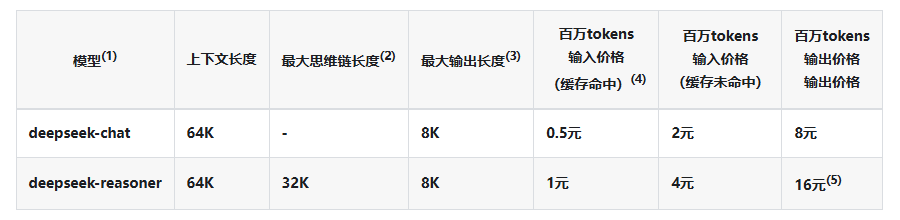
**公司自建大模型或购买第三方API服务费用预研**

**一、第三方API 横向对比**

**1.1 deepseek 官方API**



deepseek-reasoner 模型为新模型 **DeepSeek-R1**

**优点：**

6710亿参数 ，推理能力强

**缺点：**

目前基本不可用阶段，太卡，接口慢

**1.2 阿里云百炼API**

**DeepSeek-R1**

DeepSeek-R1 在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能较高，能力较强。

计费详情

模型调用-输入¥0.004/千Token

模型调用-输出¥0.016/千Token

免费额度

模型免费额度

仅适用于本模型专属的免费额度

免费额度：9990724/10000000

最近到期免费额度：9990724/10000000

最近到期时间：2025-07-26

**缺点**

推理速度稍慢 ，不能满足 2分钟必须出报告的需求 （主要出现二次调用大模型），时间上基本满足不来二次调用推理需求

**DeepSeek-V3**

为自研 MoE 模型，671B 参数，激活 37B，在 14.8T token 上进行了预训练，在长文本、代码、数学、百科、中文 能力上表现优秀。

计费详情

模型调用-输入¥0.002/千Token

模型调用-输出¥0.008/千Token

免费额度

模型免费额度

仅适用于本模型专属的免费额度

免费额度：1000000/1000000

最近到期免费额度：1000000/1000000

最近到期时间：2025-07-26

**缺点**

推理速度稍慢 ，API慢，时间上基本满足不来二次调用推理需求

**1.3 Qwen2.5系列32B模型 （目前提测阶段，在用）**

相较于 Qwen2，Qwen2.5 获得了显著更多的知识，并在编程能力和数学能力方面有了大幅提升。此外，新模型在指令执行、生成长文本、理解结构化数据（例如表格）以及生成结构化输出特别是 JSON 方面取得了显著改进。

计费详情

模型调用-输入¥0.002/千Token

模型调用-输出¥0.006/千Token

模型训练¥0.03/千Token

**优点**

可以满足推理 ，也可以满足API速度

满足根据用户的测量数据进行趋势报告总结

**心脏+ 准备上线的趋势日报、周报**

**缺点**

注册号的比较早，免费额度没有，目前个人账户先用后付费模式

**二、测试环境自建**

目前测试环境，有两个 3060显卡 11G显存

**DeepSeek-R1 1.5B & 7B**

**优点：**

速度比较快

1. 日常简单咨询 可以满足
2. 垂直知识库+简单问题 可以满足

不上传文件模式下，使用的是 DeepSeek-R1 1.5B

公司内部自建 DeeSeek-R1 聊天工具

https://ai-test.995120.cn/chat/aZ95TZoWXxXnbgfn

**缺点**

推理复杂任务无法满足，比如 文本转SQL ，无法满足

推理出用户生理数据，总结趋势，无法满足

**三、生产环境自建**

**3.1 向量知识库**

dify + bge-m3 向量知识库 搭建最低配置



**点击图片可查看完整电子表格**

**云服务器报价**

@陈海波

|  |  |
| --- | --- |
| **规格** | 价格/年 |
| 规格: GPU加速型 | pi2.2xlarge.4 | 8vCPUs | 32GiB | GPU显卡: 1 \* NVIDIA Tesla T4 / 1 \* 16GiB  系统盘: 通用型SSD, 40GiB | **¥25,719.53** |

**3.2 大模型**

**DeepSeek-R1 32B （最优选）**

需要最低配置



**点击图片可查看完整电子表格**

**云服务器报价**

华为云

@陈海波

|  |  |
| --- | --- |
| 规格 | 价格/年 |
| 规格: GPU加速型 | pi1.4xlarge.4 | 16vCPUs | 64GiB | GPU显卡: 2 \* NVIDIA P4 / 2 \* 8G  系统盘: 通用型SSD, 40GiB | **¥45,830.60** |

**自建，机房**

买显卡

显存 ≥ 24GB（如 RTX 3090 或 A5000）\*2

（8000~15000） \*2