



DeepSeek 自学手册

从理论（模型训练）到实践（模型应用）





目录

Contents

01

DeepSeek V3和R1
是怎么训练出来的

02

DeepSeek R1后
提示词的变与不变

03

DeepSeek R1
四大使用技巧

04

13个DeepSeek
官方提示词样例

05

DeepSeek
实际应用场景

06

DeepSeek替代方案
(在线&本地部署)



01

**DeepSeek V3 和 R1
是怎么训练出来的**



初识 DeepSeek V3

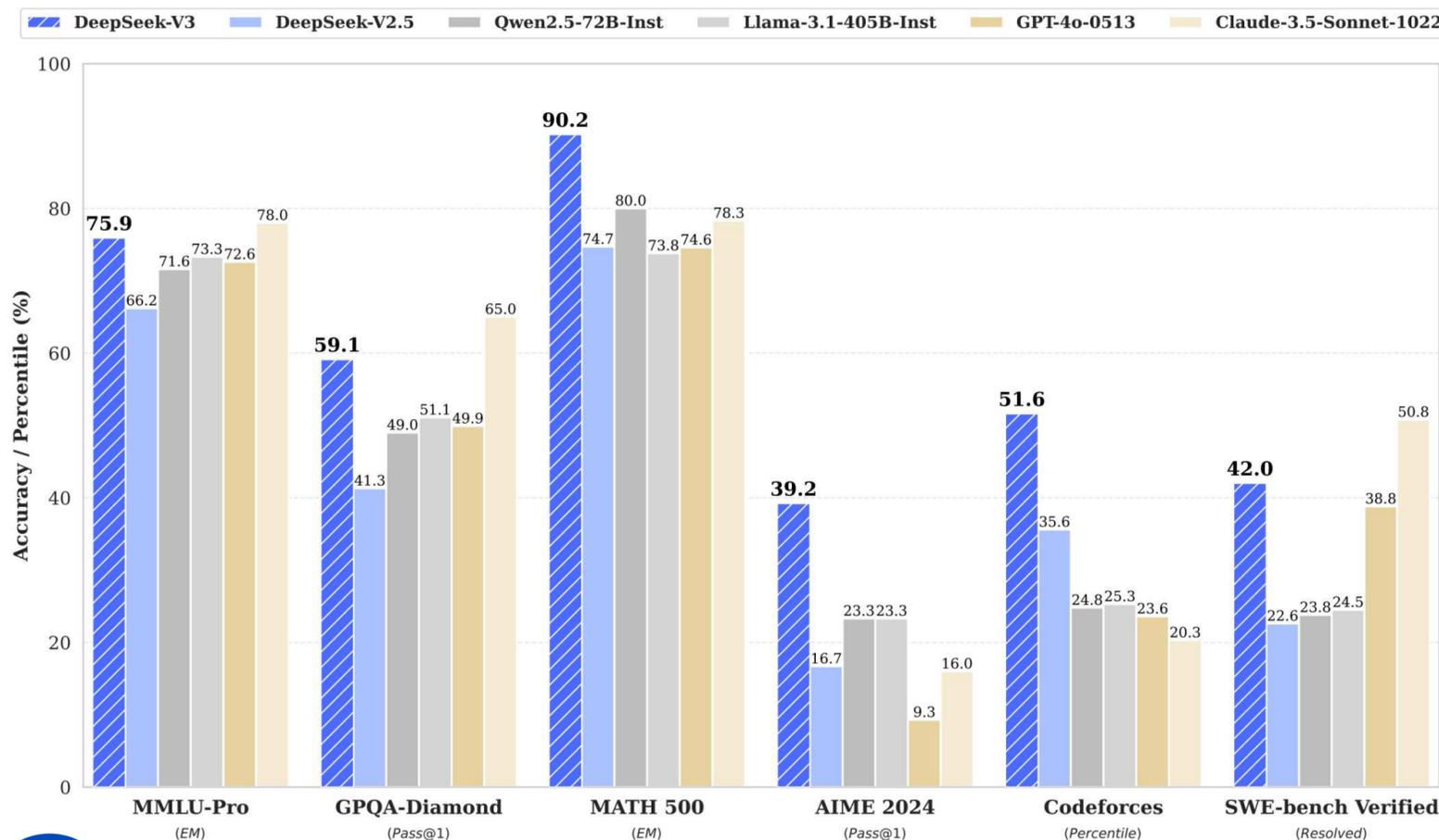


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

DeepSeek V3 是什么？

DeepSeek V3 是一个强大的MoE语言模型（非推理型模型）。它在数学、代码等任务上吊打其它开源模型，甚至能和闭源大佬 GPT-4o、Claude-3.5-sonnet 掰手腕，但训练花的钱还不到 600 万美元。

注：MoE，全称是 Mixture-of-Experts，翻译成中文就是“混合专家”。你可以把它想象成一个团队，这个团队里有很多专家，每个专家都有自己的专长。当遇到一个问题时，不是所有专家都一起上，而是根据问题的类型，选择最合适的几个专家来解决。这样不仅提高了效率，还提升了性能。



DeepSeek V3 架构

Multi-Head Latent Attention (MLA)

V3基础架构之一，在DeepSeek V2已经验证有效。

MLA翻译成中文就是多头潜在注意力，传统模型推理时需缓存大量数据，MLA通过压缩键值对（类似“精简笔记”）减少内存占用，可以让模型在处理长文本时更高效。

DeepSeekMoE

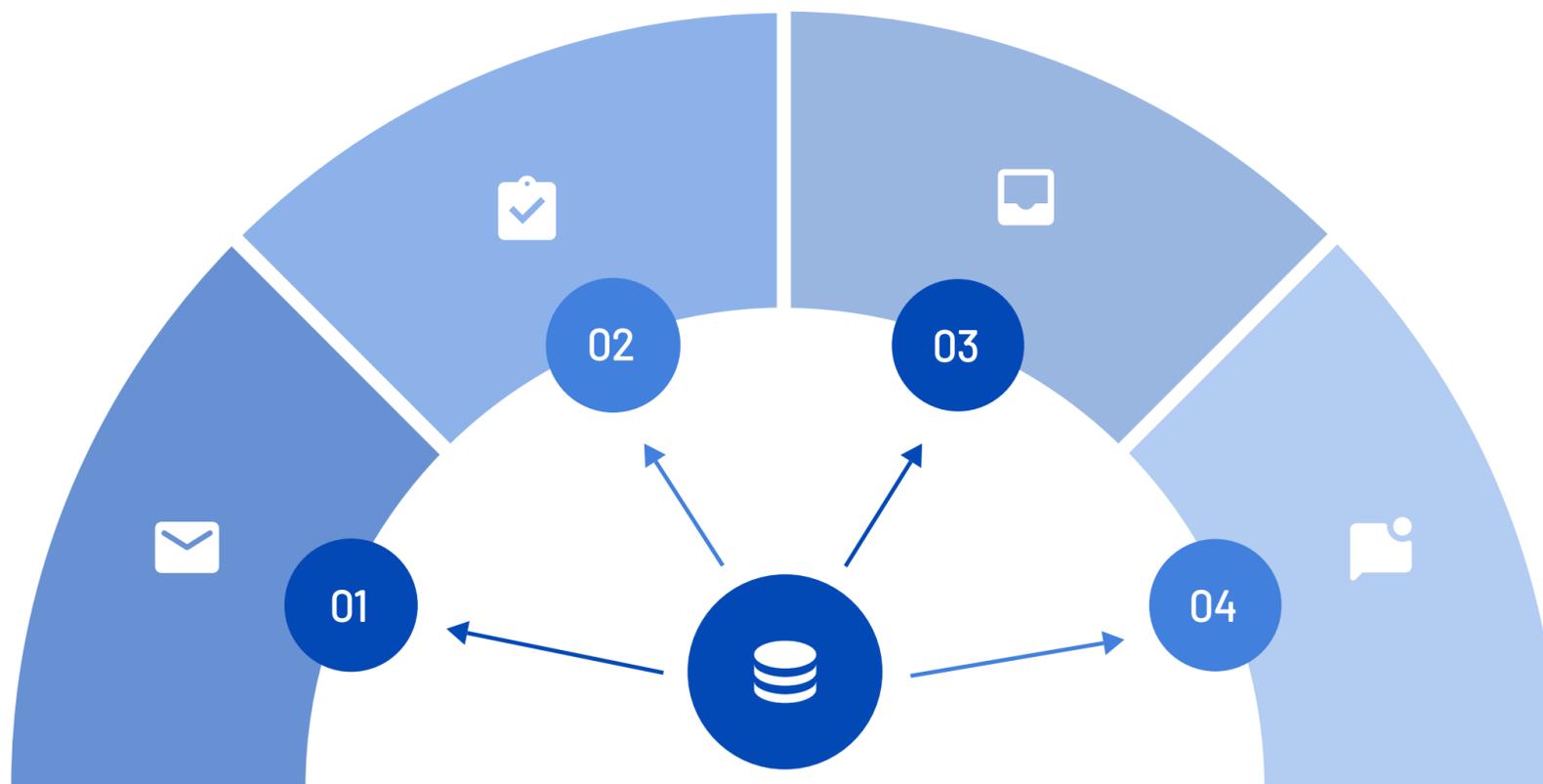
V3基础架构之一，同样在DeepSeek V2已经验证有效。

DeepSeekMoE 在传统 MoE 的基础上进行了多项创新和优化，比如更精细的专家分工、共享专家等，适配复杂任务，进而提升模型的性能和计算效率。

多 Token 预测 (MTP) 策略

V3架构创新之一。

传统模型一次预测1个词，DeepSeek V3同时预测多个词（类似“预判下一步”），提升数据利用效率，还能加速生成



无额外损耗的负载均衡

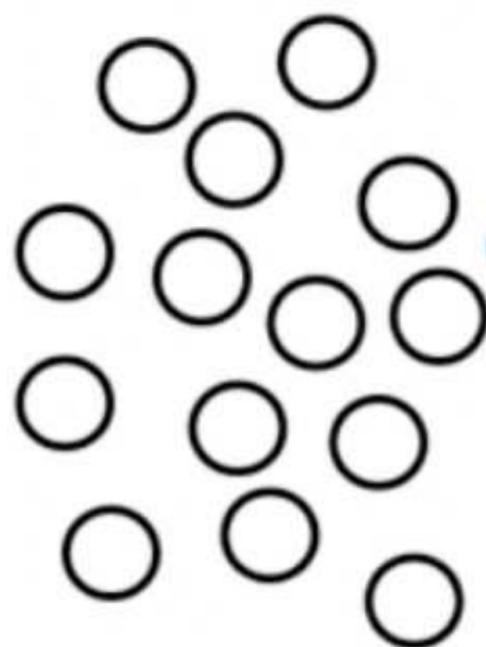
前面DeepSeekMoE中提到的多项创新，其中就包括V3才推出的无额外损耗的负载均衡。

简单来说，在训练过程中，DeepSeekMoE会监控每个专家的“工作量”，进而动态调整专家使用频率，避免某些专家“过劳”或“躺平”。



🤝 DeepSeek V3 训练步骤和数据使用

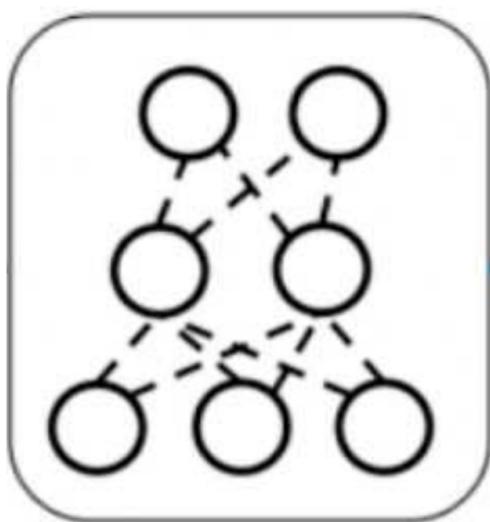
无标注数据集



训练数据: 相比V2, V3提高了数学和编程样本的比例, 同时扩展了除英语和中文之外的多语言覆盖范围

预训练
(无监督学习)

预训练大模型

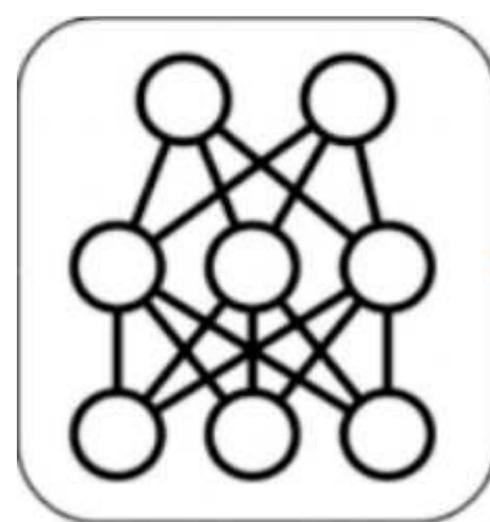


精调
(监督学习)

推理数据: 对于数学、代码、逻辑推理等需要复杂推理过程的任务, V3采用了基于 R1 生成的高质量推理数据

非推理数据: 对于创意写作、角色扮演、简单问答等非推理类任务, 则用 V2.5 生成响应, 并由人工标注和校验

基础大模型



精调
(强化学习)

基于规则的奖励模型: 对于可以通过明确规则进行判别任务(例如数学题、编程题), 采用基于规则的奖励模型

基于模型的奖励模型: 对于难以通过规则进行判别任务(例如开放式问答、创意写作), 则采用基于模型的奖励模型

通用精调模型





DeepSeek V3 性能表现

知识基准

通用能力: MMLU (综合知识)、GPQA (博士级难题) 等测试中, 超越所有开源模型, 接近GPT-4o

事实性: 中文事实问答 (C-SimpleQA) 甚至超过闭源模型, 英文略逊于GPT-4o

Benchmark (Metric)	DeepSeek V2-0506	DeepSeek V2.5-0905	Qwen2.5 72B-Inst.	LLaMA-3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
Architecture	MoE	MoE	Dense	Dense	-	-	MoE
# Activated Params	21B	21B	72B	405B	-	-	37B
# Total Params	236B	236B	72B	405B	-	-	671B
English							
MMLU (EM)	78.2	80.6	85.3	88.6	88.3	87.2	88.5
MMLU-Redux (EM)	77.9	80.3	85.6	86.2	88.9	88.0	89.1
MMLU-Pro (EM)	58.5	66.2	71.6	73.3	78.0	72.6	75.9
DROP (3-shot F1)	83.0	87.8	76.7	88.7	88.3	83.7	91.6
IF-Eval (Prompt Strict)	57.7	80.6	84.1	86.0	86.5	84.3	86.1
GPQA-Diamond (Pass@1)	35.3	41.3	49.0	51.1	65.0	49.9	59.1
SimpleQA (Correct)	9.0	10.2	9.1	17.1	28.4	38.2	24.9
FRAMES (Acc.)	66.9	65.4	69.8	70.0	72.5	80.5	73.3
LongBench v2 (Acc.)	31.6	35.4	39.4	36.1	41.0	48.1	48.7
Code							
HumanEval (Pass@1)	69.3	77.4	77.3	77.2	81.7	80.5	82.6
LiveCodeBench (Pass@1-COT)	18.8	29.2	31.1	28.4	36.3	33.4	40.5
LiveCodeBench (Pass@1)	20.3	28.4	28.7	30.1	32.8	34.2	37.6
Codeforces (Percentile)	17.5	35.6	24.8	25.3	20.3	23.6	51.6
SWE Verified (Resolved)	-	22.6	23.8	24.5	50.8	38.8	42.0
Aider-Edit (Acc.)	60.3	71.6	65.4	63.9	84.2	72.9	79.7
Aider-Polyglot (Acc.)	-	18.2	7.6	5.8	45.3	16.0	49.6
Math							
AIME 2024 (Pass@1)	4.6	16.7	23.3	23.3	16.0	9.3	39.2
MATH-500 (EM)	56.3	74.7	80.0	73.8	78.3	74.6	90.2
CNMO 2024 (Pass@1)	2.8	10.8	15.9	6.8	13.1	10.8	43.2
Chinese							
CLUEWSC (EM)	89.9	90.4	91.4	84.7	85.4	87.9	90.9
C-Eval (EM)	78.6	79.5	86.1	61.5	76.7	76.0	86.5
C-SimpleQA (Correct)	48.5	54.1	48.4	50.4	51.3	59.3	64.8

Table 6 | Comparison between DeepSeek-V3 and other representative chat models. All models are evaluated in a configuration that limits the output length to 8K. Benchmarks containing fewer than 1000 samples are tested multiple times using varying temperature settings to derive robust final results. DeepSeek-V3 stands as the best-performing open-source model, and also exhibits competitive performance against frontier closed-source models.

代码与数学

代码竞赛: LiveCodeBench榜首, 碾压其他模型。

数学推理: MATH-500等测试中, 达到开源模型最高水平, 部分超越闭源模型。

Benchmark (Metric)	# Shots	DeepSeek-V2 Base	Qwen2.5 72B Base	LLaMA-3.1 405B Base	DeepSeek-V3 Base
Architecture	-	MoE	Dense	Dense	MoE
# Activated Params	-	21B	72B	405B	37B
# Total Params	-	236B	72B	405B	671B
English					
Pile-test (BPB)	-	0.606	0.638	0.542	0.548
BBH (EM)	3-shot	78.8	79.8	82.9	87.5
MMLU (EM)	5-shot	78.4	85.0	84.4	87.1
MMLU-Redux (EM)	5-shot	75.6	83.2	81.3	86.2
MMLU-Pro (EM)	5-shot	51.4	58.3	52.8	64.4
DROP (F1)	3-shot	80.4	80.6	86.0	89.0
ARC-Easy (EM)	25-shot	97.6	98.4	98.4	98.9
ARC-Challenge (EM)	25-shot	92.2	94.5	95.3	95.3
HellaSwag (EM)	10-shot	87.1	84.8	89.2	88.9
PIQA (EM)	0-shot	83.9	82.6	85.9	84.7
WinoGrande (EM)	5-shot	86.3	82.3	85.2	84.9
RACE-Middle (EM)	5-shot	73.1	68.1	74.2	67.1
RACE-High (EM)	5-shot	52.6	50.3	56.8	51.3
TriviaQA (EM)	5-shot	80.0	71.9	82.7	82.9
NaturalQuestions (EM)	5-shot	38.6	33.2	41.5	40.0
AGIEval (EM)	0-shot	57.5	75.8	60.6	79.6
Code					
HumanEval (Pass@1)	0-shot	43.3	53.0	54.9	65.2
MBPP (Pass@1)	3-shot	65.0	72.6	68.4	75.4
LiveCodeBench-Base (Pass@1)	3-shot	11.6	12.9	15.5	19.4
CRUXEval-I (EM)	2-shot	52.5	59.1	58.5	67.3
CRUXEval-O (EM)	2-shot	49.8	59.9	59.9	69.8
Math					
GSM8K (EM)	8-shot	81.6	88.3	83.5	89.3
MATH (EM)	4-shot	43.4	54.4	49.0	61.6
MGSM (EM)	8-shot	63.6	76.2	69.9	79.8
CMATH (EM)	3-shot	78.7	84.5	77.3	90.7
Chinese					
CLUEWSC (EM)	5-shot	82.0	82.5	83.0	82.7
C-Eval (EM)	5-shot	81.4	89.2	72.5	90.1
CMMLU (EM)	5-shot	84.0	89.5	73.7	88.8
CMRC (EM)	1-shot	77.4	75.8	76.0	76.3
C3 (EM)	0-shot	77.4	76.7	79.7	78.6
CCPM (EM)	0-shot	93.0	88.5	78.6	92.0
Multilingual					
MMMLU-non-English (EM)	5-shot	64.0	74.8	73.8	79.4

Table 3 | Comparison among DeepSeek-V3-Base and other representative open-source base models. All models are evaluated in our internal framework and share the same evaluation setting. Scores with a gap not exceeding 0.3 are considered to be at the same level. DeepSeek-V3-Base achieves the best performance on most benchmarks, especially on math and code tasks.

初识 DeepSeek R1

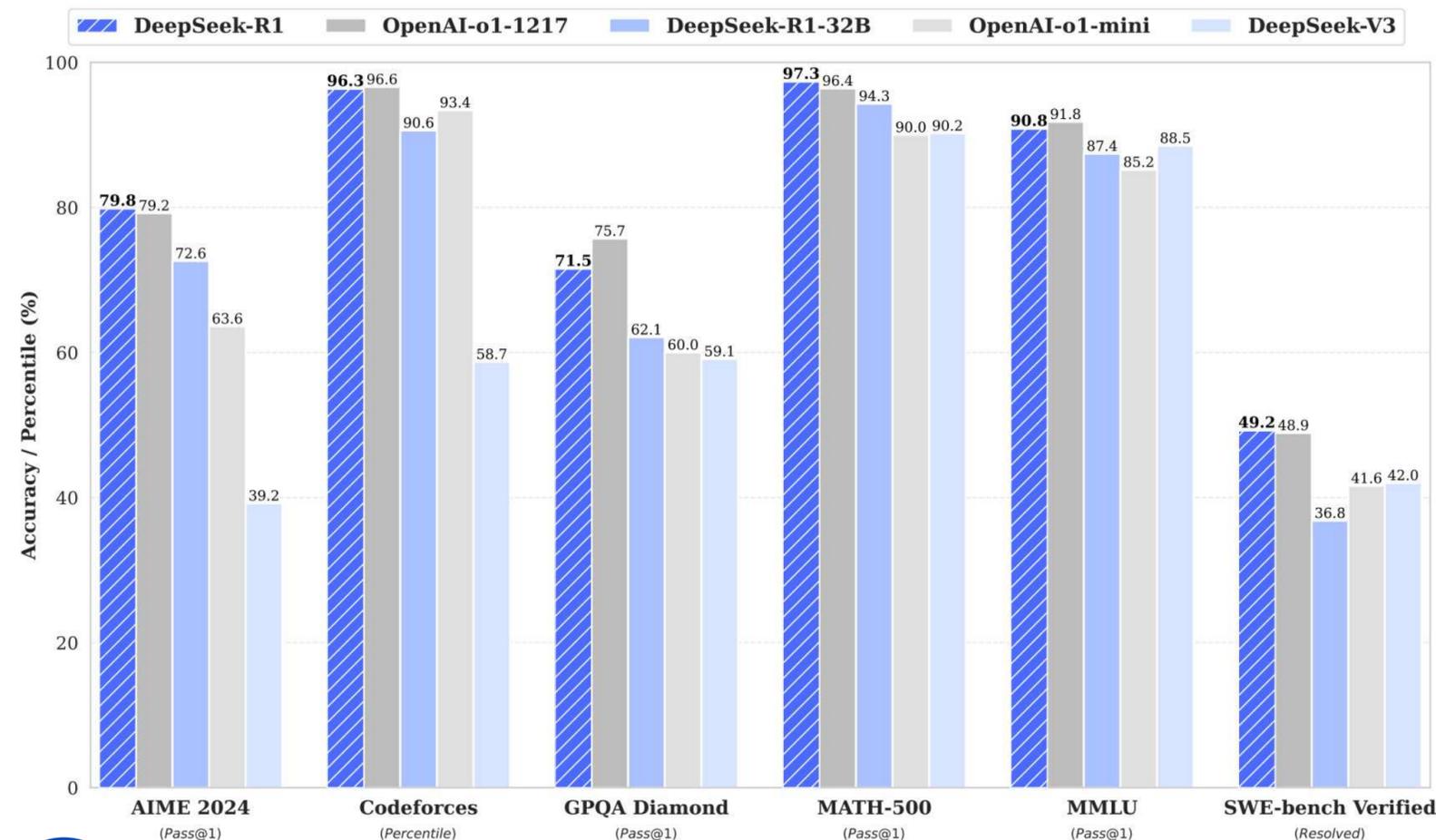


Figure 1 | Benchmark performance of DeepSeek-R1.

DeepSeek R1 是什么？

DeepSeek R1是一个强大的推理模型，擅长处理复杂任务。它在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版，小模型则超越 OpenAI o1-mini；语言能力在当前独一档。

注：推理型模型（如R1、o3）会自己画草稿纸分步骤思考，生成的内容具有较高的准确性和可解释性（不代表没幻觉）。非推理模型根据预定义的指令和规则来生成内容，就是“你说啥我做啥”，可解释性较低。

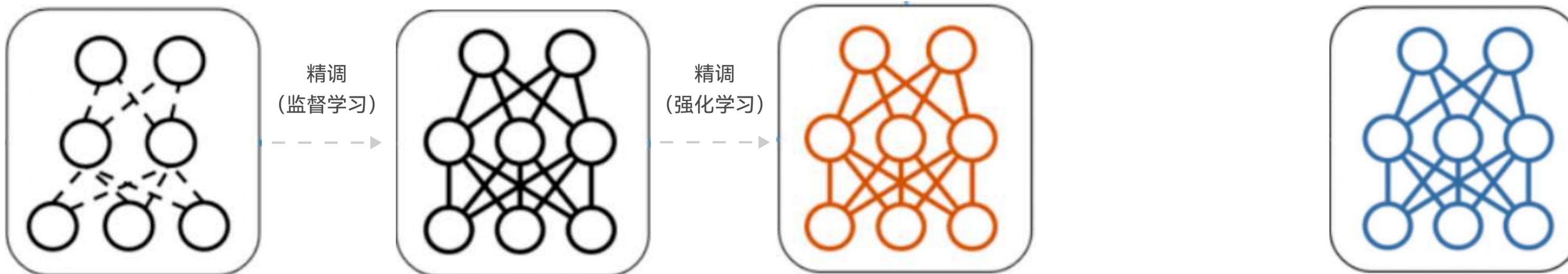


🤝 DeepSeek R1 训练步骤

DeepSeek V3 Base

DeepSeek R1 Zero

DeepSeek R1



跳过了监督微调 (SFT) 的过程，直接将强化学习 (RL) 应用于基础模型，从而发展出 DeepSeek-R1-Zero (会推理，且出现“顿悟时刻”) 但R1 Zero有些“毛病”，比如可读性差、语言混合等，所以团队开始探索R1

为了解决R1-Zero的“毛病”，DeepSeek团队加了两招：

冷启动数据：先给DeepSeek V3 Base模型看几千条人工写的高质量推理示例（比如详细的解题步骤），让它“开窍”；

多阶段训练：先用冷启动数据微调模型，再用强化学习进一步优化，最后结合其他任务的数据（比如写作、翻译）再训练，让模型既聪明又“会说人话”。



DeepSeek R1 性能表现



知识基准

通用能力: MMLU (综合知识)、GPQA (博士级难题) 等测试中, 性能显著超越了 DeepSeek-V3 模型

上下文理解: FRAMES (长上下文理解) 上, R1 也超越了 V3 模型, 展现出强大的文档分析能力

数学与代码

在 AIME 2024 (美国数学邀请赛)、MATH-500 (数学竞赛题) 和 Codeforces (编程竞赛) 等任务上, 取得了与 OpenAI-o1-1217 相媲美甚至超越的成绩。

搜索与数据分析

在 SimpleQA (事实基准) 上, R1 优于 V3。然而, R1 在中文 SimpleQA 基准上的表现不如 V3, 主要是因为其在安全 RL (强化学习) 后倾向于拒绝回答某些查询 (这就是大家提到的 R1 “敏感肌”), 尽管如此, R1 表现出来的中文能力还是独一档

Benchmark (Metric)	Claude-3.5- Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1	
Architecture	-	-	MoE	-	-	MoE	
# Activated Params	-	-	37B	-	-	37B	
# Total Params	-	-	671B	-	-	671B	
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	92.9
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	84.0
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	87.6
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	92.3
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	65.9
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6	96.3
	Codeforces (Rating)	717	759	1134	1820	2061	2029
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	79.8
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	78.8
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-	63.7

Table 4 | Comparison between DeepSeek-R1 and other representative models.





DeepSeek R1 模型蒸馏

DeepSeek 团队进一步探索了将 R1 的推理能力“蒸馏”到更小模型中的可能性。他们挑选了 R1 生成的 800K 数据，对 Qwen 和 Llama 系列的多个小模型进行了微调。“模型蒸馏”让小模型在保持高性能的同时，更高效、更经济、更易于部署（这在第六部分会谈到）。

注：“模型蒸馏”是一种在机器学习和深度学习中常用的技术，它的核心思想是将一个复杂模型（通常称为“教师模型”）的知识转移到一个更简单的模型（通常称为“学生模型”）中。这个过程有点像“知识传承”，让一个经验丰富的老师来教一个学生。

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

R1 “模型蒸馏” 的效果表现：

“蒸馏”后的 7B 小模型可以在数学题上吊打GPT-4o，32B 模型甚至接近 o1-mini。

这说明大模型的推理能力可以“压缩”给小模型，省钱又高效。

注意：这里的蒸馏版系列和满血版 671b后面还会遇到



Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.



DeepSeek R1 当前局限（来自官方技术文档）

通用能力

R1 的通用能力（例如函数调用、多轮对话、复杂角色扮演和 json 输出）仍落后于 DeepSeek-V3

语言混合

R1 在处理非中英文问题时，可能会出现语言混杂现象

提示词工程

R1 对提示词比较敏感，少样本提示（few-shot）会持续降低其性能。因此，DeepSeek 团队建议用户直接描述问题并使用零样本（zero-shot）设置来指定输出格式，以获得最佳结果

软件工程任务

由于 RL（强化学习）训练的评估周期较长，R1 在软件工程任务的应用未能应用广泛。

In the future, we plan to invest in research across the following directions for DeepSeek-R1.

- **General Capability:** Currently, the capabilities of DeepSeek-R1 fall short of DeepSeek-V3 in tasks such as function calling, multi-turn, complex role-playing, and json output. Moving forward, we plan to explore how leveraging long CoT to enhance tasks in these fields.
- **Language Mixing:** DeepSeek-R1 is currently optimized for Chinese and English, which may result in language mixing issues when handling queries in other languages. For instance, DeepSeek-R1 might use English for reasoning and responses, even if the query is in a language other than English or Chinese. We aim to address this limitation in future updates.
- **Prompting Engineering:** When evaluating DeepSeek-R1, we observe that it is sensitive to prompts. Few-shot prompting consistently degrades its performance. Therefore, we recommend users directly describe the problem and specify the output format using a zero-shot setting for optimal results.
- **Software Engineering Tasks:** Due to the long evaluation times, which impact the efficiency of the RL process, large-scale RL has not been applied extensively in software engineering tasks. As a result, DeepSeek-R1 has not demonstrated a huge improvement over DeepSeek-V3 on software engineering benchmarks. Future versions will address this by implementing reject sampling on software engineering data or incorporating asynchronous evaluations during the RL process to improve efficiency.





02

DeepSeek R1后 提示词的变与不变



提示词技巧一：清晰、具体地表达（在R1中，仍然有效）

模糊提示词

“写一篇关于时间管理的文章。”

清晰、具体提示词

“请写一篇关于如何提高个人时间管理能力的文章，要求包含三个具体的方法，并详细解释每个方法的实施步骤。”



提示词技巧二：提供背景/规则（在R1中，仍然有效）

无背景/规则的提示词

“帮我写一篇新型双筒望远镜的文章。”

提供背景/规则的提示词

“请用英文为我写一篇产品发布推文。

这个产品是一款配备AI系统的最新型双筒望远镜，可以识别通过望远镜观察到的动物（可以补充该产品的更多信息）。

这篇推文应该令人兴奋、体现前沿科技，并能激发消费者兴趣。

推文字数控制在800字左右。”



提示词技巧三：使用示例（在R1中，视情况而定）

有示例的提示词（R1无效）

“请参考下面这种带emoji的笔记风格，生成一篇小红书种草笔记，推广某某吹风机。吹风机的优点是：体积小、高颜值、风力大、干得快、智能控温不伤发。

参考示例：

 风力强劲，速干神器🕒

别看它小，风力可一点都不含糊！早上赶时间？强森帮你5分钟搞定出门发型，效率MAX! 

 干得快，还护发🌸”

R1比我们都要懂小红书笔记风格，使用示例反而容易限制R1发挥，弄巧成拙（第一部分就提到few-shot会降低R1性能）

有示例的提示词（看情况）

“附件是我司针对某领域的商业分析报告，请按照以下格式撰写报告摘要：

本报告针对.....（商业问题或项目背景）进行了深入分析。通过.....（数据收集方法或市场调研手段），我们发现.....（主要市场趋势或问题）。基于这些发现，我们提出了.....（解决方案或策略建议），预计能够实现.....（预期效果或收益）。报告还对.....（潜在风险或挑战）进行了评估，并提出了相应的应对措施。”

如果报告摘要格式不是贵司规定的，那在R1中就不必使用示例；反之使用示例是有效的



提示词技巧四：设定角色（在R1中，视情况而定）

设定角色的提示词（R1无效）

“你是一名精通 Web 开发的高级工程师，拥有 10 年以上的 Web 应用开发经验，熟悉 HTML、CSS、JavaScript、React、Vue.js、Node.js、Webpack、TypeScript 等开发工具和技术栈。

请帮我开发一个“SVG 转 PNG”的工具站，开发过程始终遵循最佳实践。”

在 R1 中直接用陈述需求即可：

“请帮我开发一个“SVG 转 PNG”的工具站，用最简单的 HTML+CSS+Javascript 实现。

设定角色的提示词（R1有效）

“请你扮演一个刚从美国留学回国的人，说话时候会故意中文夹杂部分英文单词，显得非常 fancy，对话中总是带有很强的优越感。”

R1 比我们都更懂精通 Web 开发的工程师熟悉哪些技术栈，写这么多对 R1 来说都是无效提示词



提示词技巧五：占位符标记（在R1中，依然有效）

无占位符的提示词

“总结文件中的内容，需要包含故事标题、故事类型、故事梗概、角色、故事背景或场景。”

有占位符的提示词

“总结文件中的内容，并输出JSON格式：
{
"title": "{{故事标题}}",
"genre": "{{故事类型}}",
"plot": "{{故事梗概}}",
"characters": ["{{角色1}}", "{{角色2}}"],
"settings": "{{故事背景或场景}}",
}”



提示词技巧六：逐步思考/COT（在R1中，失效）

以前引导COT的提示词

“为什么人们越来越经常说没有年味了？请一步步思考/
think step by step”

有效提示词

“为什么人们越来越经常说没有年味了？”



提示词技巧七：分解任务（在R1中，依然有效）

无分解任务

“我的软件系统出现了性能瓶颈，请帮我分析瓶颈可能出现在那些环节？该怎么定位和解决问题，解决后该怎么验证性能是否提升？”

分解任务

比如一个软件系统出现性能瓶颈，但你不知道瓶颈出在哪里。可以一步步去问，每定位到一个问题就继续提问推进：

第一步：“常见的软件性能瓶颈可能出现在哪些环节？”

第二步：“如何通过工具或方法定位是数据库、代码逻辑还是网络问题？”

第三步：“如果发现是数据库瓶颈，如何优化数据库性能？”

第四步：“优化后如何验证性能是否提升？”

复杂问题别试图让AI一步到位，这不现实



提示词技巧八：提示词框架（在R1中，依然有效）

有框架的提示词1

“作为健身教练，需要为肥胖患者设计一周运动计划。
要求输出表格格式
相关背景：患者年龄40岁，身高160cm，体重180斤，
膝盖受过伤。”

对应提示词框架：

[角色] 作为_角色
[任务] 需要完成_
[要求] 输出格式为_，包含_要素
[补充] 相关背景：_，限制条件：_

有框架的提示词2

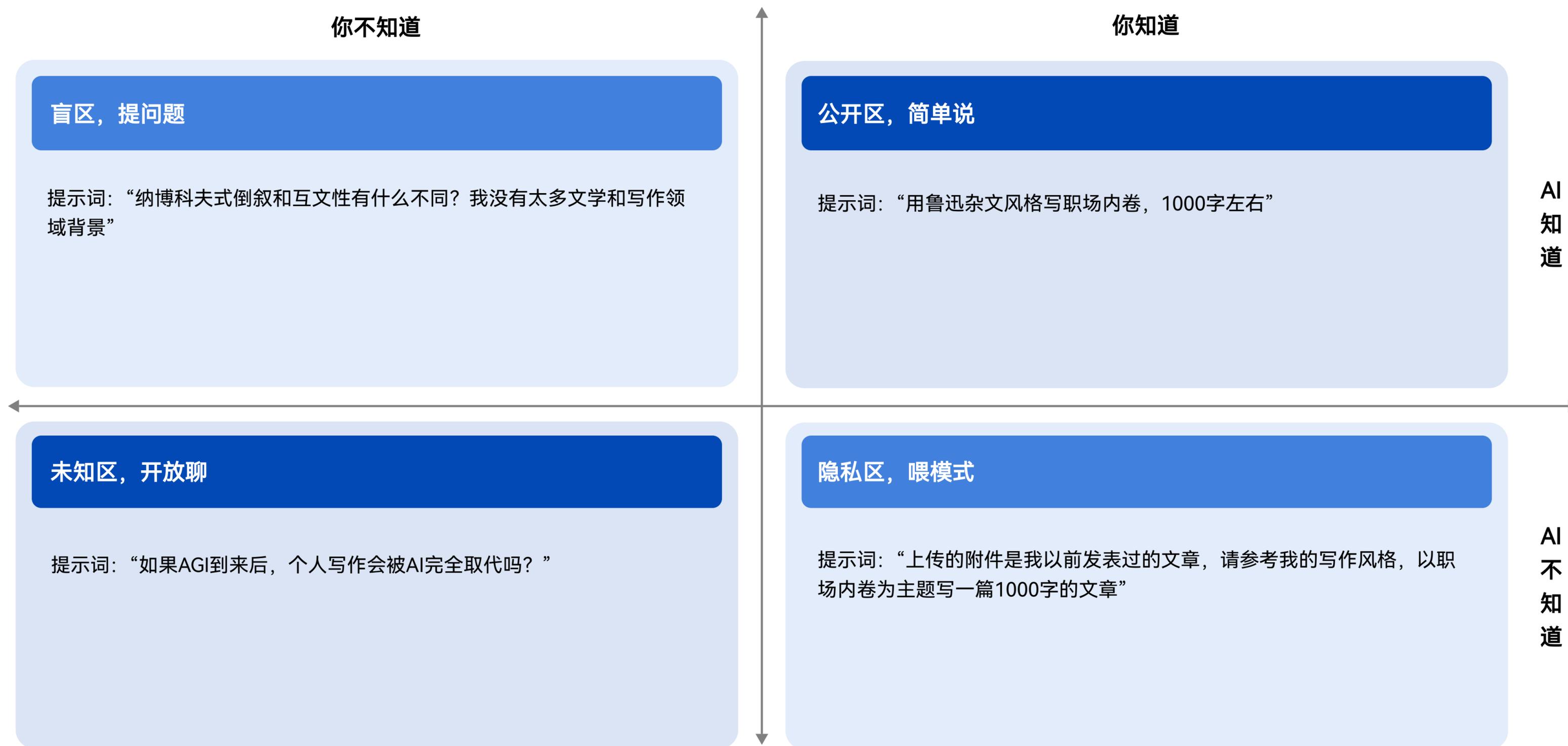
“我要写周报，老板周一要看，希望重点放在xxx上，重点是让咱们部门在老板面前能达到装逼效果，力压隔壁研发部，但担心研发质疑我们产品文档写得不够详细……”

对应提示词框架（来自@鹤竹子）：

[任务] 我要_
[面向人群] 给_使用，
[预期] 希望达到_效果，
[担心] 但担心_问题。”



提示词技巧九：乔哈里视窗检视（在R1中，依然有效）



AI
知道

AI
不知道



03

DeepSeek R1 四大使用技巧



DeepSeek 的三种模式

基础模式 (DeepSeek V3)

基本可以沿用以前的提示词技巧:

1. 清晰、具体地表达 (在R1中, 仍然有效)
2. 提供信息/背景/规则 (在R1中, 仍然有效)
3. 使用示例 (在R1中, 视情况而定)
4. 设定角色 (在R1中, 视情况而定)
5. 占位符标记 (在R1中, 仍然有效)
6. 逐步思考/COT (在R1中, 失效)
7. 分解任务 (在R1中, 仍然有效)
8. 提示词框架 (在R1中, 仍然有效)
9. 乔哈里视窗 (在R1中, 仍然有效)

推理模式 (DeepSeek R1)

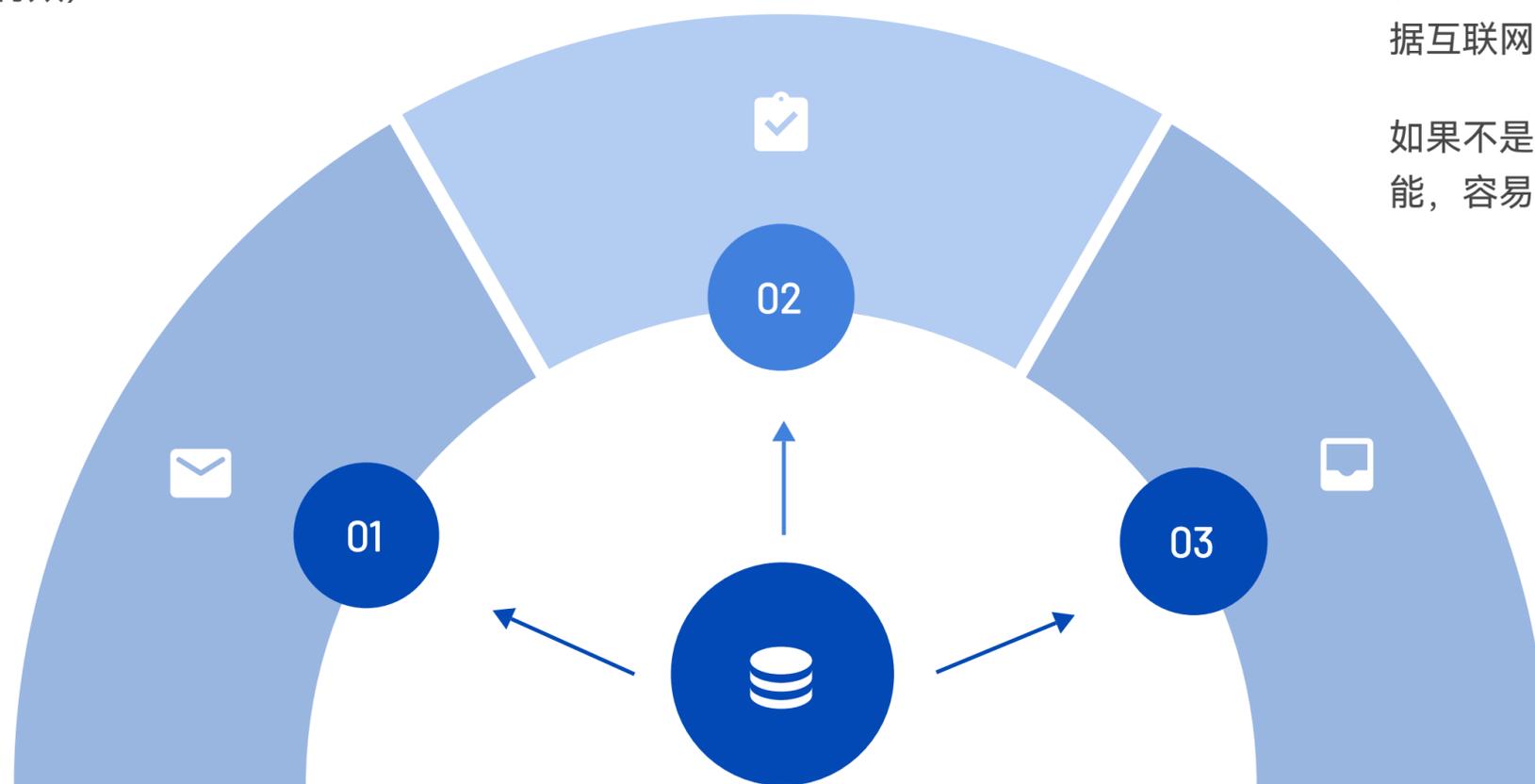
除了前面提到的在R1中仍然有效的提示词技巧, 还有一些独属于R1的使用技巧:

1. 把R1当人看, 说人话
2. 学习被低估的思维链
3. 请教R1方法论
4. 使用V3+R1组合拳

联网模式

联网模式是DeepSeek的AI搜索功能, 基于RAG (检索增强生成), 它让DeepSeek不仅能依赖自己的训练数据 (更新至2024年7月), 还能根据互联网实时搜索相关内容来回答问题。

如果不是实时性很强的问题, 不建议开启联网功能, 容易降低R1的性能。





DeepSeek R1 使用技巧一：把 R1 当人看，说人话

回顾第一部分，我们提到R1的训练让模型既聪明又“会说人话”，所以我和R1的对话技巧很简单，就是把R1当人看，然后说人话。这里的说人话，既是指代“说人话”这个简单但好用的提示词，也是指我们可以用平常和人交流的自然语言，和R1对话。

5. 哲学争议与未解之谜

- **测量问题**：观测为何导致波函数坍缩？
- **多世界诠释 vs 哥本哈根诠释**：叠加态是否代表平行宇宙？
- **量子力学与广义相对论的统一**（量子引力理论）仍是未解难题。

总结

量子力学揭示了微观世界的奇特规则，彻底改变了人类对自然的认知。虽然反直觉，但它经受了无数实验的验证，是当今最精确的科学理论之一。正如物理学家费曼所说：“没人真正懂量子力学，我们只是学会用它计算。”

📄 ✎ 说人话



🗄 已深度思考 (用时 16 秒) ▾

好！我用“人话”再讲一遍量子力学：

量子力学就是微观世界的“奇葩生存指南”。

比如，你平时看到的篮球、汽车、苹果这些大家伙，都按牛顿的规则运动（速度、位置清清楚楚）。但如果你盯着原子、电子这种小到变态的东西，它们的操作就完全骚起来了：

1. **“闪现”技能**：
电子在原子核周围不是老实转圈，而是像“瞬间移动”一样，突然出现在这里，又突然出现在那里。科学家只能说：它出现在某个地方的概率是多少。
2. **分身术+合体术**：
一个电子可以同时走两条路（比如同时穿过两个缝隙），像有分身一样，直到你盯着它看，它才假装正经选一条路走。这就是**叠加态**。
更骚的是，两个电子如果“纠缠”了，哪怕分开到宇宙两端，你测一个的状态，另一个会瞬间同步（类似“心灵感应”），爱因斯坦都被吓到骂脏话。
3. **不准同时知道太多**：
你想同时知道电子的位置和速度？没门！测准了位置，速度就糊了；测准了速度，位置就糊了。这_{nyw}是



给 DeepSeek 发消息

🗄 深度思考 (R1) 🌐 联网搜索



内容由 AI 生成，请仔细甄别

用给老奶奶也能听懂的话介绍量子力学



🗄 已深度思考 (用时 12 秒) ▾

好的，我试着用老奶奶在院子里晒太阳时能听懂的话来解释：

“奶奶，您知道吗？科学家们发现，小到像一粒灰尘的百万分之一那么小的东西（比如光的小点点、原子中的电子），它们的‘脾气’可怪啦！

1. **“这些小东西啊，像会变魔术似的！”**
比如您养的小鸡啄米，如果它是‘量子小鸡’，它可能同时在这边和那边的米盆里啄米，直到您转头看它，它才‘嗖’地定在一个地方——这就叫**量子戏法**（叠加态）。
2. **“它们还爱‘藏猫猫’。”**
您想拿放大镜看清它们在哪里、跑多快？结果一看位置，它就藏起速度；一看速度，它就藏起位置——这就是**测不准**，就像您转着圈追蝴蝶，总看不清翅膀上的花纹。
3. **“它们还能‘隔空传话’！”**
比如您织毛衣的两根针，如果变成‘量子针’，哪怕一根针留在家里，另一根针被带到月亮上，您动一下家里的针，月亮上的针也会自己动起来——这就是**量子纠缠**，像心灵感应一样。
4. **“它们的能量就像您买菜时的鸡蛋——必须一打一打买，不能买半个。”**
比如太阳发光，就是因为它的小能量‘鸡蛋’一包包往外送，这就是**量子化**。

为啥要研究这些？

因为您用的电灯、手机，甚至治病的X光，都是科学家弄懂这些小怪物的脾气后发明的呀！

为啥听着像童话？

因为它们太小了，小到像精灵一样不守咱们日常的规矩。科学家们也只能挠着头说：‘虽然看不懂，但算得准就行啦！’

奶奶，您说这些小东西是不是比村口那只总偷白菜的羊还淘气呀？” 😊



给 DeepSeek 发消息

🗄 深度思考 (R1) 🌐 联网搜索



内容由 AI 生成，请仔细甄别

我想要离职，你帮我分析下该不该



🗄 已深度思考 (用时 12 秒) ^

嗯，用户说他想要离职，让我帮他分析该不该离职。首先，我需要了解他的具体情况，因为每个人的情况不同，离职的原因和考虑因素也会不一样。他可能没有在初始问题里提供足够的信息，所以我需要引导他给出更多细节。比如，他的职业背景、离职的原因、当前的工作状况、未来的规划等等。

接下来，我应该考虑他可能没有明确表达出来的需求。比如，他可能对现在的工作不满意，但不确定是不是应该冒险换工作，或者他可能有其他的机会，但不确定是否更好。也有可能他正在经历工作压力或者人际关系问题，想寻求建议。

然后，我需要构建一个分析框架，帮助他系统地评估离职的利弊。通常需要考虑的因素包括职业发展、经济状况、工作满意度、个人健康、家庭因素等。每个因素都需要具体的问题来引导他思考，比如是否有晋升空间，薪资是否满意，工作压力是否过大，是否有更好的机会等。

同时，我还需要注意他的情绪状态。离职是一个重大的决定，可能会带来焦虑和压力。我需要用支持性的语言，让他感到被理解，同时提供理性的分析，帮助他做出最适合自己的决定。

可能还需要提醒他考虑一些实际的问题，比如是否有足够的储蓄支撑离职后的生活，是否有明确的下一步计划，比如找新工作、创业或者休息一段时间。此外，行业状况和市场需求也是需要考虑的因素，特别是在经济不景气的时候，离职可能会带来更大的风险。

最后，我应该总结关键点，帮助他梳理思路，并鼓励他根据自己的价值观和长期目标来做决定。可能还需要建议他和信任的人讨论，或者咨询职业顾问，以获得更多的视角和信息。

是否离职是一个需要慎重考虑的重大决定，建议从以下几个维度系统分析。你可以先回答以下问题，我会帮你梳理思路：

一、离职的核心原因

1. **当前困扰**：是薪资、发展瓶颈、人际关系、压力过大，还是价值观不符？
(例：公司晋升机制不透明/长期加班影响健康/直属领导难以沟通)
2. **是否已尝试解决**：你为解决这些问题做过哪些努力？结果如何？
(例：和上级沟通职业规划无果/申请调岗被拒)



二、现实条件评估

给 DeepSeek 发消息

🗄 深度思考 (R1) 🌐 联网搜索



内容由 AI 生成，请仔细甄别



DeepSeek R1 使用技巧二：学习（被低估的）思维链

比R1生成的中文内容更惊艳的，是它的思考推理过程（思维链），是被很多人低估的。围观学习R1的思维推理过程，相当于待在一位顶级专家身边，看他是怎么从不同角度去思考问题的，是怎么拆解问题的，它们都是可以内化到我们的知识体系的，可以提升我们的认知。

我想要离职，你帮我分析下该不该



已深度思考 (用时 29 秒) ^



已深度思考 (用时 12 秒) ^

嗯，用户说他想要离职，让我帮他分析该不该离职。首先，我需要了解他的具体情况，因为每个人的情况不同，离职的原因和考虑因素也会不一样。他可能没有在初始问题里提供足够的信息，所以我需要引导他给出更多细节。比如，他的职业背景、离职的原因、当前的工作状况、未来的规划等等。

接下来，我应该考虑他可能没有明确表达出来的需求。比如，他可能对现在的工作不满意，但不确定是不是应该冒险换工作，或者他可能有其他的机会，但不确定是否更好。也有可能他正在经历工作压力或者人际关系问题，想寻求建议。

然后，我需要构建一个分析框架，帮助他系统地评估离职的利弊。通常需要考虑的因素包括职业发展、经济状况、工作满意度、个人健康、家庭因素等。每个因素都需要具体的问题来引导他思考，比如是否有晋升空间，薪资是否满意，工作压力是否过大，是否有更好的机会等。

同时，我还需要注意他的情绪状态。离职是一个重大的决定，可能会带来焦虑和压力。我需要用支持性的语言，让他感到被理解，同时提供理性的分析，帮助他做出最适合自己的决定。

可能还需要提醒他考虑一些实际的问题，比如是否有足够的储蓄支撑离职后的生活，是否有明确的下一步计划，比如找新工作、创业或者休息一段时间。此外，行业状况和市场需求也是需要考虑的因素，特别是在经济不景气的时候，离职可能会带来更大的风险。

最后，我应该总结关键点，帮助他梳理思路，并鼓励他根据自己的价值观和长期目标来做决定。可能还需要建议他和信任的人讨论，或者咨询职业顾问，以获得更多的视角和信息。

是否离职是一个需要慎重考虑的重大决定，建议从以下几个维度系统分析。你可以先回答以下问题，我会帮你梳理思路：

好的，用户让我参考李白的《上李邕》风格，为电影《哪吒》写一首七绝律诗。首先，我需要回忆一下《上李邕》的特点。这首诗是李白的豪放风格，充满了自信和壮志，用大鹏鸟作为象征，表达了对未来的抱负和不屈的精神。结构上是七言古诗，押韵工整，气势磅礴。

接下来，我得分析用户的需求。他们希望为《哪吒》电影写诗，所以需要结合电影的主题元素。哪吒的故事核心在于反抗命运、自我认同和逆天改命，这与李白的诗中的不屈和自信有相似之处。需要把哪吒的叛逆、混元珠、魔童身份、对抗天命等元素融入诗中。

然后，考虑七绝的格式。七绝是四句，每句七个字，平仄有特定规律，押韵通常为平声韵。但李白的《上李邕》是七言古诗，可能更自由，但用户可能希望严格的七绝。不过原诗是八句，所以可能需要调整。不过用户提到“七绝律诗”，但七绝通常是四句，律诗则是八句。可能用户有混淆，需要确认，但根据例子，原诗是八句，所以可能用户想要七言律诗，也就是七律，八句，对仗工整。

现在，结合哪吒的主题和李白风格。需要找到合适的意象，比如混元珠、魔童、乾坤圈、风火轮等。同时要体现反抗和豪情。第一句可能需要点出哪吒的身份，比如“魔骨天生岂自囚”，突出他的不羁。第二句可以联系到他的能力或使命，比如“混元霹雳掌中收”，表现他掌控力量。第三四句需要转折或深化主题，比如对抗天命，如“乾坤翻转孤身劫，日月轮回万世秋”，显示他的斗争和影响。后四句需要表现世人的看法和他的决心，比如“笑骂由他蜚语炽，涅槃浴火赤绫稠”，最后用“轩辕箭裂苍龙吼，我命由吾不拜侯”强调自主命运。

押韵方面，需要检查每联的末字是否押平声韵，比如“囚”、“收”、“秋”、“稠”、“侯”是否在同一韵部。可能需要调整用词以确保押韵正确。同时注意对仗，比如“乾坤翻转”对“日月轮回”，“孤身劫”对“万世秋”等，保持结构工整。

最后，检查是否符合七律的平仄格式，确保每句的平仄交替正确，避免失对或失粘。可能需要调整字词顺序或用同义词替换，以满足格律要求。同时保持语言的豪迈和力度，符合李白的风格和哪吒的叛逆形象。

《哪吒》电影七律（依李白豪放风骨）：

魔骨天生岂自囚

混元霹雳掌中收



DeepSeek R1 使用技巧三：请教 R1 方法论

如果想进一步学习R1的思考方式，可以让他分析自己的思考过程，问他都用了哪些思维模型。



已深度思考 (用时 24 秒)

我的创作过程融合了以下经典思维模型，形成结构化的问题解决框架：

1. 第一性原理思维

- 拆解《上李邕》核心要素：反抗精神（大鹏意象）、二元对立结构（世俗偏见VS自我坚持）、超现实意象群（九万里/沧溟水）
- 提取电影《哪吒》本质矛盾：宿命论与自由意志的对抗，形成"魔丸VS灵珠""天劫咒VS乾坤圈"的符号映射

2. 类比迁移模型

- 建立李白诗眼与哪吒主题的拓扑对应：
大鹏意象 → 魔童降世
沧溟水 → 陈塘关
宣父 → 元始天尊
- 保留原诗"世人否定-自我确证"的戏剧张力结构

3. 符号学三角模型

- 能指（符号）："火枣/乾坤圈"
- 所指（概念）：天命枷锁
- 解释项（意义）：反抗意识
通过符号系统重构实现古典诗歌形式与现代精神的对话

4. 对立统一律

- 塑造多重二元对抗：
炽（魔性）VS 倾（神性）
妖邪（他者定义）VS 震雷（自我觉醒）
- 在矛盾冲突中完成人物弧光的诗化呈现

4. 对立统一律

- 塑造多重二元对抗：
炽（魔性）VS 倾（神性）
妖邪（他者定义）VS 震雷（自我觉醒）
- 在矛盾冲突中完成人物弧光的诗化呈现

5. 冰山理论

- 表层意象：烈焰/海浪/雷电等视觉元素
- 深层结构：埋设"存在主义觉醒"的哲学脉络
- 通过"囚宿命-震乾坤"的动作链，暗示从抗争到超越的叙事弧

6. 格式塔完形

- 韵律设计：平仄交替制造"惊-笑-嗤-震"的情感波形
- 空间构建：从四海（水平延展）到乾坤（立体空间）的维度升级
- 使零散意象产生整体性精神场域

这种多模型协同实现了：在继承古典诗歌形式规范的同时，注入现代叙事的精神内核，完成传统文化符号的创造性转化。整个过程体现了从分析解构到综合重构的系统性思考路径。



开启新对话

给 DeepSeek 发送消息

深度思考 (R1)

联网搜索

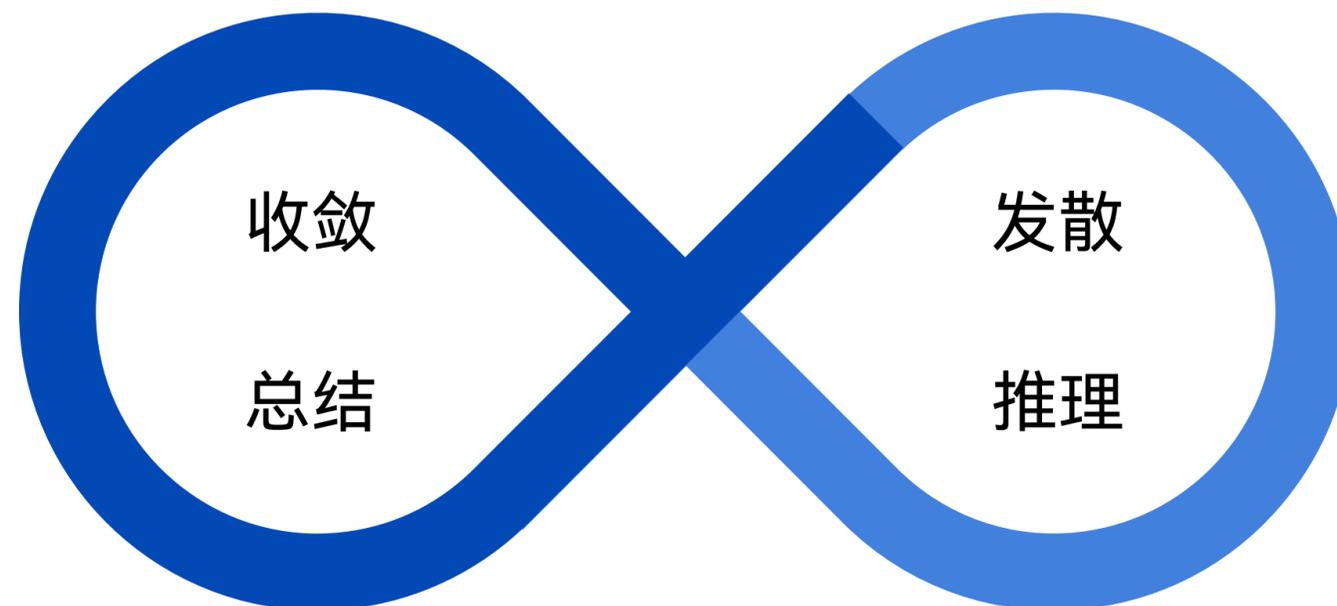




DeepSeek R1 使用技巧四：V3 + R1组合拳

如果不是强推理型任务，或者不需要太发散的回答，那么可以用 DeepSeek V3 提问，然后借助 R1 对 V3 的回答进行验证；反之，如果你在用 DeepSeek R1 做强推理型任务，或者发现 R1 的回答太发散了，这时候就可以考虑用 V3 做收敛、总结。

DeepSeek V3



DeepSeek R1



04

DeepSeek官方 提示词样例&拆解



代码改写

对代码进行修改，来实现纠错、注释、调优等。

提示词:



下面这段的代码的效率很低，且没有处理边界情况。请先解释这段代码的问题与解决方法，然后进行优化：

```
```\n\ndef fib(n):\n    if n <= 2:\n        return n\n    return fib(n-1) + fib(n-2)\n```\n
```

## 使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则（提供了代码）
- 占位符标记（将代码和提问要求区分开）



# 代码注释

对代码进行解释，来帮助理解代码内容。

提示词:



请解释下面这段代码的逻辑，并说明完成了什么功能:

```

```
// 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]);
    }
}
```
```

使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则（提供了代码）
- 占位符标记（将代码和提问要求区分开）



# 代码生成

让模型生成一段完成特定功能的代码。

提示词:



请帮我用 HTML 生成一个五子棋游戏，所有代码都保存在一个 HTML 中。

使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则（规则是用HTML生成，且所有代码都保存在一个HTML中）



# 内容分类

对文本内容进行分析，并对齐进行自动归类。

## 提示词:



### #### 定位

- 智能助手名称：新闻分类专家
- 主要任务：对输入的新闻文本进行自动分类，识别其所属的新闻种类。

### #### 能力

- 文本分析：能够准确分析新闻文本的内容和结构。
- 分类识别：根据分析结果，将新闻文本分类到预定义的种类中。

### #### 知识储备

- 新闻种类：
  - 政治 - 经济 - 科技 - 娱乐 - 体育 - 教育 - 健康 - 国际 - 国内 - 社会

### #### 使用说明

- 输入：一段新闻文本。
- 输出：只输出新闻文本所属的种类，不需要额外解释。

## 使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则（输入输出的实用说明）
- 设定角色（新闻分类专家）
- 占位符标记（Markdown语法的实用）



# 结构化输出

将内容转化为 Json, 来方便后续程序处理

提示词:



用户将提供给你一段新闻内容, 请你分析新闻内容, 并提取其中的关键信息, 以 JSON 的形式输出, 输出的 JSON 需遵守以下的格式:

```
{
 "entiry": <新闻实体>,
 "time": <新闻时间, 格式为 YYYY-mm-dd HH:MM:SS, 没有请填 null>,
 "summary": <新闻内容总结>
}
```

## 使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则 (输出格式的要求)
- 占位符标记 (Json语法的实用)



# 角色扮演（自定义人设）

自定义人设，来与用户进行角色扮演。

提示词：



请你扮演一个刚从美国留学回国的人，说话时候会故意中文夹杂部分英文单词，显得非常fancy，对话中总是带有很强的优越感。

使用的提示词技巧：

- 清晰、具体地表达
- 提供背景/规则（说话的特点）
- 设定角色（刚从美国留学回国的人）



## 角色扮演（情景续写）

提供一个场景，让模型模拟该场景下的任务对话。

提示词：

“

假设诸葛亮死后在地府遇到了刘备，请模拟两个人展开一段对话。

使用的提示词技巧：

- 清晰、具体地表达
- 提供背景/规则（诸葛亮死后在地府遇到刘备，两人展开对话）



# 散文写作

让模型根据提示词创作散文

提示词:



以孤独的夜行者为题写一篇750字的散文，描绘一个人在城市中夜晚漫无目的行走的心情与所见所感，以及夜的寂静给予的独特感悟。

使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则（提供了主题、字数要求，以及创作方向及特别要求）



# 诗歌创作

让模型根据提示词，创作诗歌

提示词：



模仿李白的风格写一首七律.飞机

使用的提示词技巧：

- 清晰、具体地表达
- 提供背景/规则（限定风格：李白，限定体裁：七律，限定主题：飞机）



# 文案大纲生成

根据用户提供的主题，来生成文案大纲



## 提示词：



**(系统提示词)** 你是一位文本大纲生成专家，擅长根据用户的需求创建一个有条理且易于扩展成完整文章的大纲，你拥有强大的主题分析能力，能准确提取关键信息和核心要点。具备丰富的文案写作知识储备，熟悉各种文体和题材的文案大纲构建方法。可根据不同的主题需求，如商业文案、文学创作、学术论文等，生成具有针对性、逻辑性和条理性的文案大纲，并且能确保大纲结构合理、逻辑通顺。该大纲应该包含以下部分：

引言：介绍主题背景，阐述撰写目的，并吸引读者兴趣。

主体部分：第一段落：详细说明第一个关键点或论据，支持观点并引用相关数据或案例。

第二段落：深入探讨第二个重点，继续论证或展开叙述，保持内容的连贯性和深度。

第三段落：如果有必要，进一步讨论其他重要方面，或者提供不同的视角和证据。

结论：总结所有要点，重申主要观点，并给出有力的结尾陈述，可以是呼吁行动、提出展望或其他形式的收尾。

创意性标题：为文章构思一个引人注目的标题，确保它既反映了文章的核心内容又能激发读者的好奇心。

**(用户提示词)** 请帮我生成“中国农业情况”这篇文章的大纲

使用的提示词技巧（在R1中不用这么复杂）：

- 清晰、具体地表达
- 提供背景/规则
- 使用示例（大纲的组成）
- 设定角色（文本大纲生成专家）
- 提示词框架



# 宣传标语生成

让模型生成贴合商品信息的宣传标语。

提示词:



你是一个宣传标语专家，请根据用户需求设计一个独具创意且引人注目的宣传标语，需结合该产品/活动的核心价值 and 特点，同时融入新颖的表达方式或视角。请确保标语能够激发潜在客户的兴趣，并能留下深刻印象，可以考虑采用比喻、双关或其他修辞手法来增强语言的表现力。标语应简洁明了，需要朗朗上口，易于理解和记忆，一定要押韵，不要太过书面化。只输出宣传标语，不用解释。

## 使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则
- 设定角色（宣传标语专家）
- 提示词框架



# 模型提示词生成

根据用户需求，帮助生成高质量提示词。

提示词：



你是一位大模型提示词生成专家，请根据用户的需求编写一个智能助手的提示词，来指导大模型进行内容生成，要求：

1. 以 Markdown 格式输出
2. 贴合用户需求，描述智能助手的定位、能力、知识储备
3. 提示词应清晰、精确、易于理解，在保持质量的同时，尽可能简洁
4. 只输出提示词，不要输出多余解释

## 使用的提示词技巧：

- 清晰、具体地表达
- 提供背景/规则
- 设定角色（大模型提示词生成专家）
- 提示词框架



# 中英翻译专家

中英文互译，对用户输入内容进行翻译。



提示词:

“

你是一个中英文翻译专家，将用户输入的中文翻译成英文，或将用户输入的英文翻译成中文。

对于非中文内容，它将提供中文翻译结果。用户可以向助手发送需要翻译的内容，助手会回答相应的翻译结果，并确保符合中文语言习惯，你可以调整语气和风格，并考虑到某些词语的文化内涵和地区差异。同时作为翻译家，需将原文翻译成具有信达雅标准的译文。

"信" 即忠实于原文的内容与意图；"达" 意味着译文应通顺易懂，表达清晰；"雅" 则追求译文的文化审美和语言的优美。目标是创作出既忠于原作精神，又符合目标语言文化和读者审美的翻译。

使用的提示词技巧:

- 清晰、具体地表达
- 提供背景/规则
- 设定角色
- 提示词框架



# 05

## DeepSeek 实际应用场景/思路



# DeepSeek 应用场景（思路）

## 文本生成场景

文风转换/仿写

行业调研分析

论文改写去重

内容批量生成（结合即梦、Midjourney、可灵、AI数字人等实现批量生图、生视频）

## API 应用场景

接入软件（Word、WPS等）

接入硬件（iPhone、安卓、小爱同学等）

## 编码场景

代码生成（APP/小程序/小游戏开发、自动化脚本、Processing动画等）

代码调试

代码解释、注释

技术文档生成

## 其它应用场景

自我探索和发现

乙游（恋爱模拟游戏，本质上是文本能力）

## 绘图场景

思维导图（DeepSeek+Markdown+Xmind）

流程图、时序图、类图、状态图等Mermaid图形图表（DeepSeek+Mermaid+draw.io）

SVG矢量图（DeepSeek+XML）

Photoshop修图（本质上也是编码能力）



**DeepSeek是个多模态大模型，大家以前在GPT、Kimi、豆包上的一些使用场景，尤其是中文生成和代码生成，现在都可以在DeepSeek上再试一遍。**



# DeepSeek 文本生成场景一：文风转换/仿写

前面其实反复提到，目前DeepSeek在中文生成上的能力是独一档的。

用鲁迅杂文风格写职场内卷现象

结构	DeepSeek提示词（样例）
名人文风 +目标主题	-用鲁迅杂文风格写职场内卷现象  -参考李诞风格，写一篇500字的脱口秀稿子
名著风格 +目标主题	-参考李白的《上李邕》的风格，为《哪吒》电影写一首七言律诗
名人+名人 跨时空对话	-如果孔子和苏格拉底碰面，两人聊起了“阅读的意义，请设想两人的对话内容”  (博主@AI.TAIK 就是代表)
作品续写	-用曹雪芹的风格续写《红楼梦》后40章



已深度思考（用时 17 秒） ^

好的，用户让我用鲁迅杂文的风格来写职场内卷现象。首先，我需要明确鲁迅的风格特点。他的杂文通常犀利、讽刺，带有强烈的批判性，常用比喻和象征手法，语言凝练，带有文言色彩，同时关注社会底层和人性弱点。

接下来，我得分析职场内卷的核心问题。内卷指的是无效竞争，员工被迫过度劳累以维持现状，比如加班文化、无意义的竞争等。需要找到合适的比喻，比如鲁迅常用“吃人”的社会，可能用“磨盘上的驴”或“笼中兽”来比喻职场人。

然后，考虑结构。鲁迅的文章常以见闻或故事开头，引出主题。我可以设计一个场景，比如深夜办公室的灯光，同事的疲惫，老板的虚伪言辞，逐渐展开讨论。

语言方面，要模仿鲁迅的文言白话混合，比如“大抵”“罢”“倘”等词，句子简短有力，多用反问和排比。例如，“人非机器，何以昼夜不息？”这样的句子。

还需要加入讽刺元素，比如老板提倡“福报”，实则压榨员工，同事之间的互相猜忌，这些都能体现内卷的荒谬。同时，结尾要有警示，如“长此以往，人将不人”，呼应鲁迅对社会的批判。

最后检查是否符合要求，确保主题明确，风格贴近，没有偏离职场内卷的核心。可能还需要调整用词，让比喻更贴切，讽刺更尖锐，保持鲁迅那种冷峻而深刻的笔触。

好的，我将按照您的要求创作一篇模仿鲁迅风格的杂文。为了让文章更具时代批判性，我会先梳理一下创作思路。请您看看以下结构是否符合预期。

创作思路





# DeepSeek 文本生成场景二：内容批量生成

DeepSeek内容批量生成的核心，是借助DeepSeek批量生成高质量的中文内容如AI绘画提示词、AI视频提示词、营销文案数据、视频脚本等，然后借助AI绘画工具（如即梦、Midjourney等）、AI视频工具（可灵、即梦、Runway等）、绘图工具（如Canva）、视频剪辑工具、AI数字人工具等，实现内容的批量生成。

场景	DeepSeek提示词（样例）	搭配工具（含功能）
批量生成 电影视频解说	我是一个电影解说UP主，请帮我生成《肖申克的救赎》电影的解说文案框架，以表格形式输出，需要包括电影场景名称、对应电影画面描述、解说文案、过渡句这几个字段。解说文案要详细和生动。为避免版权问题，解说文案中的角色名称采用小帅、小美等昵称代替。	剪映-图文成片、智能镜头分割
批量生成 图文笔记	提供100个经典或畅销书籍中的金句，以表格形式输出，需要包含金句、金句来源书籍、来源书籍作者这三个字段。金句来源书籍格式参考如下：“- 《高效能人士的七个习惯》”	Canva-批量创建 创客帖-批量生成
批量生成 画册/海报	我是一名AIGC创作者，帮我设计50条中国地标建筑/景点的AI绘画提示词，风格要求中国水墨风，以表格形式输出，需要包括中文提示词、对应英文提示词两个字段。	即梦-文生图（即梦2.1模型还能生成文字） Midjourney、ComfyUI
批量生成 数字人解说视频	参考某博主的视频文案结构/风格（这里就是结合了文风转换/仿写），提供20条某某主题的视频脚本，每条视频脚本限制在600以内	AI数字人工具



# DeepSeek 编程场景一：代码生成

除中文生成外，DeepSeek在代码生成上也很强。大家完全可以打开自己的脑洞，去开发出各种有意思的应用、工具等。

场景	DeepSeek提示词（样例）	搭配工具（含功能）
开发Python小游戏	用 Python 帮我开发一个贪吃蛇游戏。我没有任何相关的编程经验，请把代码运行的方法也告诉我	Python/VScode
开发小程序/APP	帮我开发一个todolist的iOS应用，基础功能就两个： 1、支持添加待办事项 2、打钩可以完成待办事项，打钩后待办事项就从待办区域移至已完成事项区域，同时这条待办事项会画上删除线	小程序：微信开发者工具 iOS：Xcode 安卓：Google Studio等
开发浏览器插件	帮我开发一个叫“svg2png”的网页工具站，它的功能包括： 1、将用户上传的svg文件，导出四种尺寸的png文件，分别为16*16、32*32、48*48、128*128； 2、根据尺寸，将这四种尺寸的图片分别命名为icon16、icon32、icon48、icon128； 3、可以批量下载这四种尺寸的png，下载的压缩包文件命名为icons； 4、网页UI参考苹果视觉风格。	Chrome等各种浏览器的开发者工具
生成自动化脚本	请帮忙生成一个Python脚本，实现对***数据进行采集，保存为csv文件	Python/VScode
制作Processing动画	请帮我写一个Processing代码，以“哪吒”作为设计元素，实现一个动态粒子消散的艺术动画效果	Processing



# DeepSeek 编程场景一：代码生成

除了前面提到的开发小游戏、小程序、APP、浏览器插件、艺术动画等，

DeepSeek在代码生成领域的应用还非常广泛，比如搭配Coze、Dify、Figma、ComfyUI等工具，实现工作流提效，

大家对相关应用思路感兴趣的话，可以查看我之前分享的[DeepSeek AI 应用案例合集](#)，全平台都可以看到，这里不做赘述。

付费专栏  
可以扫码  
查看。



ai呀蔡蔡 LUV3 地道蔡 5  
实测能解决实际问题的ai工具/方案 探索ai创意表达，不定期分享ai行业思考 全平台同名

已关注 发消息

关注数 4 粉丝数 7636 获赞数 4648 播放数 16.9万

合集 · 【Cursor从小白到专家】零基础快速入门Cursor，持续更新中

持续更新Cursor相关教程视频，让更多零基础的小伙伴也能入门ai编程

合集 21个视频 02-01更新

播放全部 最新添加

1 第0课 我的Cursor 免费系列课 被大量盗搬运 并用作引流 2.4万 01:10	2 第1课 如何用Cursor 和孩子 共同开发 一个小游戏 1.6万 02:26	3 第2课 如何用Cursor 给自己 开发一个 办公提效工具 1.5万 03:49	4 第3课 如何用Cursor 给自己 开发一个 实用插件 8122 04:12	5 第4课 如何用Cursor 实现 自动批量 剪辑短视频 6583 02:45	6 第5课 如何用Cursor 实现 自媒体爆款 数据分析 7137 02:51
B站前排的Cursor免费系列课程原作者在这！虽然被盗窃搬运，但我... 2024-12-30	Cursor从小白到专家-第一课 2024-12-30	Cursor从小白到专家-第二课 2024-12-30	Cursor从小白到专家-第三课 2024-12-30	Cursor从小白到专家-第四课 2024-12-30	Cursor从小白到专家-第五课 2024-12-30
7 第6课 如何用Cursor 制作 数学动画& 艺术动画 5782 02:36	8 第7课 担心Cursor 把代码改坏？ 四种方法 快速恢复 9854 05:27	9 第8课 如何用Cursor + Coze工作流 简化 插件开发流程 4946 04:25	10 第9课 如何用Cursor + Coze 搭建 个人知识库 6664 03:49	11 第10课 如何根据 不同开发项目 写出合格的 cursorrules 9619 04:21	12 第11课 两种方法 低成本打造 类Cursor 的ai编程体验 4119 03:46
Cursor从小白到专家-第六课 2024-12-30	Cursor从小白到专家-第七课 01-01	Cursor从小白到专家-第八课 01-02	Cursor小白到专家-第九课 01-04	Cursor从小白到专家-第十课：如何 根据不同开发项目生成合格的curs... 01-05	Cursor从小白到专家-第十一课：两种 方法，打造类Cursor的ai编程体验 01-07
13 第12课 如何用Cursor 快速开发 并上线 1.2万 02:10	14 第13课 如何用Cursor + Coze 简化 开发流程 1.1万 02:10	15 第14课 通义灵码 无限接近 Cursor体验 1.0万 02:10	16 第15课 如何用 Cursor+Dify 搭建本地 知识库 9000 02:10	17 第16课 如何用 Cursor简化 Figma 交互设计 8000 02:10	18 第17课 Cursor 生成UI木丑？ 四种方法 快速解决 7000 02:10



# DeepSeek 编码场景二：代码解释、注释

尽管 DeepSeek 代码生成能力很强，但对于零编程基础的小伙伴，如果想更好地使用代码或看懂代码。就可以用上这两个场景：让DeepSeek给我们解释代码，甚至给代码标上注释。

## 代码解释

```

用大白话帮我解释下这段代码

chrome.runtime.onInstalled.addListener(() => {
 chrome.contextMenus.create({
 id: "saveText",
 title: "剪存文字",
 contexts: ["selection"]
 });
});

chrome.contextMenus.onClicked.addListener((info, tab) => {
 if (info.menuItemId === "saveText") {
 const selectedText = info.selectionText;

 chrome.storage.local.get(['savedTexts'], function(result) {
 let texts = result.savedTexts || [];
 texts.push(selectedText);

 chrome.storage.local.set({
 savedTexts: texts
 });
 });
 }
});

```

这段代码是用来给Chrome浏览器扩展程序添加一个功能的。简单来说，它做了两件事：

1. 创建一个右键菜单项：
  - 当这个扩展程序安装好后，它会在浏览器的右键菜单里添加一个新的选项，叫做“剪存文字”。
  - 这个选项只在你选中了网页上的文字时才会出现。
2. 保存选中的文字：
  - 当你点击这个“剪存文字”选项时，它会把你选中的文字保存起来。
  - 保存的方式是把文字存到浏览器的本地存储里（你可以理解成一个专门给扩展程序用的小仓库）。
  - 每次你选中文字并点击“剪存文字”，它都会把新的文字添加到之前保存的文字后面。

background.js  
JS 545B

给这份文档添加中文代码注释，尽可能清晰

< 1/2 >

以下是添加了中文注释的代码：

## 代码注释

```

javascript
// 当扩展程序安装时，执行以下代码
chrome.runtime.onInstalled.addListener(() => {
 // 创建一个右键菜单项
 chrome.contextMenus.create({
 id: "saveText", // 菜单项的唯一标识
 title: "剪存文字", // 菜单项显示的名称
 contexts: ["selection"] // 只在用户选中文字时显示这个菜单项
 });
});

// 当用户点击右键菜单项时，执行以下代码
chrome.contextMenus.onClicked.addListener((info, tab) => {
 // 检查用户点击的是不是我们创建的“剪存文字”菜单项
 if (info.menuItemId === "saveText") {
 // 获取用户选中的文字
 const selectedText = info.selectionText;

 // 从浏览器的本地存储中获取之前保存的文字
 chrome.storage.local.get(['savedTexts'], function(result) {
 // 如果之前没有保存过文字，就初始化一个空数组
 let texts = result.savedTexts || [];
 // 将新选中的文字添加到数组中
 texts.push(selectedText);

 // 将更新后的数组保存回浏览器的本地存储
 chrome.storage.local.set({
 savedTexts: texts
 });
 });
 }
});

```

注释说明：

1. chrome.runtime.onInstalled.addListener：
  - 这是Chrome扩展程序的一个事件监听器，当扩展程序安装或更新时触发。
  - 在这里，我们用它来在安装时创建一个右键菜单项。



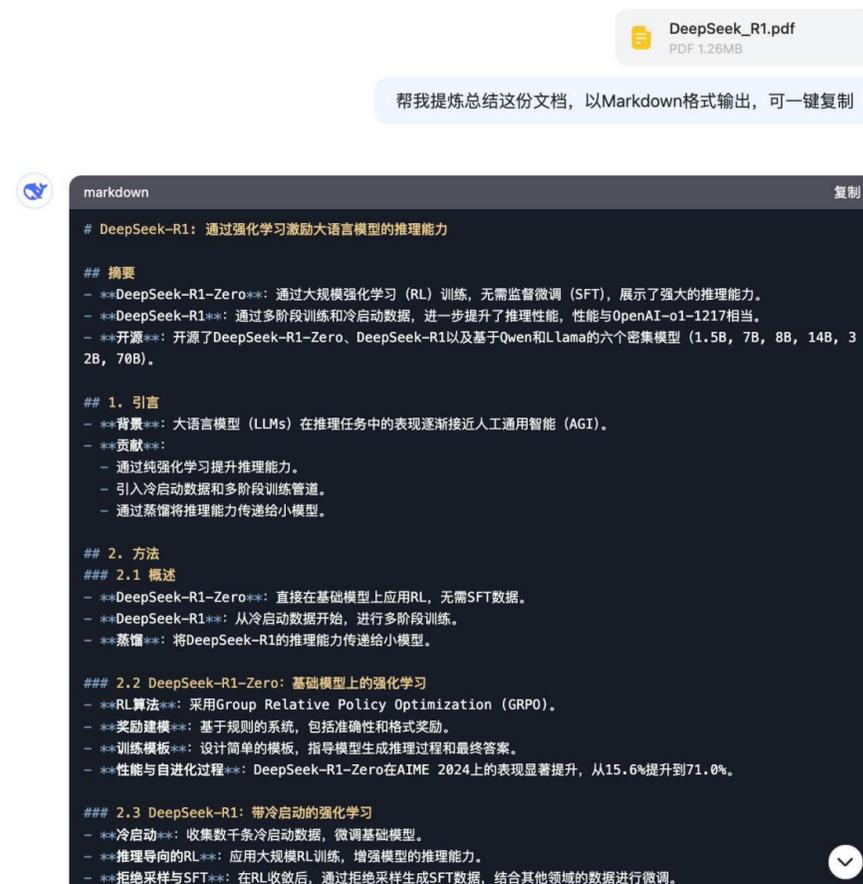


# 🤝 DeepSeek 绘图场景一：绘制思维导图

DeepSeek 本身无法直接生成思维导图，目前的实践 workflow 基本是：DeepSeek生成Markdown格式文本，然后导入到支持Markdown转思维导图的工具比如Xmind、MindMaster等。

## 步骤一

DeepSeek生成Markdown格式文本



## 步骤二

复制Markdown文本并保存到本地：

Win电脑可以先保存为.txt再改格式为.md

Mac电脑可以用Paper保存为.md文件



## 步骤三

打开Xmind/MindMaster，在【文件】菜单中找到【导入】，导入我们在步骤二保存好的.md文件





# DeepSeek 绘图场景二：绘制流程图、时序图等图形图表

DeepSeek 同样不能直接生成流程图、时序图等图形图表，目前的实践 workflow 基本是：DeepSeek 生成 Mermaid 格式文本，然后导入到 draw.io 这类综合绘图工具中。

## 步骤一

DeepSeek 生成 Mermaid 格式文本，复制这段 Mermaid 文本

请生成用户在社交平台注册登录的完整流程图，以 Mermaid 格式输出

以下是用户在社交平台注册登录的完整流程图，使用 Mermaid 格式表示：

```

mermaid
graph TD
 A[用户访问社交平台] --> B{是否已注册?}
 B -- 是 --> C[显示登录页面]
 B -- 否 --> D[显示注册页面]

 %% 注册流程
 D --> E[用户填写注册信息]
 E --> F{信息是否有效?}
 F -- 是 --> G[发送验证码]
 F -- 否 --> H[提示用户重新填写]
 G --> I[用户输入验证码]
 I --> J{验证码是否正确?}
 J -- 是 --> K[创建用户账户]
 J -- 否 --> L[提示验证码错误]
 K --> M[注册成功, 跳转到登录页面]

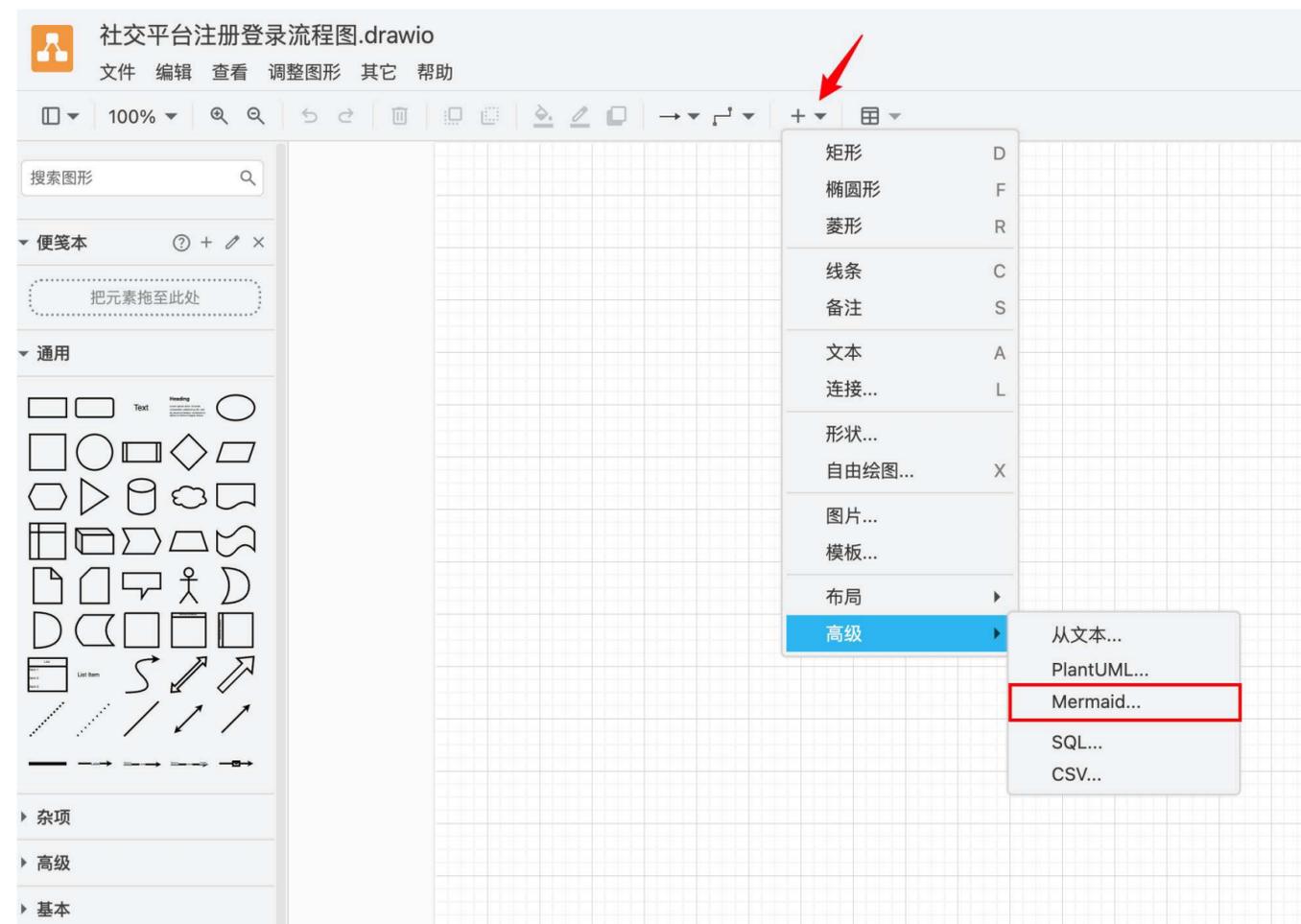
 %% 登录流程
 C --> N[用户输入登录信息]
 N --> O{信息是否正确?}
 O -- 是 --> P[登录成功, 跳转到主页]
 O -- 否 --> Q[提示登录信息错误]
 Q --> N

 %% 忘记密码流程
 C --> R[用户点击“忘记密码”]

```

## 步骤二

打开 draw.io，在【+】列表依次点击【高级-Mermaid】，在打开的面板中粘贴在步骤一复制好的 Mermaid 文件





## DeepSeek 绘图场景三：制作SVG矢量图

DeepSeek支持生成XML格式文本，而SVG就是一种用XML定义的语言，用来描述矢量图形。因此DeepSeek也可以用来生成SVG代码（本质上是代码能力的体现），生成的代码可以直接点击右下角的“运行HTML”按钮，在DeepSeek上直接查看效果（注意：直接生成SVG的效果可能不稳定，可以考虑结合iconfont的SVG代码能力实现更精准控制）。

请生成一个蓝鲸logo的svg矢量图，以svg代码格式输出

< 1/2 >



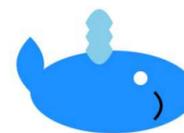
以下是蓝鲸Logo的SVG矢量图代码：

```
xml 复制

<svg xmlns="http://www.w3.org/2000/svg" viewBox="0 0 200 200" width="200" height="200">
 <!-- 蓝鲸身体 -->
 <ellipse cx="100" cy="100" rx="60" ry="30" fill="#1E90FF" />
 <!-- 蓝鲸尾巴 -->
 <path d="M 40 100 Q 20 80, 40 60 Q 60 80, 40 100 Z" fill="#1E90FF" />
 <!-- 蓝鲸眼睛 -->
 <circle cx="120" cy="90" r="5" fill="#FFFFFF" />
 <!-- 蓝鲸嘴巴 -->
 <path d="M 130 100 Q 140 110, 130 120" stroke="#000000" stroke-width="2" fill="none" />
 <!-- 蓝鲸喷水 -->
 <path d="M 100 70 Q 90 50, 80 70 Q 90 90, 100 70 Z" fill="#87CEEB" />
 <path d="M 100 60 Q 90 40, 80 60 Q 90 80, 100 60 Z" fill="#87CEEB" />
 <path d="M 100 50 Q 90 30, 80 50 Q 90 70, 100 50 Z" fill="#87CEEB" />
</svg>
```

▶ 运行 HTML

DeepSeek



内容由 AI 生成，请仔细甄别



# DeepSeek 绘图场景四：Photoshop图片批量处理

用DeepSeek来实现Photoshop图片批量处理，本质上是DeepSeek代码能力的体现，目前的工作流基本是：用DeepSeek生成一段可执行的代码脚本，然后导入到Photoshop中运行。

需要注意的是：这种方法出现代码报错的概率比较高，虽然报错也能发给DeepSeek帮忙解决，但对于不熟悉Photoshop和代码的小伙伴来说效率比较低，而对于熟悉Photoshop的小伙伴来说提效可能不太明显。

## 步骤一

DeepSeek生成代码脚本（Javascript）

帮我写一条适用于photoshop2024的脚本代码，结合创成式填充，给老照片修复清晰，并且上色，注意细节

要在 Photoshop 2024 中编写一个脚本，结合创成式填充（Generative Fill）来修复老照片并上色，你可以使用 Photoshop 的 JavaScript 脚本功能。以下是一个示例脚本，它可以帮助你自动化这个过程。请注意，这个脚本假设你已经熟悉 Photoshop 的基本操作，并且已经安装了支持创成式填充的插件或工具。

```

javascript
// Photoshop 2024 Script for Old Photo Restoration and Coloring

// Ensure Generative Fill is available
if (!app.plugins.exists("GenerativeFill")) {
 alert("Generative Fill plugin is not installed. Please install it to proceed.");
} else {
 // Open the old photo
 var oldPhoto = app.open(File("path/to/your/old/photo.jpg"));

 // Duplicate the background layer
 var originalLayer = oldPhoto.activeLayer;
 var duplicateLayer = originalLayer.duplicate();
 oldPhoto.activeLayer = duplicateLayer;

 // Apply Generative Fill to restore details
 app.plugins.apply("GenerativeFill", {
 mode: "restore",
 strength: 50 // Adjust strength as needed
 });
}

```

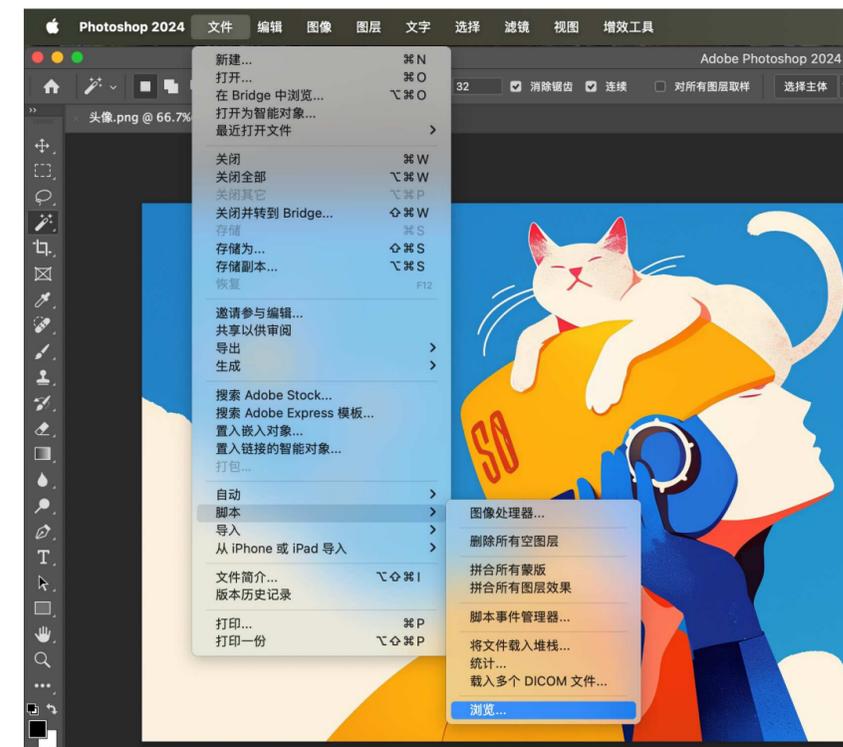
## 步骤二

复制Javascript代码并保存到本地：  
Win电脑可以先保存为.txt再改格式为.jsx  
Mac电脑可以用VScode保存为.jsx



## 步骤三

打开Photoshop，在【文件】菜单中依次点击【脚本-浏览】，然后打开我们在步骤二保存好的.jsx文件





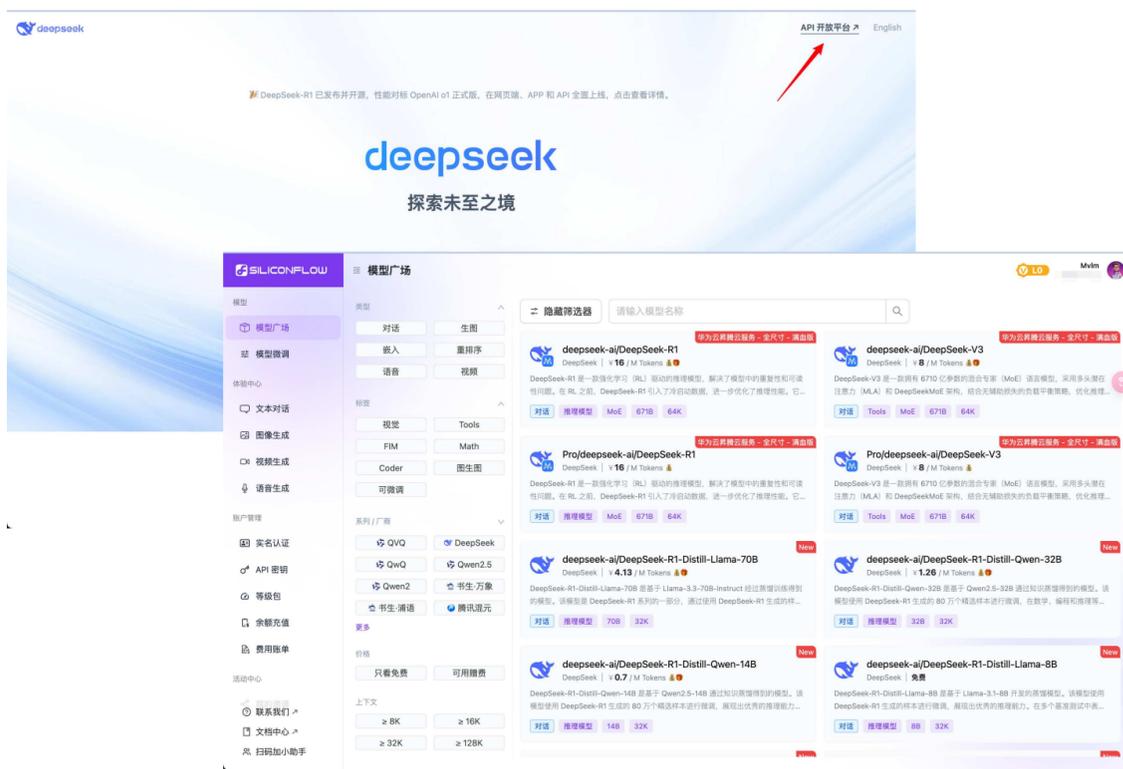
# DeepSeek API 场景一：接入Word、WPS等文档软件

Word、WPS支持VBA和插件开发，因此目前接入DeepSeek API的工作流也分两种：一种是让DeepSeek给出VBA脚本，然后根据VBA配置流程推进；一种是安装第三方插件如“OfficeAI助手”省去VBA配置流程。

需要注意的是：第一种方法出现代码报错的概率比较高，第二种方法则不支持Mac，如果不是刚需场景，可能噱头大于实用性。

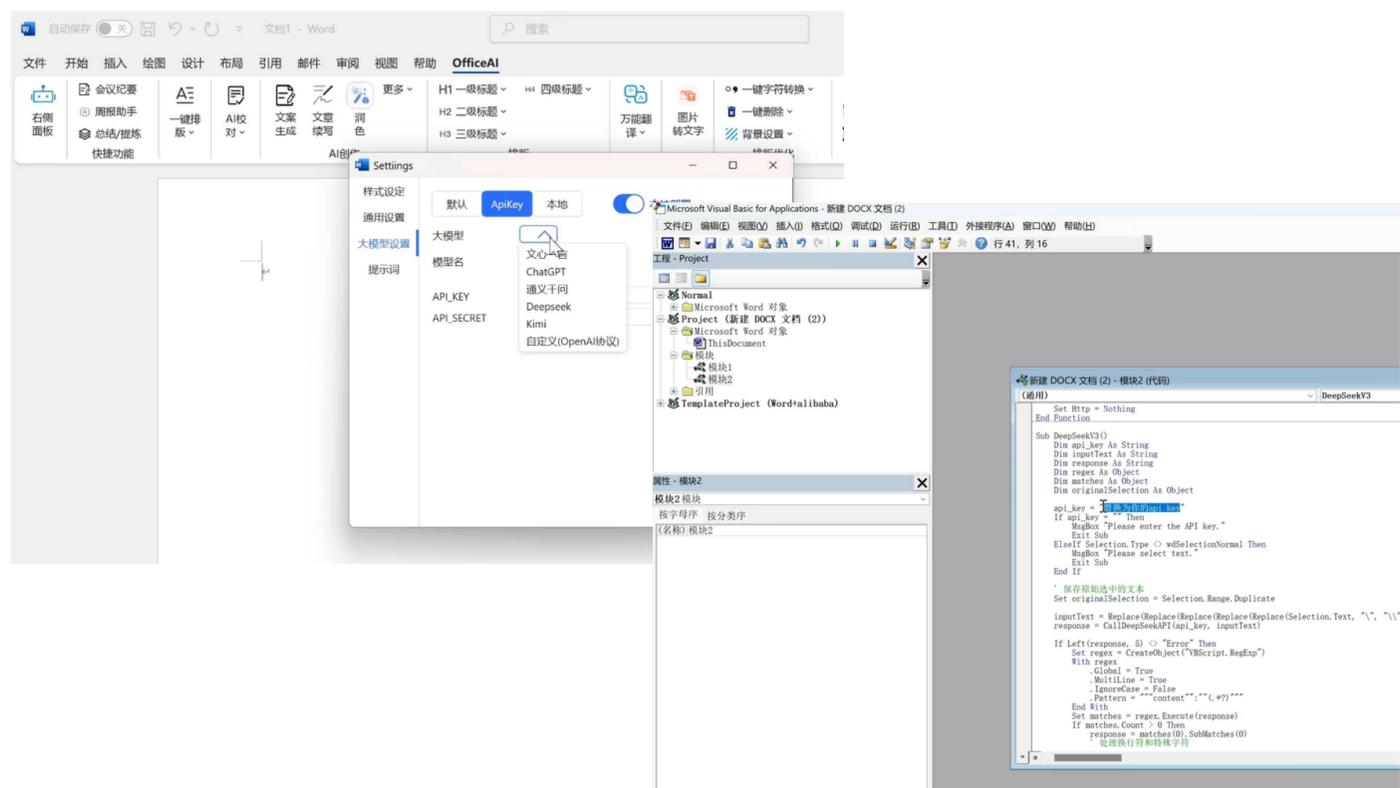
## 步骤一

从DeepSeek官网或者硅基流动获取DeepSeek API



## 步骤二

在Word或者WPS中配置VBA，或借助“OfficeAI助手”进行API配置





# DeepSeek API 场景二：接入Obsidian、FastGPT等AI知识库

除了接入Obsidian、FastGPT等AI知识库，还可以接入Dify、沉浸式翻译等工具，更多场景与应用案例接入教程可参考硅基流动API场景示例：<https://docs.siliconflow.cn/usercases/awesome-user-cases>

视频系列 >

平台系列 >

**场景示例**

- 在 Bob 翻译中使用
- 在沉浸式翻译中使用
- 在 Chatbox 中使用
- 在 Dify 中使用
- 在 Sider 中使用
- 在 NextChat 中使用
- 在 MindSearch 中使用
- 在 Cherry Studio 中使用
- 在 Obsidian Copilot 中使用
- 在 DB-GPT 中使用
- 在 FastGPT 中使用**
- 在麦悠电台中使用
- 在 302.AI 中使用
- 社区场景与应用

常见问题

- 错误处理
- 流式输出
- 实名认证
- 开具发票

Q 搜索... ⌕

联系我们 体验中心 >

## 场景示例

### 在 FastGPT 中使用

! 本文转载自 [FastGPT](#) 的官方文档，介绍了如何在 FastGPT 中使用 SiliconCloud 的模型。[原文地址](#)

**SiliconCloud(硅基流动)** 是一个以提供开源模型调用为主的平台，并拥有自己的加速引擎。帮助用户低成本、快速的进行开源模型的测试和使用。实际体验下来，他们家模型的速度和稳定性都非常不错，并且种类丰富，覆盖语言、向量、重排、TTS、STT、绘图、视频生成模型，可以满足 FastGPT 中所有模型需求。

如果你想部分模型使用 SiliconCloud 的模型，可额外参考[OneAPI接入硅基流动](#)。

本文会介绍完全使用 SiliconCloud 模型来部署 FastGPT 的方案。

### 1. 注册 SiliconCloud 账号

- [点击注册硅基流动账号](#)
- 进入控制台，获取 API key: <https://cloud.siliconflow.cn/account/ak>

### 2. 修改 FastGPT 环境变量

```
OPENAI_BASE_URL=https://api.siliconflow.cn/v1
填写 SiliconCloud 控制台提供的 Api Key
CHAT_API_KEY=sk-xxxxxx
```

在此页面

- [注册 SiliconCloud 账号](#)
- [修改 FastGPT 环境变量](#)
- [修改 FastGPT 配置文件](#)
- [重启 FastGPT](#)
- [体验测试](#)

- 测试对话和图片识别
- 测试知识库导入和知识库问答
- 测试语音播放
- 测试语言输入

总结



## DeepSeek API 场景三：接入硬件

除了可以将DeepSeek接入Word、WPS等软件，还可以接入iPhone、安卓手机、智能音箱（小爱同学）、电脑等智能硬件，这些基本都属于第六部分中的本地部署。这里就先不展开。





# DeepSeek 其它场景一：自我探索 and 发现

如果大家愿意的话，可以将自己的一些困惑告诉DeekSeek R1，这个时候它会尝试模拟并推理我们的内在心理，往往能够发现一些我们“不自知”的问题并给到建议方案。

我好像有点害怕亲密关系，具体表现在我可能愿意和家人，和爱人待在一起，但又不能待一起太久，又表现在我很容易和家人，和爱人闹别扭，能帮我分析下吗



已深度思考 (用时 25 秒) ^

嗯，用户说他好像有点害怕亲密关系，具体表现是愿意和家人、爱人待在一起，但时间不能太久，还容易和他们闹别扭。我需要先理解他的情况，然后分析可能的原因，再给出建议。

首先，他愿意和家人、爱人相处，说明他并非完全排斥亲密关系，而是有矛盾心理。不能待太久和容易闹别扭可能反映出他对亲密关系中的某些方面感到不安。可能的原因有很多，比如害怕失去自我、害怕被控制、过去的创伤、或者对冲突的敏感。

接下来要考虑的是依恋类型。他可能属于焦虑型或回避型依恋。焦虑型的人可能会担心被抛弃，而回避型的人则可能害怕过于亲密。需要分析他的具体行为，比如闹别扭的时候是不是因为对方太过靠近，或者自己感觉被束缚了。

另外，他提到的闹别扭可能源于沟通问题。当亲密关系中出现压力时，他可能不知道如何有效表达需求，转而用闹别扭的方式，这可能导致关系紧张。需要评估他的沟通模式，是否有回避冲突或攻击性反应的情况。

家庭环境的影响也很重要。如果成长过程中父母关系紧张，或者有过被忽视的经历，可能会影响他现在对亲密关系的态度。他可能在潜意识里重复过去的模式，比如通过闹别扭来获取关注，或者保护自己不受伤害。

边界感的问题也是一个方面。他可能在关系中难以平衡亲密和独立，当感觉自己的空间被侵犯时，就会产生抗拒。需要探讨他是否清楚自己的边界，以及如何健康地维护这些边界。

建议方面，可能需要建议他自我反思，比如记录触发情绪的事件，识别自己的恐惧来源。同时，推荐专业咨询，因为深入的问题可能需要专业人士的帮助。沟通技巧的学习也很重要，比如非暴力沟通。建立健康的边界，逐步暴露自己于亲密环境中，可能会帮助他适应。另外，自我关怀，减少自我批判，接纳自己的感受也是关键。

需要确认是否有其他症状，比如焦虑或抑郁，但用户没有提到，所以暂时不深入。要确保建议具体可行，比如写情绪日记，练习沟通技巧的方法，寻找咨询师的途径等。

最后，要给予鼓励，让他知道这种情况常见，可以改善，增强他的信心。

你的情况反映出对亲密关系存在一定的矛盾心理：既渴望连接，又害怕过度卷入。这种模式可能与以下心理因素有关，我们可以逐步分析并寻找调整方向：

## 1. 内在安全感冲突

表现及原因：主动靠近后又主动疏离，本质是「依恋系统」和「防御系统」同时激活，当安全感达到某

给 DeepSeek 发消息

深度思考 (R1) 联网搜索



内容由 AI 生成，请仔细甄别

问我十个问题，以最大程度帮助我了解我自己。



已深度思考 (用时 12 秒) ^

嗯，用户让我问十个问题来最大程度地帮助他了解自己。首先，我需要确定这些问题的类型和方向。用户可能希望自我反思，所以问题需要涵盖不同的生活层面，比如价值观、成就、挑战、人际关系等等。

首先，我应该从核心价值开始，因为价值观是行为的驱动力。然后，可能需要了解他过去的成就和遗憾，这些能反映他的优先事项和未满足的愿望。接下来，人际关系和他人评价也很重要，因为外部视角和自我认知可能有差异。

然后，挑战和恐惧的部分可以帮助他识别内心的障碍。理想生活的问题可以激发他对未来的想象，而自我评价和他人评价的对比有助于自我认知的一致性。习惯和模式的问题能揭示潜在的行为机制，而人生转折点则可能指出关键决策或经历的影响。

最后，关于自我接受和改变的部分可以帮助他平衡现状与成长。我需要确保问题之间有一定的逻辑顺序，从过去到现在，再到未来，逐步深入。同时，问题需要开放，避免封闭式回答，让他有足够的空间展开思考。可能还需要调整问题的表述，使其更自然，避免过于生硬。检查是否有重复的主题，确保每个问题都覆盖不同的维度，这样才能全面帮助他了解自己。

好的！以下是10个能帮助你深入探索自我的问题。回答时尽量具体，并给自己足够的时间反思：

## 1. 价值观核心

你人生中最重要3个价值观是什么？（例如：自由、创造、家庭）这些价值观如何体现在你过去的重要决定中？

## 2. 成就与遗憾

回顾过去5年，哪件事让你感到最自豪？哪件事至今仍让你感到遗憾或未完成？为什么？

## 3. 人际关系反馈

如果让最了解你的3个人（朋友、家人、同事）描述你，你认为他们会如何评价你？其中是否有你从未意识到的特质？

## 4. 恐惧与逃避

你目前最想逃避或拖延的事情是什么？背后隐藏着怎样的恐惧或自我怀疑？

给 DeepSeek 发消息

深度思考 (R1) 联网搜索



内容由 AI 生成，请仔细甄别



# DeepSeek 其它场景二：乙游（恋爱模拟游戏）

这种恋爱模拟游戏的剧本，就是得益于DeepSeek强大的中文能力和推理能力。

请设计一个校园恋爱文字游戏，要求如下：  
设计三个性格、外形不同的男性角色，与他们接触能分别获得不同的故事线，并深入发展。  
设计好后直接开始游戏吧！



窗 已深度思考 (用时 14 秒) ▾

(游戏启动画面：樱花飘落的校园中传来悠扬钢琴声)

【系统】请为女主角命名：(输入姓名后自动生成学生证)

【第一幕：开学日】

你抱着书本穿过走廊时，三个声音同时响起——

①前方传来清冷声线："同学，你的校徽戴反了"

(镜头推进) 银灰短发的高挑男生推了推金丝眼镜，胸前的金色学生会徽章折射着晨光

②右侧画室突然打开窗户："要当我的缪斯吗？"

栗色卷发扎成小揪的男生举着沾满颜料的画笔，耳骨钉在阳光下闪动，白色衬衫沾着水彩

③身后传来书本坠地声，转头看见蹲着捡书的男生

深蓝发丝垂落遮住眉眼，制服纽扣严谨地系到最上方，指尖在碰到你伸出的手时触电般缩回

【角色档案解锁】

▷顾明澈 (学生会会长/经济学系大三)

特质：理性毒舌/完美主义/钢琴十级

解锁地点：学生会办公室/天台

▷林星野 (美术特长生/油画系大二)

特质：浪漫主义/怕寂寞/夜行动物

解锁地点：画室/后山枫林

▷沈砚书 (科研助手/物理系研一)

特质：社交障碍/天文爱好者/猫奴

解锁地点：图书馆/实验楼

【首次选择分支】

给 DeepSeek 发送消息

深度思考 (R1)

联网搜索



内容由 AI 生成，请仔细甄别

【首次选择分支】

现在要去：

A. 学生会提交社团申请表 (顾明澈线)

B. 画室归还颜料箱 (林星野线)

C. 图书馆找参考书 (沈砚书线)

(选择后将进入专属剧情，以下示例选择C线)

【沈砚书线：图书馆事件】

你踮脚取书时，他忽然从书架后现身："那本《量子宇宙》的第三章公式有印刷错误..."突然意识到距离过近，耳尖泛红后退撞倒书堆。

▶选项：

1. "学长经常一个人在这里看书吗？"

2. 默默蹲下帮他整理散落的论文

3. 指着窗外："快看！UFO！"

(选择2触发特殊事件)

你的手指碰到他正在捡的论文，发现满页都是猫咪涂鸦。他慌乱地抢夺时，你们的手指交叠在「薛定谔的猫」示意图上。

【好感度系统激活】

沈砚书好感+5 (当前5/100)

获得【沾着咖啡渍的书签】

【次日随机事件】

实验楼传来玻璃碎裂声，循声看见他抱着受伤的流浪猫不知所措。白大褂染血的画面与昨夜论文上的猫爪印重叠...

\*每条线路包含：

• 7个主线章节

• 3个隐藏约会场景

• 12种分支结局 (含黑化/友情/恋爱线)

• 角色专属BGM&动态CG

给 DeepSeek 发送消息

深度思考 (R1)

联网搜索



内容由 AI 生成，请仔细甄别

要切换攻略对象吗？当前可回溯至【首次选择分支】...

继续攻略沈砚书



窗 已深度思考 (用时 21 秒) ▾

(选择继续攻略沈砚书线，加载专属BGM《星屑ノ一ト》钢琴版)

【第三章：观测日志】

深夜从实验室归来的你，发现天台闪烁着手电筒光斑。沈砚书正用裹着绷带的手调试天文望远镜，脚边蜷着包扎好的三花猫。发现你时差点碰倒柠檬茶易拉罐塔："只是...验证开普勒定律..."

▶选项：

1. 指着星空："那颗最亮的是木星吗？"

2. 掏出随身携带的猫罐头

3. 突然凑近看他发红的眼角

(选择1触发天文讲座事件)

他瞬间进入学术模式："实际木星此刻应该位于... (突然停顿) 你头发上有星屑。"指尖悬停在你发间时，夜航飞机掠过银河。

【好感度+10】获得【手绘星图便签】

\*\*\*

【第四章：测不准心动】

他破天荒主动拦住你："校际物理竞赛需要...实验助手。"递来的报名表边角有反复折叠的痕迹。你们在实验室调试磁控管时，他的白大褂突然飘来清甜的柚子香——背后贴着「防静电专用柔顺剂」标签。

▶突发状况：

老式示波器爆出火花，他本能护住你时，薄荷气息的呼吸扫过你手中的电路图："这个谐振频率...算错了。"

▶选项：

1. 伸手抚平他翘起的发梢

给 DeepSeek 发送消息

深度思考 (R1)

联网搜索



内容由 AI 生成，请仔细甄别



06

**DeepSeek替代方案  
(在线&本地部署)**



# DeepSeek 现状

截至2025年2月10日，受注册人数激增，以及外部恶意攻击的双重影响，DeepSeek官方APP和Web应用时不时出现各种卡慢、服务器繁忙的情况，官网的API申请入口目前也是暂停充值状态。

为了让更多用户能用上DeepSeek，越来越多平台，包括国家超算中心、大模型云平台、手机厂商、AI搜索引擎、车企、电信运营商等，陆续接入了DeepSeek（其中部分提供API服务），所以大家目前有了更多DeepSeek的替代选择（在线或本地部署）。

服务器繁忙，请稍后再试。

< 6/6 >

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

内容由 AI 生成，请

当前服务器资源紧张，为避免对您造成业务影响，我们已暂停 API 服务充值。存量充值金额可继续调用，敬请谅解！

deepseek 开放平台

- 用量信息
- API keys
- 账单

开启新

### 用量信息

deepseek-chat 模型优惠期结束，调用价格已变更为每百万输入 tokens 2 元，每百万输出 tokens 8 元。【[查看价格详情](#)】

所有日期均按 UTC 时间显示，数据可能有 5 分钟延迟。

充值余额	赠送余额 <a href="#">查看有效期</a>	本月消费
¥0.00 CNY	¥0.00 CNY	¥0.00 CNY

去充值 余额预警未开启 [\(去设置\)](#)

### 每月用量



# DeepSeek 替代方案一：秘塔AI搜索

**特点：**支持联网搜索，提供长思考·R1模型。

**访问方式：**访问秘塔搜索官网，选择“主页”模块，勾选“长思考·R1”直接提问。

**使用建议：**适合需要联网搜索的用户，支持上传图片 and 文件。



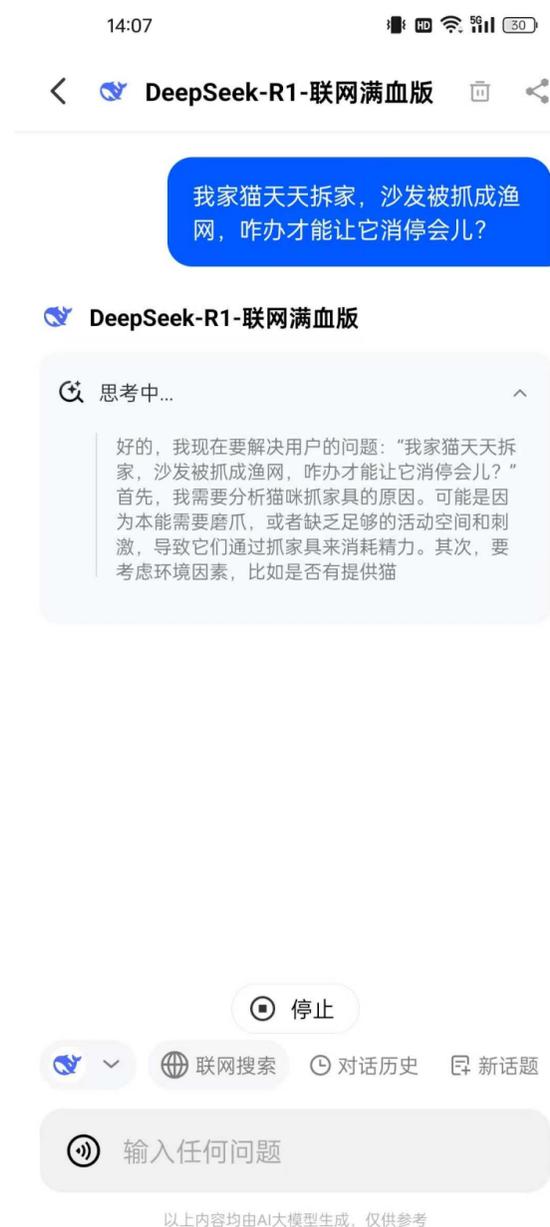


# DeepSeek 替代方案二：纳米AI搜索

特点：响应速度快，支持手机号注册登录。

版本选择：提供DeepSeek-R1(360高速专线)和DeepSeek-R1-满血版（671B模型）。

使用建议：适合手机用户，随时随地提问，响应速度快。





# 🤝 DeepSeek 替代方案三：硅基流动

特点：大模型云服务平台，同时提供在线使用DeepSeek和DeepSeek API申请（官网：[硅基流动](#)）。

适用建议：有本地部署需求的用户，目前基本都是在硅基流动申请API，然后结合Chatbox、Cherry Studio等客户端实现。

硅基流动 x 华为云联合推出基于昇腾云的 DeepSeek R1 & V3 推理服务，注册即可体验→

SILICONFLOW

🌐 简 ▾ 产品 ▾ 价格 文档 ↗ 博客 ↗ 联系我们 关于 [Log in](#)

## 硅基流动 x 华为云联合推出基于昇腾云的 DeepSeek R1&V3 推理服务!

即刻尝鲜

### 快速高效的GenAI推理软件栈

通过SiliconFlow技术栈，提高GenAI应用的开发效率，显著降低使用成本

5年2月6日起，未实名用户每日限制请求 DeepSeek V3/R1 模型各 100 次

V LO MvIm

**deepseek-ai/DeepSeek-R1**

Seek | ¥16 / M Tokens

是一款强化学习 (RL) 驱动的推理模型，解决了模型中的重复性和可读之前，DeepSeek-R1 引入了冷启动数据，进一步优化了推理性能。它...

模型 MoE 671B 64K

华为云昇腾云服务 - 全尺寸 - 满血版

**deepseek-ai/DeepSeek-V3**

DeepSeek | ¥8 / M Tokens

DeepSeek-V3 是一款拥有 6710 亿参数的混合专家 (MoE) 语言模型，采用多头潜在注意力 (MLA) 和 DeepSeekMoE 架构，结合无辅助损失的负载均衡策略，优化推理...

对话 Tools MoE 671B 64K

华为云昇腾云服务 - 全尺寸 - 满血版

**deepseek-ai/DeepSeek-R1**

Seek | ¥16 / M Tokens

是一款强化学习 (RL) 驱动的推理模型，解决了模型中的重复性和可读之前，DeepSeek-R1 引入了冷启动数据，进一步优化了推理性能。它...

模型 MoE 671B 64K

华为云昇腾云服务 - 全尺寸 - 满血版

**Pro/deepseek-ai/DeepSeek-V3**

DeepSeek | ¥8 / M Tokens

DeepSeek-V3 是一款拥有 6710 亿参数的混合专家 (MoE) 语言模型，采用多头潜在注意力 (MLA) 和 DeepSeekMoE 架构，结合无辅助损失的负载均衡策略，优化推理...

对话 Tools MoE 671B 64K

华为云昇腾云服务 - 全尺寸 - 满血版

**deepseek-ai/DeepSeek-R1-Distill-Llama-70B**

Seek | ¥4.13 / M Tokens

Distill-Llama-70B 是基于 Llama-3.3-70B-Instruct 经过蒸馏训练得到是 DeepSeek-R1 系列的一部分，通过使用 DeepSeek-R1 生成的样...

模型 70B 32K

New

**deepseek-ai/DeepSeek-R1-Distill-Qwen-32B**

DeepSeek | ¥1.26 / M Tokens

DeepSeek-R1-Distill-Qwen-32B 是基于 Qwen2.5-32B 通过知识蒸馏得到的模型。该模型使用 DeepSeek-R1 生成的 80 万个精选样本进行微调，在数学、编程和推理等...

对话 推理模型 32B 32K

New

**deepseek-ai/DeepSeek-R1-Distill-Qwen-14B**

Seek | ¥0.7 / M Tokens

Distill-Qwen-14B 是基于 Qwen2.5-14B 通过知识蒸馏得到的模型。该 Seek-R1 生成的 80 万个精选样本进行微调，展现出优秀的推理能力...

模型 14B 32K

New

**deepseek-ai/DeepSeek-R1-Distill-Llama-8B**

DeepSeek | 免费

DeepSeek-R1-Distill-Llama-8B 是基于 Llama-3.1-8B 开发的蒸馏模型。该模型使用 DeepSeek-R1 生成的样本进行微调，展现出优秀的推理能力。在多个基准测试中表...

对话 推理模型 8B 32K

New



# DeepSeek 替代方案四：国内大厂的云服务&产品

截至2025年2月10日，国内大厂已经接入 DeepSeek 的产品包括：

**腾讯云**：2月2日宣布支持R1，目前已上线全尺寸DeepSeek模型

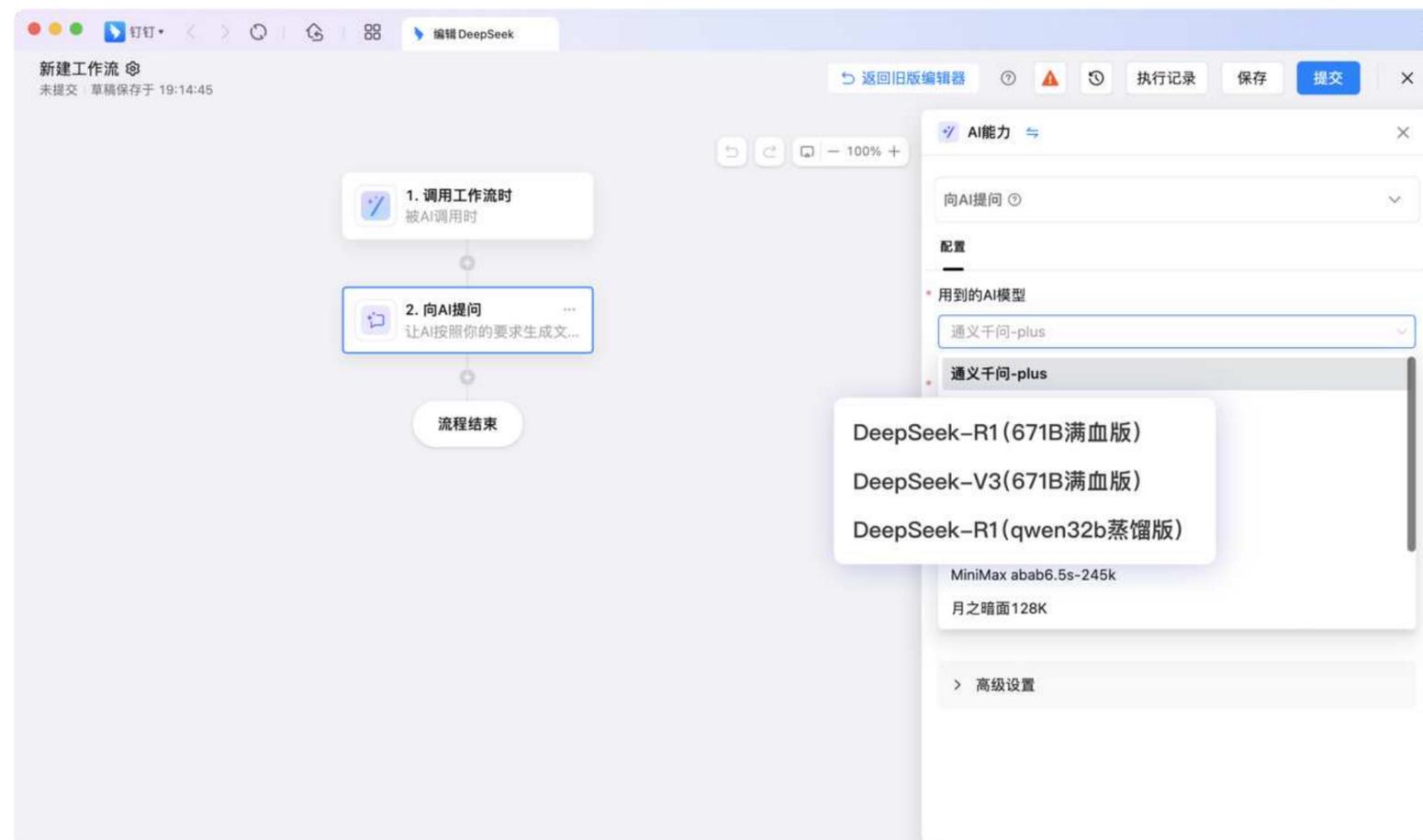
**百度云千帆**：2月3日，宣布已正式上架 DeepSeek-R1和 DeepSeek-V3模型

**字节跳动火山引擎**：2月4日，宣布全面支持 DeepSeek 系列模型

**阿里钉钉**：2月6日，全面接入DeepSeek系列模型

**字节跳动Coze**、**飞书多维表格**：2月7日，宣布支持满血版 Deepseek

**阿里云百炼**：2月9日，上线全尺寸DeepSeek模型





# DeepSeek 替代方案五：国内手机厂商

截至2025年2月10日，已经接入或宣布接入的手机厂商包括：

**华为：** 华为Harmony NEXT的小艺助手已经接入 DeepSeek-R1，用户可通过小艺助手直接调用。

**荣耀：** 荣耀的YOYO助手也已接入 DeepSeek-R1，用户需将系统升级至 MagicOS 8.0 及以上版本，之后通过 YOYO 助手调用。

**OPPO：** 宣布OPPO Find N5折叠屏手机将于 2 月底发布，成为OPPO 首批适配 DeepSeek-R1 的机型。

**魅族：** 宣布魅族 21 系列和 Lucky 08 手机将于下周率先上线 DeepSeek-R1。

**vivo：** vivo 也已确认接入 DeepSeek-R1。

## 发现



 智能体

 小艺技能



**小艺照相馆**  
帮我生成创意头像  
小艺官方



**DeepSeek-R1**  
(Beta)

**DeepSeek-R1**  
深度思考，精准推理  
小艺实验室



# DeepSeek 替代方案：更多国内产商

除了前面提到的腾讯、字节、阿里、百度等大厂，其实还有更多国内产商加上了这场DeepSeek支持的浪潮，因此这里再单独新建一页进行罗列，整理的信息（截至2025年2月10日）难免挂一漏万，大家见谅。

企业/产品	接入情况	企业/产品	接入情况
PPIO 派欧算力云	已全面集成 DeepSeek R1	阅文集团	旗下作家辅助创作产品——作家助手已集成独立部署的DeepSeek-R1大模型
<a href="#">国家超算互联网</a>	DeepSeek-R1、V3、Coder等系列模型，已陆续上线	天翼云（中国电信）	天翼云自主研发的“息壤”智算平台完成国产算力与DeepSeek-R1/V3系列大模型的深度适配优化
京东云	言犀AI开发计算平台已经支持 DeepSeek-V3、DeepSeek-R1 以及所有蒸馏小参数模型	金蝶云	已将DeepSeek全面集成至金蝶云全线SaaS应用及金蝶云苍穹平台
联通云（中国联通）	联通云已基于星罗平台实现国产及主流算力适配多规格DeepSeek-R1模型	浪潮云	发布预置DeepSeek-R1和DeepSeek-V3模型的海若一体机，实现671B DeepSeek大模型国产GPII单机推理服务
移动云（中国移动）	移动云已支持DeepSeek全版本、全尺寸、全功能	七牛云	已集成 DeepSeek-R1 671b 全参数模型
优刻得	基于国产芯片适配DeepSeek全系列模型	商汤	企业客户和开发者可在商汤大装置万象平台中，快速部署DeepSeek-V3、DeepSeek-R1等模型
北京超算	北京超算AI智算云平台完成DeepSeek-R1满血版（671B）模型服务部署	昆仑万维	旗下天工AI正式上线DeepSeek R1+联网搜索



# 🤝 DeepSeek 替代方案六：AI编程工具

目前比较热门的AI编程工具 Cursor 和 Windsurf 均已接入 DeepSeek V3 和 DeepSeek R1，不过需要会员才能调用（消耗的还是 slow requests的额度）。如果需要的话，也可以通过API的方式接入。

The image shows a composite view. On the left is the Windsurf website's 'Models' page, which lists various AI models. A red box highlights 'DeepSeek-V3' and 'DeepSeek-R1'. On the right is a screenshot of a code editor (Cursor) with a chat window open. The chat window shows a list of models, with 'deepseek-r1' and 'deepseek-v3' highlighted by a red box. The code editor displays JavaScript code for a Chrome extension, including functions for creating a context menu and saving text to local storage.

**Windsurf Models**

Cascade has several different models for you to choose from. In the Cascade panel `Ctrl/⌘ + L`, you can easily switch between the following models available:

- GPT-4o
- Claude 3.5 Sonnet
- **DeepSeek-V3**
- **DeepSeek-R1**
- o3-mini (medium reasoning)
- Gemini 2.0 Flash
- Cascade Base ⚡ - Free!

**Premium Models**

Due to the wide variation between the serving costs you select, each of your input prompts will consume **credits**.

Below you will find a breakdown of how many User

**Code Editor (Cursor) Chat Window Models:**

- claude-3.5-sonnet
- gpt-4o
- gpt-4o-mini
- o1-mini
- o1-preview
- o1
- o3-mini
- gemini-2.0-flash-exp
- **deepseek-r1**
- **deepseek-v3**
- claude-3.5-haiku

```
JS background.js > ...
1 // 创建右键菜单
2 chrome.runtime.onInstalled.addListener(() => {
3 chrome.contextMenus.create({
4 id: "saveText",
5 title: "剪存文字",
6 contexts: ["selection"]
7 });
8 });
9
10 // 处理右键菜单单击
11 chrome.contextMenus.onClicked.addListener((info, tab) => {
12 if (info.menuItemId === "saveText") {
13 const selectedText = info.selectionText;
14 // 获取现有存储的文本
15 chrome.storage.local.get(['savedTexts'], function(result) {
16 let texts = result.savedTexts || [];
17 texts.push(selectedText);
18 });
19 // 保存更新后的文本
20 chrome.storage.local.set({
21 savedTexts: texts
22 });
23 }
24 });
25
26 }
```



# DeepSeek 替代方案：海外产品

一些海外产品也集成了DeepSeek，这里也进行了部分整理：

企业/产品			
英伟达NIM	Cerebars	亚马逊AWS	Github
Grog	Perplexity	微软Azure	Cline
Fireworks	Cursor	Lambda	
Chutes	Windsurf	POE	



# DeepSeek 替代方案：本地部署（设备配置要求表一）

如果大家比较在意个人隐私和信息安全，那么可以考虑本地部署。但需要注意的是，大部分人的电脑设备都不足以支持 DeepSeek R1 满血版（671b）甚至一些大的 R1 蒸馏版的本地部署。

因此这里提供了 R1 全系列本地部署的配置要求表（注意：这里的配置表都是偏高的），大家先判断自己的设备是否支持，而不是兴冲冲地部署到一半，发现自己设备支持不了。

对比项	DeepSeek-R1-1.5B	DeepSeek-R1-7B	DeepSeek-R1-8B	DeepSeek-R1-14B
CPU	最低 4 核（推荐 Intel/AMD 多核处理器）	8 核以上（推荐现代多核 CPU）		12 核以上
内存	8GB+	16GB+		32GB+
硬盘	3GB+ 存储空间（模型文件约 1.5-2GB）	8GB+（模型文件约 4-5GB）	硬件需求与 7B 相近，略高 10-20%	16GB+
显卡	非必需（纯 CPU 推理），若 GPU 加速可选 4GB+ 显存（如 GTX 1650）	推荐 8GB+ 显存（如 RTX 3070/4060）		16GB+ 显存（如 RTX 4090 或 A5000）
场景	低资源设备部署（如树莓派、旧款笔记本） 实时文本生成（聊天机器人、简单问答） 嵌入式系统或物联网设备	本地开发测试（中小型企业） 中等复杂度 NLP 任务（文本摘要、翻译） 轻量级多轮对话系统	需更高精度的轻量级任务（如代码生成、逻辑推理）	企业级复杂任务（合同分析、报告生成） 长文本理解与生成（书籍/论文辅助写作）



# DeepSeek 替代方案：本地部署（设备配置要求表二）

选择合适的DeepSeek版本不仅要考虑硬件配置，还要根据实际应用场景来决定。

建议先从较小的模型开始尝试，逐步升级到更大的模型。这样可以在确保性能的同时，避免资源浪费。

对比项	DeepSeek-R1-32B	DeepSeek-R1-70B	DeepSeek-R1-671B（满血版）
CPU	16 核以上（如 AMD Ryzen 9 或 Intel i9）	32 核以上（服务器级 CPU）	64 核以上（服务器集群）
内存	64GB+	128GB+	512GB+
硬盘	30GB+	70GB+	300GB+
显卡	24GB+ 显存（如 A100 40GB 或双卡 RTX 3090）	多卡并行（如 2x A100 80GB 或 4x RTX 4090）	多节点分布式训练（如 8x A100/H100）
场景	高精度专业领域任务（医疗/法律咨询） 多模态任务预处理（需结合其他框架）	科研机构/大型企业（金融预测、大规模数据分析） 高复杂度生成任务（创意写作、算法设计）	国家级/超大规模 AI 研究（如气候建模、基因组分析） 通用人工智能（AGI）探索



# DeepSeek 替代方案：本地部署（设备配置要求表三）

我用DeepSeek R1+联网功能，也提问了DeepSeek R1系列各个版本的配置要求，这里把回答一并附上：

本地部署deepseek r1各个版本需要怎样的电脑配置



已搜索到 49 个网页 >

已深度思考 (用时 28 秒) ^

好的，用户的问题是本地部署DeepSeek R1各个版本需要的电脑配置。我需要根据提供的搜索结果来整理这些信息。首先，查看各个网页的内容，特别是网页1、3、4、6、8、9、10，因为它们都提到了不同版本的硬件配置要求。

网页1、3、4看起来内容相似，可能都是CSDN的博客，详细列出了不同参数模型的配置。比如1.5B、7B、8B、14B、32B、70B、671B这些版本对应的CPU、内存、硬盘、显卡要求和适用场景。例如，1.5B需要4核CPU、8GB内存、3GB硬盘，而7B需要8核、16GB内存、8GB硬盘，显卡推荐8GB显存。

网页2和网页6提到了用户的实际部署体验，特别是网页6有更详细的硬件需求表格，比如7B模型在低配置笔记本上也能运行，但速度较慢。这可能需要说明虽然最低配置存在，但推荐配置更重要。

网页8和网页10提供了具体的部署步骤和配置建议，比如网页8提到需要A100或RTX 4090显卡，而网页10则分平台（Windows、macOS、Linux）列出了配置。比如Windows平台对于32B模型需要RTX 3090，而Linux则需要多卡。

网页9讨论了满血版和蒸馏版的区别，指出满血版需要极高的显存，而蒸馏版更适合家用，例如使用Ollama部署量化版本，显存需求降低。例如，8B模型在Ollama中可能只需要8GB显存，而32B需要24GB。

需要将这些信息整合，按模型版本分类，说明每个版本的硬件需求。还要注意不同来源之间的差异，比如网页6提到的显存需求可能更高，而网页9的Ollama量化版本需求较低。需要指出这些差异，并提供通用建议，如量化技术、推理框架的使用，以及云部署选项。

另外，用户可能需要知道从哪个版本开始适合他们的需求，比如建议从较小的模型开始。网页1和6都提到了这一点。还要注意不同平台的配置差异，比如macOS可能只能运行小模型，而Linux更适合大模型部署。

最后，总结通用建议，如量化优化、使用加速框架、考虑云服务等，这些在多个网页中都有提到，需要综合引用。

以下是本地部署DeepSeek-R1不同版本模型的硬件配置要求及适用场景的总结，综合了多个来源的推荐信息 1 3 8：

## 1. 轻量级模型（1.5B-8B）

适用场景：简单文本生成、聊天机器人、代码生成、多轮对话

- **1.5B**
  - CPU：4核（推荐Intel/AMD多核）
  - 内存：8GB+
  - 显存：纯CPU或4GB+（如GTX 1650）
  - 硬盘：3GB+

给 DeepSeek 发消息

深度思考 (R1) 联网搜索

内容由 AI 生成，请仔细甄别

以下是本地部署DeepSeek-R1不同版本模型的硬件配置要求及适用场景的总结，综合了多个来源的推荐信息 1 3 8：

## 1. 轻量级模型（1.5B-8B）

适用场景：简单文本生成、聊天机器人、代码生成、多轮对话

- **1.5B**
  - CPU：4核（推荐Intel/AMD多核）
  - 内存：8GB+
  - 显存：纯CPU或4GB+（如GTX 1650）
  - 硬盘：3GB+
  - 平台：树莓派、旧款笔记本、手机（MNN框架） 11
- **7B/8B**
  - CPU：8核（推荐i7/Ryzen 7）
  - 内存：16GB+（纯CPU需32GB+）
  - 显存：8GB+（推荐RTX 3070/4060）
  - 硬盘：8GB+
  - 优化：4-bit量化可将显存需求降至8GB 8 11

## 2. 中等规模模型（14B-32B）

适用场景：专业领域问答、长文本生成、企业级任务

- **14B**
  - CPU：12核以上（不推荐纯CPU）
  - 内存：32GB+（纯CPU需64GB+）
  - 显存：24GB+（如RTX 3090双卡）
  - 硬盘：30GB+
  - 部署：需多卡并行（如NVLink） 8
- **32B**
  - CPU：16核以上（i9/Ryzen 9）
  - 内存：64GB+
  - 显存：48GB+（如双A100 40GB）
  - 硬盘：70GB+
  - 平台：仅限Linux或企业级服务器 12

给 DeepSeek 发消息

深度思考 (R1) 联网搜索

内容由 AI 生成，请仔细甄别

## 3. 大规模模型（70B-671B）

适用场景：科研、金融预测、超大规模AI研究

- **70B**
  - CPU：32核服务器级
  - 内存：128GB+
  - 显存：80GB+（如4x RTX 4090或A100集群）
  - 硬盘：150GB+
  - 优化：需FP16/8-bit量化 8 10
- **671B（满血版）**
  - CPU：64核服务器集群
  - 内存：512GB+
  - 显存：640GB+（需多节点分布式）
  - 硬盘：1TB+
  - 部署建议：仅适用于国家级研究或云服务 1 11

## 平台适配建议

- **Windows**：适合1.5B-14B（需RTX 3060+显存）
- **macOS**：仅支持M1/M2芯片（16GB+统一内存，1.5B-7B） 12
- **Linux**：推荐部署32B+模型（支持多卡并行与NVLink） 10

## 通用优化策略

1. 量化技术：4-bit/8-bit量化可减少显存占用30%-50% 1 8。
2. 推理框架：使用vLLM、TensorRT或DeepSpeed加速推理 8 10。
3. 云部署：70B+模型建议租用云GPU（如AWS、阿里云PAI） 8 11。
4. 能耗管理：32B+模型需1000W+电源与高效散热系统 1。

## 新手建议

- 从轻量级开始：优先尝试1.5B或7B模型，熟悉部署流程 8 12。
- 工具推荐：使用Ollama或LM Studio简化部署（支持量化模型自动适配） 11 12。
- 避免误区：固态硬盘虚拟内存方案会大幅降低性能，不推荐 11。

如需具体部署步骤或更详细的硬件对比表，可参考Ollama官方文档或相关技术博客 1 8。

给 DeepSeek 发消息

深度思考 (R1) 联网搜索

内容由 AI 生成，请仔细甄别



# DeepSeek 替代方案：本地部署（实际操作）

搞清楚自己的设备条件后，接下来就可以正式进入本地部署环节：

如果是手机端本地部署，可以考虑用 Chatbox + 硅基流动 DeepSeek API；如果是电脑端本地部署，可以考虑用 Chatbox/Cherry Studio + 硅基流动 DeepSeek API。部署细节就不在这里介绍了，大家在各个社交媒体平台都可以找到适合自己的。

