资料截至
2026-10-30

模型简报 · DeepSeek

DeepSeek V4.1-Flash

2026 年 9 月 10 日首发，10 月 17 日更新为 1017 版。这份简报按官方模型卡和 API 文档，看它的结构、成绩和价格。

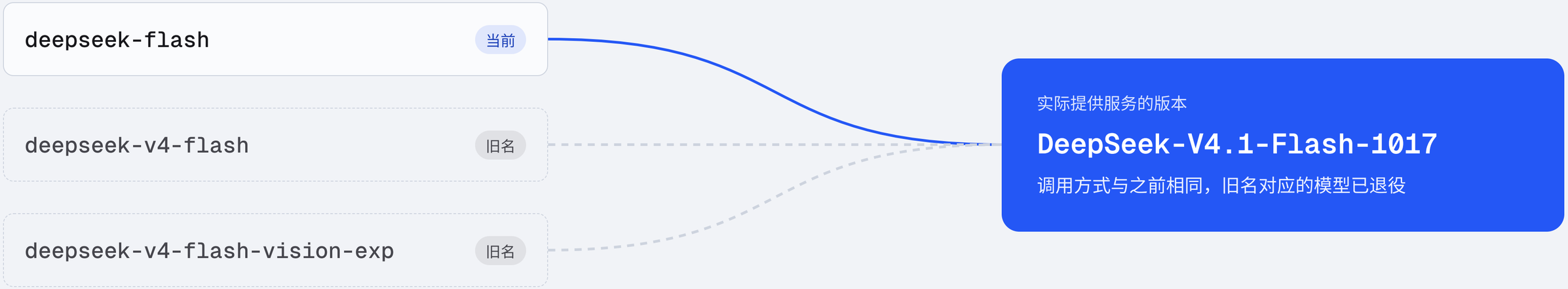
4.1

总参数	上下文	许可
552B	1M tokens	MIT

先核对名字

三个 API 名，同一个模型

官方名称是 DeepSeek-V4.1-Flash，API 里叫 deepseek-flash。两个 V4-Flash 旧名仍可调用，但已转到新版本，按 Flash 价计费。



来源 · DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）；DeepSeek API 文档 · 新闻（api-docs.deepseek.com，2026-10-17）

— 规格

大模型，小激活

总参数 552B，但每个 token 只用其中一小部分：预填充时激活 8B，解码时 16B。

deepseek-ai · 开放权重

DeepSeek-V4.1-Flash

文本 + 图像输入 | 工具调用 | 思考模式 | MIT 许可

552 B 总参数

16 B 解码时激活

层数	40 层：20 层编码 + 20 层解码
专家	1 个共享 + 384 个路由专家
注意力	压缩稀疏注意力 CSA2
预训练	45T tokens 多模态数据
后训练	SFT → 强化学习 → 在线蒸馏
许可	MIT，覆盖代码与权重

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）

混合专家

每个 token 只走 6 个专家

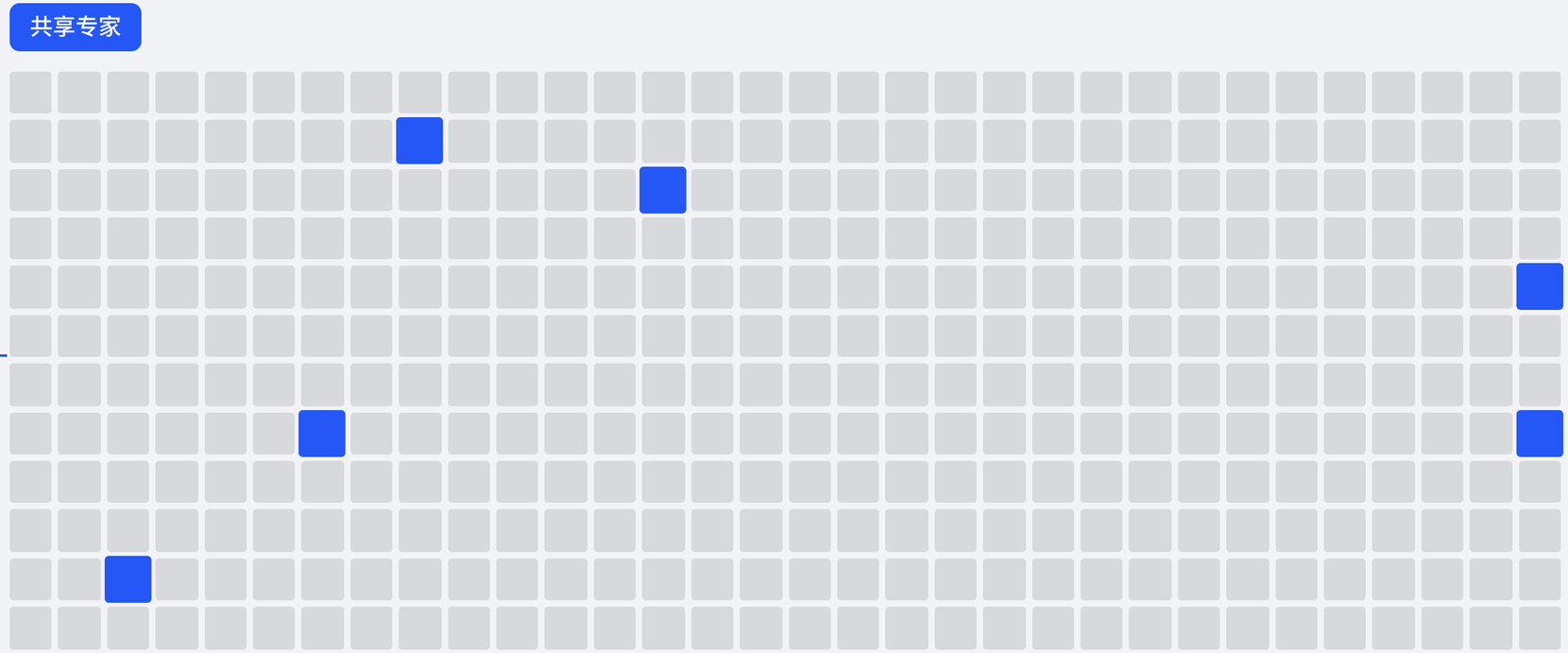
每个 MoE 层有 384 个路由专家和 1 个共享专家。路由器为每个 token 挑 6 个，共享专家每次都参与。

函数
第 1 个 token

边界
换一组专家

条件
每层都重新选

Router



384 个路由专家，每个 token 激活 6 个。亮格位置为示意

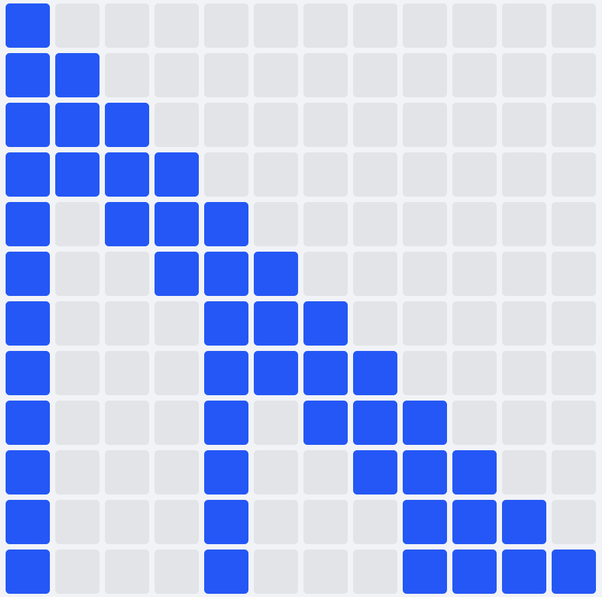
来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）

注意力

不是每个词都要看全部

稀疏注意力让每个 query 只读一部分 key，长上下文的显存和算力随之下降。左图是通用示意，不是 CSA2 的真实模式。

示意图 48/144



全注意力

每个 query 读全部 key，开销随长度平方增长

因果

生成时只读前面的 token，少一半

局部窗口

只读最近的几个 token

稀疏（CSA2）

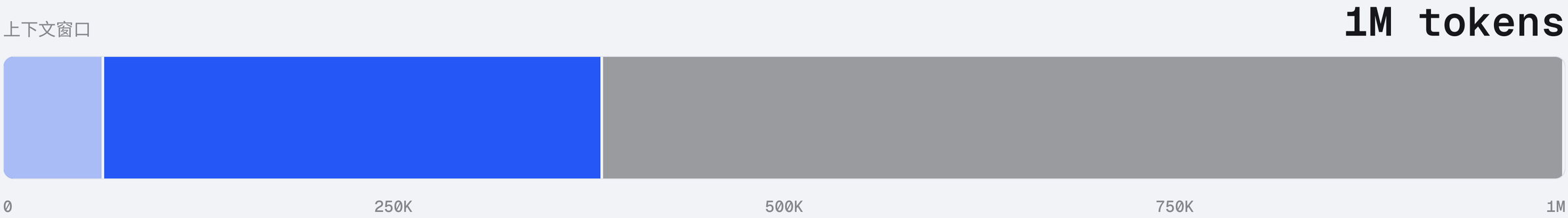
每层可设 Full、Reindex 或 Reuse 模式；KV 缓存约 890 字节/token，为 V4-Flash 的 1/4

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）

— 上下文

1M 窗口，单次最多输出 384K

稀疏注意力先在 64K 长度上训练，预训练进行到 34T tokens 时把上下文扩展到 1M。



64K
稀疏注意力的训练长度

384K
API 单次最大输出

1M
API 上下文长度上限

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）；DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）

— 计费单位

按 token 计价，分三种

同一次请求里，命中缓存的输入、未命中的输入和模型输出各有单价。下面的切分只是示意。

示意：切分方式仅作演示

22 tokens

你是 1	一名 2	代码 3	审查 4	助手 5	。 6	请 7	检查 8	这段 9	函数 10	的 11	边界 12	条件 13	？ 14	第 15

12 16	行 17	缺少 18	空 19	值 20	判断 21	。 22

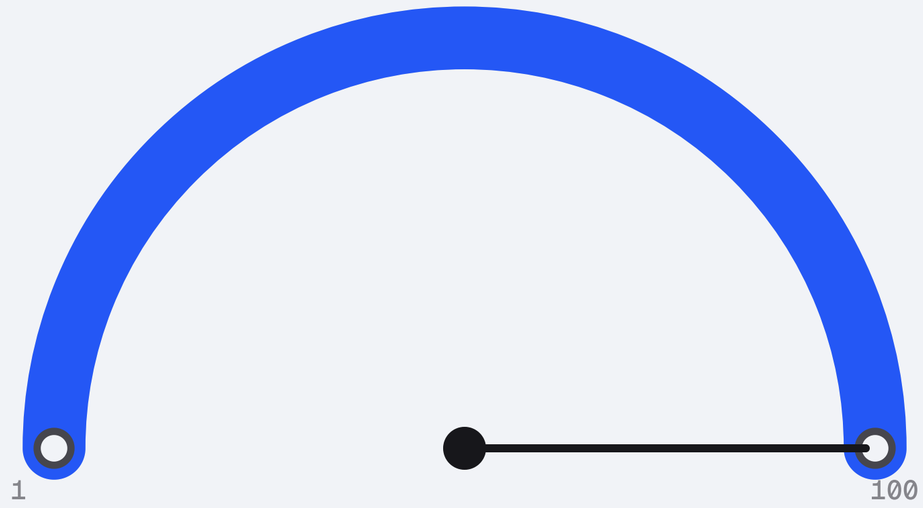
重复的前缀：可命中缓存　　新输入：按未命中价　　模型输出：单独计价

来源 · DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）

推理强度

想多久，可以调

API 默认开启思考模式，也可以关闭。思考强度可在 1 到 100 之间设定。



100
reasoning_effort

1 最低档
可调范围的下限；官方没有公布各档的具体表现

100 最高档
官方指令模型评测都在这一档完成，temperature 1.0

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）；DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）

— 成绩 · 智能体

智能体与编程任务排在前面

官方对比表里，这六项 V4.1-Flash 都是最高分。均为 DeepSeek 自测，最高推理强度。

	Opus-5.0	GPT-5.6 Sol	K3	GLM-5.3	V4-Pro	V4.1-Flash
Terminal-Bench 2.1 Pass@1	89.1	88.8	88.3	88.2	87.9	**90.6**
DeepSWE v1.1 Resolved	74.0	73.0	67.5	66.9	62.7	**74.2**
CyberGym Pass@1	–	84.5	80.0	84.5	83.3	**88.1**
AutomationBench Pass@1	50.3	45.8	46.7	48.8	43.2	**54.8**
HLE w/ tools Pass@1	63.6	–	59.8	62.5	60.0	**63.9**
Agent's Last Exam Pass@1	28.6	26.7	27.6	28.5	25.7	**31.8**

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）。– 为未报告

— 成绩 · 差距

知识问答和更难的终端任务还有差距

同一张表里，这几项 V4.1-Flash 落后于最强的对手，但多数高于自家 V4-Pro。

	Opus-5.0	GPT-5.6 Sol	K3	GLM-5.3	V4-Pro	V4.1-Flash
GPQA Diamond Pass@1	93.4	94.1 •	92.9	88.1	92.4	**90.9**
HLE Pass@1	56.3 •	44.5	43.5	–	–	**36.8**
Terminal-Bench 3.0 Pass@1	43.3 •	34.4	17.7	28.3	11.8	**30.0**
Terminal-Bench 4.0 Pass@1	51.8 •	39.9	12.6	37.9	12.4	**31.2**
ProgramBench Almost@1	37.0 •	23.0	17.5	19.0	15.5	**20.3**
ExploitGym Pass@1	22.1	33.7 •	–	15.0	5.4	**15.3**

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）。HLE 中 GLM-5.3 与 V4-Pro 只有纯文本子集成绩，不可直接比较，记为 –

— 价格

Flash 各项单价比 V4-Pro 低七成以上

缓存命中的输入最便宜，闲时 0.003 美元每百万 tokens。高峰时段所有单价翻倍。

美元 / 百万 tokens	Flash 闲时	Flash 高峰	V4-Pro 闲时	V4-Pro 高峰
输入 命中缓存	$0.003	$0.006	$0.022	$0.044
输入 未命中缓存	$0.15	$0.3	$0.66	$1.32
输出	$0.6	$1.2	$1.98	$3.96

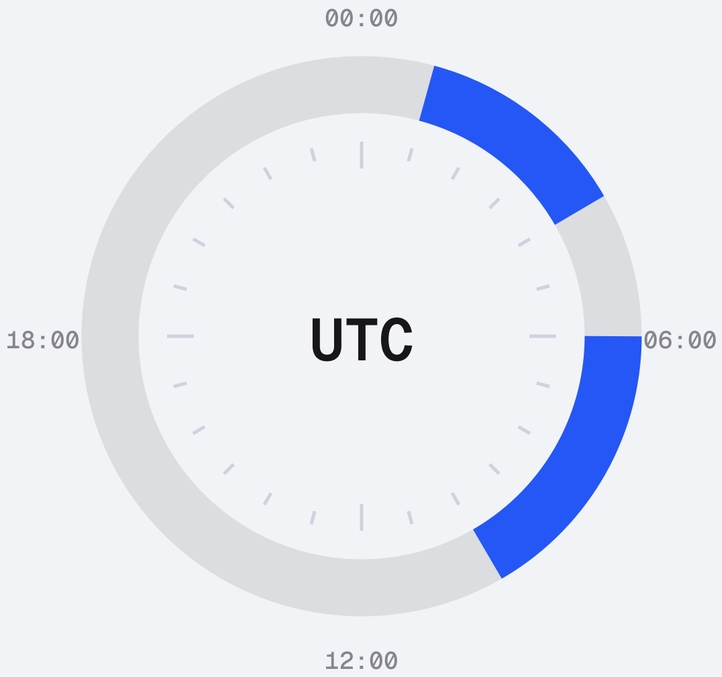
条长按对数刻度

来源 · DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）

高峰时段

每天两段高峰

价格分闲时和高峰两档。错开高峰调用，同样的请求便宜一半。



高峰价 闲时价

高峰时段

UTC 01:00–04:00、06:00–10:00，即北京时间 09:00–12:00、14:00–18:00

只在工作日

周一至周五，中国法定节假日除外；其余时间都按闲时价

差价

Flash 与 V4-Pro 的高峰价都是闲时价的 2 倍

来源 · DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）

— 小结

记住四件事

一个名字

API 名 deepseek-flash，当前由 V4.1-Flash-1017 提供服务。

稀疏结构

552B 总参数，每个 token 用 6 个路由专家，解码时激活 16B。

强在智能体

Terminal-Bench 2.1 为 90.6，DeepSWE v1.1 为 74.2，官方表中最高。

看准时段

闲时输出 0.6 美元每百万 tokens，工作日高峰翻倍。

来源 · DeepSeek-V4.1-Flash 模型卡，Hugging Face（huggingface.co/deepseek-ai，2026）；DeepSeek API 文档 · 模型与价格（api-docs.deepseek.com，2026-10-30 查阅）