《读懂 Token》延伸学习资料

> 读完这本书，接着学什么？这里整理了 99 条免费、公开的学习资料，都来自官方或作者本人发布的渠道。每条都写了对应书里哪一章、为什么值得看、国内能不能直接打开。
>
> 国内能否打开是按网站域名做的判断，没有在国内逐个实测：能直接打开 27 条，一般能打开（可能较慢）40 条，需要特殊网络 12 条，不确定 20 条。所有资料在 2026-11-04 核对过；网站 pingfan.me/token/learn/ 上的内容相同，并会持续更新。

先看这 10 条

没时间看完全部？先看这 10 条：全部国内能直接打开、中文或中文字幕，按建议的顺序排。括号里是它在下面完整列表中的序号。

1. **Token 可视化工具（本书配套）**（第 30 条）
   先亲手粘一段话进去，看它被切成多少 Token、要花多少钱，书里讲的道理马上就有了手感。
2. **【官方双语】GPT 是什么？直观解释 Transformer（深度学习第 5 章）**（第 1 条）
   用动画看懂「一个词一个词往下预测」，对应第 3 章。
3. **【官方双语】直观解释注意力机制，Transformer 的核心（深度学习第 6 章）**（第 2 条）
   看懂注意力在算什么，就明白上下文为什么越长越贵。
4. **DeepSeek：Token 用量计算**（第 31 条）
   官方给的中英文换算经验值和离线分词器，自己验证第 2 章。
5. **DeepSeek：模型与价格**（第 52 条）
   看真实价格表：输入输出分开算、缓存命中便宜得多、闲时更便宜，对应第 4 章。
6. **阿里云百炼：深度思考模型的用法**（第 51 条）
   官方写明思考内容按输出 Token 计费，看懂推理模型为什么贵。
7. **阿里云百炼：文生文 Prompt 指南**（第 45 条）
   六要素框架加真实案例，照着就能改出更好的提示词，对应第 7 章。
8. **Kimi：Prompt 最佳实践**（第 46 条）
   篇幅短、很实用，教你控制输出长度、长对话先做摘要省上下文。
9. **智谱：提示词工程**（第 47 条）
   三条原则里有一条专门讲怎么减少 AI 编造，对应第 6 章。
10. **人工智能产业发展研究报告（2025 年）**（第 90 条）
    国内权威机构的全景盘点，想看行业大局的读者读这一份就够。

按章节找资料

读到哪一章，就看对应的几条，数字是下面完整列表中的序号。

章节	推荐资料
前言 Token，21 世纪的新型能源货币	91. 人工智能算力基础设施赋能研究报告（2025 年） 92. 2026 年人工智能指数报告（AI Index 2026） 89. 在谷歌规模上测量 AI 服务的环境影响
第 1 章 为什么每个人都该懂点 Token	30. Token 可视化工具（本书配套） 1. 【官方双语】GPT 是什么？直观解释 Transformer（深度学习第 5 章） 3. 生成式 AI 导论 2024
第 2 章 Token 到底是什么	31. DeepSeek：Token 用量计算 14. Happy-LLM：从零开始构建大模型 66. 分词器让不同语言之间产生不公平
第 3 章 AI 是怎么一个字一个字生成回答的	1. 【官方双语】GPT 是什么？直观解释 Transformer（深度学习第 5 章） 2. 【官方双语】直观解释注意力机制，Transformer 的核心（深度学习第 6 章） 18. 68 Transformer【动手学深度学习 v2】
第 4 章 你的钱花在哪里了：Token 的成本真相	52. DeepSeek：模型与价格 49. 阿里云百炼：上下文缓存 50. 阿里云百炼：Batch 批处理接口
第 5 章 不同主流模型的 Token 有何差异	48. 阿里云百炼：模型调用价格 85. FrugalGPT：降低大模型使用成本、同时提升效果 97. Artificial Analysis
第 6 章 Token 机制的 6 个推论	51. 阿里云百炼：深度思考模型的用法 83. 大语言模型幻觉综述 82. 别为 2+3 想那么多：o1 类模型的过度思考问题
第 7 章 Token 实战指南	45. 阿里云百炼：文生文 Prompt 指南 46. Kimi：Prompt 最佳实践 28. 提示工程指南（Prompt Engineering Guide 官方中文版）
第 8 章 少花钱也能用好 AI	57. 腾讯混元：生文计费概述 58. 智谱：上下文缓存 86. LLMLingua：压缩提示词以加速推理
第 9 章 你现在学的东西，5 年后还用得上吗	90. 人工智能产业发展研究报告（2025 年） 92. 2026 年人工智能指数报告（AI Index 2026） 95. 大模型推理价格下降很快，但各类任务降得不均衡

完整列表

一、入门：零基础也能看懂

不需要技术背景。先看这些，对大模型怎么工作、Token 从哪里来会有直观感受。

1. 【官方双语】GPT 是什么？直观解释 Transformer（深度学习第 5 章）

3Blue1Brown 官方 B 站账号 · 视频，约 26 分钟 · 中英双语字幕 · 对应第 2、3 章 · 国内：能直接打开

用动画演示一句话怎么被切成 Token、变成向量，再一个词一个词往下预测，是理解「预测下一个 Token」最直观的视频。

B 站：https://www.bilibili.com/video/BV13z421U7cs/

2. 【官方双语】直观解释注意力机制，Transformer 的核心（深度学习第 6 章）

3Blue1Brown 官方 B 站账号 · 视频，约 25 分钟 · 中英双语字幕 · 对应第 3 章 · 国内：能直接打开

讲清楚「注意力」到底在算什么，看完就能明白为什么上下文越长、算得越慢、越贵。

B 站：https://www.bilibili.com/video/BV1TZ421j7Ke/

3. 生成式 AI 导论 2024

李宏毅（台湾大学） · 课程 · 中文讲授（网页为繁体） · 对应第 1、3、5、6 章 · 国内：能直接打开

华语圈口碑最好的生成式 AI 入门课。第 10 讲专门讲语言模型怎么「文字接龙」，第 6 到 8 讲讲大模型是怎么训练出来的。课程视频放在 YouTube 上。

课程网页：https://speech.ee.ntu.edu.tw/~hylee/genai/2024-spring.php

官方视频（YouTube）：https://www.youtube.com/channel/UC2ggjtuuWvxrHHHiaDH1dlQ/playlists

4. 生成式 AI 与机器学习导论 2025 秋

李宏毅（台湾大学） · 课程 · 中文讲授（网页为繁体） · 对应第 2、3、6 章 · 国内：能直接打开

最新一版，网页写明不需要先备知识，包括「从语言模型到 AI Agent」「解析生成式 AI 内部运作机制」等讲。

课程网页：https://speech.ee.ntu.edu.tw/~hylee/GenAI-ML/2025-fall.php

5. Generative AI for Beginners（官方中文版）

微软 · 教程，21 课 · 简体中文 · 对应第 5、7 章 · 国内：一般能打开，可能较慢

第 1 到 5 课依次讲大模型是什么、怎么比较不同模型、提示词的基础和进阶写法，结构清楚。

GitHub：https://github.com/microsoft/generative-ai-for-beginners/blob/main/translations/zh-CN/README.md

6. Intro to Large Language Models（一小时大模型科普）

Andrej Karpathy（OpenAI 创始成员） · 视频，约 1 小时 · 英文 · 对应第 1、3、6 章 · 国内：需要特殊网络

用一小时讲清大模型是什么、怎么训练、能做什么，是零基础入门的经典讲座。

YouTube：https://www.youtube.com/watch?v=zjkBMFhNj_g

7. Deep Dive into LLMs like ChatGPT

Andrej Karpathy · 视频，长篇 · 英文 · 对应第 2、3、6 章 · 国内：需要特殊网络

面向普通人，从分词、预测下一个 Token 一路讲到幻觉和推理模型，几乎覆盖本书主线。

YouTube：https://www.youtube.com/watch?v=7xTGNNLPyMI

8. How I use LLMs

Andrej Karpathy · 视频 · 英文 · 对应第 5、7、8 章 · 国内：需要特殊网络

顶级专家演示自己日常怎么挑模型、什么时候开思考模式、怎么管理上下文，实用性很强。

YouTube：https://www.youtube.com/watch?v=EWvNQjAaOHw

9. The Illustrated Transformer（图解 Transformer）

Jay Alammar · 图文 · 英文 · 对应第 3 章 · 国内：一般能打开，可能较慢

被无数课程引用的图解经典，一张张图把 Transformer 拆开讲。

原文：https://jalammar.github.io/illustrated-transformer/

10. The Illustrated GPT-2（图解 GPT-2）

Jay Alammar · 图文 · 英文 · 对应第 3 章 · 国内：一般能打开，可能较慢

专门图解 GPT 这类模型如何一个一个生成 Token，正好对应本书第 3 章。

原文：https://jalammar.github.io/illustrated-gpt2/

11. Generative AI exists because of the transformer

英国《金融时报》视觉新闻团队 · 交互式图解 · 英文 · 对应第 2、3 章 · 国内：不确定

主流媒体做的滚动式交互图解，从分词讲到注意力，适合完全不懂技术的读者。

原文：https://ig.ft.com/generative-ai/

12. Why language models hallucinate（语言模型为什么会幻觉）

OpenAI，2025 年 9 月 · 文章（附论文） · 英文 · 对应第 6 章 · 国内：需要特殊网络

观点很好懂：训练和评测时「蒙一个答案」比「说不知道」更划算，于是模型学会了自信地编。读完会明白幻觉不是偶然的毛病。

文章：https://openai.com/index/why-language-models-hallucinate/

论文：https://arxiv.org/abs/2509.04664

13. Tracing the thoughts of a large language model（追踪大模型的「思考」）

Anthropic · 文章 · 英文 · 对应第 3、6 章 · 国内：不确定

用「显微镜」看模型内部，发现它会提前规划押韵、有时会编造推理过程，帮助理解为什么不能全信模型的自述。

原文：https://www.anthropic.com/research/tracing-thoughts-language-model

二、进阶：想多懂一点原理

适合读完全书、想把原理弄得更明白的读者。有的需要一点编程基础，已在说明里标出。

14. Happy-LLM：从零开始构建大模型

Datawhale · 开源教程 · 中文 · 对应第 2、3、5 章 · 国内：一般能打开，可能较慢

从语言处理基础讲到 Transformer、注意力、预训练，第 5 章还带着训练一个分词器。项目写明完全免费。

在线阅读：https://datawhalechina.github.io/happy-llm/

GitHub：https://github.com/datawhalechina/happy-llm

15. 大模型基础（So-Large-LM）

Datawhale · 开源教程 · 中文 · 对应第 3、5、9 章 · 国内：一般能打开，可能较慢

以斯坦福 CS324 和李宏毅课程为基础，系统讲大模型的能力、架构、混合专家模型、数据和训练，适合决策者建立整体认识。

在线阅读：https://datawhalechina.github.io/so-large-lm/

16. Hugging Face LLM 课程（官方中文版）

Hugging Face · 在线课程 · 中文 · 对应第 2、3 章 · 国内：需要特殊网络

开源 AI 社区的官方课程。讲分词器的几节对比了按词、按字、按子词三种切法，以及 BPE 等算法，正好补充第 2 章。

课程首页：https://huggingface.co/learn/llm-course/zh-CN/chapter1/1

分词器一节：https://huggingface.co/learn/llm-course/zh-CN/chapter2/4

BPE 一节：https://huggingface.co/learn/llm-course/zh-CN/chapter6/5

17. 《动手学深度学习》第 10 章「注意力机制」

《动手学深度学习》作者团队 · 在线书 · 中文 · 对应第 3 章 · 国内：能直接打开

从注意力讲到多头注意力、自注意力和 Transformer，配有可运行代码，在线免费读。

本章：https://zh.d2l.ai/chapter_attention-mechanisms/index.html

Transformer 一节：https://zh.d2l.ai/chapter_attention-mechanisms/transformer.html

18. 68 Transformer【动手学深度学习 v2】

跟李沐学AI（李沐本人 B 站账号） · 视频 · 中文 · 对应第 3 章 · 国内：能直接打开

配合上一条的书看，作者亲自讲解。

B 站：https://www.bilibili.com/video/BV1Kq4y1H7FL/

19. Transformer 论文逐段精读

跟李沐学AI · 视频，约 1 小时 27 分 · 中文 · 对应第 3 章 · 国内：能直接打开

一段一段带读《Attention Is All You Need》，适合想看原始论文、又怕看不懂的读者。

B 站：https://www.bilibili.com/video/BV1pu411o7BE/

20. Let's build the GPT Tokenizer

Andrej Karpathy · 视频，约 2 小时 · 英文 · 对应第 2 章 · 国内：需要特殊网络

讲分词器的权威视频，解释了为什么中文等非英语语言更费 Token、为什么模型数不清单词里的字母。

YouTube：https://www.youtube.com/watch?v=zduSFxRajkE

21. Neural Networks: Zero to Hero

Andrej Karpathy · 视频课程（需要会 Python） · 英文 · 对应第 2、3 章 · 国内：不确定

跟着从零写出一个小 GPT，想真正搞懂原理的读者可以挑战。

课程主页：https://karpathy.ai/zero-to-hero.html

Let's build GPT：https://www.youtube.com/watch?v=kCc8FmEb1nY

22. Building effective agents（构建高效的智能体）

Anthropic · 文章 · 英文 · 对应第 4、8 章 · 国内：不确定

业内引用最广的智能体设计指南，核心观点是能用简单流程就别上复杂智能体，直接关系到省 Token。

原文：https://www.anthropic.com/research/building-effective-agents

23. Effective context engineering for AI agents（上下文工程）

Anthropic · 文章 · 英文 · 对应第 3、4、7 章 · 国内：不确定

讲清「上下文窗口是稀缺资源」，以及怎样只把最有用的 Token 放进去，是提示词工程的升级版。

原文：https://www.anthropic.com/research/effective-context-engineering-for-ai-agents

24. How we built our multi-agent research system

Anthropic · 文章 · 英文 · 对应第 4 章 · 国内：不确定

官方实战复盘，说明多智能体系统为什么 Token 消耗会成倍增加，以及在什么情况下值得。

原文：https://www.anthropic.com/research/multi-agent-research-system

25. Reasoning models don't always say what they think

Anthropic · 文章 · 英文 · 对应第 6 章 · 国内：不确定

研究发现推理模型写出的「思考过程」不一定是它真实的理由，提醒读者别把思维链当成可信的证明。

原文：https://www.anthropic.com/research/reasoning-models-dont-say-think

26. LLM Powered Autonomous Agents

Lilian Weng（翁荔） · 文章 · 英文 · 对应第 4、6 章 · 国内：一般能打开，可能较慢

华人研究者写的智能体经典综述，把规划、记忆、工具使用三块讲得系统清楚。

原文：https://lilianweng.github.io/posts/2023-06-23-agent/

27. Prompt Engineering

Lilian Weng（翁荔） · 文章 · 英文 · 对应第 6、7 章 · 国内：一般能打开，可能较慢

系统梳理少样本、思维链等提示方法背后的研究依据，适合想知道「为什么有效」的读者。

原文：https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/

28. 提示工程指南（Prompt Engineering Guide 官方中文版）

DAIR.AI · 文档网站 · 中文 · 对应第 6、7 章 · 国内：不确定

开源社区维护最久的提示词百科，零样本、少样本、思维链、自我一致性等技巧集中讲全，每种都有例子。

中文版：https://www.promptingguide.ai/zh

29. Learn Prompting 提示工程指南（官方中文版）

Learn Prompting · 教程 · 中文 · 对应第 7 章 · 国内：不确定

从基础一路讲到可靠性和提示词攻击，例子多，适合职场人系统补课。基础内容免费。

中文版：https://learnprompting.org/zh-Hans/docs/introduction

三、动手：自己试一试

能直接在浏览器里玩的工具，以及跟着做的开源项目。

30. Token 可视化工具（本书配套）

本书作者 · 在线工具 · 中文 · 对应第 2、4 章 · 国内：能直接打开

粘贴一段话，看它在 DeepSeek、通义千问、GPT-4o 三个模型里各被切成多少 Token、要花多少钱；还能模拟「聊得越久越贵」。所有计算在浏览器里完成。

本站：https://pingfan.me/token/visualizer/

31. DeepSeek：Token 用量计算

DeepSeek 官方文档 · 文档，附离线分词器 · 中文 · 对应第 2 章 · 国内：能直接打开

官方给出了中英文字符和 Token 的换算经验值，还提供可以下载的离线分词器和图片 Token 估算方法，可以自己验证书里讲的「中文和英文费 Token 不一样」。

官方文档：https://api-docs.deepseek.com/zh-cn/quick_start/token_usage

32. Kimi：计算 Token 接口

Kimi（月之暗面）官方文档 · 接口文档 · 中文 · 对应第 2、8 章 · 国内：能直接打开

发请求之前先估算会用多少 Token，支持带图片的内容。适合会一点编程的读者。

官方文档：https://platform.kimi.com/docs/api/estimate

33. OpenAI Tokenizer（官方分词器页面）

OpenAI · 在线工具 · 英文界面，可输入中文 · 对应第 2 章 · 国内：需要特殊网络

粘贴一段中文和一段英文对比 Token 数，可以切换不同代 GPT 的分词规则。

工具：https://platform.openai.com/tokenizer

34. Tiktokenizer

开发者 dqbd · 在线工具 · 英文界面，可输入中文 · 对应第 2 章 · 国内：需要特殊网络

可以切换多种模型的分词规则，彩色显示每个 Token 和它的编号，Karpathy 在分词器视频里就用它演示。

工具：https://tiktokenizer.vercel.app/

35. The Tokenizer Playground

Xenova（Hugging Face） · 在线工具 · 英文界面 · 对应第 2 章 · 国内：需要特殊网络

对比各个开源模型的分词差异，看同一句中文在不同模型里被切成几块。

工具：https://huggingface.co/spaces/Xenova/the-tokenizer-playground

36. Transformer Explainer

佐治亚理工学院 Polo Club · 交互工具 · 英文界面 · 对应第 3 章 · 国内：一般能打开，可能较慢

在浏览器里跑一个真实的小模型，输入一句话就能看它怎么算出下一个词的概率，还能拖动「温度」滑块看回答怎么变。

工具：https://poloclub.github.io/transformer-explainer/

37. LLM Visualization（大模型 3D 可视化）

Brendan Bycroft · 交互工具 · 英文界面 · 对应第 3 章 · 国内：不确定

用 3D 动画带你走进模型内部，一步步看一个 Token 怎么变成下一个 Token。

工具：https://bbycroft.net/llm

38. 面向开发者的 LLM 入门教程（llm-cookbook）

Datawhale · 开源教程，有 PDF · 中文 · 对应第 7 章 · 国内：一般能打开，可能较慢

吴恩达 11 门大模型短课程的中文整理和代码复现，包括《ChatGPT Prompt Engineering for Developers》，是想看吴恩达课程的中文读者的首选。

在线阅读：https://datawhalechina.github.io/llm-cookbook/

GitHub：https://github.com/datawhalechina/llm-cookbook

39. 动手学大模型应用开发（llm-universe）

Datawhale · 开源教程（需要会一点 Python） · 中文 · 对应第 5、8 章 · 国内：一般能打开，可能较慢

带着做一个个人知识库助手，过程中对比调用文心、星火、智谱等国内模型。

在线阅读：https://datawhalechina.github.io/llm-universe/

40. 开源大模型食用指南（self-llm）

Datawhale · 开源教程（需要 Linux 基础） · 中文 · 对应第 5、8 章 · 国内：一般能打开，可能较慢

教你在本地部署和微调通义千问等开源模型，体会「自己跑模型」和「按 Token 买 API」的成本差别。

GitHub：https://github.com/datawhalechina/self-llm

41. Hello-Agents：从零开始构建智能体

Datawhale · 开源教程 · 中文 · 对应第 4 章 · 国内：一般能打开，可能较慢

从智能体概念讲到 Coze、Dify 等低代码平台，再到自己写智能体框架，帮你看清智能体为什么会「烧」Token。

在线阅读：https://datawhalechina.github.io/hello-agents/

42. minbpe：极简分词器代码

Andrej Karpathy · 开源项目（需要会 Python） · 英文 · 对应第 2 章 · 国内：一般能打开，可能较慢

配合分词器视频一起用，几百行代码就能自己训练一个分词器。

GitHub：https://github.com/karpathy/minbpe

43. nanochat

Andrej Karpathy · 开源项目 · 英文 · 对应第 3、9 章 · 国内：一般能打开，可能较慢

项目口号是「100 美元能买到的最好的 ChatGPT」，直观感受训练一个小聊天模型要花多少钱。代码免费，自己训练要租显卡。

GitHub：https://github.com/karpathy/nanochat

44. Anthropic 交互式提示词教程

Anthropic · 练习教程 · 英文 · 对应第 7 章 · 国内：一般能打开，可能较慢

官方出的分章节练习，一边改提示词一边看效果。

GitHub：https://github.com/anthropics/prompt-eng-interactive-tutorial

四、各家官方文档：真实的计费和用法

书里讲到的计费、缓存、批处理、思考模式和提示词写法，各家官方文档里都有真实规则。国内厂商的文档都是中文、能直接打开。价格变动很频繁，以打开时看到的为准。

45. 阿里云百炼：文生文 Prompt 指南

阿里云 · 官方文档 · 中文 · 对应第 7 章 · 国内：能直接打开

给出「背景、目的、风格、语气、受众、输出」六要素框架，附有 HR 助手、销售对话提取等真实案例。

官方文档：https://help.aliyun.com/zh/model-studio/prompt-engineering-guide

46. Kimi：Prompt 最佳实践

Kimi（月之暗面） · 官方文档 · 中文 · 对应第 7、8 章 · 国内：能直接打开

短而实用，还提到用段落数、要点数限制输出长度比限制字数更好控制，以及长对话怎么先做摘要省上下文。

官方文档：https://platform.kimi.com/docs/guide/prompt-best-practice

47. 智谱：提示词工程

智谱 AI · 官方文档 · 中文 · 对应第 6、7 章 · 国内：能直接打开

讲三件事：指令要清晰具体，给参考资料以减少编造，把复杂任务拆成小任务。

官方文档：https://docs.bigmodel.cn/cn/guide/platform/prompt

48. 阿里云百炼：模型调用价格

阿里云 · 官方文档 · 中文 · 对应第 4、5 章 · 国内：能直接打开

输入和输出分开计价，并按单次输入长度分档；还写了音频怎么换算成 Token，是「语音也有 Token」的实证。

官方文档：https://help.aliyun.com/zh/model-studio/billing-for-model-studio

49. 阿里云百炼：上下文缓存

阿里云 · 官方文档 · 中文 · 对应第 4 章 · 国内：能直接打开

讲清显式缓存和隐式缓存怎么计费、命中后便宜多少，附有现成算例。

官方文档：https://help.aliyun.com/zh/model-studio/context-cache

50. 阿里云百炼：Batch 批处理接口

阿里云 · 官方文档 · 中文 · 对应第 4 章 · 国内：能直接打开

官方写明批处理费用只有实时调用的一半，对应第 4 章「批处理」一节。

官方文档：https://help.aliyun.com/zh/model-studio/batch-interfaces-compatible-with-openai

51. 阿里云百炼：深度思考模型的用法

阿里云 · 官方文档 · 中文 · 对应第 4、6 章 · 国内：能直接打开

写明思考内容按输出 Token 计费，还讲了怎么限制思考长度，正好对应「推理模型为什么贵」。

官方文档：https://help.aliyun.com/zh/model-studio/deep-thinking

52. DeepSeek：模型与价格

DeepSeek · 官方文档 · 中文 · 对应第 4、5 章 · 国内：能直接打开

缓存命中和未命中价格相差很大，空闲时段比高峰时段便宜，是讲「用缓存、错峰用」的好例子。

官方文档：https://api-docs.deepseek.com/zh-cn/quick_start/pricing

53. DeepSeek：上下文硬盘缓存

DeepSeek · 官方文档 · 中文 · 对应第 4 章 · 国内：能直接打开

缓存默认开启，返回结果里能看到命中和没命中的 Token 数，帮你理解为什么「开头不变」能省钱。

官方文档：https://api-docs.deepseek.com/zh-cn/guides/kv_cache

54. DeepSeek：思考模式

DeepSeek · 官方文档 · 中文 · 对应第 6 章 · 国内：能直接打开

可以开关思考、选择思考强度，对应第 6 章讲的思考 Token。

官方文档：https://api-docs.deepseek.com/zh-cn/guides/thinking_mode

55. 火山方舟（豆包）：模型价格

火山引擎 · 官方文档 · 中文 · 对应第 4、5 章 · 国内：能直接打开

在线推理和批量推理分开报价，还有缓存命中、缓存存储费用，以及视频生成按 Token 计费，覆盖面最广。

官方文档：https://docs.volcengine.com/docs/82379/1544106

56. 百度千帆：模型服务计费

百度智能云 · 官方文档 · 中文 · 对应第 4、5 章 · 国内：能直接打开

输入和输出分开计价，部分模型分高峰和空闲时段，或按输入长度分档。

官方文档：https://cloud.baidu.com/doc/qianfan/s/wmh4sv6ya

57. 腾讯混元：生文计费概述

腾讯云 · 官方文档 · 中文 · 对应第 4、8 章 · 国内：能直接打开

写明按 Token 数计费、调用失败不计费，并有免费额度说明。

官方文档：https://cloud.tencent.com/document/product/1729/97731

58. 智谱：上下文缓存

智谱 AI · 官方文档 · 中文 · 对应第 4 章 · 国内：能直接打开

有算例说明缓存能省多少，还建议重复的开头要足够长才能命中缓存。

官方文档：https://docs.bigmodel.cn/cn/guide/capabilities/cache

59. Kimi：模型推理价格说明

Kimi（月之暗面） · 官方文档 · 中文 · 对应第 4、5 章 · 国内：能直接打开

输入、缓存命中、输出分别计价，缓存写入按保留时长计费。

官方文档：https://platform.kimi.com/docs/pricing/chat

60. Claude 文档：提示工程（官方中文版）

Anthropic · 官方文档 · 中文 · 对应第 7 章 · 国内：不确定

模型厂商亲自写的提示词写法，有中文版可以直接读。

概述：https://platform.claude.com/docs/zh-CN/build-with-claude/prompt-engineering/overview

最佳实践：https://platform.claude.com/docs/zh-CN/build-with-claude/prompt-engineering/claude-prompting-best-practices

61. Claude 文档：缓存、Token 计数、批处理、上下文窗口、扩展思考（官方中文版）

Anthropic · 官方文档 · 中文 · 对应第 3、4、6、8 章 · 国内：不确定

本书讲到的缓存、批处理、上下文窗口、思考 Token，在这里能看到真实规则和折扣比例。

提示缓存：https://platform.claude.com/docs/zh-CN/build-with-claude/prompt-caching

Token 计数：https://platform.claude.com/docs/zh-CN/build-with-claude/token-counting

批处理：https://platform.claude.com/docs/zh-CN/build-with-claude/batch-processing

上下文窗口：https://platform.claude.com/docs/zh-CN/build-with-claude/context-windows

扩展思考：https://platform.claude.com/docs/zh-CN/build-with-claude/extended-thinking

62. OpenAI 文档：提示词、缓存、批处理、推理模型，以及用 tiktoken 数 Token

OpenAI · 官方文档 · 英文 · 对应第 2、4、6、7、8 章 · 国内：需要特殊网络

可以和其他几家对照着看怎么计费、怎么打折、推理模型为什么贵；Cookbook 一篇教你用代码数 Token。

提示词：https://developers.openai.com/api/docs/guides/prompt-engineering

提示缓存：https://developers.openai.com/api/docs/guides/prompt-caching

批处理：https://developers.openai.com/api/docs/guides/batch

推理模型：https://developers.openai.com/api/docs/guides/reasoning

用 tiktoken 数 Token：

https://developers.openai.com/cookbook/examples/how_to_count_tokens_with_tiktoken

63. Google Gemini 文档：提示词设计、Token 计算、上下文缓存

Google · 官方文档 · 英文 · 对应第 2、4、7 章 · 国内：需要特殊网络

第三家大厂的写法，Token 计算一页讲了图片、音频、视频怎么换算成 Token。

提示词设计：https://aistudio.google.com/docs/prompting-strategies

Token 计算：https://aistudio.google.com/docs/tokens

上下文缓存：https://aistudio.google.com/docs/caching

五、深度阅读：原始论文

书里讲到的很多概念，最早都出自这些论文。不必读懂公式，看看引言、配图和结论就有收获。阅读难度标在类型里。

64. 用子词单元翻译生僻词（BPE 的出处）

Sennrich、Haddow、Birch，2016 · 论文 · 难度中 · 英文 · 对应第 2 章 · 国内：一般能打开，可能较慢

今天几乎所有大模型的切词方法 BPE 都源于这篇。读完就明白 Token 为什么既不是字也不是词，而是按出现频率拼出来的碎片。

arXiv：https://arxiv.org/abs/1508.07909

65. SentencePiece：与语言无关的分词工具

Kudo、Richardson，2018 · 论文 · 难度中 · 英文 · 对应第 2 章 · 国内：一般能打开，可能较慢

讲清了中文、日文这类不用空格隔开的语言是怎么被切成 Token 的。

arXiv：https://arxiv.org/abs/1808.06226

66. 分词器让不同语言之间产生不公平

Petrov 等，NeurIPS 2023 · 论文 · 难度易到中 · 英文 · 对应第 2、4 章 · 国内：一般能打开，可能较慢
用实测数据说明同一句话换成不同语言，Token 数可以相差好几倍，价格、速度、能用的上下文长度都跟着受影响。
arXiv：https://arxiv.org/abs/2305.15425

67. 所有语言的费用都一样吗？商用大模型时代的分词问题

Ahia 等，EMNLP 2023 · 论文 · 难度易到中 · 英文 · 对应第 2、4 章 · 国内：一般能打开，可能较慢
直接从按 Token 计费的 API 账单算账，证明非英语用户往往花更多钱、得到更差的效果，很适合企业决策者看。
arXiv：https://arxiv.org/abs/2305.13707

68. Attention Is All You Need（Transformer 原始论文）

Vaswani 等，2017 · 论文 · 难度难 · 英文 · 对应第 3 章 · 国内：一般能打开，可能较慢
当今所有大模型的源头。看看引言和配图，就知道书里「注意力」这个词是从哪儿来的。配合前面李沐的逐段精读视频更好懂。
arXiv：https://arxiv.org/abs/1706.03762

69. 语言模型是无监督的多任务学习者（GPT-2）

Radford 等，OpenAI，2019 · 技术报告 · 难度中 · 英文 · 对应第 3 章 · 国内：需要特殊网络
第一次让人看到，只靠「预测下一个 Token」训练出来的模型，竟然顺带学会了翻译、问答、写摘要。
PDF：https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

70. 语言模型是少样本学习者（GPT-3）

Brown 等，OpenAI，2020 · 论文 · 难度中 · 英文 · 对应第 3、7 章 · 国内：一般能打开，可能较慢
提出在提示词里给几个例子，模型就能照着做，这是第 7 章「举例子」写法的理论源头。
arXiv：https://arxiv.org/abs/2005.14165

71. 迷失在中间：语言模型怎样使用长上下文

Liu 等，TACL 2024 · 论文 · 难度易 · 英文 · 对应第 3、7 章 · 国内：一般能打开，可能较慢
发现模型对开头和结尾的信息记得最牢，中间的容易被忽略。马上能学到一招：重要的话放在提示词的首尾。
arXiv：https://arxiv.org/abs/2307.03172

72. 神经语言模型的规模定律

Kaplan 等，OpenAI，2020 · 论文 · 难度中 · 英文 · 对应第 3、9 章 · 国内：一般能打开，可能较慢
用曲线说明模型越大、数据越多、算力越多，效果就按规律变好，这是行业大建数据中心的理论依据，投资者尤其值得一看。
arXiv：https://arxiv.org/abs/2001.08361

73. 训练算力最优的大语言模型（Chinchilla）

Hoffmann 等，DeepMind，2022 · 论文 · 难度中 · 英文 · 对应第 4、5 章 · 国内：一般能打开，可能较慢

指出很多模型参数多、读的数据少，并给出每个参数大约配 20 个训练 Token 的经验配比。由此能理解，小一点但训练更充分的模型往往更便宜也更好用。

arXiv：https://arxiv.org/abs/2203.15556

74. 把神经网络的知识蒸馏出来

Hinton、Vinyals、Dean，2015 · 论文 · 难度中 · 英文 · 对应第 4 章 · 国内：一般能打开，可能较慢

「蒸馏」这个概念就出自这里：让大模型当老师，教出一个又小又便宜的学生模型，是第 4 章蒸馏一节的源头。

arXiv：https://arxiv.org/abs/1503.02531

75. DeepSeek-V3 技术报告

DeepSeek-AI，2024 · 技术报告 · 难度难 · 英文 · 对应第 4、5 章 · 国内：一般能打开，可能较慢

讲了混合专家架构怎么做到每次只启用一小部分参数，以及训练成本，是理解国产模型为什么能把 Token 价格打下来的一手材料。可以只看引言和训练成本部分。

arXiv：https://arxiv.org/abs/2412.19437

76. Chatbot Arena：靠人类投票评测大模型

Chiang 等，2024 · 论文 · 难度易到中 · 英文 · 对应第 5 章 · 国内：一般能打开，可能较慢

媒体常引用的大模型排行榜就是这样产生的：让用户盲比两个回答再投票。看完就知道排行榜能信几分。

arXiv：https://arxiv.org/abs/2403.04132

77. 思维链提示能激发大模型的推理能力

Wei 等，Google，2022 · 论文 · 难度易 · 英文 · 对应第 6、7 章 · 国内：一般能打开，可能较慢

证明让模型写出解题步骤能显著提高数学和逻辑题的正确率，代价是多花 Token，这正是「思考 Token」这笔账的起点。

arXiv：https://arxiv.org/abs/2201.11903

78. 大语言模型是零样本推理者（「让我们一步一步想」）

Kojima 等，2022 · 论文 · 难度易 · 英文 · 对应第 6、7 章 · 国内：一般能打开，可能较慢

只在提示词里加一句「Let's think step by step」，准确率就大幅提升，是最有名的一句话提示词技巧的出处。

arXiv：https://arxiv.org/abs/2205.11916

79. 自洽性提升思维链推理（Self-Consistency）

Wang 等，ICLR 2023 · 论文 · 难度易到中 · 英文 · 对应第 6 章 · 国内：一般能打开，可能较慢

让模型多想几遍、按多数票取答案，正确率更高，是用更多 Token 换更高准确率的典型例子。

arXiv：https://arxiv.org/abs/2203.11171

80. 测试时多花算力可能比扩大参数更有效

Snell 等，2024 · 论文 · 难度中到难 · 英文 · 对应第 4、6、9 章 · 国内：一般能打开，可能较慢

解释了推理模型背后的经济逻辑：与其把模型做大，不如让它回答时多想一会儿，这也是 Token 消耗量暴涨的原因之一。

arXiv：https://arxiv.org/abs/2408.03314

81. DeepSeek-R1：用强化学习激发大模型的推理能力

DeepSeek-AI，2025 · 论文 · 难度中 · 英文 · 对应第 4、6 章 · 国内：一般能打开，可能较慢

国产推理模型的一手论文，讲了模型如何自己学会长思考、思考过程越来越长，以及「顿悟时刻」，是理解思考 Token 的最佳中国案例。

arXiv：https://arxiv.org/abs/2501.12948

82. 别为 2+3 想那么多：o1 类模型的过度思考问题

Chen 等，2024 · 论文 · 难度易到中 · 英文 · 对应第 6、8 章 · 国内：一般能打开，可能较慢

发现推理模型连简单题也会写上千个思考 Token。据此可以判断：简单任务别用推理模型，能省下不少钱。

arXiv：https://arxiv.org/abs/2412.21187

83. 大语言模型幻觉综述

Huang 等（哈尔滨工业大学等），2023 · 综述 · 难度中 · 英文 · 对应第 6 章 · 国内：一般能打开，可能较慢

系统梳理了幻觉有哪些类型、从哪里来、怎么缓解，作者团队来自国内，是这个方向引用最多的综述之一。

arXiv：https://arxiv.org/abs/2311.05232

84. 提示词报告：提示工程技术的系统综述

Schulhoff 等，2024 · 综述 · 难度易到中 · 英文 · 对应第 7 章 · 国内：一般能打开，可能较慢

把几十种提示词技巧分门别类整理成一本「词典」，适合当工具书查。

arXiv：https://arxiv.org/abs/2406.06608

85. FrugalGPT：降低大模型使用成本、同时提升效果

Chen、Zaharia、Zou，2023 · 论文 · 难度易到中 · 英文 · 对应第 5、8 章 · 国内：一般能打开，可能较慢

提出「模型级联」：先让便宜模型试答，答不好再交给贵模型，是企业选型和省钱最实用的思路之一。

arXiv：https://arxiv.org/abs/2305.05176

86. LLMLingua：压缩提示词以加速推理

Jiang 等（微软亚洲研究院），2023 · 论文 · 难度中 · 英文 · 对应第 8 章 · 国内：一般能打开，可能较慢

证明提示词里很多 Token 删掉也不影响效果，直接对应书里「少说废话就是省钱」。

arXiv：https://arxiv.org/abs/2310.05736

87. 耗电的处理过程：瓦特在推高 AI 部署成本吗?

Luccioni 等，FAccT 2024 · 论文 · 难度易到中 · 英文 · 对应前言、第 9 章 · 国内：一般能打开，可能较慢

实测不同 AI 任务每次推理的耗电量，发现生成类任务耗能高得多，把 Token 和电费直接联系了起来。

arXiv：https://arxiv.org/abs/2311.16863

88. AI 有多饿？大模型推理的能耗、水耗与碳足迹

Jegham 等，2025 · 论文 · 难度易到中 · 英文 · 对应前言、第 9 章 · 国内：一般能打开，可能较慢

估算主流模型每次查询要耗多少电、多少水，推理模型和长回答明显更高，能更具体地理解「Token 是能源货币」。

arXiv：https://arxiv.org/abs/2505.09598

89. 在谷歌规模上测量 AI 服务的环境影响

Elsworth 等，Google，2025 · 论文 · 难度中 · 英文 · 对应前言、第 9 章 · 国内：一般能打开，可能较慢

大厂首次公开单次 Gemini 文本请求的能耗、碳排放和用水，并说明一年里效率提升了几十倍，是「Token 越来越便宜」背后的能源证据。

arXiv：https://arxiv.org/abs/2508.15734

六、行业报告与数据看板

想了解行业全局、价格走势和大家实际在用什么模型，看这些。

90. 人工智能产业发展研究报告（2025 年）

中国信息通信研究院，2026 年 1 月 · 研究报告（PDF） · 中文 · 对应前言、第 1、9 章 · 国内：能直接打开

国内权威机构对 2025 年 AI 产业的全景盘点，覆盖模型、智能体、智算基础设施和行业应用，适合企业决策者和投资者了解国内大局。

PDF：https://www.caict.ac.cn/kxyj/qwfb/bps/202602/P020260202487301304903.pdf

91. 人工智能算力基础设施赋能研究报告（2025 年）

中国信息通信研究院，2025 年 11 月 · 研究报告（PDF） · 中文 · 对应前言、第 9 章 · 国内：能直接打开

讲智算中心怎么建、利用率如何、怎样赋能产业，适合关注算力投资的读者。

PDF：https://www.caict.ac.cn/kxyj/qwfb/ztbg/202511/P020251106555844142999.pdf

92. 2026 年人工智能指数报告（AI Index 2026）

斯坦福大学 HAI · 年度报告，有官方中文版 · 中文 / 英文 · 对应第 1、5、9 章 · 国内：不确定

全球 AI 最权威的年度体检报告，有官方中文版可以直接读。

中文版 PDF：https://hai.stanford.edu/assets/files/hai-ai-index-2026-chinese-version-082226.pdf

报告主页：https://hai.stanford.edu/ai-index/2026-ai-index-report

93. 2025 年人工智能指数报告（AI Index 2025）

斯坦福大学 HAI · 年度报告，有官方中文版 · 中文 / 英文 · 对应第 4、9 章 · 国内：不确定

书中「GPT-3.5 水平的推理成本两年降了 280 多倍」就出自这一版。

中文版 PDF：

https://hai.stanford.edu/assets/files/hai_ai_index_report_2025_chinese_version_061325.pdf

报告主页：https://hai.stanford.edu/ai-index/2025-ai-index-report

94. 欢迎来到 LLMflation：大模型推理成本正在快速下降

Guido Appenzeller（a16z），2024 年 11 月 · 行业文章 · 英文 · 对应第 4、9 章 · 国内：不确定

提出「同等性能的大模型，成本每年降约 10 倍」，简单好记，可以和摩尔定律对照着理解。

原文：https://a16z.com/llmflation-llm-inference-cost/

95. 大模型推理价格下降很快，但各类任务降得不均衡

Epoch AI，2025 年 3 月 · 数据分析 · 英文 · 对应第 4、9 章 · 国内：不确定

按任务分别统计，达到同等水平的价格每年下降 9 倍到 900 倍不等，比「每年降 10 倍」的说法更细致。

原文：https://epoch.ai/data-insights/llm-inference-price-trends

96. 2025 年 AI 现状：100 万亿 Token 使用实证研究

OpenRouter 与 a16z，2025 年 12 月 · 行业报告 · 英文 · 对应第 4、5、9 章 · 国内：不确定

基于真实调用数据回答「全世界的 Token 都花在哪」：推理型模型处理了一半以上的 Token，编程用途占比从约一成升到一半以上，中国开源模型约占每周调用量的 13%。

报告：https://openrouter.ai/state-of-ai

97. Artificial Analysis

独立评测机构 · 数据看板 · 英文 · 对应第 5、9 章 · 国内：不确定

把各家模型的能力、价格、输出速度放在同一张图上比，选型时最常用的独立参考。

看板：https://artificialanalysis.ai/models

98. OpenRouter 模型排行

OpenRouter · 数据看板 · 英文 · 对应第 1、5 章 · 国内：不确定

展示开发者每周在各个模型上实际消耗的 Token 量，看「大家实际在用什么」。

看板：https://openrouter.ai/rankings

99. Arena 排行榜（原 LMArena）

Arena · 数据看板 · 英文 · 对应第 5 章 · 国内：不确定

用户盲测投票的排行榜，分文本、编程、图像、智能体等类别，看「大家觉得谁好用」。

看板：https://arena.ai/leaderboard