

中国科技云智能化提升

- **对话服务（尝鲜版）**
- <https://deepseek.ai4cas.cn>
- **学术优化**
- <https://academic.ai4cas.cn>
- **Deepseek API 服务**
- **联系刘宇德老师 liuyude@cstnet.cn 申请**



中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences



中国科技云智能科研助手

deepseek版

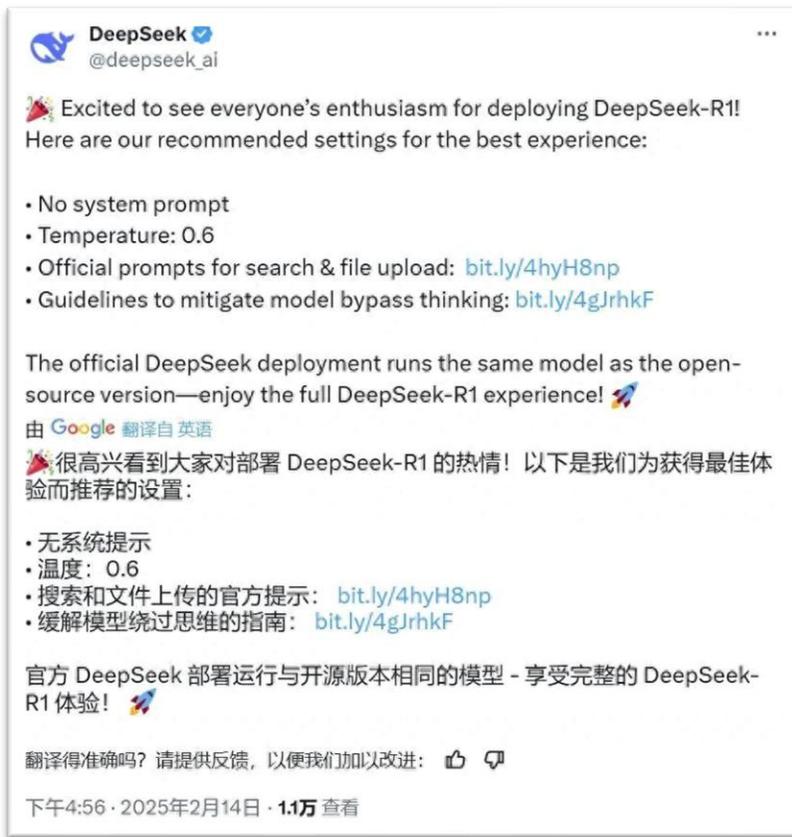
中国科学院计算机网络信息中心

陈灿 何文通

2025年2月28日

deepseek是什么？

- DeepSeek是一家专注通用人工智能（AGI）的中国科技公司，主攻大模型研发与应用。
- deepseek-R1和V3是其开源的推理模型，擅长处理复杂任务且可免费商用。
- deepseek凭借技术创新、成本优势和开源生态，推动了行业范式的转变，其成功标志着AI发展从“算力堆砌”转向“算法驱动”。



AI
+
国产
+
免费
+
开源
+
强大

deepseek主要优势

- 拥有强大的性能。在数学、代码、自然语言推理等任务上，性能比肩 OpenAI-o1 正式版
- deepseek在资源占用方面具有明显优势，降低了GPU计算量，进而提升整体吞吐量

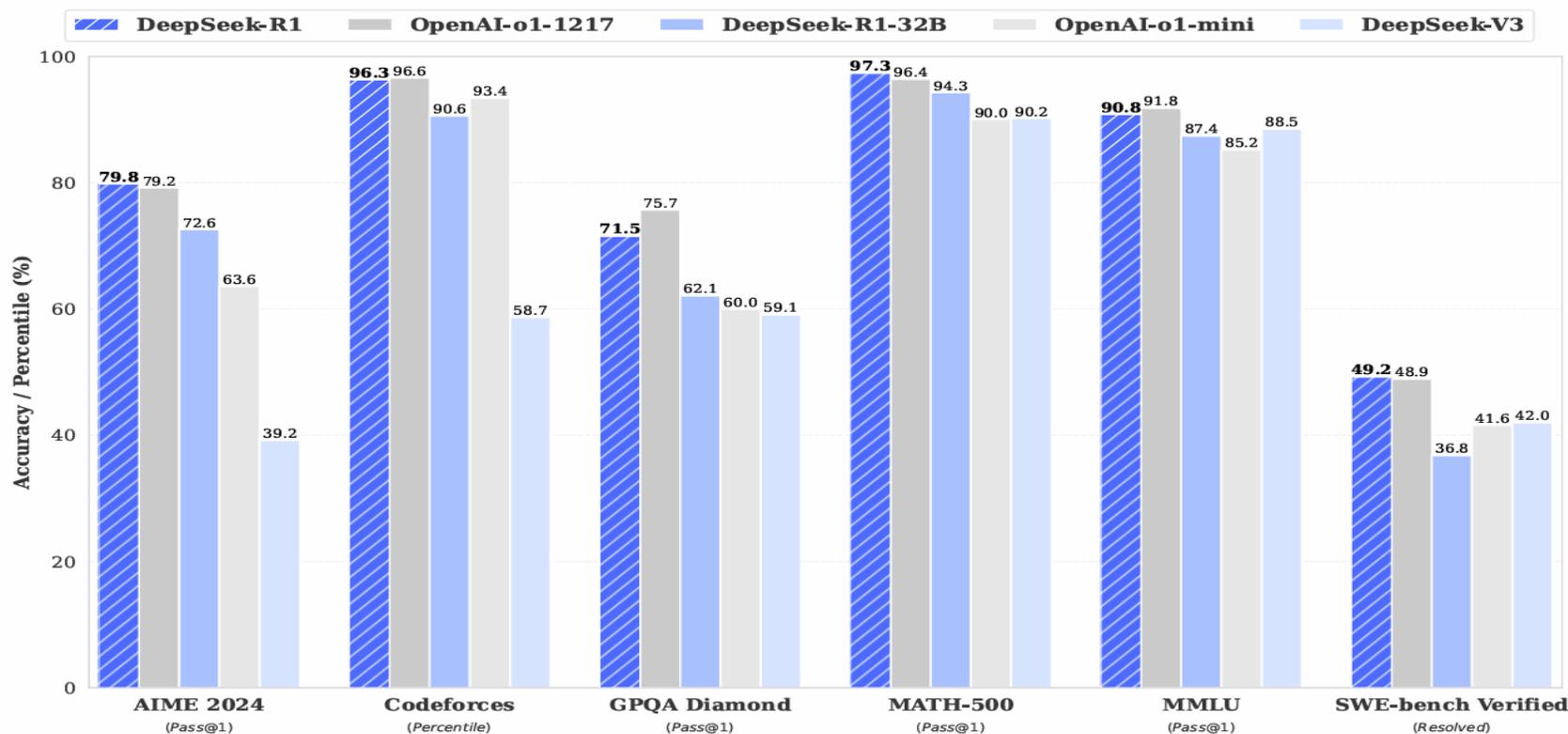


Figure 1 | Benchmark performance of DeepSeek-R1.

deepseek主要优势—更强大的思考能力

ChatGPT o3-mini-high

抚剑啸苍穹，凭栏观众生，帮我写后面两句。

已推理，持续 几秒 ^

抚剑啸望

我正在思考用户请求的后两句，即“抚剑啸苍穹，凭栏观众生”，并探索最合适的回应方式。

照应请求

我正在处理中用户请求的后两句“抚剑啸苍穹，凭栏观众生”，并寻找最合适的方式进行回应。

抚剑啸苍穹，凭栏观众生，
风雨如晦，我心依旧明。
纵横四海，谁与争锋？

- ✓ 断层级别的中文理解和表述能力，可利用其进行学术优化
- ✓ 更强大的推理能力。具有更富有逻辑的思考过程，且思考过程透明



DeepSeek-R1 思考过程

好的，用户让我续写两句诗，前两句是“抚剑啸苍穹，凭栏观众生”。首先，**我需要分析这两句的意境和结构**。抚剑和啸苍穹表现出一种豪迈、激昂的情感，可能和壮志未酬或英雄气概有关。凭栏观众生则转向一种旁观、沉思的姿态，可能是在观察世间百态。接下来，我要考虑**对仗和押韵**的问题。前两句的结构是五言，每句五个字，后两句也需要保持同样的结构。押韵方面，通常古诗押平声韵，所以最后一句的最后一个字需要押韵。原句的“穹”和“生”可能不押韵，但可能需要调整，或者后续两句押韵。然后，**思考主题的延续**。前两句既有动作又有观察，后两句可能需要表达更深的情感或哲理。比如，从个人的豪情转向对世间的感慨，或者表达对命运的看法。例如，用“擎天”来对应“抚剑”，表达支撑或挑战天地的意思。“魑魅”可能指邪恶势力，表达扫除邪恶的决心。**最后一句可能需要一种超然的态度，比如“何必问枯荣”表达对世事变迁的淡然**。还要注意对仗工整，比如“抚剑”对“擎天”，“啸苍穹”对“驱魑魅”，“凭栏”对“昂首”，“观众生”对“问枯荣”。这样结构上更对称，意境也更统一。同时，用词要符合古诗的韵味，避免现代词汇，保持古典气息。最后，检查押韵和整体意境是否连贯，确保后两句不仅结构对仗，还能深化前两句的情感，形成一个完整的诗境。可能需要调整用词，比如“擎天”是否合适，或者是否有更好的词汇表达支撑天地的意思。同时，考虑用户可能的隐藏需求，是否希望表达某种特定情感，比如无奈、豪迈或超脱，从而选择合适的词语。

DeepSeek-R1 回答

抚剑啸苍穹，凭栏观众生。
擎天驱魑魅，昂首问枯荣。

赏析：这首作品通过“擎天驱魑魅”展现匡扶天地的壮志豪情，与首句“抚剑啸苍穹”形成力量呼应。“昂首问枯荣”以哲思收尾，凭栏观世的超然视角在此升华。尾句“何必”二字举重若轻，将人世兴衰纳入苍穹俯仰之间，展现豁达胸襟。

deepseek优势—更低的资源占用率

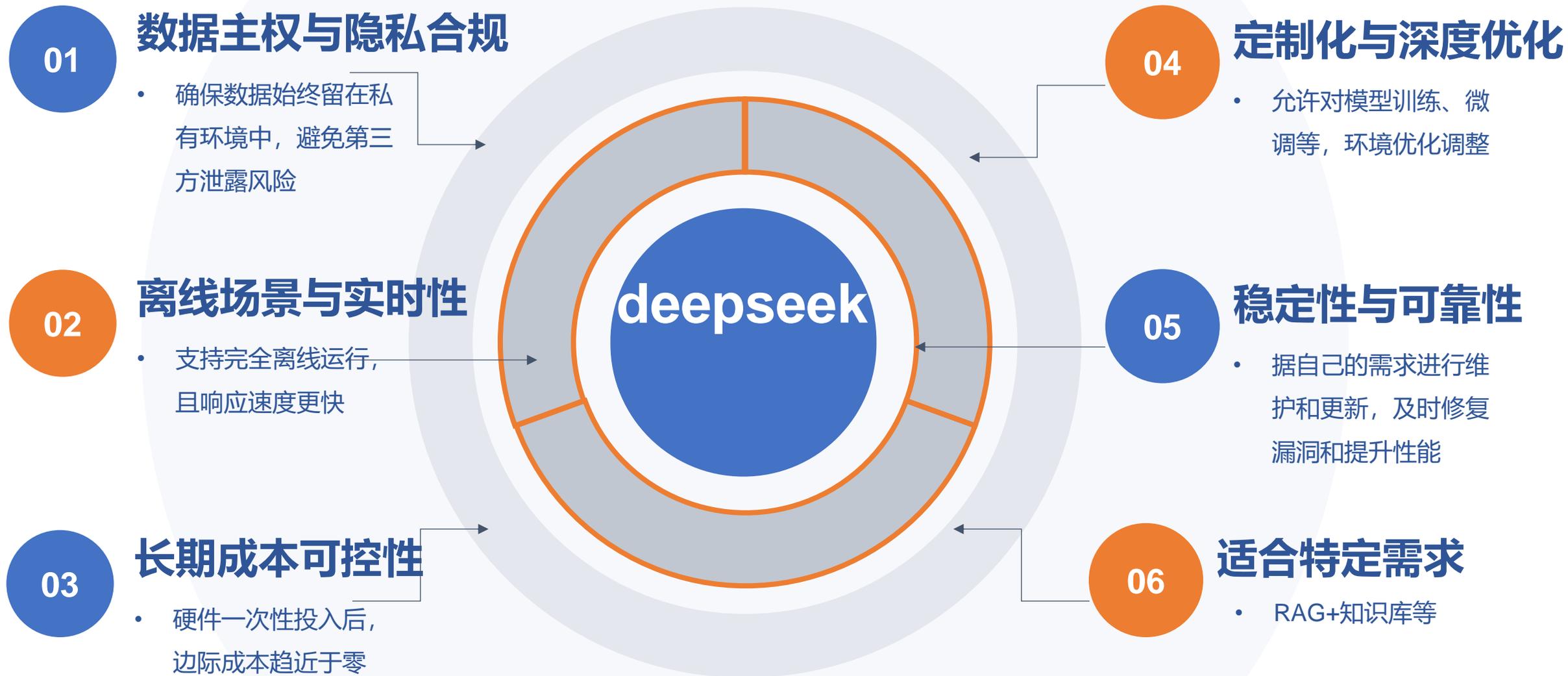
- DeepSeek采用MOE架构和MLA 低秩注意力机制，从而显著降低了计算复杂度
- 蒸馏模型性能出色，各项指标均超过了GPT-4o-0513，尤其在数学推理和代码竞赛相关任务中的表现尤为突出

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

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

蒸馏过程：
利用DeepSeek-R1模型生成QA对，
对小模型进行SFT（指令精调）

deepseek本地化部署优势



中国科技云deepseek本地化部署架构

AI应用层

ScienceDB 智能绩效评估 网络安全协作系统
中国植物DNA条形码数据库 专项研发实施进度和风险监控
等10+项目

deepseek 对话

AI学术搜索

学术优化

更多应用

(开发中)

AAI
统一身份认证

32B R1 对话
671B R1 对话

32B R1 API
671B R1 API
671B V3 API

测试中:
R1 70B

AI能力层

API开放平台 (兼容openAI接口) (部署在高安全等级云 等保三级)



deepseek

Deepseek R1和V3 32B/671B

云镜资源
代码仓库

操作系统 + CUDA + 驱动 + Ollama + Vllm

1. 服务院领导, 已为30+个团队提供API服务
2. 优化调整参数, 保障稳定运行
3. 手搓代码, 实现对话界面
4. 开发api接口, 如embedding

资源层

中国科技云一体化云平台 / 服务器统一管理和统一调度/统一监报告警

GPU

英伟达3090

英伟达4090

英伟达H100

华为910B

高速、安全、稳定的中国科技网

AI科研助手

中国科技云AI科研助手，为广大科研工作者带来了一款高效、智能的学术研究辅助工具，助力科研工作者提高学术资源检索、文献阅读分析、论文撰写编辑等效率。

- **学术优化**：能够助力用户实现论文的高品质阅读、润色与撰写，确保专业术语的精确运用与文章结构的合理调整，还支持 Python、C++ 等项目的深度剖析，助力科研人员快速把握创新精髓
- **LaTeX 公式与表格的智能识别**：用户只需上传公式或表格的图片，助手即可迅速识别并转换为 LaTeX、Markdown 或 HTML 格式，极大地提升了论文编辑的效率与便捷性



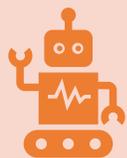
Deepseek API服务



完全兼容 OpenAI API 格式，方便各类应用的直接接入 (如 ragflow, dify, openwebui等)



提供不同版本的模型选择 (deepseek-v3:671b, deepseek-r1:671b, deepseek-r1:32b, 嵌入模型) , 后续会增加 deepseek-r1:70b 和rerank模型



使用API key进行认证，方便自研应用的接入，模型请求接口均返回请求消耗的token数，方便应用层对用户行为进行计量

API.md

2025-02-16

中国科技云 Deepseek 和相关模型服务 API 使用文档

API 文档是为开发者提供的接口调用说明,开发者可以通过 API 文档了解模型使用相关接口的调用方式,参数说明,返回结果等信息.

本文档中的全部 API 均为 OpenAI-API-Compatible, 除文档中描述的直接请求外, 均支持通过 openai sdk 调用. 只需用我们告知的 Token 替换 OPENAI_API_TOKEN 的值, 修改 BASE_URL 即可.

基本说明

BASE_URL: <https://uni-api.cstcloud.cn/v1>

Token: 联系 liuyude@cstnet.cn 获得

Models

获取模型列表

请求方式: GET

请求地址: /models

请求参数: 无

```
curl -X GET "https://uni-api.cstcloud.cn/v1/models" -H "Authorization: Bearer {Token}"
```

返回结果: 模型列表

```
{
  "object": "list",
  "data": [
    {
      "id": "model-id-0",
      "object": "model",
      "created": 1686935002,
      "owned_by": "CSTCloud"
    },
    {
      "id": "model-id-1",
      "object": "model",
      "created": 1686935002,

```

关于部署、参数、显存.....

● Deepseek模型本地部署后，如何才能稳定的提供服务

- 模型本身的大小
- 对话上下文上限
- 并发数
- **输入长度上限**

双卡4090环境 Ollama 启动 deepseek-r1 32b量化版

上下文上限	并发数	实际需要显存大小
4096	2	24 GB
16384	2	36 GB
16384	3	44 GB
32768	2	52 GB (48GB + 4GB内存)

2台8卡H100, vllm部署
deepseek-r1 671b满血版:

对话上下文上限 16384
输入长度上限 6144

目前限制在根据服务情况逐步上调中

官方部署方式 (4+40模式) :

4台8卡A100用于prefill用户输入; 40台8卡A100用于加载模型进行推理服务 (64K上下文)

模型	上下文 (输出+输出长度)	并发	循环次数	并发请求总输出速率 (tokens/s)	单个请求速率的平均	请求超时个数 (超过60s的请求)	平均首个token响应时间 (s)
/models/DeepSeek-R1	2048	1	1	21.03	21.03	0	0.63
/models/DeepSeek-R1	4096	1	1	19.31	19.31	0	1.58
/models/DeepSeek-R1	8192	1	1	17.90	17.90	0	3.11
/models/DeepSeek-R1	2048	4	1	64.32	16.43	0	2.47
/models/DeepSeek-R1	4096	4	1	10.08	2.56	3	1.08
/models/DeepSeek-R1	8192	4	1	0.00	0.00	4	0.00
/models/DeepSeek-R1	2048	8	1	91.96	12.26	0	3.40
/models/DeepSeek-R1	4096	8	1	0.00	0.00	8	0.00
/models/DeepSeek-R1	8192	8	1	0.00	0.00	8	0.00
/models/DeepSeek-R1	2048	16	1	122.72	8.76	0	5.94
/models/DeepSeek-R1	4096	16	1	0.00	0.00	16	0.00
/models/DeepSeek-R1	8192	16	1	0.00	0.00	16	0.00
/models/DeepSeek-R1	2048	32	1	72.97	2.35	23	4.24
/models/DeepSeek-R1	4096	32	1	0.00	0.00	32	0.00
/models/DeepSeek-R1	8192	32	1	0.00	0.00	32	0.00
/models/DeepSeek-R1	2048	64	1	0.00	0.00	64	0.00
/models/DeepSeek-R1	4096	64	1	0.00	0.00	64	0.00
/models/DeepSeek-R1	8192	64	1	0.00	0.00	64	0.00
/models/DeepSeek-R1	2048	96	1	0.00	0.00	96	0.00
/models/DeepSeek-R1	4096	96	1	0.00	0.00	96	0.00
/models/DeepSeek-R1	8192	96	1	0.00	0.00	96	0.00



中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences



汇报完毕！谢谢！

