Simple Deep Research

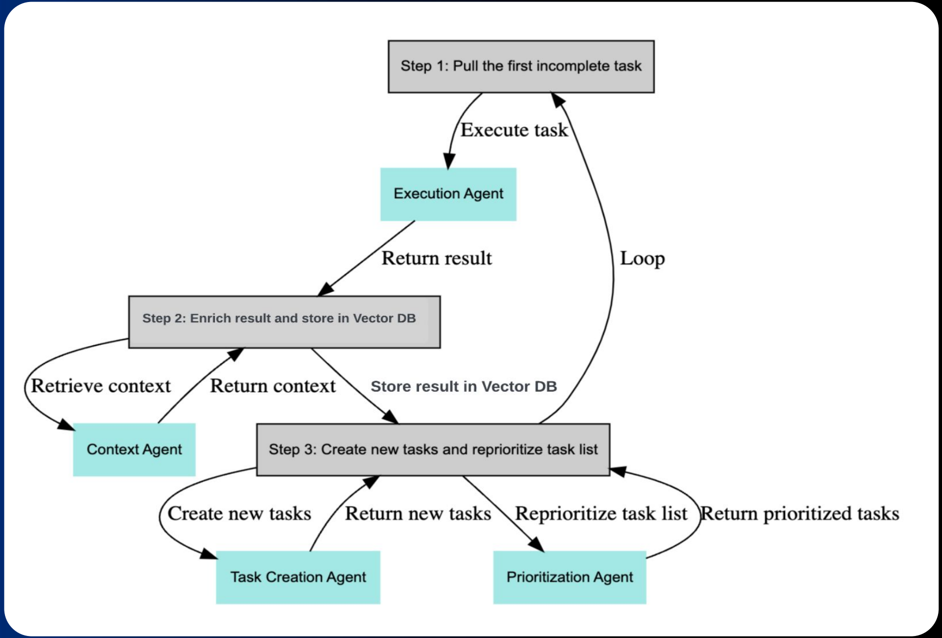
Agenda

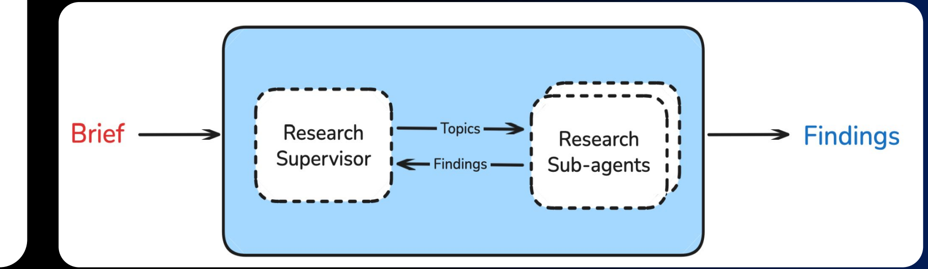
1. Motivation pg. 2
2. Related systems pg. 3
3. Simple Deep Research pg. 4
4. How to use pg. 9
5. Other things pg. 15

Motivation

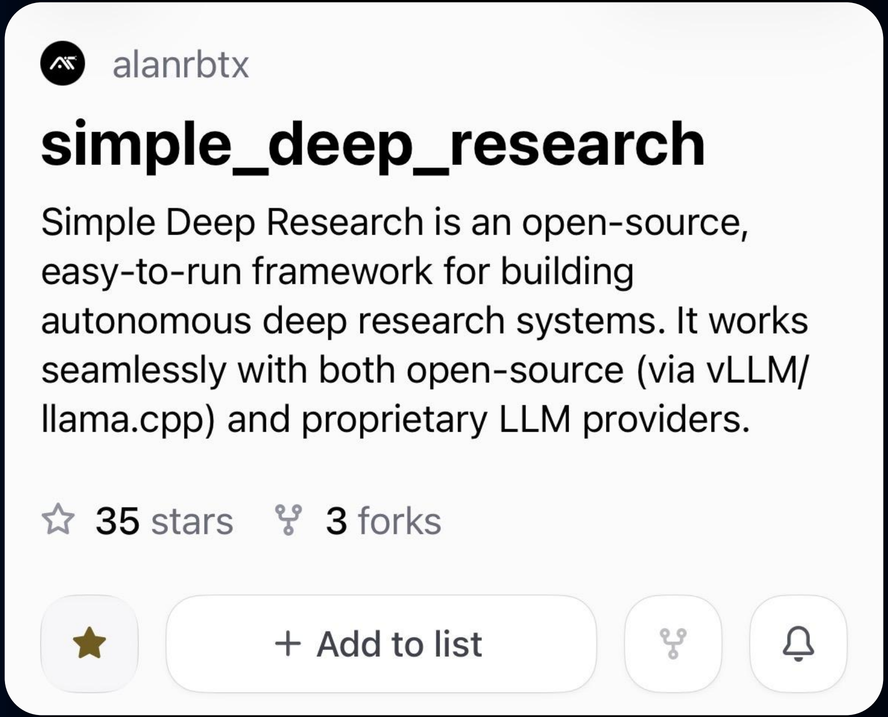
- Easy-to-Run framework
- Without heavy abstractions like LangChain or LangGraph
- Running on local devices

Related systems





Simple Deep Research



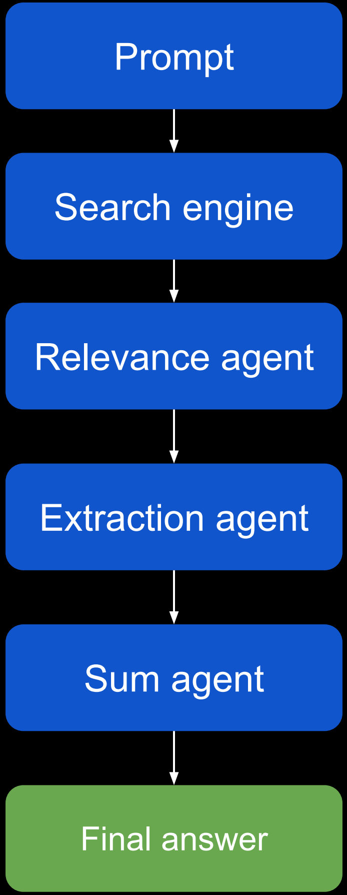
alanrbtx

simple_deep_research

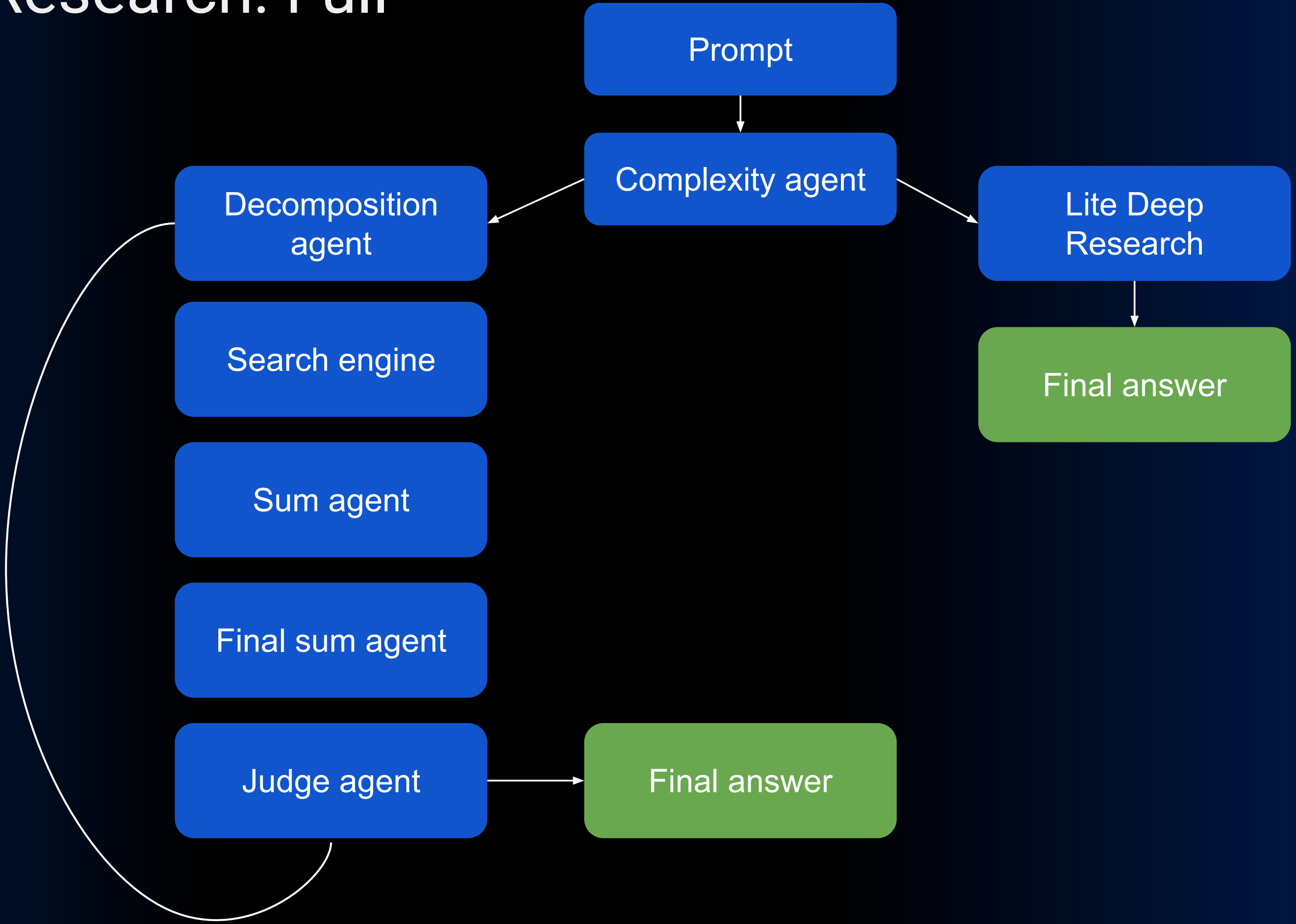
Simple Deep Research is an open-source, easy-to-run framework for building autonomous deep research systems. It works seamlessly with both open-source (via vLLM/ llama.cpp) and proprietary LLM providers.

☆ 35 stars 3 forks

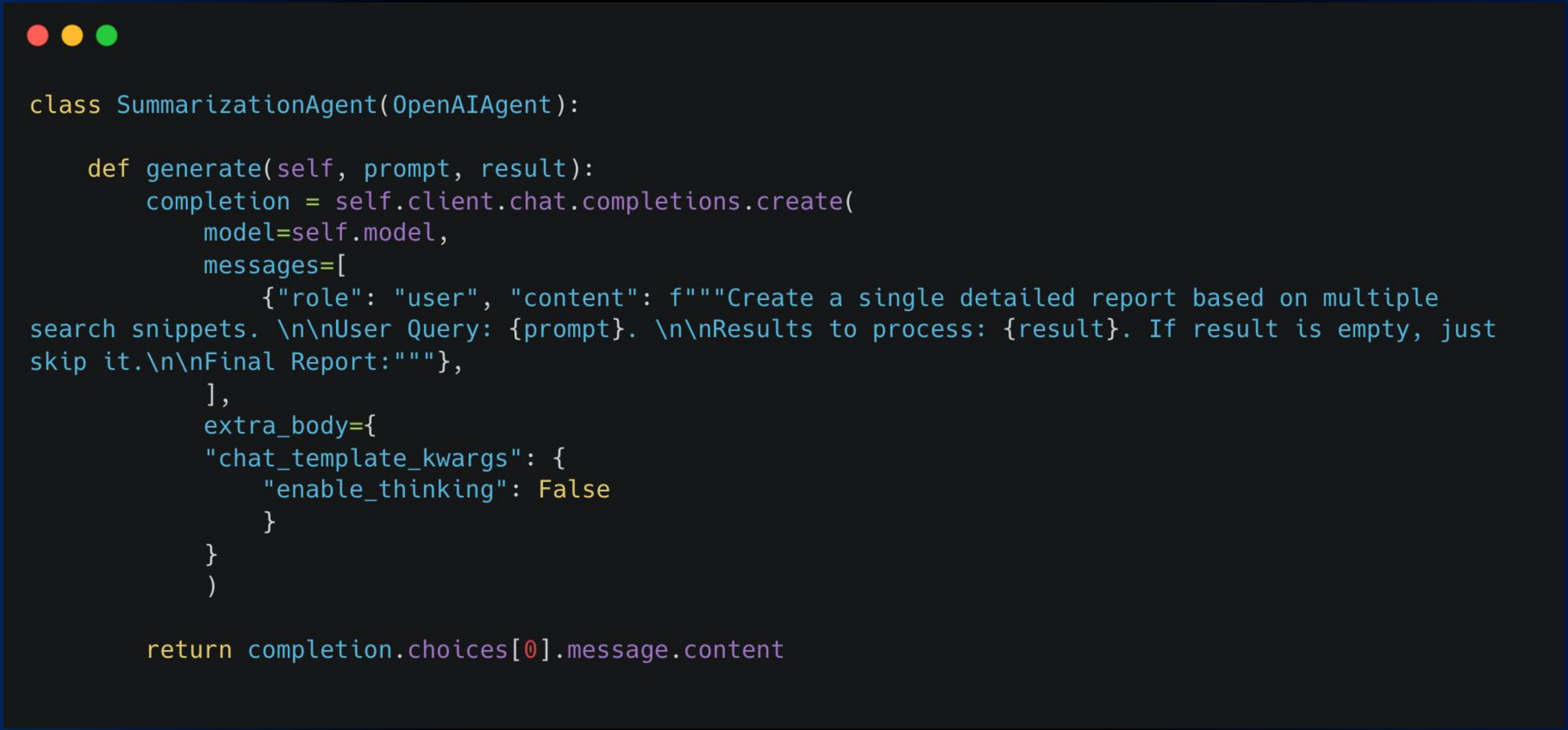
Simple Deep Research: Lite



Simple Deep Research: Full

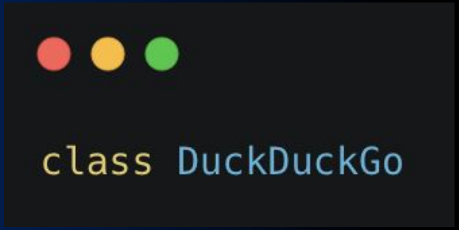


Simple Deep Research: Agents

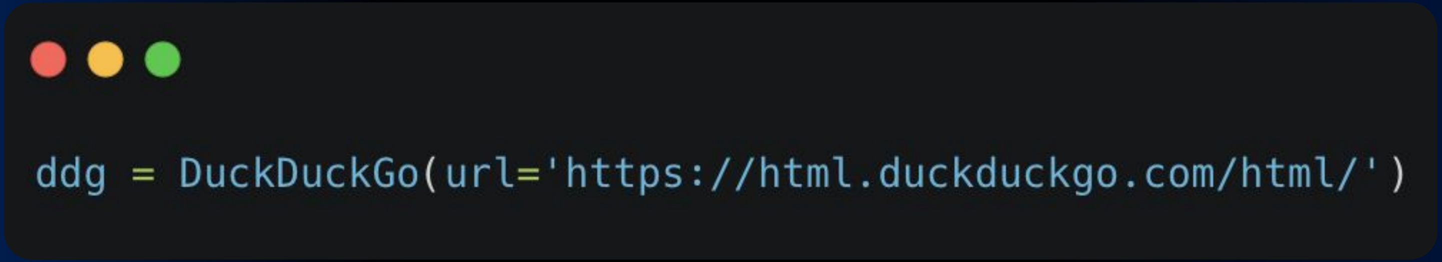


```python
class SummarizationAgent(OpenAIAgent):

    def generate(self, prompt, result):
        completion = self.client.chat.completions.create(
            model=self.model,
            messages=[
                {"role": "user", "content": f"""Create a single detailed report based on multiple search snippets. \n\nUser Query: {prompt}. \n\nResults to process: {result}. If result is empty, just skip it.\n\nFinal Report:"""},
            ],
            extra_body={
            "chat_template_kwargs": {
                "enable_thinking": False
                }
            }
            )

        return completion.choices[0].message.content
```

Simple Deep Research: Search engine



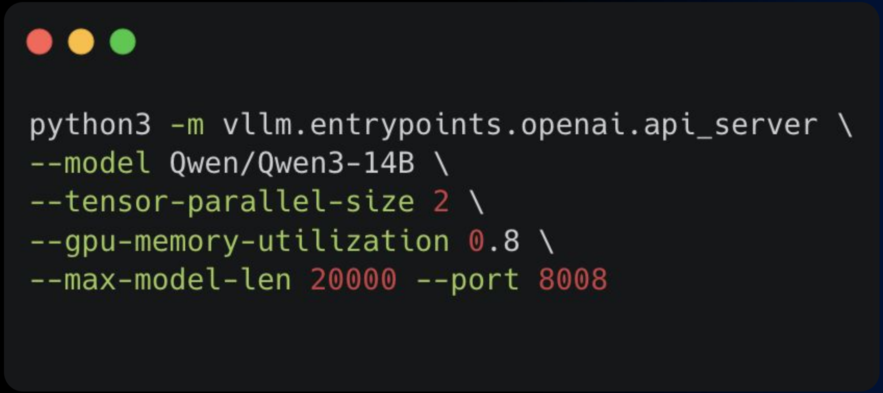
```
class DuckDuckGo
```



```
ddg = DuckDuckGo(url='https://html.duckduckgo.com/html/')
```

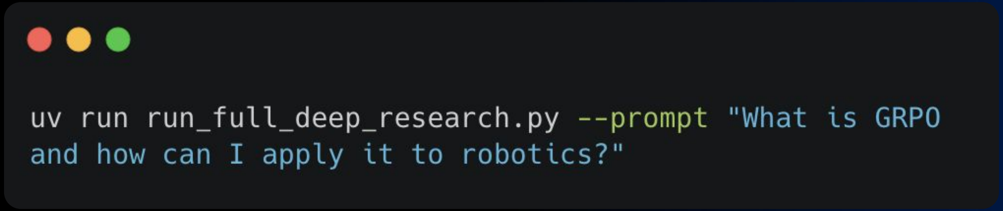
How to use

Run model with vLLM



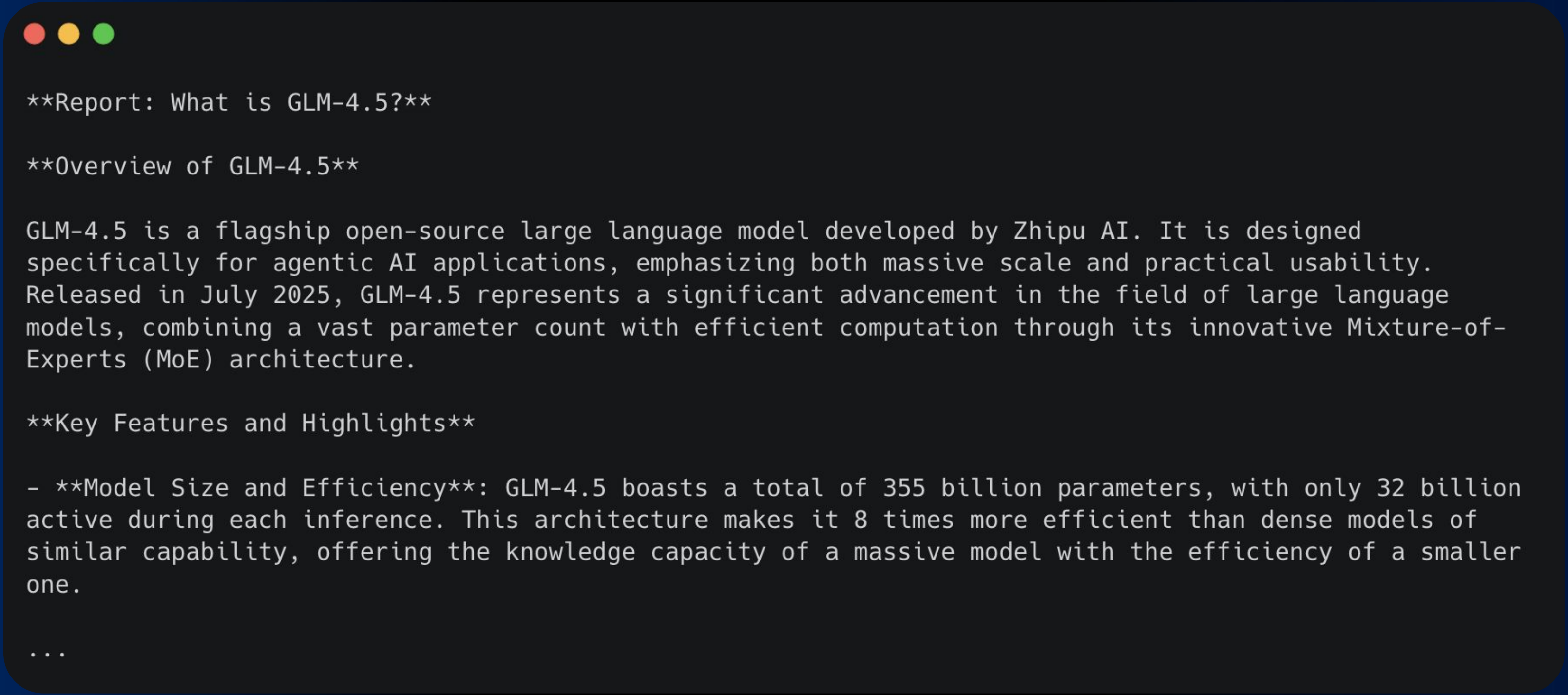
```
python3 -m vllm.entrypoints.openai.api_server \
--model Qwen/Qwen3-14B \
--tensor-parallel-size 2 \
--gpu-memory-utilization 0.8 \
--max-model-len 20000 --port 8008
```

Run Simple Deep Research



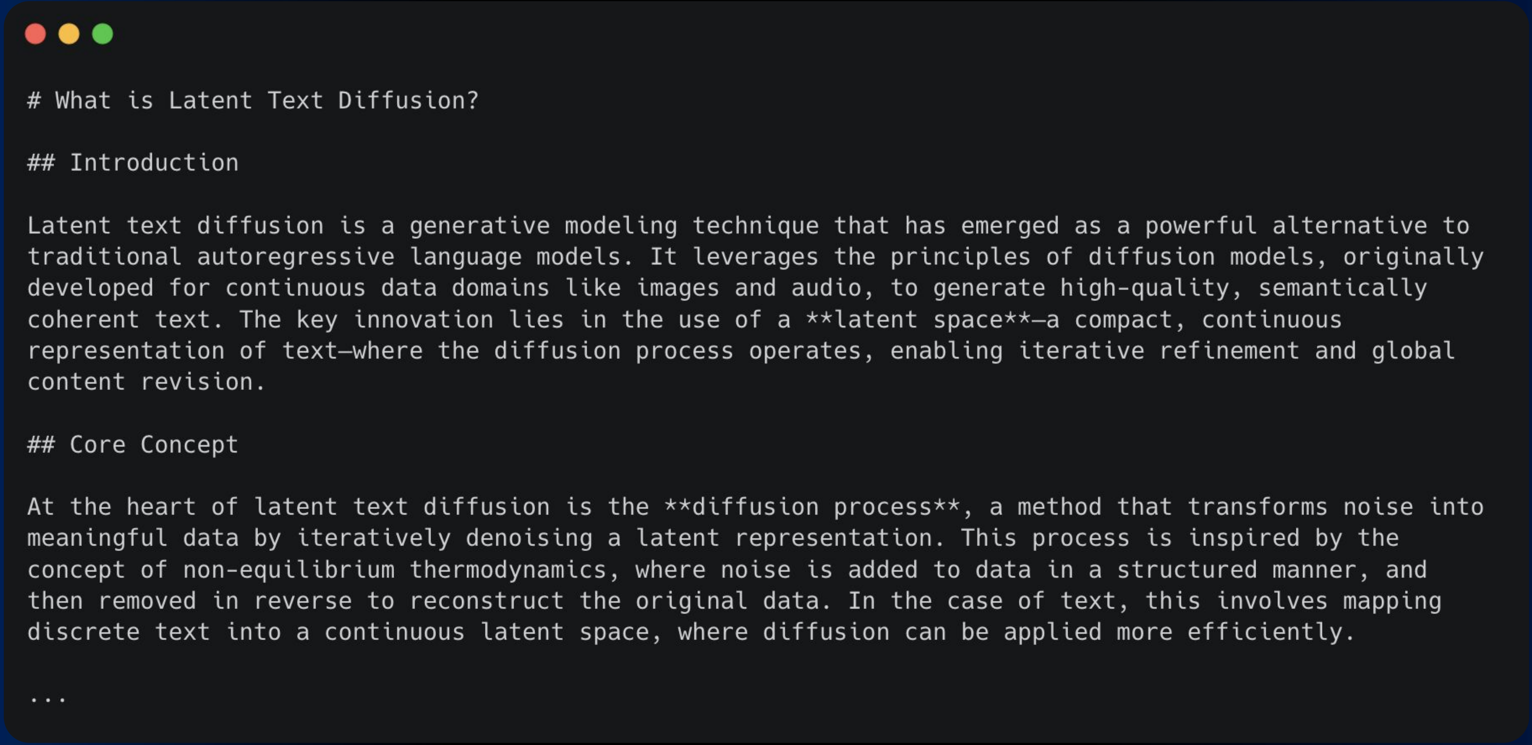
```
uv run run_full_deep_research.py --prompt "What is GRPO
and how can I apply it to robotics?"
```

Examples: Lite



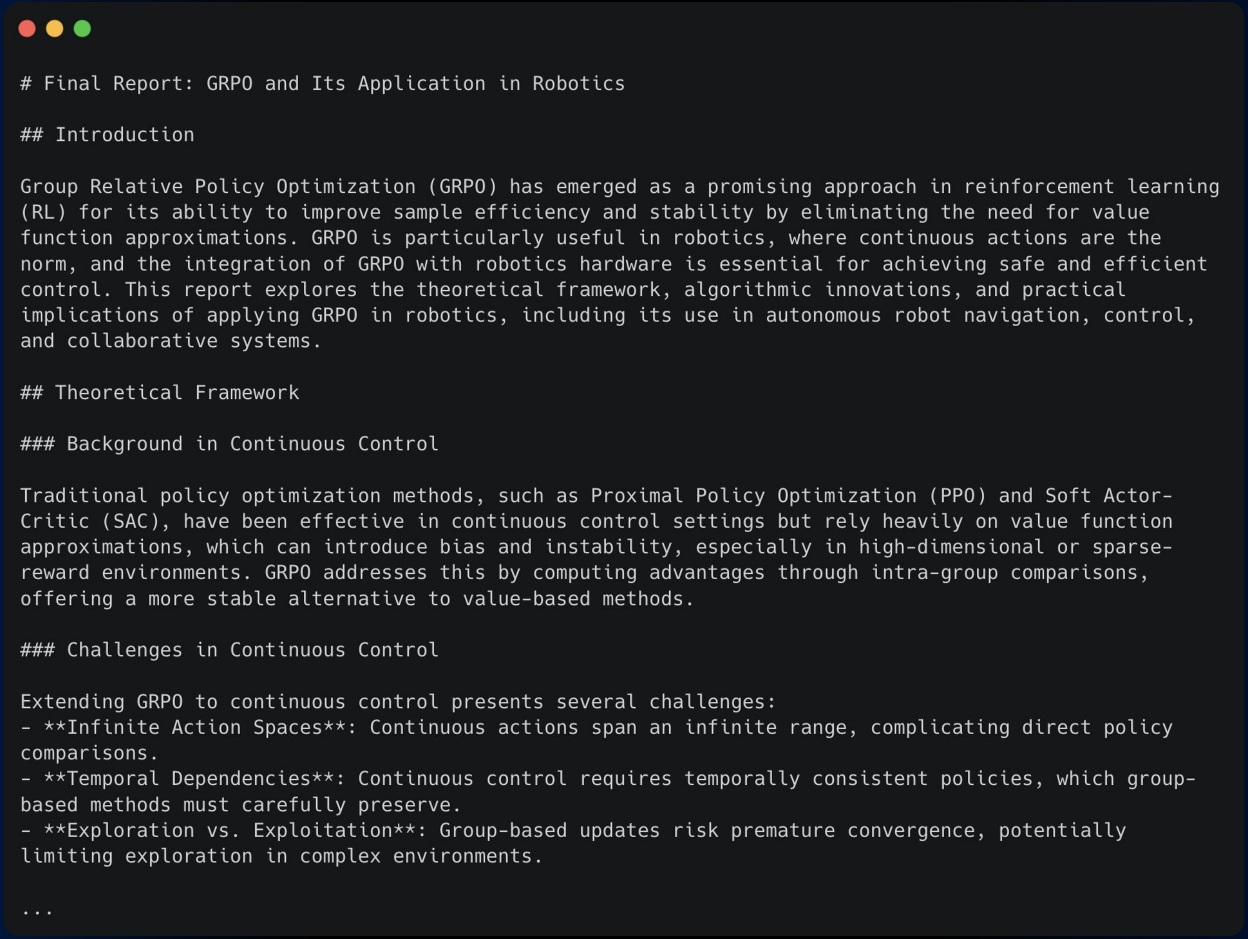
```
**Report: What is GLM-4.5?**

**Overview of GLM-4.5**

GLM-4.5 is a flagship open-source large language model developed by Zhipu AI. It is designed
specifically for agentic AI applications, emphasizing both massive scale and practical usability.
Released in July 2025, GLM-4.5 represents a significant advancement in the field of large language
models, combining a vast parameter count with efficient computation through its innovative Mixture-of-
Experts (MoE) architecture.

**Key Features and Highlights**

- **Model Size and Efficiency**: GLM-4.5 boasts a total of 355 billion parameters, with only 32 billion
active during each inference. This architecture makes it 8 times more efficient than dense models of
similar capability, offering the knowledge capacity of a massive model with the efficiency of a smaller
one.

...
```

Examples: Lite



```
# What is Latent Text Diffusion?

## Introduction

Latent text diffusion is a generative modeling technique that has emerged as a powerful alternative to
traditional autoregressive language models. It leverages the principles of diffusion models, originally
developed for continuous data domains like images and audio, to generate high-quality, semantically
coherent text. The key innovation lies in the use of a **latent space**—a compact, continuous
representation of text—where the diffusion process operates, enabling iterative refinement and global
content revision.

## Core Concept

At the heart of latent text diffusion is the **diffusion process**, a method that transforms noise into
meaningful data by iteratively denoising a latent representation. This process is inspired by the
concept of non-equilibrium thermodynamics, where noise is added to data in a structured manner, and
then removed in reverse to reconstruct the original data. In the case of text, this involves mapping
discrete text into a continuous latent space, where diffusion can be applied more efficiently.

...
```

Examples: Full



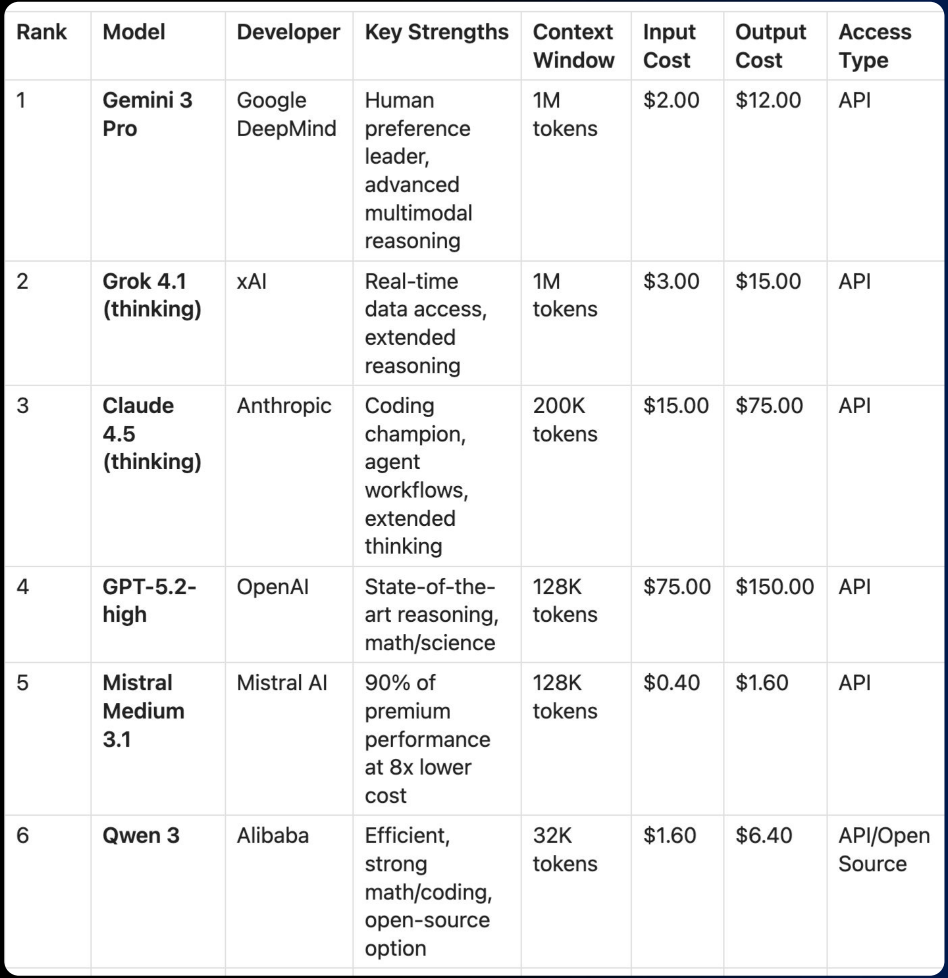
```
# Final Report: GRPO and Its Application in Robotics

## Introduction

Group Relative Policy Optimization (GRPO) has emerged as a promising approach in reinforcement learning
(RL) for its ability to improve sample efficiency and stability by eliminating the need for value
function approximations. GRPO is particularly useful in robotics, where continuous actions are the
norm, and the integration of GRPO with robotics hardware is essential for achieving safe and efficient
control. This report explores the theoretical framework, algorithmic innovations, and practical
implications of applying GRPO in robotics, including its use in autonomous robot navigation, control,
and collaborative systems.

## Theoretical Framework

### Background in Continuous Control

Traditional policy optimization methods, such as Proximal Policy Optimization (PPO) and Soft Actor-
Critic (SAC), have been effective in continuous control settings but rely heavily on value function
approximations, which can introduce bias and instability, especially in high-dimensional or sparse-
reward environments. GRPO addresses this by computing advantages through intra-group comparisons,
offering a more stable alternative to value-based methods.

### Challenges in Continuous Control

Extending GRPO to continuous control presents several challenges:
- **Infinite Action Spaces**: Continuous actions span an infinite range, complicating direct policy
comparisons.
- **Temporal Dependencies**: Continuous control requires temporally consistent policies, which group-
based methods must carefully preserve.
- **Exploration vs. Exploitation**: Group-based updates risk premature convergence, potentially
limiting exploration in complex environments.

...
```

Examples: Full

Prompt: What is the best LLM in 2026?

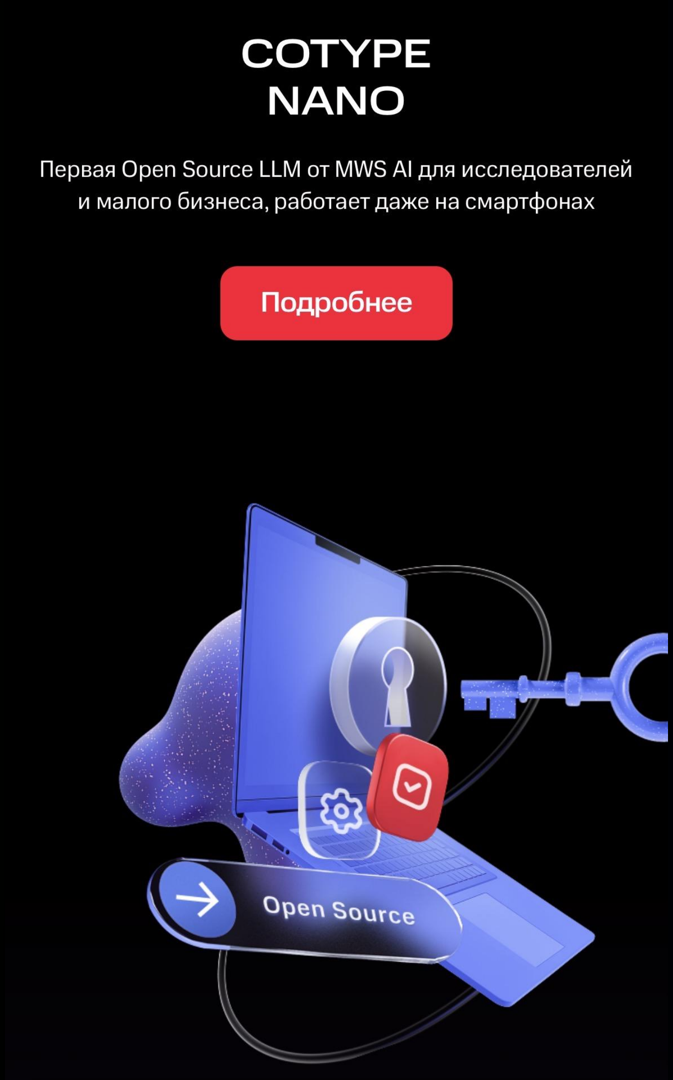
Sub-prompt-1: Top LLMs for 2026: Performance, use cases, and industry adoption

Sub-prompt-2: Best LLMs for 2026: Cost-effectiveness, accessibility, and open-source options



Rank	Model	Developer	Key Strengths	Context Window	Input Cost	Output Cost	Access Type
1	**Gemini 3 Pro**	Google DeepMind	Human preference leader, advanced multimodal reasoning	1M tokens	$2.00	$12.00	API
2	**Grok 4.1 (thinking)**	xAI	Real-time data access, extended reasoning	1M tokens	$3.00	$15.00	API
3	**Claude 4.5 (thinking)**	Anthropic	Coding champion, agent workflows, extended thinking	200K tokens	$15.00	$75.00	API
4	**GPT-5.2-high**	OpenAI	State-of-the-art reasoning, math/science	128K tokens	$75.00	$150.00	API
5	**Mistral Medium 3.1**	Mistral AI	90% of premium performance at 8x lower cost	128K tokens	$0.40	$1.60	API
6	**Qwen 3**	Alibaba	Efficient, strong math/coding, open-source option	32K tokens	$1.60	$6.40	API/Open Source

Other things



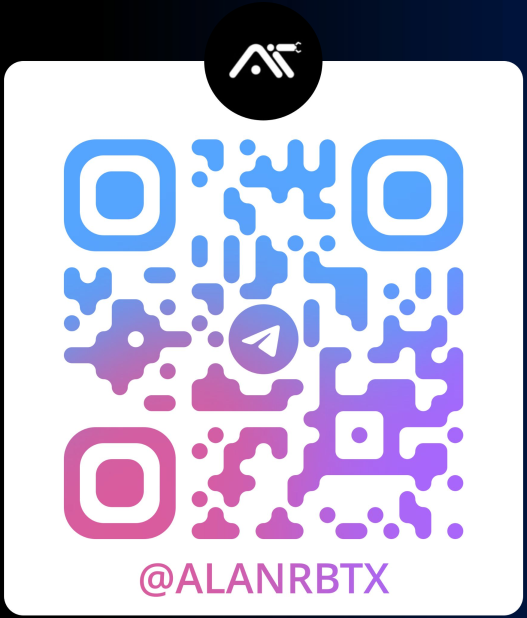
COTYPE
NANO

Первая Open Source LLM от MWS AI для исследователей
и малого бизнеса, работает даже на смартфонах

Подробнее

Simple Deep Research

Tg-channel



@ALANRBTX