《读懂 Token》参考资料（读者版）

> 本清单按章节列出书中数据的出处、各家官方价格页和值得一读的原始资料。所有出处都在 2026 年 10 月 26 日核对过；价格类页面会变，以查阅时看到的为准。网站 pingfan.me/token/ 上的「参考资料」一节内容相同。

各家官方价格页（书里的价格会过时，这里随时看最新的）

1. **DeepSeek 模型与价格**
   书中第 4 章多次用 DeepSeek V4 Flash 举例；2026 年 10 月起已换成 V4.1 Flash，分峰谷计价
   网址：https://api-docs.deepseek.com/zh-cn/quick_start/pricing/
2. **OpenAI API 价格**
   网址：https://developers.openai.com/api/docs/pricing
3. **Anthropic（Claude）API 价格**
   网址：https://platform.claude.com/docs/en/about-claude/pricing
4. **Google Gemini API 价格**
   网址：https://ai.google.dev/gemini-api/docs/pricing

前言：Token 是新的能源货币

5. **国家数据局局长刘烈宏出席国新办新闻发布会**
   国家发展改革委官网，2026-03-25。「我国日均 Token 调用量已经超过 140 万亿，相比 2024 年初的 1000 亿增长了 1000 多倍」
   网址：https://www.ndrc.gov.cn/fzggw/wld/llh/zyhd/202603/t20260325_1404351.html
6. **我国日均词元调用量超 140 万亿**
   新华社，2026-03-24
   网址：https://www.news.cn/tech/20260324/050774b26bc54f429fce2d21e284949a/c.html
7. **Token 中文名定了！**
   新华社，国家数据局官网转载，2026-03-25。官方译名「词元」
   网址：https://www.nda.gov.cn/sjj/swdt/mtsy/0325/20260325113132934906079_pc.html
8. **杨斌：确定 token 的中文译名，已经迫在眉睫了**
   原载腾讯研究院，36氪转载，2026-03-18。「模元」译法的出处
   网址：https://www.36kr.com/p/3728253539351425
9. **Stanford HAI：The 2025 AI Index Report**
   GPT-3.5 水平的推理成本，2022 年 11 月到 2024 年 10 月下降了 280 多倍
   网址：https://hai.stanford.edu/ai-index/2025-ai-index-report

第 2 章：Token 到底是什么

10. **OpenAI：What are tokens and how to count them?**
    英文里 1 个 Token 约等于 4 个字符、约 0.75 个单词
    网址：https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them
11. **OpenAI：Hello GPT-4o**
    2024-05-13。新分词器下，同一句话的中文 Token 从 34 个降到 24 个
    网址：https://openai.com/index/hello-gpt-4o/
12. **OpenAI tiktoken**
    GPT-4 用 cl100k_base（约 10 万词），GPT-4o 用 o200k_base（约 20 万词）
    网址：https://github.com/openai/tiktoken
13. **DeepSeek-V3 Technical Report**
    DeepSeek-AI，2024-12。词表约 12.8 万
    网址：https://arxiv.org/abs/2412.19437
14. **SolidGoldMagikarp (plus, prompt generation)**
    Jessica Rumbelow、Matthew Watkins，2023-02-05。「怪异 Token」现象的原始文章
    网址：https://www.lesswrong.com/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation

第 3 章：AI 是怎么一个字一个字生成回答的

15. **Attention Is All You Need**
    Vaswani 等，2017。Transformer 和注意力机制的原始论文
    网址：https://arxiv.org/abs/1706.03762
16. **Lost in the Middle: How Language Models Use Long Contexts**
    Liu 等，TACL 第 12 卷，2024。长上下文中间的内容容易被忽略
    网址：https://arxiv.org/abs/2307.03172
17. **Language Models are Few-Shot Learners（GPT-3 论文）**
    Brown 等，2020。GPT-3 的上下文窗口是 2048 个 Token
    网址：https://arxiv.org/abs/2005.14165
18. **Anthropic：Introducing Claude 3.5 Sonnet**
    2024-06-21。200K 上下文窗口
    网址：https://www.anthropic.com/news/claude-3-5-sonnet

第 4、5 章：钱花在哪里、不同模型有何差异

19. **OpenAI：Prompt caching**
    缓存命中的输入按更低价格计费
    网址：https://developers.openai.com/api/docs/guides/prompt-caching
20. **OpenAI：Batch API**
    批处理比实时调用便宜 50%
    网址：https://developers.openai.com/api/docs/guides/batch
21. **Anthropic：Batch processing**
    同样是 50% 折扣
    网址：https://platform.claude.com/docs/en/build-with-claude/batch-processing

第 6 章：Token 机制的 6 个推论

22. **Prompting Science Report 2: The Decreasing Value of Chain of Thought in Prompting**
    沃顿商学院 Generative AI Labs，2025-06-08。「一步一步思考」对不同模型的提升
    网址：https://arxiv.org/abs/2506.07142
23. **Vectara 幻觉排行榜**
    各模型做摘要时的幻觉率，持续更新
    网址：https://github.com/vectara/hallucination-leaderboard
24. **Stanford HAI：Hallucinating Law**
    2024-01-11。大模型回答具体法律问题时，幻觉率 69% 到 88%；论文见 Large Legal Fictions
    网址：https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive
    Large Legal Fictions：https://arxiv.org/abs/2401.01301
25. **CJR：AI Search Has a Citation Problem**
    哥伦比亚新闻评论，2025-03-06。八款 AI 搜索给新闻找出处，答错率最高 94%
    网址：https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php