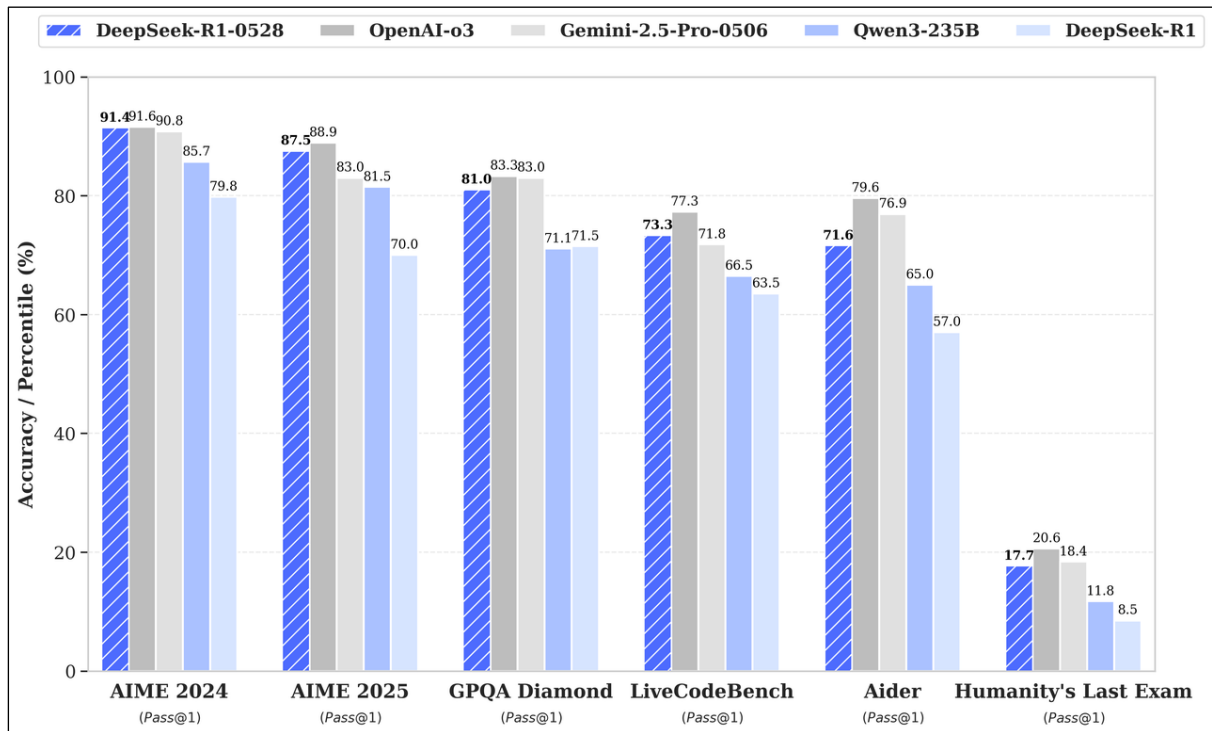
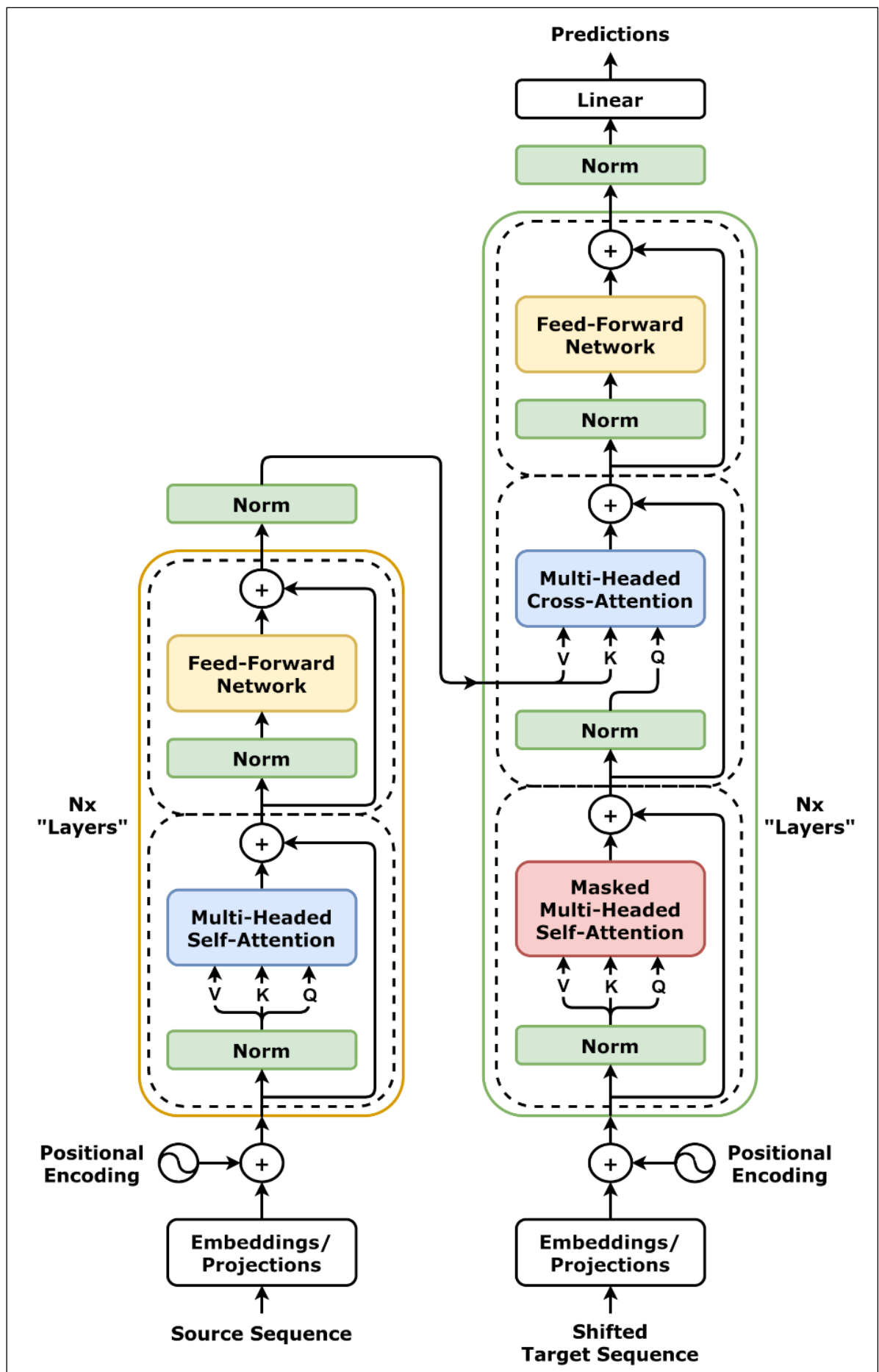
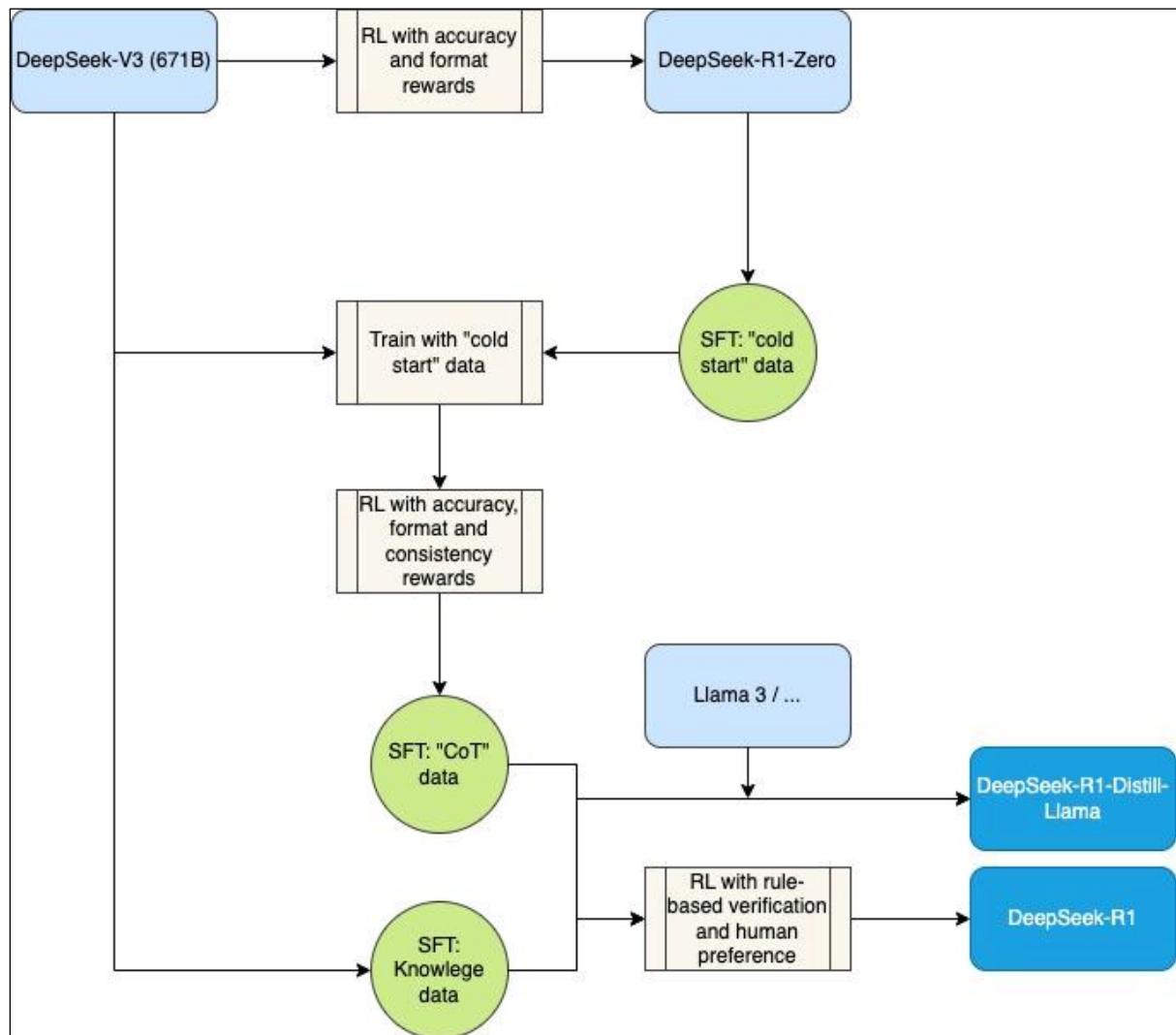
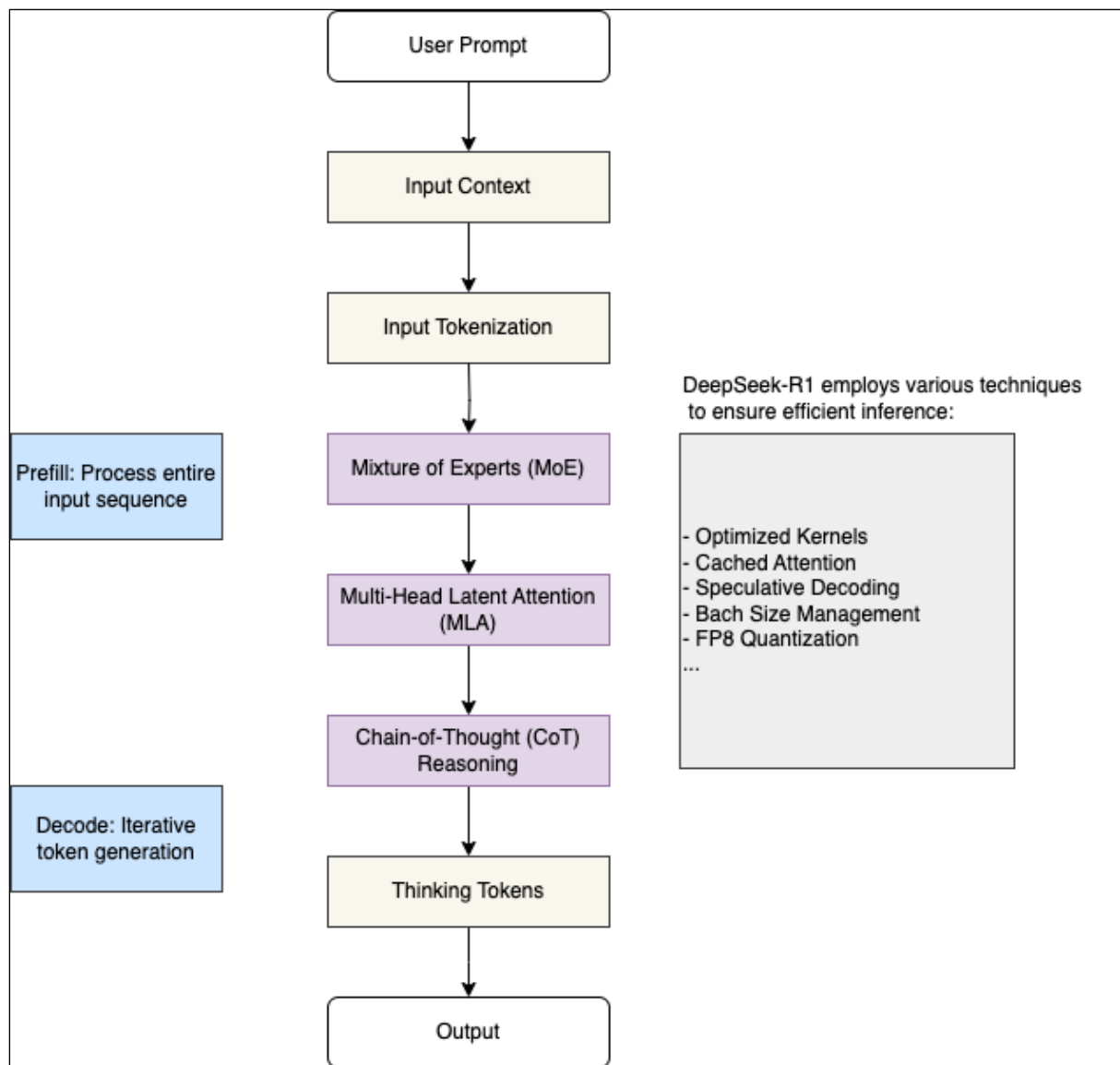


Chapter 1: What is DeepSeek?

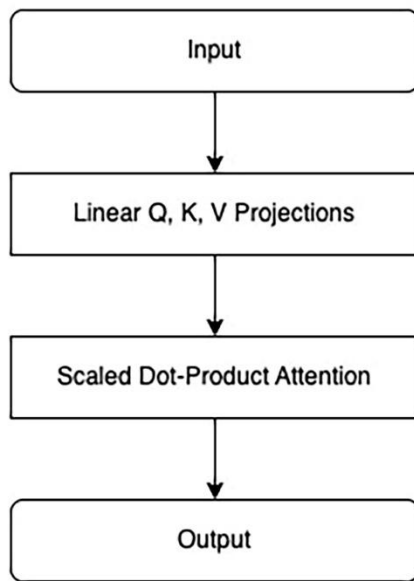




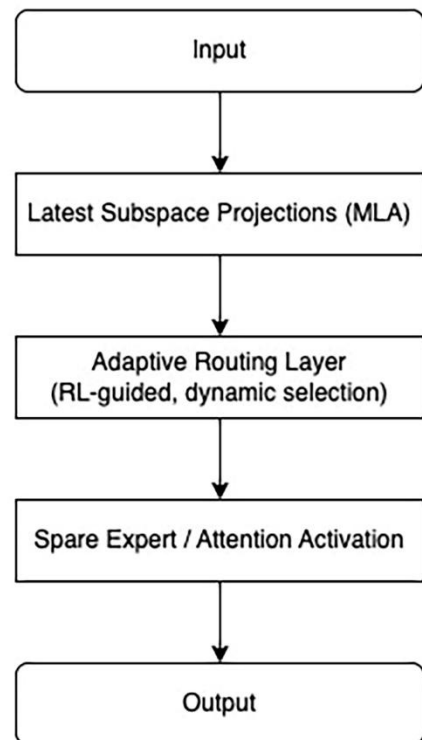




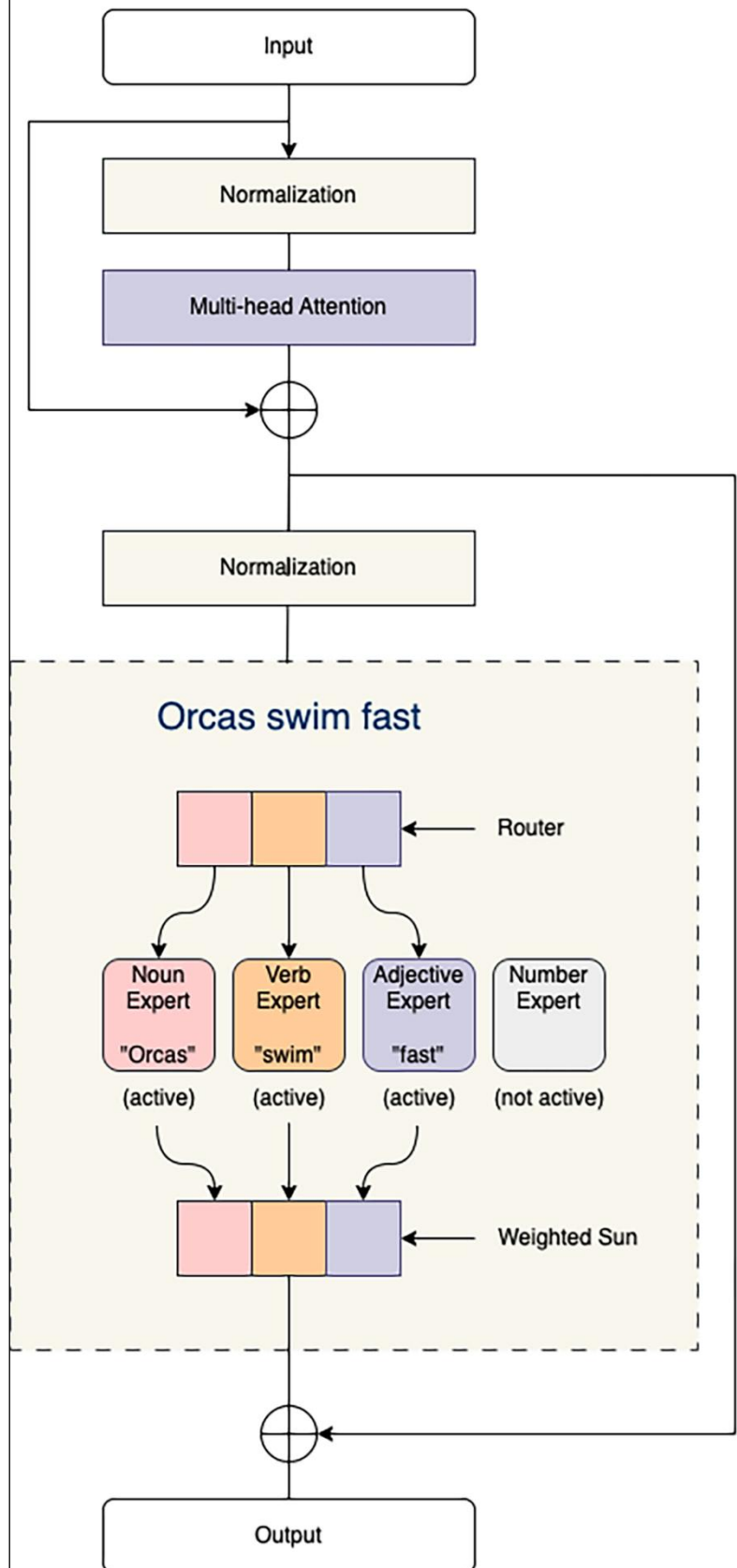
Traditional Transformer Attention (Q, K, V)



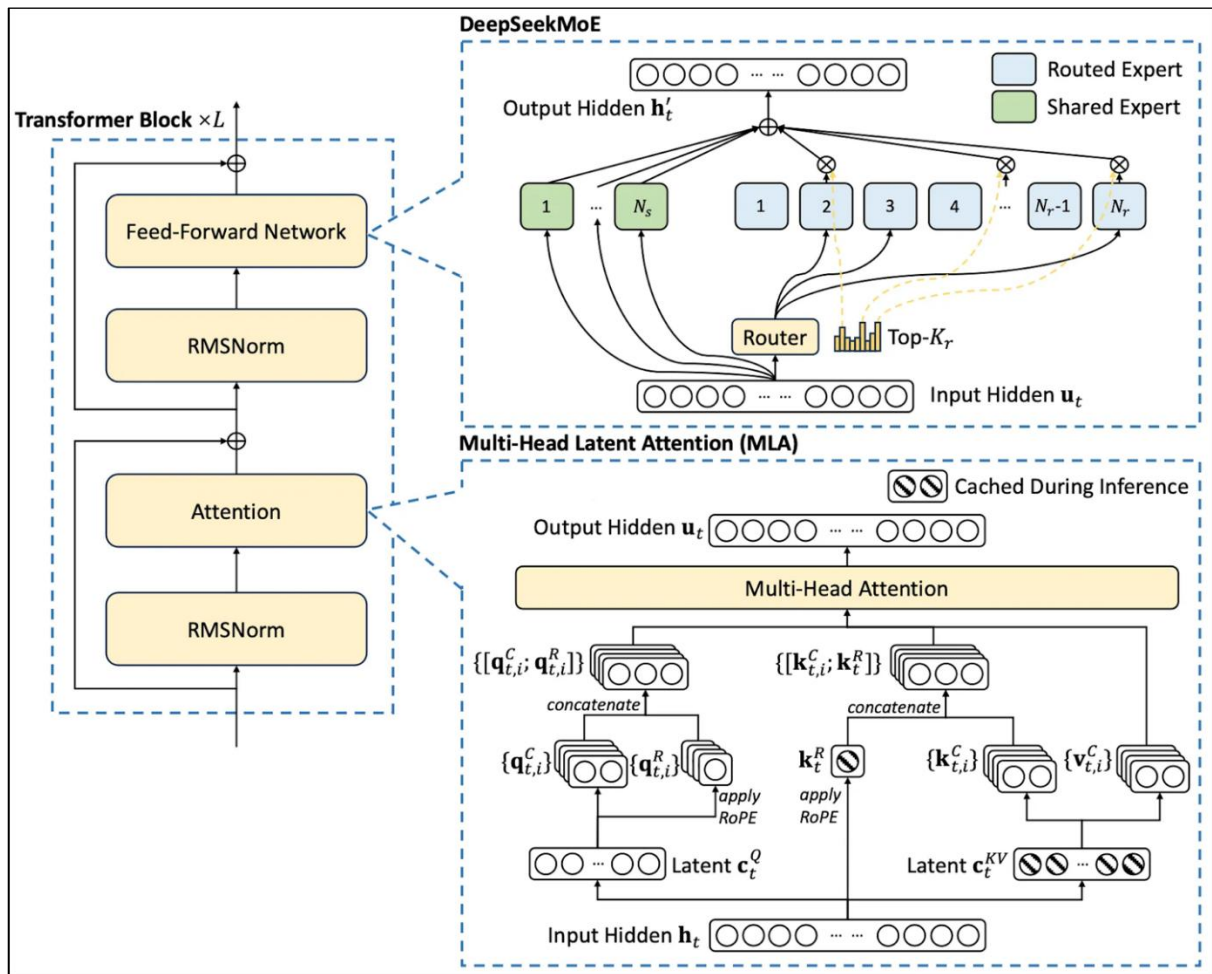
DeepSeek R1 Adaptive Attention Routing

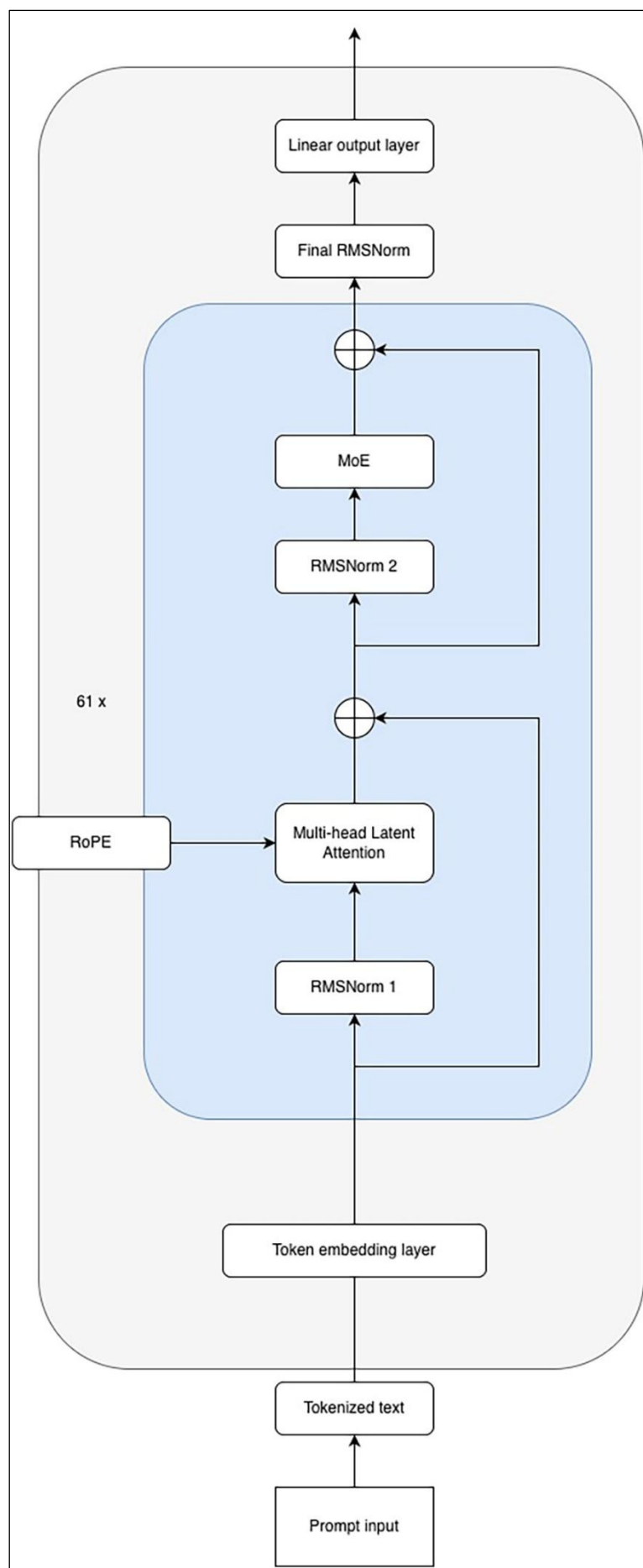


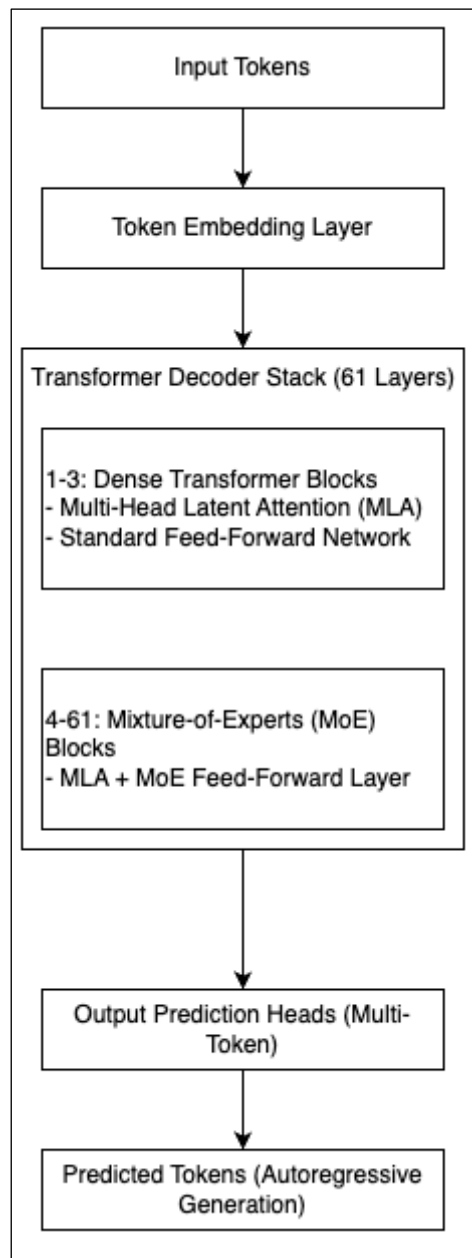
Mixture of Experts (MoE)

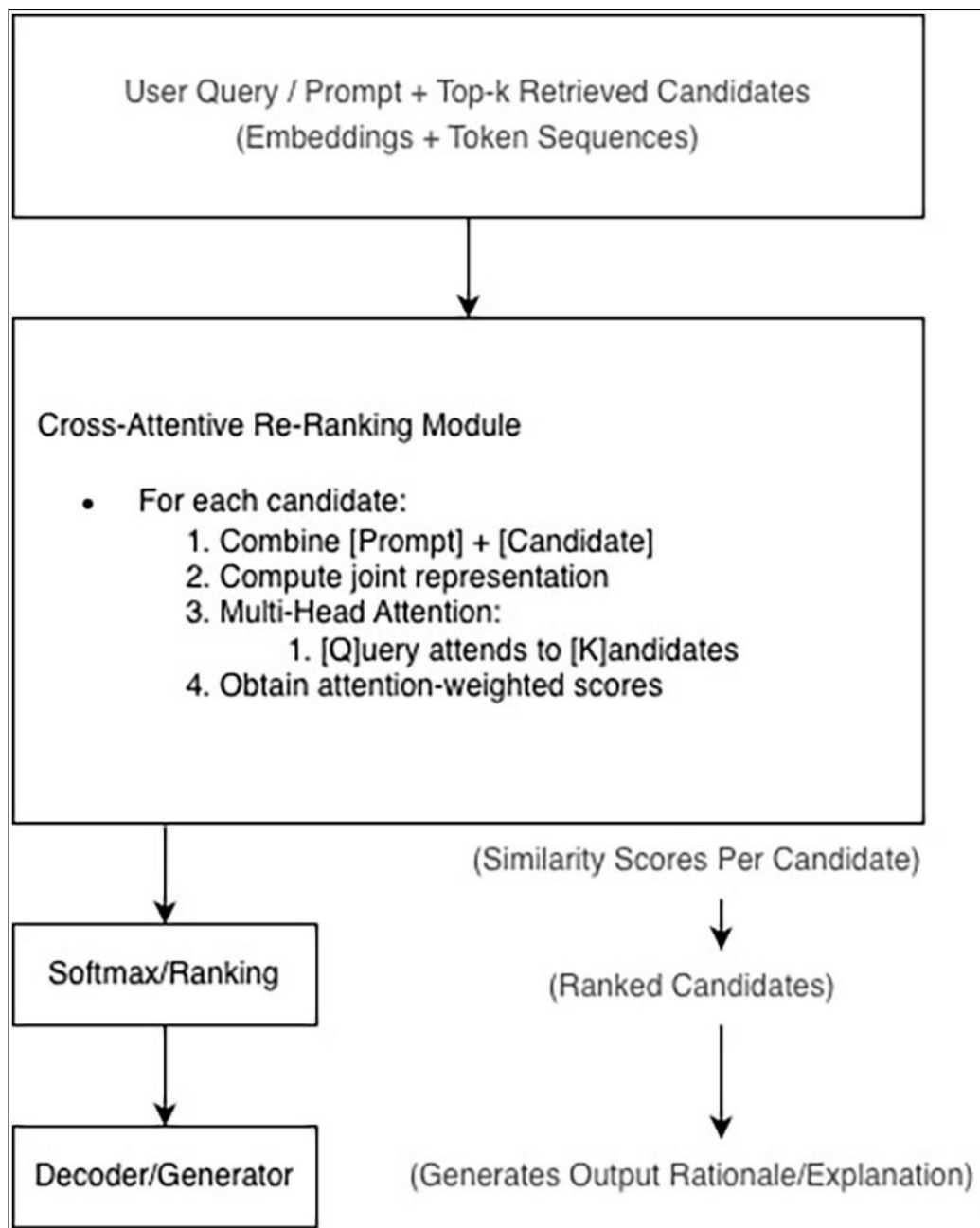


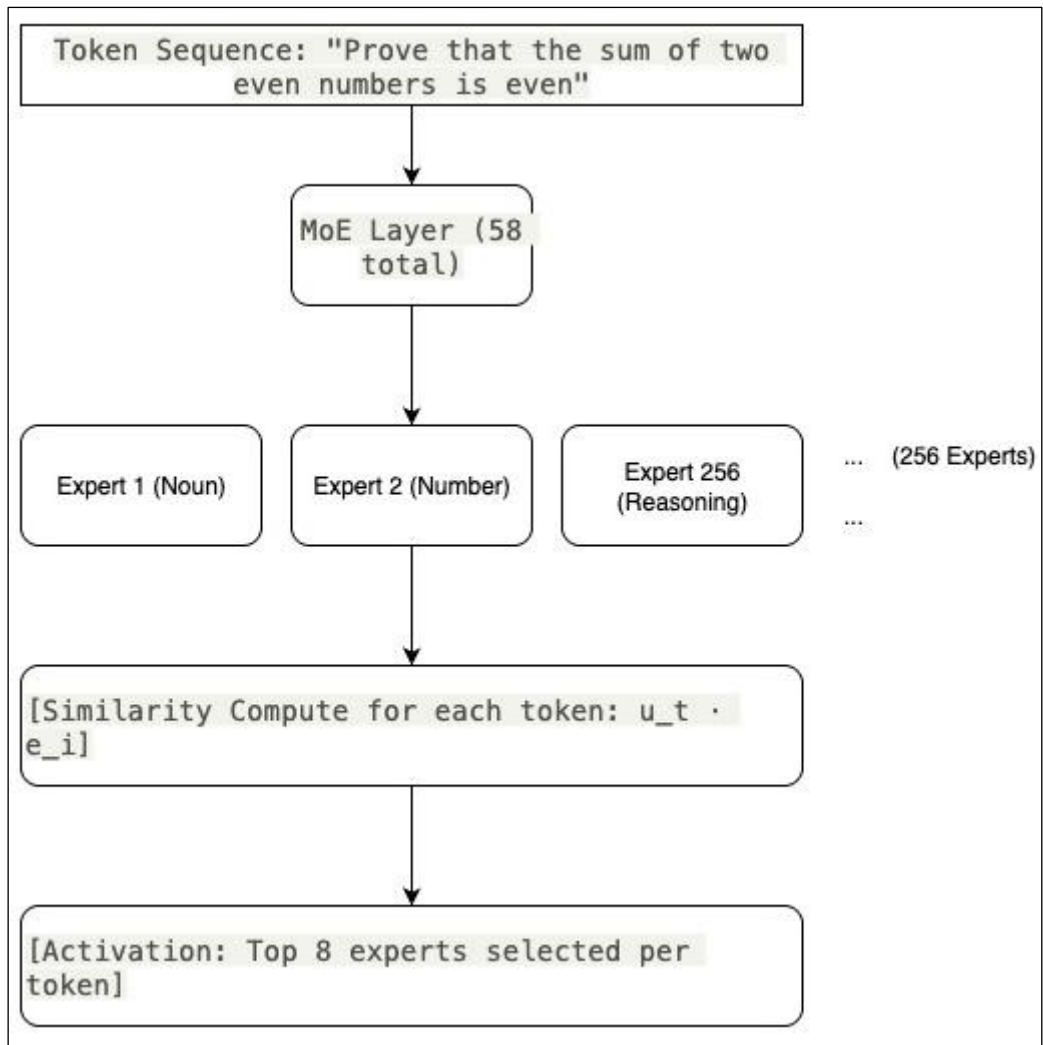
Chapter 2: Deep Dive into DeepSeek











Stage 1: Training VL Adaptor

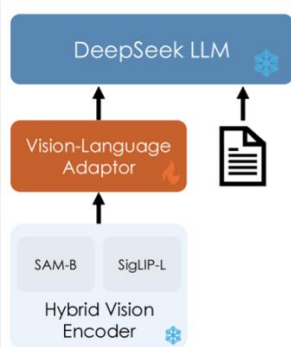
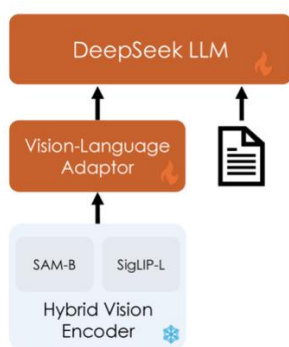


Image-Text Pairs

Stage 2: Joint VL Pre-training

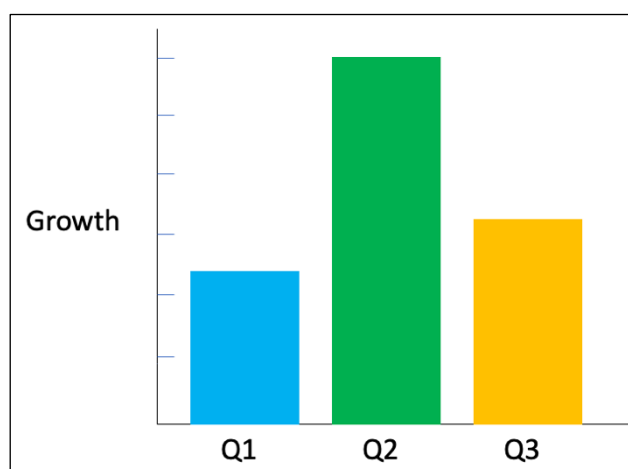


Interleaved VL +
Pure Language Sequences

Stage 3: Supervised Finetuning



VL Chat Data +
Pure Language Chat Data



Chapter 3: Prompting DeepSeek

 Mistral Medium 3.1

The use of a **black-box AI triage system** in emergency rooms (ERs)—even if empirically more accurate than human doctors—raises complex ethical, practical, and societal questions. Below is a balanced discussion of the **pros and cons**, followed by a **recommendation** based on ethical principles and real-world constraints.

Pros of Using a Black-Box AI Triage System

1. Improved Accuracy and Speed

- If the AI demonstrates superior diagnostic accuracy (e.g., lower misdiagnosis rates, faster identification of critical cases like strokes or sepsis), it could save lives and reduce harm from human error or cognitive biases (e.g., anchoring, availability bias).
- Example: Studies show AI can outperform radiologists in detecting certain conditions (e.g., pneumonia on X-rays) or predict patient deterioration hours before clinical signs appear.

2. Reduction in Human Burnout

- ER doctors often face extreme stress and fatigue, leading to errors. Offloading triage decisions to AI could reduce burnout and improve working conditions.

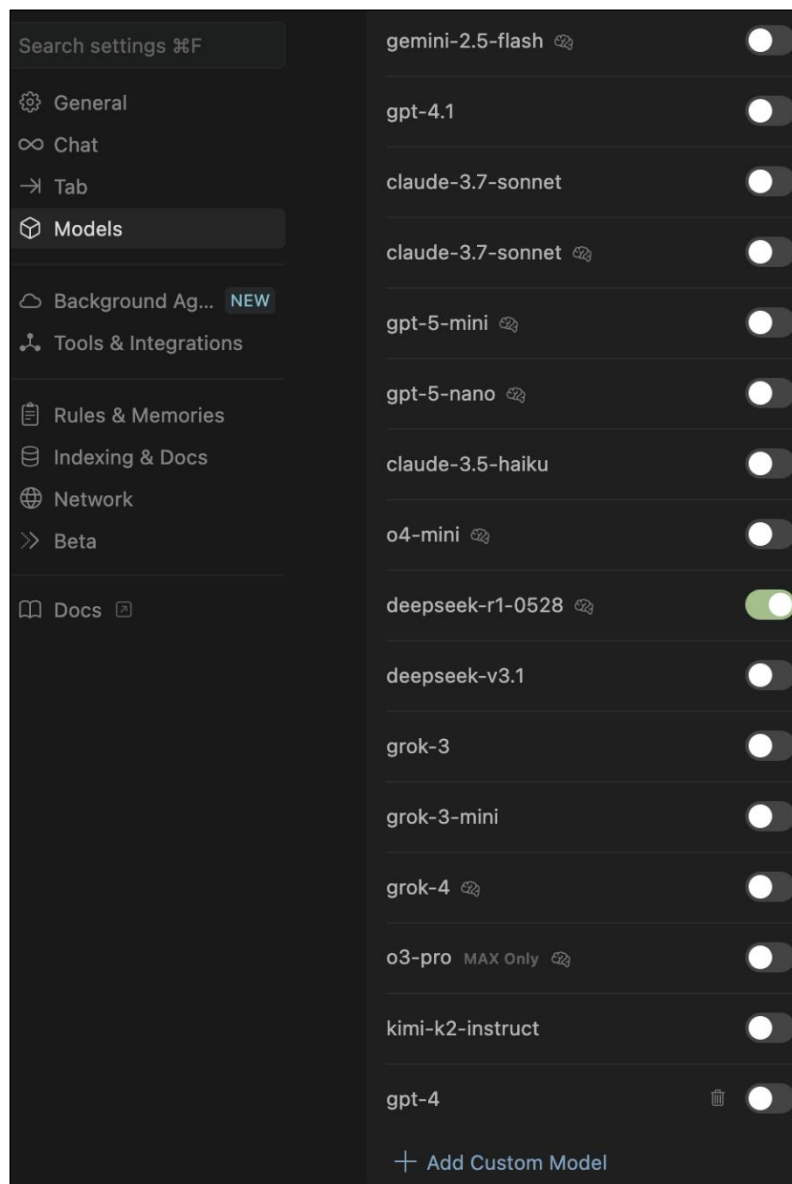
3. Scalability and Consistency

- AI can handle high patient volumes without performance degradation, ensuring consistent care quality across shifts, hospitals, or geographic regions.
- Useful in underserved areas with clinician shortages.

4. Data-Driven Insights

- AI could identify patterns humans miss (e.g., subtle interactions between symptoms, lab results, and patient history) and improve over time with more data.

Chapter 4: Using DeepSeek: Case Studies



Chapter 5: Building with DeepSeek

Engine	GitHub repo	Primary innovation	Performance optimizations	Quantization (built-in)	Distributed inference	Ease of use	Unique feature	GitHub stars	Open issues
vLLM	https://github.com/vllm-project/vllm	PageAttention with GPU-friendly KV paging	Continuous batching, fused CUDA kernels, async I/O	INT8 / INT4 (AWQ, GPTQ), SmoothQuant	Tensor & pipeline parallel	High (single-command serve)	OpenAI-compatible API; live KV-cache swap	49.2k	1.9k
Hugging Face TGI	https://github.com/huggingface/text-generation-inference	Rust + Python gRPC server	Token streaming, continuous batching, speculative decoding, FlashAttention	GPTQ, AWQ, bits-and-bytes, Marlin	Optional sharding & tensor parallel	High (Docker/CLI)	Triton backend, Prometheus metrics	10.2k	255
SGLang	https://github.com/sgl-project/sglang	Prefill/Decode disaggregation; RadixAttention	Continuous batching, expert/pipeline/tensor parallel	FP8, INT4 (AWQ/GPTQ)	Data, tensor & pipeline parallel	Medium (extra DSL)	Vision-LLM support; built-in router	15k	466
LMDeploy	https://github.com/InternLM/lmdeploy	Compression-to-serving toolkit	Persistent batch, blocked KV, split-fuse kernels	4-bit weight+KV, AWQ, SmoothQuant	Tensor parallel; K8s Helm chart	High (one executable)	Integrated quantize→deploy flow	6.5k	448
TensorRT-LLM	https://github.com/NVIDIA/TensorRT-LLM	End-to-end TensorRT graph compilation	Custom kernels, inflight batching, paged KV	FP8, FP4, INT4 (AWQ), INT8	Triton backend (multi-GPU/node)	Medium (requires TRT build)	Leader/worker orchestration	10.7k	608

Garmin Health Summary API

0.1.0OAS 3.1

/openapi.json

default

GET / Root

POST /health-summaryGet Health Summary

Get daily health summary for a specific date.
Args: request: Health summary request with date and model garmin_email: Garmin account email from header garmin_password: Garmin account password from header
Returns: Daily health summary with AI-generated insights

Parameters

Name

Description

garmin-emailrequired

Garmin email address

string

garmin-email

(header)

garmin-passwordrequired

Garmin password

string

garmin-password

(header)

Request bodyrequired

application/json

Example Value

Schema

```
{  "date": "2025-10-20",  "model": "deepseek-chat"}
```

Responses

DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249

Actions

Create batch transform job

Create endpoint

Model settings

Name

DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249

ARN

arn:aws:sagemaker:eu-central-1:427525985442:model/DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249

Creation time

8/24/2025, 10:32:51 AM

IAM role ARN

arn:aws:iam::427525985442:role/DuarteStack-DuarteBedrockSageMakerRoleFF6A8A20-qn1nIHfunLaT

Container 1

Container Name

Container 1

Image

763104351884.dkr.ecr.eu-central-1.amazonaws.com/djl-inference:0.33.0-imi15.0.0-cu128

Training job

-

Model data location

-

Mode

Single model

Environment variables

Key	Value
HF_MODEL_ID	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
OPTION_ENABLE_REASONING	true
OPTION_REASONING_PARSER	deepseek_r1
TENSOR_PARALLEL_DEGREE	max

Endpoint summary

Name
DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249-endpoint

ARN
arn:aws:sagemaker:eu-central-1:427525985442:endpoint/DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249-endpoint

URL
<https://runtime.sagemaker.eu-central-1.amazonaws.com/endpoints/DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249-endpoint/invocations>
[Learn more about the API](#)

Status
InService

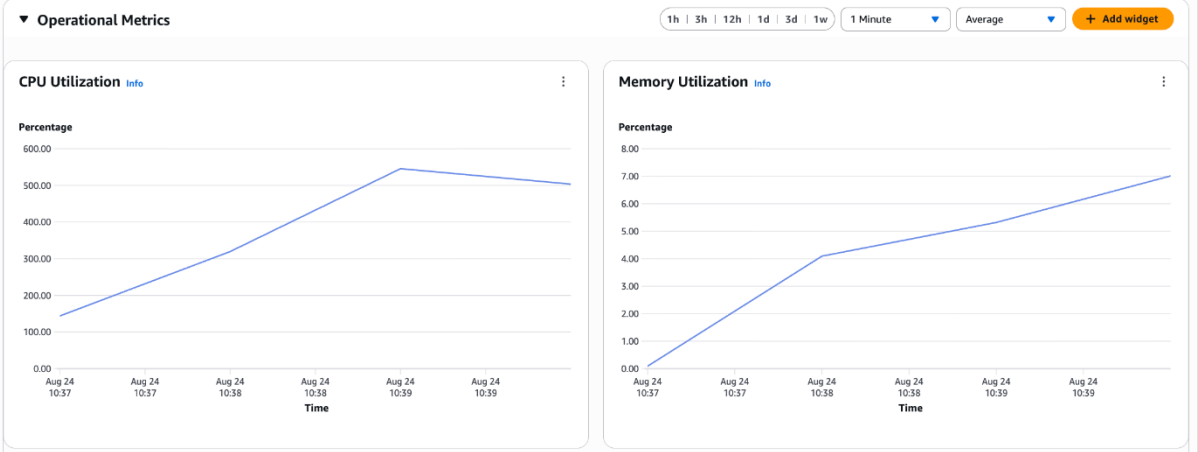
Creation time
Sun Aug 24 2025 10:32:51 GMT+0200 (Central European Summer Time)

Model container logs
[/aws/sagemaker/endpoints/DeepSeek-R1-Distill-Qwen-14B-2025-08-24-08-32-49-249-endpoint](#)

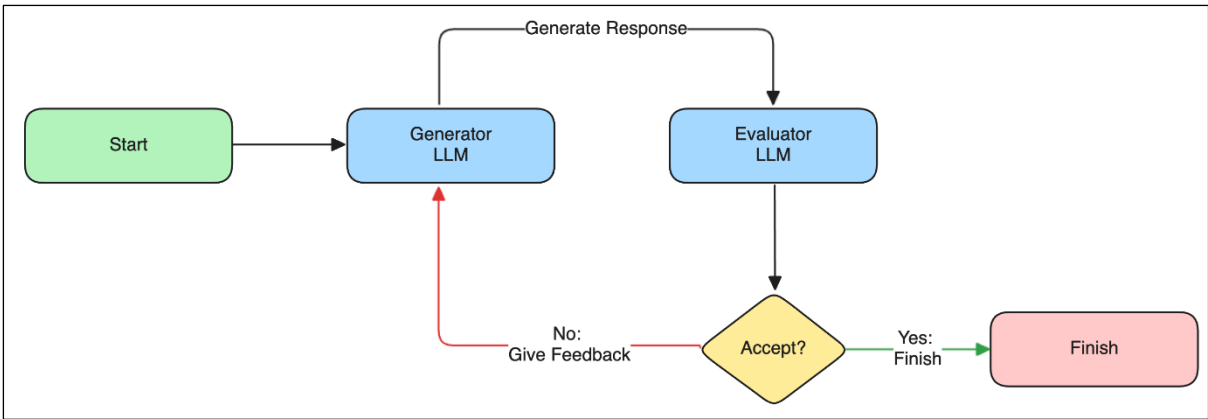
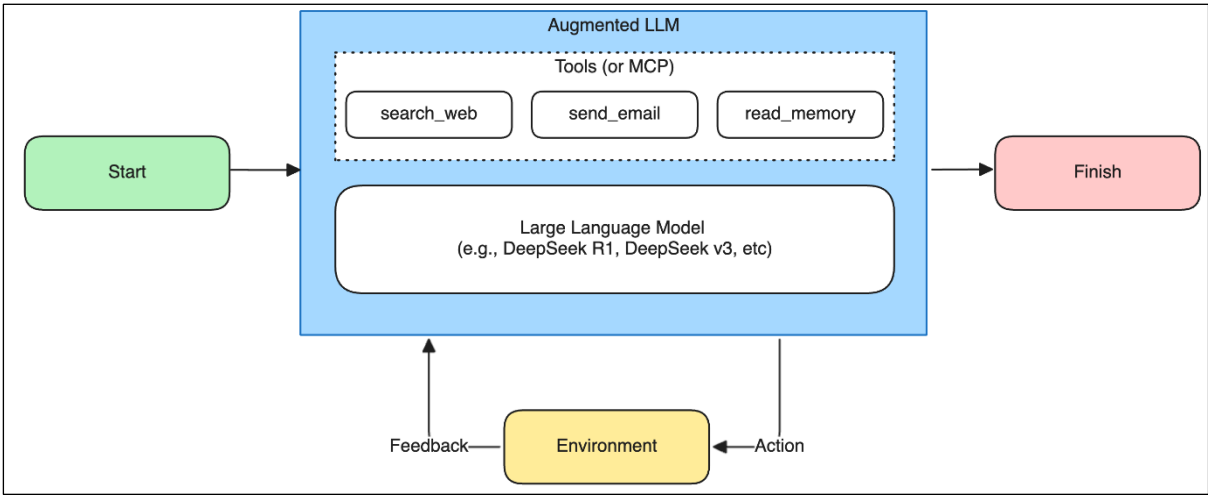
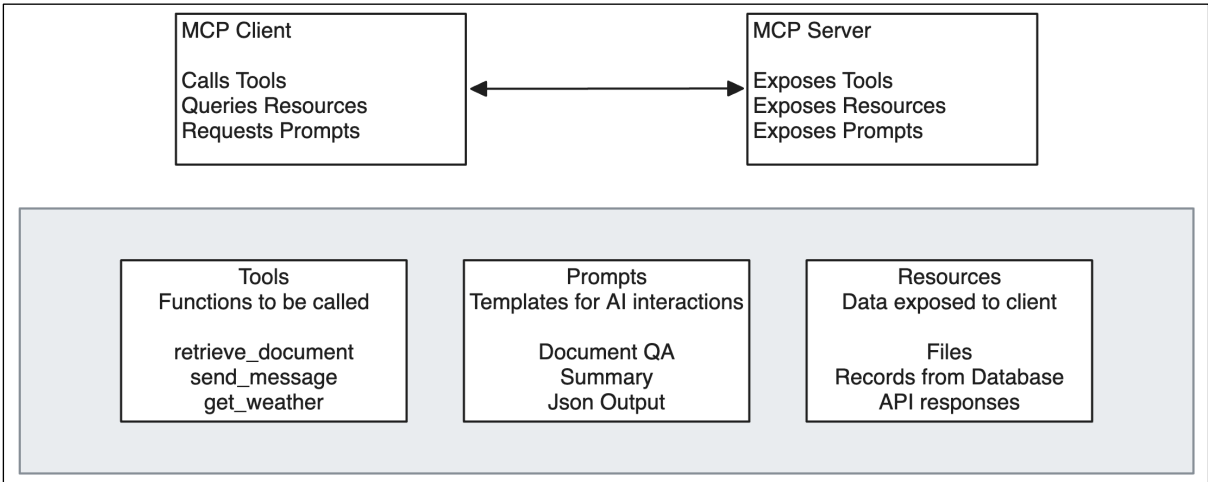
Type
Real-time

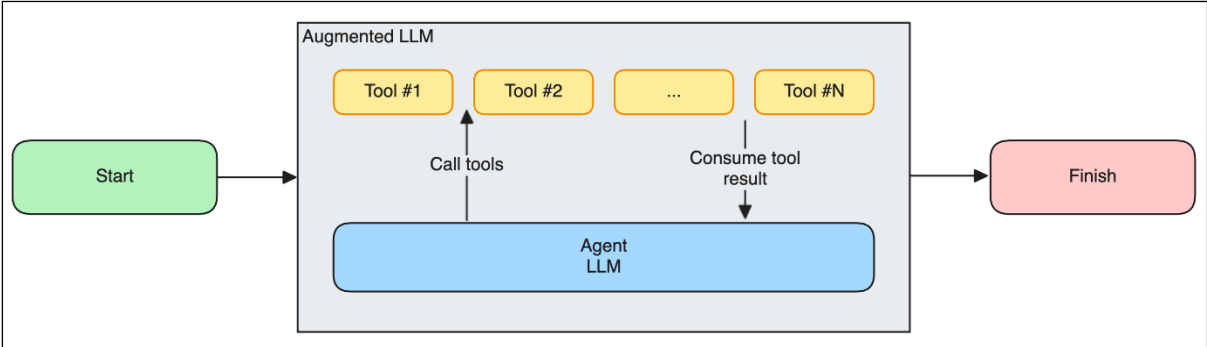
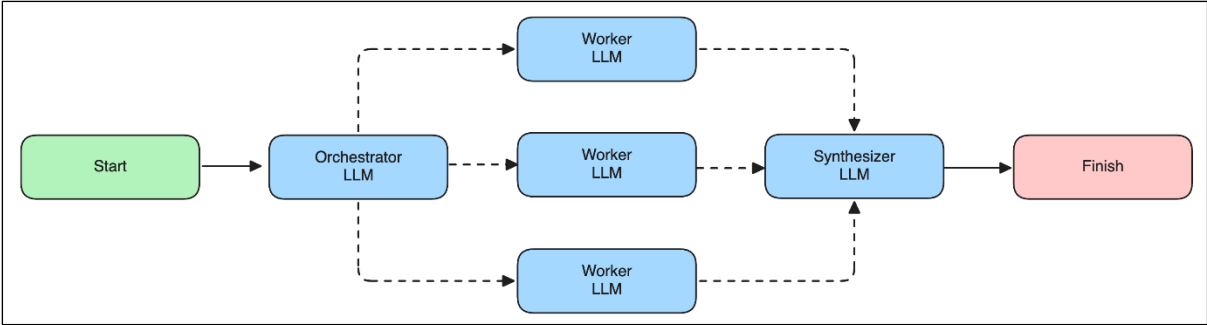
Last updated
Sun Aug 24 2025 10:41:12 GMT+0200 (Central European Summer Time)

Alarms
0 alarms

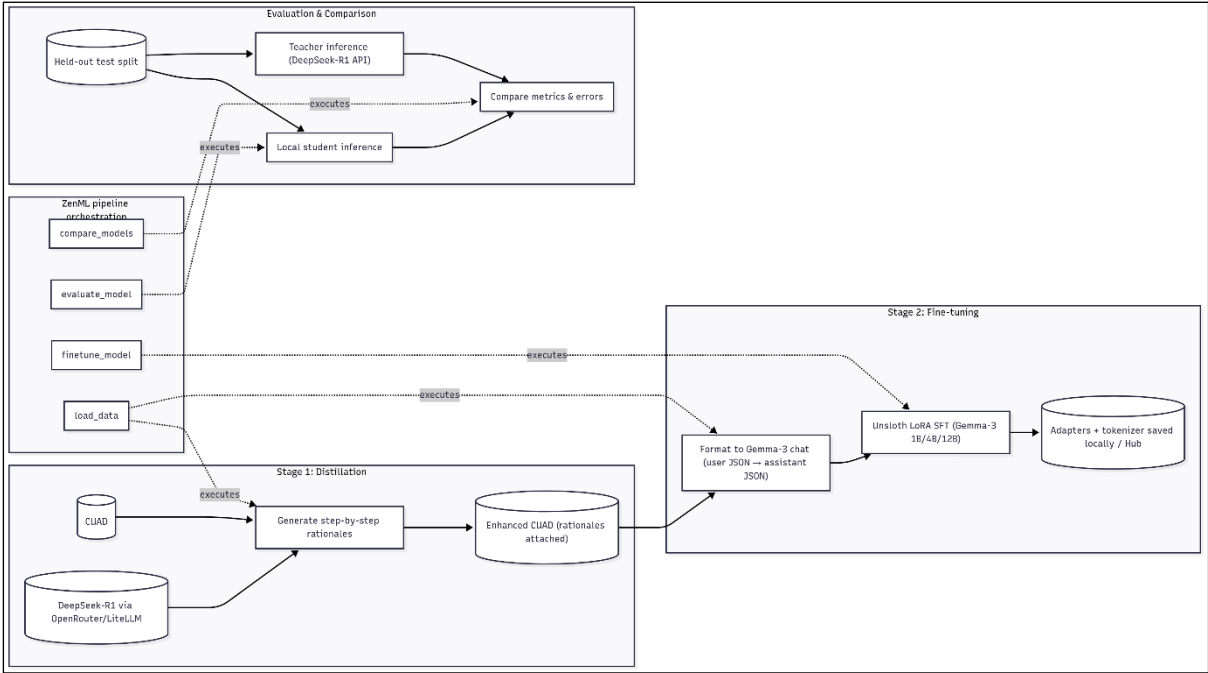
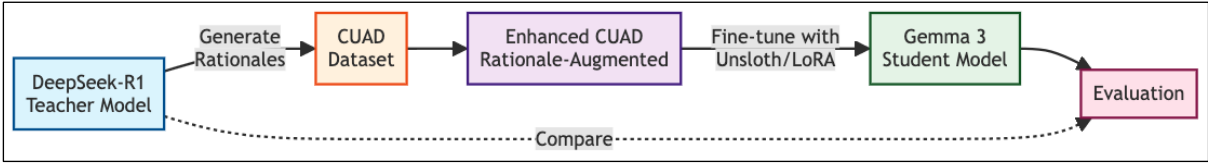


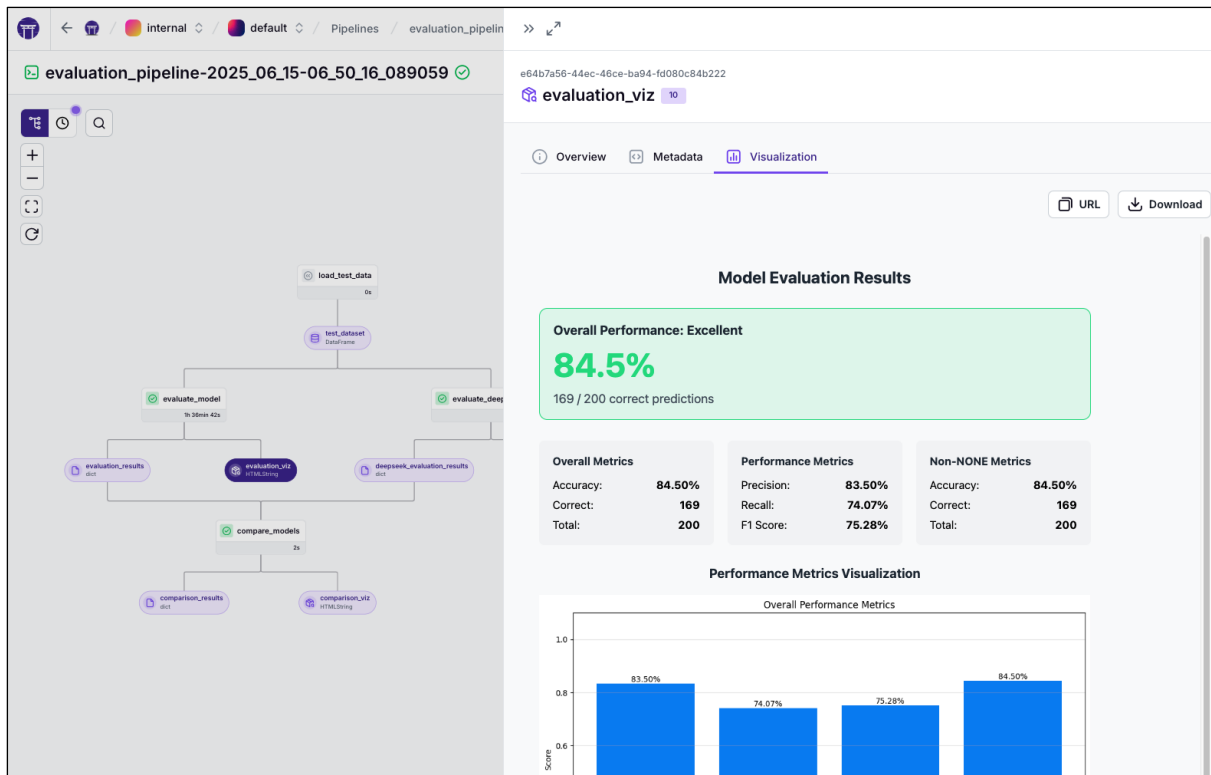
Chapter 6: Agents with DeepSeek



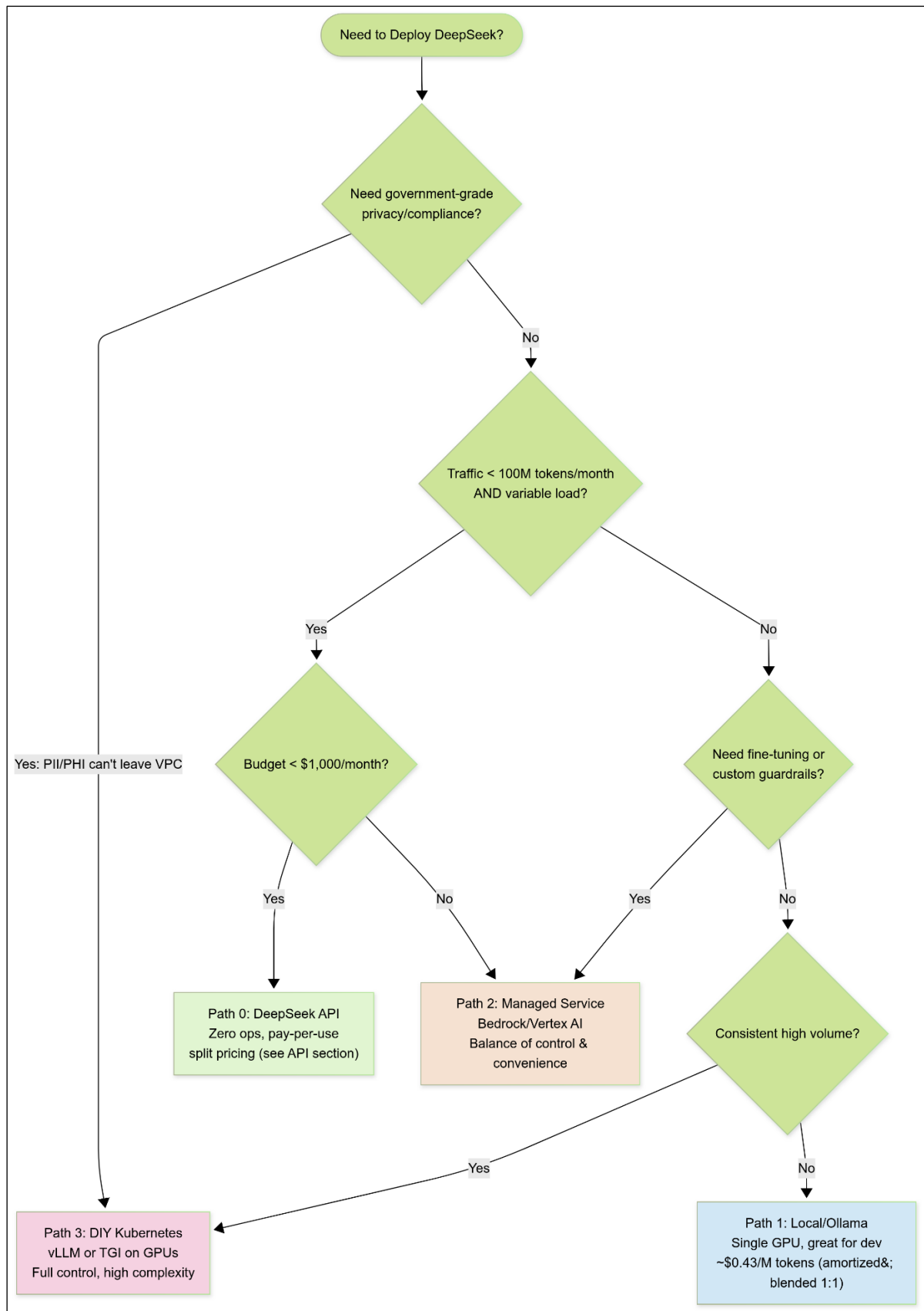


Chapter 7: DeepSeek-Driven Fine-Tuning of Gemma 3 for Legal Reasoning





Chapter 8: Deploying DeepSeek Models



Chapter 10: Appendix

Products > Workers AI > Models

Copy Page

Models

Search models

TASKS

Summarization

Text Embeddings

Text Classification

Text Generation

Object Detection

Text-to-Image

deepseek-r1-distill-qwen-32b

Text Generation • DeepSeek

DeepSeek-R1-Distill-Qwen-32B is a model distilled from DeepSeek-R1 based on Qwen2.5. It outperforms Open...

deepseek-math-7b-instruct Beta

Text Generation • DeepSeek

DeepSeekMath-Instruct 7B is a mathematically instructed tuning model derived from DeepSeekMath-...

Model catalog (211)

Discover Bedrock serverless or Marketplace models that best fit your use case. To get started using a serverless model, request access. For Marketplace models, subscribe and deploy.

Filters

Model collection

Serverless (1)

Bedrock Marketplace (8)

Providers

Amazon (12)

Anthropic (12)

Arcee AI (5)

Autogluon (1)

BRIA AI (3)

Camb.ai (1)

Cohere (7)

DeepSeek (9)

EvolutionaryScale, PBC (1)

Gretel (1)

Show 10 more

Modality

Serverless model access status

Marketplace deployments

Spotlight

Filter for a model

9 matches

Providers = deepseek

Clear filters

Model name	Version	Status	Provider name	Popularity	Deployments	Deployment type
DeepSeek-R1	v1	Available to request	DeepSeek	56	N/A	Serverless
DeepSeek-R1	4.0.5	N/A	DeepSeek	194	0	Bedrock Marketplace
DeepSeek-R1-Distill-Qwen-7B	2.0.5	N/A	DeepSeek	195	0	Bedrock Marketplace
DeepSeek-R1-Distill-Qwen-32B	2.0.5	N/A	DeepSeek	196	0	Bedrock Marketplace
DeepSeek-R1-Distill-Qwen-14B	2.0.5	N/A	DeepSeek	197	0	Bedrock Marketplace
DeepSeek-R1-Distill-Qwen-1.5B	2.0.5	N/A	DeepSeek	198	0	Bedrock Marketplace
DeepSeek-R1-Distill-Llama-8B	2.0.5	N/A	DeepSeek	199	0	Bedrock Marketplace
DeepSeek-R1-Distill-Llama-70B	2.0.5	N/A	DeepSeek	200	0	Bedrock Marketplace
DeepSeek-R1-0528	1.0.0	N/A	DeepSeek	201	0	Bedrock Marketplace

Providers for DeepSeek V3

OpenRouter routes requests to the best providers that are able to handle your prompt size and parameters, with fallbacks to maximize uptime. ⓘ

≡

Sort by

⌵

DeepInfra ⓘ US fp8	Context	Max Output	Input	Output	Latency	Throughput	Uptime	⌵
	164K	164K	\$0.38	\$0.89	0.57 s	49.96 tps		
NovitaAI ⓘ US	Context	Max Output	Input	Output	Latency	Throughput	Uptime	⌵
	64K	16K	\$0.40	\$1.30	1.09 s	25.42 tps		
Nebius AI Studio ⓘ DE fp8	Context	Max Output	Input	Output	Latency	Throughput	Uptime	⌵
	164K	164K	\$0.50	\$1.50	0.37 s	20.90 tps		
Fireworks ⚠ ⓘ US	Context	Max Output	Input	Output	Latency	Throughput	Uptime	⌵
	131K	131K	\$0.90	\$0.90	1.22 s	63.72 tps		

Models

Search settings ⌘F

⚙️ General

∞ Agents

➞ Tab

📦 Models

☁️ Cloud Agents

🔗 Tools & MCP

📄 Rules, Memories, Commands

📚 Indexing & Docs

🌐 Network

➞ Beta

📖 Docs [↗](#)

Add or search model



Composer 1



Sonnet 4.5



GPT-5 Codex



GPT-5



Haiku 4.5



Grok Code



Sonnet 4.5



GPT-5 Codex High



GPT-5 Codex Fast



[View All Models](#)

> API Keys

unsloth/DeepSeek-R1-Distill-Qwen-1.5B-GGUF

like 123 Follow Unsloth AI 10.5k

Text Generation Transformers GGUF English qwen2 deepseek qwen unsloth conversational arXiv:2501.12948 License: apache-2.0

Model card Files and versions xet Community

Train Deploy Use this model

Edit model card

See our collection for versions of Deepseek-R1 including GGUF and original formats.

Instructions to run this model in llama.cpp:

Or you can view more detailed instructions here: unsloth.ai/blog/deepseek-r1

1. Do not forget about <| User |> and <| Assistant |> tokens! - Or use a chat template formatter

2. Obtain the latest llama.cpp at <https://github.com/ggerganov/llama.cpp>

3. Example with Q8_0 K quantized cache **Notice -no-cnv disables auto conversation mode**

```
./llama.cpp/llama-cli \  
--model unsloth/DeepSeek-R1-Distill-Qwen-1.5B-GGUF/DeepSeek-R1-Distill-Qwen-1.5B-GGUF-Q8_0 \  
--cache-type-k q8_0 \  
--threads 16 \  
--prompt '<| User |>What is 1+1?<| Assistant |>' \  
-no-cnv
```

Downloads last month 19,224

GGUF

Model size 2B params Architecture qwen2 Chat template

Hardware compatibility Apple M3 Max (64 GB) x1

1-bit IQ1_S | 700 MB IQ1_M | 716 MB

2-bit IQ2_XXS | 775 MB Q2_K | 753 MB IQ2_M | 792 MB Q2_K_L | 808 MB Q2_K_XL | 886 MB

3-bit IQ3_XXS | 882 MB Q3_K_M | 924 MB Q3_K_XL | 968 MB

4-bit IQ4_XS | 1.1 GB Q4_K_M | 1.12 GB Q4_K_XL | 1.19 GB

5-bit Q5_K_M | 1.29 GB

6-bit Q6_K | 1.46 GB

8-bit Q8_0 | 1.89 GB

16-bit BF16 | 3.56 GB

Conversations

New conversation

Hello from DeepSeek!

llama.cpp

Thought Process

Hello! I'm DeepSeek-R1, an artificial intelligence assistant created by DeepSeek. I'm at your service and would be delighted to assist you with any inquiries or tasks you may have.

Type a message (Shift+Enter to add a new line)

deepseek-r1:1.5b

ollama run deepseek-r1:1.5b

📄

📄 47.6M Downloads

🕒 Updated 1 week ago

DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.

thinking

1.5b

7b

8b

14b

32b

70b

671b

Updated 1 week ago		e0979632db5a · 1.1GB
model	arch qwen2 · parameters 1.78B · quantization Q4_K_M	1.1GB
license	MIT License Copyright (c) 2023 DeepSeek Permission is hereb...	1.1kB
params	{ "stop": ["< begin_of_sentence >", "< end_of_sentence >..."	148B
template	{{- if .System }}{{ .System }}{{ end }} {{- range \$i, \$_ :=...	556B

