

# 《DeepSeek Starter Guide》

## Volume 2: Development Practice

01 Getting Started for DeepSeek Developers

02 Common Ways and Tutorials for Deploying DeepSeek

03 DeepSeek-Assisted Development Tools and Practical Examples

51CTO

51CTO智能研究院、51CTO传媒、51CTO学堂联合出品

# DeepSeek Empowers Developers

- Core Value and Three Major Pros of DeepSeek
- Five Key Decision Factors for Deploying DeepSeek

# Core Value of DeepSeek

Enterprise: Cost  
Reduction and  
Efficiency  
Enhancement

Hengrui Pharmaceuticals,  
ZEEKR

## 恒瑞医药管理总部文件

恒管总字〔2025〕10号

关于在公司内部全面开展 Deepseek 应用的  
通知

各部门、分公司、子公司:

为了积极顺应人工智能发展新趋势,推动 AI 技术在业务场景中的应用,进一步提升工作效率和管理水平,根据公司董事长指示,决定在全公司范围内开展 Deepseek 应用工作。现将有关事项通知如下:

Source: Social Media

Product:  
Experience  
Upgrade

Huawei Hey Celia, Lark



小艺  
今天发布

36%

- 1、【智能升级】小艺DeepSeek-R1智能体升级为正式版,支持联网搜索,知识库丰富时效性更好,来体验一下吧。
- 2、【体验优化】修复了“放烟花”技能的部分问题,进一步优化特效渲染体验。

版本 11.2.10.340 | 100.3 MB

Source: HarmonyOS Developers Community

Developer:  
Assisted  
Development

Code Generation, Code  
Rewriting



代码改写

对代码进行修改,来实现纠错、注释、调优等。



代码解释

对代码进行解释,来帮助理解代码内容。

Source: DeepSeek Official

# Three Major Pros of DeepSeek

High Performance

Comparable to OpenAI o1

Low Cost

Over 20 times lower than o1's

Developer-Friendly

Compatible with OpenAI's API format

CNY USD

模型 <sup>(1)</sup>	上下文长度	最大思维链长度 <sup>(2)</sup>	最大输出长度 <sup>(3)</sup>	百万tokens 输入价格 (缓存命中) <sup>(4)</sup>	百万tokens 输入价格 (缓存未命中)	百万tokens 输出价格 输出价格
deepseek-chat	64K	-	8K	0.5元	2元	8元
deepseek-reasoner	64K	32K	8K	1元	4元	16元 <sup>(5)</sup>

# Five Key Decision Factors for Deploying DeepSeek

Frequency of Use

Long-term High Frequency

Local, Cloud Hosting

Requirement of Response Latency

Ultra-low Latency

Edge Computing

Computing & Technical Resource

No Ops Team

API, Cloud Hosting

Data Sensitivity Assessment

Sensitive Data

Local, Private Cloud

Compliance Requirements

High Compliance

Local Deployment + Security Audit



# DeepSeek Deployment for Developers

- Three Common Ways for Developers to Deploy DeepSeek
  - Official API Use
  - Local Private Deployment
  - Cloud Hosting Deployment
- Other Ways for Deploying DeepSeek

# Three Common Ways to Deploy DeepSeek for Developers

Three common ways for deploying DeepSeek, with how to get started:

## Official API

**Pros:** Zero Ops cost, pay-as-you-go

**Cons:** Higher network latency

## Local Private Deployment

**Pros:** Data security, highly customizable

**Cons:** Requires GPU resources, manual Ops

## Cloud Hosting Deployment

**Pros:** Elastic scaling, hybrid cloud support

**Cons:** Dependence on Cloud providers & cross-cloud management

## Official API Use: First Call

Before using the DeepSeek API, please create an API key first.

<https://api-docs.deepseek.com/zh-cn/api/deepseek-api>

After creating the API key, you can use the following sample Python script to access the DeepSeek API.

```
# Please install OpenAI SDK first: `pip3 install openai`

from openai import OpenAI

client = OpenAI(api_key="<DeepSeek API Key>", base_url="https://api.deepseek.com")

response = client.chat.completions.create(
    model="deepseek-chat",
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "Hello"},
    ],
    stream=False
)

print(response.choices[0].message.content)
```



# Official API Use: Key Parameters

deepseek-chat

Invoking DeepSeek-V3

temperature

Parameters are set to 1.0 by default, recommended to configure based on scenarios.

deepseek-reasoner

Invoking Deepseek-R1

max\_tokens

Default maximum output length is 4K, adjustable manually to 8K

场景	温度
代码生成/数学解题	0.0
数据抽取/分析	1.0
通用对话	1.3
翻译	1.3
创意类写作/诗歌创作	1.5

Source: DeepSeek Official

# Official API Usage: Common Errors and Solutions

## 400 – Format Error

Modify the request body according to the error message

## 429 – Rate Limit Exceeded

Reasonably plan your request rate

## 401 – Authentication Failed

Check if the API key is correct

## 500 – Server Error

Retry later or contact the official support

## 402 – Insufficient Balance

Confirm your balance and recharge promptly

## 503 – Server Busy

Retry later

## 422 – Parameter Error

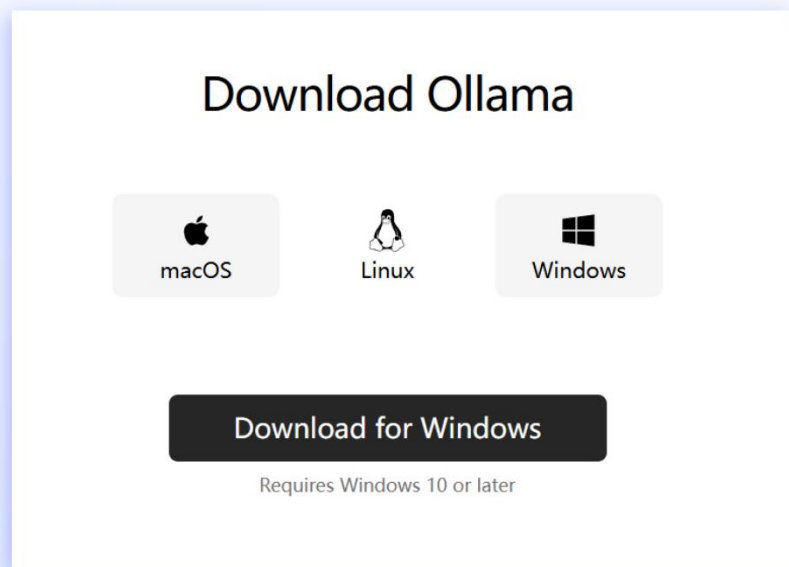
Modify relevant parameters based on the error message

## Local Private Deployment (1)

Locally deploy the distilled version of DeepSeek-R1, simplifying the deployment process through tools like Ollama, vLLM, etc.

The following tutorial uses Ollama as an example. Ollama download link: <https://ollama.com/download>

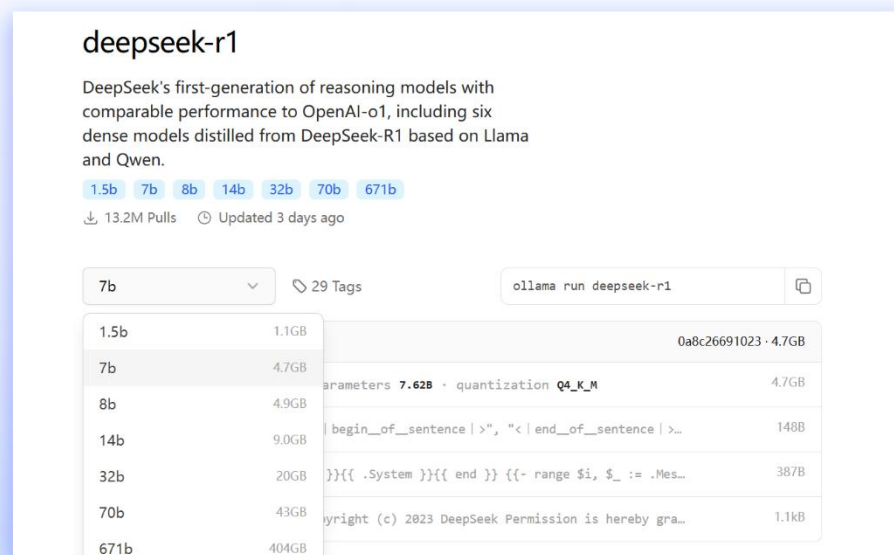
### 1. Download Ollama



Source: Ollama Official

### 2. Open the model list and download DeepSeek R1

Visit <https://ollama.com/library/deepseek-r1> to select a model with the appropriate amount of parameters



## Local Private Deployment (2)

### 3. Setup the downloaded model

Open your local command prompt and enter the following command (taking the 7B model as an example)

```
ollama pull deepseek-r1:7b
```

Use the `ollama list` command to view the storage space occupied by all local models

```
C:\Users\EDY>ollama list
NAME                ID                SIZE    MODIFIED
deepseek-r1:7b      0a8c26691023     4.7 GB  3 days ago
```

#### References:

1. <https://mp.weixin.qq.com/s/iKfE67M42ROWVCMTe6wbFA>
2. <https://mp.weixin.qq.com/s/GQ7tjmJsLr2l1dd0raD-wKg>

### 4. Enter the run command to initiate a dialogue with R1

```
C:\Users\EDY>ollama run deepseek-r1:7b
```

Output:

```
>>> 证明1+2大于2
<think>
嗯，证明“1加2大于2”，听起来有点奇怪，因为1+2等于3，而3当然大于2啊！不过可能题目表达有点问题，或者是想让我用某种特殊的方式来看待这个问题？

首先，1加2等于3。3确实比2大。这是基本的数学事实，没有争议。

也许，用户想问的是“为什么1加2大于2”，或者是不是在某种不同的情况下？比如，在某些特定的定义下，或者说是在某种特殊的系统中，“大于”可能有不同的意义？

另外，也有可能是“证明1+2>2”，这是一种不等式。让我尝试从数学的角度来分析一下。

首先，我们可以写出来：1 + 2 = 3

然后比较3和2的大小：

因为3 > 2，

所以，1 + 2 > 2

这是不是太直接了呢？可能需要更深入地解释一下为什么“大于”是这样的关系，或者有没有什么隐藏的条件？

比如，在数学中，“大于”是一个基本的概念，表示左边的数在数值上更大。在这个例子里，3确实比2大，所以1加2大于2。

也许还有其他角度可以探讨这个问题？比如说，在不同的数域中，或者结合某些运算规则的情况下，结果会不会有所不同呢？
```

## Cloud Hosting Deployment (1)

Currently, mainstream domestic cloud platforms all support the DeepSeek model, such as SiliconFlow x Huawei Cloud, Tencent Cloud, Baidu AI Cloud, etc.

The following tutorial uses SiliconFlow as an example. Address:

<https://cloud.siliconflow.cn/i/adxzw8w3af>

### 1. Register and log in to the cloud platform



欢迎登录 SiliconFlow

+86 ▾ 您的手机号

短信验证码 获取验证码

dcAlgVea

注册 / 登录

邮箱登录

Source: SiliconFlow Official

### 2. Create a new API key

Navigate to API Keys in the left sidebar and click Create



API 密钥

API密钥 是您访问 SiliconCloud 接口的凭证，且有该账户的完整权限，请您妥善保管。

新建密钥

密钥描述

请输入描述信息

关于密钥用途等的补充说明

取消 新建密钥

密钥(点击复制)

暂无数据

## Cloud Hosting Deployment (2)

### 3. Download AI development tools

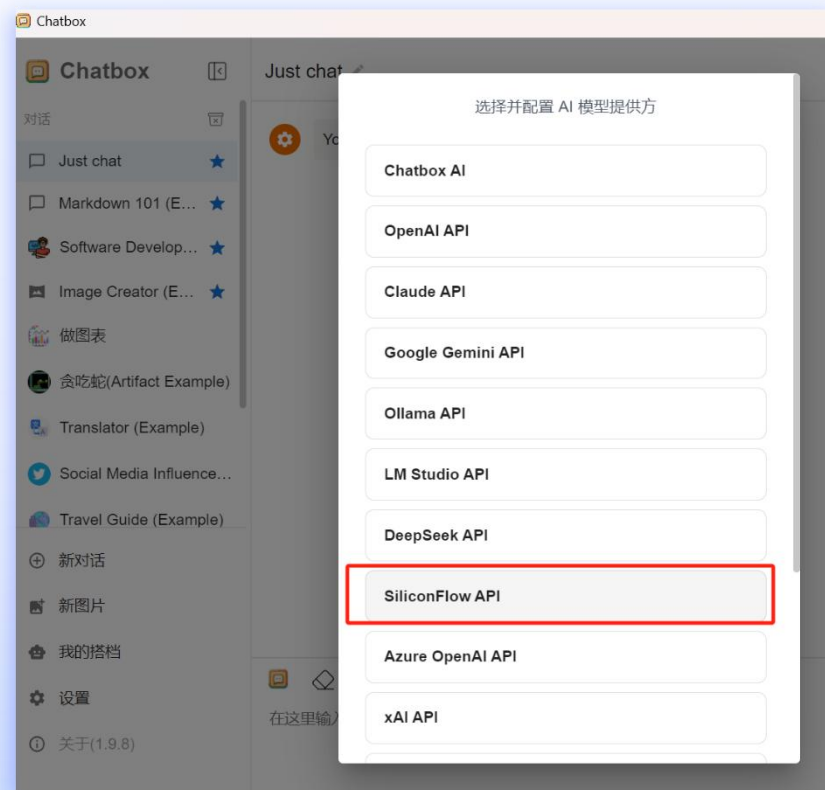
Download Cherry Studio or Chatbox AI  
The tutorial uses Chatbox AI as an example. Address:  
<https://chatboxai.app/>



Source: Chatbox AI

### 4. Install the development tool and select the model provider

Select the API you are using. SiliconFlow is highlighted in the red box below.





## Cloud Hosting Deployment (3)

## 5. Enter your API key and select the model

设置

模型 显示 对话 其他

模型提供方:

SILICONFLOW API

API 密钥

模型

deepseek-ai/DeepSeek-R1

上下文的消息数量上限

20

严谨与想象(Temperature)

0.7

严谨细致 想象发散

取消 保存

## 6. Initiate a dialogue with R1

开发者如何学习deepseek, 请帮我定制一个10天的学习计划

Thinking ↑

嗯, 用户想要一个10天的学习计划来学习DeepSeek。首先, 我需要确定DeepSeek具体指的是什么, 可能是一个框架、工具还是平台? 根据之前的模型信息, DeepSeek-R1是一个AI模型, 可能用户指的是如何学习使用或开发相关的技术。

首先, 用户的身份可能是开发者, 想学习如何应用或集成DeepSeek到自己的项目中。用户的需求是定制一个10天的计划, 所以需要结构化的内容, 分阶段进行。可能需要覆盖基础知识、API使用、实际项目集成、优化等。

用户可能没有明确说明他们的现有水平, 但作为开发者, 可能具备编程基础, 尤其是Python。需要从基础开始, 逐步深入。同时, 用户可能希望有实践环节, 每天安排实操任务, 比如安装、调用API、构建应用等。

另外, 用户可能希望计划不仅仅是理论, 还包括实际应用和项目, 这样他们能够看到成果, 保持学习的动力。需要确保每天的学习目标明确, 时间安排合理, 不会过于紧张但又有挑战性。

还需要考虑资源获取, 比如文档、GitHub仓库、社区支持等。可能需要在计划中推荐相关资源, 帮助用户自主学习。此外, 错误处理和性能优化也是开发者关心的点, 应该包含在后期阶段。

最后, 总结和项目复盘也很重要, 帮助用户巩固所学, 并规划下一步的学习路径。需要确保整个计划循序渐进, 覆盖必要的知识点, 同时保持实用性。

Source: Chatbox AI

## Other Ways to Deploy DeepSeek

Three advanced methods for deploying DeepSeek, briefly introduced as follows:

### Edge Deployment

Deploy lightweight models on edge nodes close to data sources

**Pros:** Ultra-low latency, operational during network outages

**Cons:** Models need distillation, edge hardware compatibility

### Containerized Microservices Deployment

Build a microservices cluster for models using Docker/Kubernetes

**Pros:** Resource isolation, gray release

**Cons:** Requires K8s Ops skills

### Hybrid Architecture Deployment

Combine multiple deployment ways

**Pros:** Balances security and cost, more flexible

**Cons:** Higher complexity in architecture design

# DeepSeek-Assisted Development

- Basic assisted development functions
  - Code rewriting
  - Code explanation
  - Code generation
  - Practical integration tools

# Assisted Development Example 1: Code Rewriting

Modify code to correct errors, add comments, debug, etc.

## Prompts

The following code is inefficient and does not handle edge cases. Please explain the issues and solutions, then optimize it:

```
'''
def fib(n):
    if n <= 2:
        return n
    return fib(n-1) + fib(n-2)
'''
```

This and following images are sourced from DeepSeek Official

## Sample Output

### 问题分析

这段代码实现了一个递归的斐波那契数列计算函数 `fib(n)`。虽然代码简洁，但它存在以下几个问题：

- \*\*重复计算\*\***：递归调用会导致大量的重复计算。例如，计算 `fib(5)` 时，`fib(3)` 和 `fib(2)` 会被多次计算，导致效率低下。
- \*\*边界情况处理不完善\*\***：虽然代码中处理了 `n <= 2` 的情况，但没有处理 `n < 0` 的情况。对于负数输入，函数会返回不正确的结果。
- \*\*递归深度限制\*\***：对于较大的 `n`，递归深度可能会超过 Python 的默认递归深度限制，导致 `RecursionError`。

### 解决方法

- \*\*使用记忆化 (Memoization)\*\***：通过缓存已经计算过的斐波那契数列值，避免重复计算，提高效率。
- \*\*处理边界情况\*\***：增加对 `n < 0` 的检查，并返回适当的错误信息或默认值。
- \*\*使用迭代代替递归\*\***：递归虽然简洁，但在处理大规模数据时效率较低。可以使用迭代方法来计算斐波那契数列，避免递归深度限制。

### 优化后的代码

#### 方法一：记忆化递归

```
```python
def fib(n, memo={}):
    if n < 0:
        raise ValueError("Input must be a non-negative integer")
    if n <= 2:
        return n
    if n in memo:
        return memo[n]
    memo[n] = fib(n-1, memo) + fib(n-2, memo)
    return memo[n]
```
```

## Assisted Development Example 2: Code Explanation

Explain the code for better understanding.

### Prompts

Please explain the logic of the following code and describe its functionality:

```
...  
// weight数组的大小 就是物品个数  
for(int i = 1; i < weight.size(); i++) { // 遍历物品  
    for(int j = 0; j <= bagweight; j++) { // 遍历背包容量  
        if (j < weight[i]) dp[i][j] = dp[i - 1][j];  
        else dp[i][j] = max(dp[i - 1][j], dp[i - 1][j - weight[i]] + value[i]);  
    }  
}  
...
```

### Sample Outputs

这段代码实现了一个经典的动态规划问题，通常用于解决“0/1背包问题”。下面是对代码逻辑的详细解释：

#### ### 代码逻辑解释

##### 1. \*\*初始化与遍历\*\*:

- `weight` 数组表示每个物品的重量。
- `value` 数组表示每个物品的价值。
- `bagweight` 表示背包的最大容量。
- `dp` 是一个二维数组，`dp[i][j]` 表示前 `i` 个物品在背包容量为 `j` 时的最大价值。

##### 2. \*\*外层循环\*\*:

```
```cpp  
for(int i = 1; i < weight.size(); i++) { // 遍历物品  
    ...  
    - 外层循环从第1个物品开始遍历到最后一个物品 (`weight.size()`).
```

##### 3. \*\*内层循环\*\*:

```
```cpp  
for(int j = 0; j <= bagweight; j++) { // 遍历背包容量  
    ...  
    - 内层循环从背包容量为0开始遍历到背包的最大容量 (`bagweight`).
```

## Assisted Development Example 3: Code Generation (1)

To generate code that fulfills a specific function.

### Prompts

Please generate code in xx programming language to achieve the purpose of xx:

Example: Please help me create a Gomoku game in HTML, with all code contained in a single HTML file.

### Sample Outputs

当然! 下面是一个简单的五子棋游戏的 HTML 代码。这个代码包含了一个基本的五子棋棋盘, 玩家可以通过点击棋盘上的格子来下棋。游戏会检测是否有玩家获胜。

```
```html
<!DOCTYPE html>
<html lang="zh-CN">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <title>五子棋游戏</title>
</head>
<body>
  <div id="board">
    <table border="1">
      <tr><td></td><td></td><td></td><td></td><td></td></tr>
      <tr><td></td><td></td><td></td><td></td><td></td></tr>
      <tr><td></td><td></td><td></td><td></td><td></td></tr>
      <tr><td></td><td></td><td></td><td></td><td></td></tr>
      <tr><td></td><td></td><td></td><td></td><td></td></tr>
    </table>
  </div>
</body>
</html>
```
```



## Assisted Development Example 3: Code Generation (2)

Run the generated code:

```
html
复制

boardData[y][x] = currentPlayer;
drawPiece(x, y, currentPlayer);

if (checkWin(x, y)) {
  setTimeout(() => {
    alert(`${currentPlayer === 1 ? '黑棋' : '白棋'}获胜!`);
    boardData = Array(BOARD_SIZE).fill().map(() => Array(BOARD_SIZE).fill
(0));
    initBoard();
  }, 10);
}

currentPlayer = currentPlayer === 1 ? 2 : 1;
status.textContent = `当前玩家: ${currentPlayer === 1 ? '黑棋' : '白棋'}`;
});

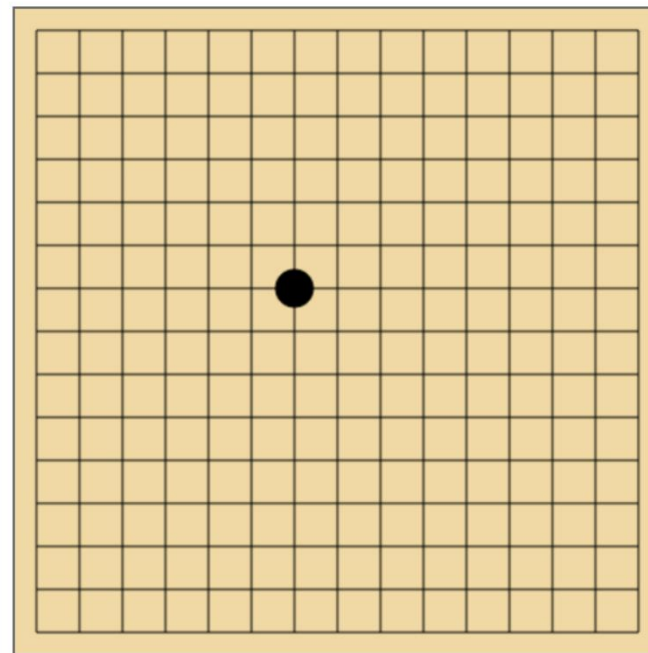
// 初始化游戏
initBoard();
</script>
</body>
</html>
```

▶ 运行 HTML

### Gomoku Game Showcase

DeepSeek

当前玩家: 白棋



# Assisted Development Example 4: Practical Integration Tools

The following frameworks have integrated the DeepSeek API, helping simplify the development process and enhance efficiency.

## AI Agent Frameworks

	<a href="#">Anda</a>	一个专为 AI 智能体开发设计的 Rust 语言框架，致力于构建高度可组合、自主运行且具备永久记忆能力的 AI 智能体网络。
	<a href="#">YoMo</a>	Stateful Serverless LLM Function Calling Framework with Strongly-typed Language Support
	<a href="#">Alice</a>	一个基于 ICP 的自主 AI 代理，利用 DeepSeek 等大型语言模型进行链上决策。Alice 结合实时数据分析和独特的个性，管理代币、挖掘 BOB 并参与生态系统治理。

## RAG Frameworks

	<a href="#">RAGFlow</a>	一款基于深度文档理解构建的开源 RAG（Retrieval-Augmented Generation）引擎。RAGFlow 可以为各种规模的企业及个人提供一套精简的 RAG 工作流程，结合大语言模型（LLM）针对用户各类不同的复杂格式数据提供可靠的问答以及有理有据的引用。
--	-------------------------	--

## Solana Frameworks

	<a href="#">Solana Agent Kit</a>	一个用于连接 AI 智能体到 Solana 协议的开源工具包。现在，任何使用 Deepseek LLM 的智能体都可以自主执行 60+ 种 Solana 操作：
--	----------------------------------	--

## Appendix: Applications Integrated with DeepSeek (Partial)

Source: DeepSeek Official

	<a href="#">Chatbox</a>	一个支持多种流行LLM模型的桌面客户端，可在 Windows、Mac 和 Linux 上使用
	<a href="#">ChatGPT-Next-Web</a>	一键获取跨平台ChatGPT网页用户界面，支持流行的LLM
	<a href="#">留白记事</a>	留白让你直接在微信上使用 DeepSeek 管理你的笔记、任务、日程和待办清单！
	<a href="#">Pal - AI Chat Client</a> (iOS, iPadOS)	一款可以在 iPhone 或 iPad 上使用的 AI 助手
	<a href="#">LibreChat</a>	LibreChat 是一个可定制的开源应用程序，无缝集成了 DeepSeek，以增强人工智能交互体验
	<a href="#">PapersGPT</a>	PapersGPT是一款集成了DeepSeek及其他多种AI模型的辅助论文阅读的Zotero插件.
	<a href="#">RSS翻译器</a>	开源、简洁、可自部署的RSS翻译器
	<a href="#">Enconvo</a>	Enconvo是AI时代的启动器,是所有AI功能的入口,也是一位体贴的智能助理.
	<a href="#">Cherry Studio</a>	一款为创造者而生的桌面版 AI 助手
	<a href="#">ToMemo (iOS, iPadOS)</a>	一款短语合集 + 剪切板历史 + 键盘输出的iOS应用，集成了AI大模型，可以在键盘中快速输出使用。
	<a href="#">Video Subtitle Master</a>	批量为视频生成字幕，并可将字幕翻译成其它语言。这是一个客户端工具, 跨平台支持 mac 和 windows 系统, 支持百度, 火山, deeplx, openai, deepseek, ollama 等多个翻译服务

# More Resources for Learning About DeepSeek and AI

## DeepSeek Official

Visit the official website to delve into the latest research findings, access source code, and official prompt example libraries:  
<https://www.deepseek.com>

## 51CTO AI Zone

Get the latest news, practical articles, useful resources, and sharing by AI experts on Deepseek and AI in the 51CTO AI.x zone:  
<https://ai.51cto.com>

## 51CTO Online Courses – DeepSeek Zone

Over 200 courses with over 1,000 hours of AI (including Deepseek) themed videos for individual and corporate learning:  
<https://b.51cto.com>

## 51CTO DeepSeek 体系化学习方案



三大主题模块, 满足不同  
岗位学习需求



技术模块体系设计分层  
进阶式学习

办公效率  
提升

《AI大模型发展地图&应用实战》

《AI构建企业新质生产力:  
DeepSeek × COZE实战课》

《Deepseek赋能办公全攻略》

《Deepseek AI赋能公文写作》

技术原理  
架构

基础:《DeepSeek大模型全栈  
开发》

进阶:《DeepSeek大模型技术  
原理与实战》

高级:《DeepSeek高级技术  
实战课程》

综合:《DeepSeek企业级实战》

综合管理  
应用

《人工智能原理、前沿介绍和应用  
分析-以 DeepSeek 为例》

《AI商业洞察:DeepSeek赋能,  
让你的决策领先一步》

《Deepseek AI大模型在营销  
领域的应用》

DeepSeek systematic learning solution launched by 51CTO Enterprise Academy



# 51CTO

帮助一亿数字化人才终身成长！



扫码关注51CTO，获取更多资料