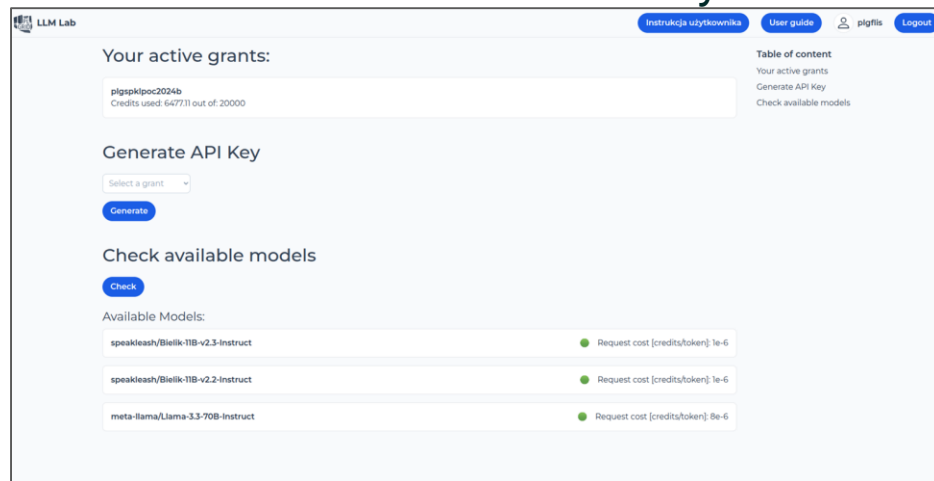
PLGrid Forge - how to apply

- Obtain PLGrid account and confirm your affiliation
- Request grant for Large Language Models resource
- Navigate to lmlab.plgrid.pl
- Select desired grant and generate API token
- DONE: Access to the inference service is now ready



The screenshot shows the PLGrid Portal interface. The main heading is "PLGrid proper grant". Below it, there is a section for "Continuous recruitment" with a message: "If you haven't found anything for yourself in our offer... create your own configuration". There are six grant cards displayed in a 2x3 grid:

- ATHENA**: CPU A100, Computation time 5,000 h.
- TRYTON**: CPU, Computation time 300,000 h, Storage Capacity 100 GB.
- ARES**: CPU, Computation time 25,000 h.
- BENZ**: CPU, Computation time 300,000 h, Storage Capacity 50 GB.
- HPCC-STORAGE**: STORAGE-01 (Ares, Athena), Capacity 100 GB.
- LARGE LANGUAGE MODELS**: Credits, Number of credits 200.



The screenshot shows the LLM Lab interface. The main heading is "Your active grants:". Below it, there is a section for "Generate API Key" with a dropdown menu for "Select a grant" and a "Generate" button. There is also a section for "Check available models" with a "Check" button. The "Available Models:" section lists three models:

- speakeash/Bleilc-11B-v2.3-Instruct (Request cost [credits/token]: 1e-6)
- speakeash/Bleilc-11B-v2.2-Instruct (Request cost [credits/token]: 1e-6)
- meta-llama/Llama-3.3-70B-Instruct (Request cost [credits/token]: 8e-6)



LLM Forge API use cases and integrations

Usage examples: bash + curl

```
#!/bin/bash
```

```
prompt="$1"  
ENDPOINT=https://llmlab.plgrid.pl/api  
source plgrid.cred  
curl $ENDPOINT/v1/chat/completions \  
-H "Content-Type: application/json" \  
-H "Authorization: Bearer $TOKEN" \  
-d "{  
  \"model\": \"speakleash/Bielik-11B-v2.3-Instruct\",  
  \"messages\": [  
    {\"role\": \"user\", \"content\": \"${prompt}\" } ]  
}" | tee resp | jq
```

```
> ./test-plg-bielik.sh "Czy warto brac kreatyne po 40 roku zycia?"  
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current  
           Dload  Upload   Total   Spent    Left  Speed  
100 3841 100 3693 100 148 361 14 0:00:10 0:00:10 ---:--- 924  
{  
  "id": "chatcmpl-6b508ace9dee424dbb8ac8459870c367",  
  "choices": [  
    {  
      "finish_reason": "stop",  
      "index": 0,  
      "logprobs": null,  
      "message": {  
        "content": "Kreatyna jest popularnym suplementem diety u\u017cywanym zar\u00f3wno przez amator\u00f3w, jak i zawodowych sp  
ortowc\u00f3w. Niezale\u017anie od wieku, niekt\u00f3re potencjalne korzy\u015bci z suplementacji kreatyn\u0105 to:\n\n1. Poprawa wydajno\u015bci  
fizycznej i si\u0142y:\n- Kreatyna pomaga w produkcji energii w kom\u00f3rkach mi\u0119\u015bninowych, co mo\u017ce prowadzi\u0107 do zwi\u0119ksze  
nia si\u0142y mi\u0119\u015bninowej i wydolno\u015bci podczas kr\u00f3tkotrwa\u0142ych, intensywnych wysi\u0142k\u00f3w, zwi\u015bkszcza treningu si\u0142owego i sprin  
tu.\n\n2. Szybsza regeneracja:\n- Kreatyna mo\u017ce sk\u00f3r\u00f3ci\u0107 czas regeneracji mi\u0119dzy intensywnymi treningami, co jest  
szczeg\u00f3lnie korzystne dla os\u00f3b \u0107wiczy\u0107cych kilka razy w tygodniu.\n\n3. Zwi\u0119kszenie masy mi\u0119\u015bninowej:\n- Suplemen  
tacja kreatyn\u0105 mo\u017ce pomóc w zwi\u0119kszeniu masy mi\u0119\u015bninowej poprzez zwi\u0119kszenie obj\u0119to\u015bci kom\u00f3rek mi\u0119\u015bninowych.\n\n4. Re  
dukcja ryzyka urazy masy mi\u0119\u015bninowej z wiekiem:\n- Z wiekiem naturalnie dochodzi do utraty masy mi\u0119\u015bninowej (sark  
openia). Kreatyna mo\u017ce pomóc w zachowaniu masy mi\u0119\u015bninowej i si\u0142y.\n\nPomimo potencjalnych korzy\u015bci, wa\u017cne jest, j  
by pamiętać o nast\u0119puj\u0105cych kwestiach:\n\n1. Bezpiecze\u0144stwo: Na og\u00f3l kreatyna jest uwa\u017cana za bezpieczny suplement, j  
ednak d\u0142ugoterminowe skutki s\u0105 wci\u0105\u017c badane. Zawsze warto konsultowa\u0107 si\u0119 z lekarzem, zwi\u015bkszcza w przypadku os\u00f3b z  
chorobami nerek, w\u0105troby lub serca.\n\n2. Indywidualne czynniki: Cho\u0107by\u015b kreatyna dzia\u0142a na wi\u0119kszo\u015b\u0107 os\u00f3b, niekt\u00f3r  
e z nich mog\u0105 nie osi\u0105gn\u0105\u0107 oczekiwanych rezultat\u00f3w ze wzgl\u0119du na indywidualne czynniki genetyczne lub metaboliczne.  
n\n3. Dob\u00f3r odpowiedniej dawki: Aby osi\u0105gn\u0105\u0107 skuteczno\u015b\u0107, dawka kreatyny powinna by\u0107 odpowiednio dostosowana. Zaz  
wyczaj zalecana dawka to 3-5 gram\u00f3w na dzie\u0144, w zale\u017cno\u015bci od masy cia\u0142a i cel\u00f3w treningowych. Mo\u017ce by\u0107 przyjmowana  
przez ca\u0142y dzie\u0144 lub podczas posi\u0142k\u00f3w.\n\n4. Hydratacja: Kreatyna wymaga du\u017cej ilo\u015bci wody w organizmie, wi\u0119c pamie  
taj o odpowiednim nawodnieniu.\n\n5. Interakcje z lekami: Je\u017celi przyjmujesz jakiegokolwiek leki, koniecznie poinform  
uj o tym lekarza, poniewa\u017c kreatyna mo\u017ce wchodzi\u0107 z nimi w interakcje.\n\n6. Inne czynniki: Pamiętaj, \u017e korzy\u015bci z  
suplementacji kreatyn\u0105 b\u0119d\u0105 widoczne, je\u017celi \u0142\u0105czy\u015b j\u0105 z odpowiednim treningiem i diet\u0105.\n\nPodsumowuj\u0105c, kreatyn  
a mo\u017ce by\u0107 korzystna dla os\u00f3b po 40. roku zycia, kt\u00f3re regularnie trenuj\u0105 i chc\u0105 zwi\u0119kszy\u0107 wydajno\u015b\u0107 fizyczn\u0105 oraz  
mas\u0119 mi\u0119\u015bninow\u0105. Zalecamy jednak konsultacj\u0119 z lekarzem przed rozpocz\u0119ciem suplementacji, aby upewni\u0107 si\u0119, \u017e jest t  
o bezpieczne i odpowiednie dla danej osoby.",  
        "role": "assistant",  
        "tool_calls": [],  
        "reasoning_content": null  
      },  
      "stop_reason": null  
    }  
  ],  
  "created": 1743724624,  
  "model": "speakleash/Bielik-11B-v2.3-Instruct",  
  "object": "chat.completion",  
  "usage": {  
    "completion_tokens": 998,  
    "prompt_tokens": 30,  
    "total_tokens": 1028,  
    "prompt_tokens_details": null,  
    "used_plgrid_credits": 0.001028  
  },  
  "prompt_logprobs": null  
}
```


Usage examples: python

```
from openai import AsyncOpenAI

key = "my-API-key-generated-in-LLM-Forge"
client = AsyncOpenAI(timeout=10, base_url="https://llmlab.plgrid.pl/api/v1", api_key=key)
res = await client.chat.completions.create(
    model="speakleash/Bielik-11B-v2.3-Instruct",
    messages=[
        {"role": "user", "content": "What is the capital of Poland?"},
    ],
    max_tokens=100,
    temperature=0.1,
    top_p=1,
    stream=True,
)
msg = ""
async for chunk in res:
    print(f"Chunk: {chunk}")
```



Integration with self-hosted chat applications - AnythingLLM example

LLM Preference

These are the credentials and settings for your preferred LLM chat & embedding provider. Its important these keys are current and correct or else AnythingLLM will not function properly.

LLM Provider



Generic OpenAI

Connect to any OpenAI-compatible service via a custom configuration



Base URL

https://llmlab.plgrid.pl/api/v1

API Key

.....

Chat Model Name

meta-llama/Llama-3.3-70B-Inst

Token context window

32000

Max Tokens

24000



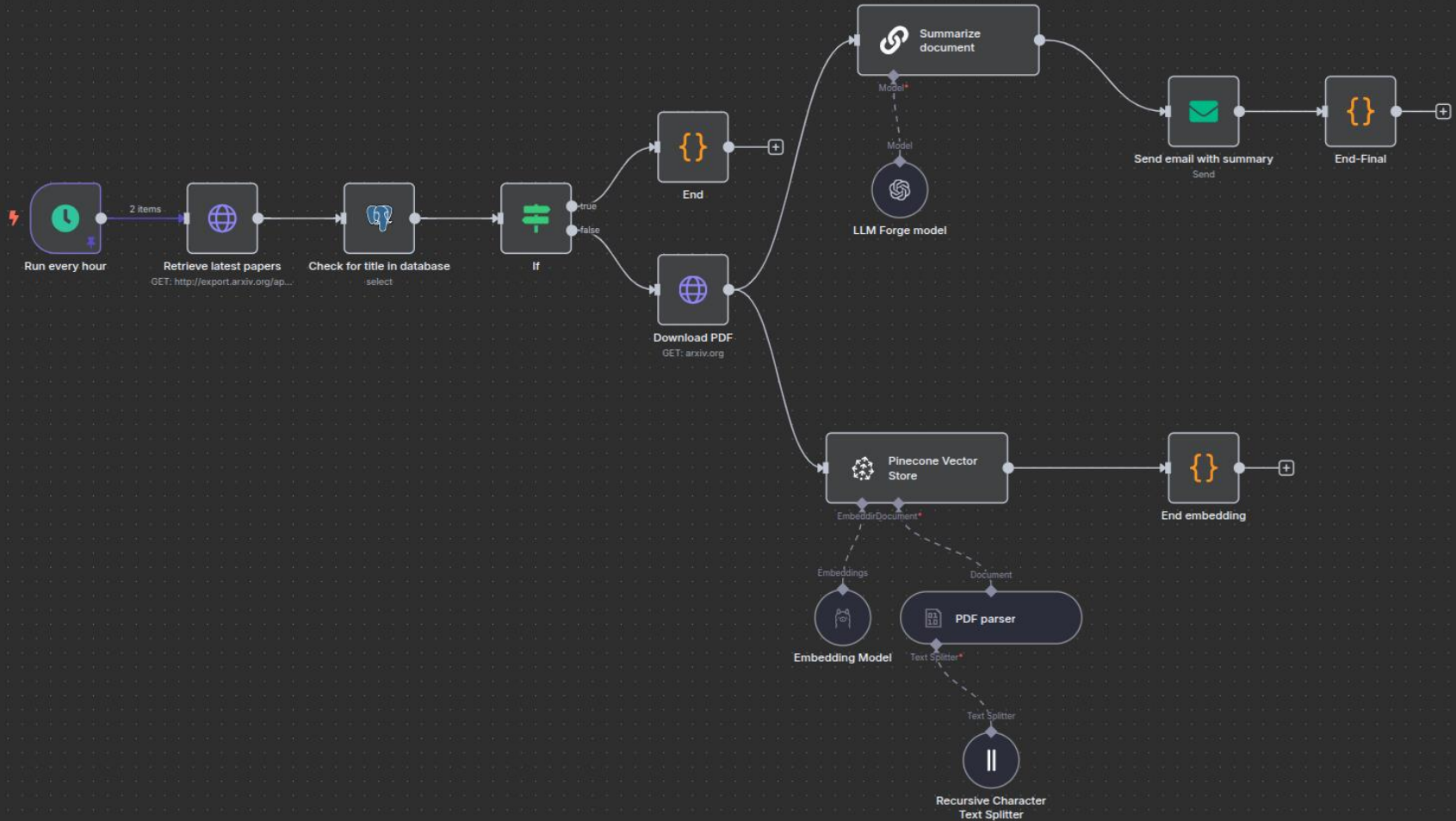
Hello! How are you and who are you?



Hello. I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI."



Integration with n8n automation software





PLGrid Forge - supported model types:

- LLM models
 - Bielik 11B
 - LLAMA 70B
 - incoming: PLuMM, DeepSeek V3
- VLM for image processing (upcoming)
 - Pixtral 12B
- speech to text model support
- embedding models

PLGrid Forge roadmap:

LLMLab:

- support for private, user provided models
- batch request processing
- on-demand model access for less popular models
- access for non-academic
 - detailed accounting
 - user defined usage quotas

A complex network diagram with a central circular pattern and a dense web of nodes and connections. The central part consists of several concentric circles, with the innermost being a solid dark blue circle, followed by several more translucent, lighter blue concentric circles. Surrounding these is a dense, intricate web of light blue lines connecting numerous small, light blue circular nodes. The nodes are distributed across the entire circular area, with some nodes extending further outwards, creating a starburst or sunburst effect. The overall appearance is that of a highly interconnected network or data structure.

Thank you for your attention!