Open in app *∧*





Generative Al · Following

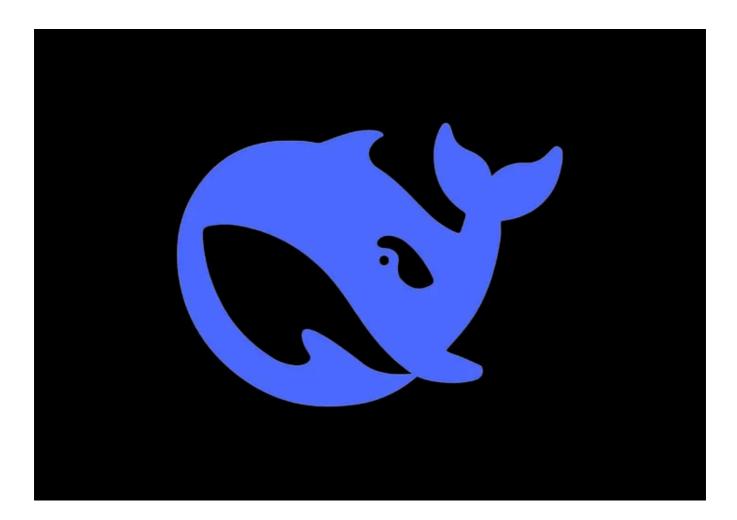
DeepSeek R1

Unlocking the Power of DeepSeek R1 Models: A Comprehensive Guide for Beginners and Experts



Naveen Krishnan Published in Generative Al 8 min read · Feb 3, 2025

Listen	Share	•••	More



Artificial Intelligence (AI) is revolutionizing industries, and the **DeepSeek R1 models** are at the forefront of this transformation. Whether you're a beginner

DeepSeek R1. 🖋 Unlocking the Power of DeepSeek R1... | by Naveen Krishnan | Feb, 2025 | Generative Al

looking to dive into the world of AI or an expert seeking to optimize your workflows, this guide will walk you through everything you need to know about DeepSeek R1 models. We'll cover hosting the model both locally and on Azure using the Model Catalog, connecting via API keys, and even how these models can accelerate AI agent implementations. Let's get started!

Table of Contents

- 1. Introduction to DeepSeek R1 Models
- 2. Why DeepSeek R1 Models?
- 3. Running DeepSeek Model Locally
- 4. Hosting DeepSeek R1 Models on Azure Using Model Catalog
- 5. Connecting to the Model Using API Keys
- 6. Explore More Samples
- 7. DeepSeek R1 Models and AI Agents
- 8. Advanced Use Cases
- 9. Conclusion

1. Introduction to DeepSeek R1 Models 🖮

DeepSeek R1 models are state-of-the-art AI models designed for a wide range of applications, from natural language processing (NLP) to computer vision. These models are pre-trained on vast datasets, making them highly versatile and capable of handling complex tasks with minimal fine-tuning.

Key Features:

- High Accuracy: Trained on diverse datasets, ensuring robust performance.
- Scalability: Can be deployed on cloud platforms like Azure for seamless scaling.

DeepSeek R1. 🖋 Unlocking the Power of DeepSeek R1... | by Naveen Krishnan | Feb, 2025 | Generative AI

• Ease of Integration: Simple API-based integration for quick deployment.

2. Why DeepSeek R1 Models? 🌞

DeepSeek R1 models stand out due to their:

- Versatility: Suitable for various industries, including healthcare, finance, and e-commerce.
- Efficiency: Optimized for fast inference, reducing latency.
- Customizability: Easily fine-tuned to meet specific business needs.

3. Running DeepSeek Model Locally

Step 1: Install Ollama

Visit <u>Ollama's download page</u> and download the installer for your operating system.

Follow the on-screen instructions to complete the installation.

Once installed, open the terminal (or command prompt) and verify the installation by typing:

```
ollama --version
```

You should see the version number if the installation was successful.

Step 2: Download and Set Up DeepSeek-R1

Open the Ollama app or use the command line to search for the DeepSeek-R1 model. There are different versions of DeepSeek-R1. Please see this <u>page</u> for versions. In this example, I am installing the 32b version.

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b	7b	8b	14b	32b	70b	671b
↓, 5.1N	A Pulls	0	Updated	10 days	ago	

7b	✓ Signature > 28 Tags	ollama run deepseek-ri	G
Updated 11 day	s ago	0a8c26	691023 · 4.7GB
model	arch qwen2 · parameters 7.62B · quantization Q4_K_M		4.7GB
params	{ "stop": ["< begin_of_sentence >", "< end_of_sentence >",		148B
template	<pre>{{- if .System }}{{ .System }}{{ end }}</pre>) {{- range \$i, \$_ := .Messages }} {{- \$la	387B
license	MIT License Copyright (c) 2023 DeepSee	Permission is hereby granted, free of ch.	1.1kB

ollama install deepseek-r1:32b

Wait for the model to download and install. This process may take some time depending on your internet speed and system performance.

Verify the installation by running:

ollama list

DeepSeek-R1 should appear in the list of installed models.

Step 3: Run DeepSeek-R1

Start the Ollama runtime by opening a terminal and typing:

ollama start

Once the runtime is active, run DeepSeek-R1 by entering:

ollama run deepseek-r1:32b

C:\>ollama run deepseek-r1:32b "can you instroduce yourself?" <think> Greetings! 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. </think> Greetings! 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.

The model will process the input and return results directly in the terminal or your connected application.

4. Hosting DeepSeek R1 Models on Azure Using Model Catalog 🥏

The Azure Model Catalog is a centralized repository for pre-trained models, including DeepSeek R1. Hosting your DeepSeek R1 model on Azure using the Model Catalog is straightforward. Here's a step-by-step guide:

Step 1: Set Up Azure Account

- Create an Azure account if you don't have one.
- Navigate to the Azure portal and set up a new resource group.

Step 2: Access the Model Catalog

- Go to the Azure Foundry -> Model Catalog in the Azure portal.
- Search for DeepSeek R1 in the catalog.

œ	← DeepSeek-R1 #	REVIEW				
Overview	Deploy					
G Model catalog						
Q1 Playgrounds	Details Existing deployments	Code samples Licens	ie in the second se			
OD Al Services	Learn more: [original model announcement]					
Build and customize	DeepSeek-R1 excels at reasoning tasks using a step-by-step training process, such as language, scientific reasoning, and coding tasks. It features 6718 total parameters with 378 active parameters, and 128k context length. DeepSeek-R1 builds on the progress of earlier reasoning-focused models that improved performance by extending Chain-of-Thought (CoT) reasoning. DeepSeek-R1 takes things					
Fine-tuning Prompt flow Assess and improve	further by combining reinforcement learning (RL) with fine-tuning on carefully chosen datasets. It evolved from an earlier version. DeepSeek-R1-Zero, which relied solely on RL and showed strong reasoning skills but had issues like hard-to-read outputs and language inconsistencies. To address these limitations, DeepSeek-R1 incorporates a small amount of cold- start data and follows a refined training pipeline that blends reasoning-oriented RL with supervised fine-tuning on curated datasets, resulting in a model that achieves state-of-the-art performance on reacting benchmarks.					
tracing PREVIEW	See more					
晉 Evaluation						
Safety + security	Data, media and la	anguages				
My assets						
Models + endpoints	Property	Description				
Data + indexes	Supported data types	Inputs	Outputs			
B Web apps		text	text			
	Supported languages	en, zh				
	Usage guidance					
	Content Filtering					
			t configuration of Azure AI Content Safety classification models to detect and prevent the output of harmful conte tions for content filtering vary when you deploy a model for production in Azure AI: learn more.	nt. Learn		

Screenshot taken by Author

Step 3: Deploy the Model

- Select the DeepSeek R1 model and click on Deploy.
- - Configure the deployment settings, including compute resources and scaling options.

Step 4: Create an Endpoint

- Once the model is deployed, create a new endpoint.
- - Note down the endpoint URL and API key for future use.

5. Connecting to the Model Using API Keys 🔑

Once your model is hosted on Azure, you can connect to it using API keys. Here's how:

Authentication using API Key: For Serverless API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service. In

DeepSeek R1. 🖋 Unlocking the Power of DeepSeek R1... | by Naveen Krishnan | Feb, 2025 | Generative AI this sample endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

Azure inferencing package includes a client for Chat Completion. A client can be authenticated using the API key. The code sample creates and authenticates a synchronous ChatCompletionsClient:

```
from azure.ai.inference import ChatCompletionsClient
from azure.core.credentials import AzureKeyCredential
# For Serverless API or Managed Compute endpoints
client = ChatCompletionsClient(
    endpoint="<ENDPOINT_URL>",
    credential=AzureKeyCredential(key)
)
```

Install dependencies: Install the Azure AI Inference SDK using pip (Requires: Python >=3.8):

```
pip install azure-ai-inference
```

Run a basic code sample: This sample demonstrates a basic call to the chat completion API. The call is synchronous.

```
import os
from azure.ai.inference import ChatCompletionsClient
from azure.ai.inference.models import SystemMessage, UserMessage
from azure.core.credentials import AzureKeyCredential
endpoint = "<ENDPOINT_URL>"
model_name = "DeepSeek-R1"
client = ChatCompletionsClient(
    endpoint=endpoint,
    credential=AzureKeyCredential(key),
)
response = client.complete(
```

```
3/14/25, 7:36 PM DeepSeek R1. & Unlocking the Power of DeepSeek R1...| by Naveen Krishnan | Feb, 2025 | Generative AI
messages=[
    SystemMessage(content="You are a helpful assistant."),
    UserMessage(content="I am going to Paris, what should I see?")
    ],
    max_tokens=2048,
    model=model_name
    )
    print(response.choices[0].message.content)
```

6. Explore More Samples

Run a multi-turn conversation:

This sample demonstrates a multi-turn conversation with the chat completion API. When using the model for a chat application, you'll need to manage the history of that conversation and send the latest messages to the model.

```
import os
from azure.ai.inference import ChatCompletionsClient
from azure.ai.inference.models import AssistantMessage, SystemMessage, UserMess
from azure.core.credentials import AzureKeyCredential
endpoint = "<ENDPOINT_URL>"
model_name = "DeepSeek-R1"
client = ChatCompletionsClient(
    endpoint=endpoint,
    credential=AzureKeyCredential(key),
)
response = client.complete(
    messages=[
        SystemMessage(content="You are a helpful assistant."),
        UserMessage(content="I am going to Paris, what should I see?"),
        AssistantMessage(content="Paris, the capital of France, is known for it
        UserMessage(content="What is so great about #1?")
    ],
    max_tokens=2048,
    model=model_namemodel=model_name
)
print(response.choices[0].message.content)
```

Stream the output:

3/14/25, 7:36 PM

DeepSeek R1. 🖋 Unlocking the Power of DeepSeek R1... | by Naveen Krishnan | Feb, 2025 | Generative AI

For a better user experience, you will want to stream the response of the model so that the first token shows up early and you avoid waiting for long responses.

```
import os
from azure.ai.inference import ChatCompletionsClient
from azure.ai.inference.models import SystemMessage, UserMessage
from azure.core.credentials import AzureKeyCredential
endpoint = "<ENDPOINT_URL>"
model_name = "DeepSeek-R1"
client = ChatCompletionsClient(
    endpoint=endpoint,
    credential=AzureKeyCredential(key),
)
response = client.complete(
    stream=True,
    messages=[
        SystemMessage(content="You are a helpful assistant."),
        UserMessage(content="I am going to Paris, what should I see?")
    ],
    max_tokens=2048,
    model=model_name
)
for update in response:
    if update.choices:
        print(update.choices[0].delta.content or "", end="")
client.close()
```

7. DeepSeek R1 Models and Al Agents 🔄

AI agents are autonomous systems that can perform tasks without human intervention. DeepSeek R1 models can significantly enhance the capabilities of AI agents by providing:

- Faster Inference: Reduced latency for real-time decision-making.
- Improved Accuracy: Higher accuracy in predictions and classifications.
- Scalability: Easily scalable to handle large volumes of data.

Example: AI Agent for Customer Support

DeepSeek R1. 🖋 Unlocking the Power of DeepSeek R1... | by Naveen Krishnan | Feb, 2025 | Generative Al

- Task: Automatically classify customer queries and route them to the appropriate department.
- Implementation: Use DeepSeek R1 models to classify queries in real-time, ensuring quick and accurate routing.

8. Advanced Use Cases 🛠

Use Case 1: Sentiment Analysis

- Description: Analyze customer reviews to determine sentiment.
- Implementation: Use DeepSeek R1 models to classify reviews as positive, negative, or neutral.

Use Case 2: Image Recognition

- Description: Identify objects in images for inventory management.
- Implementation: Deploy DeepSeek R1 models to recognize and classify objects in real-time.

Use Case 3: Fraud Detection

- **Description:** Detect fraudulent transactions in financial data.
- Implementation: Use DeepSeek R1 models to analyze transaction patterns and flag anomalies.

9. Conclusion 🎯

DeepSeek R1 models are powerful tools that can transform your AI workflows. Whether you're a beginner or an expert, this guide provides a comprehensive overview of how to host, connect, and utilize these models effectively. By leveraging **Azure Model Catalog** for hosting and API keys for connectivity, you can seamlessly integrate DeepSeek R1 models into your applications. Additionally, the integration with AI agents opens up new possibilities for automation and efficiency.

So, what are you waiting for? Dive into the world of DeepSeek R1 models and unlock the full potential of AI! 🚀

Further Reading

- [Azure Model Catalog Documentation](https://learn.microsoft.com/enus/azure/machine-learning/concept-model-catalog)
- [DeepSeek R1 on Azure AI Foundry](https://azure.microsoft.com/enus/blog/deepseek-r1-is-now-available-on-azure-ai-foundry-and-github/)

By following this guide, you'll be well-equipped to harness the power of DeepSeek R1 models and take your AI projects to the next level.

Happy coding!

Thank You!

Thanks for taking the time to read my story! If you enjoyed it and found it valuable, please consider giving it a clap (or 50!) to show your support. Your claps help others discover this content and motivate me to keep creating more.

Also, don't forget to follow me for more insights and updates on AI. Your support means a lot and helps me continue sharing valuable content with you. Thank you!



This story is published on <u>Generative AI</u>. Connect with us on <u>LinkedIn</u> and follow <u>Zeniteq</u> to stay in the loop with the latest AI stories.

Subscribe to our <u>newsletter</u> and <u>YouTube</u> channel to stay updated with the latest news and updates on generative AI. Let's shape the future of AI together!



Technology

Data Science



Following

Published in Generative AI

42K Followers · Last published 9 hours ago

All the latest news and updates on the rapidly evolving field of Generative Al space. From cutting-edge research and developments in LLMs, text-to-image generators, to real-world applications, and the impact of generative Al on various industries.



Edit	profile	

Ô

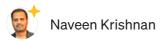
...

Written by Naveen Krishnan

158 Followers · 140 Following

Al Architect @ Microsoft. Passionate about leveraging artificial intelligence to solve real-world problems.

No responses yet



What are your thoughts?

More from Naveen Krishnan and Generative AI